Data Sheet for CT-Pub Dataset

Contact: Nafis Neehal (neehan@rpi.edu), Kristin P. Bennett (bennek@rpi.edu)

1 Data Sheet

Motivation

1. **For what purpose was the dataset created? (Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)**

   The dataset was created to address the specific task of evaluating the ability of language models (LMs) to aid in the design of clinical studies by accurately identifying baseline features of clinical trials. The motivation was to fill a gap in assessing how well AI models can determine these crucial features, which are essential for characterizing study cohorts, validating results, and estimating treatment effects in observational studies. The CT-Pub dataset aims to provide a standardized benchmark, CTBench, which facilitate the development and evaluation of AI models in this domain. This benchmark is intended to advance research on AI's role in clinical trial design, enhancing the efficacy and robustness of clinical trials.

2. **Who created this dataset (e.g., which team, research group), and on behalf of which entity (e.g., company, institution, organization)?**

   The CT-Pub dataset was created by Nafis Neehal, Bowen Wang, and Shayom Debopadhaya, Soham Dan, Keerthiram Murugesan, Vibha Anand and Kristin P. Bennett. At the time of creation, Nafis was a PhD Candidate at RPI (CS), Bowen was a Postdoc at Center of Biotechnology and Interdisciplinary Studies at RPI, Shayom was a student at Albany Medical College, Soham, Keerthiram and Vibha are research scientists and collaborators from IBM research and Kristin is a professor of Mathematical Sciences at RPI.

3. **Who funded the creation of the dataset? (If there is an associated grant, please provide the name of the grantor and the grant name and number.)**

   The dataset creation was supported by IBM Research and the Rensselaer Institute for Data Exploration and Applications.

4. **Any other comments?**

   N/A

Composition

1. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? (Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.)**

   Each instance in the dataset represents a clinical trial study. It includes various textual information about the clinical trial, such as the title, brief summary, conditions, interventions,

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

primary outcome, eligibility criteria, and baseline features collected both from the API and from related publications (see paper for details).

2. **How many instances are there in total (of each type, if appropriate)?**

   There are a total of 100 instances, each about a single clinical trial.

3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? (If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).)**

   This dataset is a randomly selected subset of the 1690 CT-Repo dataset. It includes additional information about the trials, specifically baseline features collected from publications, which are not present in the original CT-Repo dataset.

4. **What data does each instance consist of? ("Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.)**

   Each instance contains unprocessed texts for all features. See Table 1 in paper.

5. **Is there a label or target associated with each instance? If so, please provide a description.**

   For the CT-Pub dataset, the target is to predict the baseline features in the Paper_BaselineMeasures column.

6. **Is any information missing from individual instances? (If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.)**

   N/A

7. **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? (If so, please describe how these relationships are made explicit.)**

   Instances are unrelated, each instance is about a separate clinical trial.

8. **Are there recommended data splits (e.g., training, development/validation, testing)? (If so, please provide a description of these splits, explaining the rationale behind them.)**

   There are no data splits as no training/development/validation/testing is involved in our study.

9. **Are there any errors, sources of noise, or redundancies in the dataset? (If so, please provide a description.)**

   Data has been curated to the best of our ability. We believe there are no further errors (removed a few erroneous keywords as baseline features, such as - 'Continuous', see paper) or redundancies (removed a few duplicate trials).

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? (If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)**

    Dataset is self-contained.

11. **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? (If so, please provide a description.)**

No. All raw data in the dataset is from public sources (i.e. data from `clinicaltrials.gov` and publications).

12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? (If so, please describe why.)**

    N/A

13. **Does the dataset relate to people? (If not, you may skip the remaining questions in this section.)**

    N/A

14. **Does the dataset identify any subpopulations (e.g., by age, gender)? (If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.)**

    N/A

15. **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? (If so, please describe how.)**

    N/A

16. **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? (If so, please provide a description.)**

    N/A

17. **Any other comments?**

    N/A

Collection Process

1. **How was the data associated with each instance acquired? (Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.)**

   Clinical Trial MetaData was reported in `clinicaltrials.gov` accessible through API. Additional features (e.g., baseline features from publications) were collected manually through human effort.

2. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? (How were these mechanisms or procedures validated?)**

   Trial metadata were collected using publicly available API. Additional features were collected manually from each clinical trial-related publication.

3. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

   The 100 instances in the CT-Pub dataset were randomly sampled from the CT-Repo dataset.

4. **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

   All the co-authors of the paper were involved in the data collection process.

5. **Over what timeframe was the data collected? (Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news**

articles)? If not, please describe the timeframe in which the data associated with the instances was created.)

The data was collected/curated during March-April 2024. However, the clinical trials themselves have varying start and end dates, spanning several months/years.

6. **Were any ethical review processes conducted (e.g., by an institutional review board)? (If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.)**

   N/A

7. **Does the dataset relate to people? (If not, you may skip the remaining questions in this section.)**

   N/A

8. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

   N/A

9. **Were the individuals in question notified about the data collection? (If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.)**

   N/A

10. **Did the individuals in question consent to the collection and use of their data? (If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.)**

    N/A

11. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? (If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).)**

    N/A

12. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? (If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.)**

    N/A

13. **Any other comments?**

    N/A

Preprocessing/Cleaning/Labeling

1. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? (If so, please provide a description. If not, you may skip the remainder of the questions in this section.)**

   Originally, we started with around 1800 trials. After thorough preprocessing steps, including removing duplicate trials and those with missing values, we were left with 1693 trials for our final study (CT-Repo dataset). From these 1693 trials, we use 3 trials as examples for few-shot setting, and from the remaining 1690 trials we randomly selected 100 studies and created the CT-Pub dataset, which includes additional information.

2. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? (If so, please provide a link or other access point to the "raw" data.)**

Yes. Available in the same GitHub repository.

3. **Is the software used to preprocess/clean/label the instances available? (If so, please provide a link or other access point.)**

   Yes. The code is available in the GitHub repository.

4. **Any other comments?**

   N/A

Uses

1. **Has the dataset been used for any tasks already? (If so, please provide a description.)**

   The dataset has been used to benchmark State-of-the-art LLM's performance in predicting baseline features using clinical trial Metadata. We present detailed description of our experiment and data-usage throughout the paper.

2. **Is there a repository that links to any or all papers or systems that use the dataset? (If so, please provide a link or other access point.)**

   N/A - We are the first to release and use this dataset.

3. **What (other) tasks could the dataset be used for?**

   The dataset can be used for various studies, including making decisions about selecting different clinical trial design factors.

4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? (For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?)**

   N/A

5. **Are there tasks for which the dataset should not be used? (If so, please provide a description.)**

   N/A

6. **Any other comments?**

   N/A

Distribution

1. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? (If so, please provide a description.)**

   Yes, the dataset is freely available and accessible.

2. **How will the dataset be distributed (e.g., tarball on website, API, GitHub)? (Does the dataset have a digital object identifier (DOI)?)**

   Dataset is free for download at `https://github.com/nafis-neehal/CTBench_LLM`.

3. **When will the dataset be distributed?**

   The dataset is distributed as of June 2024 in its first version.

4. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? (If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.)**

   The dataset is distributed under CC0 1.0 Universal license.

5. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances? (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.)**

   Not to our knowledge.

6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? (If so, please describe these restrictions, and provide a link or other access point to any supporting documentation.)**

   Not to our knowledge.

7. **Any other comments?**

   N/A

Maintenance

1. **Who is supporting/hosting/maintaining the dataset?**

   Nafis Neehal is maintaining the dataset on his GitHub. Any further changes would be announced through the GitHub repo link.

2. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

   E-mail addresses are at the top of this document.

3. **Is there an erratum? (If so, please provide a link or other access point.)**

   Currently, no. As errors are encountered, future versions of the dataset may be released (but will be versioned). They will all be provided in the same GitHub location with proper announcements.

4. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? (If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?)**

   Same as previous.

5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? (If so, please describe these limits and explain how they will be enforced.)**

   No.

6. **Will older versions of the dataset continue to be supported/hosted/maintained? (If so, please describe how. If not, please describe how its obsolescence will be communicated to users.)**

   Yes; all data will be versioned.

7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? (If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.)**

   Errors may be submitted by emailing the authors. Further extensive augmentations may be accepted at the authors' discretion.

8. **Any other comments?**

   N/A