7th Healthcare Text Analytics Conference

HealTAC2024

Book of Abstracts

Healtex

frontiers

Data Science Institute | Lancaster University

Submissions

An Assessment on Comprehending Mental Health through Large Language Models

Mihael Arcan, David-Paul Niland, Fionn Delahunty

Lua Health, Galway, Ireland

Introduction

Progress in large language models (LLMs) has expanded their applications, yet there is a significant gap in understanding when it comes to their potential in mental health. Mental health issues affect a substantial portion of the global population, with conditions like depression and anxiety leading to significant economic losses. Previous research [1] has focused on specialised machine learning models for tasks like stress detection and depression prediction, often requiring fine-tuning or limited to predefined tasks.

Methods and Data

Within our work, we leveraged the DAIC-WOZ [1] dataset, which comprises clinical interviews based on the PHQ-4 questionnaire aimed at assessing psychological distress for depression and anxiety. The used dataset contained 28,186 training and 8,710 test observations after removing duplicates. To assess the capabilities of LLMs regarding mental health issues, we employed prompts to elicit specific responses from LLMs, with variations in lexical outputs and prompt length tailored to the PHQ-4 questionnaire. For this we employed the open source Llama2 and ChatGPT, a commercially available LLM, which was further fine-tuned on the DAIC-WOZ dataset for multiclass and binary classification tasks based on PHQ-4 scores.

Results

In our analysis (Table 1), we consolidate the optimal strategies derived from employing Llama-2's and ChatGPT's prompting and leveraging the Distil-RoBERTa transformer model, which exhibited the best performance across anxiety (GAD) and depression (PHQ) inquiries. Additionally, we gauge the predictive capabilities of XGBoost. Upon comparing Llama-2 with ChatGPT, minor advantages are discerned in favour of ChatGPT, particularly evident in the performance on GAD-2 and both PHQ questions. Lastly, when contrasting the Distil-RoBERTa transformer model with Llama-2 and ChatGPT, our investigation reveals that the former outperforms all targeted models across all GAD and PHQ inquiries concerning weighted precision, recall, and F1 scores.

Table 1. Classification results for XGBoost, Llama-2, ChatGPT and Distil-Roberta on separated questions in the PHQ-4 questionnaire (bold scores represent best result).

	GAD-1						GAD-2					
	Prec	Rec	F1	Spec	HammL	AUC-ROC	Prec	Rec	F1	Spec	HammL	AUC-ROC
XGBoost	0.45	0.55	0.48	0.64	**0.45**	0.56	0.63	0.69	0.60	0.34	**0.31**	0.51
Distil-RoBERTa	**0.57**	**0.58**	**0.56**	0.42	0.63	**0.70**	**0.69**	**0.70**	**0.68**	0.30	0.62	0.52
Llama-2 (v3)	0.38	0.33	0.33	0.68	0.67	0.52	0.56	0.23	0.27	**0.78**	0.77	0.50
ChatGPT 3.5	0.36	0.29	0.31	**0.71**	0.71	0.51	0.54	0.37	0.44	0.62	0.63	**0.53**
	PHQ-1						PHQ-2					
	Prec	Rec	F1	Spec	HammL	AUC-ROC	Prec	Rec	F1	Spec	HammL	AUC-ROC
XGBoost	0.51	0.54	0.48	0.61	**0.46**	0.55	0.52	0.53	0.48	0.66	**0.47**	0.56
Distil-RoBERTa	**0.55**	**0.55**	**0.53**	0.45	0.61	**0.71**	**0.55**	**0.57**	**0.54**	0.43	0.62	**0.73**
Llama-2 (v3)	0.43	0.29	0.30	**0.76**	0.71	0.53	0.34	0.34	0.32	0.65	0.66	0.49
ChatGPT 3.5	0.38	0.39	0.39	0.64	0.61	0.52	0.37	0.31	0.33	**0.71**	0.69	0.50

Table 2 illustrates a comparison for Distil-RoBERTa, ChatGPT 3.5, and ChatGPT 3.5 fine-tuned models. For the binary PHQ-4 classification, Mental-BERT outperforms others in precision (0.81), recall (0.84), and F1 score (0.81), with ChatGPT 3.5 fine-tuned achieving the best AUC-ROC score (0.63). Conversely, in the multiclass PHQ-4 classification, Distil-RoBERTa again exhibits best performance, achieving the highest precision (0.58), recall (0.59), F1 score (0.58), specificity (0.72), and AUC-ROC (0.63).

Table 2. Classification results for Distil-Roberta, ChatGPT and fine-tuned ChatGPT on the PHQ-4 questionnaire (bold scores represent best result for each metric).

	Binary PHQ-4					
	Prec	Rec	F1	Spec	HammD	AUC-ROC
Distil-RoBERTa	**0.82**	**0.85**	**0.82**	0.32	**0.15**	0.58
ChatGPT 3.5	0.78	0.48	0.55	**0.67**	0.52	0.57
ChatGPT 3.5 fine-tuned	0.80	0.77	0.78	0.49	0.23	**0.63**
	Multiclass PHQ-4					
	Prec	Rec	F1	Spec	HammL	AUC-ROC
Distil-RoBERTa	**0.57**	**0.59**	**0.57**	**0.69**	**0.41**	**0.61**
ChatGPT 3.5	0.39	0.39	0.39	0.62	0.61	0.51
ChatGPT 3.5 fine-tuned	0.40	0.45	0.38	0.58	0.55	0.51

Conclusion

In conclusion, mental ill-health challenges globally impact a significant portion of the population, highlighting the pressing need for effective interventions. Despite the advancements of large language models in various NLP tasks and their diverse applications, a substantial research gap persists regarding their understanding and optimisation within the realm of mental health. This study addresses this gap by conducting an initial evaluation of large language models, comparing the performance of Llama-2 and ChatGPT with further machine learning models. Leveraging Llama-2 and ChatGPT, we explore different prompting strategies and evaluate their effectiveness on the PHQ-4 questionnaire. The results indicate that transformer-based models, such as Distil-RoBERTa, consistently outperform Llama-2 and ChatGPT for all GAD and PHQ questions in terms of weighted precision, recall, and F1 score. These findings contribute valuable insights for the future development and application of language models in addressing mental health concerns. Nevertheless, with the outcomes of our study, we will further study large language models and their potential to be applied to mental health challenges. Due to the sensitive nature of mental health information, we will further analyse biases in training data and the dynamic nature of mental health that are the biggest hurdles in achieving comprehensive predictive model performance.

Study context

In this study, we analysed the DAIC-WOZ dataset to investigate mental health issues. It's important to note that while this dataset offers valuable insights, its limitations may affect the generalisability of our findings.

References

1. Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2023. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. arXiv:2307.14385 [cs.CL]
2. Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In Proceedings of the Ninth International Conference on Language Resources and Evaluation.

Automatic TNM staging classification for [^{18}F] fluorodeoxyglucose PET-CT reports for lung cancer utilising natural language processing and multi-task learning.

Stephen H. Barlow[1], Sugama Chicklore[1,2], Yulan He[3,4,5], Thomas Wagner[6], Anna Barnes[1,7*], Gary J.R. Cook[1,2*]

[1] School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK
[2]King's College London and Guy's and St. Thomas' PET Centre, St. Thomas' Hospital, London, UK
[3]Department of Informatics, King's College London, London, UK
[4]Department of Computer Science, University of Warwick, Coventry, UK
[5]Alan Turing Institute, London, UK
[6]Department of Nuclear Medicine, Royal Free Hospital, London, UK
[7] King's Technology Evaluation Centre (KiTEC), School of Biomedical Engineering & Imaging Science, King's College London, London, UK
*Equal contribution as senior authors

Introduction

Fluorodeoxyglucose positron emission tomography - computed tomography ([^{18}F]FDG PET-CT) imaging is an important component in detecting and staging lung cancer [1]. The findings of these studies are recorded in free-text reports and contain specialist terminology and language [2] that make it challenging to extract clinical information at scale. Tumour, node, metastasis (TNM) staging is widely used for guiding clinical decisions in lung cancer cases, so determining how PET-CT report findings relate to this is essential for effective treatment. Natural language processing (NLP) techniques specialise in retrieving structured information from unstructured text and are used in a variety of healthcare applications, including radiology. There has been less work extracting lung cancer staging information from PET-CT. Park et al. [3] extracted anatomic metastatic information from lung cancer reports and Nobel et al. [4] used a rule-based algorithm to extract tumour and node staging information. Neither provided extensive external validation, potentially due to the difficulties in satisfying data protection requirements. There is an opportunity to utilise transformer-based pretrained language models (PLMs) for this task and explore how their use generalises to reports from another hospital with a different reporting style. Accordingly, this study develops a transformer-based multi-task TNMu (Tumour, Node, Metastasis, uncertainty) classifier for FDG PET-CT lung cancer reports and evaluates it on internal and external data. We believe this model can be used in creating research cohorts, developing clinical alert/decision support systems, and assisting audit processes.

Methods and Data

An internal dataset of 2498 PET-CT reports for suspected or confirmed lung cancer were extracted from King's College London and Guy's and St. Thomas' PET Centre and split at a ratio of 80/10/10 (at patient level) into train/evaluation/test datasets. These were annotated by a nuclear medicine physician with 30 years of PET experience. An external test dataset of 463 reports from the Royal Free hospital was separately annotated by the internal annotator and an additional nuclear medicine physician with 14 years PET experience, with any initial label disagreements being resolved by consensus. The annotators labelled each report at document level for the presence/absence of any finding(s) which would constitute 'T', 'N' or 'M' findings as defined by *The Eighth Edition of TNM Staging for Lung Cancer* [5]. An additional binary task for 'u' (uncertainty) was also annotated and defined as positive if any of the TNM findings have ambiguity or uncertainty associated with them, otherwise it is negative. Each PET-CT accordingly has four document-level binary labels for 'T', 'N', 'M', and 'u' tasks.

We framed the problem as multi-label classification and explored the benefits of multi-task learning by using a shared PLM encoder for all tasks as opposed to separate binary classifiers. GatorTron [6] was determined to be the best performing PLM on the validation set and thus used in the final models. GatorTron was fine-tuned for five epochs with a four-neuron classification layer for the multi-task approach, and single-neuron heads for the single task classifiers. The multi-task approach was compared against the single-task classifiers, and a traditional machine learning model utilising TF-IDF (term frequency-inverse document frequency) embeddings and logistic regression classifiers for each binary task. Each approach was trained three times with three random seeds and the mean and standard deviation of the three runs is reported. The evaluation metrics used were accuracy over all four classes, and over the TNM classes (ACC_{TNMu}, $ACC_{TNM)}$, Hamming loss over TNMu classes (HL_{TNMu}) and macro average F1 scores for 'T', 'N', 'M' and 'u'.

Results

Table 1 demonstrates that the PLM-based pipelines significantly outperform the traditional machine learning approach utilising TF-IDF encodings. We also see that the multi-task model outperforms the single task models. This is most significant on the 'u' task on the external data. Interestingly all approaches find the 'N' task the easiest and the 'u' task the hardest. The 'N' task was likely easier as abnormal findings were usually self-evident with less ambiguity compared to the other tasks, and it is possible the 'u' task is hardest for the model to learn due to it being less formally defined than the TNM tasks. The multi-task model in particular offers encouraging performance over both internal and external test sets. The 'u' task's generalisation to external data particularly benefited from multi-task learning via the shared PLM encoder. Another key benefit of the multi-task approach is better computational efficiency over individual classifiers as it adds just 3 parameters (~0.00000087%) to a single-task transformer model, while performing significantly better on the 'u' task and equivalently on the TNM classes.

Dataset	Pipeline	ACC_{TNMu} ↑	ACC_{TNM} ↑	HL_{TNMu} ↓	$F1_T$ ↑	$F1_N$ ↑	$F1_M$ ↑	$F1_u$ ↑
Internal Test	Multi-task	**0.84 ± 0.01**	**0.86 ± 0.00**	**0.05 ± 0.00**	0.93 ± 0.00	0.94 ± 0.00	**0.92 ± 0.01**	**0.87 ± 0.00**
	Single task	0.80 ± 0.02	**0.86 ± 0.00**	0.06 ± 0.00	**0.95 ± 0.00**	**0.96 ± 0.00**	0.89 ± 0.02	0.85 ± 0.02
	TF-IDF	0.50 ± 0.00	0.60 ± 0.00	0.16 ± 0.00	0.69 ± 0.00	0.81 ± 0.00	0.69 ± 0.00	0.45 ± 0.00
External Test	Multi-task	**0.78 ± 0.01**	**0.83 ± 0.01**	**0.07 ± 0.00**	**0.89 ± 0.02**	**0.95 ± 0.01**	0.89 ± 0.01	**0.77 ± 0.00**
	Single task	0.73 ± 0.00	0.82 ± 0.00	0.08 ± 0.00	0.88 ± 0.00	**0.95 ± 0.00**	**0.90 ± 0.01**	0.68 ± 0.02
	TF-IDF	0.52 ± 0.00	0.61 ± 0.00	0.16 ± 0.00	0.49 ± 0.00	0.85 ± 0.00	0.76 ± 0.00	0.46 ± 0.00

Table 1. A comparison of machine learning pipelines including a multi-task approach using a shared PLM encoder (GatorTron), an ensemble of finetuned binary classifiers utilising GatorTron, and a traditional machine learning model using TF-IDF encodings and individual logistic regression classifiers for each binary class. For the single task ensembles, we calculate the 'TNMu' and 'TNM' metrics using the models trained from that random seed. Bold values represent the best performing pipeline for that metric on each test dataset. All F1 scores are macro averaged.

Conclusion

We developed a multi-task PLM-based NLP model which successfully classifies the presence/absence of TNM staging information in FDG PET-CT reports for lung cancer (and whether there is uncertainty associated with these findings) on both internal and external data. The TNM classification has potential to help with creating research cohorts, developing clinical alert systems, and assisting with audits. The uncertainty classification requires further refinement but represents a novel first step. The main limitation of this study is the amount of expert annotated data we could accrue. Only one expert was used to annotate the training data, which could potentially result in some bias towards their judgement in the models. By using an additional expert annotator and an agreement process on the external data, we hope that some of these concerns are mitigated due to the generalisation of the models on this external dataset.

Study Context

The data use and collection was approved by UK Research Ethics Committee (UK IRAS 228790) as part of Guy's Cancer Cohort (ref: 18/NW/0297)[7]. The authors acknowledge financial support from: EPSRC Research Council, part of the EPSRC DTP, Grant Ref: [EP/T517963/1], the Cancer Research UK National Cancer Imaging Translational Accelerator (C1519/A28682), and the Wellcome/Engineering and Physical Sciences Research Council Centre for Medical Engineering at King's College London (WT 203148/Z/16/Z). The authors declare there are no conflicts of interest in this work.

References

1. Farsad M. FDG PET/CT in the Staging of Lung Cancer. Current radiopharmaceuticals. 2020;13(3):195-203.

2. Patel Z, Schroeder JA, Bunch PM, Evans JK, Steber CR, Johnson AG, et al. Discordance Between Oncology Clinician–Perceived and Radiologist-Intended Meaning of the Postradiotherapy Positron Emission Tomography/Computed Tomography Freeform Report for Head and Neck Cancer. JAMA Otolaryngology–Head & Neck Surgery. 2022;148(10):927-34.

3. Park HJ, Park N, Lee JH, Choi MG, Ryu J-S, Song M, et al. Automated extraction of information of lung cancer staging from unstructured reports of PET-CT interpretation: natural language processing with deep-learning. BMC Medical Informatics and Decision Making. 2022;22(1):229.

4. Nobel JM, Puts S, Krdzalic J, Zegers KML, Lobbes MBI, F. Robben SG, et al. Natural Language Processing Algorithm Used for Staging Pulmonary Oncology from Free-Text Radiological Reports: “Including PET-CT and Validation Towards Clinical Use”. Journal of Imaging Informatics in Medicine. 2024;37:3-12.

5. Lababede O, Meziane MA. The Eighth Edition of TNM Staging of Lung Cancer: Reference Chart and Diagrams. Oncologist. 2018;23(7):844-8.

6. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. npj Digital Medicine. 2022;5(1):194.

7. Moss C, Haire A, Cahill F, Enting D, Hughes S, Smith D, et al. Guy’s cancer cohort – real world evidence for cancer pathways. BMC Cancer. 2020;20(1):187.

Leveraging Large Language Models to Extract Cancer Patient Experiences

Daisy Monika Lal[1], Paul Rayson[1], Erik van Mulligen[2], and Jan Kors[2]

[1]UCREL Research Centre, Lancaster University, UK
[2]Erasmus University Medical Center, Rotterdam, the Netherlands

1 Introduction

Nowadays, social media platforms have become a vital data resource for healthcare research as patients and caregivers are increasingly turning to the web for information and support [1]. People use these platforms to express their health concerns, opinions, beliefs, and preferences regarding a variety of diseases and treatment outcomes, including cancer [2, 3, 4]. Many cancer-support networks can be found on today's social media platforms, offering an anonymous environment for discussions on cancer-related challenges. Reddit has a diverse collection of cancer-related communities where patients and caregivers express their unedited thoughts, experiences, concerns, preferences, and sentiments, making it a useful tool for implementing value-based health care by eliciting patient values and preferences. However, the majority of the critical information contained in these narratives is concealed in unstructured natural language, rendering retrieving value and insight from these unorganised resources a significant challenge in the field of healthcare analytics. Large Language Models (LLMs), which have revolutionised the way we deal with unstructured data, may hold the key [5, 6]. These frameworks can be useful tools for mining hidden gems from the digital landscape of unstructured content [7, 8, 9, 10, 11, 12]. To fully harness the capabilities and potential of LLMs, smart prompts need to be designed to produce precise outcomes while filtering out extraneous details [13]. These prompts instruct and direct the LLMs to perform a specific task [14]. In this study, we used LLMs to analyse a subset of cancer-related Reddit posts using zero-shot prompt engineering, a technique that draws on natural language prompts to provoke desirable responses from the LLMs in the absence of labeled data. It hinges on the notion that LLMs can utilise their previous learning to comprehend unfamiliar tasks [15].

2 Methods and Data

In this work, we conducted qualitative research on prompt engineering for LLMs (GPT3.5 [16], ChatGLM3 [17], and Llama2 [18]) in the clinical setting without any base model training. Our approach involved utilising patient narratives sourced from Reddit, encompassing 2000 posts related to breast, prostate, and melanoma cancer. We employ these LLMs to extract themes (including symptoms, treatment encounters, emotional reactions, coping strategies, social support,

and interactions with healthcare professionals), values (comprising beliefs, principles, or priorities influencing decision-making), preferences (reflecting an individual's subjective choices or desires regarding medical care, treatment options, or healthcare encounters), and concerns (specific issues, worries, or focal points expressed or prioritised by patients) summarising the experiences of cancer patients. Our experiments were based on zero-shot prompt engineering [14, 19], which involved creating prompts with both basic and contextually augmented templates [20, 21], described as:

$$\mathcal{P}_{basic} = TextData + \mathcal{P}_{Ques} + OutputConstraint \tag{1}$$

$$\mathcal{P}_{aug} = TextData + \mathcal{P}_{Context} + \mathcal{P}_{Ques} + OutputConstraint \tag{2}$$

where, $TextData$ refers to an online post generated by a cancer patient on Reddit, $\mathcal{P}_{Ques}$ refers to the precise enquiry that the LLMs are tasked with answering, $\mathcal{P}_{Context}$ refers to the contextual information provided to the LLM to enhance its understanding of the input text (e.g., what the text is about, what is meant by "values, preferences, concerns, and themes" etc.), and $OutputConstraint$ is any constraint imposed on the format of the desired output.

3 Results

In the zero-shot scenario, we conducted an evaluation by analysing a series of 10 posts corresponding to each type of cancer. We utilised human annotators to assess the effectiveness of both basic and augmented prompts. The evaluation involved the examination of LLM-generated text quality in response to the prompts, employing the following quality metrics: 1) Fluency, measuring the smoothness and coherence of the text; 2) Relevance, assessing the alignment of the text with the given prompt or context; 3) Coherence, evaluating the logical and cohesive narrative of the text; 4) Grammatical Correctness, identifying any grammatical errors within the text; and 5) Diversity, scrutinising the range of outputs produced by the model for different prompts. Our findings indicate that zero-shot prompting generates promising outcomes when applied to cancer-specific data, although there are still challenges to overcome.

4 Conclusion

We use LLMs and zero-shot prompt engineering to analyse unstructured online accounts of cancer patients.[1] We conduct experiments using two prompt templates, basic and augmented. The augmented prompt directs the LLM toward the intended output by providing a context-sensitive comprehension of the imposed inquiry. Our findings show that LLMs generate effective responses when extracting themes, preferences, values, and concerns from patient narratives. Furthermore, guiding the LLMs with contextually enhanced cues improves their level of understanding and produces more accurate outcomes.

[1] Our research forms part of the 4D PICTURE project (https://4dpicture.eu/) which is funded from the EU research and innovation programme HORIZON Europe 2021 under grant agreement 101057332 and by the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number 10041120.

5 Study Context

The large scale language analysis of sentiment and emotions expressed in open or closed online forums, particularly related to sensitive topics such as cancer treatment requires ethical approval, and we have been granted approval for secondary data analysis of previously analysed datasets. The research presented here in this paper is part of a larger multilingual multinational research project, and each partner will apply it in their own organisation or country to replicate our analysis. The overall aim of the research is to improve the cancer patient journey, and ensure personal preferences are understood and respected during treatment discussions with medical professionals, thereby supporting treatment and care choices, at each stage of disease or treatment.[2]

References

[1] Kent EE, Rowland JH, Northouse L, Litzelman K, Chou WYS, Shelburne N, et al. Caring for caregivers and patients: research and clinical priorities for informal cancer caregiving. Cancer. 2016;122(13):1987-95.

[2] Bender JL, Jimenez-Marroquin MC, Jadad AR. Seeking support on facebook: a content analysis of breast cancer groups. Journal of medical Internet research. 2011;13(1):e1560.

[3] Bender JL, Jimenez-Marroquin MC, Ferris LE, Katz J, Jadad AR. Online communities for breast cancer survivors: a review and analysis of their characteristics and levels of use. Supportive Care in Cancer. 2013;21:1253-63.

[4] Domínguez M, Sapiña L. "Others Like Me". An approach to the use of the internet and social networks in adolescents and young adults diagnosed with cancer. Journal of Cancer Education. 2017;32:885-91.

[5] Percha B. Modern clinical text mining: a guide and review. Annual review of biomedical data science. 2021;4:165-87.

[6] Xiao W, Jing L, Xu Y, Zheng S, Gan Y, Wen C, et al. Different data mining approaches based medical text data. Journal of Healthcare Engineering. 2021;2021.

[7] Wadhwa S, DeYoung J, Nye B, Amir S, Wallace BC. Jointly extracting interventions, outcomes, and findings from RCT reports with LLMs. In: Machine Learning for Healthcare Conference. PMLR; 2023. p. 754-71.

[8] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nature medicine. 2023;29(8):1930-40.

[9] Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digital Medicine. 2023;6(1):210.

[2] Funder and ethics approval details are redacted for submission.

[10] Goel A, Gueta A, Gilon O, Liu C, Erell S, Nguyen LH, et al. Llms accelerate annotation for medical information extraction. In: Machine Learning for Health (ML4H). PMLR; 2023. p. 82-100.

[11] Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Network Open. 2023;6(11):e2343689-9.

[12] Liu S, McCoy AB, Wright AP, Carew B, Genkins JZ, Huang SS, et al. Leveraging large language models for generating responses to patient messages. medRxiv. 2023:2023-07.

[13] Russe MF, Reisert M, Bamberg F, Rau A. Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning. RoFo: Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin. 2024.

[14] Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing. arXiv preprint arXiv:230908008. 2023.

[15] Tringale M, Stephen G, Boylan AM, Heneghan C. Integrating patient values and preferences in healthcare: a systematic review of qualitative evidence. BMJ open. 2022;12(11):e067268.

[16] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.

[17] Zeng A, Liu X, Du Z, Wang Z, Lai H, Ding M, et al. Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:221002414. 2022.

[18] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.

[19] Nazary F, Deldjoo Y, Di Noia T. ChatGPT-HealthPrompt. Harnessing the Power of XAI in Prompt-Based Healthcare Decision Support using ChatGPT. In: European Conference on Artificial Intelligence. Springer; 2023. p. 382-97.

[20] Alhamed F, Ive J, Specia L. Using Large Language Models (LLMs) to Extract Evidence from Pre-Annotated Social Media Data. In: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024); 2024. p. 232-7.

[21] Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, et al. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2024;8(1):1-32.

Label-To-Text-Transformer: Generating Synthetic Medication Prescriptions

Samuel Belkadi[1], Nicolo Micheletti[1], Lifeng Han[1], Warren Del-Pinto[1], and Goran Nenadic[1]

[1]Department of Computer Science, University of Manchester, UK

1 Introduction

Access to real-world medication prescriptions is essential for medical research and healthcare quality improvement [1, 2, 3]. However, access to such data is often limited due to the sensitive nature of the information. This is particularly challenging for training and fine-tuning clinical Natural Language Processing (NLP), as often non-clinical staff would need access to confidential healthcare data. In response to these challenges, this study harnesses NLP methodologies to generate synthetic medication prescriptions. The use of this synthetic data alongside, or in place of, real medical data can alleviate challenges associated with accessing and employing sufficient data for NLP research [4].

2 Methods and Data

We have developed a novel task-specific model architecture, the Label-To-Text-Transformer (LT3), crafted to generate synthetic medication prescriptions. For example, given a medication "*docusate sodium*" we would expect it to generate a prescription such as "*docusate sodium 100 mg Capsule Sig: One (1) Capsule PO BID (2 times a day) as needed for constipation.*".

Based on the Transformer's architecture [5], LT3 is trained on a set of around 2K medication prescriptions collected from a subset of the MIMIC-III (Medical Information Mart for Intensive Care) database [6, 7] that aligns with the n2c2 2018 shared task data on adverse drug events and medication extraction with gold labels [8].

3 Results

We evaluate the synthetic data based on similarity to reference, intra-diversity and the suitability for downstream named-entity recognition (NER) training. The results in Table 1 demonstrate that LT3 achieves notably better results than fine-tuned T5 models across different scores.

To measure diversity, we have implemented a score that measures the Jaccard similarity score of the generated outputs of our models: a lower score indicates better diversity. The results show

Table 1: Quantitative evaluation of LT3 vs T5 (fine-tuned) on the test set.

	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
2 T5 Small	71.75	76.16	66.24	75.55	0.70
T5 Base	71.98	76.28	66.30	75.45	0.70
T5 Large	69.89	75.07	65.19	74.22	0.68
LT3	**78.52**	**78.16**	**68.72**	**77.55**	**0.72**

a lower intra-similarity score (0.650) for the generations of LT3 than T5 (0.660), implying that LT3 produces more diverse samples.

For downstream applications, we used the synthetic data generated by LT3 for training the SpacyNER model to compare the model performance with the ones trained from real data (see 1). The LT3 data achieved comparable performance to the real data,on five labels "drug, form, frequency, route, and strength" achieving 0.96+ scores.

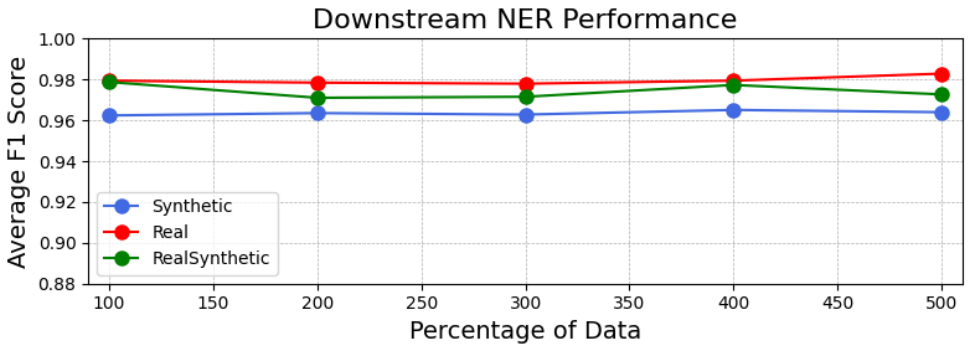


Figure 1: Average F1 score for five labels (Drug, Strength, Form, Route, Frequency) using Synthetic data, Real data, and Real+Synthetic.

4 Conclusion

To facilitate clinical NLP research and address the data privacy and restriction issues, we have developed LT3 for generating synthetic clinical data using pre-defined drug labels and related attributes from the n2c2-2018 shared task. The evaluation against the T5 model demonstrated that LT3 can generate better quality and diversity outputs. Furthermore, utilising synthetic data generated by LT3 for the NER task demonstrated its ability to effectively train SpacyNER, resulting in performances comparable to those achieved with real data.

Study context

This study used the deidentified data available through MIMIC III and have generated synthetic medication data, and therefore did not need an additional ethical approval. The work presented

here was partially supported by grants “Assembling the Data Jigsaw: Powering Robust Research on the Causes, Determinants and Outcomes of MSK Disease” (funded by the Nuffield Foundation; the views expressed are those of the authors and not necessarily the Foundation) and “Integrating hospital outpatient letters into the healthcare data space” (EP/V047949/1; funder: UKRI/EPSRC). SB and NM were partially supported by the University of Manchester student summer project via the Department of Computer Science.

References

[1] Nazari Nezhad S, Zahedi MH, Farahani E. Detecting diseases in medical prescriptions using data mining methods. BioData Mining. 2022 Nov;15(1):29. Available from: https://doi.org/10.1186/s13040-022-00314-w.

[2] Alrdahi H, Han L, Šuvalov H, Nenadic G. MedMine: Examining Pre-trained Language Models on Medication Mining. arXiv e-prints. 2023:arXiv-2308.

[3] Cui Y, Han L, Nenadic G. MedTem2.0: Prompt-based Temporal Classification of Treatment Events from Discharge Summaries. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop). Toronto, Canada: Association for Computational Linguistics; 2023. p. 160-83. Available from: https://aclanthology.org/2023.acl-srw.27.

[4] Chen J, Chun D, Patel M, Chiang E, James J. The validity of synthetic clinical data: A validation study of a leading synthetic data generator (Synthea) using clinical quality measures. BMC Medical Informatics and Decision Making. 2019 03;19.

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al.. Attention Is All You Need; 2023.

[6] Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3(1):160035.

[7] Johnson A, Pollard T, Mark R. MIMIC-III clinical database. PhysioNet; 2020.

[8] Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association : JAMIA. 2019 10;27.

Improving Biomedical Text Readability with LLMs and Controllable Attributes

Zihao Li[1], Samuel Belkadi[2], Nicolo Micheletti[2], Lifeng Han[2], Matthew Shardlow[1], and Goran Nenadic[2]

[1]Manchester Metropolitan University, UK
[2]University of Manchester, UK

1 Introduction

Biomedical literature often uses complex language and inaccessible professional terminologies. That is why simplification plays an important role in improving public health literacy and patient and public engagement. A recent community challenge on Plain Language Adaptation of Biomedical Abstracts (PLABA 2023, https://bionlp.nlm.nih.gov/plaba2023/) demonstrated the state-of-the-art in this domain. The task was to simplify biomedical abstracts sentence by sentence by either adapting it, splitting it in several sentences and then adapting them separately, or omitting it.

2 Methods and Data

As part of that challenge, we explored different models along with controllable attributes. The overall framework of our experimental design is displayed in Figure 1. In the first step, we fine-tune selected Large Language Models (LLMs) including T5, SciFive, BioGPT, and BART, apply prompt-based learning for ChatGPTs, and optimise control mechanisms on BART model. Training data distributed for the task included 750 abstracts that were simplified to result in 7,643 sentence pairs (source sentence, simplified sentence(s)).

We applied the modified control token strategy in [1] for both BART-base and BART-large models, leveraging both Wikilarge training set [2] and our split of training set from PLABA [3]. We used four attributes for control tokens (CTs) that represent 1) the syntactic complexity, 2) the lexical complexity, 3) the inverse similarity of input and output at the character level, and 4) the length ratio of input and output.

To evaluate bigger model architectures, we fine-tune FLAN-T5 XL [4] and BioGPT-Large, which have 3 billion and 1.5 billion parameters, respectively. FLAN-T5 XL is based on the pre-trained T5 model with instructions for better zero-shot and few-shot performance. To optimise training efficiency, and as our computational resources do not allow us to fine-tune the full version of these models, we employ the LoRA [5] technique, which allows us to freeze certain parameters, resulting in more efficient fine-tuning with minimal trade-offs.

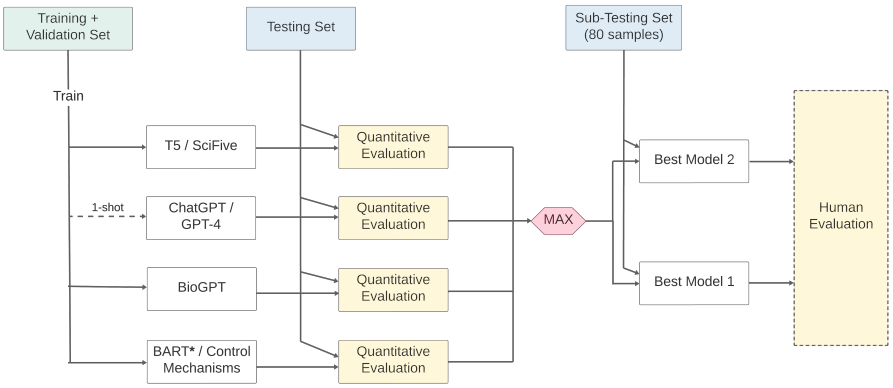


Figure 1: Model Development and Selection.

	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	SARI	BERTScore
2 T5 Small	**49.86**	**65.94**	**48.60**	**63.94**	33.38	69.58
T5 Base	43.92	64.36	46.07	61.63	44.10	**72.62**
T5 Large	43.52	64.27	46.01	61.53	*43.70*	60.39
FLAN-T5 XL (LoRA)	44.54	63.16	45.06	60.53	43.47	67.94
SciFive Base	44.91	64.67	46.45	61.89	44.27	60.86
SciFive Large	44.12	64.32	46.21	61.41	*44.38*	72.59
BART Base with CTs	21.52	56.14	35.22	52.38	46.52	50.53
BART Large with CTs	20.71	54.73	32.64	49.68	**46.54**	50.16

Table 1: Automatic Evaluations of fine-tuned T5, SciFive, and BART Models with Control Token (BARTs-w-CTs) mechanisms on the test set. FLAN-T5 XL used LoRA.

3 Results

The results are displayed in Table 1. The fine-tuned T5 Small model obtains the highest scores in both BLEU and ROUGE metrics, while the fine-tuned BART Large with CTs (BART-L-w-CTs) produces the highest SARI score. The fine-tuned T5 Base model achieved the highest BERTScore with a slightly lower SARI score (44.10), while the fine-tuned SciFive Large achieved the highest SARI score (44.38) among T5-like models.

4 Conclusion

The results presented here are encoruaging. Our submissions rank the 2nd in PLABA 2023 using the SARI score, with **BART-w-CTs** produced 2nd and 3rd highest scores on sentence-simplicity and term-simplicity. The results of human evaluation also demonstrate which of the adaptions needs to be improved.

5 Study context

This study uses publically available data and did not require ethical approval. The work has been partially funded by the grant "Integrating hospital outpatient letters into the healthcare data space" (EP/V047949/1; funder: UKRI/EPSRC).

References

[1] Li Z, Shardlow M, Hassan S. An Investigation into the Effect of Control Tokens on Text Simplification. In: Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022). Abu Dhabi, United Arab Emirates (Virtual): Association for Computational Linguistics; 2022. p. 154-65. Available from: https://aclanthology.org/2022.tsar-1.14.

[2] Zhang X, Lapata M. Sentence simplification with deep reinforcement learning. arXiv preprint arXiv:170310931. 2017.

[3] Attal K, Ondov B, Demner-Fushman D. A dataset for plain language adaptation of biomedical abstracts. Scientific Data. 2023;10(1):8.

[4] Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al.. Scaling Instruction-Finetuned Language Models. arXiv; 2022. Available from: https://arxiv.org/abs/2210.11416.

[5] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations; 2022. Available from: https://openreview.net/forum?id=nZeVKeeFYf9.

A tool for mapping medical narratives into medical ontologies in low resource settings: A case study for German

Faizan E Mustafa[1], Juan G. Diaz Ochoa[12]
[1] QuiBiQ GmbH, Stuttgart, Germany
[2] PerMediQ GmbH, Wang, Germany

Introduction

In health institutions, the extraction of information from unstructured healthcare data and clinical narratives is critical at the clinical and administrative levels (for example, to determine medical procedures or to obtain relevant information for reimbursement purposes). As an example, diseases can be coded (using the International Disease Classification ICD) using Natural Language Processing (NLP). Currently there is restricted availability of corpora and annotated data for NLP in medicine in languages other than English. We implemented a pipeline based on the generation of synthetic narratives for named entity recognition model training (to identify diseases and procedures in medical texts) and entity linking to link recognized entities to ICD and OPS (medical procedure) classifications. We deployed the trained models as a tool to assist customers identify relevant items in clinical narratives for reimbursement. The implementation has been performed in German but can be extrapolated into other languages.

Methods and Data

Real narratives are problematic from a data protection perspective. This is particularly critical in the development of AI models outside hospitals. Even if the model is deployed without a database, there is a probability that with reverse engineering critical data can be extracted from the model [1].

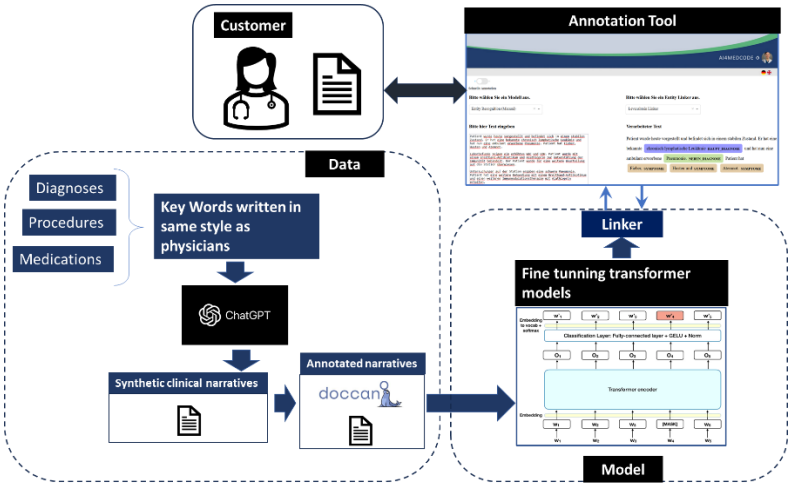


Figure 1. A pipeline for training the demonstrator models. In the diagram, a snapshot of the implemented solution is displayed, where the annotator (customer) can scan her documents and extract classifications (ICD for Diagnoses and OPS for medical procedures).

We create a synthetic dataset to train NER model without using any real patient data following the method used for the creation of the Aluminium standard [2]. We extracted medical entities (main disease, co-disease, medication, procedure) from the original narratives that are then automatically introduced in a prompt to generate the synthetic narrative through ChatGPT. The created synthetic narratives were manually annotated using Doccano and used to train NER model with SCAI-BIO/bio-gottbert-base [3] as base model for 20 epochs. We use SapBERT model finetuned on German biomedical data [4] to link the recognized entities by the NER model to ICD-10 (for diseases) and OPS ontology (for medical procedures). Finally, the trained model is deployed using PyDash on MS-Azure.

Results

NER validation results are reported for "Type" evaluation scheme [2] in Table 1. Since diagnoses usually have larger spans, they are more prone to be context dependent than procedures or medications (which are often provided in the medical narratives as single words or spans containing 2 to 3 words). Thus, validation values differ, and are often better for medical procedures and medications (with f1 > 0.5) and poorer for co-diagnoses (with f1 < 0.5).

Table 1. NER evaluation results.

	Precision	**Recall**	**F1**
Main Diagnose	0.51	0.59	0.54
Co Diagnose	0.27	0.78	0.41
Medication	0.64	0.50	0.56
Procedure	0.53	0.76	0.62

We have developed a demonstrator that identifies diagnostic and procedural elements within clinical narratives and links them to ontology entities. It occurs after a clinical narrative has been entered (see Fig. 1). The NER-recognized entities are highlighted in the text, along with their associated ontology entities. The customer can also select an appropriate ICD or OPS entity from the dashboard. This pipeline is easily adaptable to languages with limited resources.

Conclusion

We are presenting a demonstrator for a semi-automatic annotation tool for clinical narratives. In order to protect patient data, training data has been fully synthesized, in what we call a synthetic gold standard. The current implementation does not exclude the possibility to further fine tune the model using a silver standard based on bio-medical terminology (as reported by Wang et al. [5]). The method presented here allows us to train and deploy tools for text annotation in any language while protecting the privacy of original patient data. We plan to synthesize narratives using off-the-shell models like LLama2/3 / Mistral in future implementations.

Study context

This project was supported by the Ministry for Economics, Labor and Tourism from Baden-Württemberg, Germany via grant agreement number BW1_1456 (AI4MedCode).

References

[1] “Frontiers | Editorial: Transparent machine learning in bio-medicine.” Accessed: Mar. 27, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbinf.2023.1264803/full

[2] J. G. Diaz Ochoa *et al.*, “The Aluminum Standard: Using Generative Artificial Intelligence Tools to Synthesize and Annotate Non-Structured Patient Data.” Mar. 27, 2024. doi: 10.21203/rs.3.rs-3552289/v1.

[3] M. Lentzen *et al.*, “Critical assessment of transformer-based AI models for German clinical notes,” *JAMIA Open*, vol. 5, no. 4, p. ooac087, Nov. 2022, doi: 10.1093/jamiaopen/ooac087.

[4] F. E. Mustafa, C. Dima, J. G. Diaz Ochoa, and S. Staab, “Leveraging Wikidata for Biomedical Entity Linking in a Low-Resource Setting: A Case Study for German,” *Submited 6th Clin. Nat. Lang. Process. Workshop*, Mar. 2024.

[5] Y. Wang, C. Dima, and S. Staab, “[Novel] WikiMed-DE: Constructing a Silver-Standard Dataset for German Biomedical Entity Linking using Wikipedia and Wikidata,” presented at the The 4th Wikidata Workshop, Aug. 2023. Accessed: Apr. 04, 2024. [Online]. Available: https://openreview.net/forum?id=5dQ7YDSYya

NLP Enriched Research Data Extracts: An OMOP Pipeline for Producing Research Data Extracts

Kawsar Noor[1,3], Baptiste Paul Ribyere[3], Adam Sutton[2], Xi Bai[1,3], Tom Searle[2], Timothy Roberts[3], and Richard J Dobson[1,2]

[1]University College London, London, United Kingdom
[2]King's College London, London, United Kingdom
[3]University College London Hospitals, London, United Kingdom

1 Introduction

With the ongoing digitisation of UK hospital data enabling large scale evidence-based research, it remains critical to address how such data can be structured for multi partner collaboration projects and interoperability of data between systems. Whilst common data models such as the OMOP common data model address these needs, in practise they struggle to standardise the free-text portion of records on account of two things: firstly the free-text itself contains valuable clinical information, requiring extraction and structuring, and secondly the free-text, for governance purposes needs to be free of any identifiable patient data (names, addresses etc).

In this work we present and end-to-end pipeline for providing deidentified electronic healthcare records as research extracts following the OMOP standard. The pipeline has been developed at University College London Hospitals (UCLH) as part of a 'research data extracts service' available within the trust. The pipeline is comprised of two components: a component for populating the structured fields in the OMOP extract, and a components for populating the 'NLP/free text' fields. The 'NLP' tables export free-text notes and any extracted medical concepts present in the free text. At UCLH we have deployed AnonCAT [1] and MedCAT [2] for these respective tasks.

2 Methods and Data

Figure 1 depicts the end-to-end OMOP data extract service. The service starts with a consultation between the data extract services team and the researcher yielding a 'research project specification file' that captures the required cohort and associated meta data. This specification file is then fed to the automated OMOP extraction service (OMOP ES). OMOP ES subsequently interacts with two services. The first is a hospital managed research database (Caboodle) which is used to populate various structured OMOP fields such as demographic data, problem list and meta data about hospital visits etc.

The second is the CogStack platform [3] for the free-text data. OMOP ES communicates with the platform via the CogStack catalogue application through which it specifies the various types

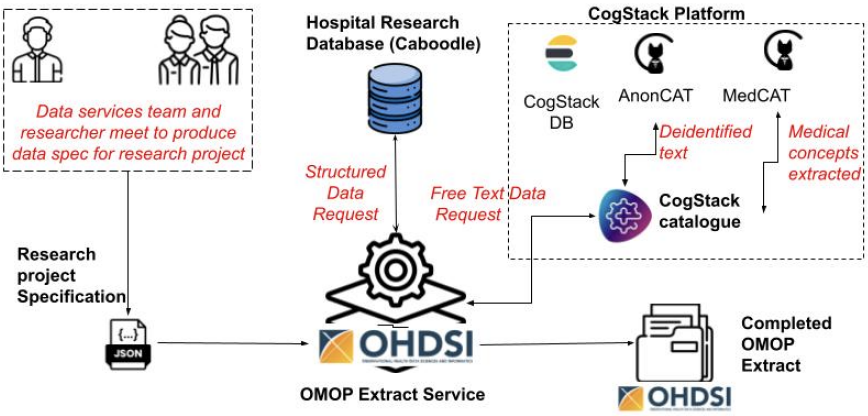


Figure 1: End-to-end OMOP and CogStack integrated research data extraction service.

of notes it requires and concepts it wants extracted. The catalogue application subsequently retrieves the free text and passes it through the UCLH trained and hosted AnonCAT service [1] for deidentidcation and thereafter MedCAT[2] for clinical concept extraction. The data is subsequently fed back to the OMOP ES for populating the NLP tables.

One the main challenges with this pipeline is validating the quality of deidenfitied data. UCLH have imposed a number of guidelines to address this including requiring the original requester to spot-check a subset of the documents. Secondly, as a separate process, every six months an approved annotator is asked to act as a motivated intruder and test the model using a combination of synthetic data and real data. These measures not only ensure AnonCAT's ongoing performance but also produces training data for further improving the model.

3 Results

To date the OMOP ES pipeline has mostly been used to provide structured data however as of this year demands for data extracts with free text have increased. Some of theses projects include the multi-centre multiple sclerosis project MS-Pinpoint [4] and the EU funded data tools for heart project [5].

4 Conclusion

We have demonstrated the feasibility of deploying an end-to-end research data extract service for providing OMOP compliant data extracts. We have also discussed the additional quality assurance measures being taken at UCLH to ensure high quality deidentifcation.

5 Study context

The OMOP ES and CogStack integration work was a collaboration between UCLH and KCH. Additionally both AnonCAT ane MedCAT models were cross trained between KCH and UCLH.

References

[1] Kraljevic Z, Shek A, Yeung J, Sheldon E, Shuaib H, Al-Agil M, et al. Validating Transformers for Redaction of Text from Electronic Health Records in Real-World Healthcare. 2023 jun:544-9.

[2] Kraljevic Z, Bean D, Mascio A, Roguski L, Folarin A, Roberts A, et al. MedCAT–medical concept annotation tool. arXiv preprint arXiv:191210166. 2019.

[3] Noor K, Roguski L. Deployment of a Free-Text Analytics Platform at a UK National Health Service Research Hospital: CogStack at University College London Hospitals. 2023:544-9.

[4] multiple sclerosis pinpoint. *MS Pinpoint*;. [Accessed 5th April 2024]. [Online]. Available from: `https://www.ms-pinpoint.com/`.

[5] Data Tools for Heart. *DataToolsforHeart*;. [Accessed 5th April 2024]. [Online]. Available from: `https://www.datatools4heart.eu/`.

Exploring GPT-4 for Fine-Grained Emotion Classification

Ratchakrit Arreerard, Scott Piao

School of Computing and Communications, Lancaster University, Lancaster, UK

1 Introduction

Mental health has been receiving an increasing attention. To assist people with mental health issue, the digital health and natural language processing communities have been exploring methods and techniques for automatically detecting mental health issues from textual data. A key technique useful for mental health analysis is the identification of emotions people express in their social media messages. Mental state is closely related to people's emotions. In fact, emotions can be the second mostly used feature for detecting mental issues on social media, particularly depression and suicide risk [1].

There exist various theories and psychological frameworks for classifying emotion categories. Some of them are coarse-grained, which divide emotions into broad main categories. For example, Ekman's scheme consists of six emotion categories [2]. Other emotion classification schemes attempt to define fine-grained emotion categories. For example, the GoEmotions scheme [3] consists of 27+1 emotion categories.

To automatically recognise emotions from textual data, several emotion classification models have been proposed and tested, such as [4, 5]. Since ChatGPT was introduced in late 2022, which is capable of classifying text [6], generative AI models have been tested for emotion classification. An important issue in this regard is, how the granularity level of the different emotion schemes can affect the emotion classification performance of generative AI. In this study, we investigate this issue by comparing the performance of ChatGPT4 for emotion classification with two different emotion schemes including Ekman's and GoEmotions schemes.

2 Methods and Data

In our experiment, we chose ChatGPT based on GPT-4 [7] as an emotion classification tool and selected GoEmotions dataset [3] as our test data. GoEmotions dataset is a collection of English Reddit messages, where each message is manually tagged with one or more emotion categories. The annotation scheme of this dataset consists of a range of finely grained emotion categories, including a total 27 emotion types and neutral category. These 27 emotion types can be grouped under the Ekman's broader six basic emotion categories.

We selected GPT-4 as the classifier because it was the latest generative AI model when we started this study. We wrote a Python script to access the OpenAI API of GPT-4, and used prompts to request GPT-4 to complete the emotion classification task. Figure 1 shows a prompt

template used in our experiment. The purpose of the prompt is to ask GPT-4 to select only one emotion category from a provided list of emotions conveyed by the given *Text*.

Figure 1: A GPT-4 prompt template.

We tested GPT-4 based on Ekman and GoEmotions schemes separately. For each scheme, we evaluated the performance of GPT-4 using an accuracy metric. First, we calculated an accuracy for each emotion category. Then we averaged the accuracy of all emotion categories, obtaining an overall accuracy for each scheme. Finally, we compared the overall accuracy between the two schemes.

3 Results

In our experiment, GPT-4 achieved an overall accuracy of 46.5% with the Ekman's scheme. It showed the best performance in identifying *fear* followed by *disgust* and *joy*, with accuracy of 70.1%, 67.1%, and 52.4% respectively. With GoEmotion scheme, GPT-4 obtained an overall accuracy of 35.6%, which is 10.9% lower than that with Ekman's scheme. GPT-4 obtained the best performance in identifying *amusement* with 80.1% accuracy, followed by *nervousness* (75%) and *disapproval* (67.2%).

We found that GPT-4 sometimes classifies the messages into classes beyond the range of provided candidate categories. With GoEmotion scheme, the number of emotions classified by GPT-4 reached 46 categories, while it produced 10 categories with Ekman's scheme. While some of them are incorrect, some others indeed capture correct emotions linked to the manual gold-standard annotation. In a couple of cases, GPT-4 suggestions appear to be even more appropriate than manual annotation. Occasionally, GPT-4 lacks understanding of context and refused to provide an answer when it determines a message as offensive.

4 Conclusion

Our experiment shows that, overall, a higher granularity level of emotion scheme negatively affects the performance of GPT-4. However, as shown in the previous section, some emotion categories of GoEmotion received higher accuracy than those of Ekman scheme. This implies that generative AI can perform better on more specifically defined narrower emotion categories than on broader basic categories. Another interesting finding is, the GPT-4 suggested 18 and 3 new categories to the GoEmotion and Ekman respectively, some of which even make more sense than human classification. This result implies a possibility that generative AI can potentially assist in designing a robust emotion classification scheme. Finally, it also raises an issue of how to accurately evaluate classification of generative AI where human produced gold-standard may not be completely reliable. For future work, we will extend our study for more emotion classification schemes and multiple emotion classification cases.

5 Study Context

This study used a subset (4,590 Reddit posts) of GoEmotions dataset [3] which is accessible at https://huggingface.co/datasets/go_emotions) and is publicly freely available. For detecting emotions, we used GPT-4 API that was accessed using a Python program. There is no ethical issue in our work.

References

[1] Malhotra A, Jindal R. Deep learning techniques for suicide and depression detection from online social media: A scoping review. Applied Soft Computing. 2022;130:109713. Available from: https://www.sciencedirect.com/science/article/pii/S1568494622007621.

[2] Ekman P. Basic emotions. Handbook of cognition and emotion. 1999;98(45-60):16.

[3] Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: A Dataset of Fine-Grained Emotions. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 4040-54. Available from: https://aclanthology.org/2020.acl-main.372.

[4] Hasan M, Rundensteiner E, Agu E. Automatic emotion detection in text streams by analyzing twitter data. International Journal of Data Science and Analytics. 2019;7:35-51.

[5] Akhtar MS, Ekbal A, Cambria E. How Intense Are You? Predicting Intensities of Emotions and Sentiments using Stacked Ensemble [Application Notes]. IEEE Computational Intelligence Magazine. 2020;15(1):64-75.

[6] Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: Jack of all trades, master of none. Information Fusion. 2023;99:101861. Available from: https://www.sciencedirect.com/science/article/pii/S156625352300177X.

[7] OpenAI. GPT-4 Technical Report; 2023. https://arxiv.org/abs/2303.08774.

Shedding light about canine and feline cancer in the UK. A text-mining approach to analyse 1,000,000 canine and feline tumour diagnoses between 2010 and 2023.

Jose Rodríguez Torres[1], Antonio Espinosa de los Monteros[1], Ángelo Santana[2], David R. Killick[3], PJ Noble[3], Alan Radford[3].

[1] Institute for Animal Health & Food Safety, Univ. of Las Palmas de Gran Canaria, Spain.
[2] Mathematics Department, Univ. of Las Palmas de Gran Canaria, Spain.
[3]Institute of Infection, Veterinary Science and Ecology, University of Liverpool, Neston, UK.

Introduction

Cancer registry surveillance systems rely on data quality across four dimensions: comparability, accuracy, completeness, and timeliness. While well-established in human medicine, such systems are lacking in veterinary medicine, necessitating efforts to develop robust surveillance tools(1).

The Small Animal Veterinary Surveillance Network (SAVSNET) was established in response to the lack of comprehensive surveillance and population health research in companion animals. Initially focused on infectious diseases and antimicrobial prescription and resistance, now it has expanded its scope to cancer surveillance in companion animals(2).

In 2021, as a pilot project, we published the largest canine and feline tumour registry(3) with more than 100,000 tumour cases reported to SAVSNET from electronic pathology records (EPRs) between mid-2018 and mid-2019 and using mostly Ms Excel and simple text-mining techniques based on key word searches.

Now, in our current project, following a similar and validated rule-based approach and using more powerful tools (Python libraries) we have curated a dataset of more than a million canine and feline tumour cases that can be updated periodically. Doing so, we are taking steps to improve both completeness and timeliness in our animal tumour registry.

Methods and Data

SAVSNET receives electronic pathology records (EPRs) from three laboratories, requiring data harmonization due to variations in structure and terminology. Utilizing Python (PY), data ordering processes were performed, including regular expressions to isolate diagnostic information and turn the dataset from animal-based to tumour-based. PY dictionaries were utilized to harmonize tumour names, breeds, and other variables. Geographical data, derived from practitioner postcodes, were mapped to rural/urban areas and NUTs regions.

Anticipating limitations of our approach, certain terms were modified to prevent undesired matches (for example, the word “lipoma” was modified to “lixpoma” in order to prevent an unwanted match with the word “lip”). Additionally, in order to dealing with false positives, reports containing strings like “no evidence", "no neoplas" or "inconclusive" were removed.

The main result was a normalized dataset of 1,101,028 tumours from 876,896 dogs and 101,722 cats reported in the UK from 2010 to 2023 including information about the kind of tumours diagnosed as well as about the animals suffering from it. Table 1 shows a row sample of the dataset.

Table 1. Tumour ID colum shows the identification of both the animal and the lesion. In this example, Tumour ID 1234.2 is a second tumour found in animal 1234.

LAB	Year	Spp.	Breed	Age	Sex	Tumour ID	Diagnosis	Location	Vet pcd	Rural urban	Area	NUTS3
VPG	2020	C	Labrador	13	M neut	1234.2	Fibrosarcoma	Thorax	N8	Urban	London	ABC

Concerning geographical data, postcodes of the submitting veterinary practitioner (used as a surrogate of owner location) were mapped to rural\urban areas as well as to NUTs regions using data from the Office from National Statistics.

Additionally, as example of the main descriptive findings, Figure 1 shows the distribution of breeds, sex and neuter status while Figure 2 displays the most frequent tumour types on urban and rural areas. Lipomas and mast cell tumours were the most frequently identified tumours consistent with former studies(4); breed distribution was also in line with former canine demographic studies(5).

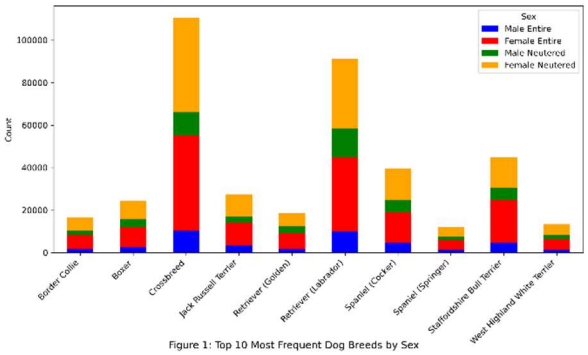


Figure 1: Top 10 Most Frequent Dog Breeds by Sex

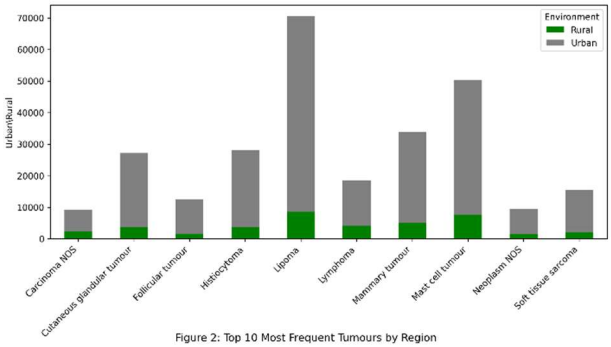


Figure 2: Top 10 Most Frequent Tumours by Region

Conclusion

This updated text-mining methodology has enabled the development of a large animal tumour dataset spanning 10 years of data from the UK pet population, optimized for easy expansion upon submission of further data.

As a secondary data-derived database, accuracy limitations and text-mining challenges, including spelling errors and context-dependent medical narratives, require careful consideration.

In terms of future research, comparison of these data with a population database will allow epidemiological studies of the UK dog population, attempting to find associations between tumour type, breed, sex or neuter status and geographical location leading to potential benefits for animals and their owners and others sharing the same environment. Additionally, the use of large language models (LLMs) and transformers presents a promising avenue for overcoming the limitations of traditional text mining methods mentioned above.

Study context

This study has been funded by Pet Plan Charitable Trust and has approval from University of Liverpool Research Ethics Committee. Authors declare no competing interests. A sample of data is available on figshare(3). We are also grateful for the support from veterinary diagnostic laboratories that routinely submit data to SAVSNET.

References

1. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. Eur J Cancer. 2009 Mar;45(5):747–55.
2. Radford A, Noble P, Pinchbeck G. REF2021 Reserch Excellence Framework. 2021. Small Animal Veterinary Surveillance Network (SAVSNET) research and surveillance initiative leading to behaviour change and improved companion animal health.
3. Rodríguez J, Killick DR, Ressel L, Espinosa de los Monteros A, Santana A, Beck S, et al. A text-mining based analysis of 100,000 tumours affecting dogs and cats in the United Kingdom. Sci Data. 2021;8(1).
4. O'Neill DG, Corah CH, Church DB, Brodbelt DC, Rutherford L. Lipoma in dogs under primary veterinary care in the UK: prevalence and breed associations. Canine Genet Epidemiol. 2018 Dec;5(1).
5. Sánchez-Vizcaíno F, Noble PJM, Jones PH, Menacere T, Buchan I, Reynolds S, et al. Demographics of dogs, cats, and rabbits attending veterinary practices in Great Britain as recorded in their electronic health records. BMC Vet Res. 2017 Jul 11;13(1).

The role of natural language processing in cancer care: a systematic scoping review with narrative synthesis

Mengxuan Sun[1], Ehud Reiter[2], Lisa Duncan[1], Rosalind Adam[1]
[1] Institute of Applied Health Sciences, School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, United Kingdom

[2] The School of Natural and Computing Sciences, University of Aberdeen, Aberdeen, United Kingdom

Introduction

In cancer care, the application of natural language processing (NLP) technology is an emerging and promising area. NLP could be used to enhance cancer care efficiency and improve patient treatment quality(1, 2), but the full potential is not yet clear. We conducted a systematic scoping review to investigate the role of NLP in cancer care, summarise the NLP methods used, analyse the results of evaluations of these technologies and identify potential limitations of the systems and their evaluations. This research aims to provide a comprehensive overview of the specific applications of NLP in cancer care and to guide future policy and research.

Methods and Data

The authors selected a scoping review to explore the scope, challenges, and complexities of using NLP in cancer care. Studies were included if they applied NLP to improve cancer care by patients or clinicians. Studies that focused on diagnosing cancer, death-related applications, studies that used cancer datasets for NLP research without a clinical application, or those using NLP solely for medical database searches were excluded. The study searched six databases with strategies that combined keywords, truncation symbols and wildcards. Four authors screened titles, abstracts, and full texts, excluding the studies that did not meet the inclusion/exclusion criteria (3).

Results

Database searching identified 4,507 titles. After screening, 29 studies were selected for inclusion. The included studies were published between 2013 and 2023, and an upward trend was observed in number of studies published over time. The most common country of origin was the United States (n=14, 48.28%), and studies were primarily published in medical journals (n=10, 34.48%). Most studies included multiple types of cancer (n=10, 34.48%).

Most NLP systems adopted feature extraction based pipelines (n=11, 37.93%) and embedding-based models (n=10, 34.48%) as system architecture. Electronic medical records were the most popular data sources of NLP systems across eleven studies (37.93%). Text classification (n=9, 31.03%) and information extraction (n=8, 27.59%) were the most common NLP tasks.

Fig 1 demonstrates an overview of the categories of NLP use in cancer care. NLP has been used in cancer care in four main ways: to assist doctors in improving clinical efficiency, for cancer risk stratification, for patient education and self-management, and to guide evidence-based treatment decisions. More studies are designed to assist doctors than helping patients. Many NLP systems have achieved over 80% accuracy, while some systems perform at a lower level of 60%. It's important to note that in medical, even small errors or language nuances can cause high-risk problems. However, only a few NLP studies have been evaluated in clinical settings.

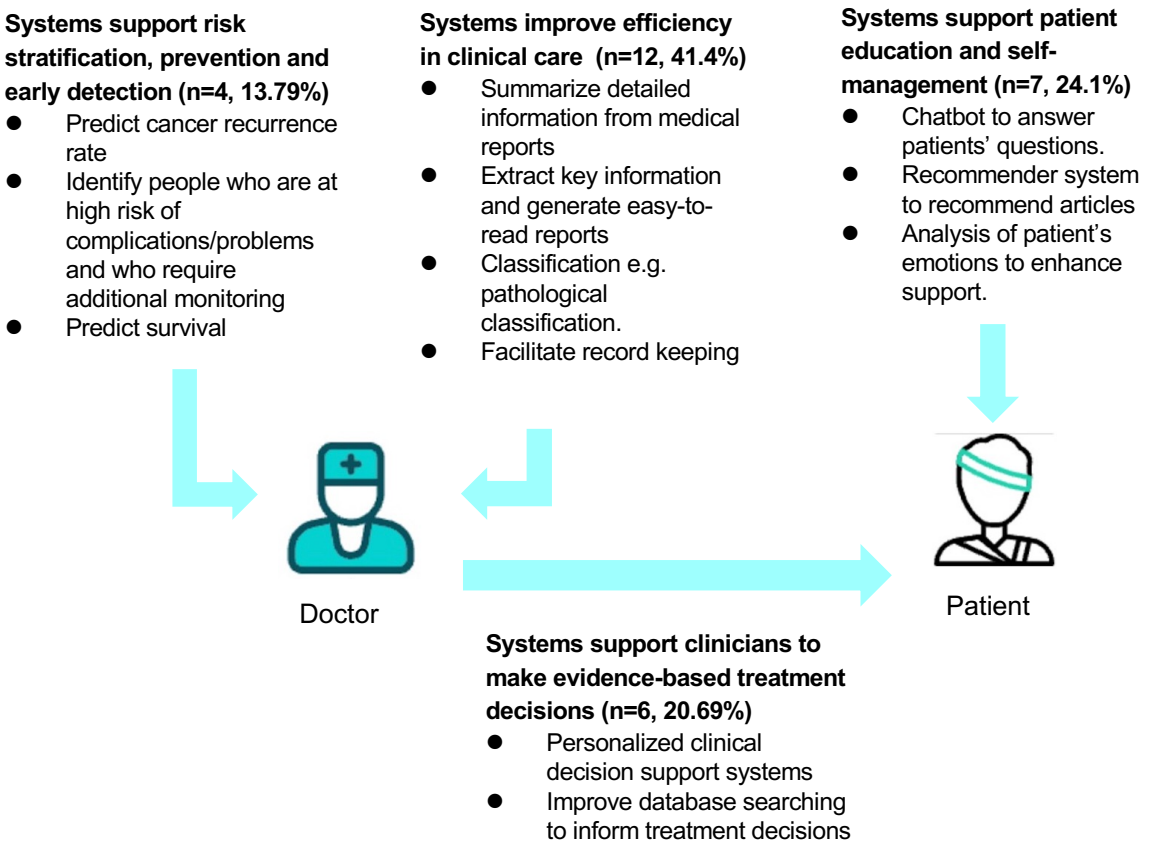


Figure1. Overview of NLP applied in Cancer Care Cases

Conclusion

NLP has made significant progress in information extraction and classification, which is conducive to automating and improving efficiency in clinical care processes. Chatbots, sentiment analysis, and decision support also have considerable potential benefits in improving patient-facing communication and support. However, there are currently limitations to implementing NLP in cancer care, including performance issues (accuracy) and a lack of evaluation in clinical practice. We strongly recommend that computer scientists, clinicians, and researchers work closely together to design processes to evaluate and implement effective systems in the clinical care pathway.

Study context

Registered Protocol: https://doi.org/10.17605/OSF.IO/G9DSR

Funding: This research was funded by the Chief Scientist Office (https://www.cso.scot.nhs.uk/)

Scottish Clinical Academic Fellowship (Grant CSO-SCAF/18/02). This grant was awarded to Rosalind Adam. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Marder SR. Natural language processing: its potential role in clinical care and clinical research. Oxford University Press US; 2022. p. 958-9.

2. Trivedi G, Dadashzadeh ER, Handzel RM, Chapman WW, Visweswaran S, Hochheiser H. Interactive NLP in clinical care: identifying incidental findings in radiology reports. Applied clinical informatics. 2019;10(04):655-69.
3. Mengxuan Sun RA, Ehud Reiter, Lisa Duncan. The Role of Natural Language Processing in Cancer Care: A systematic scoping review with narrative synthesis. OSF; 2023.

Investigating the Use of Transformer Models for Clinical Prediction Modelling – A Case Study in UK Biobank Secondary Care Data

Yusuf Yildiz[1], Goran Nenadic[1], Meghna Jani[1], David Jenkins[1]

[1] The University of Manchester, Manchester, United Kingdom

Introduction

In healthcare, there's a notable shift towards proactive risk assessment and prediction, alongside the rise of precision medicine (1). This shift has increased the number of clinical prediction models being developed. These models are often built using Electronic Health Records (EHRs). Despite their value, EHRs pose challenges due to their high dimensionality, heterogeneity, and temporal nature. Predictive modelling requires translating raw EHR data into a machine-readable format, often involving expert-defined features or deep learning methods (2,3). Deep learning techniques have shown promise in utilising EHR data for research, including the development of prediction models. Recent studies have introduced advanced approaches considering temporal order and irregular visit intervals, such as convolutional neural networks and Long Short-Term Memory networks (4). Additionally, embedding algorithms like Bidirectional Encoder Representations from Transformers (BERT) (5) have found applications in representing clinical concepts in EHR data. One of the examples, BEHRT(1), a variation of BERT, has been developed to represent patients' EHR data with temporal information.

The success of deep neural transduction models across various domains, combined with the document-like structure nature of EHRs, has inspired researchers to explore more of these models in clinical applications. This study aims to apply BEHRT to a real-world dataset to uncover implementation challenges specific to these networks and suggest potential solutions.

Methods and Data

This study utilises data from the UK Biobank, which includes anonymised records from approximately half a million patients. We extracted the secondary care data for all individuals. This includes ICD-10 codes for all diagnoses recorded in a patient's electronic health record and the associated date of entry. Age at each data entry was also mapped to the data from the baseline data available in the UK Biobank. Data were available between 2006 and 2010. The structured codes (diagnoses) for each patient were arranged in temporal order, with visits grouped into sequences.

The patient's medical history was captured through four embeddings. Disease embedding holds diagnosis information, age embedding records age at diagnosis, while segment and position embeddings encode visit sequence details. All together, they represent patient's medical journey, encompassing diagnoses, age at diagnosis, and visit order.

The workflow of this study is as follows: Patient embedding preparation, pre-training with masked language modelling and fine-tuning. This flow is illustrated with Figure 1.

The model performance is assessed with Average Precision Score (APS), Area Under the Receiving Operator and attention visualisation techniques. We fine-tuned the pre-trained model for 3 prediction tasks. The outcomes were:

- Next disease within patient's history
- Unplanned hospital readmissions within 30 days after a patient's discharge from the hospital
- Cardiovascular Disease (CVD) risk for 5 and 10 years for patients that has no history of CVD

Initial Results

The ongoing study encompasses 440,004 patients with at least one ICD-10 code in their secondary care record. Initial findings highlight the necessity of making various decisions depending on the dataset when applying these models:

- Minimum number of visits: Many studies opt to include patients with a minimum of 3 visits. This led to a 30% reduction in sample size.

- Vocabulary Size: The dataset consists of 12,221 unique clinical codes, but this count varies with different datasets, impacting the likelihood of each code occurrence. Thus, making transportability and training a model challenging.
- Temporal representation: Introducing a SEP token between visits is a common method to represent time intervals. However, not all intervals between consecutive visits are of equal length.
- Disease detail level: Diagnoses are encoded with ICD-10 codes, which can be up to 7 alphanumeric characters long. However, it's common practice in this field to retain only the first 3 digits of the code, resulting in some loss of information.

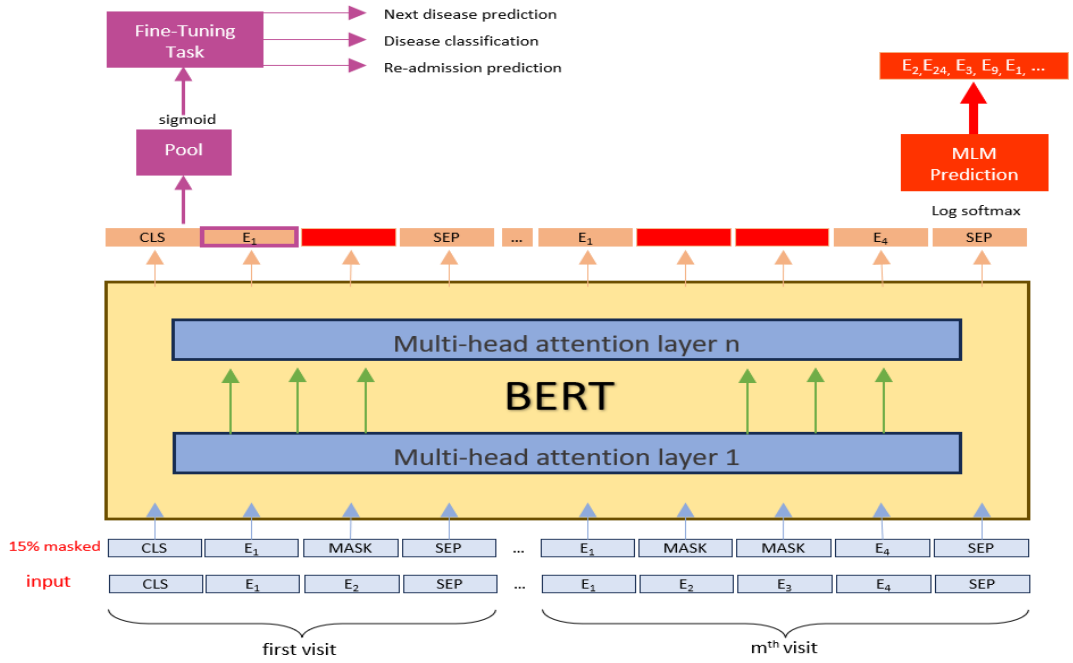


Figure 1. This figure shows how patient embedding is masked and fed into BERT architecture (Bottom to top). Next stage is pre-training with masked language modelling. The last stage is fine-tuning according to designed tasks.

Conclusion

In this research, we investigated the application of BERT, a deep neural transduction model, for clinical prediction using real-world health data. This methodology diverges from traditional clinical prediction algorithms, such as regression models, due to variations in data processing, predictor selection. While the utilisation of these models for clinical prediction shows promise and potential advantages compared to other modelling methods, numerous challenges and practical decisions require thorough investigation. This study has highlighted these challenges, and we suggest future research should concentrate on mitigating these obstacles. Specifically, establishing minimum visit criteria, including temporal information and how best to deal with the coding structure hierarchy. By doing so, transduction models can realise their full potential in healthcare prediction.

Study context

Yusuf Yildiz is funded by republic of Turkiye Ministry of National Education.

In this study UK Biobank data is used.

Code Availibilty

Code will be available on a Github repo after study's completion.

References

1. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for Electronic Health Records. Sci Rep. 2020 Apr 28;10(1):7155.

2. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating [Internet]. Cham: Springer International Publishing; 2019 [cited 2023 Oct 20]. (Statistics for Biology and Health). Available from: http://link.springer.com/10.1007/978-3-030-16399-0

3. Williams R, Kontopantelis E, Buchan I, Peek N. Clinical code set engineering for reusing EHR data for research: A review. J Biomed Inform. 2017 Jun 1;70:1–13.

4. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr: A Convolutional Net for Medical Records [Internet]. arXiv; 2016 [cited 2023 Aug 3]. Available from: http://arxiv.org/abs/1607.07519

5. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv; 2019 [cited 2023 Apr 12]. Available from: http://arxiv.org/abs/1810.04805

How Patient-Level Knowledge Graph Benefits ICD Coding

Mingyang Li, Viktor Schlegel, Goran Nenadic
University of Manchester, Manchester, UK

Introduction

ICD (International Classification of Diseases) coding is the process of allocating standardized ICD codes to diagnoses and procedures detailed in patient electronic records or paper notes. It's the translation of unstructured patient records, often presented in free text, into a structured format of coded alphanumeric data. This classification task offers advantages across various domains, including audit procedures, decision support systems and medical billing processes [1].

In the past years, the field of automated ICD coding has witnessed significant advancements, transitioning from rule-based methods to machine learning and deep learning approaches. Researchers have explored the application of advanced NLP models based on text format of data to represent the patient, such as RoBERTa, BigBird and Longformer. This paper aims to provide a more comprehensive representation of the patient's healthcare context by building a patient-level knowledge graph with rich and structured information of the patient's medical history. From the experimental result, our proposed model outperforms the baseline significantly, showing the efficacy of patient-level knowledge graph.

Methods and Data

A. Patient-Level Knowledge Graph

We utilize one of the most widely used NLP library in specific for the Healthcare domain Healthcare NLP [2] to extract the medical concepts within the notes. The output covers 14 types of entities and 14 types of relationships. Table 1 shows the statistics of 5 most common entities and relationships in dataset MIMIC-III Top-50.

Table 1. Statistics of most common 5 entities and relationships in Top-50 MIMIC-III dataset

Entity Type	Number of Unique Entities	Relatioinship	Number of Triples
problem	1438296	temporal events	2366185
treatment	663787	clinical relation	1024392
test	647029	posology	1433621
drug	15811	bodypart directions	131566
strength	1979	ade	69992

We build graph for each patient by constructing two kinds of triples <entity, relation, entity>: one links the entities and their relationship; the other one links the entity and its type. Figure 1 shows an example of a patient-level knowledge graph.

B. ICD Coding

The proposed framework is visually depicted in Figure 2. In the model, the patient is represented by both the text format note and its patient-level knowledge graph. We utilize pre-trained Roberta in the text embedding module, while graph embedding module is realized by deep GCN to capture both semantic and structured information of the graph. These two kinds of representations are concatenated before feeding into a label-wise attention layer. The output of this pipeline is the distribution of Top 50 labels in MIMIC-III.

C. Dataset and Metrics

We evaluate the performance of models on MIMIC-III Top-50 dataset (filtered data with 50 common labels). Like most evaluation methods for multi-label classification tasks, ICD coding task is typically evaluated through four standard metrics: Precision@N, Recall@N, F1 and AUC.

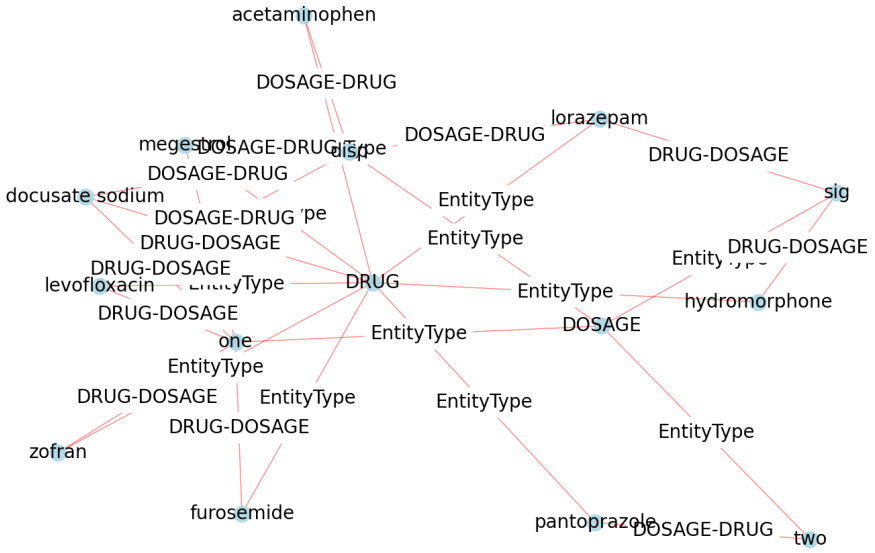


Figure 1. Patient-level Knowledge Graph

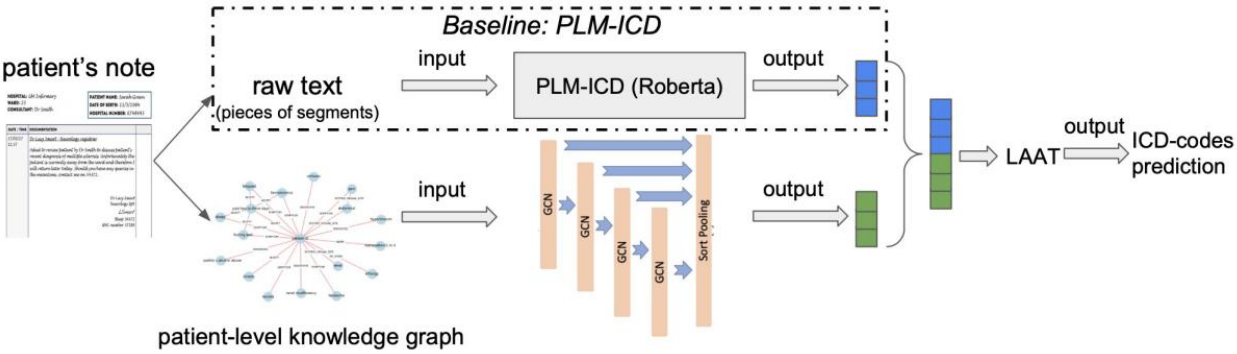


Figure 2. Framework of ICD coding with text-based and graph-based inputs

Results and Conclusion

Table 1. Performance on ICD Coding

	F1-Macro	F1-Micro	AUC-Macro	AUC-Micro	Precision@5	Recall@5
baseline	65.68	70.55	91.71	94.09	66.39	64.48
Text+Graph	67.25	71.44	92.03	94.24	66.77	64.83

The graph used in the proposed model consists of entities with four types of relationships (the most 4 common relationships). Our model performs better compared to one of the state-of-the-art work PLM-ICD [3] by around 1.8% and 1.1% in AUC-Macro and AUC-Micro metrics separately. It also achieves competitive results in other metrics. It exemplifies how patient-level knowledge graphs possess the capability to enhance the representation of individual patients, attributable to the structured nature of the graph.

Study Context

We used the Healthcare NLP library by John Snow Labs for named-entity recognition and relation extraction, leveraging its research license.

The dataset used in this work, MIMIC-III, is publicly available and requires authorization for access.

The baseline model is derived from the published work "PLM-ICD: Automatic ICD Coding with Pretrained Language Models," and its code is available on GitHub: https://github.com/MiuLab/PLM-ICD

References

1. Blundell, J., 2023. Health information and the importance of clinical coding. Anaesthesia & Intensive Care Medicine, 24(2), pp.96-98.
2. Spark NLP for Healthcare. Retrieved from https://www.johnsnowlabs.com/spark-nlp-health/.
3. Huang, C.W., Tsai, S.C. and Chen, Y.N., 2022. PLM-ICD: automatic ICD coding with pretrained language models. arXiv preprint arXiv:2207.05289.

Can GPT-3.5 Generate and Code Discharge Summaries?

Matúš Falis[1], Aryo Pradipta Gema[1], Hang Dong[2], Luke Daines[1], Siddharth Basetti[3], Michael Holder[1], Rose S Penfold[1], Alexandra Birch[1], and Beatrice Alex[1]

[1]University of Edinburgh, Edinburgh, United Kingdom
[2]University of Exeter, Exeter, United Kingdom
[3]National Health Service Highland, Inverness, United Kingdom

Introduction

Medical document coding is the task of assigning structured codes from a medical ontology – *e.g.*, the International Classification of Diseases (ICD) – to clinical documents. Human resources used in coding could be used elsewhere in patient care prompting research in Machine Learning and Natural Language Processing (NLP). Medical document coding is cast as a Large-Scale Multi-Labelled Text Classification (LMTC) task [1] in NLP. Labels within LMTC tasks display big-head long-tail distribution – few common conditions contrast with many underrepresented or absent in corpora (exacerbated by limited data availability). Performance of deep learning ICD coding models (*e.g.*, [2], [3], [4]) is negatively affected by data sparsity.

Recently, Large Language Models (LLMs), notably GPT-3.5, have become the standard for advanced NLP tasks. LLMs retain background knowledge seen in training and can utilise it when deployed, but also suffer from hallucinations [5]. LLMs' utility in medicine has recently been discussed [6, 7, 8, 9]. While using GPT-3.5 is problematic with real discharge summaries due to privacy issues, it has the potential to aid in generating synthetic data for training local models.

This study aims to assess GPT-3.5's efficacy in the context of automated ICD-10 coding and investigate its viability as: (1) A data generator for ICD-10 coding (especially for rare labels); (2) an automated ICD-coding classifier; (3) a generator producing clinically accurate and plausible synthetic data (based on input ICD-10 descriptions) from the perspectives of clinical experts.

Methods and Data

Employing GPT-3.5 we generated and coded 9,606 discharge summaries based on lists of ICD-10 code descriptions of patients with infrequent codes in the MIMIC-IV dataset [10]. Combined with the baseline training set, this formed an augmented training set. Neural coding models were trained on baseline and augmented data and evaluated on a MIMIC-IV test set. We report micro- and macro-F_1 scores on the full codeset, generation codes, and their families. Weak Hierarchical Confusion Matrices [11] determined in-family and out-of-family coding errors in the latter codesets. Coding performance of GPT-3.5 was evaluated on prompt-guided self-generated data and real MIMIC-IV data. Clinicians evaluated the clinical acceptability of the generated documents.

Results

Data augmentation results in slightly lower overall model performance but improves performance for the generation candidate codes and their families (Table 1, left), including one absent from

the baseline training data. Augmented models display lower out-of-family error rates (Table 1, right). When used for coding, GPT-3.5 identifies ICD-10 codes by their prompted descriptions in synthetic data, but underperforms on real data. Evaluators highlight the correctness of generated concepts while suffering in variety, supporting information, and narrative[1].

Table 1: A comparison between local neural network models trained on baseline ($base$) and augmented (aug) training sets and evaluated using micro- and macro-averaged F_1 scores (mi and ma respectively) on three codesets – ov (overall) on all codes present in MIMIC-IV; f on all codes within the families we chose for generation; and GPT-3.5 on candidate codes used in generation with a population of at most 100 in the baseline training set. Weak Hierarchical Confusion Matrix (WHCM) error rates are produced for codesets f and GPT-3.5. Performance on the common test set is reported using the macro-averaged proportion of errors that were Out-of-Family (OOF) and in-family (IF). The best score in each metric for each model pair (baseline versus augmented) is highlighted in **bold** (highest for F_1, lowest for WHCM).

Experiment	$F_1 \uparrow$						WHCM error $\downarrow$			
	mi_{ov}	ma_{ov}	mi_f	ma_f	mi_g	ma_g	OOF_f	IF_f	OOF_g	IF_g
CAML[2]$_{base}$	**53.65**	3.87	**38.43**	3.03	17.41	6.64	66.53	25.05	83.81	9.83
CAML$_{aug}$	53.54	**3.90**	38.41	**3.78**	**20.68**	**11.86**	**65.98**	**23.77**	**79.79**	**9.17**
LAAT[12]$_{base}$	**57.29**	**6.18**	**43.59**	4.96	**26.79**	14.48	58.57	**28.35**	74.03	12.20
LAAT$_{aug}$	57.18	6.09	43.36	**5.38**	25.70	**14.98**	**55.93**	29.78	**73.65**	**11.98**
MRCNN[13]$_{base}$	**55.66**	6.40	40.16	5.04	26.80	13.92	52.72	32.41	**69.68**	15.24
MRCNN$_{aug}$	54.69	**6.46**	**42.69**	**5.85**	**30.39**	**17.68**	**49.65**	**32.36**	70.41	**10.22**

Conclusion

While GPT-3.5 alone given our prompt setting is unsuitable for ICD-10 coding, it supports data augmentation for local neural models. Augmentation positively affects generation code families but mainly benefits codes with existing examples and reduces out-of-family errors. Documents generated by GPT-3.5 state prompted concepts correctly but lack variety and authenticity in narratives.

Context

This work is supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. HD is supported by the Engineering and Physical Sciences Research Council (EPSRC, grant EP/V050869/1), Concur: Knowledge Base Construction and Curation. RSP is a fellow on the Multimorbidity Doctoral Training Programme for Health Professionals, which is supported by the Wellcome Trust [223499/Z/21/Z]. BA is supported by the Advanced Care Research Centre at the University of Edinburgh. Our method was consulted with and approved by PhysioNet, as we merely use the descriptions of attached codes, which are not considered part of the MIMIC-IV.

[1]The results of the latter two experiments are not presented here due to the page limit

References

[1] Dong H, Falis M, Whiteley W, Alex B, Matterson J, Ji S, et al. Automated clinical coding: what, why, and where we are? NPJ digital medicine. 2022;5(1):159.

[2] Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable Prediction of Medical Codes from Clinical Text. In: Proceedings of NAACL-HLT; 2018. p. 1101-11.

[3] Dong H, Suárez-Paniagua V, Whiteley W, Wu H. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. Journal of biomedical informatics. 2021;116:103728.

[4] Kim BH, Ganapathi V. Read, attend, and code: pushing the limits of medical codes prediction from clinical notes by machines. In: Machine Learning for Healthcare Conference. PMLR; 2021. p. 196-208.

[5] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2023 mar;55(12). Available from: `https://doi.org/10.1145/3571730`.

[6] Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. New England Journal of Medicine. 2023;388(13):1233-9.

[7] Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT. Diagnostic and Interventional Imaging. 2023.

[8] Yeung JA, Kraljevic Z, Luintel A, Balston A, Idowu E, Dobson RJ, et al. AI chatbots not yet ready for clinical use. medRxiv. 2023:2023-03.

[9] Kraljevic Z, Bean D, Shek A, Bendayan R, Yeung JA, Deng A, et al. Foresight-Deep Generative Modelling of Patient Timelines using Electronic Health Records. CoRR. 2022.

[10] Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific data. 2023;10(1):1.

[11] Falis M, Dong H, Birch A, Alex B. Horses to zebras: ontology-guided data augmentation and synthesis for ICD-9 coding. In: Proceedings of the 21st Workshop on Biomedical Language Processing. Association for Computational Linguistics; 2022. .

[12] Vu T, Nguyen DQ, Nguyen A. A label attention model for icd coding from clinical text. arXiv preprint arXiv:200706351. 2020.

[13] Li F, Yu H. ICD coding from clinical text using multi-filter residual convolutional neural network. In: proceedings of the AAAI conference on artificial intelligence. vol. 34; 2020. p. 8180-7.

Developing a Common Schema for De-identification of Personal Health Identifiers in EHRs across Scotland

Arlene Casey[1], Matúš Falis, Franz S. Gruber[1], Matthew Murrell[1], Spyro Nita[1], Amy Tilbrook[1], Charlie Mayor[2], Katherine O'Sullivan[3], Kathy Harrison[1]
[1] University of Edinburgh, Edinburgh, Scotland, [2] West of Scotland Safe Haven, NHS Glasgow, Scotland, [3]DaSH, University of Aberdeen, Scotland

Introduction

De-identification of health records - the process of removing Personal Health Identifiers (PHIs) e.g. names, dates, age - to enable safe sharing of health data with researchers is an active and important research area [1]. However, the adoption of de-identification tools into organisations delivering health free-text for research still remains low in the UK. Adoption of existing schemas is challenging as health boards process data differently, and PHI entities can differ between nations meaning annotation schemas need to be modified, e.g. the use of Community Health Index number as a patient identifier in Scotland. Generalisability of tools across health record types can also be challenging due to their diversity [2].

Whilst we know privacy risks occur, there is relatively little systematic evidence about their real-world occurrence. Understanding privacy-risk prevalence combined with understanding inter-annotator disagreements helps develop better proportionate risk-based judgements, which minimise risk when approving health data extracts for research. In addition, this helps create an understanding of the most appropriate NLP approaches to support the task of de-identification of PHIs. In this study we worked across the Trusted Research Environments (TREs)[1] in Scotland to agree on a common schema for de-identification of PHIs. We apply this schema to understand how PHI entities occur within health boards and between health board regions, in two different electronic health record (EHR) types.

Methods and Data

Annotation schema development: Following a literature review of existing PHI schemas the four Scottish regional and national TREs took part in a workshop to discuss current approaches to PHI identification and agree on a common annotation schema for labelling PHIs. Annotation was undertaken in brat and eHOST depending on software approvals within the regional TREs. TRE 1 undertook pre-annotation with an existing tool using available structured data, patient first and last names, DOB and CHIs. In addition, the tool used regular expressions to annotate patterns, such as dates, postcodes, hospitals occupations, and medical or machine IDs.

Data: 2000 Discharge Summaries and 2000 X-Ray radiology reports for patients seen in the year 2022. The Discharge Summaries were selected based on the main hospitals within the regional TREs for patients 18+ who were admitted to a ward with at least a 24-hour stay.

Annotator arrangements and Scoring: TRE 1: All files were double annotated for both data sets. Annotation was undertaken by three annotators. TRE 2: All data was double annotated by two groups which consisted of a pool of 10 annotators, 5 assigned to each group, TRE 3: 20% of each data set were double annotated by a pool of four annotators. Scoring was done between pairs of annotators or the pools of annotators. We report F1 scores using strict matching. Note: although there are four regional TREs, one is not included in results reporting.

Results

Among the TREs', Radiology report human agreement of F1 scores are much higher than Discharge Summaries, as there are fewer entities within Radiology reports. Scores were generally higher in TRE 1, which could be attributed to using pre-annotation. TRE 2 had

[1] Organisations who work in partnership with NHS Health Boards to deliver access to health data for researchers

almost 50% fewer names or address-related information compared to TRE 1 and 3. This is a technical artifact of how data are processed, which allows TRE 2 to automatically remove many entities that would normally be placed within headers or footers in these EHRs.

Table 1 Results are F1 % scores listed in order for TRE 1, 2 and 3. Abbreviations: DS- Discharge Summaries, RAD – Radiology reports, GMC- General Medical Council Id, CHI – Community Index Number, UHPI-Unique Hospital Patient Number, – means no entities found,

EHR Type	First Name	Last Name	Title	Initials	Age	Dates	Year	Hospital	Ward	Occu-pation
DS	95/88/60	95/91/70	99/93/71	72/80/35	94/69/38	97/74/48	95/72/25	95/89/53	83/64/52	79/69/63
RAD	99/96/97	95/95/98	99/98/99	77/87/67	94/77/92	99/81/96	95/80/32	94/71/91	67/ 49/ -	97/68/97
EHR Type	**Address Line**	**Town**	**Country**	**PostCode**	**Phone**	**Org Name**	**Build Name**	**GMC**	**CHI**	**UHPI**
DS	80/15/13	63/77/66	67/64/80	92/92/71	83/50/53	09/45/63	0/64/17	100/08/0	100/71/68	60/0/0
RAD	0/100/-	100/40/-	67/100/92	-/100/-	100/100/98	67/62/96	-/ -/ -	98/54/99	-/ 98/ -	-/ -/ -

Highlighting Challenges: Patterns of disagreement around entity types and boundaries:

- Double first or last names, marking of dates and years inconsistently, and boundaries on age (often written as 46yo or 46F).
- Local hospitals marked as two entities i.e. town and a hospital (e.g. Fife hospital) particularly in one TRE which supported more rural areas.
- Phone numbers where they included extensions.
- All wards or clinics were annotated as we were interested in how many may be sensitive, for example, those for gender change, HIV, sexual transmitted diseases. Wards and clinics can be hard to recognise when named after clinicians or community places which caused disagreement on entity type or boundaries.
- Occupations mentions, for example, ‘the police found the patient’, disagreement occurred as to whether this occupation mattered in the context of privacy risk.

Conclusion

The outcome of our work was an agreed common de-identification schema which we applied to our data to understand PHI occurrence and how challenging agreement was. Our current work is reviewing the evidence of how and when entities occur to make policy decisions on how PHIs should be treated and consider what types of NLP tooling may work best to identify and protect patient confidentiality.

Study context

This work was supported by The Dunhill Medical Trust [grant number PF2302\2]; Research Data Scotland SDF1; DataLoch (dataloch.org) which is funded by the Data-Driven Innovation programme within the Edinburgh and South East Scotland City Region Deal. Ethical review and approvals for data access were carried out by each TRE according to their Caldicott agreements.

References

1. Kovačević A, Bašaragin B, Milošević N,Nenadić G. De-identification of clinical free text using natural language processing: A systematic review of current approaches, Artificial Intelligence in Medicine, 2024,102845,
2. Kraljevic *Z, Shek A, Yeung JA, Sheldon EJ, Shuaib H, Al-Agil M, Bai X, Noor K, Shap AD, Dobson R, Teo J.* Validating Transformers for Redaction of Text from Electronic Health Records in Real-World Healthcare, *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, Houston, TX, USA, 2023, pp. 544-549, doi: 10.1109/ICHI57859.2023.00098.

Towards one resource for drug prescription within the UK

Imane Guellil[1], Mike Holder[2], Aileen Elizabeth Stirling[3], Beatrice Alex[4], and Bruce Guthrie[5]

[1,2,4]University of Edinburgh, United Kingdom
[3]NHS Lothian

Introduction

The two key sources for almost all community-dispensed medicines in the UK are "GP-prescribed" datasets which are derived from electronic health records (EHRs), and "pharmacy-dispensed" datasets which are derived from pharmacy payment systems. In addition to having the prescriptions written in two different formats, they can be written using generic names (i.e. the accepted substance name, e.g. *paracetamol*) and others using brand names (i.e. the name given to a specific product by a drug company, e.g. *Panadol*). Also, drugs commonly come in combinations, where codes can only identify one of the drugs (e.g. combinations of thiazide diuretics and beta-blockers). The main idea of this ongoing work is to use Natural Language Processing to merge these resources.

1 Method and results

Three resources were used: GP-prescribed (CPRD Aurum EMIS[1])[1], pharmacy dispensed (the "BNF[2] Catalogue" that Dataloch provided us with), and the BNF-SNOMED mapping (associating BNF prescription to SNOMED[3] codes with a GP prescribing term). We automatically removed the dressings and bandages from EMIS and the "BNF Catalogue". We then considered a text matching (between EMIS and BNF-SNOMED) and proposed a matching algorithm for the remaining ones. This algorithm used a panoply of techniques including the use of regular expression, training a model and defining some rules for the matching. The main architecture of our approach and some results are presented in Figure 1.

2 Evaluation

To evaluate how many of the detected (matched) rows (EMIS prescription) are well mapped to their BNF Paragraph code (7th first characters of the BNF code), we select 1075 rows with

[1]derived from the EMIS GP EHR. Dataloch, and potentially other datasets, will also be used
[2]The British National Formulary[2], a reference for prescribing, dispensing and administering medicines in the UK.
[3]`https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct`

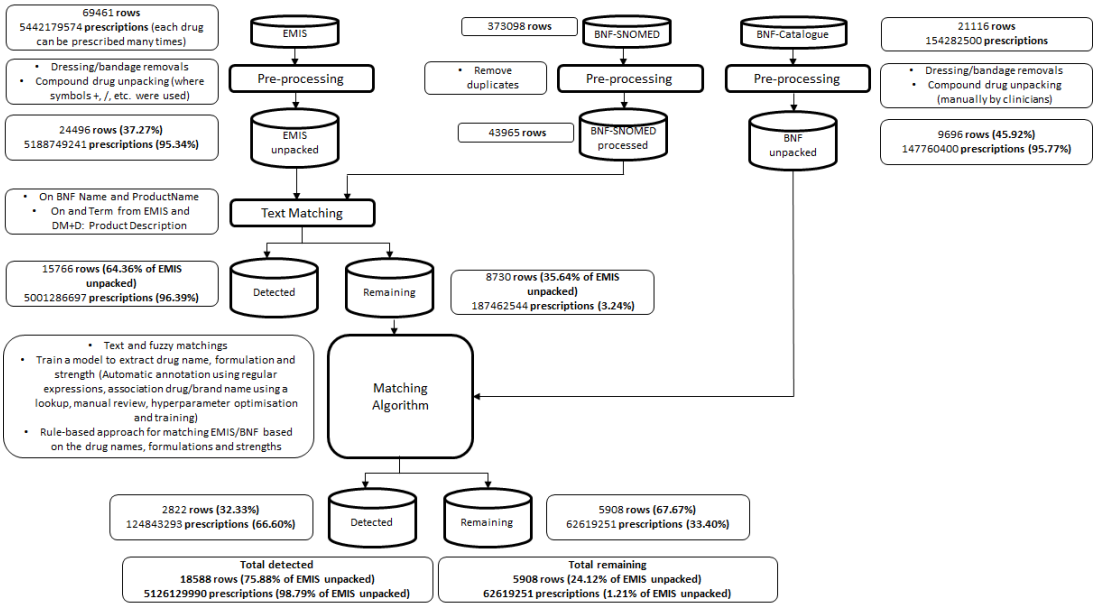


Figure 1: The main matching architecture associated with some results

a panoply of scenarios in our testing set (single drugs, compound drugs, the most prescribed drugs, etc.). Two clinician annotators manually map each prescription to its BNF Paragraph code. The annotators initially agreed on 878 rows (81.67%) corresponding to 95238835 prescriptions (90.54%). However, all the disagreements were resolved after discussion. Table 1 illustrates the different results obtained using the BNF-SNOMED, our algorithm and BNF-SNOMED + our algorithm on this gold dataset. When applied to the entire gold dataset, we can observe that 92.05% of the rows detected are correctly mapped and 92.97% of the prescriptions detected are correctly mapped using a combination of BNF-SNOMED and our algorithm (where we first detect the rows using BNF-SNOMED and apply our matching algorithm for the remaining ones).

Task	Technique used	Number of rows detected	number of prescriptions detected
Detection	BNF-SNOMED	748(69.58%)	94711938(90.04%)
	Our Algo	652(60.65%)	85410825(81.2%)
	BNF-SNOMED + Our Algo	881(81,95%)	101164897(96.18%)
Mapping	BNF-SNOMED	689(64.09%)	93443217(88.83%)
	Our Algo	622(57.86%)	82840293(78.75%)
	BNF-SNOMED + Our Algo	811(75.44%)	94051674(89.14%)

Table 1: Detection/mapping results on our gold test including 1075 rows/105187438 prescriptions

3 Conclusion

The aim of this work is to merge all of the drug-prescribing datasets for the automatic detection of adverse drug events in the future. These results are encouraging, with 98.79% of the prescriptions

correctly detected and 92.97% of these correctly mapped to their BNF Paragraph code. This work can be improved in the future by fine-tuning an OpenAI model to match the two datasets.

4 Study context

This study/project (AIM-CISC) is funded by the National Institute for Health Research (NIHR). Artificial Intelligence and Multimorbidity: Clustering in Individuals, Space and Clinical Context (AIM-CISC) grant NIHR202639. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

This research was funded by the Legal & General Group (research grant to establish the independent Advanced Care Research Centre at University of Edinburgh). The funder had no role in conduct of the study, interpretation or the decision to submit for publication. The views expressed are those of the authors and not necessarily those of Legal & General.

This research is based on CPRD Aurum (study protocol number: 21_000542).

References

[1] Clinical Practice Research Datalink. (2022). CPRD Aurum (Version 202205001). May 2022;(Version 2022.05.001). Available from: `https://doi.org/10.48329/t89s-kf12`.

[2] British National Formulary. London: British Medical Association). 2022;Edition 86. Available from: `Print`.

Developing a Common Model for De-identification of Personal Health Identifiers in EHRs across Scotland

Arlene Casey[1], Matúš Falis, Franz S. Gruber[1], Matthew Murrell[1], Spiro Nita[1], Amy Tilbrook[1], Charlie Mayor[2], Katherine O'Sullivan[3], Kathy Harrison[1]
[1] University of Edinburgh, Edinburgh, Scotland, [2] West of Scotland Safe Haven, NHS Glasgow, Scotland, [3]DaSH, University of Aberdeen, Scotland

Introduction

De-identification of health records - the process of removing Personal Health Identifiers (PHIs) e.g. names, dates, age - to enable safe sharing of health data with researchers is an active and important research area [1]. However, the adoption of de-identification tools into organisations delivering health free-text for research still remains low in the UK. Adoption of existing models is challenging as health boards process data differently, and PHI entities can differ between nations meaning annotation schemas need to be modified, e.g. the use of Community Health Index number as a patient identifier in Scotland. Generalisability of tools across health record types can also be challenging due to their diversity [2].

Whilst we know privacy risks occur, there is relatively little systematic evidence about their occurrence. Understanding privacy-risk prevalence combined with understanding inter-annotator disagreements helps develop better proportionate risk-based judgements, which minimise risk when approving health data extracts for research. In addition, this helps create an understanding of the most appropriate NLP approaches to support the task of de-identification of PHIs. In this study we worked across the Trusted Research Environments (TREs)[1] in Scotland to agree on a common model for de-identification of PHIs. Secondly, we applied this model to understand how PHI entities occur within health boards and between health board regions in two different electronic health record (EHR) types.

Methods and Data

Annotation schema development: Following a literature review of existing PHI schemas the four Scottish regional and national TREs took part in a workshop to discuss current approaches to PHI identification and agree on a common annotation schema for labelling PHIs. Annotation was undertaken in brat and eHOST depending on software approvals within the regional TREs. TRE 1 undertook pre-annotation with an existing tool using available structured data, patient first and last names, DOB and CHIs. In addition, the tool used regular expressions to annotate patterns, such as dates, postcodes, hospitals occupations, and medical or machine IDs.

Data: 2000 Discharge Summaries and 2000 X-Ray radiology reports for patients seen in the year 2022. The Discharge Summaries were selected based on the main hospitals within the regional TREs for patients 18+ who were admitted to a ward with at least a 24-hour stay.

Annotator arrangements and Scoring: TRE 1: All files were double annotated for both data sets. Annotation was undertaken by three annotators. TRE 2: All data was double annotated by two groups which consisted of a pool of 10 annotators, 5 assigned to each group, TRE 3: 20% of each data set were double annotated by a pool of four annotators. Scoring was done between pairs of annotators or the pools of annotators. We report F1 scores using strict matching. Note: although there are four regional TREs, one is not included in results reporting.

Results

Among the TREs', Radiology report human agreement of F1 scores are much higher than Discharge Summaries, as there are fewer entities within Radiology reports. Scores were generally higher in TRE 1, which could be attributed to using pre-annotation. TRE 2 had

[1] Organisations who work in partnership with NHS Health Boards to deliver access to health data for researchers

almost 50% fewer names or address-related information compared to TRE 1 and 3. This is a technical artifact of how data are processed, which allows TRE 2 to automatically remove many entities that would normally be placed within headers or footers in these EHRs.

Table 1 Results are F1 % scores listed in order for TRE 1, 2 and 3. Abbreviations: DS- Discharge Summaries, RAD – Radiology reports, GMC- General Medical Council Id, CHI – Community Index Number, UHPI-Unique Hospital Patient Number, – means no entities found,

EHR Type	First Name	Last Name	Title	Initials	Age	Dates	Year	Hospital	Ward	Occu-pation
DS	95/88/60	95/91/70	99/93/71	72/80/35	94/69/38	97/74/48	95/72/25	95/89/53	83/64/52	79/69/63
RAD	99/96/97	95/95/98	99/98/99	77/87/67	94/77/92	99/81/96	95/80/32	94/71/91	67/ 49/ -	97/68/97
EHR Type	**Address Line**	**Town**	**Country**	**PostCode**	**Phone**	**Org Name**	**Build Name**	**GMC**	**CHI**	**UHPI**
DS	80/15/13	63/77/66	67/64/80	92/92/71	83/50/53	09/45/63	0/64/17	100/08/0	100/71/68	60/0/0
RAD	0/100/-	100/40/-	67/100/92	-/100/-	100/100/98	67/62/96	-/ -/ -	98/54/99	-/ 98/ -	-/ -/ -

Highlighting Challenges: Patterns of disagreement around entity types and boundaries:

- Double first or last names, marking of dates and years inconsistently, and boundaries on age (often written as 46yo or 46F).
- Local hospitals marked as two entities i.e. town and a hospital (e.g. Fife hospital) particularly in one TRE which supported more rural areas.
- Phone numbers where they included extensions.
- All wards or clinics were annotated as we were interested in how many may be sensitive, for example, those for gender change, HIV, sexual transmitted diseases. Wards and clinics can be hard to recognise when named after clinicians or community places which caused disagreement on entity type or boundaries.
- Occupations mentions, for example, ‘the police found the patient’, disagreement occurred as to whether this occupation mattered in the context of privacy risk.

Conclusion

The outcome of our work was an agreed common de-identification model which we applied to our data to understand PHI occurrence and how challenging agreement was. Our current work is reviewing the evidence of how and when entities occur to make policy decisions on how PHIs should be treated and consider what types of NLP tooling may work best to identify and protect patient confidentiality.

Study context

This work was supported by The Dunhill Medical Trust [grant number PF2302\2]; Research Data Scotland SDF1; DataLoch (dataloch.org) which is funded by the Data-Driven Innovation programme within the Edinburgh and South East Scotland City Region Deal. Ethical review and approvals for data access were carried out by each TRE according to their Caldicott agreements.

References

1. Kovačević A, Bašaragin B, Milošević N,Nenadić G. De-identification of clinical free text using natural language processing: A systematic review of current approaches, Artificial Intelligence in Medicine, 2024,102845,
2. Kraljevic *Z, Shek A, Yeung JA, Sheldon EJ, Shuaib H, Al-Agil M, Bai X, Noor K, Shap AD, Dobson R, Teo J.* Validating Transformers for Redaction of Text from Electronic Health Records in Real-World Healthcare, *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, Houston, TX, USA, 2023, pp. 544-549, doi: 10.1109/ICHI57859.2023.00098.

A Privacy Risk Dashboard for Clinical Free-text

Franz S. Gruber[1], Matúš Falis[1], Amy Tilbrook[1], Arlene Casey[1]
[1] University of Edinburgh, Edinburgh, Scotland

Introduction

Providers of health data for research face significant challenges in extracting and linking complex data and safeguarding access by approved researchers. Risk assessments are a key component to ensure that data extracts are processed correctly and do not contain any identifiable patient information. Currently a significant burden is placed on data service teams – including data analysts and information governance specialists – to carry out risk assessment and minimise the risk of patients being identified. Whilst comprehensive systems are in use to support data service teams already, these are primarily manual when it comes to reviewing clinical free-text for privacy risks. This presents significant restrictions on being able to release clinical free-text for research as the process to manually review and check for potential privacy risks is challenging, and therefore very time consuming.

Our work focused on delivering a prototype dashboard that can bring together and visualise privacy risks found in free-text clinical records. The purpose of the dashboard was to support Information Governance (IG) and data analysts within Trusted Research Environments (TRE) at the data access stage in better understanding risks present in a data cohort and to enable proportionate risk-based decisions about data access.

Methods and Data

Initial mock-up designs were done and discussed amongst the team, which included representatives from across the Scottish Safe Havens and facilitated by a specialist design company. Whilst the initial purpose of the prototype is to provide a visual means to explore risks present in a cohort, there are plans for the future to use the App to track and audit the de-risking decisions. For privacy risk assessment, our public involvement activities have emphasised that the use of semi-automation tools should be followed by human checks and monitored using audit trails. Our design process captured the full workflow but we have implemented only the visualising and exploring of privacy risks in the current version, with development done using R Shiny App (see Figure 1 for an example of what this full workflow could look like). Our reasoning behind using R is this is the most universally accepted tool in use within Safe Havens in Scotland not requiring additional approval levels beyond existing mechanisms.

Results

The dashboard tool allows for a cohort to be uploaded, assuming it has been appropriately labelled for risks (see Figure 2 for an example of the tool). The dashboard can be used to understand the types of reports that are in the cohort, sex, age range, ethnicity, and Scottish Index of Multiple Deprivation (SIMD) details along with the years range of the reports. Each of these variables can be filtered to rearrange views of the data and of the potential risks. The risks can be viewed based on the filter selection. The report also provides a summary table of these risks which the user can drill down to individual risk types and see how these occur together in a report. The user can also view individual reports with their risk counts as well as rearrange the view to look on a per patient basis and understand cumulative risk. This support Information Governance teams to identify any risk-mitigations steps necessary.

Conclusion

Our privacy risk tool provides a prototype for how we can visualise privacy risks in free text and support those responsible for making decisions about making data extracts available for research. It is a tool that will be built upon in the future and we can see this extended to cohort requests to include both structured and unstructured clinical notes.

Figure 1 Example of workflow of the Risk Management Dashboard

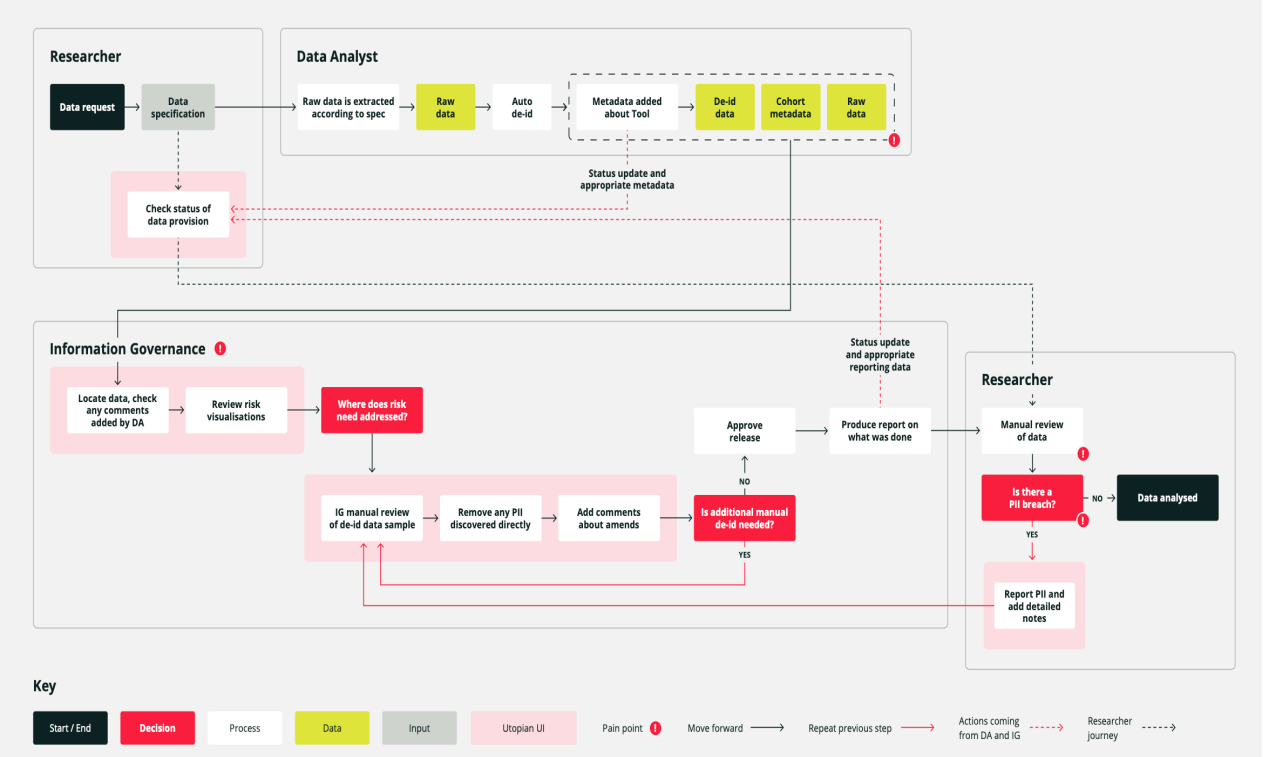
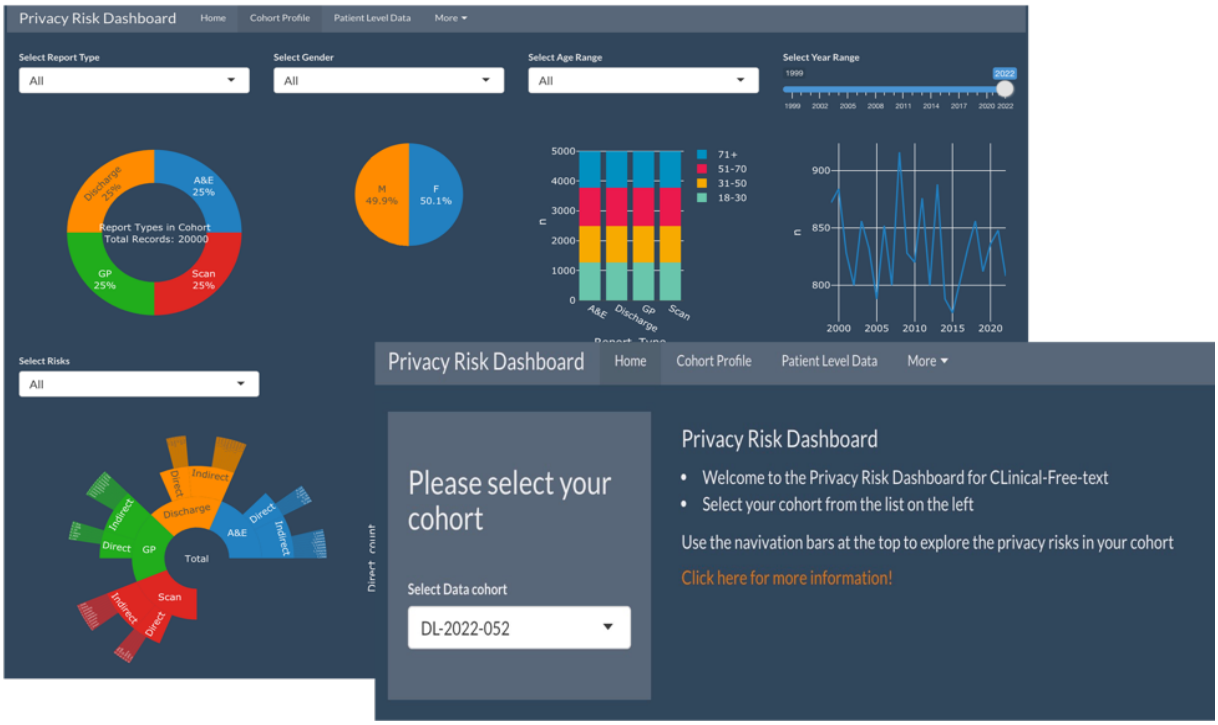


Figure 2 Prototype Privacy Risk Dashboard



Study Context
This work was supported by The Dunhill Medical Trust [grant number PF2302\2]; UK Research & Innovation MC_PC_23005; DataLoch (dataloch.org) which is funded by the Data-Driven Innovation programme within the Edinburgh and South East Scotland City Region Deal.

The challenge of teasing out language in veterinary electronic healthcare records

Katharine Anderson[1,2], Sean Farrell[1,2], Robert Christley[3], PJ Noble[1], Gina Pinchbeck[1]

[1] University of Liverpool, UK
[2] University of Durham, UK
[3] Dogs Trust, UK

Introduction

Electronic health records can provide a critical resource for evaluation of important behavioural problems in dogs with significant implications for canine and human welfare. Valuable information about canine behaviour is often hidden in free-text notes and to extract this at scale, some form of automation is required. Clinical language can be diverse within records and in particular cases, such as that in canine aggression, cases can be complicated and nuanced. Furthermore, capturing critical features about behaviour such as context and setting requires subtle analysis of language. This research, therefore, explored this issue in greater detail, addressing methodological considerations for extracting behavioural data from canine veterinary consultation notes, using canine aggression as an example.

Methods and Data

This study used electronic health records collected through the Small Animal Surveillance Network (SAVSNET); a syndromic surveillance, where participating veterinary surgeons record clinical free-text and treatment information at the end of each consultation which is sent to, and securely stored by, SAVSNET for analysis. Based on a case definition of aggressive behaviour (specifically that the owner reports an aggressive behaviour displayed in context or setting away from the clinic, towards any person or other animal, for any function (e.g. fear, resource guarding etc), subtle pattern matching using regular expressions was applied to the clinical free text. Design of the regular expression required multiple iterations, incrementally incorporating key words related to aggression in order to capture these records in the data, until a final version was created:

```
ag{1,}res{1,}ion|ag{1,}res{1,}ive|(ag{1,}res{1,}ion|behaviour|ag{1,}res{1,}ive)\Wtowards|defen
sive|pos{1,}es{1,}ive|lunge|lunged|lunging|growl|growls|growling|\bdomina(nt|nce|te)|bit\W(a\
sperson|people|a\sdog|child.*|son|family\smember|man|her|him|owner|O\b|friend|dad|strang
er|someone|neighbour|mother\b)|has\Wbit(ted|ten|\s)|resource\W(guarding|guard|guarde[dr])
|(guarding|guard|guarded)\W(behaviour|around|food|toys)|(nervous|fear)\W(ag{1,}res{1}ive|a
g{1,}res{1,}ion)|aggressive\Wbehaviour|(?<!not)\Waggressive|aggressive\W(?!treatment)|ag
gressive\W(?!towards\Wme)|protective\W(of|behaviour|over|with)|territorial|provoked|unprov
oked|attacking|\bnip\b|nipped|grumpy\Wwith
```

Results

The regular expression returned 121,191 records in total, approximately 1300 ‘cases’ per 100,000 records. Manual screening of 100 records demonstrated poor performance of the regular expression with low accuracy and specificity with only 39% of returned records matching the case definition. For example, wording such as aggressive returned semantically unrelated meanings such as aggressive disease or treatment; and bitten or bit often returned cases where the patient themselves were bitten by something, rather than they had bitten another. Further, many scenarios related to aggression occurring within the consultation itself only during examination and not being reported in the context defined by the case definition. Following manual assessment of filtered cases we found owners often sought help from vets

for both dog-directed and human-directed aggression with human-directed aggression cases more commonly resulting in euthanasia, compared to other management practices such as pain trials, neutering and calming supplements.

Conclusion

An initial evaluation of regular expression searching to assist in identifying patients that display aggressive behaviour to other animals or people outside the veterinary clinic demonstrated that this approach could capture large volumes of records but lacked specificity, such that critical contextual features could not be excluded, and key words (‘aggressive’) were often used outside the behavioural context. Having manually annotated a large group of records matching our case definition, we estimate the full database of over 10 million records may contain up to 25,000 relevant records. We are now using our annotated records as training data for large language model-based tools in order to extract specific records at scale.

Study context

The SAVSNET project has ethical approval from the University of Liverpool Research Ethics Committee (RETH001081, previously RETH000964). Research projects using SAVSNET data must pass through an internal process known as a Data Access and Publication Panel (DAPP). DAPP approval was received covering the work outlined in this research. Funding for this research was provided through a Dogs Trust Canine Welfare Grant.

Data, Dialogue, and Design: Patient and Public Involvement and Engagement for Natural Language Processing with Real-World Cancer Data

Wuraola Oyewusi[1], Eliana Vasquez Osorio[1,3], Goran Nenadic[1], Issy MacGregor[2], and Gareth Price[1,3]

[1]The University of Manchester, Manchester, UK
[3]The Christie NHS Foundation Trust, Manchester,UK
[2]NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust,UK

1 Introduction

Traditional clinical trials are crucial for gathering evidence for cancer treatment but often struggle with the underrepresentation of patient populations, limiting generalizability. Real-World Data (RWD) collected during routine patient care offers an alternative [1]. Unlike trials with narrow criteria, RWD allows learning from a broader patient range. However, it is estimated that around 80% of RWD exists as unstructured free-text medical notes[2]. Natural Language Processing (NLP) can extract insights from this data but working with free-text data in healthcare presents unique challenges, including privacy concerns.To ensure ethical and trustworthy NLP applications in cancer research, Patient and Public Involvement and Engagement (PPIE) is crucial.PPIE studies have addressed different health topics including medical free text[3].This work explores the process and findings from implementing a PPIE for using NLP on cancer medical notes.

2 Methods and Data

PPIE actively involves patients, carers, and the public in shaping research. [4]We conducted a PPIE and Figure 1 illustrates the four key steps we followed in designing our PPIE event on NLP for cancer medical notes.

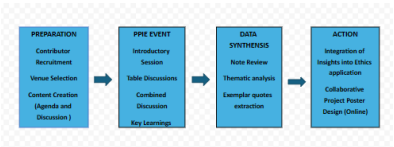


Figure 1: PPIE Methodology for NLP on cancer medical notes.

Preparation: In collaboration with a local research network, we recruited participants, selected a venue, and developed discussion questions centered on data use, consent, and communication.
PPIE Event: We conducted a two-hour discussion event with thirteen participants, comprising nine cancer survivors/caregivers and four researchers. The discussion focused on data use, research participation, and the application of NLP to medical notes.
Data Synthesis: Facilitators analyzed the discussion notes to identify key themes and representative quotes.
Action: We integrated insights from the PPIE event into the ethics application and collaboratively designed a project communication poster with some of the contributors, highlighting the significance of patient inclusion and advocacy in NLP research.

3 Results

The PPIE event provided insights into patient perspectives on the application of NLP to cancer medical notes.The discussion revolved around three key areas: data use, research participation, and communication.Table 1 shows the themes and exemplar quotes from some the questions asked about each point.

Table 1: Themes, Sub-Themes, Questions, and Exemplar quotes from the PPIE event on NLP for cancer medical notes.

Theme	Sub-Theme	Question	Exemplar quotes
Data Use	Data Accuracy	How accurate is the data? Accuracy vs privacy	"My GP has very detailed notes on me electronically. I am not concerned at all" "The data is as good as the recorder"
	Data Completeness	How is data about patients treated at multiple centers included?	"Previous cancer history is important would be missed? = partial results/accuracy"
Research Participation	Vote on how to obtain consent	Voting Result	National Opt-Out: 6/9 Project Specific Opt-Out: 3/9
Research Communication	Inclusion	Why is the study covering only a few years as treatments can change – will the data be out of date?	"2020 - 2014 Old DATA not Current"
	Research info dissemination	Is there provision for non-English speakers?	"What to do if you want to be involved? / How not to be involved?"

4 Conclusion

In our work, we present insights from a PPIE Event focused on the application of NLP to real-world cancer medical notes. While data privacy remains crucial, contributors emphasized balancing information availability with privacy protection. The majority (66.6%) favored a National Opt-Out consent model, and clear communication and inclusivity were highlighted. These insights enhance our research design as we proceed with NLP applications in cancer notes under ethical approval

5 Study context

This study bridges AI research with real world cancer medical notes and patient advocacy. It is also part of ethics application for a larger project on NLP to optimise cancer care funded by the Medical Research Council(MRC)

References

[1] Tyler N, Giles S, Daker-White G, McManus B, Panagioti M. A patient and public involvement workshop using visual art and priority setting to provide patients with a voice to describe quality and safety concerns: Vitamin B12 deficiency and pernicious anaemia. Health Expectations. 2021 Feb;24(1):87-94.

[2] Benedum CM, Sondhi A, Fidyk E, Cohen AB, Nemeth S, Adamson B, et al. Replication of Real-World Evidence in Oncology Using Electronic Health Record Data Extracted by Machine Learning. Cancers (Basel). 2023;15(6):1853.

[3] Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell JA. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. Journal of Medical Ethics. 2020;46:367 377. Available from: https://api.semanticscholar.org/CorpusID:218909917.

[4] Bertelsen N, Dewulf L, Ferrè S, et al. Patient Engagement and Patient Experience Data in Regulatory Review and Health Technology Assessment: A Global Landscape Review. Therapeutic Innovation & Regulatory Science. 2024;58(1):63-78.

Multimodal LLM for Computer Assisted Intervention: Human in the Loop with Eye Gaze of Radiologists

Yunsoo Kim[1]

[1]University College London, London, United Kingdom

1 Introduction

In recent years, the integration of Vision-Language Models (VLMs) with Large Language Models (LLMs) holds great promise of a new era for Medical Image Computing (MIC), extending its applications to handle intricate multimodal data. This multimodal LLM revolutionizes the interpretation of medical images and reports by facilitating comprehensive analysis without requiring extensive fine-tuning [1, 2, 3]. This versatility of VLMs has shown promising results in various downstream tasks, from generating radiology findings automatically to answering visual queries and correcting radiology reports based on CXR images [4, 5, 6, 7].

Despite these advancements, the effectiveness of these models as standalone tools in real-world clinical scenarios remains uncertain, given the inherent complexity of medical imaging data characterized by its variability and contextual intricacies. Addressing this challenge necessitates innovative methodologies, such as the adoption of a human-in-the-loop (HITL) framework. This collaborative approach fosters synergy between AI systems and clinical expertise to enhance diagnostic accuracy, reliability, and interpretability. HITL systems in radiology have demonstrated superior diagnostic accuracy compared to either radiologists or AI operating in isolation [8, 9].

However, current models' human-computer interaction remains somewhat limited, predominantly focused on single-modality applications in computer vision (CV) [10, 11]. To elevate human-computer interaction to a multimodal level, I propose a novel HITL approach for computer assisted intervention leveraging eye gaze data.

The proposed HITL method integrates textual prompts enriched with eye fixation data and image prompts overlaid with eye gaze pattern heatmaps into the VLM framework. This multimodal approach aims to advance human-centred AI research in medical image computing by integrating artificial and human intelligence. The proposed research will test the effectiveness of radiologist's expertise in enhancing the model's performance in various medical image analysis tasks.

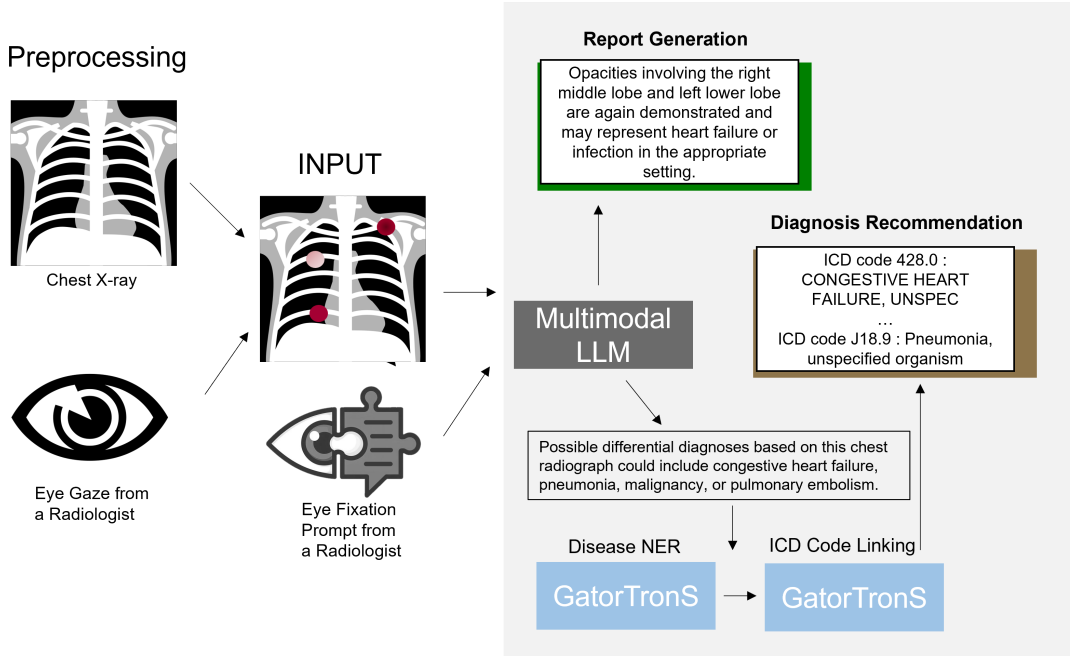


Figure 1: Overview of HITL multimodal LLM with Eye Gaze

2 Methods and Data

Figure 1 summarises the overview of the proposed HITL method for disease diagnosis recommendation and report generation tasks.

2.1 Eye Gaze Dataset and its Utilisation

We make use of the posterior to anterior (PA) view images sourced from the MIMIC-Eye dataset [12]. This dataset includes 3,689 chest X-ray images and we use the EyeGaze subset data as each image is paired with a list of diagnoses. Consequently, our evaluation dataset comes down to 994 chest X-ray images. The raw eye gaze data serves as the basis for generating heat maps on top of the CXR image. Each gaze point is drawn as a red dot and its opacity reflects the frequency of gazes. Also, the top 1 eye gaze fixation data is converted to textual prompts with the information of duration and relative position.

2.2 Diagnosis Recommendation

In the diagnosis recommendation task, the multimodal LLM suggests potential diagnoses based on the input chest X-ray frontal view image. These diagnoses typically involve categorization using codes like the International Statistical Classification of Diseases (ICD) codes from the World Health Organization (WHO). While LLMs could directly output ICD codes, we chose a cautious approach with named entity recognition (NER) and entity linking due to the risk of hallucinations. To refine the LLM's output, we finetuned a GatorTronS model with the BC5CDR and

NCBI-disease datasets for disease entity recognition [13, 14, 15]. Subsequently, we finetuned the GatorTronS model for entity linking using the SapBERT approach [16]. This workflow aims to accurately align extracted disease entities with their corresponding ICD codes, thereby ensuring the reliability of the diagnosis recommendations by the multimodal LLM.

The evaluation metric is the F1 score, calculated with diagnosis predictions at the ICD code level after disease entity recognition and alignment. Precision is calculated by dividing the number of correct predictions by the total number of predictions made, while recall is determined by comparing the correct predictions to the total number of relevant diseases for the patient. These recall and precision values are then utilized in the computation of the F1 score.

2.3 Report Generation

A standard radiology report includes a "Findings" section, which describes the observations made from the image. These reports are crucial for accurate diagnosis and treatment planning. The task of report generation aims to produce the "Findings" section based on the images. The evaluation metric for this task is METEOR [17], recognised for its robust assessment of semantic similarity and lexical variations using WordNet. This approach considers unigram mapping and introduces a penalty function for discrepancies in word order, thus providing a more nuanced evaluation. To compute the score, we employed HuggingFace's `evaluate` package.

2.4 Evaluation Model and Setting

We selected **CXR-LLaVA** for our study as it has been trained with various chest X-ray-related datasets including ones from MIMIC for report generation and differential diagnosis [4]. Also, its vision encoder is ViT-L patch 16 at resolution 224 trained with chest X-ray images.

For evaluation, we implemented a zero-shot approach for both tasks, a batch size of 1, and a temperature parameter of 0. A temperature of 0 was chosen to minimize the randomness in the generated text produced by the model. The maximum length of the model's responses for each task was determined based on the expected length of the response: 192 for the diagnosis recommendation and 320 for the report generation. This setting was done to maintain consistency and efficiency across our experiments.

3 Results

The evaluation results indicate that incorporating eye gaze information from radiologists does not contribute significantly to all the tasks. The only case it helped was the diagnosis recommendation. Specifically, utilizing eye gaze heat map image prompts led to a notable improvement, resulting in a 0.9 higher F1 score compared to the default baseline method. This finding underscores the potential efficacy of incorporating eye gaze information, even when the model was not initially trained with such additional data.

However, other prompts, including the top 1 fixation text prompt, showed a decrease in performance for diagnosis recommendation tasks. Even for report generation tasks, all the prompts showed a decrease in performance. Still, we observe a different order of prompting methods as the top 1 fixation prompt yielded the highest score, surpassing all other prompts except for the

baseline method. Future studies on other methods of incorporating gaze data will validate the effect of expertise inclusion.

Table 1: Evaluation Results. Diagnosis recommendation evaluation metric is F1. Report generation evaluation metric is METEOR.

Method	**Diagnosis Recommendation**	**Report Generation**
Baseline	5.5	28.5
Top 1 Fixation Prompt	5.0	27.8
Gaze Heat Map	6.4	27.7
Fixation+Gaze Heat Map	5.2	27.3

4 Conclusion

The results of our study reveal that eye gaze information exhibits varied impacts across different tasks within medical image analysis. Despite these variations, our findings underscore the potential utility of leveraging eye gaze data to enhance the performance of machine learning models in clinical decision support systems. Our study also aligns with our objective of exploring innovative approaches to improve the interpretability and effectiveness of AI-driven diagnostics in healthcare settings.

Moving forward, further research could delve deeper into the finetuning effect of eye gaze patterns as well as tailored model architecture for the radiologist's expertise. Additionally, future investigations could extend our work to include other modalities such as voice dictation data of the radiologist to further enhance the method of human-computer interaction. Exploring the applicability of radiologist's expertise data across diverse imaging modalities such as MRI or CT could also broaden the scope of our findings and enhance the generalizability of our approach.

Despite the promising outcomes, our study has certain limitations. For instance, our evaluation sample size is quite small, limiting the generalizability of our findings. Also, the number of models tested in the work is limited as there are very few models that are trained for both diagnosis recommendation and report generation. We plan to train our own model for these two tasks soon.

5 Study context

This study was conducted in strict compliance with the data usage agreements outlined by PhysioNet for the use of the MIMIC. Adhering to these agreements ensured that all patient data remained secure and confidential throughout the research process.

References

[1] Li Y, Liu Y, Wang Z, Liang X, Liu L, Wang L, et al. A Comprehensive Study of GPT-4V's Multimodal Capabilities in Medical Imaging. medRxiv. 2023:2023-11.

[2] OpenAI. GPT-4; 2023. Available from: https://www.openai.com/gpt-4.

[3] Team G, Anil R, Borgeaud S, Wu Y, Alayrac JB, Yu J, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:231211805. 2023.

[4] Lee S, Youn J, Kim M, Yoon SH. CXR-LLaVA: Multimodal Large Language Model for Interpreting Chest X-ray Images. arXiv preprint arXiv:231018341. 2023.

[5] Wu J, Kim Y, Keller EC, Chow J, Levine AP, Pontikos N, et al. Exploring Multimodal Large Language Models for Radiology Report Error-checking. arXiv preprint arXiv:231213103. 2023.

[6] Wu J, Kim Y, Wu H. Hallucination Benchmark in Medical Visual Question Answering. arXiv preprint arXiv:240105827. 2024.

[7] Yildirim N, Richardson H, Wetscherek MT, Bajwa J, Jacob J, Pinnock MA, et al. Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications for Radiology. arXiv preprint arXiv:240214252. 2024.

[8] Patel BN, Rosenberg L, Willcox G, Baltaxe D, Lyons M, Irvin J, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. npj Digital Medicine. 2019;2(1):111. Available from: https://doi.org/10.1038/s41746-019-0189-7.

[9] Calisto FM, Santiago C, Nunes N, Nascimento JC. BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions. Artificial Intelligence in Medicine. 2022;127:102285.

[10] Ma C, Zhao L, Chen Y, Wang S, Guo L, Zhang T, et al. Eye-gaze-guided vision transformer for rectifying shortcut learning. IEEE Transactions on Medical Imaging. 2023.

[11] Ji C, Du C, Zhang Q, Wang S, Ma C, Xie J, et al. Mammo-Net: Integrating Gaze Supervision and Interactive Information in Multi-view Mammogram Classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2023. p. 68-78.

[12] Hsieh C, Ouyang C, Nascimento JC, Pereira J, Jorge J, Moreira C. MIMIC-Eye: Integrating MIMIC Datasets with REFLACX and Eye Gaze for Multimodal Deep Learning Applications. 2023.

[13] Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digital Medicine. 2023;6(1):210.

[14] Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics. 2014;47:1-10.

[15] Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database. 2016;2016.

[16] Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. arXiv preprint arXiv:201011784. 2020.

[17] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization; 2005. p. 65-72.

Exploring Training Methods for Medical LLMs

Yunsoo Kim[1], Jinge Wu[1], and Honghan Wu[1]

[1]University College London, London, United Kingdom

1 Introduction

The large language models (LLMs) has shown outstanding ability to comprehend medical text and generate medically accurate information [1, 2]. This advancement has led to the emergence of chatbots with LLMs such as ChatGPT, offering unprecedented opportunities to facilitate clinical decision-making processes [3, 4, 5]. Moreover, the introduction of open-source medical LLMs has mitigated privacy concerns associated with patient data by opening the possibility of local implementation, thus broadening the scope of these technologies within hospital settings [6, 7, 8]. The training dataset used to train these medical LLMs remains relatively smaller when compared to that of general domain models. Previous studies have addressed this challenge through continued pretraining or supervised finetuning with instruction-tuning datasets. However, the specific impact of these different training methods on medical domain-specific performance remains understudied.

In our study, we propose to fill this gap by curating a dataset from MedlinePlus[1] and constructing pretraining (PT), supervised finetuning (SFT), and direct policy optimization (DPO) datasets to explore the training effect of these training methods on various tasks within the biomedical domain. MedlinePlus includes a medical encyclopedia and texts about drugs and genetics and has no privacy issue in MIMIC data or no positive research data issue seen in PubMed articles. In this preliminary study, we aim to highlight the optimal training strategies for maximizing performance in medical LLMs with a small training dataset.

2 Methods and Data

2.1 Models and Training Data

Llama2 We use Llama2 HuggingFace weights released on the Hugging Face model repository [9]. 7B model without chat optimization is used in this work. **Mistral** We use Mistral-7B-v0.1 weight released on the Hugging Face model repository [10]. The model size is known to be 7.24B parameters, and this is slightly larger than **Llama2-7B**, 6.74B. **Phi-2** We use Phi-2 model weight released on the Hugging Face model repository [11]. It has 2.78B parameters and is the smallest model in our paper.

[1]https://medlineplus.gov/

The PT dataset is the original MedlinePlus text data, the SFT dataset is based on the GPT4 reformatting of the MedlinePlus text data into question and answering data, and the DPO dataset further includes GPT4's response for the reformatted question as the rejected response.

2.2 Evaluation

For multiple-choice questions (**MCQ**), we use medical subjects within MMLU, which we refer to MMLU-Medical [12]. We also use MedQA[13] and MedMCQA [14] for medical license exam questions. For short answer question answering (**QA**), we used a subset of MedQuAD [15] and bioASQ [16]. For the document classification task (**CLS**), we used the LitCovid dataset [17]. For the summarisation task (**SUM**), we use MeQSum [18].

3 Results

The evaluation result suggests that pretraining improves the performance of the model in most of the cases. Mistral model for MCQ and Phi-2 model for QA did not show improvement in performance with any of the training methods we tested. Supervised finetuning and direct policy optimization enhanced the performance in summarisation and short answer question answering, while pre-training improved the performance of classification and MCQ tasks.

Table 1: Evaluation Result. Classification accuracy of the next token is used for MCQ. F1 score is used for CLS. ROUGE-L is used for QA and SUM [19]. For MCQ and SUM, we report the average score of multiple datasets.

Model	MCQ	SUM	CLS	QA
LLaMA2-7B	35.02	4.59	26.73	9.08
LLaMA2-7B+PT	**35.60**	5.30	**41.06**	10.71
LLaMA2-7B+SFT	33.93	**8.37**	30.96	**16.23**
LLaMA2-7B+DPO	22.47	7.04	26.89	12.42
Mistral	**53.81**	4.59	42.09	10.03
Mistral+PT	50.54	3.60	**46.61**	9.29
Mistral+SFT	29.61	7.92	7.10	**15.92**
Mistral+DPO	27.52	**8.00**	32.84	10.43
Phi2	40.77	3.68	35.30	**9.53**
Phi2+PT	**42.72**	3.48	**37.01**	9.26
Phi2+SFT	42.68	**10.66**	7.52	8.65
Phi2+DPO	35.26	3.18	33.67	7.53

4 Conclusion

The result suggests that different methods have a varied impact depending on the model and downstream tasks. Pre-training helps with deterministic tasks, while supervised fine-tuning and direct policy optimisation help with generative tasks. Still, as our study is tested with only one data source, we will resolve this limitation by collecting more training data and exploring the effect of training data size as well as the diversity of the training data. Additionally, future research will focus on investigating combinations of training methods to determine the synergistic effects.

5 Study context

There are no major ethical concerns raised in this work. All the LLMs were used for research purpose only.

References

[1] Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, et al. Towards generalist biomedical AI. arXiv preprint arXiv:230714334. 2023.

[2] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:230509617. 2023.

[3] OpenAI. ChatGPT; 2023. Available from: `https://chat.openai.com/chat`.

[4] Team G, Anil R, Borgeaud S, Wu Y, Alayrac JB, Yu J, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:231211805. 2023.

[5] Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digital Medicine. 2024;7(1):20.

[6] Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. Clinical Camel: An Open-Source Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. arXiv preprint arXiv:230512031. 2023.

[7] Kweon S, Kim J, Kim J, Im S, Cho E, Bae S, et al. Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes. arXiv preprint arXiv:230900237. 2023.

[8] Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al. Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:231116079. 2023.

[9] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.

[10] Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, et al. Mistral 7B. arXiv preprint arXiv:231006825. 2023.

[11] Microsoft. Phi-2: The surprising power of small language models;. Available from: `https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/ (accessed8Feburary2024)`.

[12] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring massive multitask language understanding. arXiv preprint arXiv:200903300. 2020.

[13] Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences. 2021;11(14):6421.

[14] Pal A, Umapathi LK, Sankarasubbu M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Conference on Health, Inference, and Learning. PMLR; 2022. p. 248-60.

[15] Ben Abacha A, Demner-Fushman D. A question-entailment approach to question answering. BMC bioinformatics. 2019;20:1-23.

[16] Krithara A, Nentidis A, Bougiatiotis K, Paliouras G. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. Scientific Data. 2023;10(1):170.

[17] Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic acids research. 2021;49(D1):D1534-40.

[18] Abacha AB, Demner-Fushman D. On the summarization of consumer health questions. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 2228-34.

[19] Lin CY. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out; 2004. p. 74-81.

Automated Generation of Hospital Discharge Summaries Using Clinical Guidelines and Large Language Models

Simon Ellershaw[1], Christopher Tomlinson[1,2,3], Oliver Burton[4], Thomas Frost[1], John Gerrard Hanrahan[4,5], Danyal Z Khan[4,5], Hugo Layard Horsfall[4,5], Mollie Little[4], Evaleen Malgapo[1], Joachim Starup-Hansen[4], Jack Ross[4], George Woodward[4], Martinique Vella-Baldacchino[6], Kawsar Noor[1,2,3], Anoop D Shah[1,2,3], and Richard JB Dobson[1,2,3,7]

[1]Institute of Health Informatics, University College London, UK
[2]National Institute for Health and Care Research Biomedical Research Centre, University College London Hospitals National Health Service Foundation Trust, UK
[3]Health Data Research UK, UK
[4]University College London Hospitals NHS Foundation Trust, UK
[5]Wellcome/EPSRC Centre for Interventional and Surgical Sciences, University College London, UK
[6]MSK Lab, Imperial College London, UK
[7]Department of Biostatistics and Health Informatics, King's College London, UK

1 Introduction

A clinician must write a discharge summary at the end of every patient's hospital stay. The summary communicates to the post-hospital care team what has happened to the patient during their hospital stay and their ongoing care plan [1]. However, this manual process adds to clinicians' workloads and can be of varying quality [2].

Therefore, the automation of this process using machine learning models has been proposed as a solution [3]. Current state-of-the-art approaches [4] fine-tune encoder-decoder models [5] to map a set of clinician notes to a discharge summary. However, this supervised approach faces challenges due to the limited training data, extended length of clinician notes and variable ground truth quality [6].

Recently, the scaling of the training and size of natural language auto-regressive transformers has led to a new class of models known as large language models (LLMs) [7]. LLMs have shown the ability to learn from a few examples, accept inputs over 100,000 words and attain state-of-the-art performance on several benchmark tasks, including text summarisation [8, 9]. Such model properties could solve several problems currently faced in the automatic generation of discharge summaries.

This work presents an LLM-based discharge summary generator tested on full clinical notes and evaluated by clinicians. Our key contribution is the use of clinical guidelines to prompt the

LLM with the desired format and content of a summary instead of learning this from the data.

This aligns with the first author's PhD investigating potential use cases for LLMs in healthcare. The aim of this initial study was to show the capabilities of this new class of models to perform a real-world clinical task that was out of the scope of the previous generation of natural language processing approaches.

2 Methods and Data

We converted guidelines from the UK's Royal College of Physicians London (RCP) [10], see Fig 2, to a JSON schema. We excluded the medication section, which requires the non-trivial merging of structured e-prescribing data with the extraction of the reasons for any medication changes from the clinical notes.

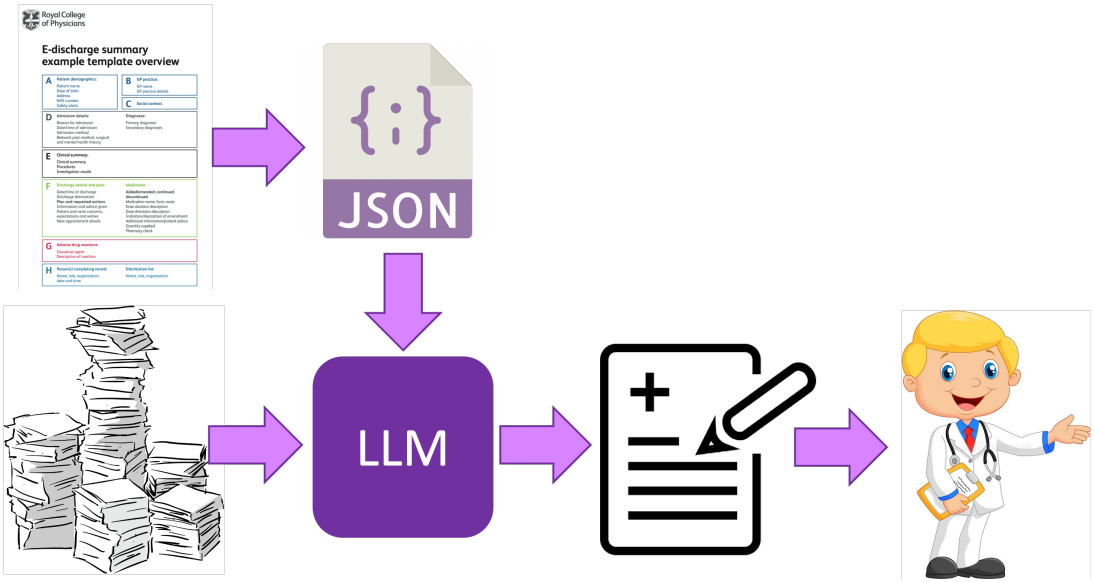


Figure 1: Shows the proposed method, which combines discharge summary guidelines and physician notes into an LLM prompt in order to produce a discharge summary for review by a clinician.

Following this, we created a fixed prompt of a system message containing the JSON schema and a one-shot example generated from an exemplar RCP discharge summary [10]. For full details of this process see Appendix 2.

To test the efficacy of the method, we used the freely-available MIMIC-III v1.4 dataset [11, 12, 13]. We filtered the notes table for hospital admissions for which a discharge summary exists and so could be generated. For these admissions, we concatenated and time-ordered the patient's physician notes. Other note sources, such as nursing notes and imaging reports, were not included to limit the input size. Extraneous characters and artefacts from the anonymity process were then removed and the note was deduplicated by keeping only the first occurrence of a line of text.

For our experiments, we used GPT-4-turbo version 1106-Preview [14], with temperature=0, due to its strong benchmark performance [8] and 128k context window, which allowed all sets

of tested physician notes to be accepted in a single query.

One round of qualitative evaluation was performed with a clinician using a sample of 5 hospital admissions. We used this feedback to adjust the description of a select number of fields. For a complete list see Table 2.

We evaluated the final system using a team of 11 UK-qualified doctors and physician associates with prior experience writing discharge summaries. After reading the physician notes and clinical guidelines, the clinicians were asked to evaluate the number of times the following errors occurred for each discharge summary field: missed severe, missed minor, additional hallucination and additional not relevant. A missed error was categorised as severe if it had the potential to meet the NHS England [15] definition of medium to severe levels of harm. Each clinician evaluated five summaries, of which one was duplicated with another clinician to allow the calculation of inter-annotator agreement.

3 Results

53 discharge summaries were generated and evaluated. The median input physician notes length after de-duplication was 4996 tokens and the fixed prompt was 5057 tokens, measured using the cl100k_base tokeniser [16]. The median inference time was 40.59s at a median API cost of $0.12. The model extracted 25.07% of the generated elements verbatim from the input physician notes. For a further breakdown of these metrics, see Table 3.

We found the median number of errors per summary to be 7, with the error proportions to be 36.28% missed severe, 27.44% missed minor, 14.55% added hallucination and 21.73% added not relevant. One summary failed to conform to the JSON schema. We calculated the percentage agreement between annotators, see Eqn 8, to be 59.72%.

To calculate the performance metrics in Table 1, we used Equations 1-7, defining a missed error as a false negative and an addition error as a false positive. Table 4 shows a per-field view of the same results. The GP Practice section is excluded from the analysis, as the GP is not a role in the American healthcare system, and so the section was never filled.

Section	Recall	Precision	Acc
Admission Details	0.90	0.95	0.85
Allergies And Adverse Reaction	0.98	1.00	0.98
Clinical Summary	0.76	0.92	0.71
Diagnoses	0.84	0.94	0.80
Discharge Details	0.93	0.96	0.89
Patient Demographics	1.00	0.84	0.84
Plan And Requested Actions	0.90	0.88	0.80
Social Context	0.96	0.88	0.84
Macro Average	0.91	0.92	0.84
Micro Average	0.86	0.92	0.81

Table 1: Recall, precision and accuracy metrics per section for discharge summaries generated from MIMIC-III notes as evaluated by clinicians.

The correlation of input note and generated summaries token length with respect to accuracy is $r_s = -0.23$ ($p = 0.1$) and $r_s = -0.15$ ($p = 0.31$) respectively. Where r_s is the Spearman's

correlation coefficient chosen due to the non-normal distribution of the variables [17]. In both cases, this shows a weak negative correlation, visualised in Fig 4, but the null hypothesis that this is due to random chance cannot be rejected.

4 Conclusion

While the metrics in Table 1 show promise for many fields, safety-critical errors, such as missed severe and hallucinations, highlight the challenges in using LLMs for discharge summarisation and the need for clinician-in-the-loop review at the point of use. However, this in turn poses the risk of automation bias arising over time.

Notably, a strength of LLMs lies in their understanding and ability to manipulate text. Hence, the model's performance on extraction tasks such as listing allergies is near perfect. However, tasks requiring clinical judgment are more challenging. For example, the RCP guidelines state that "only investigations which the GP is likely to monitor either of the health condition or associated with medication" are to be included in the summary. This subjective inclusion criteria led to inconsistent recording of similar tests.

Future work to improve the methodology presented could include further rounds of clinician-led prompt iteration. This may fix error cases such as the miss-classification of conditions between the social context, past medical history, and secondary diagnosis fields. Also, improved prompting could improve the differentiation between the last actions to take in hospital and post-hospital care. Furthermore, although the size of LLM's context windows has dramatically increased, it is possible that a retrieval augmented generation approach could lead to improvements in accuracy and generation speed on longer sets of notes.

The evaluation of this work was limited to a single centre's ICU data due to data availability, in scale due to the labour-intensive nature of clinical evaluation and the low inter-annotator agreement metric, which shows the variability of clinical review for this task. Therefore, the development of a clinically grounded, scalable and systematically repeatable evaluation framework is vital for future work.

A current barrier to adoption is the need for a closed-source third-party LLM to comply with a healthcare provider's data governance framework or appropriate computing infrastructure to host an open-source LLM on-premises.

The key strength of this work is that, to the author's knowledge, it is the first to show the effectiveness of using clinical guidelines to prompt LLMs for administrative medical tasks, such as discharge summarisation. This overcomes the main limitations of supervised approaches, namely the need for large labelled datasets and the inherent biases encoded in training on real-world data of variable quality.

5 Social Context

The collection of patient information and creation of the MIMIC-III research resource was previously reviewed by the Institutional Review Board at the Beth Israel Deaconess Medical Center, which granted a waiver of informed consent and approved the data-sharing initiative [11]. No additional specific ethics board approval was required for this project. Access to the MIMIC-

III dataset requires an approval process, including mandatory data ethics training. All authors, including clinical evaluators, undertook this process.

The PhysioNet Credentialed Data Use Agreement [18], which governs the use of the MIMIC-III dataset, explicitly prohibits sharing access to the data with third parties. Therefore, in line with MIMIC's guidance on the use of third-party LLMs [19] all GPT-4 queries were made using Azure OpenAI service whilst being opted out of the human review of the data.

Concerning reproducibility, we cannot openly share the generated summaries and evaluations due to the terms of the MIMIC dataset license. However, the code to produce the summaries is open-sourced (`https://github.com/simonEllershaw/llm-discharge-summaries`), allowing a MIMIC-credentialed user to reproduce the summaries evaluated in this work. Similarly, all data analysis scripts are also released.

SE is supported by a UCL UKRI Centre for Doctoral Training in AI-enabled Healthcare studentship (EP/S021612/1). ADS is supported by research grants from EPSRC (EP/Y018087) and NIHR (AI_AWARD01864). CT is supported by a UCL UKRI Centre for Doctoral Training in AI-enabled Healthcare studentship (EP/S021612/1), a MRC Clinical Top-Up, a studentship from the NIHR Biomedical Research Centre at University College London Hospital NHS Trust, and the Health Data Research UK Phenomics and Prognostic Atlas Theme.

References

[1] Kind AJ, Smith MA. Documentation of Mandated Discharge Summary Components in Transitions from Acute to Subacute Care. Agency for Healthcare Research and Quality (US), Rockville (MD); 2008. Available from: `http://europepmc.org/books/NBK43715`.

[2] Rattray NA, Sico JJ, Cox LM, Russ AL, Matthias MS, Frankel RM. Crossing the communication chasm: challenges and opportunities in transitions of care from the hospital to the primary care clinic. The Joint Commission Journal on Quality and Patient Safety. 2017;43(3):127-37.

[3] Patel SB, Lam K. ChatGPT: the future of discharge summaries? The Lancet Digital Health. 2023;5(3):e107-8.

[4] Pal K, Bahrainian SA, Mercurio L, Eickhoff C. Neural Summarization of Electronic Health Records. arXiv preprint arXiv:230515222. 2023.

[5] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:191013461. 2019.

[6] Searle T, Ibrahim Z, Teo J, Dobson RJ. Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models. Journal of Biomedical Informatics. 2023;141:104358.

[7] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.

[8] Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M, et al. Holistic evaluation of language models. arXiv preprint arXiv:221109110. 2022.

[9] Anthropic. Model Card and Evaluations for Claude Models; 2023. Accessed: 2024-01-01. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

[10] Royal College of Physicians. Improving discharge summaries – learning resource materials; 2021. Accessed: 2023-12-28. https://www.rcplondon.ac.uk/guidelines-policy/improving-discharge-summaries-learning-resource-materials.

[11] Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3(1):1-9.

[12] Johnson A, Pollard T, Mark R. MIMIC-III Clinical Database (version 1.4); 2016. Accessed: 2024-01-02. PhysioNet. https://doi.org/10.13026/C2XW26.

[13] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. circulation. 2000;101(23):e215-20.

[14] OpenAI. New models and developer products announced at DevDay; 2023. Accessed: 2023-12-28. https://openai.com/blog/new-models-and-developer-products-announced-at-devday.

[15] NHS England National Patient Safety Team. Policy guidance on recording patient safety events and levels of harm; 2023. Accessed: 2023-12-28. https://www.england.nhs.uk/long-read/policy-guidance-on-recording-patient-safety-events-and-levels-of-harm/.

[16] OpenAI. https://github.com/openai/tiktoken/tree/main; 2021. Accessed: 2024-01-27. https://github.com/openai/tiktoken/tree/main.

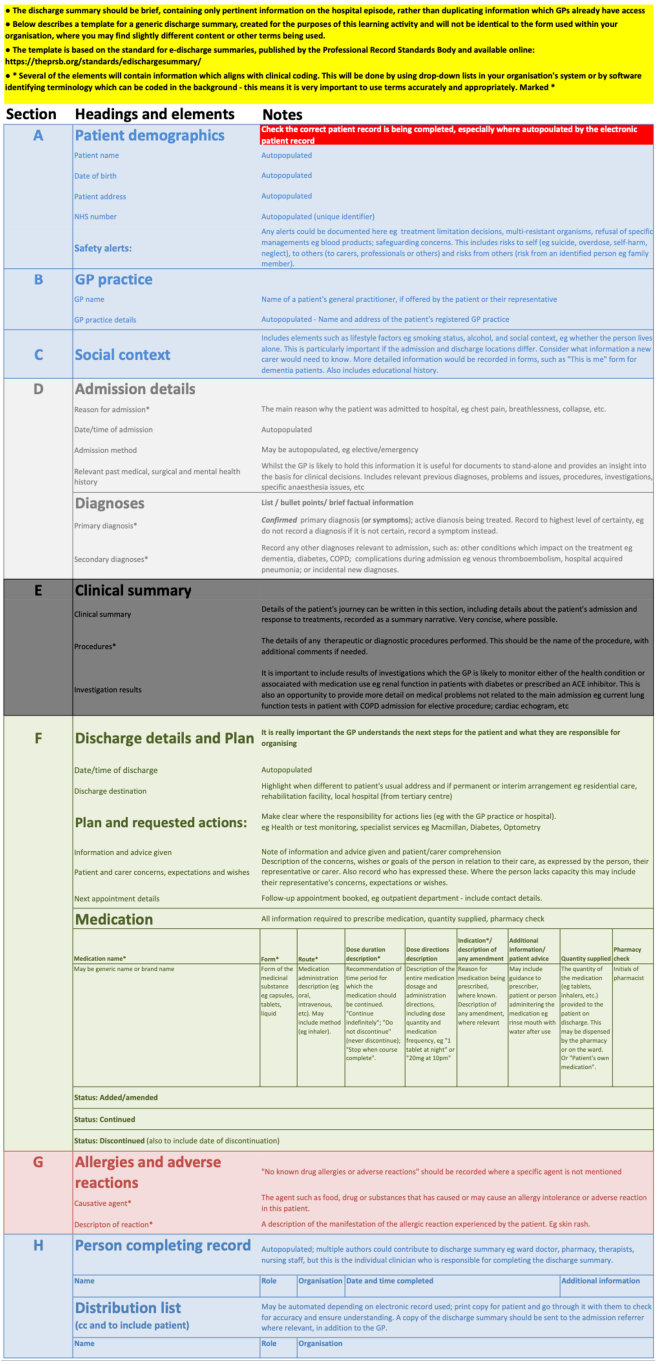
[17] Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal. 2012;24(3):69-71.

[18] PhysionNet. PhysioNet Credentialed Health Data License 1.5.0; 2023. Accessed: 2024-05-01. https://physionet.org/content/mimiciii/view-license/1.4/.

[19] PhysionNet. Responsible use of MIMIC data with online services like GPT; 2023. Accessed: 2024-05-01. https://physionet.org/news/post/415.

[20] OpenAI. Text generation models- Chat Completions; 2023. Accessed: 2024-10-01. https://platform.openai.com/docs/guides/text-generation/chat-completions-api.

A Appendix 1- Royal College Of Physician Guidelines



- The discharge summary should be brief, containing only pertinent information on the hospital episode, rather than duplicating information which GPs already have access
- Below describes a template for a generic discharge summary, created for the purposes of this learning activity and will not be identical to the form used within your organisation, where you may find slightly different content or other terms being used.
- The template is based on the standard for e-discharge summaries, published by the Professional Record Standards Body and available online: https://theprsb.org/standards/edischargesummary/
- * Several of the elements will contain information which aligns with clinical coding. This will be done by using drop-down lists in your organisation's system or by software identifying terminology which can be coded in the background - this means it is very important to use terms accurately and appropriately. Marked *

Section	Headings and elements	Notes
A	**Patient demographics**	Check the correct patient record is being completed, especially where autopoulated by the electronic patient record
	Patient name	Autopopulated
	Date of birth	Autopopulated
	Patient address	Autopopulated
	NHS number	Autopopulated (unique identifier)
	Safety alerts:	Any alerts could be documented here eg treatment limitation decisions, multi-resistant organisms, refusal of specific managements eg blood products; safeguarding concerns. This includes risks to self (eg suicide, overdose, self-harm, neglect), to others (to carers, professionals or others) and risks from others (risk from an identified person eg family member).
B	**GP practice**	
	GP name	Name of a patient's general practitioner, if offered by the patient or their representative
	GP practice details	Autopopulated - Name and address of the patient's registered GP practice
C	**Social context**	Includes elements such as lifestyle factors eg smoking status, alcohol, and social context, eg whether the person lives alone. This is particularly important if the admission and discharge locations differ. Consider what information a new carer would need to know. More detailed information would be recorded in forms, such as "This is me" form for dementia patients. Also includes educational history.
D	**Admission details**	
	Reason for admission*	The main reason why the patient was admitted to hospital, eg chest pain, breathlessness, collapse, etc.
	Date/time of admission	Autopopulated
	Admission method	May be autopopulated, eg elective/emergency
	Relevant past medical, surgical and mental health history	Whilst the GP is likely to hold this information it is useful for documents to stand-alone and provides an insight into the basis for clinical decisions. Includes relevant previous diagnoses, problems and issues, procedures, investigations, specific anaesthesia issues, etc
	Diagnoses	**List / bullet points/ brief factual information**
	Primary diagnosis*	***Confirmed*** primary diagnosis (**or symptoms**); active dianosis being treated. Record to highest level of certainty, eg do not record a diagnosis if it is not certain, record a symptom instead.
	Secondary diagnoses*	Record any other diagnoses relevant to admission, such as: other conditions which impact on the treatment eg dementia, diabetes, COPD; complications during admission eg venous thromboembolism, hospital acquired pneumonia; or incidental new diagnoses.
E	**Clinical summary**	
	Clinical summary	Details of the patient's journey can be written in this section, including details about the patient's admission and response to treatments, recorded as a summary narrative. Very concise, where possible.
	Procedures*	The details of any therapeutic or diagnostic procedures performed. This should be the name of the procedure, with additional comments if needed.
	Investigation results	It is important to include results of investigations which the GP is likely to monitor either of the health condition or assocaiated with medication use eg renal function in patients with diabetes or prescribed an ACE inhibitor. This is also an opportunity to provide more detail on medical problems not related to the main admission eg current lung function tests in patient with COPD admission for elective procedure; cardiac echogram, etc
F	**Discharge details and Plan**	**It is really important the GP understands the next steps for the patient and what they are responsible for organising**
	Date/time of discharge	Autopopulated
	Discharge destination	Highlight when different to patient's usual address and if permanent or interim arrangement eg residential care, rehabilitation facility, local hospital (from tertiary centre)
	Plan and requested actions:	Make clear where the responsibility for actions lies (eg with the GP practice or hospital). eg Health or test monitoring, specialist services eg Macmillan, Diabetes, Optometry
	Information and advice given	Note of information and advice given and patient/carer comprehension
	Patient and carer concerns, expectations and wishes	Description of the concerns, wishes or goals of the person in relation to their care, as expressed by the person, their representative or carer. Also record who has expressed these. Where the person lacks capacity this may include their representative's concerns, expectations or wishes.
	Next appointment details	Follow-up appointment booked, eg outpatient department - include contact details.
	Medication	All information required to prescribe medication, quantity supplied, pharmacy check

Medication name*	Form*	Route*	Dose duration description*	Dose directions description	Indication*/ description of any amendment	Additional information/ patient advice	Quantity supplied	Pharmacy check
May be generic name or brand name	Form of the medicinal substance eg capsules, tablets, liquid	Medication administration description (eg oral, intravenous, etc). May include method (eg inhaler).	Recommendation of time period for which the medication should be continued. "Continue indefinitely"; "Do not discontinue" (never discontinue); "Stop when course complete".	Description of the entire medication dosage and administration directions, including dose quantity and medication frequency, eg "1 tablet at night" or "20mg at 10pm"	Reason for medication being prescribed, where known. Description of any amendment, where relevant	May include guidance to prescriber, patient or person administering the medication eg rinse mouth with water after use	The quantity of the medication (eg tablets, inhalers, etc.) provided to the patient on discharge. This may be dispensed by the pharmacy or on the ward. Or "Patient's own medication".	Initials of pharmacist
Status: Added/amended								
Status: Continued								
Status: Discontinued (also to include date of discontinuation)								

G	**Allergies and adverse reactions**	"No known drug allergies or adverse reactions" should be recorded where a specific agent is not mentioned
	Causative agent*	The agent such as food, drug or substances that has caused or may cause an allergy intolerance or adverse reaction in this patient.
	Descripton of reaction*	A description of the manifestation of the allergic reaction experienced by the patient. Eg skin rash.
H	**Person completing record**	Autopopulated; multiple authors could contribute to discharge summary eg ward doctor, pharmacy, therapists, nursing staff, but this is the individual clinician who is responsible for completing the discharge summary.

Name	Role	Organisation	Date and time completed	Additional information
Distribution list (cc and to include patient)	May be automated depending on electronic record used; print copy for patient and go through it with them to check for accuracy and ensure understanding. A copy of the discharge summary should be sent to the admission referrer where relevant, in addition to the GP.			
Name	**Role**	**Organisation**		

Figure 2: A copy of the RCP crib sheet outlining their guidelines for discharge summary writing [10].

A Appendix 2-LLM Prompt

To form the LLM prompt, firstly, we take guidelines written by the RCP, see Fig 2, [10] and using the title and description of each section convert this to a JSON schema shown in Listing 1. We excluded the medication section, which requires the non-trivial merging of structured e-prescribing data with the extraction of the reasons for any medication changes from the clinical notes. The schema's required and title fields are redundant and removed to reduce input length.

Listing 1: RCP-based discharge summary JSON schema. For presentation purposes, only the patient_demographics section is shown.

```
{
  "description": "The discharge summary should be brief, containing only
      pertinent information on the hospital episode, rather than
      duplicating information which GPs already have access to in their
      own records.",
  "type": "object",
  "properties": {
    "patient_demographics": {
      "$ref": "#/definitions/PatientDemographics"
    },
    ...
  }
  "definitions": {
    "AdmissionDetails": {
      "type": "object",
      "properties": {
        "reason_for_admission": {
          "description": "The main reason why the patient was admitted to
              hospital, eg chest pain, breathlessness, collapse, etc. This
               should be symptoms and not the diagnosis.",
          "type": "string"
        },
        "admission_method": {
          "description": "Eg elective/emergency",
          "type": "string"
        },
        "relevant_..._history": {
          "description": "Whilst the GP is likely to hold this information
               it is useful for documents to stand-alone and provides an
              insight into the basis for clinical decisions. Includes
              relevant previous diagnoses, problems and issues, procedures
              , investigations, specific anaesthesia issues, etc",
          "type": "array",
          "items": {
            "type": "string"
          }
        }
      }
    },
    ...
  }
}
```

Next, we convert an exemplar discharge summary from the RCP guidelines to JSON according to the schema. The accompanying physician notes are formatted and de-duplicated using the same method as outlined in the methodology sections for the MIMIC physician notes. Together, the RCP JSON schema, one-shot prompt and set of input physician notes form the input prompt, as shown in Fig 3.

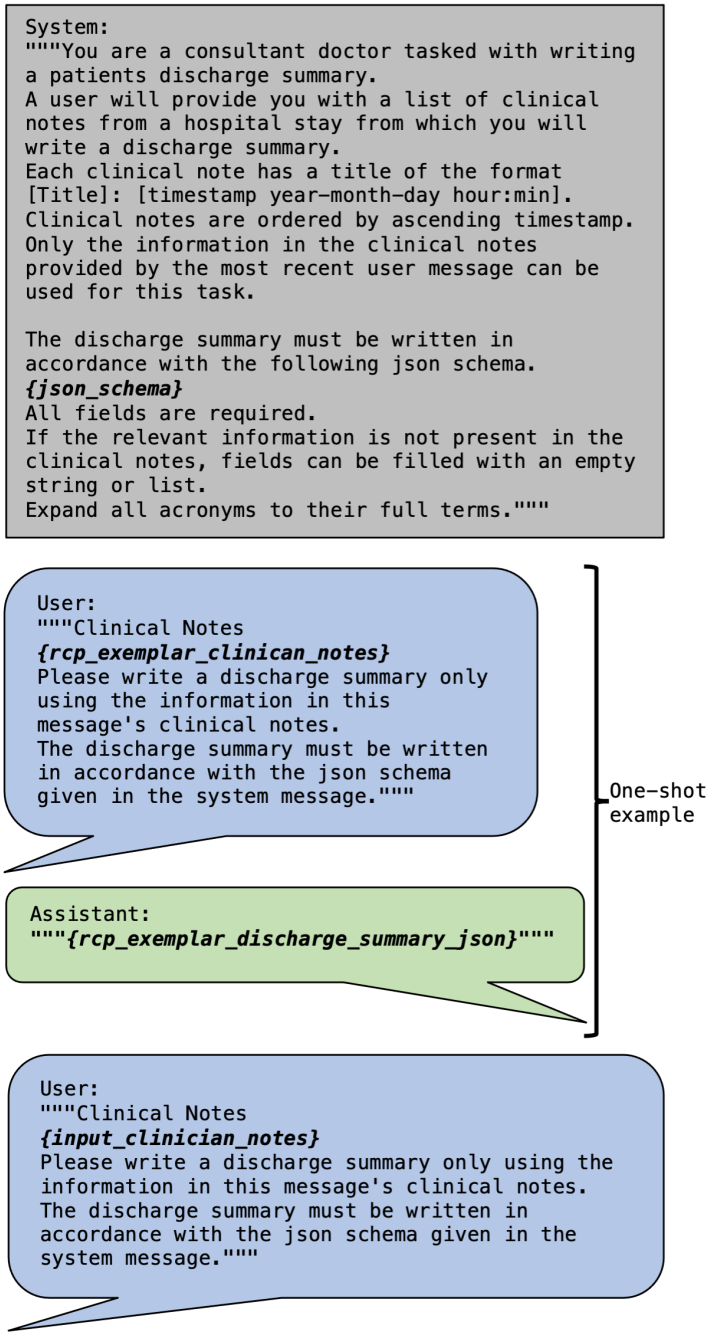


Figure 3: The GPT-4-turbo [14] prompt used in this work. Contained in bold braces are the variables produced by the processes outlined in the methodology section. System, user and assistant refer to the different roles used by OpenAI's chat completions API [20].

Section	Field	Change to Description
Admission Details	Reason For Admission	Added- “This should be symptoms and not the diagnosis."
	Admission Method	Remove- “May be autopopulated"
Diagnoses	Secondary Diagnoses	Added- “Do not include diagnoses made before this hospital admission."
Clinical Summary	Procedures	Added- “Do not include procedures performed before this hospital admission."
	Investigation Results	Added- “, chest x-ray, mri scan, etc. Each investigation is a separate element in the list."
PlanAndRequestedActions	Post Discharge Plan and Requested Actions	Added- Do not include jobs that are still to be done in hospital before discharge."
	Next Appointment Details	Added- "Note date and contact details if available."

Table 2: A table showing the alterations made to the field descriptions of the RCP discharge summary JSON schema after 1 round of clinical evaluation.

B Appendix 3- Metric Equations

In order to calculate the performance metrics shown in Tables 1 and 4, we first defined the evaluation of each field as a 4-dimensional vector (sum missing severe errors, sum missing minor errors, sum additional hallucination errors, sum additional not relevant errors).

From this definition we calculated the number of additional errors for a given field f summed across all generated summaries as the number of false positives, FP_f and likewise for missing errors and false negatives FN_f. The number of positive predictions for a field, P_f, is defined as either the length of list type fields or the number of sentences for string type fields. Therefore, the number of true positives, TP_f, for a field f is

$$TP_f = P_f - FP_f \tag{1}$$

From this and given that true negatives do not exist in this framework, the field's precision, p_f, recall, r_f, F1, $F1_f$ and accuracy, acc_f scores, can be calculated,

$$p_f = \frac{TP_f}{TP_f + FP_f}, \tag{2}$$

$$r_f = \frac{TP_f}{TP_f + FN_f}, \tag{3}$$

$$F1_f = 2 \times \frac{p_f \times r_f}{p_f + r_f}, \tag{4}$$

$$acc_f = \frac{TP_f}{TP_f + FP_f + FN_f}. \tag{5}$$

We found the average precision scores by averaging across all fields

$$p_{macro} = \frac{1}{|p|} \sum_f p_f. \tag{6}$$

Or by first pooling across fields

$$p_{micro} = \frac{\sum_f TP_f}{\sum_f TP_f + \sum_f FP_f}. \tag{7}$$

Similar equations hold for averaging recall, F1 and accuracy.

To calculate the inter-annotator agreement for the set of all doubly evaluated field, f, we defined two 2-D vector (FN_{f1}, FP_{f1}) and (FN_{f2}, FP_{f2}) one for each evaluator. FN and FP were chosen as they are the evaluation defined inputs to Eqn 7. A_o was then calculated as

$$A_o = \frac{\sum_f \delta\{(FN_{f1}, FP_{f1}), (FN_{f1}, FP_{f1})\}}{|f|} \tag{8}$$

where the δ function is defined as

$$\delta_{a,b} = \begin{cases} 1, a = b \\ 0, a \neq b. \end{cases} \tag{9}$$

C Appendix 4- Additional Results

	Percentile			
	25th	50th	75th	Max
De-Duplicated Physician Note Length / Tokens	2793	4996.	8772	95682
Output Note Length / Tokens	705	807	884	1234
Inference Time / secs	33.41	40.60	48.61	125.95
Inference Cost / $	0.10	0.12	0.16	1.04

Table 3: Table of system properties when tested on MIMIC-III notes. The fixed prompt length is 5057 tokens. We calculated token lengths using cl100k_base tokenizer [16]

.

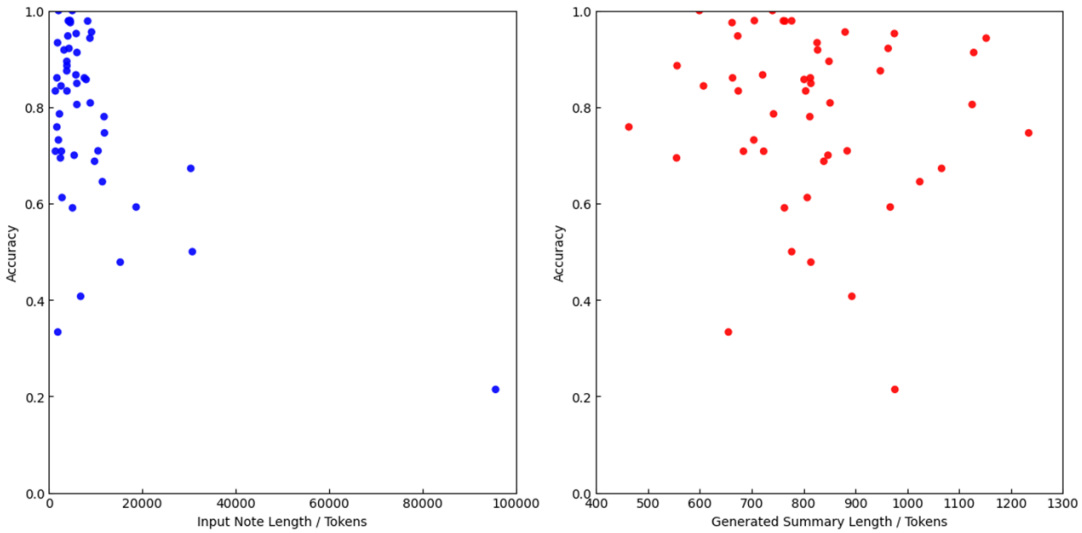


Figure 4: Scatter plots showing the weak negative correlation of input note and generated summary length with accuracy. We calculated token lengths using cl100k_base tokenizer [16]

Section	Field	Mean Number of Elements	Proportion of Blank Values	Recall	Precision	F1	Acc
Admission Details	Admission Method	1.00	0.00	0.93	0.96	0.94	0.89
	Reason For Admission	1.00	0.00	0.79	0.92	0.85	0.74
	Relevant Past Medical And Mental Health History	8.34	0.08	0.91	0.95	0.93	0.87
Allergies And Adverse Reaction	Causative Agent	1.87	0.00	0.98	1.00	0.99	0.98
	Description Of Reaction	1.87	0.09	0.98	1.00	0.99	0.98
Clinical Summary	Clinical Summary	4.28	0.00	0.71	0.98	0.82	0.70
	Investigation Results	4.30	0.04	0.75	0.86	0.80	0.67
	Procedures	2.36	0.28	0.87	0.94	0.91	0.83
Diagnoses	Primary Diagnosis	1.00	0.00	0.83	0.94	0.88	0.79
	Secondary Diagnoses	3.45	0.13	0.84	0.94	0.89	0.80
Discharge Details	Discharge Destination	1.00	0.00	0.93	0.96	0.94	0.89
Patient Demographics	Safety Alerts	1.74	0.72	1.00	0.84	0.91	0.84
Plan And Requested Actions	Information And Advice Given	1.40	0.55	0.98	0.80	0.88	0.79
	Next Appointment Details	1.00	0.72	1.00	0.89	0.94	0.89
	Patient And Carer Concerns Expectations And Wishes	1.25	0.62	0.89	0.83	0.86	0.75
	Post Discharge Plan And Requested Actions	7.89	0.00	0.88	0.90	0.89	0.80
Social Context	Social Context	2.89	0.17	0.96	0.88	0.91	0.84
Macro Average				0.90	0.92	0.90	0.83
Micro Average				0.86	0.92	0.89	0.81

Table 4: Evaluation metrics per discharge summary field, including mean number of elements and proportion of blank values per field as well as recall, precision, F1 and accuracy.

ArcTEX – a precise clinical data enrichment model to support real world evidence studies

Joseph Cronin, Keiran Tait, Jamie Wallis, Robert Dürichen
Arcturis Data, Oxford, United Kingdom

Introduction

Many leading pharmaceutical companies enhance their clinical development and post-market launch strategies through the integration of real-world data. Often these real-world evidence studies depend on the availability of several specific biomarker values. However, a significant hurdle lies in the fact that often this information resides within unstructured textual formats, impeding direct accessibility for analysis. Further, not all hospital environments are comfortable sharing redacted clinical notes due to the risk of revealing personally identifiable information, making analysis in the hospital environment preferable. However, this is often restricted by a resource constrained environment. To overcome this, we developed ArcTEX (Arcturis Text Enrichment and EXtraction), a light-weighted question-answering (QA) model to extract biomarker information from unstructured clinical reports. Compared to other baseline models, we demonstrate that ArcTEX can extract disease relevant information, including biomarker results from unstructured pathology reports. Further advantages of the approach are 1) high flexibility, as the model requires only a few training samples to be adapted to other biomarkers, 2) high robustness, as confidence scores can be used to identify misclassified samples, and 3) the model can be executed on CPUs and ensures that no patient identifiable information can be extracted which makes it ideal for usage in hospital environments.

Methods and Data

Data: The used dataset consists of 77,693 anonymised English pathology reports from Oxford University Hospital Foundation Trust (OUH). Reports were available in digital form; no OCR was required. They originate from patients with at least one of the following seven oncology areas: lung, pancreatic, renal, breast, ovarian, endometrial, or liver. The length of the reports varies between 20-4015 characters (average: 1084). A subset of 243 reports were annotated for 14 biomarkers (e.g. *p53, er, pr, her2, mmr, tumour grade*), to which we refer to as the finetuning dataset. Around 200 reports were annotated for each biomarker additionally as extended validation & test (EVT) datasets.
Methodology: The approach consists of 2 stages, a QA stage, and a biomarker specific classification stage. Different models have been evaluated as QA model, including RoBERTa[1], BioBERT [2], PubMedBERT [3], Flan-T5 [4], or Falcon [5]. Based on initial results, a BioBERT model finetuned on squad-QA dataset [6] was selected as an acceptable compromise between accuracy and computational complexity. The model was further optimised on the finetuning dataset. In contrast to generative approaches, the model returns only a subsection of the original text containing the most relevant answer to the question. Stage 2 classifies answers to a biomarker specific class by converting the answers into an embedding vector using a sentence embedding (SE) model and classifying the class using an additional classification head. The outcome of stage 2 is only the predicted class, ensuring that no patient data can be leaked, even if it was present in the extracted text of stage 1. The approach requires, for each biomarker, a label file listing possible classes and a few example answers for each class. The SE model is trained using the *setfit* approach [7], which uses a contrastive learning scheme. Additionally, the SE model is optimised through an unsupervised domain adaptation stage using denoising autoencoders as proposed by Wang et al. [8] on all available reports. We refer to the ArcTEX model using a finetuned BioBERT model in stage 1 and a domain adapted *setfit* classifier in stage 2. Other baseline methods use different QA models in stage 1 and a non-domain adapted *setfit* classifier.

Evaluation: The approach is evaluated in 2 experiments (*exp*). *Exp1*: Aims to investigate the impact of different QA models, finetuning of the QA model, and the relevance of the unsupervised pretraining (domain adaption) of the SE model (stage 2) on the prediction accuracy. Biomarker p53 was used as an example. Evaluation was done on 50 randomly selected reports as a test set from the EVT dataset. We simulated an iterative human-in-the-loop approach to investigate by how much the model performance could be further increased. Therefore, the model was also evaluated on a validation set (50 randomly selected reports of the EVT dataset distinct from the test set). The worst 5 classified samples (either misclassified and/or samples with the lowest confidence score) were added to the training set in the next iteration (the validation set was replenished by 5 randomly selected examples of the remaining EVT dataset). This procedure was repeated for 5 iterations, adding up to 25 additional training samples. The whole evaluation was repeated 10 times to estimate the robustness of the approach. *Exp2:* Investigates the performance of the approach across 5 different biomarkers (P53, MSH6, Grade, FIGO, MMR) and tumour grading for a BERT model and the ArcTEX model using the same evaluation scheme as in exp1.

Results

Fig. 1 shows the mean accuracy and standard deviation of the different models for exp 1. Results indicate that mean accuracy for the different non-finetuned, non-domain adapted BERT type models is between 84.8-88% at iteration 0. The avg. accuracy can be increased by finetuning of a BioBERT model to 92.2% and by performing additional domain adaptation (ArcTEX model) to 93.6%. Compared to the baseline models, adding additional challenging training samples does improve the performance only slightly, indicating that little to no further training data is required. This is confirmed by experiment 2 for other biomarkers.

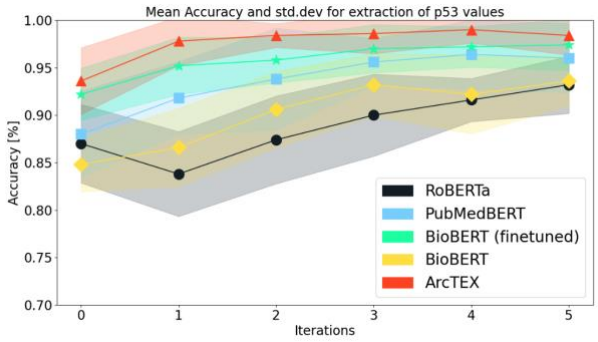


Figure 1: Experiment 1 – Mean accuracy and standard deviation for different QA models to extract the status of biomarker p53 from unstructured reports.

Table 1: Experiment 2 - mean accuracy (standard deviation) for BioBERT and ArcTEX model for different biomarkers at iteration 0 and 5.

Marker	BioBERT		ArcTEX	
	iter. 0	*iter. 5*	*iter. 0*	*iter. 5*
P53	84.8 (2.9)	93.6 (2.6)	93.6 (3.5)	**98.4 (2.1)**
MSH6	92.8 (3.4)	96.5 (2.1)	98.2 (2.0)	**99.0 (2.5)**
Grade	69.4 (7.9)	98.0 (1.6)	92.6 (3.0)	**98.2 (2.0)**
FIGO	89.8 (3.3)	98.6 (1.3)	95.6 (2.8)	**99.0 (1.1)**
MMR	81.6 (5.6)	95.4 (5.2)	96.6 (3.8)	**99.6(0.84)**

Conclusion

We demonstrate that through unsupervised domain adaption and finetuning, the ArcTEX model can extract biomarker values and disease relevant information with high accuracy. This approach does not rely on large language models, making it suitable for hospital environments. In the next stage, the model will be evaluated in a hospital environment and further validated by clinical experts.

Study context

This work uses anonymised data collected by NHS Trusts as part of routine care. We believe that the safe, transparent, and ethical use of anonymised patient data is vital to improve health and care for everyone, and we would like to thank Oxford University Hospitals NHS Foundation Trust for their contribution.

References

[1] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Jul. 2019, Accessed: Apr. 08, 2024. [Online]. Available: https://arxiv.org/abs/1907.11692v1

[2] J. Lee *et al.*, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/BIOINFORMATICS/BTZ682.

[3] Y. Gu *et al.*, “Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing,” *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, Oct. 2021, doi: 10.1145/3458754.

[4] H. W. Chung *et al.*, “Scaling Instruction-Finetuned Language Models,” Oct. 2022, Accessed: Apr. 08, 2024. [Online]. Available: https://arxiv.org/abs/2210.11416v5

[5] E. Almazrouei *et al.*, “The Falcon Series of Open Language Models,” Nov. 2023, Accessed: Apr. 08, 2024. [Online]. Available: https://arxiv.org/abs/2311.16867v2

[6] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 2383–2392, Jun. 2016, doi: 10.18653/v1/d16-1264.

[7] L. Tunstall *et al.*, “Efficient Few-Shot Learning Without Prompts,” Sep. 2022, Accessed: Apr. 08, 2024. [Online]. Available: http://arxiv.org/abs/2209.11055

[8] K. Wang, N. Reimers, and I. Gurevych, “TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning,” *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, pp. 671–688, Apr. 2021, doi: 10.18653/v1/2021.findings-emnlp.59.

Development of Guidelines for Annotating Medication-Related Incident Reports

Huda M. Alshammari[1,2], Denham Phipps[1], Penny Lewis[1], Haifa Alrdahi[1], Riza Batista-Navarro[1],
[1] University of Manchester, Manchester, United Kingdom
[2] Northern Border University, Arar, Saudi Arabia

Introduction

Medication errors pose a significant threat to patient safety worldwide [1-4]. Reporting systems, endorsed by influential bodies like the Institute of Medicine and the World Health Organisation, have collated extensive incident reports globally. However, the narrative format of these reports obstructs their transformation into actionable insights [5]. Natural Language Processing (NLP) and Machine Learning (ML) hold promise in extracting crucial information from medical records, with existing tools demonstrating the ability to identify entities in electronic health records efficiently [6-10]. Despite this potential, NLP's application in learning from incident reports remains underutilised [5,11].
Establishing data annotation guidelines is paramount as a precursor to developing NLP models for incident reports, ensuring uniformity across various documents. Currently, only one project focuses on annotating medication-related incident data, yielding a substantial corpus from Japanese reports [12,13]. This study aims to expand research efforts to improve learning from medication-related incidents, emphasising the importance of developing comprehensive annotation guidelines to enhance patient safety.

Methods and Data

The dataset comprised incident reports from the National Health Service England's Controlled Drug Reporting System. It included incidents reported in 2021 or 2022 related to prescribing, dispensing, or administering medication (N = 1373). To safeguard privacy, identifiable information was anonymized by the custodian. Pre-processing involved removing unnecessary structured data. Notably, text normalisation and sentence tokenisation were intentionally omitted to preserve text nuances in the NLP model. This decision prioritises learning from raw data to maintain contextual understanding and consistent model performance evaluation, avoiding the risk of hindrance from excessive normalisation and tokenisation.

The brat [14] platform was chosen for annotation to maintain consistency. It is an open-source, web-based, user-friendly text annotation tool. A subset of incident reports (N = 100) was selected due to adequate text for annotation. This subset was converted to plain text documents for use in brat. An annotation file in the standoff format, required by brat, was generated. It indicates entity locations, relations, and event participation attributes within the text.

The annotation process consists of four distinct tasks: named entity recognition; entity relation identification; event extraction; and event attribute annotation. Manual annotation using the brat tool was conducted by the lead author, with frequent research team meetings to check annotation quality and discuss ambiguous cases.

An initial draft of the annotation guidelines was developed after extensive literature review and in collaboration with NLP researchers. The guidelines were continuously updated throughout the annotation process, building upon the initial guidelines to produce the final version of the medication-related incident report annotation guidelines.

Results

Following the annotation process, the Medication-Related Incident Report Annotation (MRIRA) guidelines were created to facilitate the recognition of named entities and extraction of events in medication-related incident reports. They define entity types and provide detailed annotation instructions with examples. The guidelines include six general entities (e.g., individuals, locations, time), eight domain-specific entities (e.g., medication details), and define two patient-specific entity types along with nine relation types between entities, see an example for the annotation in the Figure 1.

Before annotation, five main event types, aligned with the reporting system categories, were identified. Event argument roles were established during annotation based on report text. Additional categories were later added to extract crucial information about actions to mitigate harm or prevent incidents. Argument roles for these actions mirrored main events. The guidelines also include two event attribute categories: three for intent and actuality annotation, and one for labelling negated events (see Table 1).

After annotating 30 of the 100 reports, stable guidelines were set; no new entity types were needed. The guidelines also address the handling of ambiguous cases that may arise during annotation.

Table 1. Overview of MRIRA Guidelines' Entity and Event Types and Other Categories.

Category	Annotation Labels
Entity types General	People, location, time, artifact, knowledge, function.
Patient specific	Age group, gender.
Domain specific	Drug name, dosage form, strength or amount, dose, frequency, route of administration, duration, medical condition.
Entities relations	Has dose, has frequency, has form, has route, has duration, has strength or amount, has time, has, at.
Main event types	Prescribing, transcription, dispensing, administration, monitoring.
Supplementary event types	Corrective action, preventive action, underlying and contributing factor, error outcome.
Event arguments roles	Agent, subject, receiver, when, where.
Event attributes	Intended & actual, not intended & actual, intended & not actual, negated.

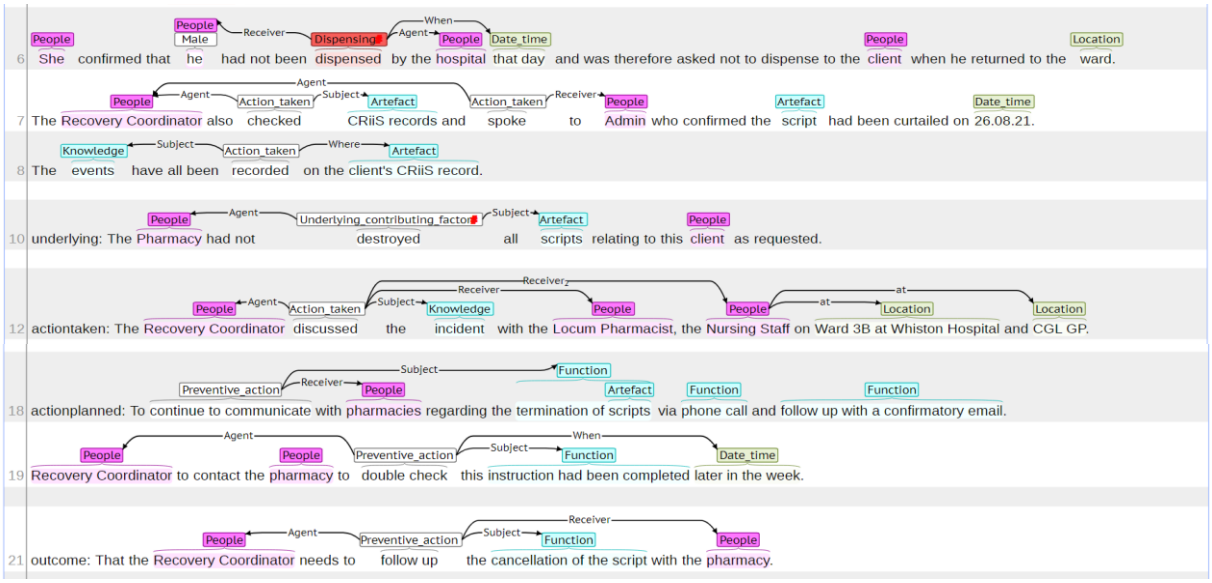


Figure 1. Example of annotation by using brat.

Conclusion

This study presents structured guidelines for annotating medication incident reports, facilitating NLP and ML automation in healthcare incident analysis. However, limitations exist. The guidelines were tailored to a specific dataset, possibly restricting their broader applicability. Additionally, subjective annotation may introduce variability despite aiming for consistency. Future research should adapt the guidelines to diverse datasets, integrate advanced NLP for automation, and conduct inter-annotator agreement tests to assess reliability and validity.

Study Context

This study forms part of a PhD dissertation entitled "Improving of Organisational Learning from Medication Related Incident Data" approved by the University of Manchester UREC, United Kingdom and funded by Northern Border University, Saudi Arabia.

References

1. Reporting systems endorsed by the Institute of Medicine and WHO have collated millions of incident reports. Patient safety incident reporting and learning systems: technical report and guidance. (World Health Organization, 2020)
2. Project to collect medical near-miss/adverse event information: project details and how to participate [Iryō jiko jōhō shūshū-tō jigyō jigyō no naiyō to sanka hōhō]. (The Japan Council for Quality Health Care, 2022).
3. Cooper, J. et al. Nature of blame in patient safety incident reports: mixed methods analysis of a national database. Ann. Fam. Med. 15, 455–461, https://doi.org/10.1370/afm.2123 (2017).
4. Wang, Y., Coiera, E., Runciman, W. & Magrabi, F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. BMC Med. Inform. Decis. Mak. 17, 84, https://doi.org/10.1186/s12911-017-0483-8 (2017).
5. Patient safety incident reporting and learning systems: technical report and guidance. (World Health Organization, 2020)
6. Wong, A., Plasek, J. M., Montecalvo, S. P. & Zhou, L. Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. Pharmacotherapy 38, 822–841, https://doi.org/10.1002/phar.2151 (2018).
7. Wang, Y., Coiera, E. & Magrabi, F. Can unified medical language system–based semantic representation improve automated identification of patient safety incident reports by type and severity? J. Am. Med. Inform. Assoc. 27, 1502–1509, https://doi.org/10.1093/jamia/ocaa082 (2020).
8. Wong, Z. S. Y., So, H. Y., Kwok, B. S., Lai, M. W. & Sun, D. T. Medication-rights detection using incident reports: A natural language processing and deep neural network approach. Health Inform. J. 26, 1460458219889798, https://doi.org/10.1177/1460458219889798 (2019).
9. Wong, Z. S. Y. Statistical classification of drug incidents due to look-alike sound-alike mix-ups. Health Inform. J. 22, 1–17, https://doi.org/10.1177/1460458214555040 (2014).
10. Thompson, P., Daikou, S., Ueno, K. et al. Annotation and detection of drug effects in text for pharmacovigilance. J Cheminform 10, 37 (2018). https://doi.org/10.1186/s13321-018-0290-y
11. Global patient safety action plan 2021–2030: towards eliminating avoidable harm in health care. (World Health Organization, 2022).
12. Wong, Z.S.Y., Waters, N., Liu, J. *et al.* A large dataset of annotated incident reports on medication errors. *Sci Data* **11**, 260 (2024). https://doi.org/10.1038/s41597-024-03036-2
13. Zhang, H. K., Sasano, R., Takeda, K. & Wong, Z. S. Y. Development of a medical incident report corpus with intention and factuality annotation. *LREC 2020* 4578–4584 (2020).
14. https://brat.nlplab.org/

Advancing Clinical Language Representation: Leveraging Semantic Cues in Clinical narrative

Lena AlMutair[1,2], Eric Atwell[1], and Nishant Ravikumar[1]

[1]School of Computing, University of Leeds, Leeds, UK
[2]Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

1 Introduction

Enhancing clinical language representation through advanced embeddings is crucial for healthcare and medical research. Current healthcare information retrieval often overlooks contextual nuances and relies on structured data or keyword matching, impeding semantic relevance. To overcome this, our research aims to build robust information retrieval systems by enhancing embeddings through clinical note training and incorporating semantic cues from medical concepts or ICD codes.

This research is vital for improving patient care and supporting downstream tasks like clinical trial matching[1, 2, 3], drug repurposing[4], clinical decision support [5, 6]. Utilizing open datasets like MIMIC-III[7] promotes transparency and reproducibility. Challenges in obtaining labeled clinical data include cost, time-consuming procedures, and privacy regulations. Recent studies, including Memarzadeh [8] and Zhang's [9], advance patient similarity assessments and prediction tasks using EHR data. While Wei and Eickhoff's work aligns with ours in the context of information retrieval for clinical notes [10], it differs in approach. They emphasize recommending pertinent PubMed articles to support the clinical notes of a patient[10]. Automating ICD code assignment using deep learning can significantly impacts the quality of healthcare information retrieval[11]. Incorporating semantic cues aligning ICD codes with clinical notes can enhance information retrieval and medical natural language inference tasks, promising advancements in healthcare information retrieval and medical research.

2 Methods and Data

In our study, we utilized the MIMIC-III dataset [7] and integrated various cues, including concepts and ICD codes, to characterize clinical notes. We evaluated the effectiveness of combining concept embeddings with ClinicalBert embeddings [12] across different settings to enhance note retrieval. These settings ranged from using only clinical narrative (setting 1), combining embeddings into a single vector (setting 2), using multiple concpt embeddings for aggregated similarity (setting 3), notes with concepts (setting 4), and measuring similarity with the query. Concept extraction was facilitated by the MedCAT tool [13].

We leveraged ICD codes to augment embeddings, utilizing a Siamese network approach [14] with ClinicalBERT[12]/BlueBERT [15] and a contrastive loss function to maximize similarity between query q and relevant note n (Equation 1). Our approach was inspired by the work of Henderson et al. [16]. These negative samples, denoted as n_j for all $i \neq j$, serve to encourage the model to maximize the dissimilarity between q and n_j. We assessed the impact on the accuracy of the MedNLI dataset derived from MIMIC-III [17], evaluating accuracy pre- and post-training to measure the significance of incorporating ICD codes.

$$L(q, n_i) = \max(0, \delta + sim(q, n_j) - sim(q, n_i)) \quad (1)$$

$sim(x, y)$ represents the similarity function between embeddings x and y.

Overall, our study involved comprehensive evaluations of different cues for characterizing clinical notes, with a focus on improving information retrieval and medical natural language inference tasks.

3 Results

In routine healthcare, practitioners often concentrate on individual cases. We explore the value of medical concepts for language representation by applying top 20 concepts as key queries related to prevalent diseases in the MIMIC-III dataset [7], validated on a patient with numerous records. Setting thresholds on the number of queries, such as popular ICD codes, optimizes relevance and practicality, as seen in various medical contexts [8, 10].We utilize pretrained ClinicalBERT to compute BERTScore across four settings (see Sec 2). Setting 4 yields the highest score among the settings as depicted in Table 1, indicating the effectiveness of combining concept with narrative embeddings in capturing language nuances. We assess language representation in the Medical Natural Language Inference (MedNLI) task to evaluate embedding quality for future Information Retrieval (IR) systems. Initially, with various pretrained models, we achieve accuracy rates of 86.66% and 89.31% with 'Our (BlueBERT)' and 'Our (Large BlueBERT),' respectively. Upon retraining embeddings with ICD codes, our model, ClinicNarr, achieves a higher accuracy of 90.5%, demonstrating the effectiveness of this integration. Additionally, we report results in Table 2 with a batch size of 32 over four epochs.

This study evaluates various models' performance in capturing semantics within a dataset [18]. We achieve a correlation coefficient of 0.72 with coder ratings and a low p-value, indicating significant improvement over previous results [19], highlighting the model's effectiveness in enhancing semantic similarity tasks.

Table 1: BERTscore system evaluation for 4 settings on the top 20 concepts in MIMIC-III.

Setting	**F1**	**Precision**	**Recall**
Setting 1	0.630	0.590	0.677
Setting 2	0.630	0.582	0.688
Setting 3	0.673	0.610	0.751
Setting 4	0.699	0.648	0.770

Table 2: Comparison of Approaches and Accuracyon MedNLI

Approach	**Accuracy**
Our (ClinicalBERT)	83.4%
Our (BlueBert)	**86.66%**
Our (Large BlueBERT)*	**89.31%**
BLUE (Base)[15]	84%
CIinicNarr	**90.5%**
Handcrafted Features[17]	52%
InferSent [17]	73.5%
LongTransformer[20]	84%

* with 24 layers

4 Conclusion

This research evaluates cues' efficacy in enhancing semantic understanding within clinical datasets. Initially, incorporating concepts, particularly in setting 4, showed promise. Enriching embeddings with ICD codes further improved semantic representation, crucial for future information retrieval systems. Inclusion of semantic features enhanced embeddings' downstream performance on MedNLI and correlation with expert

Table 3: Semantic Similarity Correlations with Medical practitioner Ratings

Model	Correlation		p-value	
	Physician	Coder	Physician	Coder
ClinicalBERT	0.46	0.51	0.0001	0.0036
BlueBert	0.61	0.65	0.0003	0.0002
Our fine-tuned on MedNLI	0.64	0.72	0.0003	6.91e-06

ratings. Challenges in accessing labeled data and computing resources highlight areas for future development.

5 Study context

This study utilized the publicly available MIMIC-III dataset [7], available at (https://physionet.org/content/mimiciii/1.4/), which adheres to the ethical standards and IRB approvals obtained by its creators. The authors declare no conflicts of interest.

References

[1] Mc Cord KA, Hemkens LG. Using electronic health records for clinical trials: Where do we stand and where can we go? Cmaj. 2019;191(5):E128-33.

[2] Jin Q, Wang Z, Floudas CS, Sun J, Lu Z. Matching patients to clinical trials with large language models. arXiv preprint arXiv:230715051. 2023.

[3] Goldstein BA. Five analytic challenges in working with electronic health records data to support clinical trials with some solutions. Clinical Trials. 2020;17(4):370-6.

[4] Liu R, Wei L, Zhang P. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. Nature machine intelligence. 2021;3(1):68-75. Available from: https://doi.org/10.1038/s42256-020-00276-w.

[5] Mills S. Electronic health records and use of clinical decision support. Critical Care Nursing Clinics. 2019;31(2):125-31.

[6] Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine. 2020;3(1):17. Available from: https://doi.org/10.1038/s41746-020-0221-y.

[7] Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database [Journal Article]. Scientific data. 2016;3(1):1-9.

[8] Memarzadeh H, Ghadiri N, Samwald M, Lotfi Shahreza M. A study into patient similarity through representation learning from medical records. Knowledge and Information Systems. 2022 12;64(12):3293-324. Available from: https://doi.org/10.1007/s10115-022-01740-2.

[9] Zhang J, Kowsari K, Harrison JH, Lobo JM, Barnes LE. Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record. IEEE Access. 2018;6:65333–65346. Available from: http://dx.doi.org/10.1109/ACCESS.2018.2875677.

[10] Wei X, Eickhoff C. Embedding Electronic Health Records for Clinical Information Retrieval. CoRR. 2018;abs/1811.05402. Available from: http://arxiv.org/abs/1811.05402.

[11] Kaur R, Ginige JA, Obst O. A Systematic Literature Review of Automated ICD Coding and Classification Systems using Discharge Summaries. CoRR. 2021;abs/2107.10652. Available from: https://arxiv.org/abs/2107.10652.

[12] Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings [Journal Article]. arXiv preprint arXiv:190403323. 2019.

[13] Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit [Journal Article]. Artificial Intelligence in Medicine. 2021;117:102083.

[14] Chicco D. Siamese Neural Networks: An Overview [Journal Article]. Methods Mol Biol. 2021;2190:73-94.

[15] Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. CoRR. 2019;abs/1906.05474. Available from: http://arxiv.org/abs/1906.05474.

[16] Henderson M, Al-Rfou R, Strope B, Sung YH, Lukács L, Guo R, et al. Efficient natural language response suggestion for smart reply. arXiv preprint arXiv:170500652. 2017.

[17] Romanov A, Shivade C. Lessons from Natural Language Inference in the Clinical Domain; 2018.

[18] Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain [Journal Article]. Journal of biomedical informatics. 2007;40(3):288-99.

[19] Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission [Journal Article]. arXiv preprint arXiv:190405342. 2019.

[20] Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. Journal of the American Medical Informatics Association. 2022 11;30(2):340-7. Available from: https://doi.org/10.1093/jamia/ocac225.

Where are all the antimicrobials being used? Large Language Models for Monitoring and Adherence to Stewardship Guidelines in Veterinary Practices

Sean Farrell[1], Noura Al Moubayed[1], Alan Radford[2], Gina Pinchbeck[2], and Peter-John Mäntylä Noble[2]

[1]Department of Computer Science, Durham University, United Kingdom
[2]Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, United Kingdom

1 Introduction

Companion animals are increasingly being recognised as an important contributor to the development and transmission of antimicrobial resistant (AMR) bacteria [1, 2, 3]. AMR has emerged as a global health concern, posing significant challenges to effectively treating infectious diseases. The overuse of antimicrobials in human and animal healthcare has contributed to the emergence of resistant strains [1, 4, 5]. Antimicrobials remain amongst the most commonly prescribed pharmaceutical agents in companion animals [6]. These essential medications are frequently prescribed in small animal veterinary practices where there is a complete lack of monitoring and oversight for the rationale behind HPCIA prescriptions in companion animals in the United Kingdom (UK). Growing concerns about the effect of AMR on public health has prompted various organisations to create antimicrobial stewardship guidelines (ASGs) [7]. These aim to promote responsible use of antimicrobials among veterinarians [8]. However, there is currently no system to check adherence to these guidelines. In the UK, annual reports focus on overall antimicrobial sales data without an in-depth analysis of the reasoning behind clinicians' prescription choices, underscoring a significant oversight in monitoring antimicrobial prescribing practices.

2 Methods and Data

SAVSNET has collected Electronic Health Records (EHRs) since March 2014. The network consists of over 500 veterinary sites located throughout the UK [9]. The collected data includes species, breed, sex, neuter status, age, owner's postcode, insurance and microchipping status, any products sold or prescribed and a free-text clinical narrative recorded by the attending practitioner providing insights of the events during the consultation. The construction of a multi-label hierarchical classifier was informed by the diseases identified within the most prevalent ASG, PROTECT ME [10, 11]. From this, 39 binary sequence classification models were developed

utilising the PetBERT foundational model [12]. This hierarchical classifier model was subsequently trained on a dataset comprising 250,000 records, each one having been subjected to all 39 classifiers. We then apply this model to the entire SAVSNET dataset and compare the prescription data for a disease consult against what the ASG suggests.

3 Results

We analysed a dataset totalling 1,239,534 prescriptions for dogs and 309,321 for cats. We attained a overall F1-score performance of 88% across our 39 disease classifier, and applied this against the full SAVSNET dataset. We verified prescription data against the PROTECT ME ASGs and reported percentage of compliance in table 1

Table 1: Antimicrobial Recommendations with total count of prescriptions that adhere to the ASG with a a relative percentage of agreement with guidelines over all antimicrobials prescribed a given condition. *AC = Amoxicillin Clavulanate, FA = Fusidic acid, FQ = Fluroquinlones, TMPS = Trimethoprim/Sulphonamide*

Disease	**Guidance**	**Cats (%)**	**Dogs (%)**
Cat Bite Abscess	AC, Cefalexin FQ	14496 (89%)	1789 (61%)
Conjunctivitis	Chlortetracycline FA, Gentamicin	15365 (42%)	36215 (39%)
Cystitis	AC	5052 (28%)	10069 (85%)
Gingivitis	AC, Clindamycin Metronidazole	32845 (44%)	39592 (53%)
Hepatitis	Ampicillin AC, Cefalexin Metronizole	122 (58%)	243 (56%)
Pancreatitis	Not advised	3454 (84%)	13699 (69%)
Pneumonia	Dogs: AC, FQ Doxycycline Metronidzole Cats: AC, Doxycycline	458 (69%)	3191 (96%)
Pyoderma	Chlorhexidine	0 (0%)	2 (1%)
Rhinitis	AC	954 (30%)	1174 (59%)
UTI	AC, TMPS	4477 (47%)	18637 (71%)

4 Conclusion

Here we present our hierarchical LLM classifier as a novel methods to combat AMR, specifically through understanding antimicrobial prescribing for pets in the UK. Our analysis of over 1.2 million antimicrobial prescriptions revealed previously unknown motivations. With an 88% accuracy rate, our model supports the identification of 39 different diseases providing a robust basis for analysing antimicrobial use in line with ASGs. Our data underlines the urgent need for improved stewardship as AMR poses risks to both public health and animal welfare.

5 Study context

We thank the data providers in veterinary practice (VetSolutions, Teleos, CVS, and other practitioners). Without their support and participation, this research would not be possible. SAVSNET has obtained ethical approval from the University of Liverpool Research Ethics Committee (RETH001081). The datasets analysed during the current study are not publicly available due to issues surrounding owner confidentiality. Reasonable requests can be made to the SAVSNET Data Access and Publication Panel (savsnet@liverpool.ac.uk) for researchers who meet the criteria for access to confidential data. P-JN was funded by CWG grant from Dogs Trust (SAVSNET Agile). SF was supervised by NAM on a BBSRC funded PhD studentship.

References

[1] Rantala M, Lahti E, Kuhalampi J, Pesonen S, Järvinen AK, Saijonmaa-Koulumies L, et al. Antimicrobial resistance in Staphylococcus spp., Escherichia coli and Enterococcus spp. in dogs given antibiotics for chronic dermatological disorders, compared with non-treated control dogs. Acta Veterinaria Scandinavica. 2004;45(1):37. Available from: /pmc/articles/PMC1820999//pmc/articles/PMC1820999/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC1820999/.

[2] Trott DJ, Filippich LJ, Bensink JC, Downs MT, McKenzie SE, Townsend KM, et al. Canine model for investigating the impact of oral enrofloxacin on commensal coliforms and colonization with multidrug-resistant Escherichia coli. Journal of Medical Microbiology. 2004 5;53(5):439-43. Available from: https://www.microbiologyresearch.org/content/journal/jmm/10.1099/jmm.0.05473-0.

[3] Guardabassi L, Loeber ME, Jacobson A. Transmission of multiple antimicrobial-resistant Staphylococcus intermedius between dogs affected by deep pyoderma and their owners. Veterinary Microbiology. 2004 1;98(1):23-7. Available from: https://pubmed.ncbi.nlm.nih.gov/14738778/https://pubmed.ncbi.nlm.nih.gov/14738778/?dopt=Abstract.

[4] Schmidt VM, Pinchbeck G, McIntyre KM, Nuttall T, McEwan N, Dawson S, et al. Routine antibiotic therapy in dogs increases the detection of antimicrobial-resistant faecal Escherichia coli. Journal of Antimicrobial Chemotherapy. 2018 12;73(12):3305-16. Available from: https://academic.oup.com/jac/article/73/12/3305/5095201.

[5] Cantón R, Bryan J. Global antimicrobial resistance: from surveillance to stewardship. Part 1: surveillance and risk factors for resistance. http://dxdoiorg/101586/eri12120. 2014;10(11):1269-71. Available from: https://www.tandfonline.com/doi/abs/10.1586/eri.12.120.

[6] Singleton DA, Sánchez-Vizcaíno F, Arsevska E, Dawson S, Jones PH, Noble PJM, et al. New approaches to pharmacosurveillance for monitoring prescription frequency, diversity, and co-prescription in a large sentinel network of com-

panion animal veterinary practices in the United Kingdom, 2014–2016. Preventive Veterinary Medicine. 2018 11;159:153. Available from: /pmc/articles/PMC6193134//pmc/articles/PMC6193134/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6193134/.

[7] Allerton F, Prior C, Bagcigil AF, Broens E, Callens B, Damborg P, et al. Overview and Evaluation of Existing Guidelines for Rational Antimicrobial Use in Small-Animal Veterinary Practice in Europe. Antibiotics (Basel, Switzerland). 2021 4;10(4). Available from: https://pubmed.ncbi.nlm.nih.gov/33918617/.

[8] Davey P, Marwick CA, Scott CL, Charani E, Mcneil K, Brown E, et al. Interventions to improve antibiotic prescribing practices for hospital inpatients. The Cochrane database of systematic reviews. 2017 2;2(2). Available from: https://pubmed.ncbi.nlm.nih.gov/28178770/.

[9] Sánchez-Vizcaíno F, Jones PH, Menacere T, Heayns B, Wardeh M, Newman J, et al. Small animal disease surveillance. Veterinary Record. 2015 12;177(23):591-4. Available from: https://onlinelibrary.wiley.com/doi/full/10.1136/vr.h6174https://onlinelibrary.wiley.com/doi/abs/10.1136/vr.h6174https://bvajournals.onlinelibrary.wiley.com/doi/10.1136/vr.h6174.

[10] Farrell S, Bagcigil AF, Chaintoutis SC, Firth C, Aydin FG, Hare C, et al. A multinational survey of companion animal veterinary clinicians: How can antimicrobial stewardship guidelines be optimised for the target stakeholder? The Veterinary Journal. 2023 11:106045. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1090023323000965.

[11] BSAVA and SAMSoc, editor. BSAVA/SAMSoc Guide to Responsible Use of Antibacterials: PROTECT ME. British Small Animal Veterinary Association; 2018. Available from: https://www.bsavalibrary.com/content/book/10.22233/9781910443644.

[12] Farrell S, Appleton C, Noble PJM, Al Moubayed N. PetBERT: automated ICD-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records. Scientific Reports 2023 13:1. 2023 10;13(1):1-14. Available from: https://www.nature.com/articles/s41598-023-45155-7.

Feasibility study of 'MiADE' point of care natural language processing system: methodology and initial results

Jennifer Jiang[1,2], James Brandreth[1,2], Mairead McErlean[3], Jack Ross[3], Maisarah Amran[3], Enrico Costanza[1], Yogini Jani[1,3], Leilei Zhu[2,3], Richard Dobson[1,2], Folkert Asselbergs[1,2], Wai Keong Wong[4], Anoop D. Shah[1,2,3]

[1] University College London, London, UK
[2] UCLH / NIHR BRC Clinical and Research Informatics Unit, UCLH, London, UK
[3] University College London Hospitals NHS Foundation Trust, London, UK
[4] Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

Introduction

Well-organised medical records are essential for high quality patient care [1], but much of the information in today's records are in free text rather than in a structured form [2,3], and is difficult to use clinically or for research. Structured recording of diagnoses and other key items of clinical information is recommended by national guidance from the Professional Record Standards Body (PRSB) [4]. Widespread use of controlled clinical terminologies such as SNOMED CT enables clinical concepts to be recorded in a consistent way, but only if clinicians are able to use the system easily. It can be onerous and time-consuming for clinicians to enter detailed structured information in many EHR systems [3], and time spent on data entry can affect the human experience of clinical consultations. There is a need to facilitate data entry without impeding the clinical workflow.

MiADE (Medical Information AI Data Extractor) is a natural language processing (NLP) system developed by our team, and designed to be embedded within electronic health record (EHR) systems. It enables unstructured text to be converted into structured data in real time. SNOMED CT concepts for clinical findings extracted from the text are presented to the clinician for validation, and can be amended or corrected before being saved.

MiADE is based on the open source CogStack MedCAT named entity extraction system [5] along with a rule-based section detection algorithm to identify headings such as 'Problem list'. MiADE was implemented for a clinical evaluation study at University College London Hospitals (UCLH), and went live on 26 February 2024 for a limited set of users, with 'before' and 'after' evaluation. This paper summarises the initial informal feedback received; further findings will be presented at the conference.

Methods and Data

The MiADE evaluation consists of an inpatient and outpatient substudy (see Figure 1). Clinicians (doctors, physician associates and allied health professionals) were recruited via articles in the hospital newsletter, by presentations given at departmental and divisional meetings and by word of mouth. For inpatients, we asked clinical leads to agree to the study and recruit their ward teams.

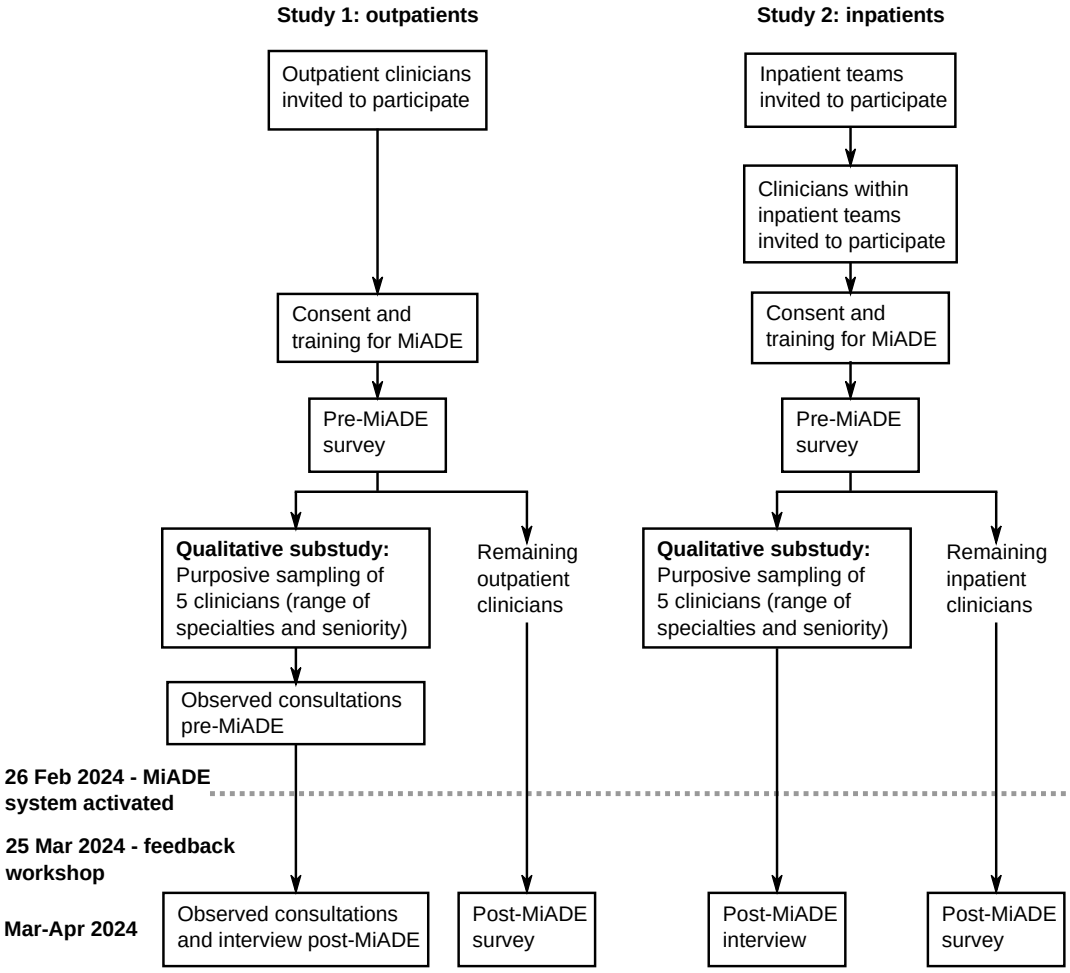
MiADE was activated for individual clinicians based on their Epic user profiles. Before MiADE activation, clinicians were asked to complete a consent form, watch a training video and complete an online questionnaire on their thoughts about structured data and their experience of the EHR system. Three clinician investigators (ADS, JR and MA) tested the MiADE system live from 14 February 2024, and it was switched on for the trial participants on 26 February. A qualitative researcher observed a sample of outpatient consultations and interviewed patients before and after MiADE activation.

Clinician participants were kept informed of the trial progress via a weekly email newsletter, and were encouraged to report feedback or issues. A hybrid feedback workshop was held on 25 March 2024, at which suggestions for improvement were discussed.

Results

Forty-five clinicians were recruited and completed the pre-MiADE questionnaire. The MiADE system could not be activated for three participants because of specific profile settings, so 42 were able to use the system. Post-MiADE surveys and observations are in progress. The feedback workshop was attended by 10 clinicians of whom 6 clinicians currently had MiADE activated. Those who had used MiADE found that it was useful for diagnoses, but there were some errors and some non-diagnosis findings (e.g. 'non-smoker') which were not useful for the problem list and should be filtered out. They also felt it was essential to embed technological improvements such as this within an education and quality improvement programme to ensure that problem lists are well curated and useful.

Figure 1. Study flow diagram



Conclusion

We successfully integrated a novel NLP system with the Epic EHR to assist the entry of structured problem lists at the point of care, and initial feedback from users was positive. Full results from the trial need to be collected and analysed, and it would be beneficial to test this approach with a larger number of users and additional clinical sites.

Study context

The study was approved by the Hampshire A Research Ethics Committee (23/SC/0221) and is registered at https://www.isrctn.com/ISRCTN58300671. The study was funded by NIHR (AI_AWARD01864) and EPSRC (EP/Y018087/1). We acknowledge the assistance of the UCLH EHR team in configuring Epic at UCLH to communicate with the MiADE system.

References

1. Weed LL. Medical records that guide and teach. N Engl J Med. 1968;278: 593–600.
2. Poulos J, Zhu L, Shah AD. Data gaps in electronic health record (EHR) systems: An audit of problem list completeness during the COVID-19 pandemic. Int J Med Inform. 2021 Jun;150:104452. doi: 10.1016/j.ijmedinf.2021.104452.
3. Kalra D, Fernando B. Approaches to enhancing the validity of coded data in electronic medical records. Prim Care Respir J. 2011;20: 4–5. doi:10.4104/pcrj.2010.00078
4. Shah AD, Quinn NJ, Chaudhry A, Sullivan R, Costello J, O'Riordan D, et al. Recording problems and diagnoses in clinical care: developing guidance for healthcare professionals and system designers. BMJ Health Care Inform. 2019;26. doi:10.1136/bmjhci-2019-100106
5. Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artif Intell Med. 2021 Jul;117:102083. doi: 10.1016/j.artmed.2021.102083.

Improving Multi-Task Text Classification Performance in Electronic Health Records

Shubham Agarwal[1], Thomas Searle[1], Anthony Shek[2], James Teo[2], and Richard Dobson[1]

[1]Department of Biostatistics & Health Informatics, King's College London, UK
[2]Guy's and St Thomas' NHS Foundation Trust, London, UK

1 Introduction

Electronic Health Records (EHRs) contain unstructured text with comprehensive records of patient interactions and health-related data which serve various purposes [1]. Clinical text classification plays a crucial role in uncovering insights hidden within clinical narratives [2]. However, the extraction of insights and subsequent information from text can be challenging due to the complexity of the data, sensitive nature of data, inconsistent and missing information and class imbalance [3][4]. Current research on clinical text classification presents a complex landscape, with contrasting results leading to uncertainty regarding the optimal choice of models. This study investigates the performance of BERT-based classification models for EHR data and evaluates the influence of class imbalance on model performance and explore techniques to address the same.

2 Methods and Data

The dataset is sourced from CogStack [5], deployed at Guy's & St Thomas' NHS Foundation Trust, and includes 1800 documents containing information for the following classification tasks:

- **Presence:** 'Hypothetical (N/A)' - 978 samples, 'Not present (False)' - 578 samples; 'Present (True)' - 7430 samples
- **Experiencer**: 'Family' - 75 samples, 'Patient' - 7908 samples, 'Other' - 1002 samples
- **Temporality**: 'Past' - 733 samples, 'Recent' - 7771 samples, 'Future' - 484 samples

The pre-trained BERT model was fine-tuned by adding a single fully connected layer to perform the classification. We investigated the effect of freezing BERT's parameters and training them, with LoRA [6] being used to optimize the training. The dataset used has severe class imbalance which poses challenge for the model's performance. This is addressed by making use of class weights, loss functions tailored for class imbalance (focal loss [7]), and 2 phase learning [8]. 2 phase learning is a training approach designed to fix the issue of the gradients being dominated by the majority class. The 2 phases in this approach are: **Phase 1 -** In this phase, all classes are down sampled to a specified value N (close to the number of samples in minority class) and training is performed with higher weights given to minority classes. Phase 1 allows the model to capture and learn the details for the minority classes. **Phase 2 -** During this stage, the model undergoes a second round of training, now on the entire dataset. In contrast to Phase 1, the class weights assigned to minority classes are relatively higher but lower compared to the initial phase. This phase allows the model to capture the finer details for all classes, leading to a more finely-tuned model.

2.1 LLM Approaches

Along with the above approaches, oversampling is performed using Mistral 7b [9] model to generate additional data samples for the minority classes to boost the performance of the model. The model was presented with 10 examples from the dataset, 8 from the minority classes and 2 from the majority classes with manual validation of the generated data was performed before adding it to the dataset. Llama 2 7b and Mistral 7b models are also used for text classification by leveraging zero-shot and few-shot learning techniques with 3 examples per class being presented for few-shot learning.

3 Results

Performance of the model is evaluated on accuracy and normalized confusion matrix. Table 1 contains the performance of all models evaluated during this experiment .

Table 1: BERT model performance for classification

Category	**Variant**	**Test**	
		Accuracy	***Confusion matrix***
Presence	Base	0.86	[0.8,0.75,0.91]
	2 phase learning	0.87	[0.87,0.84,0.9]
Experiencer	Base	0.87	[0.83,0.81,0.9]
	2 phase learning	0.9	[0.92,0.875,0.91]
Temporality	Base	0.87	[0.8,0.78,0.83]
	2 phase learning	0.87	[0.84,0.88,0.88]

The two-phase learning model demonstrates consistently high performance across all categories, especially for the minority classes (including both the first two and all categories) and significantly surpassing the baseline BERT model which is without class imbalance mitigation techniques. Given the considerable class imbalance within the dataset, employing a two-phase learning approach effectively mitigates this issue. When utilizing oversampled data, the model exhibits promising outcomes, with a notable improvement of **5-10%** observed in the confusion matrix across many classes. Furthermore, although evaluated on a limited dataset, both Llama 2 and Mistral models exhibit comparable performance in zero-shot learning, achieving an accuracy of 73%. However, during few-shot learning, Mistral surpasses Llama 2, achieving an accuracy of 86% compared to Llama 2's 76%.

4 Conclusion

The BERT model demonstrates remarkable performance in medical text classification, particularly when evaluating metrics that prioritize importance across all cases. The integration of the two-phase learning approach proves to be advantageous in effectively managing class imbalance and ensuring model's ability to be adeptly fitted across all classes. Oversampling data using LLM adds value and shows promise to address class imbalance issues.

5 Study context

This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and Kings College London. No conflict of interest declared.

References

[1] NHS. Purpose of the GP electronic health record; 2023. Accessed on March 18, 2024. Available from: https://www.england.nhs.uk/long-read/purpose-of-the-gp-electronic-health-record/.

[2] Spasic I, Nenadic G, et al. Clinical text data in machine learning: systematic review. JMIR medical informatics. 2020;8(3):e17984.

[3] Ratwani RM. Electronic health records and improved patient care: opportunities for applied psychology. Current directions in psychological science. 2017;26(4):359-65.

[4] Khushi M, Shaukat K, Alam TM, Hameed IA, Uddin S, Luo S, et al. A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access. 2021;9:109960-75.

[5] Jackson R, Kartoglu I, Stringer C, Gorrell G, Roberts A, Song X, et al. CogStack-experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. BMC medical informatics and decision making. 2018;18:1-13.

[6] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:210609685. 2021.

[7] Ross TY, Dollár G. Focal loss for dense object detection. In: proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2980-8.

[8] Lee H, Park M, Kim J. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. In: 2016 IEEE international conference on image processing (ICIP). IEEE; 2016. p. 3713-7.

[9] Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, et al. Mistral 7B. arXiv preprint arXiv:231006825. 2023.

How representative are heart failure clinical trials? A comparative study using natural language processing

Jack Wu[1], Jonathan Breeze[2], Dhruva Biswas[1,2], Sam Brown[1,2], Brian Tam To[2], Matthew Ryan[1,2], Theresa McDonagh[1,2], Daniel Bromage[1,2], Antonio Cannata[1,2], Thomas Searle[1], James Teo[1,2], Richard Dobson[1], Ajay Shah[1,2], Kevin O'Gallagher[1,2]

[1] King's College London, London, UK
[2] King's College Hospital NHS Foundation Trust, London, UK

Introduction

Natural language processing (NLP) has emerged as a powerful tool in the field of healthcare, offering the ability to efficiently analyse vast amounts of data [1]. In randomised controlled trials (RCTs), the representativeness of the participants is crucial for the generalisability of the findings. However, there has been a lack of scalable and effective methods to assess how well these trials reflect the broader patient population. This study aims to leverage the capabilities of NLP to evaluate the representativeness of heart failure (HF) RCTs cohort by comparing it to a large real-world HF cohort. Ultimately, this study seeks to uncover disparities [2] and ensure that clinical trials are as inclusive and representative as possible, which is essential for the advancement of patient-centred care.

Methods and Data

This is a retrospective cohort study of patients with HF at a large regional tertiary centre. Identifiers of patients who enrolled in 15 HF RCTs in the centre from 2012 to 2023 were manually collated using local enrolment records. The overall HF cohort was identified using NLP pipeline based on CogStack [3] and MedCAT [4] deployed in the centre. The NLP pipeline enables the integration and analysis of both structured and unstructured health data, allowing comprehensive cohort characterization. A similar NLP-based approach was recently used to assemble a HF with preserved ejection fraction (HFpEF) cohort [5]. Demographics, Index of Multiple Deprivation (IMD), comorbidities, symptoms, smoking behaviour and hospitalisation of patients were analysed. Principal component analysis (PCA) was used to visualise the 2 cohorts.

Results

We identified 11,885 patients with a diagnosis of HF. Of these, 236 were recruited into HF RCTs. Characteristics of the patients are shown in Table 1. The age (Figure 1) of participants in HF RCTs was significantly younger (66 ± 12 years) compared to the whole HF cohort (70 ± 15 years) ($p<0.001$). The proportion of females in HF RCTs was significantly lower (27%) than in the whole HF cohort (42%) ($p<0.001$). No significant differences were found in terms of ethnicity and IMD. Regarding comorbidities, compared to the whole HF cohort, myocardial infarction was more common in the HF RCTs group (56% vs 38%, $p<0.001$) while kidney disease was less common (48% vs 60%, $p<0.001$). Both cohorts show similar rates of symptoms, smoking behaviour, and hospitalization. Distributions of the 2 cohorts using PCA are shown in Figure 2.

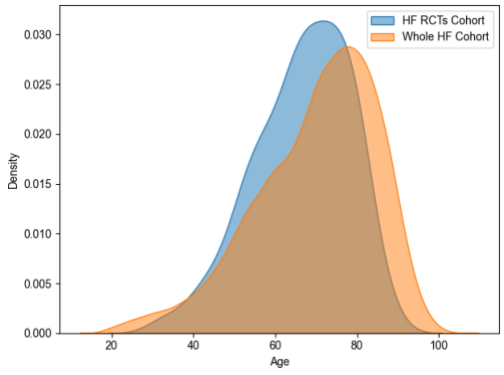


Figure 1. Kernel density estimate plot of age for the 2 cohorts

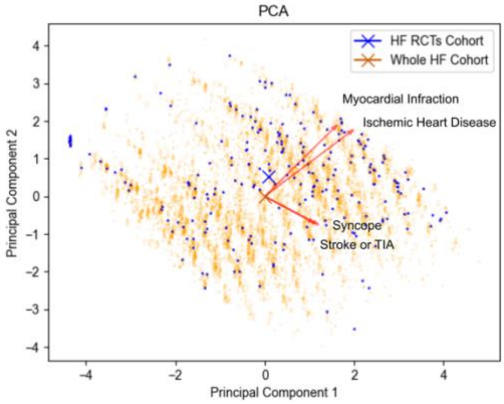


Figure 2. PCA of the 2 cohorts with loadings of the top features

Table 1. Patient Characteristics

Characteristics	HF RCTs Cohort N=236	Whole HF Cohort N=11,885	p-value
Age, Mean ± SD	66 ± 12	70 ± 15	**<0.001**
Female	63 (27%)	5,046 (42%)	**<0.001**
Ethnicity			
White	153 (65%)	7,394 (62%)	0.412
Black	51 (22%)	2,628 (22%)	0.856
Asian	8 (3%)	670 (6%)	0.137
IMD, Median [IQR]	4 [3, 6]	4 [3, 6]	0.415
Comorbidities			
Diabetes Mellitus	93 (39%)	4,118 (35%)	0.140
Hypertension	177 (75%)	9,738 (82%)	0.016
Stroke	87 (37%)	4,598 (39%)	0.565
Ischemic Heart Disease	167 (71%)	7,237 (61%)	0.001
Atrial Fibrillation	92 (39%)	5,262 (44%)	0.101
Myocardial Infarction	131 (56%)	4,465 (38%)	**<0.001**
Kidney Disease	114 (48%)	7,094 (60%)	**<0.001**
Angina	77 (33%)	2,846 (24%)	0.005
Symptoms			
Dyspnea	204 (86%)	10,993 (92%)	0.008
Dyspnea at Rest	10 (4%)	622 (5%)	0.455
Chest Pain	162 (69%)	8,192 (69%)	0.929
Dizziness	103 (44%)	5,049 (42%)	0.722
Presyncope	14 (6%)	692 (6%)	0.943
Syncope	44 (19%)	2,211 (19%)	0.986
Lifestyle			
Non-smokers	173 (73%)	8,587 (72%)	0.718
Healthcare utilization			
Hospitalization, Median [IQR]	2 [1, 5]	2 [1, 5]	0.026

Conclusion

Our NLP-based analyses revealed that HF trials at a tertiary cardiology centre were more likely to recruit patients who were younger and male compared to the broader patient population. This study shows NLP methods can be used to assess the representativeness of RCTs. The same approach can be applied across different domains within cardiovascular medicine, and potentially, for any medical condition. For future work, we plan to study how the inclusion/exclusion criteria could skew the distribution and develop metrics of representativeness that will further quantify the degree to which clinical trials reflect the broader patient population. This can also enhance patient recruitment by incorporating local cohort characteristics into randomisation strategies.

Study context

This project operated under London South-East Research Ethics Committee approval (18/LO/2048) granted to the King's Electronic Records Research Interface (KERRI) which did not require written informed patient consent. This study complies with the Declaration of Helsinki.

References

1. Bean, D. M., Kraljevic, Z., Shek, A., Teo, J., & Dobson, R. J. (2023). Hospital-wide Natural Language Processing summarising the health data of 1 million patients. *PLOS Digital Health*, 2(5), e0000218.
2. Mihan, A., & Van Spall, H. G. (2024). Interventions to enhance digital health equity in cardiovascular care. *Nature Medicine*, 1-3.
3. Jackson, R., Kartoglu, I., Stringer, C., Gorrell, G., Roberts, A., Song, X., ... & Dobson, R. (2018). CogStack-experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. *BMC medical informatics and decision making*, 18(1), 1-13
4. Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., ... & Dobson, R. J. (2021). Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117, 102083.
5. Wu, J., Biswas, D., Ryan, M., Bernstein, B. S., Rizvi, M., Fairhurst, N., ... & O'Gallagher, K. (2023). Artificial intelligence methods for improved detection of undiagnosed heart failure with preserved ejection fraction. *European Journal of Heart Failure*.

Annotation of Outpatient Letters to Estimate Prevalence and Misclassification of Musculoskeletal Disease

Warren Del-Pinto[1], Jenny Humphreys[1], Meghna Jani[1], Prajwal Khairnar[2], Ana Aldana[2], Robyn Hamilton[2], Karim Webb[2], Goran Nenadic[1], and William G. Dixon[1]

[1]University of Manchester, United Kingdom
[2]The Northern Care Alliance NHS Foundation Trust, United Kingdom

1 Introduction

This paper presents an annotation schema for capturing musculoskeletal (MSK) diagnoses described in hospital rheumatology outpatient letters (OPLs). The schema defines an annotation task to provide datasets to evaluate NLP tools and address the following clinical questions:

1. What is the prevalence and incidence of musculoskeletal (MSK) disease in the population?
2. How many patients have been misclassified, with respect to categories of MSK disease, in primary care (GP surgeries) when compared to secondary care (hospital OPLs)?

The design of the schema aims to minimise the time cost of the task for clinicians, and avoid unnecessarily exposing annotators to the complexity of the terminologies being mapped to, while faithfully capturing the information required to address questions (1) and (2).

2 Annotation Task and Guidelines

The annotation task maps diagnoses in clinical text to a representation using SNOMED CT together with the HL7 FHIR information model. The diagnoses must be mapped such that the clinical questions (1) and (2) above can be addressed, where the classification of a patient into a category of MSK disease is based upon the resulting SNOMED CT and HL7 FHIR encoding.

We follow the principles proposed in [1], such as the delineation between *core* and *supporting* concepts. Therefore, the annotation schema must define: (i) *Core concepts*: the codes in SNOMED CT corresponding to the central focus of annotation, (ii) *Supporting concepts*: contextual qualifiers to interpret core concepts for the clinical task and (iii) how each of these should be interpreted when mapping from text to terminologies during the annotation task. An example of the template, applied to the domain of MSK diagnoses, is shown in Figure 1.

In this work, core concepts are SNOMED CT Clinical Findings. The annotation task is to identify each separate core concept within a diagnosis text and provide a corresponding Clinical Finding code, each of which is treated as a separate diagnosis. For each diagnosis, the annotators provide qualifiers (supporting concepts): *Verification Status* (Confirmed, Probable, Possible or Refuted), *Family History?* (Yes or No) and *Mentioned Date*. Verification Status indicates the

text	SNOMED_desc	SNOMED_ID	Verification Status	Family History?	Mentioned Date
...attended with AxSpA...	Axial Spondyloarthritis	723116002	Confirmed	No	
...AxSpA occurring in family, sent for...	Axial Spondyloarthritis	723116002	Confirmed	Yes	
...signs of ankylosing spondylitis, examined in 2016...	Ankylosing spondylitis	9631008	Possible	No	XX/XX/2016

Figure 1: Example of the template and annotations for MSK diagnoses. SNOMED_ID and SNOMED_desc correspond to the core concept. Green columns indicate qualifiers.

certainty of the diagnosis, and Family History indicates the subject of the diagnosis, i.e., the patient or a family member. These enable the identification of *definite* cases of MSK disease for patients in the population (prevalence), while the Mentioned Date is used to estimate incidence.

The template is designed to allow flexible mapping. For example, Verification Status can be mapped to the HL7 FHIR Condition resource, corresponding to the Condition.verificationStatus field, while Family History can be represented using the SNOMED CT Situation hierarchy.

3 Application

The annotated dataset will be used to categorise patients under one of five categories of MSK disease: axial spondyloarthritis (AxSpA), psoriatic arthritis (PsA), rheumatoid arthritis (RA), osteoarthritis (OA) and inflammatory arthritis (IA). Reference sets for each category were constructed by rheumatologists and an informatician familiar with SNOMED CT. Datasets to be annotated have been prepared with a pre-annotation step using a fine-tuned MedCAT [2] model, which identified and mapped the core concepts (Clinical Findings). Annotators will check the quality of these codes, suggesting alternatives if necessary, and annotate the qualifiers. Annotation quality will be evaluated using a graph-based evaluation over SNOMED CT [3] and qualitative measures on the suitability of the annotated data in addressing the clinical questions.

4 Conclusion

This work presents an annotation schema for capturing diagnoses from hospital OPLs, which is an essential step in producing high quality datasets focused on a particular clinical task. These datasets will be used to evaluate existing NER tooling and to address key epidemiological questions, both directly and by enabling the ongoing development and improvement of NLP tooling.

5 Study Context

HRA and Research Ethics Committee approval was obtained for this work, and is funded as part of the “Assembling the Data Jigsaw" project[1], funded by the Nuffield Foundation’s[2] Oliver Bird Fund and Versus Arthritis[3]. The views expressed are those of the authors and not necessarily the funders. All data annotation is performed in an NHS secure data environment.

[1] https://blogs.manchester.ac.uk/centre-for-epidemiology/assembling-the-data-jigsaw

[2] www.nuffieldfoundation.org

[3] www.versusarthritis.org

References

[1] Schulz S, Del-Pinto W, Han L, Kreuzthaler M, Aghaei S, Nenadic G. Towards principles of ontology-based annotation of clinical narratives. In: Proceedings of the International Conference on Biomedical Ontologies (ICBO); 2023. .

[2] Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. Artificial intelligence in medicine. 2021;117:102083.

[3] Del-Pinto W, Demetriou G, Jani M, Patel R, Gray L, Bulcock A, et al. Exploring the Consistency, Quality and Challenges in Manual and Automated Coding of Free-text Diagnoses from Hospital Outpatient Letters. arXiv preprint arXiv:231110856. 2023.