Responsible NLP Checklist

Paper title: *AutoCT: Automating Interpretable Clinical Trial Prediction with LLM Agents*
Authors: *Fengze Liu, Haoyu Wang, Joonhyuk Cho, Dan Roth, Andrew Lo*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ **A. Questions mandatory for all submissions.**

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☒ A2. Did you discuss any potential risks of your work?
We did not provide a dedicated discussion of risks because our work is limited to publicly available and anonymized clinical trial metadata from TrialBench (Chen et al., 2024). Our system does not access or process personally identifiable information (PII), patient records, or sensitive clinical notes. As such, the risks of data misuse or harm to individuals are minimal. Moreover, our framework is intended strictly for research purposes and not for direct clinical deployment. Nevertheless, we acknowledge that any predictive system for clinical trials may be misinterpreted as offering medical advice; to mitigate this, we clearly position our work as a methodological contribution for research and benchmarking, not for decision-making in ongoing trials.

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☑ B1. Did you cite the creators of artifacts you used?
4.2

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
No. We did not explicitly discuss licenses or terms of use because our work relies solely on publicly available datasets (TrialBench (Chen et al., 2024)) and standard machine learning libraries with well-established open-source licenses. The dataset itself is already distributed under research-use terms, and we did not modify or redistribute it. For the artifacts we create (trained models, intermediate features, and code), we intend to release them under an open-source license upon acceptance, which will be clearly documented at the time of release.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes. Our use of the TrialBench dataset is fully consistent with its intended purpose, as it was explicitly released for research on clinical trial outcome prediction. We did not repurpose the data beyond this scope. For the artifacts we create (feature sets, models, and code), we specify their intended use for research purposes only, in line with the original access conditions of the dataset. We explicitly do not recommend or permit their use in clinical practice or decision-making outside research contexts.

☑ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
Yes. The TrialBench dataset is derived entirely from publicly available, de-identified sources such as ClinicalTrials.gov and PubMed, which do not contain personally identifiable information or offensive content. As such, no additional anonymization was required. We carefully verified that our use of the dataset did not involve sensitive attributes or patient-level identifiers, and all experiments were conducted strictly at the aggregated trial level.

N/A B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
(left blank)

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
4.2

☑ **C. Did you run computational experiments?**

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4.1

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4.4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
Yes. We primarily relied on standard Python packages such as scikit-learn for classical machine learning model training and evaluation. We used the default implementations of models to focus the evaluation on the feature construction process rather than extensive hyperparameter optimization. We report this explicitly in Limitations section.

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

N/A D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
(left blank)

N/A D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
(left blank)

N/A D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
(left blank)

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☒ N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(left blank)

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☒ N/A E1. If you used AI assistants, did you include information about their use?
(left blank)