Overview of the CT-DEB'26 Shared Task on Predicting Dosing Errors in Interventional Clinical Trials

Sohrab Ferdowsi[1], **Félicien Hêche**[1], **Anthony Yazdani**[1], **Edward Choi**[2], **Sara Sansaloni-Pastor**[3], **Douglas Teodoro**[1]
[1] Department of Radiology and Medical Informatics, University of Geneva, Switzerland
[2] KAIST, Republic of Korea
[3] Actelion Pharmaceuticals Ltd, Basel, Switzerland

Abstract

Dosing errors represent an important source of medication-related risk in interventional clinical trials, potentially affecting both participant safety and the validity of study outcomes. Despite their importance, systematic methods for predicting dosing error risk from trial design information remain largely unexplored. To address this gap, we organized the *Clinical Trial Dosing Error Benchmark 2026* (CT-DEB'26) shared task, hosted at the CL4Health workshop at LREC 2026. The task focuses on predicting the risk of dosing errors in interventional clinical trials using heterogeneous information extracted from ClinicalTrials.gov, including structured protocol metadata and long-form textual descriptions. The released benchmark dataset contains over 42,000 clinical trial records spanning multiple study phases and therapeutic areas, annotated with binary labels indicating a significant high rate of dosing errors. Participants were asked to develop ML models capable of estimating trial-level dosing error risk, evaluated primarily using the ROC-AUC metric under strong class imbalance. The shared task was conducted in two phases and attracted 15 submissions in the development stage and 4 submissions in the final evaluation phase. This paper provides an overview of the shared task, describing the dataset construction, evaluation protocol, and participating systems. In addition, we present a schema-aware CatBoost baseline that leverages structured trial metadata and simple textual statistics, achieving ROC-AUC scores of 0.8606 and 0.8624 on the Phase 1 and Phase 2 leaderboards, respectively. We further summarize the approaches proposed by participating teams, which explore both feature-engineering pipelines and transformer-based text representations. The results highlight the importance of structured trial design variables and hybrid modeling strategies combining tabular and textual information. Finally, we discuss limitations of the benchmark and outline future directions for applying natural language processing and ML to improve medication safety in clinical trial design.

Keywords: clinical trials, medication errors, dosing errors, shared task, clinical NLP, machine learning

1. Introduction

1.1. Clinical Trials

Interventional clinical trials (CTs) constitute the cornerstone of pharmaceutical research and development, providing the evidentiary basis for assessing the safety and efficacy of medicinal products and interventions. Once the drug discovery research and pre-clinical animal-based studies are successful, for a candidate medication to reach the market, regulatory agencies require various phases of clinical studies on human volunteers, each designed to rigorously evaluate specific aspects of the drug's performance and safety profile.

Various regulations such as the (U.S. Food and Drug Administration, 2024), the (European Parliament and Council of the European Union, 2014), as well as the (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), 2016) require careful planning of the CTs prior to their execution in what is known as CT protocols (Cipriani and Barbui, 2010). Despite these efforts and the oversight by regulatory agencies, CTs remain costly, lengthy, and specifically highly failure-prone processes. Various studies estimate that no more than 14% of candidate drugs successfully progress from phase I to market authorization (Wong et al., 2018; Sun et al., 2022), while CT execution constitutes a major portion of the very costly (approximately 1.3B$ on the average) drug R&D life cycle (Wouters et al., 2020; Mulcahy et al., 2025). These failures are often attributed not only to the lack of drug efficacy, but also to operational and procedural deficiencies in trial execution (Martin et al., 2017). As a result, identifying and mitigating such inefficiencies early in the trial lifecycle is of critical importance for cost reduction of clinical studies and hence the drug prices and eventually the sustainability of the healthcare systems, as well as for the safety of participants of these studies.

Towards optimizing CT design, large public repositories such as ClinicalTrials.gov (CTGov) (National Library of Medicine (US), 2000) provide access to the past trial protocols and outcomes, enabling data-driven methods to estimate trial-level risks and to identify protocol design factors associated with operational failure. This is furthermore motivated by the advances and success stories of methods based on Machine Learning (ML) and Artificial Intelligence (AI) in healthcare. To get a

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 234–244
12 May 2026.

better overview of the application of these methods to trial optimization, the reader is referred to the recent scoping review by (Teodoro et al., 2025), where various categories of risk prediction models are reviewed.

1.2. Dosing Errors

On the other hand, medication errors, and in particular dosing errors, are among the major sources of adverse events imposing substantial burden on healthcare systems (Elliott et al., 2021; Rouhani et al., 2018; Hodkinson et al., 2020; Tariq et al., 2018). Within the context of clinical studies, medication errors can compromise trial validity by biasing efficacy estimates (Vrijens et al., 2024), jeopardize participant safety (Tariq et al., 2025), and lead to regulatory non-compliance or trial termination (European Medicines Agency, 2023).

Medication errors are defined as failures in the treatment process that lead to, or have the potential to lead to, harm to the patient (Aronson, 2009). Examples of dosing-related medication errors include incorrect dose amount, frequency, timing, or administration route.

While medication safety has been extensively studied in post-marketing and routine clinical care settings, comparatively little attention has been paid to the systematic analysis and prediction of dosing errors occurring during clinical research itself.

Prior work has investigated trial outcome prediction and failure risk estimation using protocol-level information (Ferdowsi et al., 2023, 2021), as well as the prediction of adverse drug events from clinical trial data (Yazdani et al., 2025), highlighting the value of data-driven approaches for assessing trial-level risk. However, as emphasized by the recent scoping review of (Hêche et al., 2026), ML-based studies of medication errors have almost exclusively focused on errors occurring in routine clinical practice, leaving a notable gap with respect to errors arising in the conduct of CTs. Addressing this gap is particularly challenging due to the heterogeneity of trial designs, the rarity of documented dosing errors, and the prevalence of unstructured textual information in trial protocols and reports.

1.3. Shared Task

To foster community-driven progress on this problem, we organized the *Clinical Trial Dosing Error Benchmark 2026* (CT-DEB'26), an open shared task aimed at predicting the risk of dosing errors in interventional clinical trials using information available prior to trial initiation. The shared task leverages a large, publicly released dataset derived from the ClinicalTrials.gov registry, combining structured trial design variables with long-form textual descriptions extracted from registry entries and protocol documents. Participants were invited to develop machine learning models that, given only pre-initiation protocol information, estimate the probability that a trial will exhibit an elevated rate of dosing errors during its execution. System performance was evaluated primarily using the area under the receiver operating characteristic curve (ROC-AUC), among other complementary metrics.

CT-DEB'26 was hosted on the CodaBench platform (Xu et al., 2022) and conducted in two phases, supporting model development and subsequent evaluation on a fully held-out test set. The challenge was organized as part of the CL4Health workshop at LREC 2026 and aims to establish a reusable benchmark for studying dosing error prediction in clinical trials.

This paper provides an overview of the CT-DEB'26 shared task. We first describe the dataset construction process and the prediction task formulation in Section 2. We then present the challenge setup in Section 3, including the task definition and evaluation protocol. Section 4 summarizes the submitted systems and presents a baseline model. Finally, Section 5 discusses the main findings and directions for future research.

2. Dataset

2.1. Data Source and Trial Selection

The CT-DEB'26 shared task is based on the `ct-dosing-errors-benchmark` dataset (Hêche et al., 2026) derived from completed and terminated interventional clinical trials registered at the CTGov registry (National Library of Medicine (US), 2000), which was the main source of the raw data used due to its scale, standardized reporting, as well as its regulatory relevance. The dataset was curated to support trial-level prediction of dosing error risk using only information available prior to the trial initiation, thereby reflecting realistic deployment conditions for prospective risk assessment.

Trials were selected according to predefined inclusion criteria, including interventional study design, availability of reported results, and a documented completion date. Restricting the dataset to completed or terminated trials ensures that dosing error outcomes can be derived from finalized adverse event reports, while enabling temporally coherent dataset splitting. A total of 42,112 trials satisfied these criteria and were included in the full dataset.

The dataset is publicly released on the Hugging Face Hub[1]. A detailed account of the dataset

[1]The shared-task dataset is available at https://huggingface.co/datasets/sssohrab/ct-dosing-errors-benchmark.

Feature	Avail.	Mean	Min	Max
enrollmentCount	100%	263.81	1	864,493
numArms	100%	2.33	0	43
numInterventions	100%	2.48	1	44
numLocations	100%	22.88	0	1,611

Table 1: Summary statistics of numerical features.

construction pipeline, feature engineering, labeling methodology and splitting strategy is provided in the companion paper (Hêche et al., 2026). Below we provide a summary as is relevant for the shared task.

2.2. Features

For each trial, a heterogeneous set of features was extracted, comprising numerical, categorical, and textual variables. Structured features describe core aspects of trial design and logistics, such as trial phase, enrollment size, intervention configuration, allocation and masking strategies, sponsor and oversight information, and characteristics of treatment arms and interventions.

In addition to structured variables, the dataset includes multiple free-text fields drawn from registry entries and protocol-related documents, such as brief summaries, detailed descriptions, arm and intervention descriptions, and, when available, text extracted from protocol PDF files. Approximately 42% of trials contain long-form textual descriptions, which often encode nuanced procedural details relevant to medication administration and dosing. As a result, effective utilization of unstructured text constitutes a central challenge of the benchmark.

To ensure consistency and reproducibility, all features were validated and parsed using structured data schemas with tools such as the Pydantic library (Pydantic AI), or Python's standard data types, such as the enumerated data structures (Enum) (Python Software Foundation). Notably, this results in categorical variables constraint to predefined value sets and would later facilitate ML-related feature engineering. These schemas were created according to the structures defined in CT-Gov's instructions.

Tables 1, 2, 3, and 4 provide summary statistics for the numerical, textual, categorical, and list-valued categorical features, respectively.

Notably, among the textual fields, `protocolPdfText` is both the sparsest and by far the longest, reflecting the partial availability of full protocol documents and the substantial additional procedural detail they contain when present.

Overall, the dataset combines structured numerical and categorical protocol descriptors with multiple free-text fields of highly variable length, yielding a multimodal benchmark for trial-level dosing error prediction.

2.3. Outcome Labeling and Dataset Splits

The prediction target is a binary variable indicating whether a trial exhibited an unusually high rate of dosing errors during its execution. Labels were derived from adverse event (AE) reports available in the ClinicalTrials.gov registry, which are coded using the Medical Dictionary for Regulatory Activities (MedDRA, version 27.1) (MedDRA Maintenance and Support Services Organization, 2025).

To identify dosing-related events, MedDRA categories associated with overdosing, underdosing, and medication-use errors were used as entry points in the MedDRA hierarchy. The descendant terms under these categories were reviewed by clinical pharmacology experts to retain only concepts specifically related to dosing errors. The resulting set of curated terms was then matched against the reported adverse events of each clinical trial.

For each study, dosing-related events were aggregated across all trial arms to estimate a trial-level dosing error rate based on the number of affected participants relative to the population at risk. A trial was labeled as positive when the lower bound of the 95% Wilson confidence interval of this rate exceeded a predefined threshold. Given the rarity of dosing errors and the pharmacovigilance practice of treating low-frequency events as potential safety signals (Beninger, 2020), a conservative threshold of 0.01% was adopted.

Using this procedure, approximately 4.6% of trials in the dataset were labeled as positive, resulting in a highly imbalanced classification problem that reflects the low prevalence of dosing errors in clinical research. A detailed description of the labeling workflow is provided in the companion dataset paper (Hêche et al., 2026).

The dataset was partitioned into training, validation, and test splits using a chronological strategy based on trial completion dates. This approach mitigates duration-related selection bias and provides a realistic estimate of prospective model performance. For the shared task, participants were provided with labeled training data and a validation set with masked labels during Phase 1, followed by an unlabeled test set for final evaluation in Phase 2. Table 5 summarizes key statistics of the dataset.

more detailed version with additional risk categories is available at https://huggingface.co/datasets/ds4dh/ct-dosing-errors. The source code used to reproduce these datasets is available at https://github.com/ds4dh/CT-dosing-errors.

Field	Availability (%)	Mean chars	Median chars	Median words
allocation	99.62%	8.46	10	1
armDescriptions	99.40%	388.24	223	37
briefSummary	100%	502.40	338	50
conditions	100%	39.21	23	3
conditionsKeywords	63.94%	81.17	59	7
detailedDescription	61.93%	1,610.83	1,151	175
interventionDescriptions	100%	286.18	177	27
interventionModel	99.67%	9.34	8	1
interventionNames	100%	82.72	63	8
locationDetails	94.38%	1,310.26	123	22
protocolPdfText	42.33%	223,247.14	186,043	26,122

Table 2: Summary statistics of textual fields.

Feature	Avail.	#Cat.	Categories
healthyVolunteers	99.93%	2	FALSE, TRUE
masking	99.82%	5	NONE, SINGLE, DOUBLE, TRIPLE, QUADRUPLE
oversightHasDmc	87.82%	2	FALSE, TRUE
primaryPurpose	99.06%	10	TREATMENT, PREVENTION, DIAGNOSTIC, ECT, SUPPORTIVE_CARE, SCREENING, HEALTH_SERVICES_RESEARCH, BASIC_SCIENCE, DEVICE_FEASIBILITY, OTHER
sex	99.98%	3	ALL, FEMALE, MALE

Table 3: Features with categorical values.

Feature	Avail.	#Cat.	Categories
armGroupTypes	99.40%	6	EXPERIMENTAL, ACTIVE_COMPARATOR, PLACEBO_COMPARATOR, SHAM_COMPARATOR, NO_INTERVENTION, OTHER
interventionTypes	100%	11	DRUG, DEVICE, BIOLOGICAL, PROCEDURE, RADIATION, BEHAVIORAL, GENETIC, DIETARY_SUPPLEMENT, COMBINATION_PRODUCT, DIAGNOSTIC_TEST, OTHER
phases	99.99%	6	NA, EARLY_PHASE1, PHASE1, PHASE2, PHASE3, PHASE4

Table 4: Features with list of categorical values.

Statistic	Value
Total trials	42,112
Training set	29,478
Validation set	6,316
Test set	6,318
Positive label prevalence	4.62%
Trials with protocol PDF text	42%

Table 5: Overview of the CT-DEB dataset.

3. CT-DEB'26

The Clinical Trial Dosing Error Benchmark 2026 (CT-DEB'26) was organized as an open shared task with the goal of benchmarking ML approaches for predicting dosing error risk in interventional clinical trials. The challenge was hosted on the CodaBench platform[2] and is part of the CL4Health workshop at LREC 2026. Emphasis was placed on reproducibility, transparency of methodology, and realistic evaluation of prospective trial-level risk prediction.

3.1. Task Definition

Participants were tasked with predicting, for each clinical trial protocol in the dataset, the probability that the trial would exhibit an elevated rate of dosing errors during its execution. Formally, given a feature vector $\mathbf{x}_i$ describing trial i, models were required to estimate a scalar probability $\hat{y}_i \in [0, 1]$ corresponding to the likelihood of the positive class, wheras the prediction target was a binary label as described in Section 2.

Submissions consisted exclusively of trial-level probability estimates, one per trial, identified by the unique ClinicalTrials.gov identifier (`nctid`). Participants were free to use any modeling approach, including classical ML, deep learning, or hybrid methods, provided that predictions were generated locally and only inference outputs were submitted for evaluation.

3.2. Challenge Phases

CT-DEB'26 was conducted in two sequential phases. During Phase 1 (validation phase), participants were provided with labeled training data and a validation set in which the target labels were masked. This phase was intended to support model development, hyperparameter tuning, and exploratory analysis, while preventing leaderboard overfitting. Fifteen teams submitted valid predictions during this phase.

Phase 2 (test phase) evaluated generalization performance on a held-out test set with no released labels. Leaderboards remained hidden during the active phase and were revealed only after the submission deadline. Four teams submitted final predictions during Phase 2, two of which (AlSmadi, 2026; Hamnett et al., 2026) subsequently

[2] https://www.codabench.org/

contributed full technical papers to the CL4Health workshop.

3.3. Evaluation Protocol

Submitted predictions were evaluated by comparing the estimated probabilities against the ground truth labels on the hidden validation or test sets. The primary evaluation metric used for leaderboard ranking was the area under the receiver operating characteristic curve (ROC-AUC), which evaluates the ability of a model to rank positive instances above negative ones independently of a fixed decision threshold. Given the strong class imbalance of the dataset, additional metrics such as sensitivity, specificity, macro-F_1, balanced accuracy, and Brier score were also reported to provide complementary perspectives on model performance.

In addition to ROC-AUC, several secondary metrics were reported for descriptive purposes, including macro-averaged F_1 score, sensitivity, specificity, and balanced accuracy. These metrics provide complementary insights into model behavior under different operating points, particularly with respect to minority-class detection.

All submissions were executed locally by participants, and only prediction files in a standardized CSV format were uploaded to the platform. No training or inference was performed on the CodaBench servers.

4. Results

4.1. Baseline

An official baseline was established to provide a reference point for the CT-DEB'26 shared task and to contextualize participant submissions. The baseline model is based on a CatBoost classifier (Dorogush et al., 2018) trained on the structured components of the dataset, with explicit handling of heterogeneous feature types derived from the dataset schema.

Schema-aware feature grouping: The dataset released for the shared task exposes a structured schema through the Hugging Face dataset interface, which specifies the semantic type of each feature (e.g., numerical value, categorical label, or sequence of labels). The baseline pipeline automatically parses this schema and partitions the input variables into four groups: (i) numerical features (e.g., enrollment counts, number of arms, and number of interventions), (ii) scalar categorical features corresponding to controlled vocabularies (e.g., primary purpose, masking strategy, or participant sex), (iii) list-valued categorical features representing multi-label protocol attributes (e.g., trial phases, intervention types, or arm group types), and (iv) free-text fields extracted from registry entries and protocol documents. This automatic routing ensures that each feature type is processed using a transformation compatible with its underlying structure.

Feature preprocessing. Numerical variables are used directly as scalar inputs. Scalar categorical variables are encoded using one-hot representations derived from their predefined vocabularies. List-valued categorical variables are transformed using multi-hot encoding, where each possible category is represented by a binary indicator reflecting its presence within a given trial. Textual fields are intentionally processed using a minimal representation rather than deep linguistic models: each string-valued field is converted into a single numeric feature corresponding to its character length. This applies both to long-form narrative fields (e.g., trial summaries or protocol descriptions) and shorter string-valued registry attributes. This design choice keeps the baseline lightweight while emphasizing the predictive signal contained in structured trial metadata rather than relying on semantic modeling of protocol text.

Model training: The resulting tabular feature matrix is used to train a CatBoost classifier (Dorogush et al., 2018), a gradient-boosted decision tree model designed for heterogeneous tabular data. CatBoost was selected because of its strong empirical performance on structured datasets and its ability to handle categorical features efficiently. The model is trained on the official training split, with predictions expressed as probabilities for the positive class (elevated dosing error risk). These probabilities are used directly for evaluation using the ROC-AUC metric specified by the shared task.

The baseline therefore establishes a strong tabular reference point against which more complex multimodal approaches incorporating deep text representations can be evaluated.

Baseline performance: Despite its relatively simple design, the CatBoost baseline achieved strong performance and ranked first in both phases of the challenge. On the Phase 1 (validation) leaderboard, the baseline obtained an ROC-AUC of 0.8606, with a macro-F_1 score of 0.6291 and a balanced accuracy of 0.7062. On the Phase 2 (test) leaderboard, it achieved an ROC-AUC of 0.8624, with a macro-F_1 score of 0.6276 and a balanced accuracy of 0.7291. These results indicate that a schema-aware tabular learning approach already captures a substantial portion of the signal relevant for trial-level dosing error risk prediction, highlighting the importance of structured trial design variables.

4.2. Submissions

Participation overview: The shared task attracted 15 official submissions during Phase 1 and 4 submissions during Phase 2.

Leaderboard summary: In Phase 1, several submissions achieved ROC-AUC scores close to that of the baseline, with the best non-baseline entry reaching 0.8599.

In Phase 2, the highest-ranking non-baseline submission achieved an ROC-AUC of 0.8466, followed by a submission with a ROC-AUC of 0.8231.

Tables 6 and 7 present the official leaderboards for Phases 1 and 2, respectively, sorted by ROC-AUC.

Submissions with accompanying papers: Two submissions were accompanied by technical papers presented at the CL4Health workshop, offering complementary methodological perspectives.

ALSmadi (AL-Smadi, 2026): This submission adopts a feature-engineering–driven approach centered on gradient-boosted decision trees (LightGBM) trained on a high-dimensional representation of trial text. The authors construct a 3,451-dimensional feature space combining: (i) sparse lexical features (TF-IDF word unigrams and character n-grams), (ii) dense sentence embeddings (all-MiniLM-L6-v2) (Reimers and Gurevych, 2019; Wang et al., 2020), (iii) domain-inspired pattern features capturing dosing units, routes of administration, dosing frequency expressions, dose adjustments, and related cues, and (iv) scalar probability scores produced by fine-tuned transformer classifiers (BiomedBERT (Gu et al., 2021) and DeBERTa-v3 (He et al., 2021)). The model is trained using a stratified 5-fold ensemble with class-weighted loss to address the severe class imbalance of the dataset. Through systematic ablation experiments, the authors show that sentence embeddings contribute the largest individual performance gain, while sparse lexical features remain strongly complementary despite the presence of modern embedding methods. Their analysis further reveals that feature selection can improve predictive performance: restricting the model to the top 500–1000 most informative features yields slightly higher cross-validation ROC-AUC than using the full feature space, suggesting that aggressive feature selection acts as a regularization mechanism in the high-dimensional setting. Excluding the baseline, the corresponding participant ranked first in both phases (ROC-AUC: 0.8599 in Phase 1; 0.8466 in Phase 2).

Hamnett et al. (Hamnett et al., 2026): This submission investigates domain-specific transformer encoders as the primary mechanism for representing protocol text and compares several biomedical language models, including ClinicalBERT (Alsentzer et al., 2019), PubMedBERT (Gu et al., 2021), BioBERT (Lee et al., 2020), and MedCPT (Jin et al., 2023). The authors embed multiple CT textual fields (e.g., brief summary, detailed description, arm and intervention descriptions) using mean-pooled transformer representations, then combine these embeddings with one-hot encoded categorical trial design variables (including intervention model, allocation, oversight, and phase). The resulting representations are evaluated using a range of downstream classifiers and neural architectures—including logistic regression and support vector classifiers, gradient-boosting models (XGBoost/LightGBM), and feed-forward networks with residual connections. The authors also explore the role of class weighting in mitigating the strong class imbalance present in the dataset. Across their experiments, BioBERT yielded the strongest text representations among the evaluated encoders, whereas concatenating embeddings from multiple transformer models did not provide additive gains in their experimental setting. Excluding the baseline, the corresponding submission ranked second in Phase 2 (ROC-AUC: 0.8231).

Methodological observations: Although the two submissions adopt different modeling paradigms, they share several common design choices. First, both approaches combine textual representations with structured trial metadata derived from ClinicalTrials.gov registry records.

Second, both studies confirm that classical ML models such as gradient boosting remain highly competitive for this task, achieving performance comparable to more complex neural architectures when trained on sufficiently informative representations. Finally, both systems emphasize the complementary roles of lexical and semantic text features: sparse textual cues (e.g., domain-specific vocabulary or dosing expressions) and dense contextual embeddings capture different aspects of protocol descriptions and together contribute to improved predictive performance. These findings illustrate the diversity of viable modeling strategies for the CT-DEB'26 task and suggest that hybrid approaches combining structured trial metadata with semantically rich text representations constitute a promising direction for future research.

Observed performance characteristics: Across both phases, submissions exhibited diverse operating behaviors, with some models favoring conservative predictions with very high specificity and others prioritizing sensitivity to detect a larger fraction of trials with elevated dosing error risk.

Submission	ROC-AUC	F1 (macro)	Sensitivity	Specificity	Bal. Acc.	Brier score	Rank
Baseline	**0.8606**	**0.6291**	0.4950	0.9174	0.7062	0.0747	-
(AL-Smadi, 2026)_475872	0.8599	0.5166	0.0301	**0.9990**	0.5146	**0.0434**	1
IDEAMCVG_459849	0.8561	0.5101	**0.8930**	0.6791	**0.7860**	0.1848	2
ahajighasem_449348	0.8561	0.5101	**0.8930**	0.6791	**0.7860**	0.1848	2
kzan25_460627	0.8561	0.5101	**0.8930**	0.6791	**0.7860**	0.1848	2
uccaeid_480109	0.8561	0.5101	**0.8930**	0.6791	**0.7860**	0.1848	2
mshamani_459802	0.8555	0.5204	0.8528	0.7037	0.7783	0.1885	6
5Gmodels_459893	0.8495	0.5076	0.8729	0.6789	0.7759	0.2001	7
MLtests_465716	0.8490	0.5425	0.7960	0.7497	0.7728	0.1629	8
msha25_465724	0.8489	0.5206	0.8294	0.7088	0.7691	0.1791	9
TCVG_465640	0.8460	0.5362	0.8227	0.7346	0.7787	0.1875	10
alha35_465681	0.8460	0.5362	0.8227	0.7346	0.7787	0.1875	10
mehla40_465693	0.8460	0.5362	0.8227	0.7346	0.7787	0.1875	10
IAU CV team_465636	0.8460	0.5362	0.8227	0.7346	0.7787	0.1875	10
farhankhan_462617	0.8426	0.5157	0.7960	0.7078	0.7519	0.1796	14
yanhh_zry_474851	0.5000	0.4879	0.0000	1.0000	0.5000	0.0473	15

Table 6: Official Phase 1 (validation) leaderboard for CT-DEB'26, sorted by ROC-AUC.

Submission	ROC-AUC	F1 (macro)	Sensitivity	Specificity	Bal. Acc.	Brier score	Rank
Baseline	**0.8624**	**0.6276**	0.5676	0.8906	0.7291	0.0865	-
(AL-Smadi, 2026)_530456	0.8466	0.5072	0.0235	**0.9963**	0.5099	**0.0466**	1
(Hamnett et al., 2026)_510369	0.8231	0.5010	**0.8588**	0.6529	**0.7559**	0.1838	2
MLtests_511064	0.5000	0.4862	0.0000	1.0000	0.5000	0.0538	3
5Gmodels_511131	0.5000	0.4862	0.0000	1.0000	0.5000	0.0538	3

Table 7: Official Phase 2 (test) leaderboard for CT-DEB'26, sorted by ROC-AUC.

This diversity is visible in the leaderboard. For example, the (AL-Smadi, 2026) submission achieves extremely high specificity but very low sensitivity in both phases (Phase 1 sensitivity: 0.0301; Phase 2 sensitivity: 0.0235), indicating a highly conservative strategy that flags very few trials as high risk. In contrast, the (Hamnett et al., 2026) submission favors sensitivity (Phase 2 sensitivity: 0.8588) at the expense of specificity (0.6529), reflecting a more aggressive detection strategy that identifies most positive trials while producing more false positives. While ROC-AUC served as the primary ranking metric, the additional evaluation measures provide complementary perspectives on model behavior under different operating points. Overall, the results illustrate that CT-DEB'26 supports meaningful comparison of modeling strategies ranging from schema-aware tabular baselines to text-centric and hybrid approaches.

4.3. Discussion and Lessons Learned

The results of the CT-DEB'26 shared task highlight several important observations regarding the prediction of dosing errors from clinical trial protocols.

First, the strong performance of the CatBoost baseline demonstrates that a substantial portion of the predictive signal is already captured by structured trial metadata and high-level protocol descriptors. Despite relying only on simple statistics derived from textual fields and schema-aware processing of tabular variables, the baseline achieved the highest ROC-AUC in both phases. This suggests that features related to trial design (e.g., intervention types, masking strategy, study phase, and enrollment characteristics) contain meaningful information about the operational complexity of a trial and therefore about the potential risk of dosing errors.

Second, the submissions illustrate complementary modeling paradigms for leveraging the rich textual information contained in trial documentation. The feature-engineering approach of ALSmadi (AL-Smadi, 2026) demonstrates that carefully constructed lexical and semantic representations combined with gradient-boosted models can effectively capture dosing-related patterns in protocol descriptions. In contrast, the work of Hamnett et al. (Hamnett et al., 2026) emphasizes the role of biomedical transformer models for producing contextualized text embeddings. Despite these methodological differences, both approaches confirm that hybrid models combining structured trial variables with textual representations provide the most promising direction for this task.

Third, the leaderboard results highlight the importance of considering multiple evaluation metrics when analyzing system behavior under strong

class imbalance. While ROC-AUC served as the primary ranking metric because it evaluates the ability of models to discriminate between positive and negative trials independently of a fixed decision threshold, the additional metrics provide important complementary perspectives. For example, some submissions achieved high ROC-AUC values while exhibiting extremely low sensitivity, indicating highly conservative prediction strategies that identify only a small fraction of trials with elevated dosing error risk. Conversely, other systems achieved higher sensitivity but substantially lower specificity, leading to larger numbers of false positive predictions. The Brier score further provides insight into the calibration quality of probabilistic predictions by measuring how closely predicted probabilities align with observed outcomes. Such calibration properties may be particularly relevant for potential downstream use of these models in risk assessment or decision-support settings.

Despite these encouraging results, several limitations of the current benchmark should be acknowledged. The dataset is derived from publicly available ClinicalTrials.gov registry entries and associated protocol documents, which provide only a partial view of the full operational complexity of clinical trials. In practice, dosing errors may arise from detailed procedural instructions, investigator decisions, or site-level deviations that are not fully captured in registry metadata. Moreover, the binary formulation of the task collapses potentially heterogeneous types of dosing errors into a single label, whereas real-world scenarios may require distinguishing between different categories of protocol deviations (e.g., incorrect dose amount, timing deviations, or administration route errors). As a result, the benchmark should be viewed primarily as a proxy task for studying trial-level risk signals rather than as a complete operational detection system.

Another challenge concerns the difficulty of interpreting model predictions in a regulatory or clinical quality assurance context. While gradient-boosting models allow feature importance analysis, the influence of textual features remains difficult to interpret at a semantic level. Developing explainable approaches that identify specific protocol elements or linguistic patterns associated with dosing error risk would be an important step toward practical deployment.

These observations suggest several directions for future research. One promising avenue is the application of more advanced natural language processing techniques to extract structured dosing instructions directly from full protocol documents. For example, information extraction systems could identify dosage amounts, schedules, administration routes, and protocol modification rules from long-form protocol PDFs with non-uniform structures, enabling more fine-grained modeling of dosing complexity. Another direction involves the systematic study of feature importance and model explainability, including techniques such as SHAP analysis (Lundberg and Lee, 2017) or attention-based attribution methods, to better understand which aspects of trial design contribute most strongly to predicted risk. Finally, future datasets could incorporate richer supervision signals, including multi-class taxonomies of medication errors or annotations of specific protocol deviations, enabling models that move beyond binary risk prediction toward more actionable clinical insights.

Overall, the CT-DEB'26 shared task provides a first benchmark for studying dosing error prediction from clinical trial protocols and highlights both the promise and the challenges of applying ML to this safety-critical problem.

5. Conclusions

This paper presented an overview of the CT-DEB'26 shared task on predicting dosing errors in interventional clinical trials. The task introduced a new benchmark dataset derived from ClinicalTrials.gov that combines structured trial design variables with long-form textual protocol descriptions. By framing dosing error prediction as a ML problem, the challenge aimed to stimulate research at the intersection of clinical trial methodology, medication safety, and natural language processing.

The shared task attracted a diverse set of modeling approaches. While the official baseline relied on a schema-aware tabular learning pipeline using structured trial metadata, participating teams explored complementary strategies incorporating richer textual representations, including large-scale feature engineering pipelines and transformer-based biomedical language models. The results demonstrate that structured trial design features already provide substantial predictive signal for identifying trials at risk of dosing errors, while textual representations can provide complementary information about protocol details and dosing instructions.

Another important direction concerns model interpretability and explainability. In order for predictive models to be used in clinical trial planning or quality assurance settings, it is essential to understand which aspects of trial design contribute most strongly to predicted dosing error risk. Techniques for feature attribution and explainable ML may therefore play a key role in translating predictive models into actionable insights for trial designers and regulatory stakeholders.

Finally, advances in natural language processing offer opportunities to extend this benchmark by extracting more detailed dosing-related infor-

mation from full protocol documents. Methods for structured information extraction from long protocol PDFs could enable richer representations of dosing schedules, administration rules, and protocol modifications. Such developments would move beyond trial-level risk prediction toward automated identification of potential dosing issues during the early design stages of clinical studies.

Overall, CT-DEB'26 represents a step toward the development of data-driven methods for improving medication safety in clinical research. We hope that the dataset, baseline models, and shared task results will provide a useful foundation for future work on ML methods supporting safer and more efficient clinical trial design.

6. Acknowledgments

This work was supported by Innosuisse - the Swiss Innovation Agency - grant number 114.721 IP-ICT.

7. Bibliographical References

Mohammad AL-Smadi. 2026. Automated detection of dosing errors in clinical trial narratives: A multi-modal feature engineering approach with lightgbm. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma de Mallorca, Spain. European Language Resources Association (ELRA).

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Jeffrey K Aronson. 2009. Medication errors: definitions and classification. *British journal of clinical pharmacology*, 67(6):599–604.

Paul Beninger. 2020. Signal management in pharmacovigilance: a review of activities and case studies. *Clinical therapeutics*, 42(6):1110–1129.

Andrea Cipriani and Corrado Barbui. 2010. What is a clinical trial protocol? *Epidemiology and Psychiatric Sciences*, 19(2):116–117.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Rachel Ann Elliott, Elizabeth Camacho, Dina Jankovic, Mark J Sculpher, and Rita Faria. 2021. Economic analysis of the prevalence and clinical and economic burden of medication error in england. *BMJ Quality & Safety*, 30(2):96–105.

European Medicines Agency. 2023. Guideline for the notification of serious breaches of regulation (eu) no 536/2014 or the clinical trial protocol.

European Parliament and Council of the European Union. 2014. Regulation (eu) no 536/2014 of the european parliament and of the council of 16 april 2014 on clinical trials on medicinal products for human use, and repealing directive 2001/20/ec. Official Journal of the European Union, L 158/1.

Sohrab Ferdowsi, Nikolay Borissov, Julien Knafou, Poorya Amini, and Douglas Teodoro. 2021. Classification of hierarchical text using geometric deep learning: the case of clinical trials corpus. *arXiv preprint arXiv:2110.15710*.

Sohrab Ferdowsi, Julien Knafou, Nikolay Borissov, David Vicente Alvarez, Rahul Mishra, Poorya Amini, and Douglas Teodoro. 2023. Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study. *Patterns*, 4(3).

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. In *ACL*.

Leon Hamnett, Favour Igwezeke, Joseph Itopa, and Mary Adetutu Adewunmi. 2026. Detecting dosing errors in clinical trials using domain-specific transformer embeddings and classification models. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma de Mallorca, Spain. European Language Resources Association (ELRA).

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*.

Félicien Hêche, Sohrab Ferdowsi, Anthony Yazdani, Sara Sansaloni-Pastor, and Douglas Teodoro. 2026. Early risk stratification of dosing errors in clinical trials using machine learning. *arXiv preprint arXiv:2602.22285*.

Alexander Hodkinson, Natasha Tyler, Darren M Ashcroft, Richard N Keers, Kanza Khan, Denham Phipps, Aseel Abuzour, Peter Bower, Anthony Avery, Stephen Campbell, et al. 2020. Preventable medication harm across health care settings: a systematic review and meta-analysis. *BMC medicine*, 18:1–13.

Félicien Hêche, Sohrab Ferdowsi, Anthony Yazdani, et al. 2026. Machine learning for medication error detection: A scoping review. *Research Square*. Research Square preprint, Version 1. Available at: https://doi.org/10.21203/rs.3.rs-8919709/v1.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). 2016. Integrated addendum to ich e6(r1): Guideline for good clinical practice e6(r2). ICH Harmonised Guideline.

Qiao Jin, Xin Yuan, Sheng Wang, Yifan Zhang, Jiayuan He, and Zhiyuan Liu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. In *ACL*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Linda Martin, Melissa Hutchens, and Conrad Hawkins. 2017. Trial watch: clinical trial cycle times continue to increase despite industry efforts. *Nature Reviews Drug Discovery*, 16(3):157–158.

MedDRA Maintenance and Support Services Organization. 2025. Medical dictionary for regulatory activities (meddra). https://www.meddra.org/. Accessed: 2025-09-12.

Andrew Mulcahy, Stephanie Rennane, Daniel Schwam, Reid Dickerson, Lawrence Baker, and Kanaka Shetty. 2025. Use of clinical trial characteristics to estimate costs of new drug development. *JAMA Network Open*, 8(1):e2453275.

National Library of Medicine (US). 2000. Clinicaltrials.gov. Bethesda (MD): National Institutes of Health (US).

Pydantic AI. Data validation using python type hints. https://github.com/pydantic/pydantic.

Python Software Foundation. enum — support for enumerations. https://docs.python.org/3/library/enum.html. Accessed: 2026-01-07.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*.

Mahsa Rouhani, Maryam Mousavi, Mohammad Mahdi Kooshyar, Soodabeh Shahidsales, Yasha Makhdoumi, Amir Amirabadi, and Sepideh Elyasi. 2018. Application of a chemotherapy standard form in patients with breast cancer: comparison of private and public centers. *Jundishapur Journal of Natural Pharmaceutical Products*, 13(3).

Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. 2022. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12(7):3049–3062.

Rayhan A Tariq, Rishik Vashisht, Ankur Sinha, and Yevgeniya Scherbak. 2018. Medication dispensing errors and prevention.

Rayhan A. Tariq, Rishik Vashisht, Ankur Sinha, and Yevgeniya Scherbak. 2025. *Medication Dispensing Errors and Prevention*. StatPearls Publishing, Treasure Island (FL).

Douglas Teodoro, Nona Naderi, Anthony Yazdani, Boya Zhang, and Alban Bornet. 2025. A scoping review of artificial intelligence applications in clinical trial risk assessment. *medRxiv*, pages 2025–01.

U.S. Food and Drug Administration. 2024. Title 21, code of federal regulations, part 312: Investigational new drug application. Code of Federal Regulations. Accessed via the Electronic Code of Federal Regulations (eCFR).

Bernard Vrijens, Antoine Pironet, and Eric Tousset. 2024. The importance of assessing drug exposure and medication adherence in evaluating investigational medications: ensuring validity and reliability of clinical trial results. *Pharmaceutical Medicine*, 38(1):9–18.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*.

Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. 2018. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286.

Olivier J Wouters, Martin McKee, and Jeroen Luyten. 2020. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, 323(9):844–853.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Anthony Yazdani, Alban Bornet, Philipp Khlebnikov, Boya Zhang, Hossein Rouhizadeh, Poorya Amini, and Douglas Teodoro. 2025. An evaluation benchmark for adverse drug event prediction from clinical trial results. *Scientific Data*, 12(1):424.