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.
Masking Type-Care Provider	Indicates whether care providers administering interventions are unaware of assignments to prevent bias in care delivery .
Masking Type-Investigator	Indicates whether investigators assessing outcomes are unaware of intervention assignments to prevent assessment bias.
Masking Type-Outcomes Assessor	Indicates whether individuals evaluating outcomes are unaware of intervention assignments to prevent bias in outcome evaluation.
Masking Type-Participant	Indicates whether participants are unaware of intervention assignments 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_retention	Indicate whether samples of material from research participants are retained in a biorepository:- None Retained: No samples retained- Samples With DNA: Samples retained, with potential for DNA extraction (e.g., frozen tissue, whole blood)- Samples Without DNA: Samples retained, no potential for DNA extraction (e.g., fixed tissue, plasma)
brief_summary/textblock	A short lay-public description of the clinical study, including a brief statement of the hypothesis.
brief_title	A short lay-public title of the clinical study, including participants, condition, and intervention(s) where possible.
condition	Disease or health condition names provided by the sponsor/investigator to describe the medical issues studied.

condition_browse/mesh_term	Standardized Medical Subject Headings (MeSH) terms assigned to studied conditions , auto-mapped by NLM algorithm.
detailed_description/textblock	Extended protocol description with technical information (more detailed than the brief summary).
eligibility/criteria/textblock	Inclusion and exclusion criteria in bulleted lists to assist participant screening.
eligibility/gender	Participant sex eligibility:- All: No sex restriction- Female: Only female participants - Male: Only male participants
eligibility/gender_description	Descriptive information about gender criteria if eligibility is sex-based.
eligibility/healthy_volunteers	Indication of whether healthy volunteers (without the studied disease/condition) are permitted (Yes/No).
eligibility/maximum_age	Numerical value for the minimum age a participant must meet for eligibility (if applicable).
eligibility/minimum_age	Numerical value for the maximum age a participant can be for eligibility (if applicable).
eligibility/sampling_method	Sampling approach:- Probability Sample : Exclusively random process (e.g., simple random, stratified random)- Non-Probability Sample: Non-random process (e. g., convenience sampling)
enrollment	Estimated or actual total number of participants to be enrolled in the clinical study.
has_expanded_access	Whether investigational products are available via expanded access:- Yes: Available - No: Not available- Unknown: When the

	responsible party is not the sponsor/ manufacturer
icdcode	List of ICD-10 codes for targeted diseases, auto-mapped from the condition field.
intervention/description	Public details of the intervention (e.g., dosage form, dose, frequency) to distinguish it from similar interventions.
intervention/intervention_name	Brief descriptive name of the intervention (non-proprietary if available; otherwise, use a descriptive identifier).
intervention/intervention_type	General type of intervention (e.g., drug, device, behavioral).
intervention_browse/mesh_term	Standardized MeSH terms assigned to interventions, auto-mapped by NLM algorithm.
ipd_info_type-Analytic Code	Indicates whether the analytic code for data analysis is available for sharing (Yes/No/ Undecided).
ipd_info_type-Clinical Study Report (CSR)	Indicates whether the Clinical Study Report is available for sharing.
ipd_info_type-Informed Consent Form (ICF)	Indicates whether the Informed Consent Form is available for sharing.
ipd_info_type-Statistical Analysis Plan (SAP)	Indicates whether the Statistical Analysis Plan is available for sharing.
ipd_info_type-Study Protocol	Indicates whether the Study Protocol is available for sharing.
keyword	Words/phrases describing the protocol, preferably using MeSH-controlled vocabulary .
location/facility/address/city	Name(s) of the city/cities where clinical trial facilities are located.

location/facility/address/city-Aging	Proportion of population aged ⩾65 in the trial city (supplemented by ChatGPT).
location/facility/address/city-GDP	Gross Domestic Product (GDP) of the trial city (supplemented by ChatGPT).
location/facility/address/city-Population	Total population of the trial city (supplemented by ChatGPT).
number_of_arms	Maximum number of arms in the trial (for multi-period studies, use the highest count across phases).
oversight_info/has_dmc	Indicates whether a Data Monitoring Committee (DMC) is appointed (Yes/No).
oversight_info/is_fda_regulated_device	Indicates whether the study involves an FDA-regulated device (Yes/No).
oversight_info/is_fda_regulated_drug	Indicates whether the study involves an FDA-regulated drug (Yes/No).
patient_data/sharing_ipd	Plan to share individual participant data (IPD):- Yes: Planned- No: Not planned- Undecided : Not yet determined
phase	Clinical trial phase (for drug/biological products):- N/A: Non-applicable (e.g., device studies)- Early Phase 1: Exploratory microdose studies- Phase 1-4: Traditional phase definitions
responsible_party/responsible_party_type	Type of responsible party:- Sponsor: Entity initiating the study- Principal Investigator : Individual designated by the sponsor - Sponsor-Investigator: Individual both initiating and conducting the study
smiless	List of SMILES strings for drugs used, auto -mapped from intervention/intervention_ name.
sponsors/lead_sponsor/agency_class	Category of the lead sponsor (e.g., NIH, Industry, U.S. Federal, Other).

study_design_info/allocation	Method of participant assignment to arms:- N/A: Single-arm trial- Randomized: Assigned by chance- Nonrandomized: Assigned non-randomly (e.g., physician choice)
study_design_info/intervention_model	Strategy for assigning interventions:- Single Group Assignment: Single-arm trial - Parallel Assignment: Multiple groups in parallel- Crossover Assignment: Participants receive alternating interventions- Factorial Assignment: Evaluates interventions alone and in combination- Sequential Assignment : Assignments based on study milestones
study_design_info/masking	Parties masked to intervention assignments (e.g., participants, investigators).
study_design_info/masking_description	Supplementary details about masking procedures.
study_design_info/masking_num	Number of masked parties (derived from masking field count).
study_design_info/observational_model	Primary strategy for participant identification /follow-up (e.g., cohort, case-control).
study_design_info/primary_purpose	Main objective of interventions (Treatment, Prevention, Diagnostic, etc.).
study_design_info/time_perspective	Temporal relationship of observations to enrollment:- Retrospective: Data collected before enrollment- Prospective: Data collected after enrollment- Cross-sectional: Data collected at a single timepoint
study_type	Nature of the investigation:- Interventional (clinical trial): Prospective assignment of interventions- Observational: Assessment of pre-defined groups without intervention assignment- Patient Registry: Observational study registered in PRS- Expanded Access: Investigational product access for ineligible patients

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).