Fine-Tuning Foundational AI Models to Code Diagnoses from Veterinary Health Records

Adam Kiehl[2], Mayla R. Boguslav[1,5], David Kott[1], G. Joseph Strecker[2], Nadia Saklou[3], Tracy Webb[3], Terri Ward[4], Michael Kirby[1]

[1] Colorado State University (CSU) Data Science Research Institute
[2] CSU College of Veterinary Medicine and Biomedical Sciences (CVMBS) Research IT
[3] CSU CVMBS
[4] CSU Veterinary Teaching Hospital
[5] Southern California Clinical and Translational Science Institute

Background

One Health data linkages, combining human, veterinary, and environmental data create an enhanced data resource that can be used to study the characteristics of diseases such as obesity[1] and diabetes[3] at a household level. The OMOP Common Data Model (CDM), as actively demonstrated by the OHDSI Electronic Animal Health Records working group, is compatible with veterinary clinical data and thus serves as a potential mechanism for facilitating One Health linkages. In collaboration with the University of Colorado Anschutz Medical Campus (AMC), Colorado State University (CSU) is engaged in efforts to establish One Health data sharing.

Clinical diagnoses at CSU are recorded by clinicians in free-text medical summary documents that are not easily codable automatically by traditional rules-based means the way data from other clinical domains (e.g., drug exposure) often are. Artificial intelligence-based natural language processing (NLP) can facilitate automatic coding of diagnoses at scale when human coding resources are scarce. Previous research into automated clinical text coding produced models such as DeepTag[9] (2018) and VetTag[15] (2019) that were trained on a hand-labeled set of records from CSU's legacy electronic health records (EHRs) and could apply a corpus of 42 and 4,522 SNOMED-CT[10] diagnosis codes, respectively, to veterinary medical summaries with impressive accuracy. Since VetTag's[15] creation, the number of publicly available pre-trained foundational clinical large language models (LLMs) has burgeoned[2,4,5,6,7,8,11,12,13,14], thus providing a variety of new possibilities for improving AI-based clinical coding.

Our work aims to expand upon previous research in the field of automated clinical coding by leveraging CSU's extensive set of labeled veterinary records to fine-tune an assortment of publicly available foundational LLMs on the downstream task of diagnosis coding. Given a free-text veterinary clinical visit summary, a fine-tuned LLM is tasked to determine a set of relevant SNOMED-CT[10] diagnosis codes that can be appropriately (as compared with an expert manual coding) applied to code the record.

Methods

CSU's fine-tuning dataset is derived from 246,473 labeled records from the CSU Veterinary Teaching Hospital (VTH) legacy EHR system. At the CSU VTH, medical summaries are the main clinical document, containing much of the relevant information about a visit (e.g., presenting complaint, history, assessment, physical exam, etc.). For this study, only the *diagnosis* and *assessment* sections were used, as they were identified as the most clinically relevant sections to the task of diagnosis coding. The medical summaries used were manually coded by the VTH Medical Records team resulting in 7,739 distinct SNOMED-CT[10] diagnosis codes spanning visits from 2012 to 2019. Data were split into training, validation, and test splits.

Several human foundational models (GatorTron[14] – 2022, MedicalAI ClinicalBERT[12] – 2023, and medAlpaca[6] – 2023), veterinary foundational models (VetBERT[7] – 2020 and PetBERT[5] – 2023), and non-clinical models (BERT[4] – 2018, RoBERTa[8] – 2019, and GPT-2[11] – 2019) were selected for fine-tuning (see Table 1). All models were fine-tuned under a consistent training framework that included a batch size of 32, a binary cross-entropy loss function, an AdamW optimizer, scheduled optimization with 5,000 warmup steps and a plateau learning rate of $3*10^{-5}$ (constant learning rate after warmup), and an early stopping criterion with five epoch patience (number of epochs without improved results required to trigger an early stop). All models' pooler layers were allowed to update, and dropout (rate of 0.25) and classifier layers were appended to each. Fine-tuning was performed on CSU's Riviera high performance computing cluster featuring GPU nodes containing four Tesla A100 GPUs, each with 80 gigabytes of RAM.

Model	Source	Pre-Training Data	Parameters
GatorTron[14]	Univ. of Florida, NVIDIA	University of Florida Health (2.9M notes), MIMIC-III, PubMed, Wikipedia (91B words)	3.9B
MedicalAI ClinicalBERT[12]	Fudan Univ., Beijing Univ. of Posts and Telecommunications	Zhongshan Hospital, Qingpu Hospital (1.2B words)	135M
medAlpaca[6]	Univ. of California, Berkeley, Technical Univ. Munich, Berlin Univ. of Applied Sciences & Technology, Aachen Univ.	Anki flashcards, Stack Exchange, Wikidoc, other Q&A	6.6B
VetBERT[7]	Asia-Pacific Centre for Animal Health, Univ. of Melbourne	VetCompass (15M notes, 1.3B tokens)	108M
PetBERT[5]	Durham Univ., Lancaster Univ., Univ. of Liverpool	United Kingdom veterinary EHRs (5.1M notes, 500M words)	108M
BERT Base[4]	Google AI	BooksCorpus (800M words), Wikipedia (2.5B words)	108M
BERT Large[4]	Google AI	Same as BERT	335M
RoBERTa[8]	Facebook AI	Same as BERT with CC-News (63M articles), OpenWebText, Stories	125M
GPT-2 Small[11]	OpenAI	WebText (8M documents)	125M
GPT-2 XL[11]	OpenAI	Same as GPT-2	1.6B
*VetTag[15]**	*Stanford Univ., Tsinghua Univ., Colorado State Univ.*	*Private specialty veterinary group (1M notes)*	*42M*

Table 1. Foundational clinical LLMs, author institutions, and sources of pre-training data.

Model performance on this multi-label classification problem was gauged using weighted macro (class-wise) averages of F1, precision, and recall, as well as record-wise exact match rate. We compared all fine-tuned models to the current state-of-the-art model in veterinary clinical coding, VetTag[15].

Results

We achieved state-of-the-art results on the task of veterinary clinical coding to the largest set of diagnoses (see Table 2). GatorTron[14] fine-tuned for this task achieved the best results with an average weighted F1 score of 74.9 and an exact match rate of 51.6% on a held-out test set (10% of the labeled CSU VTH data). Among tested models, GatorTron[14] was second largest and was pre-trained on the largest volume of clinical text. However, smaller models[4,5,7,8,11,12] and models trained on less or no clinical text performed comparably (e.g., medAlpaca[6] was not trained on any clinical text). All models demonstrated vastly improved performance when fine-tuned as compared to when used out-of-the-box (these results are not shown). Overall, the precision of all models was found to be higher than recall. Finally, it was shown that strong model performance can be achieved using smaller amounts of fine-tuning data than were available for this study (see Figure 1).

Model	F1	Precision	Recall	Exact Match	Fine-Tune Time
GatorTron[14]	**74.9** ± 0.29	**80.8** ± 1.14	**71.8** ± 1.00	**51.6** ± 0.29	21.9 ± 2.87 hrs
MedicalAI ClinicalBERT[12]	68.7 ± 0.57	78.0 ± 1.15	63.9 ± 0.29	45.2 ± 0.86	1.5 ± 0.14 hrs
medAlpaca[6]	67.1 ± 0.90	79.1 ± 0.66	61.7 ± 1.00	41.6 ± 1.46	14.1 ± 0.14 hrs
VetBERT[7]	69.5 ± 0.87	78.7 ± 0.86	64.7 ± 0.90	46.5 ± 0.66	2.9 ± 0.38 hrs
PetBERT[5]	69.4 ± 2.00	77.6 ± 1.88	65.3 ± 2.17	46.4 ± 1.41	2.7 ± 0.50 hrs
BERT Base[4]	68.5 ± 1.93	77.5 ± 0.57	63.9 ± 3.19	45.8 ± 1.27	3.6 ± 0.57 hrs
BERT Large[4]	70.4 ± 0.14	78.4 ± 0.29	66.2 ± 0.38	47.2 ± 0.57	7.4 ± 1.15 hrs
RoBERTa[8]	67.7 ± 4.88	76.5 ± 3.02	63.3 ± 5.41	44.6 ± 4.81	2.6 ± 0.87 hrs
GPT-2 Small[11]	68.3 ± 1.51	78.3 ± 1.17	63.4 ± 1.74	44.6 ± 1.08	4.1 ± 0.87hrs
GPT-2 XL[11]	71.7 ± 1.62	80.4 ± 0.76	67.2 ± 2.01	47.8 ± 1.74	15.2 ± 1.58 hrs
*VetTag[15]**	*66.2*	*72.1*	*63.1*	*26.2*	*Unknown*

Table 2. Fine-tuning results computed on three test sets to allow for the generation of 95% confidence intervals. *: Top results from the VetTag[15] paper using 4,577 SNOMED-CT diagnosis codes with hierarchy-enriched target sets.

Conclusion

CSU presents an enhanced automated clinical diagnosis coding tool that utilizes a larger training dataset and an expanded set of SNOMED-CT[10] diagnosis codes, relative to previous studies[9,15]. We demonstrate that the use foundational LLMs for clinical coding tasks can circumvent the need for massive pre-training datasets and further illustrate that even limited amounts of labeled fine-tuning data can be leveraged to achieve strong model performance. This can serve as a guide for other institutions, with either extensive or limited data and computational resources, who are interested in performing similar coding tasks to inform their OMOP CDM instances.

The results of this study not only contribute to the improvement of the quality of veterinary electronic health records by offering more accessible methods for automated coding but also advance both human and animal health research by paving the way for more integrated and comprehensive health databases that span species and institutions.

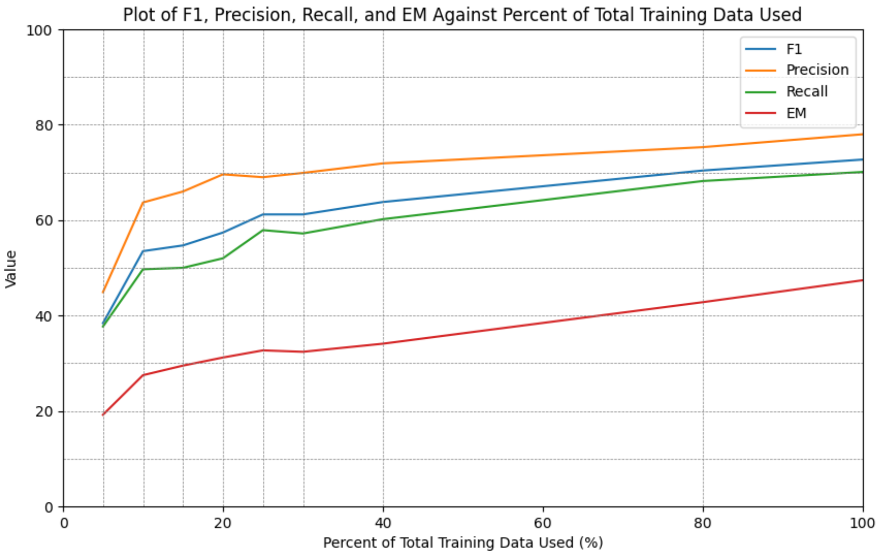

Figure 1. GatorTron[14] results when fine-tuning was performed using only fractions of the entire available training set.

References

1. Chandler, M., Cunningham, S., Lund, E., Khanna, C., Naramore, R., Patel, A., & Day, M. (2017). Obesity and associated comorbidities in people and companion animals: A one health perspective. Journal of comparative pathology, 156 (4), 296–309.
2. Chen, Z., Hernandez-Cano, A., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., K¨opf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M.-A., Jaggi, M., & Bosselut, A. (2023). Meditron- 70b: Scaling medical pretraining for large language models.
3. Delicano, R. A., Hammar, U., Egenvall, A., Westgarth, C., Mubanga, M., Byberg, L., Fall, T., & Kennedy, B. (2020). The shared risk of diabetes between dog and cat owners and their pets: Register based cohort study. bmj, 371.
4. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.
5. Farrell, S., Appleton, C., Noble, P.-J. M., & Al Moubayed, N. (2023). Petbert: Automated icd-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records. Scientific Reports, 13 (1), 18015.
6. Han, T., Adams, L., Papaioannou, J.M., Grundmann, P., Oberhauser, T., Loser, A., Truhn, D., & Bressem, K. (2023). MedAlpaca–an open-source collection of medical conversational ai models and training data*. arXiv preprint arXiv:2304.08247*.
7. Hur, B., Baldwin, T., Verspoor, K., Hardefeldt, L., & Gilkerson, J. (2020). Domain adaptation and instance selection for disease syndrome classification over veterinary clinical notes. Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, 156–166.
8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: a robustly optimized bert training approach. *arXiv preprint arXiv:1907.11692.*
9. Nie, A., Zehnder, A., Page, R. L., Zhang, Y., Pineda, A. L., Rivas, M. A., Bustamante, C. D., & Zou, J.

(2018). Deeptag: Inferring diagnoses from veterinary clinical notes. NPJ digital medicine, 1 (1), 60.

10. of Medicine, N. L. (2024). Snomed ct united states edition.
11. Radford, A, Wu, J, Child, R, Luan, D, Amodei, D, Sutskever, I, et al. (2019). Language models are unsupervised multitask learners. OpenAI blog 2019; 1(8):9.
12. Wang, G., Liu, X., Ying, Z., Yang, G., Chen, Z., Liu, Z., Zhang, M., Yan, H., Lu, Y., Gao, Y., et al. (2023). Optimized glycemic control of type 2 diabetes with reinforcement learning: A proof-of-concept trial. Nature Medicine, 29 (10), 2633–2642.
13. Xie, Q., Chen, Q., Chen, A., Peng, C., Hu, Y., Lin, F., Peng, X., Huang, J., Zhang, J., Keloth, V., He, H., Ohno-Machido, L., Wu, Y., Xu, H., & Bian, J. (2024). Me llama: Foundation large language models for medical applications.
14. Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). A large language model for electronic health records. npj Digital Medicine, 5 (1), 194.
15. Zhang, Y., Nie, A., Zehnder, A., Page, R. L., & Zou, J. (2019). Vettag: Improving automated veterinary diagnosis coding via large-scale language modeling. NPJ digital medicine, 2 (1), 35.