Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Clinical Trial Duration Prediction Dataset: A Resource for AI Research

Yuta Nonomiya

nonomiya.yuta@omu.ac.jp

Graduate School of Medicine, Osaka Metropolitan University

Muhammad Nabeel Asim

German Research Center for Artificial Intelligence GmbH (DFKI)

Yume Takahashi

Graduate School of Medicine, Osaka Metropolitan University

Koichi Kise

Graduate School of Informatics, Osaka Metropolitan University

Ayumi Shintani

Graduate School of Medicine, Osaka Metropolitan University

Andreas Dengel

German Research Center for Artificial Intelligence GmbH (DFKI)

---

data-descriptor

Keywords: Clinical Trial, Clinical Trial Duration Prediction, Machine Learning, Artificial Intelligence, Clinical Trial Dataset, AI-supported Clinical Trial

Posted Date: February 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8852939/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

The Clinical Trial Duration Prediction Dataset: A Resource for AI Research

Yuta Nonomiya[1,2*], Muhammad Nabeel Asim[2], Yume Takahashi[1], Koichi Kise[3], Ayumi Shintani[1], Andreas Dengel[2,4]

[1*]Department of Medical Statistics, Graduate School of Medicine, Osaka Metropolitan University, Osaka, 545-8585, Japan.
[2]German Research Center for Artificial Intelligence GmbH (DFKI), Kaiserslautern, 67663, Germany.
[3]Department of Core Informatics, Graduate School of Informatics, Osaka Metropolitan University, Sakai, 599-8531, Japan.
[4]Department of Computer Science, Rhineland-Palatinate Technical University of Kaiserslautern-Landau, Kaiserslautern, 67663, Germany.

*Corresponding author(s). E-mail(s): nonomiya.yuta@omu.ac.jp;

Abstract

Early forecasting of clinical trials duration supports planning, optimized resource allocation, and cost control. Developing effective AI predictors remains challenging, requiring collaboration between medical and computational expertise. Although a publicly available dataset exists for this task, it is not well-suited for development of an AI-driven predictor capable of accurately predicting clinical trials duration. This limitation arises from its lack of critical features that significantly influence trials length, as well as the presence of incompletely and inappropriately selected trials. To empower development of AI predictors, we introduce the Clinical Trial Duration Prediction (CTDP) dataset, derived from 81,943 completed drug trials across 1,520 ICD-10 categories. The CTDP dataset contains 51 features on clinical trial characteristics including phase, purpose, condition, eligibility, enrollment, arm information, design, endpoint, site, sponsor, and summary that together influence trial duration. CTDP dataset development through collaboration between clinical trial experts and AI researchers will enable the creation of robust and accurate AI-driven applications for predicting clinical trial durations.

Keywords: Clinical Trial, Clinical Trial Duration Prediction, Machine Learning, Artificial Intelligence, Clinical Trial Dataset, AI-supported Clinical Trial

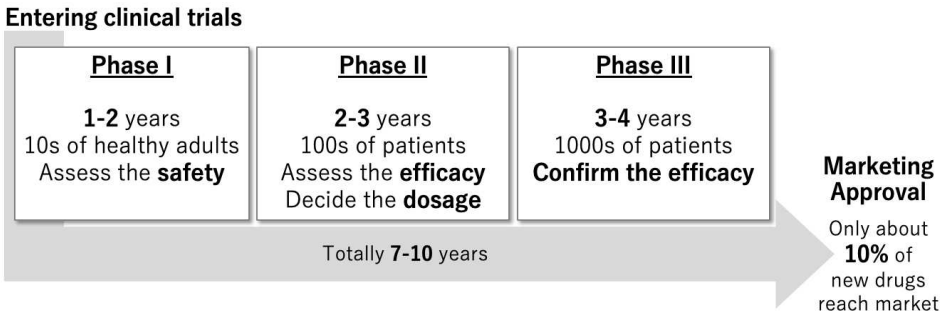


Fig. 1 Overview of typical clinical trial phases.

1 Background & Summary

Clinical trials are prospective intervention studies with human subjects to investigate experimental drugs, new treatments, medical devices, or clinical procedures, under rigorously specified conditions[1]. In drug development, they are typically conducted in sequential phases: Phase I (1–2 years) primarily evaluates safety, Phase II (2–3 years) tests proof of concept, and Phase III (3–4 years) confirms efficacy[2–4]. Finally, only about 10% of investigational drugs that entered Phase I trials are approved for marketing[4], after which Phase IV studies provide post-marketing surveillance[2](summarized in Fig. 1). Taken together, the entire process often requires 7–10 years, and more than 80% of trials exceed their planned timelines[5, 6]. Trial delays adversely affect both patients waiting for new therapies and pharmaceutical companies themselves[7–9]. A recent study estimates that each day of delay leads to approximately $500,000 lost prescription drug or biologic sales, and the direct daily cost of conducting a clinical trial is approximately $40,000 per day for Phase II and III clinical trials[10]. In this context, early forecasting of clinical trial duration facilitates enhanced strategic planning, optimized resource allocation, and a significant reduction in overall costs and risks[11–15].

The development of AI-driven applications for predicting the duration of clinical trials is currently an active and rapidly evolving area of research.[11–14]. Long et al. develop an AI capable of predicting duration of Phase I lymphoma clinical trials using machine learning methods in a limited domain[11]. Yue et al. are the first to formulate this challenge as a machine learning problem applicable to general clinical trials and propose a novel hierarchical attention mechanism to capture key features information within clinical trials[12]. Patrone et al. compare the predictive performance of their neural network model-estimated trial durations with that of sponsor-estimated durations for oncology trials[14]. Cho et al. apply both survival analysis and machine learning models to this task[13]. The establishment of a publicly available benchmark dataset is essential to ensure reproducibility and fair comparison among AI predictors. However, most prior studies[11–14] rely on in-house datasets, making systematic comparison across models difficult.

At present, TrialBench[15] is the only publicly available dataset specifically constructed for clinical trial duration prediction. While this dataset represents an important step toward benchmarking AI models in this domain, its design reflects a broader scope that does not fully align with the requirements of prospective duration prediction for drug development trials.

Specifically, TrialBench includes ongoing studies for which actual trial duration is not yet defined, as well as interventional studies that do not involve drug interventions, such as device- or behavior-based trials. In addition, several trial-level attributes that are known to influence trial timelines—such as primary endpoint characteristics, planned enrollment size, and explicit arm–intervention structures—are not consistently available in a structured form.

From a modeling perspective, TrialBench adopts random data splits across trials, which may result in temporal information leakage when models are evaluated on studies conducted earlier than those included in the training set. Such a setting differs from the practical use case in which trial duration is predicted prospectively based solely on information available at the time of trial initiation.

These characteristics limit the direct applicability of TrialBench for developing and evaluating AI models intended to support real-world trial planning and operational decision-making in drug development contexts.

Following the need of public dataset, the paper in hand introduces the Clinical Trial Duration Prediction (CTDP) dataset, a resource designed to support systematic development and evaluation of AI predictors. Importantly, the dataset is designed through a collaboration between clinical trial experts and AI researchers, mobilizing expertise from both domains to resolve these limitations and create a resource suitable for both reproducible AI model development and clinically meaningful trial planning.
Main characteristics of the CTDP dataset are as follows:

- Scale: 81,943 completed or terminated drug trials across 1,520 ICD-10 categories.
- Scope of features: 51 features covering trial phase and purpose, condition, eligibility criteria, planned enrollment, arm and intervention, design, endpoint, site, sponsor, and summary.
- Arm–intervention structure: hierarchical relationships between trial arms and interventions are retained.
- LLM-derived information: outcome- and intervention-related data extracted and structured from free-text using large language models (LLMs).
- Temporal data split: partitioning of data based on calendar time to enable prospective evaluation and to prevent information leakage.

By making this dataset publicly available, we aim to provide a foundation for reproducible research and to enable systematic comparison among methods for predicting clinical trial duration.

2 Methods

The following sub-sections illustrate the data source and eligible studies, target label, input features and data split. In addition, we include domain-specific notes titled **Clinical trial perspective** to explain, where appropriate, the rationale for methodological choices from the viewpoint of practical trial operations.

2.1 Data source and eligible studies

On March 5, 2025, we downloaded XML files of all registered clinical studies from ClinicalTrials.gov (CTG) [16](https://classic.clinicaltrials.gov/AllPublicXML.zip). We extracted all clinical studies categorized as "Interventional study", defined as clinical trials in any phase, and whose duration could be calculated. Additionally, unlike TrialBench[15], we applied stricter eligibility criteria: trials had to be marked as "Completed" or "Terminated" and include at least one intervention labeled as "Drug" or "Biological" in the Intervention Type field of CTG. In this study, such trials were collectively referred to as drug trials. This definition was adopted to ensure a reproducible, label-based selection rule using the original registry metadata. Because the distinction between drugs and biologics is not always consistently represented in the source data, these two categories were grouped together and treated as a single class of pharmacological interventions.

Clinical trial perspective. Studies registered in CTG are classified as either *Interventional* or *Observational*. Interventional studies involve prospective assignment of participants to interventions, whereas observational studies collect data without intervention assignment. By definition, clinical trials are categorized as "Interventional study". It is also critical to distinguish between ongoing and completed studies. For trials that are still in progress, the study periods recorded in CTG are usually only the planned duration defined by the *sponsor* (the individual, company, or organization responsible for initiating and managing the trial). It should also be noted that CTG includes interventional trials not limited to drug development; device, behavioral, or procedural interventions are also registered. Because our primary interest lies in drug development, we added the condition that at least one *arm* (a predefined participant group assigned to receive specific interventions) involve a drug intervention. However, CTG does not disclose whether a given trial was specifically conducted for drug development purposes, which represents an inherent limitation when using CTG as a data source. As a minimal safeguard, we additionally conditioned on phase information to better approximate the set of drug development trials.

2.2 Ethics statement

This study used only publicly available registry data from ClinicalTrials.gov. The dataset contains trial-level information on study design and characteristics and does not include any individual-level or participant-derived data, nor any identifiable private information. The study did not involve human subjects and does not constitute human subjects research. Ethics approval and informed consent were not required.

Table 1 Trial Duration by Phase

Phase	Dataset	n	Min	Q1	Median	Q3	Max
Phase 1	Train	15,821	1	3	8	26	248
	Test	6,130	1	3	7	16	69
Phase 1/2	Train	3,276	1	16	30	50	265
	Test	993	1	11	20	32	69
Phase 2	Train	18,946	1	13	25	43	411
	Test	4,503	1	10	17	28	72
Phase 2/3	Train	1,660	1	14	26	44	247
	Test	550	1	7	14	23	64
Phase 3	Train	13,522	1	13	23	39	306
	Test	2,689	1	10	17	28	70
Phase 4	Train	11,199	1	12	21	36	287
	Test	2,654	1	7	15	25	72
Total	Train	64,424	1	9	21	38	441
	Test	17,519	1	6	13	24	72

All duration values are reported in months.

Durations were calculated as the interval between the enrollment start date and the primary completion date as reported in ClinicalTrials.gov.

The dataset includes the full range of observed values without exclusion or truncation.

2.3 Target label

The prediction target in CTDP is the actual trial duration, defined as the time from the enrollment of the first participant to the collection of the final primary endpoint measurement for the last participant. This value was calculated using the enrollment start date and the primary completion date as reported in CTG. Table 1 shows the distribution of trial durations by phase in the CTDP dataset.

Clinical trial perspective. While previous studies[11–15] have often used the interval between the trial start date and the overall completion date, we focus on the primary completion date. In clinical development practice, regulatory decisions are typically based on primary endpoint results, making the primary completion date more directly linked to marketing authorization. In contrast, the overall completion date may be influenced by secondary outcomes or extended follow-up, which are less relevant for regulatory approval. Therefore, the primary completion date provides a more consistent and meaningful target for duration prediction.

2.4 Features

The CTDP dataset has totally 51 features: 12 categorical, 20 numeric, 16 free-text, and 3 other type features. Table 2 provides a detailed correspondence between the

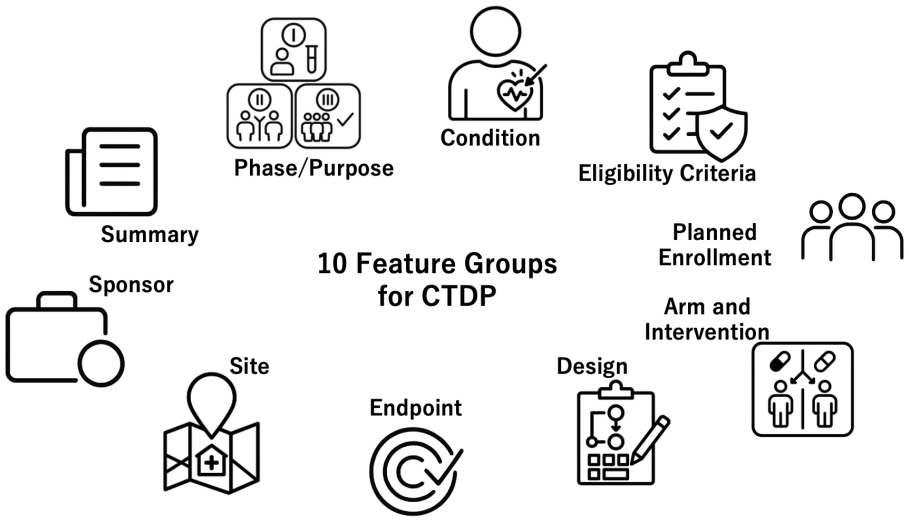


Fig. 2 Overview of the ten semantic feature groups included in the CTDP dataset. These groups correspond to the structured and free-text trial characteristics described in Table 2 and are used as inputs for clinical trial duration prediction.

data type grouping and the semantic grouping (Fig. 2) adopted in CTDP. While the data type classification offers a computational perspective, it does not fully capture the operational meaning of each feature in clinical trials. Therefore, in the following subsections we reorganize the features according to their semantic groups—phase and purpose, condition, eligibility criteria, planned enrollment, arm and intervention, design, endpoint, site, sponsor, and summary—and describe them in greater detail.

Clinical trial perspective. In practice, the overall duration of a clinical trial can be broadly separated into two components: the *recruitment period* and the *follow-up period*. This relationship can be conceptually expressed as:

$$\text{Trial duration} \approx \text{Recruitment period} + \text{Follow-up period}.$$

The recruitment period may be further expressed as

$$\text{Recruitment period} = \frac{\text{Planned enrollment}}{\text{Recruitment speed}},$$

where planned enrollment specifies the number of participants required, and recruitment speed represents the aggregate effect of all other factors. The specific contributions of individual feature groups to recruitment speed are described in subsequent sections.

The follow-up period depends on the type of primary endpoint. For fixed-duration endpoints (e.g., biomarker measurement at 12 months), all participants are observed for a constant interval defined in the protocol. For time-to-event outcomes (e.g., overall

Table 2 Clinical trial features categorized by data type and semantic group

	Data type				
Semantic group	**Categorical**	**Numeric**	**Free-text**	**Other**	**# feat.**
Phase and Purpose	`phase` ($6^{\dagger}$) `purpose` ($9^{\dagger}$)				2 (15)
Condition		`cond._count`	`cond.`	`icd10`	3 (3)
Eligibility Criteria	`gender` ($3^{\dagger}$) `gender_based` ($1^{\ddagger}$) `healthy_vol.` ($1^{\ddagger}$)	`maximum_age` `minimum_age`	`criteria`		6 (8)
Planned Enrollment		`enrollment`			1 (1)
Arm and Intervention	`int._type` ($10^{\ddagger}$)	`int._count` `n_arm`: – Experimental – Active – Placebo – Sham – No Int. – Other	`int._name`	`arms` `smiles`*	11 (20)
Design	`allocation` ($4^{\dagger}$) `assign_type` ($5^{\dagger}$) `masking` ($4^{\ddagger}$)	`masking_count`			4 (14)
Endpoint		`time_frame_y`* `measure_count`	`time_frame` `measure`		4 (4)
Site	`main_country` ($38^{\dagger}$) `us_included` ($1^{\ddagger}$)	`country_count` `address_count`	`country` `address`		6 (43)
Sponsor	`sponsor_class` ($4^{\dagger}$)	`n_collaborator`: – Industry – NIH – US – Other			5 (8)

Summary			`brief_summary`
			`brief_title`
			`official_title`
			`keyword`
			`description`:
			– detailed
			– intervention
			– outcome
			– assign_type
			– masking
					9 (9)

# feat.	12 (86)	20	16	3	51 (125)

Numbers in parentheses indicate the number of classes.
Abbreviation: cond., condition; int., intervention; feat., features.
† One-hot (exactly one class must take value 1);
‡ Multi-hot (any number of classes, including zero or multiple, can take value 1).
* Features generated from free-text descriptions using a large language model.

survival), the follow-up continues until the occurrence of events, with the analysis cut-off applied if the planned time frame is not exceeded:

$$\text{Follow-up period} = \begin{cases} T_{\text{frame}}, & \text{if the endpoint is fixed-duration} \\ \min(T_{\text{last_event}}, T_{\text{frame}}), & \text{if the endpoint is time-to-event,} \end{cases}$$

where T_{frame} denotes the predefined observation interval specified in the protocol, and $T_{\text{last_event}}$ is the time to event for the last participant.

It should also be recognized that not all trials proceed to completion as originally planned. Some are discontinued early for various reasons, including *futility* (lack of efficacy), *efficacy stopping* (overwhelming benefit observed at interim analysis), *safety concerns*, or *operational issues* such as insufficient enrollment. These scenarios result in truncated durations and represent an additional source of variability that must be considered when modeling trial length.

No prior studies[11–15] have correctly considered the full range of these features(Table 3). For example, some earlier works excluded key features such as endpoints[12, 14, 15], design[12–14], or planned enrollment[12, 15], or relied on post hoc values (e.g., actual enrollment[11, 14]), which are not usable for prospective prediction.

2.4.1 Phase and purpose

The dataset includes a `phase` field, covering trial phases (I, II, III, IV) as well as combined phases (I/II, II/III), and a `purpose` field, covering nine categories such as treatment, prevention, diagnostic, supportive care, screening, health services research, device feasibility, basic science, and other. Each category was one-hot encoded.

Clinical trial perspective. The phase of a trial is directly related to its typical duration: Phase I studies often involve fewer participants and shorter follow-up, whereas Phase III studies require large-scale enrollment and longer observation to demonstrate efficacy. The purpose of the study also affects trial length; for example, prevention and screening trials generally require longer follow-up to observe incidence rates, while treatment trials may conclude more quickly if early endpoints such as biomarker response are used.

2.4.2 Condition

The `condition` field in CTG describes the disease area or medical problem under investigation, which can include healthy population. We extracted the raw free-text condition terms and processed them into their corresponding the 10th revision of the International Classification of Diseases (ICD-10) codes via an API from US National Library of Medicine[17] as done by Yue et al.[12]. The total number of distinct conditions was recorded as `condition_count`.

Clinical trial perspective. The therapeutic area has a substantial influence on trial duration. For instance, oncology trials typically require long-term follow-up to capture survival outcomes, whereas dermatology or infectious disease trials may conclude more quickly due to short-term endpoints. Moreover, trials involving rare diseases

Table 3 Comparison of features

Study	Phase and Purpose	Condition	Eligibility Criteria	Planned Enrollment	Arm and Intervention	Design	Endpoint	Site	Sponsor	Summary
Long et al. (2024)[11]	N/A[1]	×	×	×[2]	∼	∼	∼	∼	✓	×
Yue et al. (2024)[12]	∼	✓	✓	×	∼	×	×	×	×	×
Cho et al. (2025)[13]	✓	∼	∼	✓	∼	×	∼	✓	×	×
Patrone et al. (2025)[14]	✓	✓	×	×[2]	∼	×	×	✓	✓	×
Chen et al. (2024)[15]	✓	✓	✓	×	∼	∼	×	✓	✓	✓
CTDP dataset	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

✓: explicitly used as a structured or clearly extractable feature; ∼: partially used, used in limited form, or indirectly considered; ×: not used.

[1]Not applicable due to the study scope.

[2]Actual values are used. They are not available at prediction time.

often face recruitment challenges because of limited patient availability, prolonging the recruitment period.

2.4.3 Eligibility criteria

Our dataset includes free-text `criteria` as they are recorded in CTG. In addition, we extracted structured features including `gender`, acceptance of `healthy_volunteers`, and `minimum/maximum_age` limits.

Clinical trial perspective. *Eligibility criteria* define who can participate in a clinical trial and thereby regulate the pool of potential participants. Narrower criteria, such as restricted age ranges or exclusion of comorbidities, reduce the number of eligible participants and can slow recruitment. Broader criteria allow faster enrollment but may increase heterogeneity within the study population. The inclusion of healthy volunteers is particularly relevant for early-phase studies, especially those involving safety, pharmacokinetic, or pharmacodynamic assessments, where participants without underlying medical conditions are typically required. Moreover, eligibility criteria also shape the underlying risk profile of enrolled participants. A higher risk or lower risk population affects the required length of follow-up for time-to-event outcomes. Consequently, eligibility criteria influence both the pace of recruitment and the duration of follow-up.

2.4.4 Planned enrollment number

This feature represents the number of participants that were originally planned to be enrolled in the trial. Unlike some previous studies[11, 14] that used actual enrollment numbers, which are not available at the time of trial initiation, we focused on the planned enrollment. Because the XML files from CTG do not directly provide this information, we extracted it from the HTML pages of each individual trial under the structured label "Original Enrollment" on the "Researcher View" tab (e.g., https://clinicaltrials.gov/study/{NCTID}?tab=table). The extracted values were parsed and stored as a numeric feature.

Clinical trial perspective. Planned enrollment directly determines the scale of participant recruitment and is a principal component of the recruitment period. Larger enrollment targets generally require longer recruitment period unless offset by broader eligibility criteria, higher disease prevalence, or greater site capacity.

2.4.5 Arm and intervention

This feature group captures both the hierarchical structure and summary statistics of study arms and interventions.

Unlike previous studies [11–15] that typically treat arms and interventions as flat, unlinked attributes, our dataset preserves the original hierarchical relationship: each study arm is explicitly linked to its corresponding intervention(s). This structure is stored as a list of dictionaries in JSON format, where each dictionary includes the arm label and its assigned interventions (e.g., [{ "arm": "Experimental", "interventions": ["Drug A"] }, { "arm": "Placebo Comparator", "interventions": ["Placebo"] }]).

In addition to the structured representation, we also included flat features derived from the arm and intervention data. These comprise the number of arms, the distribution of arm types, the total number of interventions, and intervention categories.

Clinical trial perspective. In clinical trials, participants are assigned to a predefined group—sometimes a single group—receiving a specific treatment regimen, referred to as an *arm*. Preserving the structure of arms and their linked interventions provides a faithful representation of trial design, as it reflects how treatments are organized and compared. The complexity of this structure and the choice of comparators can delay recruitment through participants' preferences and increased operational complexity.

2.4.6 Design

Our dataset includes the method of `allocation` (e.g., randomized or non-randomized), the `intervention_model` (e.g., parallel assignment, crossover assignment), and the use of `masking` (participant, care provider, investigator, outcomes assessor). In addition, we compute the number of masked roles, representing how many parties were blinded in the trial.

Clinical trial perspective. The design of a clinical trial refers to the overall framework that determines how participants are allocated to arms, how interventions are administered, and how outcomes are assessed. Randomization and blinding generally increase complexity, as blinded trials with multiple masked roles demand more logistical coordination than open-label studies. Likewise, crossover designs extend the observation window because of repeated treatment periods and washout intervals. These aspects of design can therefore contribute to longer trial durations through both increased operational demands and prolonged follow-up requirements.

2.4.7 Endpoint

Because our task defines trial duration based on the primary completion date, we include only primary endpoints as input features. Each primary endpoint description in CTG typically specifies both the clinical outcome being measured and the associated time frame (e.g., "Change in blood pressure at week 12"). In CTDP, we retain these descriptions as structured free-text features, preserving the full detail of both outcome and time frame.

To further enhance usability, we applied a large language model, Gemini 2.0 Flash [18] to parse these free-text descriptions into *time frame* (observation window specified in the protocol).

Clinical trial perspective. The follow-up period in a clinical trial is generally determined by the specified time frame. Its interpretation, however, varies according to the nature of the endpoint. For time-to-event endpoints, the time frame typically represents the maximum observation period during which the occurrence of predefined events is monitored. In contrast, for fixed-time endpoints, the follow-up duration is conceptually constant because the outcome is measured at predefined time points, although minor operational variability may occur in practice.

Table 4 Country or Region Group

Area	Country/Region Group
North America	United States, Canada
Central America and Caribbean	Mexico, Others
South America	Argentina, Brazil, Others
Oceania	Australia, Others
East Asia	China, Hong Kong and Macau, Taiwan Japan, Korea
Southeast Asia	Singapore, Thailand, Others
Central and South Asia	India, Pakistan, Others
West Asia	Israel, Turkey, Others
Europe	Austria, Belgium, Denmark, France, Germany Italy, Netherlands, Norway, Spain, Sweden Switzerland, United Kingdom, Others
Africa	Egypt, Others

2.4.8 Site information

The dataset includes multiple features derived from the site information recorded in ClinicalTrials.gov, which are relevant to the operational and clinical aspects of trial conduct. Specifically, we extracted the `site_count` (the total number of participating institutions), the `country_count` (the number of distinct countries in which sites are located), a binary feature `us_included` (indicating whether the United States is among the participating countries), and a categorical variable `main_country` (representing the country with the largest number of participating sites). To enhance generalizability, countries with few registered trials were grouped into broader regional categories as summarized in Table 4. These features collectively characterize the geographical distribution and scale of the study network.

Clinical trial perspective. The geographical and institutional distribution of study sites affects multiple dimensions of trial duration. First, operational complexity increases in multicenter or multinational studies due to additional layers of regulatory review, ethics approval, and coordination across institutions. Second, recruitment speed is strongly influenced by site location and capacity: trials conducted in regions with higher disease prevalence or larger eligible populations generally achieve faster enrollment, whereas those relying on rare-disease centers or countries with limited infrastructure may experience slower recruitment. Third, event rates—the frequency with which clinical endpoints occur—can vary geographically due to differences in background risk, healthcare systems, or standard of care. Trials in regions with lower event incidence require longer follow-up to reach the prespecified number of events. Thus, site-related characteristics shape not only the logistical complexity but also the pace of recruitment and the duration of follow-up, jointly determining the overall timeline of a clinical trial.

2.4.9 Sponsor information

The dataset includes sponsor-related features, reflecting the organizational characteristics that may influence clinical trial timelines. Specifically, we extracted the `sponsor_type`, a categorical feature representing the organization type of the lead sponsor (Industry, NIH, U.S. Federal Agency, or Other), and four numeric variables indicating the `collaborator_count` for each sponsor category (Industry, NIH, U.S. Federal Agency, and Other). These structured features capture the diversity of trial leadership and support, allowing analysis of how different sponsor types contribute to the efficiency and scale of trial execution.

Clinical trial perspective. The sponsor plays a central role in determining the pace and quality of clinical trial conduct. Industry-sponsored trials often benefit from greater financial resources, established site networks, and professional project management, which can shorten recruitment and follow-up periods. In contrast, investigator-initiated or academic trials may have more limited infrastructure and rely on grant-based funding cycles, often resulting in slower initiation and extended timelines.

2.4.10 Summary description

We extracted a set of narrative fields that provide free-text descriptions of various aspects of the trial. These fields, although unstructured, may contain useful contextual information for predicting trial duration. Specifically, we included the following text fields from the XML files: brief summary, brief title, official title, detailed description, intervention description, primary outcome description, intervention model description, masking description, keyword.

These features were retained in their original form for potential use in natural language processing–based models. While not all prior studies incorporated such narrative components, we preserved them to enable future work that may leverage semantic content for potential use in downstream prediction tasks.

2.5 Additional LLM-derived features

Because most of the fields in CTG are recorded in a free-text form, some natural language processing was required before these data could be used in AI predictors. In this work, we utilized the Gemini 2.0 Flash[18] via API to convert raw free-text fields into structured features. These outputs were provided as examples of pre-processing. Their plausibility was assessed through a manual review of randomly sampled trials, as described in the Technical Validation section.

2.5.1 Endpoint time frame

CTG allows trial sponsors to describe the timing of endpoint assessments in natural language, often in the "Time Frame" field associated with each primary outcome. To use this information systematically, we asked the LLM to extract the duration from enrollment to the time when the primary endpoint is measured for each arm. This value was returned in days and stored as a numeric feature. If multiple primary endpoints

existed, we selected the one with the longest time frame, reflecting the assumption that trial completion is determined by the last primary assessment. The prompt used for this extraction is shown in the supplemental material 1A.

2.5.2 Arm and intervention

Each arm is assigned one or more interventions; however, in CTG, the correspondence between arms and interventions is sometimes recorded incorrectly, and in some cases, information corresponding to multiple arms is described within a single arm. In addition, interventions are occasionally registered under regimen names, or multiple drugs are grouped together and recorded as a single intervention, making direct matching of intervention names to drug databases infeasible. In this dataset, to address these recording inconsistencies related to arms and interventions, we restructured the arm information using an LLM.

Even after restructuring, interventions may include the names of active ingredients, their dosages, as well as non-pharmacological interventions, such as medical devices or procedures. We mapped these interventions to their corresponding SMILES strings using the PubChem API when available in the PubChem[19] database. For intervention names that could not be mapped, we examined whether they represented non-effective comparators, such as placebo, vehicle, or sham. When an intervention name indicated an active treatment, we extracted the names of the active ingredients from the intervention text and attempted to map the extracted names to their SMILES strings again. All steps other than SMILES mapping were performed by using an LLM. The prompt used for this extraction is shown in the supplemental material 1B-E.

2.6 Data split

To evaluate the generalizability of the models developed in CTDP, we followed the temporal split strategy used by Yue et al.[12]. Specifically, we used trials that were completed before 2019 as the training set (n = 64,424) and trials that started in or after 2019 as the test set (n = 17,519). By separating the data according to calendar time rather than randomly, we simulate real-world deployment conditions and assess the robustness of the model to the temporal distribution shift. This setting better reflects the practical use case of predicting durations for future or ongoing trials based on historical data.

Importantly, temporal splitting avoids the risk of information leakage inherent in random splits. For instance, when evaluating a model on a trial conducted in 2023 using a training set that includes data from 2024 (e.g., other studies involving the same newly developed drug), the model may benefit from future knowledge that would not have been available in a real-world setting. By using a time-based split, we prevent such unrealistic advantage and ensure a more faithful estimation of prospective performance.

3 Data Record

We provide the training and test datasets on a GitHub repository (https://github.com/Nonokyu/CTDP) as separate parquet files (four files for training and two files for testing). For model development and evaluation, the parquet files for training and testing should be loaded and concatenated separately.

Sample Python 3 code for viewing the data records is also available in the same GitHub repository.

4 Technical Validation

As this dataset was derived from a publicly available registry (ClinicalTrials.gov), its quality depends on the correctness of the data submitted by trial sponsors. Accordingly, we acknowledge that some values may contain inaccuracies that cannot be fully corrected by us. In light of this limitation, we present several materials to facilitate the assessment of the technical soundness and practical usability of our dataset.

4.1 Summary statistics

First, we provide proportions of missing values for each feature group and unique counts of selected features to assess data completeness and the practical usability of individual features (see Table 4). Features originally included in CTG were more likely to be missing in trials registered at earlier time points. Even after filtering clinical trials to include only those flagged in CTG as involving at least one drug intervention, some trials still lacked any drug or biologic interventions. From the perspective of SMILES strings, we were unable to map intervention names to corresponding SMILES strings registered in PubChem in 16.09% and 30.66% of trials in the training and test datasets, respectively. This is mainly because interventions registered at later time points may still be confidential, and their actual chemical names are therefore not disclosed. In addition, recent trials increasingly evaluate biologics—which are not small molecules and cannot be represented as SMILES strings.

4.2 Quality of LLM-derived features

Second, for features derived from unstructured text using an LLM, conducting an exhaustive validation is infeasible. Therefore, we focused on assessing the plausibility of the extracted values and characterizing common failure modes through a manual review of 100 randomly sampled trials from each dataset. This review was independently conducted by Y.N. and Y.T., both of whom have formal training in medicine and clinical trial design. In cases where the assessments differed between the two reviewers, discrepancies were resolved through discussion.

4.2.1 Endpoint time frame

For this feature, we assessed whether the LLM correctly extracted the primary endpoint time frame as numeric values with appropriate time units.

Table 5 Missing ratio of and unique counts

Feature	Train (n=64,424)	Test (n=17,519)	Unique counts
Phase and Purpose			
phase	0.00% (0)	0.00% (0)	
purpose	2.58% (1,660)	0.00% (0)	
Condition	0.00% (0)	0.00% (0)	
condition	0.00% (0)	0.00% (0)	24055/18846/9084
icd10	8.13% (5235)	13.35% (2339)	12521/11345/8655
Eligibility Criteria			
criteria	0.01% (6)	0.00% (0)	
gender	0.00% (0)	0.00% (0)	
healthy volunteers	0.00% (0)	0.00% (0)	
maximum age	0.00% (0)	0.00% (0)	
minimum age	0.00% (0)	0.00% (0)	
Planned Enrollment	5.99% (3,860)	0.03% (6)	
Arm and Intervention			
arms	0.00% (0)	0.00% (0)	
intervention	0.00% (0)	0.00% (0)	77303/59656/21111
drug or biologics	0.18% (118)	0.23% (40)	47533/36296/15458
SMILES	16.09% (10365)	30.66% (5371)	7387/6289/3552
Design			
masked role	1.53% (985)	0.00% (0)	
masking count	0.32% (205)	0.00% (0)	
allocation	0.00% (0)	0.00% (0)	
intervention model	0.41% (261)	0.00% (0)	
Endpoint	0.47% (305)	0.00% (0)	
Site information			
country	8.85% (5703)	0.01% (1)	178/166/147
address	8.88% (5722)	0.01% (1)	
Sponsor information			
sponsor	0.00% (0)	0.00% (0)	
collaborator	0.00% (0)	0.00% (0)	
Descriptions			
brief summary	0.00% (0)	0.00% (0)	
brief title	0.00% (0)	0.00% (0)	
official title	1.88% (1212)	0.00% (0)	
detail	42.61% (27,452)	41.36% (7,246)	
intervention	12.07% (7,779)	0.01% (1)	
primary outcome	34.97% (22,530)	12.15% (2,128)	
intervention model	97.72% (62,957)	72.69% (12,734)	
masking	98.55% (63,492)	84.35% (14,777)	
keyword	33.98% (21,893)	51.14% (8,959)	

Missing values are summarized as percentages (counts). Unique counts are reported for the total, training, and test datasets, in the order of total/training/test.

Table 6 Result of plausibility check for endpoint time frame

Error category	Description	Train	Test
Multiple time points	Multiple primary endpoint time frames were described, and the longest relevant time frame was not correctly identified	1	3
Relative expressions	Time frames described using relative formulations, for which the LLM extracted only relative expressions without performing the required calculations to obtain absolute durations	0	1
Correct extraction	Primary endpoint time frame was correctly extracted and converted into a numeric value	99	96

Accurate extraction of the primary endpoint observation time failed in 1 of the 100 trials in the training dataset and in 4 of the 100 trials in the test dataset. In all cases, these failures occurred in trials in which multiple time points for primary endpoint assessment were described (Table 6). In addition, the LLM occasionally failed in cases that required temporal calculations, such as computing time points defined as "6 months after the final visit."

4.2.2 Arm and intervention

For this feature, we assessed two aspects: whether the LLM successfully restructured the arm–intervention relationships, and whether it accurately extracted compound names from intervention names to enable mapping to the corresponding SMILES strings via the PubChem API. Arm restructuring was considered successful if the effective number of arms was correctly identified and each arm was consistently assigned the appropriate intervention(s).

In the training dataset, 12 of the 100 trials failed either in arm restructuring or in mapping drug interventions to SMILES strings. Among these, six failures were attributable to issues in the original source data. Of the remaining trials, one failed due to unsuccessful arm restructuring, while the remaining five failed because of unsuccessful extraction of active ingredient names from intervention names. In the test dataset, 8 of the 100 trials failed either in arm restructuring or in mapping drug interventions to SMILES strings. Among these, two failures were attributable to issues in the original source data. Of the remaining trials, two failed due to unsuccessful arm restructuring, while the remaining four failed because of unsuccessful extraction of active ingredient names from intervention names (Table 7).

In addition, during the manual review, we identified a certain number of trials in which the arm information recorded in CTG contained inconsistencies, such as placebo arms being assigned active drug interventions, interventions described elsewhere in the trial records but not assigned to any arm, or multiple arms being conflated within a single arm entry. Even in these cases, we confirmed that the LLM-based processing successfully restructured the arm–intervention relationships according to the predefined criteria.

Table 7 Result of plausibility check for arm-intervention structure

Error category	Description	Train	Test
Source data inconsistency	Inconsistencies or ambiguities in the original ClinicalTrials.gov records regarding arm or intervention information	6	2
Arm misidentification	Incorrect identification of the number of arms or incorrect separation of arm boundaries	1	2
Active ingredient extraction failure	Failure to correctly extract active ingredient names from intervention descriptions	5	4
Correct restructuring and mapping	Arm–intervention relationships were correctly restructured and SMILES strings were successfully assigned where applicable	88	92

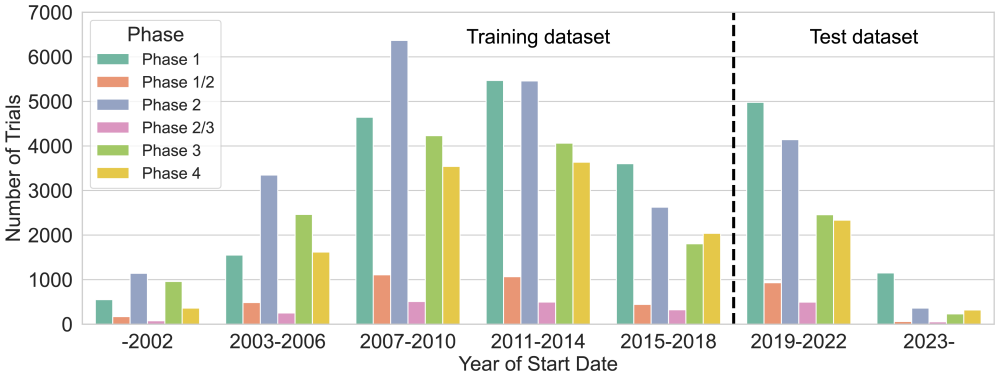


Fig. 3 Distribution of included clinical trials by start year and phase. The dashed vertical line indicates the temporal split between the training dataset (trials completed before 2019) and the test dataset (trials starting in or after 2019).

4.3 Trials included in CTDP

Third, CTDP differs from several existing public datasets in its inclusion criteria and target use case. In particular, it focuses on completed interventional drug trials to support the prediction of actual trial duration.

Specifically, we retained only interventional drug trials that were marked as either "Completed" or "Terminated". Trials that were ongoing or that did not involve drug interventions were excluded. These inclusion criteria ensure that the target duration is well defined and aligned with the intended use case of prospective duration prediction.

Figure 3 shows the distribution of included trials by start year and phase, as well as the temporal split used to construct the training and test datasets. Trials completed before 2019 were assigned to the training set, whereas trials starting in or after 2019 were assigned to the test set. This time-based split reflects the practical scenario of predicting durations for future trials using historical data.

Finally, even if future studies identify additional predictive features not included in our current dataset, researchers can retrieve the corresponding original trial records using the NCT identifiers provided. This ensures that new features can be reconstructed or appended consistently and that model performance can be compared on an identical set of clinical trials. This design choice supports reproducibility, facilitates incremental improvements, and enables fair benchmarking across studies.

5 Data Availability

The training and test datasets are publicly available at https://github.com/Nonokyu/CTDP.

6 Code Availability

Sample Python 3 code for viewing the data records is available at https://github.com/Nonokyu/CTDP.

References

[1] Yin, G.: Clinical Trial Design: Bayesian and Frequentist Adaptive Methods. John Wiley & Sons, New York (2012)

[2] Kandi, V., Vadakedath, S.: Clinical trials and clinical research: A comprehensive review. Cureus **15**(2), 35077 (2023) https://doi.org/10.7759/cureus.35077

[3] Martin, L., Hutchens, M., Hawkins, C.: Clinical trial cycle times continue to increase despite industry efforts. Nat. Rev. Drug Discov. **16**(3), 157 (2017) https://doi.org/10.1038/nrd.2017.21

[4] Wong, C.H., Siah, K.W., Lo, A.W.: Estimation of clinical trial success rates and related parameters. Biostatistics **20**(2), 273–286 (2019) https://doi.org/10.1093/biostatistics/kxx069

[5] Brøgger-Mikkelsen, M., Ali, Z., Zibert, J.R., Andersen, A.D., Thomsen, S.F.: Online patient recruitment in clinical trials: Systematic review and meta-analysis. J. Med. Internet Res. **22**(11), 22179 (2020) https://doi.org/10.2196/22179

[6] Desai, M.: Recruitment and retention of participants in clinical studies: Critical issues and challenges. Perspect. Clin. Res. **11**(2), 51–53 (2020) https://doi.org/10.4103/picr.PICR_6_20

[7] Philipson, T.J., Sun, E.: Cost of caution: The impact on patients of delayed drug approvals. Project FDA Report 2, Manhattan Institute for Policy Research, New York (June 2010). Project FDA Report No. 2. https://www.manhattan-institute.org/html/fda_02.htm

[8] Lai, J., Forney, L., Brinton, D.L., Simpson, K.N.: Drivers of start-up delays in global randomized clinical trials. Ther. Innov. Regul. Sci. **55**(1), 212–227 (2021) https://doi.org/10.1007/s43441-020-00207-2

[9] Bieske, L., Zinner, M., Dahlhausen, F., Truebel, H.: Critical path activities in clinical trial setup and conduct: How to avoid bottlenecks and accelerate clinical trials. Drug Discov. Today **28**(10), 103733 (2023) https://doi.org/10.1016/j.drudis.2023.103733

[10] Smith, Z.P., DiMasi, J.A., Getz, K.A.: New estimates on the cost of a delay day in drug development. Ther. Innov. Regul. Sci. **58**(5), 855–862 (2024) https://doi.org/10.1007/s43441-024-00667-w

[11] Long, B., Lai, S.-W., Wu, J., Bellur, S.: Predicting phase 1 lymphoma clinical trial durations using machine learning: An in-depth analysis and broad application insights. Clin. Pract. **14**(1), 69–88 (2024) https://doi.org/10.3390/clinpract14010007

[12] Yue, L., Xing, S., Li, J., Islam, M.Z., Xia, B., Chen, J., Fu, T.: Trialdura: Hierarchical attention transformer for interpretable clinical trial duration prediction. In: Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. BCB '24. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3698587.3701434

[13] Cho, J., Xu, Q., Wong, C.H., Lo, A.W.: Predicting clinical trial duration via statistical and machine learning models. Contemp. Clin. Trials Commun. **45**, 101473 (2025) https://doi.org/10.1016/j.conctc.2025.101473

[14] Patrone, C., Mahmmdla, J., Seth, R., Raghupathy, G.D.: Neural network models to predict clinical trial completion. Foresight (Int. J. Appl. Forecast.) (76), 15–20 (2025)

[15] Chen, J., Hu, Y., Cai, M., Lu, Y., Wang, Y., Cao, X., Lin, M., Xu, H., Wu, J., Cao, X., Sun, J., Li, Y., Glass, L., Huang, K., Zitnik, M., Fu, T.: TrialBench: Multi-modal AI-ready datasets for clinical trial prediction. Sci. Data **12**(1), 1564 (2025) https://doi.org/10.1038/s41597-025-05680-8

[16] US National Library of Medicine: ClinicalTrials.gov. https://clinicaltrials.gov/. Accessed: 2025-03-05

[17] US National Library of Medicine: Clinical Table Search Service: API for ICD-10-CM (version 3). https://clinicaltables.nlm.nih.gov/apidoc/icd10cm/v3/doc.html. Accessed: 2025-03-05

[18] Google: Gemini 2.0 Flash. Large multimodal model (2024). https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash Accessed 2025-11-06

[19] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E.: Pubchem 2025 update. Nucleic Acids Res. **53**(D1), 1516–1525 (2024) https://doi.org/10.1093/nar/gkae1059

Author Contributions

Y.N. and M.N.A. conceived and designed the study. Y.N. conducted the dataset curation. Y.N. and Y.T. contributed to the clinical interpretation of the data and performed manual validation. K.K., A.S., and A.D. provided methodological guidance and supervised the study. Y.N. and M.N.A. wrote the main manuscript text. All authors reviewed and approved the final manuscript.

Funding

This work was supported by JST BOOST, Japan Grant Number JPMJBS2401 and JST ASPIRE, Japan Grant Number JPMJAP2403.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementalmaterial1.pdf