A Comprehensive Guide to TrialBench: Multi-Modal AI-Ready Datasets for Clinical Trial Prediction

1. Document Overview

This document delves into the specific meanings of **Features** and **Task Labels** within the datasets and tasks of **TrialBench**, open-source multi-modal AI-ready datasets for clinical trial prediction. It serves to clarify the definitions and application contexts of each field, enabling users to effectively leverage the datasets for research on crucial areas such as trial duration, patient dropout, serious adverse events, and mortality prediction. The content is organized into two distinct sections—**Feature Descriptions** and **Task Label Descriptions**—presented in tabular format for easy comparison of names and explanations.

This setup supports seamless integration with related resources like the TrialBench toy samples and the AI4Trial codebase for running relevant experiments.

Our project homepage: https://github.com/ML2Health/ML2ClinicalTrials .

2. Feature Descriptions

Feature Name	Description
Active Comparator Arm Number	Number of arms receiving an established active treatment to compare against the experimental group.
Behavioral Intervention Number	Number of behavioral interventions, such as counseling or psychotherapy.
Biological Intervention Number	Number of interventions involving biological products or vaccines.
Combination Product Intervention Number	

	Number of interventions involving combination products (e.g., drug - device combinations).
Device Intervention Number	Number of device - related interventions, including real or sham devices.
Diagnostic Test Intervention Number	Number of diagnostic test - related interventions, including imaging or in vitro tests.
Dietary Supplement Intervention Number	Number of interventions involving dietary supplements, such as vitamins or minerals.
Drug Intervention Number	Number of drug - related interventions, including both active drugs and placebos.
Experimental Arm Number	Number of arms assigned to receive experimental interventions being tested for efficacy or safety.
Genetic Intervention Number	Number of genetic - level interventions, such as gene transfer or stem cell therapy.
MaskingType - Care Provider	Indicates whether the care providers administering the interventions are unaware of the assignments, aiming to prevent bias in the delivery of care.
MaskingType - Investigator	Indicates whether the investigators assessing the outcomes are unaware of the intervention assignments, aiming to prevent assessment bias.
MaskingType - Outcomes Assessor	Indicates whether the individuals assessing the outcomes are unaware of the intervention assignments, aiming to prevent bias in outcome evaluation.
MaskingType - Participant	Indicates whether the participants in the clinical trial are unaware of the intervention assignments, aiming to prevent bias in self - reported outcomes.

No Intervention Arm Number	Number of arms where participants receive no intervention, serving as observational controls.
Other Arm Number	Number of arms with interventions that do not fit standard classifications.
Other Intervention Number	Number of interventions that do not fall under standard categories.
Placebo Comparator Arm Number	Number of arms receiving placebo treatments used to compare against active interventions.
Procedure Intervention Number	Number of interventions involving surgical or procedural operations.
Radiation Intervention Number	Number of interventions involving radiation therapy.
Sham Comparator Arm Number	Number of arms receiving sham interventions that mimic treatment procedures without therapeutic effect.
biospec_descr/textblock	Specify all types of biospecimens to be retained (e.g., whole blood, serum, white cells, urine, tissue).
brief_summary/textblock	A short description of the clinical study, including a brief statement of the clinical study's hypothesis, written in language intended for the lay public.
brief_title	A short title of the clinical study written in language intended for the lay public. The title should include, where possible, information on the participants, condition being evaluated, and intervention(s) studied .
condition	Disease or health condition names provided by the study sponsor or investigator to describe the medical issues being studied in the clinical trial.

condition_browse/mesh_term	A list of standardized Medical Subject Headings (MeSH) terms assigned to the conditions studied in the clinical trial. These terms are generated by the National Library of Medicine (NLM) using an internal algorithm that maps the free - text condition entries to the closest matching MeSH terms.
detailed_description/textblock	Extended description of the protocol, including more technical information (as compared to the Brief Summary), if desired.
eligibility/criteria/textblock	"A limited list of criteria for selection of participants in the clinical study, provided in terms of inclusion and exclusion criteria and suitable for assisting potential participants in identifying clinical studies of interest. Use a bulleted list for each criterion below the headers ""Inclusion Criteria"" and "" Exclusion Criteria""."
eligibility/gender	"The sex of the participants eligible to participate in the clinical study. ""Sex"" means a person's classification as male or female based on biological distinctions.- All : Indicates no limit on eligibility based on the sex of participants- Female: Indicates that only female participants are being studied- Male: Indicates that only male participants are being studied"
eligibility/gender_description	If eligibility is based on gender, provide descriptive information about Gender criteria.
eligibility/healthy_volunteers	Indication that participants who do not have a disease or condition, or related conditions or symptoms, under study in the clinical study are permitted to participate in the clinical study. Yes/No.
eligibility/maximum_age	

	The numerical value, if any, for the minimum age a potential participant must meet to be eligible for the clinical study.
eligibility/minimum_age	The numerical value, if any, for the maximum age a potential participant can be to be eligible for the clinical study.
eligibility/sampling_method	"Indicate the method used for the sampling approach and explain in the Detailed Description. .- Probability Sample: Exclusively random process to guarantee that each participant or population has specified chance of selection, such as simple random sampling, systematic sampling, stratified random sampling, cluster sampling, and consecutive participant sampling- Non - Probability Sample: Any of a variety of other sampling processes, such as convenience sampling or invitation to volunteer"
eligibility/study_pop/textblock	A description of the population from which the groups or cohorts will be selected (for example, primary care clinic, community sample, residents of a certain town).
enrollment	The estimated total number of participants to be enrolled (target number) or the actual total number of participants that are enrolled in the clinical study.
has_expanded_access	"Whether there is expanded access to the investigational product for patients who do not qualify for enrollment in a clinical trial. Expanded Access for investigational drug products (including biological products) includes all expanded access types under section 561 of the Federal Food, Drug , and Cosmetic Act: (1) for individual participants, including emergency use; (2) for intermediate - size participant populations; and (3) under a treatment IND or treatment

	protocol. - Yes: Investigational product is available through expanded access- No : Investigational product is not available through expanded access- Unknown: If the responsible party is not the sponsor of the clinical trial and manufacturer of the investigational product"
icdcode	"A list of ICD - 10 codes representing the diseases targeted in the clinical trial. These codes are mapped from the ""condition" " field using standard disease - to - ICD mapping. "
intervention/description	Details that can be made public about the intervention, other than the Intervention Name(s) and Other Intervention Name(s) , sufficient to distinguish the intervention from other, similar interventions studied in the same or another clinical study. For example, interventions involving drugs may include dosage form, dosage, frequency, and duration.
intervention/intervention_name	A brief descriptive name used to refer to the intervention(s) studied in each arm of the clinical study. A non - proprietary name of the in...

3. Task Label Descriptions

Task Label Name	**Description**
trial - approval - forecasting outcome	Binary label indicating trial approval based on primary outcome success (derived from `overall_status`).
trial - duration - forecasting_completion_ date	Predicted completion date of the clinical trial .
trial - duration - forecasting_month	

	Predicted trial duration in months (calculated from start/completion dates).
trial - duration - forecasting_start_date	Start date of the clinical trial.
trial - duration - forecasting_time_day	Predicted trial duration in days (calculated from start/completion dates).
trial - duration - forecasting_year	Predicted trial duration in years (calculated from start/completion dates).
trial - failure - reason - identification_failure_ reason	Categorical label for primary failure reason (safety, efficacy, poor enrollment, other).
serious - adverse - event - forecasting_Y/N	Binary label for occurrence of serious adverse events (based on clinical_results).
serious - adverse - event - forecasting_serious _adverse_rate	Proportion of participants experiencing serious adverse events (subjects_affected / subjects_at_risk).
drug - dose - prediction_(Min Max Avg)	Log - transformed (base - 10) and scaled (0 - 3) values for minimum, maximum, and average daily drug dosage (mg/day).
patient - dropout - event - forecasting_Y/N	Binary label for patient dropout (based on difference between started/completed counts).
patient - dropout - event - forecasting_ dropout_rate	Dropout rate (number of not completed participants / number of started participants).
mortality - event - prediction_Y/N	Binary label for mortality - related adverse events (based on mortality events in clinical _results).
mortality - event - prediction_mortality_rate	Mortality rate (sum of subjects_affected / subjects_at_risk for mortality events).