Title

Automated Extraction of Patient-Centered Outcomes After Breast Cancer Treatment: An Open-Source Large Language Model—Based Toolkit

Journal

JCO Clinical Cancer Informatics, 8(8)

ISSN

2473-4276

Authors

Luo, Man

Trivedi, Shubham

Kurian, Allison W

et al.

Publication Date

2024-08-01

DOI

10.1200/cci.23.00258

HHS Public Access

Author manuscript

JCO Clin Cancer Inform. Author manuscript; available in PMC 2025 August 01.

Published in final edited form as:

JCO Clin Cancer Inform. 2024 August ; 8: e2300258. doi:10.1200/CCI.23.00258.

Automated Extraction of Patient-Centered Outcomes following Breast Cancer Treatment: An Open-Source Large Language Model-Based Toolkit

Man Luo[a], **Shubham Trivedi**[a], **Allison W. Kurian**[b], **Kevin Ward**[d], **Theresa H.M. Keegan**[e], **Daniel Rubin**[c], **Imon Banerjee**[a,f,*]

[a]Department of Radiology, Mayo Clinic, Phoenix, Arizona, USA

[b]Departments of Medicine and of Epidemiology & Population Health, Stanford University School of Medicine, Palo Alto, California, USA

[c]Department of Biomedical Data Science, Radiology, and Medicine, Stanford University School of Medicine, Palo Alto, California, USA

[d]Rollins School of Public Health, Emory University, Atlanta, Georgia, USA

[e]Department of Internal Medicine, UC Davis School of Medicine, Sacramento, California, USA

[f]School of Computing and Augmented Intelligence, Arizona State University, Tempe, Arizona, USA

Abstract

PURPOSE -—Patient-Centered Outcomes (PCO) are pivotal in cancer treatment, as they directly reflect patients' quality of life (QoL). While multiple studies suggest that factors impacting breast cancer-related morbidity and survival are influenced by treatment side effects and adherence to long-term treatment, such data is generally only available on a smaller scale or from a single center. The primary challenge with collecting these data is that the outcomes are captured as free text in clinical narratives written by clinicians.

METHODS -—Given the complexity of PCO documentation in these narratives, computerized methods are necessary to unlock the wealth of information buried in unstructured text notes that often document PCOs. Inspired by the success of LLMs, we examined the adaptability of three LLMs: GPT-2, BioGPT, and PMC-LLaMA, on PCO tasks across three institutions, Mayo Clinic, Emory University Hospital (EUH), and Stanford University. We developed an open-source framework for fine-tuning LLM that can directly extract the five different categories of PCO from the clinic notes.

RESULTS -—We found that these LLMs without fine-tuning (zero-shot) struggle with challenging PCO extraction tasks, displaying almost random performance, even with some task-specific examples (few-shot learning). The performance of our fine-tuned, task-specific models is notably superior com- pared to their non-fine-tuned LLM models. Moreover, the fine-tuned GPT-2 model has demonstrated a significantly better performance than the other two larger LLMs.

* Corresponding author Banerjee.Imon@mayo.edu (Imon Banerjee).

CONCLUSION -—Our discovery indicates that although LLMs serve as effective general-purpose models for tasks across various domains, they require fine-tuning when applied to the clinician domain. Our proposed approach has the potential to lead more efficient, adaptable models for PCO information extraction, reducing reliance on extensive computational resources while still delivering superior performance for specific tasks.

Keywords

Large language models (LLM); few-shot generalization; patient center outcome (PCO); cancer treatment; side-effects

1. Introduction

Breast cancer treatment-related physical and mental health outcomes are not always detectable by laboratory diagnostic tests, but many can be gathered only through patient communications and are rarely documented in population-wide cancer registries.[1] There are two broad types of documentation strategies for patient-specific outcomes, including side effects: (1) *patient-reported outcomes (PRO)* — health outcomes directly reported by the patient in their language either via survey or questionnaire; and (2) *patient-centered outcomes (PCO)* — defined as “outcomes” that can only be tracked through patient-caregiver communication”, as defined by the PCORI.[2] Studies[3] show that routine collection of PRO can have a positive impact on patient-provider communication, (shared) decision-making, and symptom management. However, the collection of relevant PRO on time has traditionally been an extremely labor-intensive, and thus, it is difficult to scale the collection of PRO in routine clinic workflows. Some studies[4,5,6] have assessed the feasibility of monitoring PRO among oncology patients, and their findings suggest that monitoring PRO outside of clinic visits can reduce the severity and frequency of adverse symptoms and decrease the number of emergency room visits and hospitalizations.[7]

However, engaging oncology patients in such routine traditional monitoring activities (e.g., via apps or phone calls) is resource-intensive, and it only enables the collection of limited information from adherent patients. PCO, unlike PRO, are documented in a free-text format by clinicians (e.g., physicians, nurses) in nonverbal, verbal, face-to-face, or non-face-to-face methods.[8] Given the complexity of PCO documentation, computerized methods, including machine learning and natural language processing (NLP), are necessary to unlock the wealth of information buried in unstructured textual notes that often document PCO. A few rule-based and supervised NLP systems were proposed to extract PCO from clinical notes of cancer patients;[9,10] however such methods have limited generalizability when applied to different institutions, given variability in linguistic expression.

Large Language Models (LLMs) are universally built upon the Transformer architecture,[11] incorporating self-attention and multi-head attention mechanisms. This innovative design empowers LLMs to master con- textual representation in the free-text, essential for understanding the nuances of natural language. LLM[12] can be leveraged for the automatic interpretation of free-text clinical narratives by exploiting distributional semantics and contextual learning to provide adequate generalizability by addressing linguistic

variability.[13,14] The trajectory of clinical information extraction is aligned with the evolution of LMs from fine-tuning approaches with language models (LMs)[15,16]to zero-shot[17,18] and few-shot[19] approaches without any downstream task training. While LLMs have shown promise in general clinical information extraction, their application to PCO extraction remains underexplored, marking our study as a novel contribution to the field. Our work comprehensively investigates LLMs on PCO tasks by fine-tuning, zero-shot, and few-shots approaches.

To ensure confidentiality and compliance with privacy regulations, we focus open-source models such as LLama-2 models[20] Based on Llama-2, researchers have adapted these models to the biomedical or clinical domains.[21,22,23] We explore the application of state-of-the-art LLMs for PCO extraction tasks without any specific fine-tuning, using one generic model, GPT-2, and two extensively trained models in the biomedical field: BioGPT and PMC-LLaMA. We aim to extract five common side effects of breast cancer treatments - fatigue, nausea, anxiety, depression, and lymphedema. We also devised an efficient LLM fine-tuning paradigm to extract complex PCO data from clinic notes. We evaluated the performance of the model on both in-domain (Mayo Clinic) and out-of-domain (Emory University Hospital (EUH) and Stanford University) data.

2. Material and Methods

Figure 1 represents the pipeline of our proposed LLM modeling which parsed all the clinic notes of the patients as input and extracted the five targeted PCOs from the free text clinic notes for each encounter.

2.1. Annotation protocol

We aim to extract common treatment-related side effects following breast cancer therapy from multiple types of clinical notes, including nursing notes and oncology notes (see Supplementary text). We first created a detailed annotation protocol for each category by discussing with breast oncologists how we should classify text as attributing the PCO to breast cancer treatment (see Supplement table 1, 2). To create the “ground truth” for PCOs, three expert clinical readers manually reviewed the free-text notes and annotate if each PCO is positive or negative.

2.2. Text snippet extraction

To establish the vocabulary for the PCO extraction task, we compiled the following two complementary dictionaries: the target term list, which was a publicly available terminology program (Clinical Event Recognizer) extended with 46 additional terms by leveraging a combination of vocabularies from OBO-approved ontologies - MedDRA, NCIT, and Mental Health Management Ontology (Supplementary table 2); and the modifier list, which was a list of modifier terms, including negations (e.g., no, rule out), temporality (e.g., history, current), family (e.g., mother, sister) and discussion (e.g., risk of, may introduce). Finally, a keyword-based sentence retrieval method was applied to each clinic note, which selected only the sentences that contained at least one of the PCO-related terms as a named entity and generated a text snippet by combining the sentences extracted from the whole notes as

the PCO documentation can span across multiple sentence. We dropped all PCO that had historical temporality or discussion modifiers.

2.3. Pretrained Language Models

We investigate three LLMs; all of them are based on decoder-only models, which emphasize the use of self-attention mechanisms to process sequences in a unidirectional or auto-regressive manner and predict one-word token at a time. It leverages information only from previously seen tokens to predict the next token, and this will be fed back to the model input to predict a new token until the stopping criteria are satisfied, such as the end-of-sentence token is predicted or the maximum generation length is reached. Selection criteria of these three models was twofold. Firstly, we assess the significance of domain-specific pretraining on the performance of the PCO extraction task. GPT-2[24] serves as a non-domain-specific model since it has not been trained on clinical datasets. In contrast, BioGPT[25] and PMC-Llama have undergone extensive pretraining on biomedical literature and clinical notes data. Secondly, we explore how the size of a model influences its performance on the PCO extraction task. The models vary significantly in size: GPT-2 (0.22 billion parameters) is the smallest model, BioGPT is mid-sized (2.7 billion parameters), and PMC-Llama[22] is the largest (7 billion parameters). Our approach ensures a balanced examination of both domain- specific pretraining and model size, enabling us to draw nuanced conclusions about their respective impacts on the PCO extraction task. We describe each model in detail in Supplementary text and Supplementary Table 3 presents the comparison of these models.

2.4. In-context learning of LLMs

LLMs have in-context learning ability, which refers to the model's ability to understand and perform tasks based on the context provided in the input, without prior specific training on those tasks.[26] This is achieved by interpreting and adapting to the patterns or examples included in the input prompt, allowing the model to generate relevant responses or perform specific tasks based on this immediate context. There are two types of in-context learning that we will describe in the following.

Zero-shot.—In the zero-shot situation, a prompt is constructed as a question using a predefined template: "Based on the input text, does the patient have X?" Here, 'X' is substituted with each PCO (e.g., 'fatigue', 'depression', 'anxiety', 'nausea', 'lymphedema'), followed by a clinical note serving as the 'input text'. Given the prompt and the *input text*, the model will generate an answer, if the answer is "Yes", it indicates that the presence of 'X' is confirmed in the clinical note. Conversely, a "No" response signifies that 'X' is not present/confirmed.

Few-shot.—The prompt is the same as in the zero-shot learning case, but for each input text, we provide a few examples of the same task to the LLM. Each example is composed of an input text and a response ("Yes" or "No"). We use internal data (from Mayo) as the pool of in-context examples (see Supplementary Figure 1). Furthermore, we study a range of in-context examples to see the effects of the number of examples in the context. To avoid label bias, we evenly sample the examples for each label, for example, if we target 4 in-context examples, we will select 2 examples with a response of "Yes" and 2 examples

with a response of “No”. Previous work has also shown that the model tends to predict the same answer as the examples that are closer chronologically to the input.[27] To avoid such bias, in each inference, we randomly shuffle the order of the examples. Lastly, we select the in-context example from the same type of prompt. For example, if the prompt is about “Fatigue”, then, we only select examples that are also about “Fatigue”.

2.5. Fine-tuning of GPT-2 and BioGPT

Unlike in-context learning, where the trainable parameters of the models are not changed, here we update the model parameters by fine-tuning the GPT-2 and BioGPT models on the internal training dataset. We optimize the loglikelihood of the ground truth answer, mathematically, $\zeta = \sum_{i=1}^{K} logP(a_i|h, a_{:i})$ where K is the number of tokens in answer **a**, a_i is the i^{th} token in **a**, and a_0 corresponds to a special beginning of sequence (BOS) token, **h** is the input sequence token. Specifically, since the label of this task is either “Yes” or “No”, K is always 1. We use the same inference strategy discussed in the previous section. We initialize the model with the pre-trained weight of gpt-2 and BioGPT, the learning rate is set 2e-5, and the total training epoch is 10, batch size of 16. The optimizer chosen for this task was AdamW.[28]

2.6. Efficient Parameters Fine-tuning of PMC-LLaMA

Even the smallest LLama model (7B) is too big to be trained on most academic hardware. We apply LoRA (Localized Reweighting of Attention),[29] one of the state-of-the-art EPFT techniques, to train our model. We fine-tune PMC-Llama using the same Mayo Clinic training dataset that is utilized for the full fine-tuning of GPT-2 or BioGPT and applied the learning rate of 1e-4, the linear learning rate scheduler, the total training epoch of 5, and the batch size of 128. Such hyper-parameters are adapted from Alpaca-lora[1].

2.7. Inference Process

We studied three auto-regression models that generate the next tokens by conditioning on the input text. Theoretically, we use greedy search meth- ods[30] to generate the answer, meaning that every generated token is chosen based on the largest probability of the entire vocabulary. However, if we use auto-regressive generalization methods, the model tends to predict free-form answers (e.g. instead of answering “No”, the model generates a sentence “based on the information you provide, the patient does not have fatigue because ...”), which causes difficulty in the evaluation process since the performance will be dependent on post-processing of model’s answers. Thus, instead of generalization, following previous works,[31] we get the probabilities of the two labels “Yes” and “No” by first computing their individual loss values, then applying negative log-likelihood to these values, and finally using the *argmax* function to ascertain the model’s final prediction.

[1] https://github.com/tloen/alpaca-lora

3. Result

3.1. Cohort

We obtained IRB approvals with waivers of informed consent at each site for analyzing clinical notes for breast cancer patients. Table 1 presents the overall characteristics for the Mayo Clinic, Emory University Hospital (EUH), and Stanford University breast cancer cohorts. Supplementary Table 4 presents the training and validation datasets; only Mayo Clinic data are used for training and internal validation and Stanford and EUH datasets are used for external evaluation. We calculated the agreement (Cohen's kappa) between two annotators using 50 Mayo clinic notes and observed an agreement of 0.85. All data were annotated internally within the institutional firewall following the same annotation protocol (Sec. 2.1). The performance of the LLM models was assessed on both internal and external test sets where we calculated Precision, Recall, and F1 scores using the optimal operating point based on the Youden Index[32] derived from each ROC curve (see Supplementary Figure 2,3,4).

3.2. Zero-shot and Few-shots Model Performance

For in-context learning, we evaluate BioGPT, GPT2, and the PMC- Llama performances, and the results in terms of AUC are presented in Figure 2. First, when observing the zero-shot inference on these pre-trained LLMs, indicated by the 0 examples (x-axis), it is evident that the AUC scores frequently fall below 0.5 and the models are more inclined to classify the text as "No" rather than "Yes" in most cases. This might indicate that models have an inherent bias that represents the real-world distribution where negative PCO recordings are appearing much more than the positive ones.

Second, it is often observed that the performance in a few-shot context surpasses that of the zero-shot scenario. Nevertheless, increasing the quantity of demonstration examples for PCOs does not consistently lead to enhanced performance which could be due to the wide variety of ways that PCOs are described. Given the strict vocabulary for describing lymphedema, it shows consistent improvement across few-shot learning but the trend does not continue for the EUH as there is only a single positive case. Third, despite some variations in their performances, BioGPT and PMC- LLaMA showcase relatively comparable results. This similarity occurs notwithstanding the fact that PMC-LLama has been trained on a larger dataset and possesses a greater number of parameters. Finally, the overall zero-shot and few-shot outcomes for both models are inadequate, signifying that the PCO tasks pose a significant challenge to the LLMs. This underscores the need for domain-specific fine-tuning to enhance performance for complicated targeted information extraction tasks.

3.3. Fine-tuned Model Performance

In-domain Performance.—Table 2 presents the in-domain (Mayo test set) performance of three different LLM models. Among all models, the GPT-2 model achieves the highest performance across all metrics, despite possessing the least number of trainable parameters (0.22B). Conversely, the PMC-LLaMA model, despite being the largest model (7B), shows the worst results. This contrast suggests that an efficient parameter tuning approach (see Sec.

2.6) can tailor a model to a specific task, but it might not be as effective as fully fine-tuning a smaller model. Delving into the performance of GPT-2, we observe that the recall value is high while the precision value is moderate. This reflects that the model can capture most of the positive PCO cases which is very important in the clinical domains, on the other hand, the model predicts some false positive cases.

Out-of-domain Performance.—Table 2 highlights the performance of the models on independent data (Emory and Stanford test sets). Overall, both GPT-2 and BioGPT achieve better performance than PMC-LLaMA, and in terms of recall for PCO, all three models demonstrate satisfactory performance on the independent test sets. On EUH data, GPT-2 is better than BioGPT, with the best performance on two categories out of five (Fatigue, Anxiety). Three models show relative poor precision on 'Anxiety' as the models are identified anxiety which is not particularly related to breast cancer treatment outcome, e.g. *'Patient is anxious about the surgery'*. On Stanford data, BioGPT is better than GPT-2, with the best performance on three category out of five (Fatigue, Depression, Nausea). Nevertheless, GPT-2 is compatible with BioGPT on four categories except for 'depression'. To summarize, BioGPT and GPT-2 achieve similar out-of-domain performance, even though GPT-2 is much smaller than BioGPT, which demonstrates that fine-tuning a relative small model on domain-specific tasks is an efficient approach.

3.4. Qualitative Analysis

We visualize the attention map of a set of true-positive and false-positive examples of GPT-2 as being the model with the best performance on both internal and external test sets (Fig. 3). The line in the figures shows the attention between the positive word prediction 'Yes' and all the other words in the input text snippet and wider line width represents a higher attention value. As seen from true positive examples, the model is assigning more attention to the PCO-related words (e.g. fatigue, depression) and their confirmation (e.g. patient, about, procedure). False-positives primarily resulted due to PCO documentation in clinical notes that are not caused by cancer treatment side-effects, e.g. *'she is anxious to speak with the gynecologist.'* where the model is attending to the word 'anxious' but not to 'speak with' and missing the context of the PCO. Such instances, known as hard-negatives, present significant challenges for only identifying breast cancer treatment related PCOs.

4. Discussion

We designed a fine-tuning framework for LLMs for extracting treatment- related side effects following breast cancer therapy from multiple types of clinical notes. We compared the performance of light-weight (GPT2), middle-weight (BioGPT), and heavy- weight (LLaMA) LLMs for the same extraction task on the in-domain and out-of-domain test sets. We released our extraction code with the academic open-source license in Github[2] for community use to receive some feedback. To our knowledge, this is the first study that reports LLM frameworks that can extract breast cancer treatment-related side effects from

[2] https://github.com/imonban/pco_extraction_man

clinic notes and appear to generalize to data from independent institutions. Such a tool could be extremely useful for understanding breast cancer treatment-related side-effects.

Our experimental findings highlight that, although LLMs have shown re- markable capabilities in general domains, fine-tuning remains crucial for specialized tasks in the clinical domain. Moreover, fine-tuning enables smaller language models to rival larger ones in performance. Considering computational resource constraints, fully fine-tuning a smaller model might be more effective than partially fine-tuning a larger one. The fine-tuning approach also demonstrated that LLM models trained on data from one institution could effectively generalize to data from other institutions, although performance varies among PCO.

We also demonstrated that with the increasing number of in-context examples, the model performance increases for some PCOs. However, it comes at the cost of longer inference time and larger GPU memory. An interesting observation is that for some PCO that have limited variation in terms describing them, such as lymphedema, in-context prompt-based learning with more samples continuously improves performance; however, with more generic terms that can be described less specifically (e.g., fatigue), we do not observe the same improvement.

Our LLM-based targeted information extraction framework directly reads the temporal sequence of long clinic notes and extracts the targeted 5 PCOs from the free text. Fig. 4 shows a sample patient with 161 clinic notes with 36 notes that mentioned any PCO recording and 9 positive occurrences, and most of the positive PCO is recorded during surgery and after the start of hormone therapy.

For the PCO extraction task, there is a potential for use of simpler methods like TF-IDF with logistic regression to perform this task. However, their effectiveness is limited by the small training data sets,[33] leading us to choose LLMs with zero-shot and fine-tuning. Previous work[34] indicates that without fine-tuning, LLMs can’t independently make treatment recommendations but can assist oncologists. Our findings suggest that for clinical applications, LLMs must be fine-tuned to aid in decision-making, aligning with previous literature that underscores their supportive role in healthcare. We’ve used attention maps to highlight keywords in clinical notes, a technique to demystify transformer models (Fig. 3), and SHAP (Shapley Additive exPlanations)[35] can be another way to interpret the model prediction.[36]

Limitation.

The model has only been trained on five focused PCO labels in a controlled environment. We will extend the architecture to include additional PCO labels, such as suicidal ideation and pain. Moreover, our study focuses on extracting binary labels for each PCO, we plan to expand our analysis to include the severity of each PCO to enable more personalized healthcare interventions. To address the PCO label imbalance, we implemented down-sampling of the negative instances through random selection. Although this approach reduces bias towards predicting the negative label, it also limits the diversity of negative examples in our training set, potentially overlooking varied negative patterns. In future,

we plan to add data augmentation as a potential remedy that automatically generates more positive instances, enabling us to maintain a larger dataset without sacrificing negative data comprehensiveness. Lastly, our best fine-tuned model, GPT-2 achieves modest precision and resulted false positives, so the output may still requires manual review.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This work is supported by NIH/NCI, U01 CA269264–01-1, *'Flexible NLP toolkit for automatic curation of outcomes for breast cancer patients'* (PI: Banerjee). Stanford Oncoshare database was supported by Breast Cancer Research Foundation, the Susan and Richard Levy Gift Fund, the Suzanne Pride Bryan Fund for Breast Cancer Research, the Jan Weimer Junior Fac- ulty Chair in Breast Oncology, the Regents of the University of California's California Breast Cancer Research Program (16OB-0149 and 19IB- 0124), the BRCA Foundation, the G. Willard Miller Foundation, and the Biostatistics Shared Resource of the NIH-funded Stanford Cancer Institute (P30CA124435).

References

1. Schmidt Martina E, Scherer Sophie, Wiskemann Joachim, and Steindorf Karen. Return to work after breast cancer: The role of treatment- related side effects and potential impact on quality of life. European journal of cancer care, 28(4):e13051, 2019. [PubMed: 31033073]
2. Karamfilov Theodor, Konrad Helga, Karte Kerstin, and Wollina Uwe. Lower relapse rate of botulinum toxin a therapy for axillary hyperhidrosis by dose increase. Archives of dermatology, 136(4):487–490, 2000. [PubMed: 10768647]
3. Yang Luanyi Y, Manhas Domnick S, Howard A Fuchsia, and Olson RA. Patient-reported outcome use in oncology: a systematic review of the impact on patient-clinician communication. Supportive Care in Cancer, 26:41–60, 2018. [PubMed: 28849277]
4. Weaver Andrew, Young Annie M, Rowntree J, Townsend N, Pearson S, Smith J, Gibson O, Cobern W, Larsen M, and Tarassenko L. Application of mobile phone technology for managing chemotherapy-associated side- effects. Annals of Oncology, 18(11):1887–1892, 2007. [PubMed: 17921245]
5. Paladino Andrew J, Anderson Janeane N, Krukowski Rebecca A, Waters Teresa, Kocak Mehmet, Graff Carolyn, Blue Ryan, Jones Tameka N, Buzaglo Joanne, Vidal Gregory, et al. Thrive study protocol: a randomized controlled trial evaluating a web-based app and tailored messages to improve adherence to adjuvant endocrine therapy among women with breast cancer. BMC Health Services Research, 19:1–12, 2019. [PubMed: 30606168]
6. Kearney Nora, McCann Lisa, Norrie John, Taylor Lesley, Gray Peter, McGee-Lennon Mari- lyn, Sage Meurig, Miller Morven, and Maguire Roma. Eval- uation of a mobile phone-based, advanced symptom management system (asyms©) in the management of chemotherapy-related toxicity. Supportive Care in Cancer, 17:437–444, 2009. [PubMed: 18953579]
7. Cleeland Charles S, Gonin Rene, Hatfield Alan K, Edmonson John H, Blum Ronald H, Stewart James A, and Pandya Kishan J. Pain and its treatment in outpatients with metastatic cancer. New England Journal of Medicine, 330(9):592–596, 1994. [PubMed: 7508092]
8. Sonn Geoffrey A, Sadetsky Natalia, Presti Joseph C, and Litwin Mark S. Differing perceptions of quality of life in patients with prostate cancer and their doctors. The Journal of urology, 189(1):S59–S65, 2013. [PubMed: 23234635]
9. Forsyth Alexander W, Barzilay Regina, Hughes Kevin S, Lui Dickson, Lorenz Karl A, Enzinger Andrea, Tulsky James A, and Lind- vall Charlotta. Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. Journal of pain and symptom management, 55(6):1492–1499, 2018. [PubMed: 29496537]

10. Lindvall Charlotta, Lilley Elizabeth J, Zupanc Sophia N, Chien Isabel, Udelsman Brooks V, Walling Anne, Cooper Zara, and Tulsky James A. Natural language processing to assess end-of-life quality indicators in can- cer patients receiving palliative surgery. Journal of Palliative Medicine, 22(2):183–187, 2019. [PubMed: 30328764]
11. Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser L- ukasz, and Polosukhin Illia. Attention is all you need. Advances in neural information processing systems, 30, 2017.
12. Kojima Takeshi, Gu Shixiang Shane, Reid Machel, Matsuo Yutaka, and Iwasawa Yusuke. Large language models are zero-shot reasoners, 2022. URL https://arxiv.org/abs/2205.11916.
13. Chiang Chia-Chun, Luo Man, Dumkrieger Gina, Trivedi Shubham, Chen Yi- Chieh, Chao Chieh-Ju, Schwedt Todd J, Sarker Abeed, and Banerjee Imon. A large language model-based generative natural language pro- cessing framework finetuned on clinical notes accurately extracts headache frequency from electronic health records. medRxiv, 2023.
14. Li Hanzhou, Moon John T, Iyer Deepak, Balthazar Patricia, Krupinski Elizabeth A, Bercu Zachary L, Newsome Janice M, Banerjee Imon, Gichoya Judy W, and Trivedi Hari M. Decoding radiology reports: Potential ap- plication of openai chatgpt to enhance patient understanding of diagnostic reports. Clinical Imaging, 2023.
15. Wei Qiang, Ji Zongcheng, Si Yuqi, Du Jingcheng, Wang Jingqi, Tiryaki Firat, Wu Stephen, Tao Cui, Roberts Kirk, and Xu Hua. Relation ex- traction from clinical narratives using pre-trained language models. In AMIA annual symposium proceedings, volume 2019, page 1236. American Medical Informatics Association, 2019.
16. Yang Xi, Chen Aokun, PourNejatian Nima, Shin Hoo Chang, Smith Kaleb E, Parisien Christopher, Compas Colin, Martin Cheryl, Costa Anthony B, Flores Mona G, et al. A large language model for electronic health records. NPJ Digital Medicine, 5(1):194, 2022. [PubMed: 36572766]
17. Sezgin Emre, Hussain Syed-Amad, Rust Steve, and Huang Yungui. Ex- tracting medical information from free-text and unstructured patient- generated health data using natural language processing methods: fea- sibility study with real-world data. JMIR Formative Research, 7:e43014, 2023. [PubMed: 36881467]
18. Choi Hyeon Seok, Song Jun Yeong, Shin Kyung Hwan, Chang Ji Hyun, and Jang Bum-Sup. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiation Oncology Journal, 41(3):209, 2023. [PubMed: 37793630]
19. Agrawal Monica, Hegselmann Stefan, Lang Hunter, Kim Yoon, and Sontag David. Large language models are few-shot clinical information extractors. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1998–2022, 2022.
20. Touvron Hugo, Lavril Thibaut, Izacard Gautier, Martinet Xavier, Lachaux Marie- Anne, Lacroix Timothée, Rozière Baptiste, Goyal Naman, Hambro Eric, Azhar Faisal, et al. Llama: Open and efficient foundation lan- guage models. arXiv preprint arXiv:2302.13971, 2023.
21. Han Tianyu, Adams Lisa C, Papaioannou Jens-Michalis, Grundmann Paul, Oberhauser Tom, Löser Alexander, Truhn Daniel, and Bressem Keno K. Medalpaca–an open-source collection of medical conversational ai models and training data. arXiv preprint arXiv:2304.08247, 2023.
22. Wu Chaoyi, Zhang Xiaoman, Zhang Ya, Wang Yanfeng, and Xie Weidi. Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454, 2023.
23. Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. Chat- doctor: A medical chat model fine-tuned on llama model using medical domain knowledge. arXiv preprint arXiv:2303.14070, 2023.
24. Zhu Yukun, Kiros Ryan, Zemel Rich, Salakhutdinov Ruslan, Urta- sun Raquel, Torralba Antonio, and Fidler Sanja. Aligning books and movies: To- wards story-like visual explanations by watching movies and reading books. In The IEEE International Conference on Computer Vision (ICCV), December 2015.
25. Luo Renqian, Sun Liai, Xia Yingce, Qin Tao, Zhang Sheng, Poon Hoifung, and Liu Tie-Yan. Biogpt: generative pre-trained transformer for biomedi- cal text generation and mining. Briefings in bioinformatics, 23(6):bbac409, 2022.

26. Kaplan Jared, McCandlish Sam, Henighan Tom, Brown Tom B, Chess Ben- jamin, Child Rewon, Gray Scott, Radford Alec, Wu Jeffrey, and Amodei Dario. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.
27. Zhao Zihao, Wallace Eric, Feng Shi, Klein Dan, and Singh Sameer. Cal- ibrate before use: Improving few-shot performance of language models. In International Conference on Machine Learning, pages 12697–12706. PMLR, 2021.
28. Loshchilov Ilya and Hutter Frank. Decoupled weight decay regularization. In International Conference on Learning Representations, 2018.
29. Hu Edward J, Wallis Phillip, Allen-Zhu Zeyuan, Li Yuanzhi, Wang Shean, Wang Lu, Chen Weizhu, et al. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2021.
30. Scholak Torsten, Schucher Nathan, and Bahdanau Dzmitry. Picard: Pars- ing incrementally for constrained auto-regressive decoding from language models. arXiv preprint arXiv:2109.05093, 2021.
31. Min Sewon, Lyu Xinxi, Holtzman Ari, Artetxe Mikel, Lewis Mike, Han- naneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demon- strations: What makes in-context learning work? arXiv preprint arXiv:2202.12837, 2022.
32. Ruopp Marcus D, Perkins Neil J, Whitcomb Brian W, and Schisterman Enrique F. Youden index and optimal cut-point estimated from obser- vations affected by a lower limit of detection. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 50(3):419–430, 2008.
33. Elmarakeby Haitham A, Trukhanov Pavel S, Arroyo Vidal M, Riaz Irbaz Bin, Schrag Deborah, Van Allen Eliezer M, and Kehl Kenneth L. Em- pirical evaluation of language modeling to ascertain cancer outcomes from clinical text reports. BMC bioinformatics, 24(1):328, 2023. [PubMed: 37658330]
34. Benary Manuela, Wang Xing David, Schmidt Max, Soll Dominik, Hilfenhaus Georg, Nassir Mani, Sigler Christian, Knödler Maren, Keller Ulrich, Beule Dieter, et al. Leveraging large language models for decision support in personalized oncology. JAMA Network Open, 6(11):e2343689–e2343689, 2023.
35. Lundberg Scott M and Lee Su-In. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, 2017.
36. Lee Siryeol, Lee Juncheol, Park Juntae, Park Jiwoo, Kim Dohoon, Lee Joohyun, and Oh Jaehoon. Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage. The American Journal of Emergency Medicine, 77:29–38, 2024. [PubMed: 38096637]

Context Summary

Key objectives:

Develop an open-source natural language processing system to extract to extract five common side effects of breast cancer treatments - fatigue, nausea, anxiety, depression, and lymphedema, from free-text clinic notes.

Knowledge generated:

Although Large Language Models (LLM) have shown re- markable capabilities in general domains, fine-tuning remains crucial for specialized tasks such as side-effect extraction in the clinical domain. Moreover, fine-tuning enables smaller LLM to rival larger ones in performance for the complex extraction task.

Relevance:

Patient-Centered Outcomes (PCOs) are vital in cancer treatment as they offer significant insights into the quality of life (QoL) of patients. Studies indicates that treatment side effects impact morbidity and survival rates in breast cancer and influence adherence to prolonged treatment regimens. However, PCOs are often difficult to access as they are embedded within clinical notes.

Large Language Models (LLMs) have shown promise in clinical data extraction tasks. This study presents strategies, including fine-tuning and other advanced approaches, that are essential for comprehending complex language structures. Zero-shot learning is insufficient for these tasks, underscoring the need for more tailored methods to effectively extract and utilize PCOs from clinical documentation.

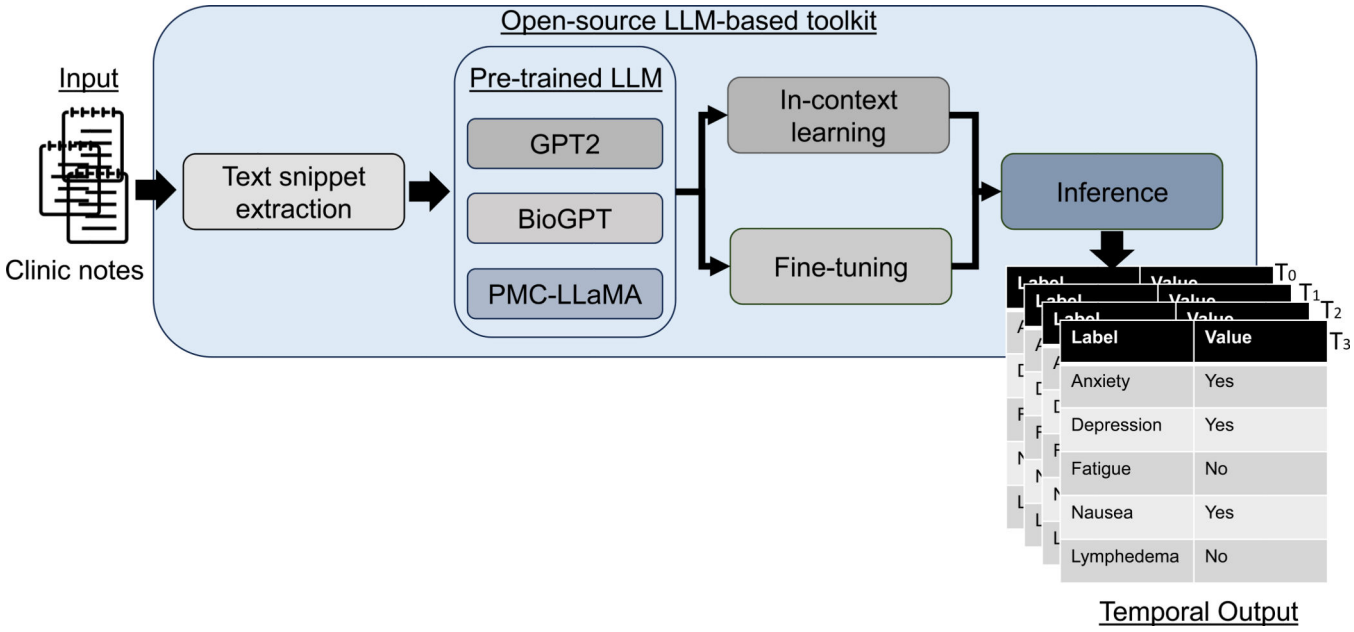

Figure 1.

LLM based PCO extraction pipeline. *(Input)* Free-text clinic notes, *output* PCO extracted from each encounters at different timepoints (T_n).

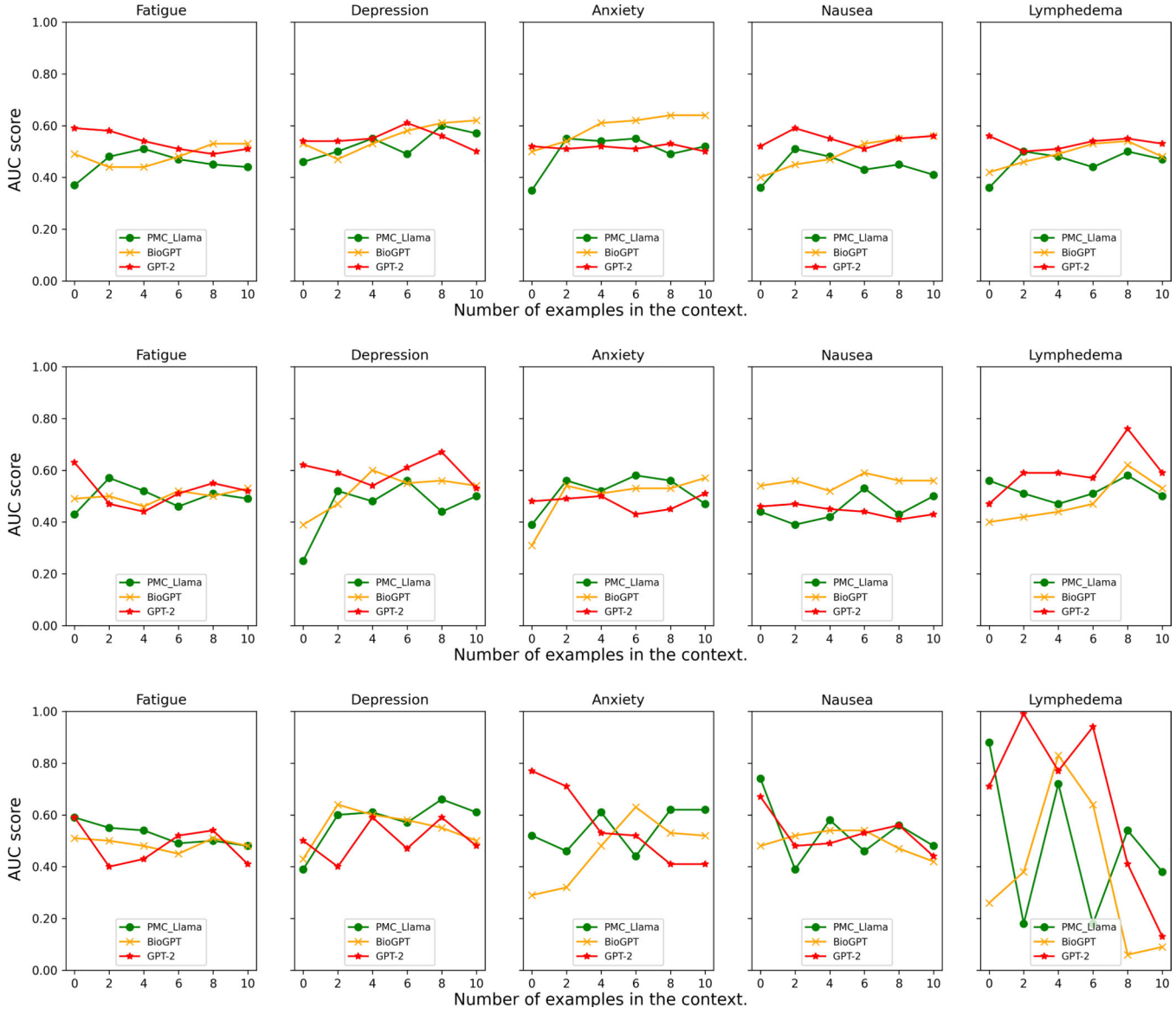

Figure 2.

The AUC score of the PMC-LLama2, BioGPT, and GPT-2 models on the PCO task - (top) Mayo test set, (middle) Stanford dataset, and (bottom) on Emory dataset. The X-axis represents the number of in- context examples and the Y-axis shows the corresponding AUC value.

Figure 3.

Attention map visualization of a true positive (top) and false positive (bottom) example of identifying each PCO using the fine-tuned GPT-2 model.

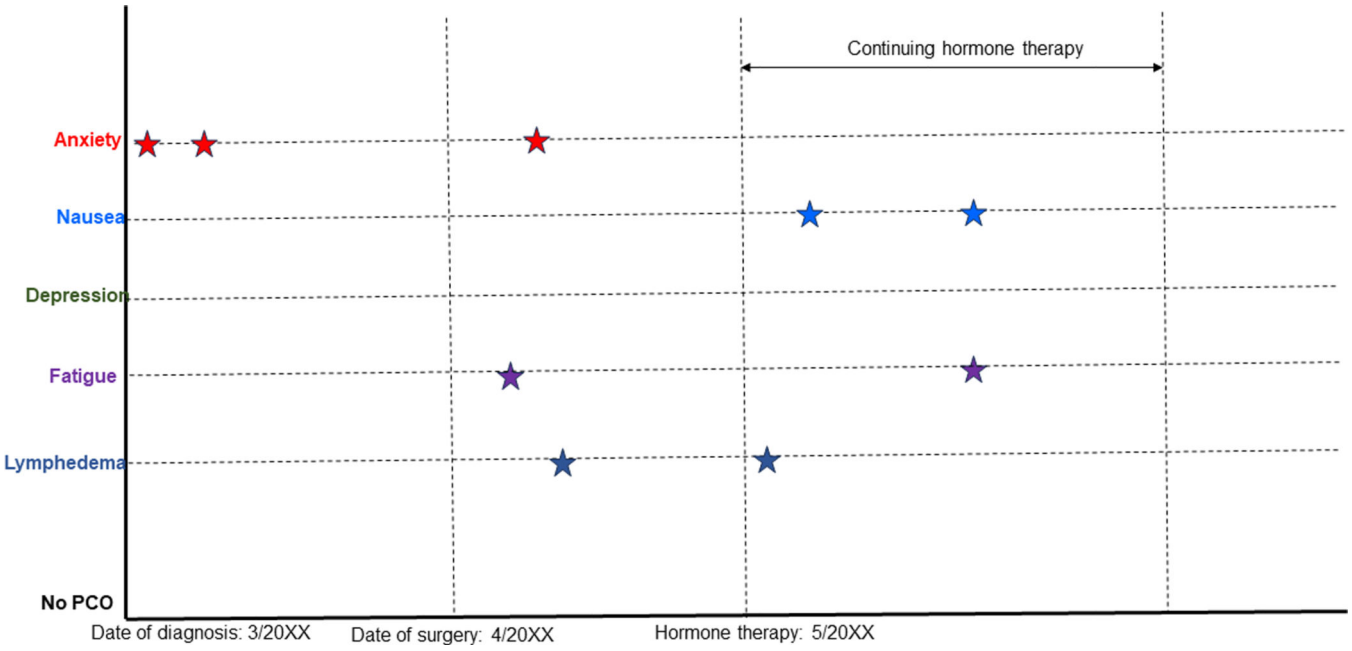

Figure 4.
Derived PCO timeline for a sample patient with 116 clinic notes

Table 1:

Cohort characteristics: Mayo (internal) and Stanford and Emory (EUH) (external).

	Mayo clinic (%)	**EUH (%)**	**Stanford (%)**
Total patients	26,692	33,077	8,956
No. of clinic notes/patient	6,516,013	2,846,987	1,065,400
Average length of notes (no. of words)	844 (+/− 762)	756 (+/− 643)	1023 (+/− 232)
Age - mean and std	65 (+/− 20) yrs	50 (+/− 10)	54 (+/−13)
Race			
White	14,170 (93)	15546 (47)	6,726(75)
Black	343 (2.5)	17,200(52)	325(4)
Asian	245 (1.5)	331 (1%)	1,353(15)
American Indian	82 (0.5)	0	17(1)
Unknown	361 (2.5)	0	486(5)
Ethnicity			
Hispanic	326 (2.5)	661 (2%)	842(9)
Non Hispanic	14,547 (95)	32,000(97%)	7649(85)
Unknown	328 (2.5)	416(1.2%)	465(5)
Nuclear grade			
I (well differentiated)	4,003 (26.5)	-	1609 (18)
II (moderately differentiated)	6,683 (44)	-	3,363(38)
III (poorly differentiated)	4,191 (27.5)	-	3,984(44)
Unknown	324 (2)	-	–
Generation chemotherapy			
Generation 1	258 (2)	-	-
Generation 2	1,397 (9)	-	-
Generation 3	4,231 (28)	-	-
No chemotherapy	9,315 (61)	-	-
Train/Test Data Split (note level)			
Train	3924	-	-
Test	1450	474	525

Table 2:

AUROC, Precision, Recall, and F1 scores of three models. The optimal operating point is chosen based on the ROC. 95% confidence interval is calculated using bootstrapping.

Cohort	PCO	GPT-2				BioGPT				PMC-Llama			
		AUC	Precision	Recall	F1	AUC	Precision	Recall	F1	AUC	Precision	Recall	F1
Mayo in-domain	Fatigue	**0.97**	0.66	0.94	0.72	0.96	0.65	0.91	0.71	0.90	0.60	0.84	0.62
		[0.97,0.98]				[0.95,0.98]				[0.88, 0.93]			
	Depression	**0.99**	0.65	0.98	0.72	**0.99**	0.61	0.96	0.67	0.91	0.55	0.87	0.54
		[0.98,0.99]				[0.95,0.98]				[0.88,0.93]			
	Anxiety	**0.97**	0.61	0.92	0.65	0.94	0.61	0.89	0.65	0.81	0.56	0.78	0.57
		[0.97,0.98]				[0.93,0.98]				[0.78,0.85]			
	Nausea	**0.97**	0.65	0.91	0.70	0.95	0.60	0.89	0.63	0.92	0.58	0.85	0.59
		[0.96,0.98]				[0.94, 0.96]				[0.90,0.94]			
	Lymphedema	**0.97**	0.66	0.95	0.73	**0.97**	0.62	0.93	0.66	0.88	0.57	0.83	0.59
		[0.97,0.98]				[0.96,0.98]				[0.85,0.93]			
EUH out-of-domain	Fatigue	**0.95**	0.94	0.97	0.95	0.94	0.91	0.95	0.93	0.92	0.88	0.91	0.89
		[0.95,0.96]				[0.94,0.97]				[0.91,0.95]			
	Depression	0.90	0.88	0.92	0.90	**0.92**	0.90	0.95	0.92	0.89	0.80	0.86	0.81
		[0.89,0.94]				[0.92,0.94]				[0.88,0.91]			
	Anxiety	**0.86**	0.62	0.92	0.65	0.85	0.61	0.91	0.63	0.80	0.57	0.80	0.57
		[0.84,0.88]				[0.83,0.87]				[0.77,0.82]			
	Nausea	0.97	**0.98**	0.94	0.96	0.98	0.93	0.94	0.94	0.96	0.85	0.90	0.87
		[0.96,0.98]				[0.97,0.99]				[0.95,0.97]			
	Lymphedema	**1.0**	1.00	1.00	1.00	1.0	1.00	1.00	1.00	1.0	1.00	1.00	1.00
		[1.0,1.0]				[1.0,1.0]				[1.0,1.0]			
Stanford out-of-domain	Fatigue	0.94	0.78	0.89	0.82	**0.96**	0.71	0.91	0.75	0.90	0.69	0.85	0.72
		[0.92,0.98]				[0.94,0.97]				[0.85,0.94]			
	Depression	0.82	0.74	0.85	0.78	**0.97**	0.65	0.94	0.70	0.88	0.60	0.85	0.61
		[0.74,0.93]				[0.95,0.98]				[0.80,0.95]			
	Anxiety	**0.95**	0.78	0.93	0.83	0.94	0.63	0.88	0.67	0.84	0.60	0.81	0.61
		[0.94,0.98]				[0.92,0.97]				[0.77,0.92]			
	Nausea	0.91	0.75	0.87	0.79	**0.96**	0.76	0.91	0.81	0.92	0.64	0.85	0.66
		[0.88,0.97]				[0.94,0.98]				[0.86,0.96]			
	Lymphedema	**0.92**	0.61	0.90	0.64	0.91	0.58	0.83	0.59	0.88	0.58	0.81	0.59
		[0.88,0.97]				[0.85,0.96]				[0.79,0.97]			