WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Structured vs. Unstructured Data for Clinical Trial Risk Prediction: A Comparative Machine Learning Study

Hiten Mehta[1], Gurshaan Singh[2], Yash Sharma[3], Ashmit Agrawal[4]

University Institute of Engineering, Chandigarh University, Punjab, India
Email: hiten.cuchd@gmail.com, gurshaan@tiwanas.me, yashyashrgp@gmail.com, ashmit1357@gmail.com

ABSTRACT:

Forecasting the outcome of clinical trial is a long standing challenge in pharmaceutical research and healthcare management. This paper presents a systematic evaluation of machine learning approaches for three-tier risk categorization of clinical trials, comparing models trained on structured administrative metadata against a model trained on unstructured free text. Using 50,000 records from the louisbrulenaudet/clinical-trials dataset on hugging face, originally sourced from Clinical-Trials.gov - three logistic regression classifiers were developed and evaluated: a baseline structured model, a class-balanced structured model and text-based model using TF-IDF vectorization. Risk labels - low, medium, and high, were derived from each trial's operational status field. The TF-IDF model achieved superior performance on balanced evaluation metrics attaining a balanced accuracy of 0.5123 and a macro F1-score of 0.4959, compared to the 0.3335 and 0.2365 for the unweighted structured baseline. Coefficient analysis also displayed readable textual cues in accordance with every risk category. These findings indicate that unstructured trial text represents predictive information not reflected in structured metadata, and that neglecting imbalance in the classes when training a classifier can significantly reduce its utility.

Index Terms - clinical trial's risk, machine learning, TF-IDF, NLP, class imbalance, logistic regression

I. INTRODUCTION

Clinical trials are the evidencing process by which we put new drugs and devices to test regarding their safety and effectiveness. These researches are among the most resource consuming undertakings in medical research, a single trial can cost approximately millions of dollars and take over years to complete. Despite such massive investments, numerous research works are never completed. Voluntary withdrawals, legal problems, dismissals and numerous other factors do not only lead to an irreparable financial loss, but also lead to delays in delivering treatments to the patients who require them [1].

Early flagging of rish-risk trials is directly valuable to sponsors, regulators and investigators. Soon indications can result in protocol redesign, resource reallocation or more active monitoring plans prior to a trail reaching a dead end. Unfortunately, most risk assessment approaches are largely manual, relying on expert judgement, heuristic scoring systems, and retrospective case analysis that do not scale to current requirements of the world [2]. Standard structured fields such as sponsor category, enrollment count, study type are easy to work upon but only provide a high-level overview.

Besides the structured information, clinical trials also contain rich textual description with brief summaries, eligibility criteria and listed medical conditions. These narratives explain the scientific rationale of study, clinical complexity of the population being studied, and investigator's intent. These factors may or may not play a role in trial's success, but have not been widely used in predictive modeling research [3].

Recently, advances in the natural language processing have made it feasible to extract and use the text data effectively at scale. Bag-of-words techniques like TF-IDF are easily interpretable while using less computational resources, whereas the transformer based models offer richer semantic representations, but require much more computational resources [4], [5]. For a basic comparison among data modalities, we prefer a simple and transparent approach as it isolates the informational contribution of data without confounding it with model complexity.

The primary question being addressed by the study is, to what extent can the unstructured textual data improve clinical trial risk classification over structured administrative features? We also examine other things like the impact of class imbalance on the model's behaviour and analysis of interpretability of coefficient-based keyword. We also evaluate three logicstic regression models with a structured baseline, class balanced structured model, and a TF-IDF text model.

The contributions of this work are:

(a) A systematic and controlled comparison of structured and text-based feature sets for trial-level risk stratification using a large public dataset.

(b) Empirical evidence that the TF-IDF representations of trial narratives outperform the structured metadata on balanced evaluation metrics.

(c) Demonstration that uncorrected class imbalance leads to very high raw accuracy while producing classifiers that ignore minority classes.

(d) A lexical signature analysis which was interpretable and identified lexical signatures based on each tier of risk.

II. RELATED WORK

The use of machine learning on ClinicalTrials.gov data has been steadily increasing in the last two years. The initial research was aimed at forecasting

the success or failure of the trial based on structured factors: phase, therapeutic area, type of sponsor and history of previous approval. Wong et al. [2] and Siah et al. [1] showed that these features do have meaningful predictive signal, phase and therapeutic area being two of the most significant predictive variables. These studies established a structured-data baseline for the problem but did not exploit the textual content available in trial registrations.

NLP has been applied in adjacent clinical contexts. Weng et al. [3] examined the formal representation of eligibility criteria, showing that free-text clinical criteria follow systematic linguistic patterns amenable to automated extraction and classification. Studies on adverse event reports and clinical notes [6] have demonstrated that unstructured medical text contains extractable safety signals — a finding that motivates similar analysis of trial-level narratives.

Work on predicting trial completion or toxicity using computational methods has employed a range of feature types. Gayvert et al. [7] developed a data-driven framework for predicting clinical trial toxicity using structural and target-based drug properties. Artemov et al. [8] combined transcriptomic profiles with trial metadata in a deep learning framework for outcome prediction. These studies primarily operated in the structured or molecular data regime, largely leaving free-text trial descriptions unexplored.

Transformer based language models have since been utilized in clinical text with great success. Clinical BERT [4] was trained on clinical notes to predict readmission to hospitals; Bio BERT [5] - pre-trained on PubMed abstracts as well as on PMC full texts demonstrated consistent gains over bag-of-words baselines on biomedical NLP tasks such as named entity recognition and relation extraction. These findings indicate that more benefits might be obtained even than TF-IDF by way of rich textualization of clinical trial records.

Nevertheless, in the literature, a systematic comparison of structured and unstructured modalities, specifically of risk stratification at trial level, with the granularity of public registries and available modelling tools, has not been conducted. This paper fills that gap.

III. DATASET AND PREPROCESSING

A. Data Source

The dataset that we choose is louisbrulenaudet/clinical-trials dataset, from the Hugging Face platform [9]. This data is in the form of trial registration documents which can be found at ClinicalTrials.gov [10], the publicly accessible registry of clinical studies at the NIH, and one of the largest such documentaries in the world. Each row consists of a combination of constant structured fields and text filled fields, making it fully fit for a cross modality comparison.

B. Sampling Strategy

The entire dataset was streamed to eliminate the need to load the entire corpus in memory in one go. A subsample of 50,000 trial records was selected one at a time and transformed into a pandas DataFrame to be analyzed further onwards. The sample size in this case was selected to ensure that both statistical representativeness and computational tractability are maintained in a typical cloud computing setting.

C. Target Variables: Risk Level Determination

There is no pre-classified risk variable in the dataset in the start. The target variable, risk level, was decided based on the overall status field that captures the current recruitment and operational status of each trial using the following mapping:

- ***Low Risk:*** Trials with status COMPLETED. These experiments did not end in disastrous failure of operation.
- ***Medium Risk:***

RECRUITING, ACTIVE_NOT_RECRUITING, NOT_YET_RECRUITING or ENROLLING_BY_INVITATION statuses. These are continuous experiments that have moderate operational uncertainty.

- ***High Risk:*** The rest of the statuses, namely TERMINATED, WITHDRAWN, SUSPENDED, WITHHELD, and UNKNOWN, which reflect premature termination or indeterminate status, both reflect high risk.

The resultant distribution of classes will be presented in Table I and plotted in Fig. 1. As demonstrated, Low Risk trials make up the majority class with 54.9, which forms a moderate class balance that directly affects the design of classifiers.

TABLE I: DISTRIBUTION OF DERIVED RISK LEVELS (N = 50,000)

Risk Level	*Count*	*Percentage*
Low Risk	27,440	54.9%
High Risk	11,822	23.6%
Medium Risk	10,738	21.5%

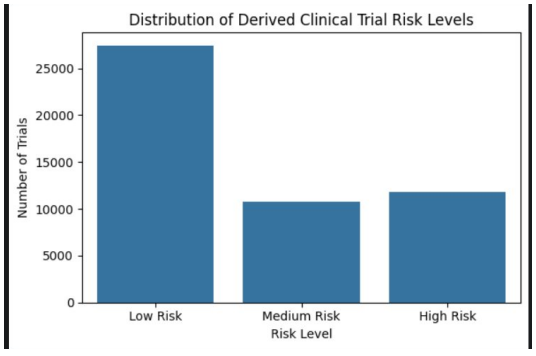


Fig. 1. Giving out derived clinical trial risks in the 50,000 record sample. The majority class is the Low Risk (COMPLETED) trials.

Fig. 2 shows the one-to-one correspondence between every raw overall_status value and the risk category associated with it, and the prevalence of COMPLETED trials and the proportion of active and failed status codes under the Medium and High-Risk levels.

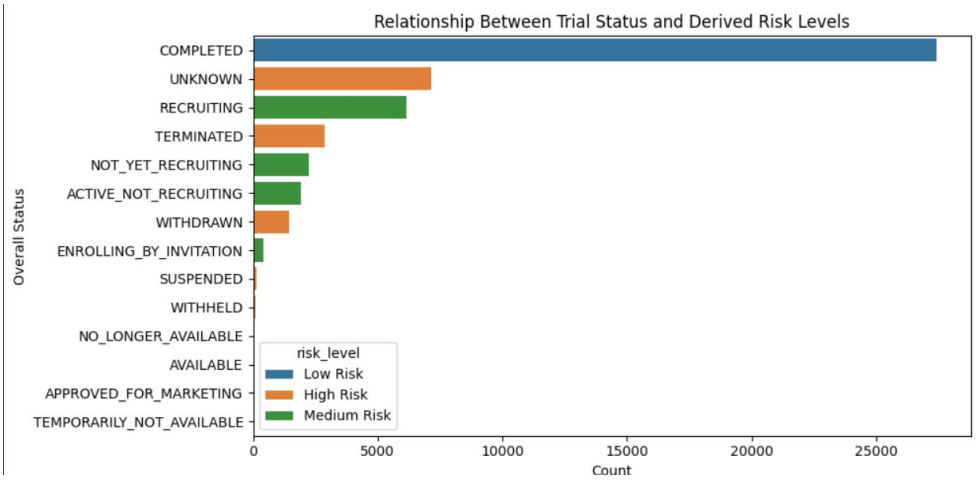


Fig. 2. Correlation between the trial status (raw) values and calculated risk levels

Fig. 3 demonstrates the way the risk levels are distributed in the three types of studies in the dataset. Interventional trials represent the highest absolute number of all three risk groups, and the Expanded Access trials are nearly all of High Risk, which is again in line with their exploratory, non-standard nature.

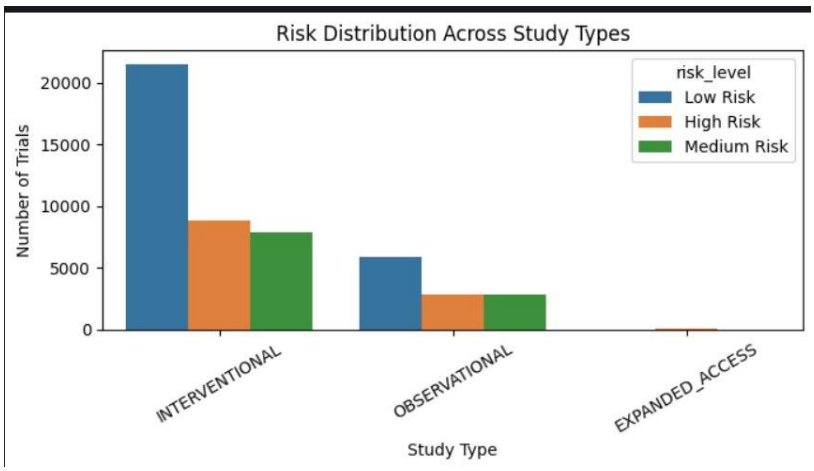


Fig. 3. The distribution of the level of risk involving different types of studies. The majority of the three risk categories are dominated by interventional trials; Expanded Access trials are nearly all High Risk.

D. Phase Normalization

Raw form in the phases field was very heterogeneous, with values represented as Python list objects (e.g. [PHASE1, PHASE2]) or plain strings or None values and the special value [NA]. These representations were further reduced through a hierarchical normalization process to five categories, including Phase 1, Phase 2, Phase 3, Phase 4, and Unknown. The order of precedence was descending phase order; all the entries that contained None, [NA], or unresolvable values were put under Unknown. Fig. 4 illustrates the distribution of raw phase labels prior to the application of this procedure, which revealed the label fragmentation which inspired it.

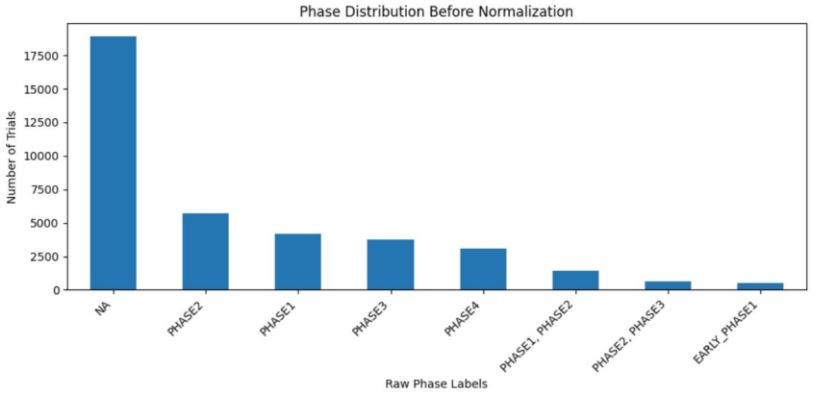


Fig. 4. Raw phase labels Distribution before normalization.

The clean distribution is given in Fig. 5. The Unknown category prevails with 61.3 which is an observational and non-interventional studies that do not have any formal phase designation. The normalized distribution is also reported in Table II.

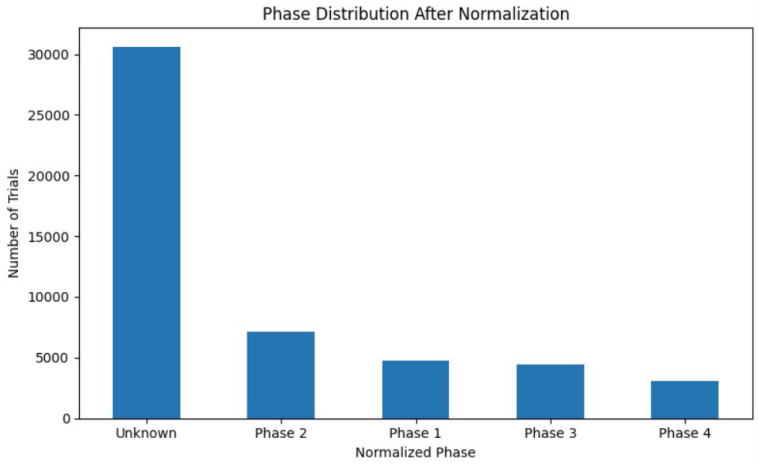


Fig. 5. Normalized trial phases distribution after consolidation.

TABLE II: TRIAL PHASE DISTRIBUTION AFTER NORMALIZATION (N = 50,000)

Phase	*Count*	*Notes*
Unknown	30,631	Missing or unspecified phase
Phase 2	7,142	2nd-stage human clinical testing
Phase 1	4,714	First-in-human safety/dosage
Phase 3	4,419	Large-scale efficacy and safety
Phase 4	3,094	Post-market surveillance

The impact of the normalization process is given in Fig. 6, which compares the counts of raw labels and compiled labels. The eight heterogeneous raw categories are reduced to five clean canonical phases, which are evident. Instead of considering Unknown missing data, it was included in all the models as an informative category.

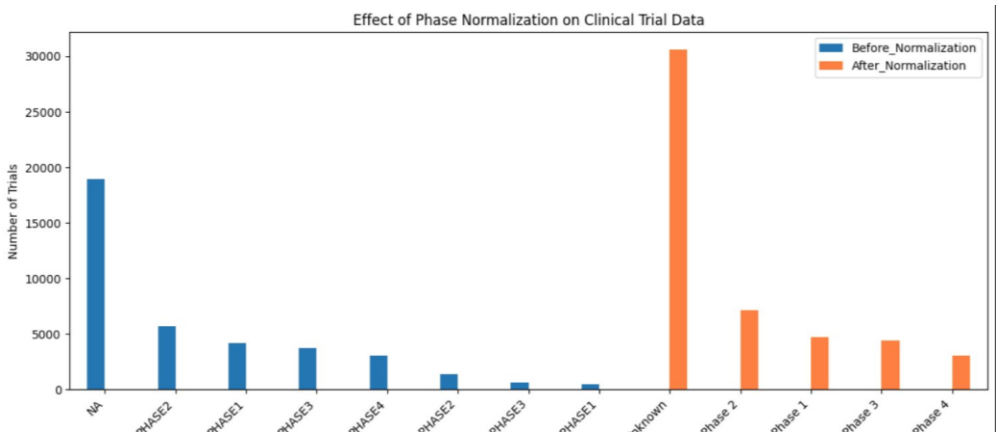


Fig. 6. Effect of phase normalization: before (blue) vs. after (orange). The raw labels that were fragmented are grouped into five semantically consistent categories.

In order to determine whether the phase information represents risk-relevant information, a cross-tabulation heatmap of the risk level proportions per normalized phase is provided in Fig. 7.

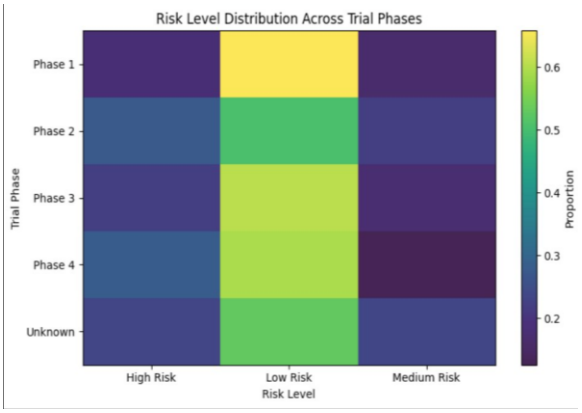


Fig. 7. Heatmap showcasing risk level proportions for each normalized trial phase.

IV. METHODOLOGY

A. Research Design

Three logistic regression classifiers were trained and evaluated under a common experimental framework: (1) a baseline structured model with no class imbalance correction; (2) a class-balanced structured model using inverse-frequency weighting; and (3) a text-based model trained on TF-IDF representations of free-text trial fields. All models were evaluated on the same stratified train-test split.

B. Structured Feature Engineering

Ten structured features were selected based on conceptual relevance to clinical trial risk and practical availability in the dataset: phase_normalized, study_type, enrollment_count, sex, healthy_volunteers, minimum_age, maximum_age, lead_sponsor_class, sampling_method, and oversight_has_dmc.

Missing values in numerical columns (enrollment_count, minimum_age, maximum_age) were imputed using column-wise medians to limit the influence of outliers. Missing categorical values were replaced with the string Unknown, treating absence of information as a distinct and potentially informative category. Categorical features were one-hot encoded using pandas' get_dummies function with drop_first=True to prevent multicollinearity.

C. Text Feature Engineering

To create a single text input(called text_blob) for the text based model, we concatenated three fields, brief_summary, eligibility_criteria and conditions because they collectively capture the scientific rationale, participant selection logic, and the medical context of a trial. All of these affect weather the trial will complete successfully or not. The fields with null values were replaced with empty string before concatenation.

Then, we used scikit-learn's TfidfVectorizer to convert the text to numerical vectors. The configuration was set to use 30,000 most important words, to include both unigrams(single words) and bigrams(pair of words), to ignore common english words, to disregard unusual words that appear in less than 5 trials and to disregard the words that appear in more than 90% trials. As a result, we got a sparse matrix which was directly used as input for text classifier.

D. Model Training

We used logistic regression as the base algorithm for all three models because of its computational efficiency, interpretability and multiclass performance. This made it suitable for this study where we wanted to isolate the contribution of feature type rather than model architecture. We used the scikit-learn's implementation of LogisticRegression, configured to handle multiple classes (multinomial objective), allowed maximum iteration of about 1000-3000 depending upon the feature space size and used parallel processing to speed up computation(n_jobs = -1).

For the baseline structured model, we did not use any class weighting. This allowed it to reflect the natural distribution of traning errors, but for the other models, balanced structured and text model, class_weight=balanced was set to compute class weights in inverse proportion to class frequencies. This helped the model minimise errors on the minority class during training by making such mistakes more costly.

E. Train-Test Split and Evaluation

The dataset was divided into 80% training and 20% test set using stratified sampling (stratify=y, random_state=42), where the class distribution of each set was preserved. Fig. 8 shows the class distribution in the training set. Here, the Low Risk majority class (label 1) accounts for about 55% of training instances. Therefore, the imbalance correction techniques are important to ensure that model performs effectively on minority classes.

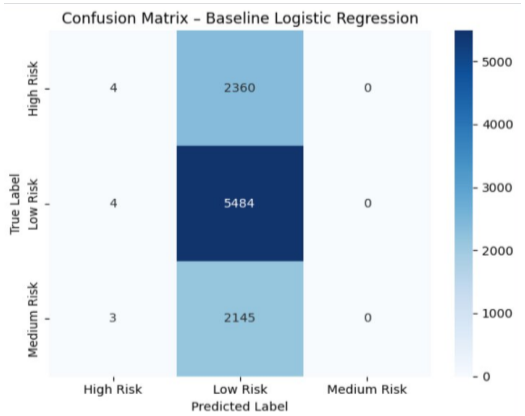


Fig. 8. Class distribution in training set (encoded labels: High Risk = 0, Low Risk = 1, Medium Risk = 2).

For the text-based model, TF-IDF matrices were fitted on the training set before applying to test set in order to minimize data leakage.

The evaluation of the model's performance was done using four metrics - standard accuracy, balanced accuracy (mean per-class recall), macro F1 score (unweighted mean f1 scores per class) and weighted f1 scores. Balanced accuracy and macro F1 were used as the main indicators due to the class imbalance in the dataset. Additionally, confusion matrices were made for each model to analyze classification patterns per class.

F. Keyword Extraction

To support model interpretability, the fitted logistic regression coefficients for the TF-IDF model were analyzed. For each risk class, the twenty vocabulary terms with the highest positive coefficients were extracted and visualized. These terms represent the most discriminative lexical features for each class.

V. RESULTS

A. Overall Model Performance

Table III shows how well the three models did on all four metrics. The Structured Baseline had the highest raw accuracy (0.5489), but the Text-Based TF-IDF model did much better than both structured models on the two metrics that are best for imbalanced classification: balanced accuracy and macro F1.

TABLE III: COMPARATIVE MODEL PERFORMANCE SUMMARY

Model	*Acc.*	*Bal. Acc.*	*Macro F1*	*Wtd. F1*
Structured (Baseline)	0.5489	0.3335	0.2365	0.3891
Structured (Balanced)	0.4227	0.4062	0.3932	0.4385
Text-Based TF-IDF	0.5252	0.5123	0.4959	0.5336

B. Structured Baseline Model

The baseline model got an accuracy of 0.5489 when it was trained on structured features without fixing the class imbalance. But when you look at the data by class, you can see that this number is mostly due to the fact that one class is larger than the others. The model got a recall of 1.00 for the Low Risk class, but it got almost no recall for the High Risk and Medium Risk classes. This meant that almost every test instance was put in the majority class. The macro F1-score of 0.2365 and the balanced accuracy of 0.3335 show that this didn't work. This behavior is directly validated in Figure 9, the confusion matrix of the baseline model. Almost all test instances go to the Low Risk column, no matter what the true label is. The number of counts of the High Risk and Medium Risk diagonals is small.

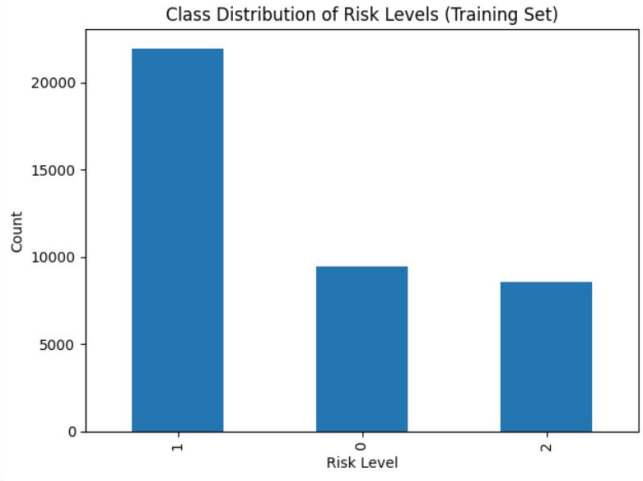


Fig. 9. Confusion matrix of Structured Baseline model. Almost all the cases are foretold as Low Risk and degenerate majority-class behavior is established.

C. Structured Balanced Model

The option of class_weight = balanced had a significant influence on the model. The raw accuracy went down to 0.4227, but the balanced accuracy went up to 0.4062 and the macro F1 went up to 0.3932. That is a 66% relative improvement of macro F1 compared to the baseline. Each of the classes was much more fairly recalled: High Risk 0.33, Medium Risk 0.43, and Low Risk 0.46. The balanced model's confusion matrix, shown in Fig. 10, shows a much more sspread-out prediction pattern is, with correct classifications showing up in all three risk categories. This contrasts with the baseline where its predictions were mostly in the Low Risk category.

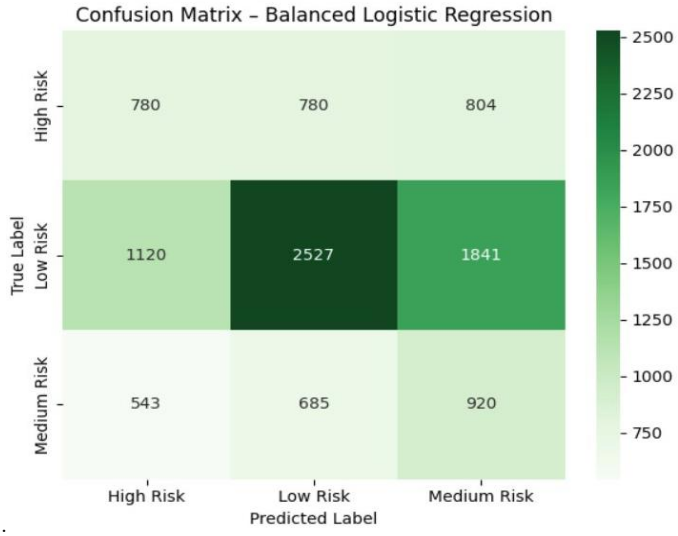


Fig. 10. Confusion matrix of Structured Balanced model. Class weighting yields a fairer prediction distribution than the baseline.

D. Text-Based TF-IDF Model

The raw accuracy of the text model was 0.5252, slightly lower than the structured baseline, but the balanced performance of the text model was much improved. The balanced accuracy is 0.5123, 26 and 54 percent better than the balanced structured model and unweighted baseline, respectively. The macro F1 score of 0.4959 was also better than both structured options.

The F1 scores for each class were 0.35 (High Risk), 0.62 (Low Risk) and 0.52 (Medium Risk). The recall for Medium Risk was especially high at 0.59, which means that the text of currently active trials shows detectable linguistic patterns. This is supported by the confusion matrix in Fig. 11: the text model has a more uniform distribution of the predictions among all three classes. High Risk (904), Low Risk (3,070) and Medium Risk (1,278) diagonal counts are all substantial.

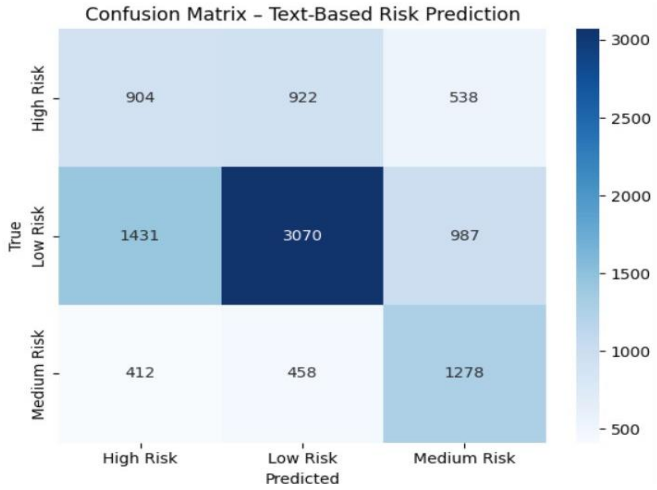


Fig. 11. Confusion matrix for the Text-Based TF-IDF model. The most balanced prediction distribution of the three models, with meaningful correct classifications across all tiers.

E. Keyword Analysis

The coefficient analysis of the TF-IDF model found different lexical patterns for each risk class. Figure 12 shows the most important risk-indicative terms for the High Risk and Low Risk groups. High Risk terms include words that are related to serious or specialized medical conditions, such as colonization, vertigo, morphine, adenomyosis, and amputation. The phrase "expanded access" also falls into this category. It means investigational treatment that is not part of the standard trial pathway. Low-risk terms like "purpose study," "dsm iv," "included," and "hypothesis" show the finalized, retrospective language that is typical of finished trials.

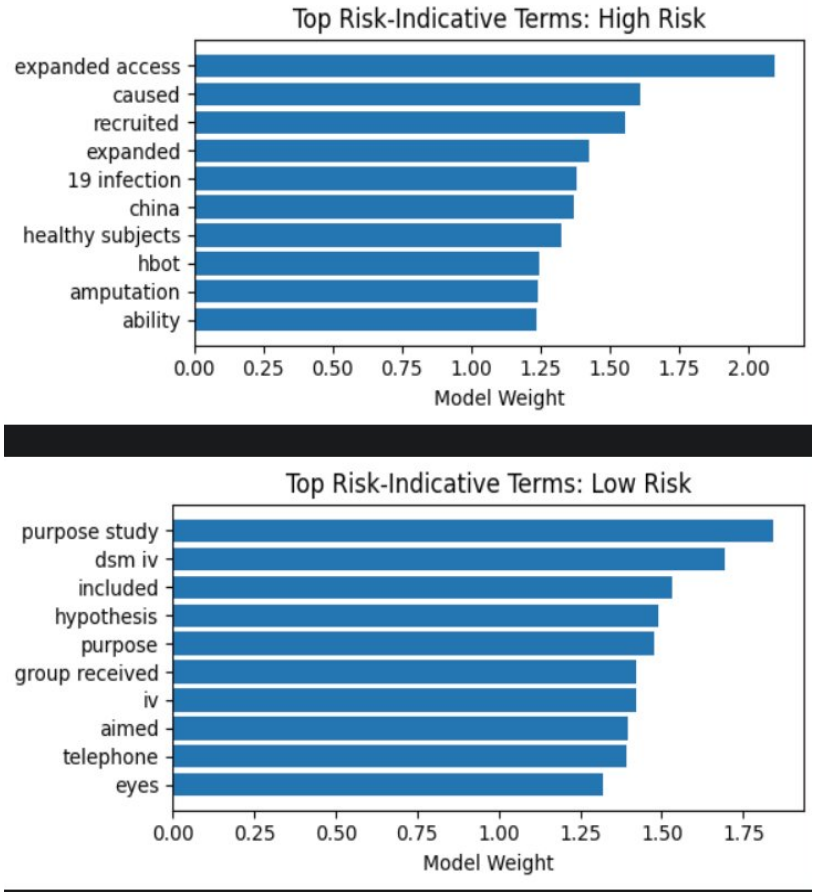


Fig. 12. Top risk-indicative TF-IDF terms for High Risk (top) and Low Risk (bottom) classes, ranked by logistic regression coefficient magnitude.

Fig. 13 presents the analogous visualization for Medium Risk. The dominant terms - aims, goal, goal clinical, main questions, aims answer - constitute a prospective, objective-oriented vocabulary consistent with active recruiting trials. Together, these class-specific lexical profiles provide a degree of face validity to the model's predictions and suggest that the learned associations are clinically grounded rather than spurious.

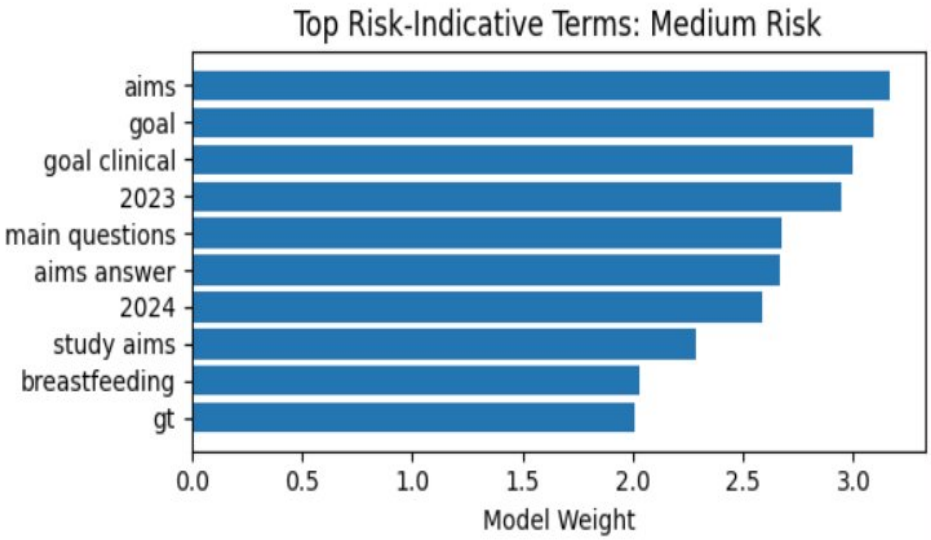


Fig. 13. Top risk-indicative TF-IDF terms for the Medium Risk class. Prospective, goal-oriented vocabulary dominates, consistent with actively recruiting trials.

VI. DISCUSSION

A. Interpretation of Key Findings

The conclusion of it all is that the TF-IDF text features are considerably beating up the structured metadata in the balanced evaluation matrices. This has a practical and theoretical implication. In fact, this means that the predictive risk indicators are imbedded in the manner the investigators are articulating their trials: the scientific purposes, the target patient population, and the medical issues under consideration. This means theoretically that the categorical field and the numerical field which the administrative data bases rank highest do not point fully to the things which make the trials distinctive in the terms

of risk of failure.

The failure of the baseline structured model is educational. The uncooked correctness of 0.5489 is quite natural, and it is predominately a predictor of the majority group. This is an established breakdown of unbalanced classification [11]. It demonstrates that the use of standard accuracy is not a preferable method to assess performance in the cases when the distributions of classes are unequal. This is made obvious at first glance when you consider the balance accuracy and macro F1: scores of 0.3335 and 0.2365 indicate that the model has not been taught any meaningful decision boundaries regarding any classification besides the plurality.

B. The Role of Class Imbalance

The two structured models comparison indicates the influence of the weighting of the classes on the outcomes. With a much more simple reweighting scheme even imbalance can be dealt with and the model becomes much more successful in finding minority-class trials. Relative gain of the balanced structured model of 66% of disparity in macro F1, on top of the baseline, using the same number of features, indicates how poor uncorrected imbalance can negatively impact on the classifier.

However, the even balanced structured model has not shown such good performances on the main metrics as compared to text model. This is to demonstrate the structured features employed despite its relevance do not deliver adequate discriminate data in a combination that can be utilized to deliver dependable three-class risk stratification. Trials in general are characterized by metadata level features. They will document the nature of trial, but fail to document clinical and operating information that seemingly leads to diverse levels of risk.

C. Interpretability and Clinical Plausibility

The analysis of keywords give qualitative justification to the validity of the text model(Figs. 12 and 13). Correlation of High Risk vocabulary to some of the more serious conditions and access programs in effect such as adenomyosis, amputation and the expanded access suggests that the model has been trialed or that it has been more complex to operate which also generates increased operational risk. The retrospective orientation of the Low Risk vocabulary is appropriate to the reality that closed trials are in the past: they are presently described using the past tense and report about the findings of the discovery. The words used in the vocabularies to represent the status of RECRUITING or ACTIVE_NOT_RECRUITING of that group are in the Medium risk language.

Such trends are clinically feasible and alleviate the fear that the text model is being based on spurious associations. They also create a possible solution to the creation of decision-support tools to monitor the trials with ease.

D. Limitations

There are a couple of issues that we ought to know about this study. Risk classification scheme is rule based and indicates one of the methods of connecting the operational status to risk. The other mappings can be rationalized; separating those that are TERMINATED (which can be extremely suggestive of safety concerns) and those that are WITHDRAWN (which may be extremely suggestive of logistical or funding concerns) may provide a more clinically refined target variable.

Second, TF-IDF treats text as a bag of terms and discards syntactic structure, word order, and semantic meaning. Domain-adapted transformer models such as BioBERT [5] or ClinicalBERT [4] might extract considerably richer representations from the same textual content. The performance gap between TF-IDF and more advanced language models on this specific task remains an open question.

Third, this study did not combine modalities. A multimodal architecture that fuses structured and textual features — perhaps through feature concatenation or a stacking ensemble — may outperform either modality in isolation, and represents a natural direction for follow-on work.

Fourth, all data originate from a single registry (ClinicalTrials.gov), which focuses predominantly on US-registered studies. Generalizability to other jurisdictions or to industry-proprietary data remains untested.

Finally, logistic regression is a linear classifier and may not capture complex nonlinear interactions among features. Gradient-boosted trees or support vector machines might

achieve higher performance on the same structured feature set.

E. Implications for Practice

For trial sponsors and regulatory bodies, these results suggest that automated text-screening tools could flag high-risk registrations at the time of submission, enabling proactive interventions before operational failure occurs. For researchers building predictive models of trial outcomes, the findings argue for including textual features rather than relying exclusively on the structured metadata that is most convenient to extract. More broadly, the study reinforces the value of open-access clinical trial registries as training data sources for machine learning applications in drug development and clinical research.

VII. CONCLUSION

This study compared three logistic regression classifiers for clinical trial risk stratification — a structured baseline, a class-balanced structured model, and a text-based TF-IDF model — on 50,000 records from ClinicalTrials.gov. Across the balanced evaluation metrics most appropriate for an imbalanced three-class problem, the text model was consistently superior, achieving a balanced accuracy of 0.5123 and a macro F1-score of 0.4959 against respective scores of 0.3335 and 0.2365 for the unweighted structured baseline.

There are two other secondary findings that are also important. The misleadingly high raw accuracy of the structured baseline (0.5489) concealed an almost total inability to recognize trials of the minority-class, which was only removed when class weighting was introduced, highlighting the danger of relying on accuracy as a leading measure in an unequal environment. The coefficient structure of the text model depicted in Fig. 12 and 13 generated clinically reasonable risk-indicative vocabulary and hence bestowed some interpretive validity to its forecasts.

Future directions should include studies on transformer-based text representations (e.g., BioBERT, ClinicalBERT), multimodal feature fusion, as well as more specific risk taxonomies which distinguish between different forms of trial failure. Future validation on withheld registration data and third-party data would further ascertain the extrapolation of the findings. The increasing availability of large-scale clinical trial registrations represents a significant, underutilized resource for machine learning applications aimed at improving the efficiency and safety of the drug development pipeline.

REFERENCES

[1] K. W. Siah et al., "Predicting drug approvals: The Novartis data science and artificial intelligence challenge," Patterns, vol. 2, no. 8, p. 100312, Aug. 2021. doi: 10.1016/j.patter.2021.100312

[2] C. H. Wong, K. W. Siah, and A. W. Lo, "Estimation of clinical trial success rates and related parameters," Biostatistics, vol. 20, no. 2, pp. 273–286, Apr. 2019. doi: 10.1093/biostatistics/kxx069

[3] C. Weng, S. W. Tu, I. Sim, and R. Richesson, "Formal representation of eligibility criteria: A literature review," J. Biomed. Inform., vol. 43, no. 3, pp. 451–467, Jun. 2010. doi: 10.1016/j.jbi.2009.12.004

[4] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," arXiv preprint arXiv:1904.05342, 2019.

[5] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020. doi: 10.1093/bioinformatics/btz682

[6] R. Harpaz, W. DuMouchel, P. LePendu, A. Bauer-Mehren, P. Ryan, and N. H. Shah, "Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system," Clin. Pharmacol. Ther., vol. 93, no. 6, pp. 539–546, Jun. 2014. doi: 10.1038/clpt.2013.24

[7] K. M. Gayvert, N. S. Madhukar, and O. Elemento, "A data-driven approach to predicting successes and failures of clinical trials," Cell Chem. Biol., vol. 23, no. 10, pp. 1294–1301, Oct. 2016. doi: 10.1016/j.chembiol.2016.07.023

[8] A. V. Artemov et al., "Integrated deep learned transcriptomic and structure-based predictor of clinical trials outcomes," bioRxiv, 2016. doi: 10.1101/095653

[9] L. Brulenaudet, "louisbrulenaudet/clinical-trials [Dataset]," Hugging Face, 2023. [Online]. Available: https://huggingface.co/datasets/louisbrulenaudet/clinical-trials

[10] U.S. National Library of Medicine, "ClinicalTrials.gov," 2023. [Online]. Available: https://clinicaltrials.gov

[11] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009. doi: 10.1109/TKDE.2008.239

Appendix A: RISK LEVEL DERIVATION LOGIC

The following mapping was applied programmatically to every record prior to modelling:

```
def assign_risk(status):
    if status == "COMPLETED":
        return "Low Risk"
    elif status in ["RECRUITING", "ACTIVE_NOT_RECRUITING",
                    "NOT_YET_RECRUITING", "ENROLLING_BY_INVITATION"]:
        return "Medium Risk"
    else:
        return "High Risk"
```

Appendix B: TF-IDF VECTORIZER CONFIGURATION

max_features : 30,000 *(top terms by corpus frequency)*

ngram_range : (1, 2) *(unigrams and bigrams)*

stop_words : 'english' *(standard stop-word removal)*

min_df : 5 *(terms in < 5 documents excluded)*

max_df : 0.9 *(terms in > 90% of documents excluded)*

Appendix C: TOP RISK-INDICATIVE TERMS BY CLASS (TF-IDF MODEL)

High Risk: expanded access, caused, recruited, 19 infection, china, healthy subjects, hbot, amputation, ability

Low Risk: purpose study, dsm iv, included, hypothesis, purpose, group received, iv, aimed, telephone, eyes

Medium Risk: aims, goal, goal clinical, 2023, main questions, aims answer, 2024, study aims, breastfeeding, gt