A Systematic Evaluation of Federated Learning on Biomedical Natural Language Processing

Le Peng
peng0347@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

Sicheng Zhou
zhou1281@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

Jiandong Chen
chen8111@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

Ziyue Xu
ziyuex@nvidia.com
Nvidia Corporation
Santa Clara, California, USA

Rui Zhang
zhan1386@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

Ju Sun
jusun@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

KEYWORDS

federated learning, biomedical natural language processing

ACM Reference Format:
Le Peng, Sicheng Zhou, Jiandong Chen, Ziyue Xu, Rui Zhang, and Ju Sun. 2023. A Systematic Evaluation of Federated Learning on Biomedical Natural Language Processing. *ACM/IMS J. Data Sci.* 37, 4, Article 111 (August 2023), 9 pages. https://doi.org/XXXXXXX.XXXXXXX

ABSTRACT

Language models (LMs) like BERT and GPT have revolutionized natural language processing (NLP). However, privacy-sensitive domains, particularly the medical field, face challenges to train LMs due to limited data access and privacy constraints imposed by regulations like the Health Insurance Portability and Accountability Act (HIPPA) and the General Data Protection Regulation (GDPR). Federated learning (FL) offers a decentralized solution that enables collaborative learning while ensuring the preservation of data privacy. In this study, we systematically evaluate FL in medicine across 2 biomedical NLP tasks using 6 LMs encompassing 8 corpora. Our results showed that: 1) FL models consistently outperform LMs trained on individual client's data and sometimes match the model trained with polled data; 2) With the fixed number of total data, LMs trained using FL with more clients exhibit inferior performance, but pre-trained transformer-based models exhibited greater resilience. 3) LMs trained using FL perform nearly on par with the model trained with pooled data when clients' data are IID distributed while exhibiting visible gaps with non-IID data. Our code is available at: https://github.com/PL97/FedNLP

Corresponding author: Rui Zhang, zhan1386@umn.edu; Ju Sun, jusun@umn.edu; University of Minnesota MN, USA, 55455.

2831-3194/2023/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The recent advances in deep learning have sparked the widespread adoption of language models (LMs), including prominent examples of BERT [1] and GPT [2], in the field of natural language processing (NLP). These LMs are trained on massive amounts of public text data, comprising billions of words, and have emerged as the dominant technology for various natural language processing tasks, including text classification [3, 4], text generation [5, 6], information extraction [7–9], and question answering [10], [11]. The success of LMs can be largely attributed to their ability to leverage large volumes of training data. However, in privacy-sensitive domains like medicine, data are often naturally distributed making it difficult to construct large corpora to train LMs. To tackle the challenge, the most common approach thus far has been to fine-tune pre-trained LMs for downstream tasks, using limited annotated data [12]. Nevertheless, pre-trained LMs are typically trained on text data collected from the general domain, which exhibits divergent patterns from that in the medical domain, resulting in a phenomenon known as domain shift. Biomedical texts are highly specialized, containing domain-specific terminologies and abbreviations [13]. For example, medical records and drug descriptions often include specific terms that may not be present in general language corpora, and the terms often vary among different clinical institutes. Also, biomedical data lacks uniformity and standardization across sources, making it challenging to develop NLP models that can effectively handle different formats and structures. Electronic Health Records (EHRs) from different healthcare institutions, for instance, can have varying templates and coding systems [14]. So, direct transfer learning from LMs pre-trained on the general domain can suffer a drop in performance and generalizability when applied to a different domain, as also demonstrated in the literature [15]. Therefore, developing LMs that are specifically designed for the medical domain, using large volumes of domain-specific training data, is essential. Another vein of research explores pre-training the LMs on corpora collected from the clinical domain, e.g., BlueBERT [12], and PubMedBERT [16]. These LMs were pre-trained on public biomedical dataset, e.g., PubMed literature data and Medical Information Mart for Intensive Care (MIMIC III) dataset. These LMs have shown superior performances compared to traditional deep learning methods (e.g., RNN-based methods) in various clinical NLP tasks, such as named entity recognition (NER), relation extraction (RE), and question answering [16–18]. Nonetheless, it is important

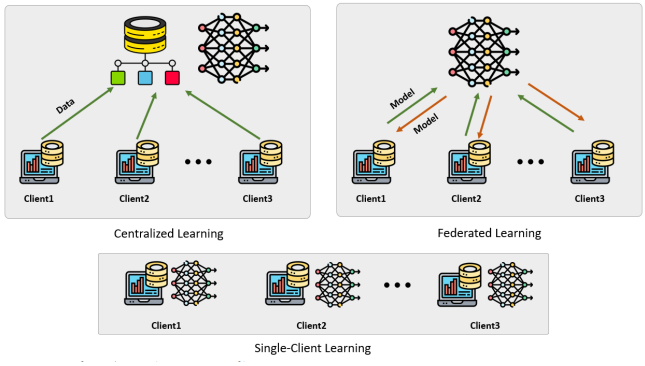

Figure 1: An overview of three learning schemes included in this study.

to highlight that the efficacy of these pre-trained medical LMs heavily relies on the availability of large volumes of task-relevant public data, which may not always be readily accessible.

All these mentioned above represent the classical centralized learning regime that involves aggregating data from distributed data sites and training a model in a single environment. However, this approach poses significant challenges in medicine, where data privacy is crucial, and data access is restricted due to regulatory concerns. Thus, in practice, people can only perform local training with their own dataset – *single-client training*. The drawback comes when the local dataset is small and often gives poor performance when evaluating an external dataset – poor generalization. To take advantage of the massively distributed data as well as improve the model generalizability, federated learning (FL) was initialized in 2017 [19] as a novel learning scheme to empower training with a decentralized environment and achieve many successes in critical domains with data privacy restrictions [20–22]. In an FL training loop, clients jointly train a shared global model by sharing the model weights or gradients while keeping their data stored locally. By bringing the model to the data, FL strictly ensures data privacy while still achieving competitive levels of accuracy and performance compared to a model trained with full data. Although FL has shown promising results in general vision and language tasks, its application in biomedical NLP is still limited. To close this gap, we conduct a comprehensive study of two representative NLP tasks to evaluate the feasibility of adopting FL with LM in biomedical NLP. Our study aims to contribute to a better understanding of the challenges and opportunities in FL for biomedical NLP and provide insights into the development of more robust and secure machine learning models for healthcare applications. Our major contribution includes:

- We confirm the effectiveness of FL in LM for biomedical NLP. In many cases, models trained using FL, specially pre-trained BERT-based models, can often match the performance of centralized learning, a significant boost compared with single-client learning.
- Larger model tends to be more resistant to the changes on FL scales. With a fixed number of data, the performance of FL models overall degrades as the clients' size increases. However, the effect diminished training on larger pre-trained models such as BERT-based models and GPT-2.
- FL closely approximates centralized training performance under IID data distribution, but a visible performance gap arises with non-IID data, varying across FL algorithms.

2 METHODS

Task The NER and RE are two established tasks for information extraction in the clinical NLP domain. Given an input sequence of tokens, the goal of NER is to identify and classify the named entities, such as diseases and genes, present in the text sequence. RE is often the follow-up task that aims for discovering the relations between pairs of named entities. For example, there is a gene-disease relation (*BRCA1-breast cancer*) that can be identified in the sentence *"Mutations of BRCA1 gene are associated with breast cancer"*. The NER and RE offer valuable insights for a range of applications, including retrieving information, constructing knowledge graphs, and answering questions. In our study, we consider RE as that: given a sentence and the position/span of the two named entities, the objective is to determine the relationship between the two named entities.

Corpora We compared federated learning with alternative training schemes on 8 biomedical NLP datasets on two NLP tasks: NER (5 corpora) and RE (3 corpora). For all NER corpora, it follows the same BIO notation to distinguish the beginning (B), inside (I), and outside (O) of entities. We adopted most of the preprocessed corpora from the paper of BioBERT [8], except for the 2018 n2c2 dataset (both NER and RE). For all the datasets, we remove duplicated notes and split the data into the train(80%), dev(10%), and test(10%). A summary of the dataset can be found in section 2 (detailed descriptions of each dataset can be found in appendix B.2).

Table 1: Copora used in this study. The data splits are counted based on the number of sentences.

Corpus	ENtity/Relation	Task	Train	Dev	Test
2018 n2c2 [23]	8 entities	NER	48727	6091	6091
BC2GM [24]	gene	NER	26006	3251	3251
BC4CHEMD [25]	drug/chem	NER	94170	11772	11771
JNLPBA [26]	gene	NER	29559	3695	3695
NCBI-disease [27]	disease	NER	10125	1266	1266
2018 n2c2 [23]	disease	RE	72786	9099	9098
EUADR [28]	gene-disease	RE	284	36	35
GAD [29]	gene-disease	RE	4097	531	512

Study design As shown in fig. 1, we explored three learning methods: 1) federated learning, centralized learning, and single-client learning. To simulate the conventional learning scenario, we varied the data scale and conducted the following experiments: centralizing all client data to train a single model (centralized learning) and training separate models on each individual client's data (single-client learning). These notations will be retained in the experimentation section.

Models To better understand the effect of network architecture on FL, we chose models with various sizes of parameters range from 20 M to 334 M, including transformer-based networks such as

Bidirectional Encoder Representations from Transformer(BERT) [1], and Generative Pre-trained Transformer (GPT), as well as classical recurrent networks BILSTM-CRF [30]. BERT-based models use a transformer encoder and incorporate bi-directional information acquired through two unsupervised tasks as a pre-training step into its encoder. Different BERT models differ in their pre-training source dataset and model size, deriving many variants such as BlueBERT [12], BioBERT [8], and Bio_ClinicBERT [31]. BILSTM-CRF [30] is the only model in our study that is not pretrained and also not built upon transformers. It is a bi-directional model designed to handle long-term dependencies, is used to be popular for NER, and uses LSTM as its backbone. We select this model in the interest of investigating the effect of federation learning on models with smaller sets of parameters.

Table 2: Models used in this study. The model is fine-tuned using pre-trained weights, if provided.

Model	Param(M)	pre-trained Source	year
BiLSTM-CRF [30]	20	-	2015
BERT [1]	109	Wiki+BooksCorpus	2018
BlueBERT [12]	334	PubMed	2019
BioBERT [8]	108	Wiki+BooksCorpus+PubMed+PMC	2020
ClinicBERT [31]	108	clinical notes	2019
GPT2	124	Wiki+news+books	2019

Training details Unless otherwise stated, our experiments adhere to a consistent set of training protocols: we set the maximum token as 512, whereby sentences exceeding this length will be trimmed, leaving a uniform length of input tokens. We considered two different learning settings: learning from independent and identically distributed (IID) data and learning from non-IID data. In the former case, we randomly split the data into k fold uniformly. For the majority of our experiments, k was chosen as 10, while we also varied the k from 2 to 10 to explore the impact of federation size. In the latter setting, we consider learning from heterogeneous data collected from different sources. This represents the real-world scenario where complex and entangled heterogeneities are co-existed. We pick the BC2GM and JNLPBA as two independent clients, both are targeting at indentifying gene entity but collected from different sources.

Reported evaluation For NER, we evaluate the model performance using F1-score at the macro average level with both strict and lenient match criteria. The strict match considers a pair as true positive when the boundary of entities exactly matches with the gold standard, while lenient match consider a pair as true positive whenever when there is an overlap between the boundaries of predicted entities and the gold standard. For both NER and RE, we repeat the experiment for three times and report the mean and standard deviation.

3 RESULTS AND DISCUSSION

FL yeilds single-client learning on NER and RE, and sometimes performs comparably with centralized training. We compare FL with centralized learning and single-client learning

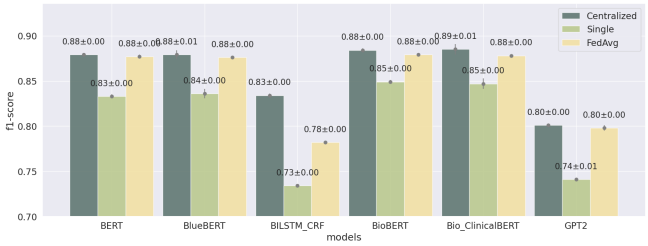

Figure 2: NER results on 2018 n2c2

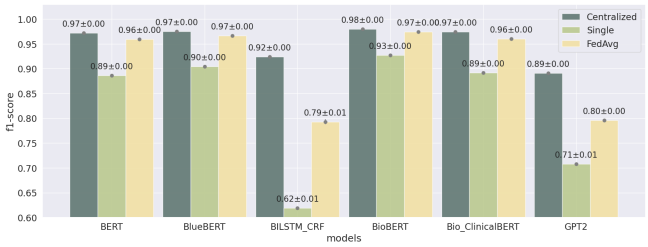

Figure 3: NER results on BC2GM

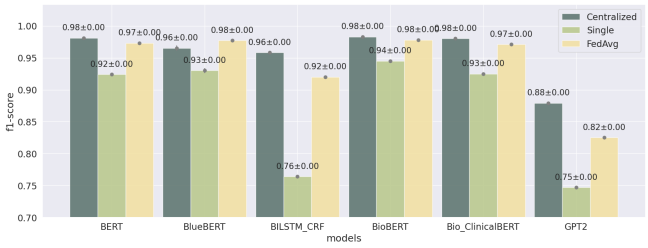

Figure 4: NER results on BC4CHEMD

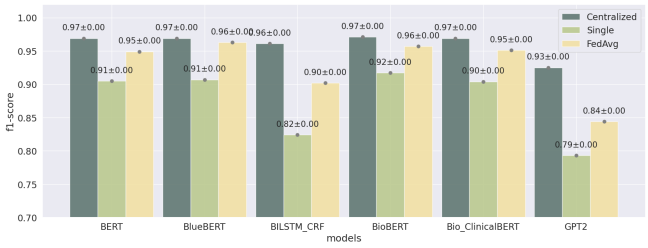

Figure 5: NER results on JNLPBA

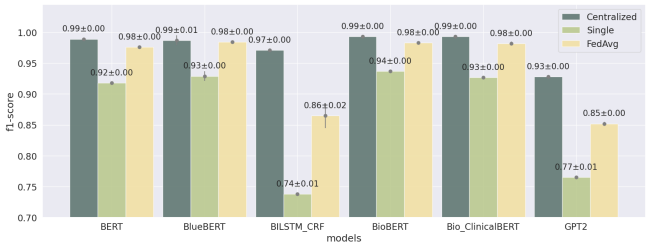

Figure 6: NER results on NCBI-disease

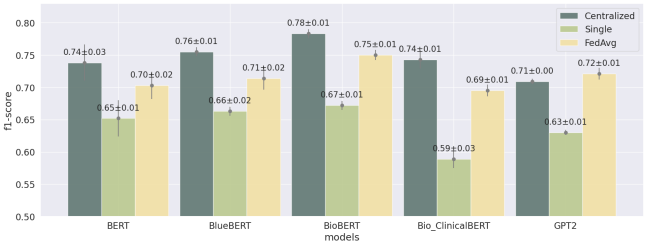

Figure 7: RE results on GAD

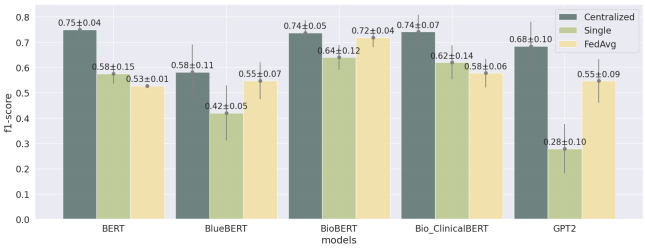

Figure 8: RE results on EUADR

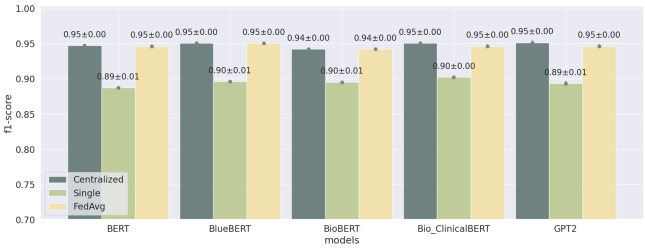

Figure 9: RE results on 2018 n2c2

using 6 models on NER (see fig. 2 ,fig. 3, fig. 4, fig. 5, and fig. 6), and RE (see fig. 7, fig. 8, and fig. 11). On both tasks, FL almost consistently outperforms single-client learning. The only exception is in fig. 8, where centralized learning yields FL when using BERT and bio_ClinicalBERT. We believe this is due to the lack of training data. As each client only owns 28 training sentences, the data distribution, although IID, is highly under-represented, making it hard for FedAvg to converge to the global minimum. Surprisingly, centralized learning achieves quite reasonable good performance even when the training data is limited (284 training sentences), confirming that transfer learning from either the general text domain (e.g., BERT and GPT2) or biomedical text domain (e.g., blueBERT, bioBERT, bio_ClinicalBERT) is beneficial to the downstream biomedical RE task. Another interesting finding is that GPT-2 model always gives inferior results compared to BERT-based models. We believe this is because GPT-2 is pre-trained on text generation tasks that only encode left-to-right attention for the next work prediction. However, this unidirectional nature prevents it from learning more about global context which limits its ability to capture dependencies between words in a sentence.

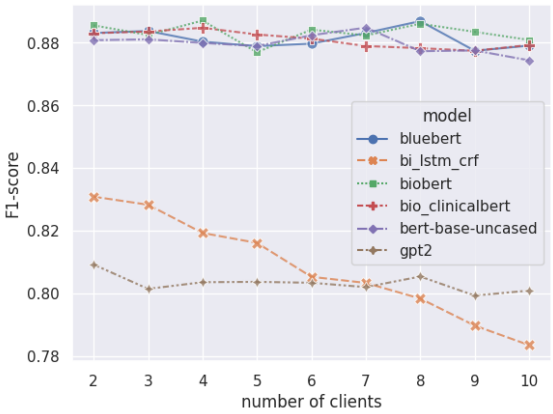

Figure 10: Performance of models with varied numbers of clients.

Larger model tends to be more resilient to the increasing scale of FL. In practice, there are two different learning scenarios. One is called the "cross-silo", where the number of participants is small but each site holds a large volume of data, while the other is called the "cross-device", where more participants are joining and each site only has a small amount of data. In the context of clinical settings, "cross-silo" typically corresponds to collaborations among major health institutions, where each site owns a large data repository. Conversely, "cross-device" scenarios may involve collaborations among smaller clinics or involve data from individualized patient devices. Here we are trying to investigate the performance of FL under the two different learning scenarios by varying the size of the clients while fixing the number of total training data. There are two interesting observations: 1) the classical BILSTM-CRF model (20M) tends to perform better in the "cross-silo" setting, but quickly degrade in performance as they enter the regime of "cross-device" region, as shown in fig. 10; 2) the transformer-based model ($\geq$ 108M), which is over 5 sizes larger, is more resilient to the change of federation scale.

FedAvg performs closely to FedProx under the non-IID setting and is almost non-distinguishable when data is IID. Many biomedical texts are very specialized as the different hospitals have unique protocols when generating medical text (i.e. sublanguage differences). Therefore, practical deployed FL systems should consider this data heterogeneity. We simulate a real non-IID scenario by emulating *BC2GM* and *JNLPBA* as two clients and jointly perform federated learning. We consider two FL algorithms including FedAvg and FedProx, both are widely deployed in practice. For comparison, we also study the performance of two algorithms under a simulated IID setting on *2018 n2c2* dataset. As summarized in table 3, we found that FedAvg performs closely with FedProx when data is non-IID distributed and is mostly undistinguishable when data is IID distributed. Although FedProx overall performs better than FedAvg, it is susceptible to the hyper-parameter μ which aims to balance the local objective function and the proximal term. To encapsulate, we see a gap between FL with centralized learning (pool) when data is non-IID distributed, and the gap closed when data from differentclients are IID distributed. Comparing across

Table 3: Comparison of FedAvg with centralized learning and Single client learning on 2018 n2c2 NER dataset using bioBERT.

method	μ	IID		non-IID	
		lenient	strict	lenient	strict
Pool	-	0.884±0.002	0.823±0.002	0.964±0.001	0.929±0.000
FedAvg	-	0.879±0.002	0.818±0.003	0.934±0.003	0.884±0.003
FedProx	1	0.855±0.003	0.790±0.005	0.880±0.001	0.772±0.002
	0.5	0.868±0.001	0.809±0.002	0.881±0.002	0.777±0.001
	0.1	0.872±0.003	0.814±0.004	0.897±0.002	0.817±0.002
	0.01	0.878±0.003	0.819±0.002	0.933±0.002	0.884±0.003
	0.001	0.880±0.002	0.820±0.001	0.944±0.002	0.901±0.002

two FL algorithms, FedProx gets better strict matching results than FedAvg, especially when data is non-IID distributed.

While seeing many promising results of FL for LMs, it is imperative to acknowledge several limitations inherent in our study: 1) we only study two FL algorithms under a non-IID setting, and we are aware that personalized FL algorithms may handle the distribution shift issues better. 2) we did not explore large language models (LLM). The concerns most come from the following aspect. First, many LLMs (e.g., GPT3 and GatorTron) are not open-source. Second, simulated federated learning in a single environment requires a lot of computation resources and a long training time. Third, for real-world deployment, LLM poses a big challenge in communication as FL typically requires frequent communication between server and clients. To address these limitations and further advance our understanding of FL for LMs, our future study will focus on the real-world implementation of FL and explore the practical opportunities and challenges. We believe our study will offer comprehensive insights into the potential of FL for LMs, which can serve as a catalyst for future research aimed at developing more effective AI systems by leveraging distributed clinical data in real-world scenarios.

4 ACKNOWLEDGEMENT

This work was in part supported by Cisco Research under award number 1085646 PO USA000EP390223. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper. URL: http://www.msi.umn.edu

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
[2] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
[3] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer, 2019.
[4] Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*, 2019.
[5] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.
[6] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
[7] Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019.
[8] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
[9] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.
[10] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*, 2019.
[11] Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1133–1136, 2019.
[12] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.
[13] Yonghui Wu, Jun Xu, Yaoyun Zhang, and Hua Xu. Clinical abbreviation disambiguation using neural word embeddings. In *Proceedings of BioNLP 15*, pages 171–176, 2015.
[14] Miriam Reisman. Ehrs: the challenge of making electronic data usable and interoperable. *Pharmacy and Therapeutics*, 42(9):572, 2017.
[15] Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, and Rui Zhang. Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association*, 29(7):1208–1216, 2022.
[16] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.
[17] Arpita Roy and Shimei Pan. Incorporating medical knowledge in bert for clinical relation extraction. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5357–5366, 2021.
[18] Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. *arXiv preprint arXiv:2010.03746*, 2020.
[19] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.
[20] Le Peng, Gaoxiang Luo, Andrew Walker, Zachary Zaiman, Emma K Jones, Hemant Gupta, Kristopher Kersten, John L Burns, Christopher A Harle, Tanja Magoc, et al. Evaluation of federated learning variations for covid-19 diagnosis using chest radiographs from 42 us and european hospitals. *Journal of the American Medical Informatics Association*, 30(1):54–63, 2023.
[21] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated Learning: Privacy and Incentive*, pages 240–254. Springer, 2020.
[22] Anh Nguyen, Tuong Do, Minh Tran, Binh X Nguyen, Chien Duong, Tu Phan, Erman Tjiputra, and Quang D Tran. Deep federated learning for autonomous driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1824–1830. IEEE, 2022.
[23] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Özlem Uzuner. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 2020.
[24] Larry Smith, Lorraine K Tanabe, Cheng-Ju Kuo, I Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, Manabu Torii, et al. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2):1–19, 2008.
[25] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17, 2015.
[26] Nigel Collier and Jin-Dong Kim. Introduction to the bio-entity recognition task at jnlpba. In *NLPBA/BioNLP*, 2004.
[27] Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.
[28] Erik M. van Mulligen, Annie Fourrier-Réglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifirò, Jan A. Kors, and Laura Inés Furlong. The eu-adr corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45 5:879–84, 2012.
[29] PubMed. The genetic association database, 2004. Accessed on Mar. 14, 2023.
[30] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991, 2015.

[31] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings. *ArXiv*, abs/1904.03323, 2019.
[32] Andrew Wen, Sunyang Fu, Sungrim Moon, Mohamed El Wazir, Andrew Rosenbaum, Vinod C Kaggal, Sijia Liu, Sunghwan Sohn, Hongfang Liu, and Jungwei Fan. Desiderata for delivering nlp to accelerate healthcare ai advancement and a mayo clinic nlp-as-a-service implementation. *NPJ digital medicine*, 2(1):130, 2019.
[33] Jianbo Lei, Buzhou Tang, Xueqin Lu, Kaihua Gao, Min Jiang, and Hua Xu. A comprehensive study of named entity recognition in chinese clinical text. *Journal of the American Medical Informatics Association*, 21(5):808–814, 2014.
[34] Kai Xu, Zhanfan Zhou, Tianyong Hao, and Wenyin Liu. A bidirectional lstm and conditional random fields approach to medical named entity recognition. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017*, pages 355–365. Springer, 2018.
[35] Sicheng Zhou, Nan Wang, Liwei Wang, Ju Sun, Anne Blaes, Hongfang Liu, and Rui Zhang. A cross-institutional evaluation on breast cancer phenotyping nlp algorithms on electronic health records. *arXiv preprint arXiv:2303.08448*, 2023.
[36] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194, 2022.
[37] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):1–12, 2020.
[38] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
[39] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.
[40] Yang Liu, Anbu Huang, Yun Luo, He Huang, Youzhi Liu, Yuanyuan Chen, Lican Feng, Tianjian Chen, Han Yu, and Qiang Yang. Fedvision: An online visual object detection platform powered by federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13172–13179, 2020.
[41] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.
[42] Yi Liu, JQ James, Jiawen Kang, Dusit Niyato, and Shuyu Zhang. Privacy-preserving traffic flow prediction: A federated learning approach. *IEEE Internet of Things Journal*, 7(8):7751–7763, 2020.
[43] Jiwei Huang, Zeyu Tong, and Zihan Feng. Geographical poi recommendation for internet of things: A federated learning approach using matrix factorization. *International Journal of Communication Systems*, page e5161, 2022.
[44] Latif U Khan, Shashi Raj Pandey, Nguyen H Tran, Walid Saad, Zhu Han, Minh NH Nguyen, and Choong Seon Hong. Federated learning for edge networks: Resource optimization and incentive mechanism. *IEEE Communications Magazine*, 58(10):88–93, 2020.
[45] Liu Yang, Ben Tan, Vincent W Zheng, Kai Chen, and Qiang Yang. Federated recommendation systems. *Federated Learning: Privacy and Incentive*, pages 225–239, 2020.
[46] Guanyu Lin, Feng Liang, Weike Pan, and Zhong Ming. Fedrec: Federated recommendation with explicit feedback. *IEEE Intelligent Systems*, 36(5):21–30, 2020.
[47] Yogesh Kumar and Ruchi Singla. Federated learning systems for healthcare: perspective and recent progress. *Federated Learning Systems: Towards Next-Generation AI*, pages 141–156, 2021.
[48] Ming Liu, Stella Ho, Mengqi Wang, Longxiang Gao, Yuan Jin, and He Zhang. Federated learning meets natural language processing: a survey. *arXiv preprint arXiv:2107.12603*, 2021.
[49] Lin BY, He C, Zeng Z, Wang H, Huang Y, Gupta R Dupuy C, Soltanolkotabi M, Ren X, and Avestimehr S. Fednlp: Benchmarking federated learning methods for natural language processing tasks. *arXiv preprint arXiv:2104.08815*, 2021.
[50] Dianbo Liu and Timothy Miller. Federated pretraining and fine tuning of bert using clinical notes from multiple silos. *ArXiv*, abs/2002.08562, 2020.
[51] Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuan tao Xie, and Weijian Sun. Feded: Federated learning via ensemble distillation for medical relation extraction. In *Conference on Empirical Methods in Natural Language Processing*, 2020.

A RELATED WORK

A.1 Natural language processing in medicine

NLP is playing an increasingly important role in the clinical domain. A significant portion of pertinent clinical information is hidden in unstructured data, NLP plays an important role in extracting target information that can aid in clinical decision-making, administrative reporting, and research [32] As two important clinical NLP tasks, NER and RE can help extract important information such as patient demographics, medical conditions, and treatments from these records, thereby facilitating efficient information extraction. Different NLP approaches have been developed in previous studies to solve the NER and RE tasks, evolving from rule-based systems to traditional machine learning methods to the latest LM fine-tuning methods. Relevant to this study, the conditional random fields (CRF) model that considers the likelihood of contextual interdependence among words is found to be an effective method for the NER task. Lei et al. developed a CRF model to extract clinical problems, procedures, laboratory tests, and medications from Chinese clinical texts, and the model achieved better performance than the baseline support vector machine model [33]. Xu, et al. proposed bidirectional long short-term memory and conditional random field (Bi-LSTM-CRF) model for medical NER tasks. The model contains three layers and relies on character-based word representations learned from the supervised corpus. It obtained a 0.802 F1 score on the NCBI Disease Corpus [34]. Zhou et al. developed the CancerBERT model using the BERT pre-training and fine-tuning pipeline and obtained an overall F1 score of 0.909 to extract eight types of cancer phenotypes from the clinical texts [15]. Though satisfied performances could be achieved, the cost of data annotation and computing resources are demanding, especially for the LM pre-training and fine-tuning methods. Roy et al. developed methods that add medical knowledge into pre-trained BERT models for clinical relation extraction tasks, and achieved state-of-the-art performance on the i2b2/VA 2010 clinical relation extraction dataset [Incorporating medical knowledge in BERT for clinical relation extraction]. Basically, two pre-training paradigms were applied for current LM pre-training, mixed-domain pre-training and domain-specific pre-training from scratch [16]. For mixed-domain pre-training, the LMs are first pre-training on the corpus of the general domain (e.g., Wikipedia), then keep pre-training on the domain-specific corpus like PubMed articles. It's assumed that general domain text is still helpful for domain-specific NLP tasks. ClinicalBERT and BlueBERT were developed under this scheme and obtained excellent performances for clinical NLP tasks. The domain-specific pretraining from scratch involves creating the vocabulary and conducting pre-training using only domain-specific texts. The PubMedBERT is representative of this scheme, and it achieves even better performance compared to the mixed-domain pre-training paradigm [16]. A recent study has shown that the LM pre-training and fine-tuning method has better generalizability than traditional machine learning and deep learning methods in terms of a clinical NER task. The performance of LM only drops 6.6% when directly evaluated on the test dataset from another clinical institute. However, the LM still needs to be further fine-tuned on the annotated data from the other clinical institute to achieve similar performances compared to LM that pre-trained and fine-tuned on local data [35]. Recently, Yang et al. developed a large LM (LLM) for EHRs, i.e., GatorTron [36]. The model was pre-trained on 90 billion words of clinical texts extracted from the EHRs. It increased the number of parameters of the clinical language model from 110 million to 8.9 billion and improved the performances of five clinical NLP tasks including NER and RE on benchmark datasets. The

GatorTron improved the generalizability of LM through pre-train on a larger corpus collected from regional clinical institutes, however, due to the data privacy restrictions, it's unfeasible to integrate corpora of different clinical institutes across the nation to pre-train the LMs. Federated learning (FL) allows data to remain decentralized, ensuring data privacy while enabling collaborative analysis across multiple healthcare institutions [37].

A.2 Federated Learning

The concept of FL was initialized by Google [38] and applied to the design of a virtual keyboard application named Gboard [39]. The typical FL framework represents a hub-spoke topology with clients acting as spokes that train the private data locally and the server acting as a hub that signals model dispatch and aggregation. Due to its mastery of privacy-preserved training with distributed data, FL has gained widespread adoption in various domains where data privacy is critical. In the domain of finance, for example, FL enables bankers to make use of distributed clients' data to help with bank openings [21], identify suspicious transactions [40], and detect financial fraud [41]. The applicability of FL has also been demonstrated in the domain of transportation, where FL can effectively leverage distributed sensory data to enhance autonomous driving [22], as well as improve traffic flow prediction using data from the government and companies [42]. Other domains including IoT [43]– [44] and recommendation systems [43], [45], [46] have also seen wide adoption of FL to utilize distributed data.

Similar to these domains, FL has gained increasing popularity in medicine. One major cause of the poor performance of machine learning models in the clinical domain is the insufficiency of data due to strict privacy and security regulations [47]. FL is promising to break through this bottleneck, it has already been implemented in many AI-aided system designs and achieved remarkable results. Peng et al. [20] collaborated with three universities in the U.S. to build an FL system for COVID-19 diagnosis and show that FL achieved boosted generalization performance compared with model training on local single institution data. Similarly, in a pancreas segmentation task, Wang et al. collect CT images from Taiwan and Japan and observed improved generalizability on model training with FL. More examples of federated learning in medical imaging applications can be found in this survey [cite]. While there is a rise of research showing great promise of applying FL in general NLP [48], [49], applications of FL in NLP for medicine are still underexplored. Existing work in FL on medical NLP is either focused on optimizing one task [50] [51] or trying to improve communication efficiency [51]. Current literature lacks a comprehensive comparison of FL on the biomedical NLP tasks across different datasets, especially based on transformed-based NLP models. Thus, our study aims to provide a comprehensive investigation of FL in medical NLP by studying several FL variants on multiple practical learning scenarios including varying sizes of the federation, different model architectures, and data heterogeneities on multiple benchmark datasets.

B DATASETS

B.1 Data Preprocessing

We adapted most of the dataset from the BioBERT paper with reasonable modifications by removing the duplicate entries and splited the data into the train(80%), dev(10%), and test(10%) datasets. To simulate a federated learning setting, we further split the train and dev set squally into 2/5/10 folds to mimic varied numbers of participation.

B.2 Details of the copura

2018 National NLP Clinical Challenges (n2c2) Shared Task [23]. 2018 n2c2 corpus contains 505 discharge summaries from the MIMIC-III clinical care database4. The goal of the task is to extract entity tags (reason, frequency, ADE, strength, duration, route, form, drug, and dosage) that indicate the presence of drug and ADE information, and relations (strength-drug, duration-drug, route-drug, form-drug, ADE-drug, Dosage-drug, reason-drug, and frequency-drug) between the entities.

BioCreative II Gene Mention Recognition (BC2GM) [24]. BC2GM Dataset collected text data related to gene information. The dataset comprises a set of sentences, and a set of gene mentions (GENE annotations) for each sentence. Some GENE annotations in a sentence may also have alternate boundaries that are judged by human annotators which can be essentially equivalent references (ALTGENE annotations). The goal of the task is to identify gene mentions in a sentence according to its start and end characters.

BioCreative IV Chemical Compound and Drug Name Recognition (BC4CHEMD) [25]. BC4CHEMD contains a total of 84,355 chemical mention annotations from 10,000 PubMed abstracts which are manually labeled by some chemistry literature experts. The goal of the task is to classify the text into multiple CEM classes: systematic, identifiers, formula, trivial, abbreviation, family, and multiple.

JNLPBA [26]. JNLPBA is orignate from the GENIA version 3.02. It is a selection of 2,000 abstracts with a controlled search on MEDLINE using the MeSH terms 'human', 'blood cells', and 'transcription factors'. The abstracts were hand-annotated to 36 terminal classes according to a small taxonomy of 48 classes based on a chemical classification.

NCBI-disease [27]. The NCBI-disease corpus is collected from 793 PubMed abstracts that are fully annotated at the disease mentions and concept level based on corresponding identifiers from either Medical Subject Headings (MeSH) or Online Mendelian Inheritance in Man (OMIM). It includes 6, 892 disease mentions, which are mapped to 790 unique disease concepts. 12% link to an OMIM identifier, while the remaining contain a MeSH identifier. In addition, 91% of mentions are described as a single disease concept, while the remaining link to a combination of concepts.

EUADR [28]. EUADR corpus was annotated for disorders, drugs, genes, and their inter-relationships. Three experts were used to annotate a set of 100 abstracts for each of the drug-disorder, drug-target, and target-disorder relations. The drug-disorder and drug-target relations were composed of 100 randomly selected abstracts from the PubMed result. For the target-disorder set, 50 abstracts were randomly selected from gene disorder, and 50 abstracts were randomly selected from SNP-disorder relation.

Table 5: Comparison of FedAvg with centralized learning and Single client learning on 2018 n2c2 NER dataset.

Model	methods	lenient	strict
BERT	Centralized	0.879±0.002	0.822±0.001
	Single	0.833±0.004	0.766±0.002
	FedAvg	0.877±0.002	0.817±0.002
BlueBERT	Centralized	0.879±0.005	0.820±0.007
	Single	0.836±0.004	0.767±0.005
	FedAvg	0.876±0.002	0.817±0.000
BILSTM-CRF	Centralized	0.834±0.002	0.783±0.002
	Single	0.734±0.001	0.667±0.006
	FedAvg	0.782±0.002	0.734±0.003
BioBERT	Centralized	0.884±0.002	0.823±0.002
	Single	0.849±0.004	0.784±0.003
	FedAvg	0.879±0.002	0.818±0.003
Bio_clincialBERT	Centralized	0.885±0.006	0.827±0.005
	Single	0.847±0.002	0.782±0.002
	FedAvg	0.878±0.001	0.815±0.001
GPT-2	Centralized	0.801±0.001	0.745±0.001
	Single	0.741±0.005	0.685±0.007
	FedAvg	0.798±0.003	0.746±0.001

Table 6: Comparison of FedAvg with centralized learning and Single client learning on BC4CHEMD NER dataset.

Model	methods	lenient	strict
BERT	Centralized	0.981±0.001	0.968±0.001
	Single	0.924±0.001	0.883±0.000
	FedAvg	0.973±0.000	0.954±0.001
BlueBERT	Centralized	0.965±0.004	0.944±0.007
	Single	0.930±0.001	0.895±0.003
	FedAvg	0.977±0.000	0.959±0.000
BILSTM-CRF	Centralized	0.958±0.001	0.934±0.001
	Single	0.764±0.002	0.669±0.007
	FedAvg	0.920±0.002	0.882±0.002
BioBERT	Centralized	0.983±0.001	0.972±0.001
	Single	0.945±0.001	0.913±0.001
	FedAvg	0.978±0.000	0.963±0.001
Bio_clincialBERT	Centralized	0.980±0.001	0.967±0.001
	Single	0.925±0.002	0.885±0.003
	FedAvg	0.971±0.001	0.953±0.001
GPT-2	Centralized	0.879±0.002	0.857±0.002
	Single	0.747±0.004	0.687±0.006
	FedAvg	0.825±0.000	0.794±0.000

Table 7: Comparison of FedAvg with centralized learning and Single client learning on JNLPBA NER dataset.

Model	methods	lenient	strict
BERT	Centralized	0.969±0.001	0.939±0.002
	Single	0.905±0.001	0.813±0.002
	FedAvg	0.949±0.001	0.896±0.001
BlueBERT	Centralized	0.969±0.001	0.940±0.003
	Single	0.907±0.001	0.817±0.003
	FedAvg	0.963±0.001	0.923±0.001
BILSTM-CRF	Centralized	0.961±0.000	0.924±0.001
	Single	0.824±0.003	0.669±0.010
	FedAvg	0.902±0.001	0.810±0.004
BioBERT	Centralized	0.971±0.000	0.943±0.001
	Single	0.917±0.001	0.828±0.002
	FedAvg	0.957±0.001	0.910±0.002
Bio_clincialBERT	Centralized	0.969±0.001	0.941±0.001
	Single	0.904±0.001	0.815±0.001
	FedAvg	0.951±0.000	0.901±0.001
GPT-2	Centralized	0.925±0.001	0.881±0.001
	Single	0.793±0.004	0.669±0.005
	FedAvg	0.844±0.001	0.748±0.001

Table 8: Comparison of FedAvg with centralized learning and Single client learning on NCBI-disease NER dataset.

Model	methods	lenient	strict
BERT	Centralized	0.989±0.001	0.973±0.000
	Single	0.918±0.003	0.842±0.003
	FedAvg	0.976±0.001	0.949±0.001
BlueBERT	Centralized	0.987±0.008	0.968±0.009
	Single	0.929±0.004	0.857±0.006
	FedAvg	0.984±0.002	0.963±0.000
BILSTM-CRF	Centralized	0.971±0.002	0.944±0.004
	Single	0.738±0.012	0.589±0.041
	FedAvg	0.865±0.020	0.767±0.035
BioBERT	Centralized	0.993±0.001	0.975±0.001
	Single	0.937±0.001	0.870±0.008
	FedAvg	0.983±0.002	0.958±0.001
Bio_clincialBERT	Centralized	0.993±0.001	0.975±0.001
	Single	0.927±0.001	0.854±0.008
	FedAvg	0.982±0.003	0.958±0.004
GPT-2	Centralized	0.928±0.002	0.904±0.002
	Single	0.765±0.012	0.684±0.013
	FedAvg	0.852±0.003	0.809±0.002

Table 9: Comparison of FedAvg with centralized learning and single client learning on relation extraction tasks.

Model	methods	2018 n2c2	EUADR	GAD
BERT	Centralized	0.947±0.001	0.750±0.040	0.738±0.028
	Single	0.887±0.008	0.576±0.154	0.652±0.010
	FedAvg	0.946±0.002	0.527±0.008	0.703±0.021
BlueBERT	Centralized	0.950±0.002	0.582±0.109	0.755±0.007
	Single	0.896±0.010	0.420±0.048	0.663±0.021
	FedAvg	0.950±0.002	0.548±0.073	0.714±0.018
BILSTM-CRF	Centralized	0.942±0.002	0.737±0.049	0.783±0.007
	Single	0.895±0.009	0.640±0.119	0.672±0.015
	FedAvg	0.942±0.002	0.718±0.037	0.750±0.008
BioBERT	Centralized	0.950±0.001	0.741±0.067	0.743±0.014
	Single	0.902±0.004	0.620±0.138	0.589±0.034
	FedAvg	0.946±0.003	0.578±0.057	0.695±0.009
Bio_clincialBERT	Centralized	0.951±0.004	0.684±0.097	0.709±0.004
	Single	0.893±0.013	0.279±0.104	0.630±0.008
	FedAvg	0.946±0.003	0.547±0.086	0.721±0.009

genetic association studies to explore gene-disease relations. It contains 10, 697 genes, 12, 774 diseases, and 74, 928 gene-disease associations.

C SUPPLEMENT RESULTS

table 4, table 5, table 6, table 7, table 8 summarizes the comparison of centralized learning, single training, and FL on NER measured by F1-score using both lenient and strict match, and results for RE can be found in table 9.

As *2018 n2c2* contains multiple concept types with varied sizes, we report a detailed performance measure for both FedAvg and FedProx.

Table 4: Comparison of FedAvg with centralized learning and Single client learning on BC2GM NER dataset.

Model	methods	lenient	strict
BERT	Centralized	0.879±0.002	0.822±0.001
	Single	0.886±0.001	0.755±0.00
	FedAvg	0.959±0.001	0.897±0.000
BlueBERT	Centralized	0.975±0.000	0.932±0.002
	Single	0.904±0.003	0.775±0.003
	FedAvg	0.966±0.001	0.919±0.002
BILSTM-CRF	Centralized	0.924±0.001	0.866±0.001
	Single	0.619±0.005	0.409±0.014
	FedAvg	0.793±0.005	0.645±0.013
BioBERT	Centralized	0.980±0.000	0.937±0.003
	Single	0.927±0.001	0.808±0.001
	FedAvg	0.974±0.001	0.922±0.000
Bio_clincialBERT	Centralized	0.974±0.001	0.933±0.001
	Single	0.892±0.001	0.765±0.004
	FedAvg	0.960±0.002	0.901±0.001
GPT-2	Centralized	0.891±0.001	0.836±0.001
	Single	0.708±0.010	0.549±0.011
	FedAvg	0.796±0.001	0.674±0.006

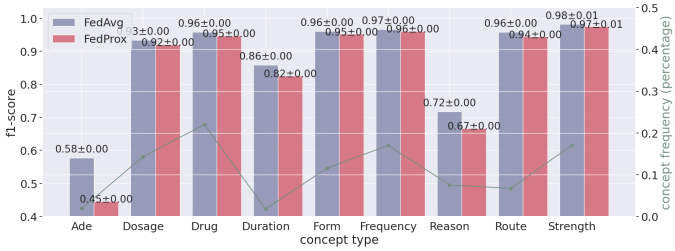

Figure 11: A comparison between FedAvg and FedProx on 2018 n2c2 NER task