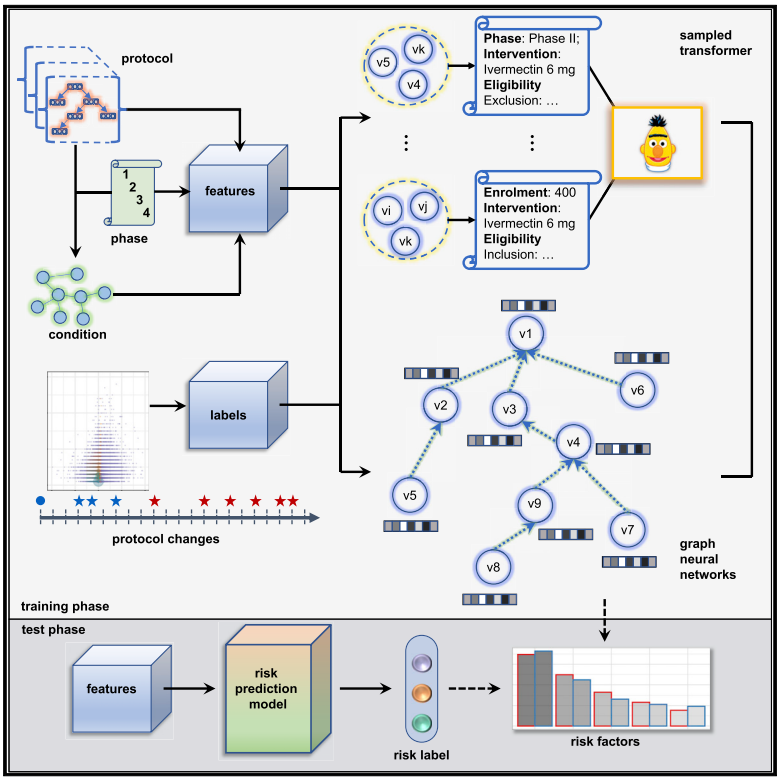
Article

Patterns

Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study

Graphical abstract



Highlights

- We perform a large-scale analysis of historical changes in CT protocols
- We propose a ternary risk assignment model based on major changes on CT protocols
- We design and evaluate transformer and GNN models for CT risk prediction
- We disclose a fine-grained benchmark dataset for CT risk prediction

Authors

Sohrab Ferdowsi, Julien Knafou, Nikolay Borissov, David Vicente Alvarez, Rahul Mishra, Poorya Amini, Douglas Teodoro

Correspondence

douglas.teodoro@unige.ch

In brief

Ferdowsi et al. propose a deep learning-based methodology to predict risk of clinical trials using the design protocol. Instead of relying on the termination status, they consider the history of major changes in the protocol to create a ternary risk label model. This approach enables fine-grained risk assessment to support risk mitigation strategies.

Ferdowsi et al., 2023, Patterns *4*, 100689
March 10, 2023
https://doi.org/10.1016/j.patter.2023.100689

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study

Sohrab Ferdowsi,[1,2] Julien Knafou,[2] Nikolay Borissov,[3,4] David Vicente Alvarez,[1,2] Rahul Mishra,[1] Poorya Amini,[3,4] and Douglas Teodoro[1,2,5,6,*]

[1]Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland
[2]Geneva School of Business Administration, HES-SO University of Applied Sciences and Arts of Western Switzerland, Geneva, Switzerland
[3]Clinical Trials Unit, University of Bern, Bern, Switzerland
[4]Risklick AG, Bern, Switzerland
[5]Swiss Institute of Bioinformatics, Lausanne, Switzerland
[6]Lead contact
*Correspondence: douglas.teodoro@unige.ch
https://doi.org/10.1016/j.patter.2023.100689

THE BIGGER PICTURE Many risk factors contributing to the low clinical trial (CT) success rate can be traced back to protocol design issues. Our study proposes using machine learning techniques that leverage large CT databases to predict protocol risks and help trial designers make more informed decisions about the design. Our work is unique in that it takes advantage of historical evolution of protocols to retrospectively derive risk-related metrics. The proposed prediction models based on deep learning architectures, such as transformers and graph neural networks, showed promising performance in predicting risk labels and can provide clues about risk aspects of the protocol related to CT failure. The goal is not only to provide protocol designers with retrospective insights but also to support risk-mitigation strategies to maximize CT effectiveness. Given their significant costs, any effective method of de-risking CTs can have a substantial benefit to the healthcare system and patient well-being.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Success rate of clinical trials (CTs) is low, with the protocol design itself being considered a major risk factor. We aimed to investigate the use of deep learning methods to predict the risk of CTs based on their protocols. Considering protocol changes and their final status, a retrospective risk assignment method was proposed to label CTs according to low, medium, and high risk levels. Then, transformer and graph neural networks were designed and combined in an ensemble model to learn to infer the ternary risk categories. The ensemble model achieved robust performance (area under the receiving operator characteristic curve [AUROC] of 0.8453 [95% confidence interval: 0.8409–0.8495]), similar to the individual architectures but significantly outperforming a baseline based on bag-of-words features (0.7548 [0.7493–0.7603] AUROC). We demonstrate the potential of deep learning in predicting the risk of CTs from their protocols, paving the way for customized risk mitigation strategies during protocol design.

INTRODUCTION

The fundamental way to assess the safety and efficacy of clinical interventions is to carry out clinical trials (CTs), i.e., multiple phases of randomized clinical studies on volunteer individuals with clear hypotheses to be tested.[1,2] For a candidate medication to reach the market, once the necessary basic drug discovery research and preclinical animal studies are successful, multiple phases of CTs are executed to prove the safety and efficacy of the intervention against a target study group. If successful

throughout all the phases, only then can a medication obtain market authorization from regulatory agencies. This process takes around 60%–70% of the average 13.8-year-long drug development cycle[3] and comprises a major portion[4] of the ever increasing cost of drug development, estimated to be around 1.3B$ on average.[5] Hence, CTs are among the major influencers of the final medication prices.[6]

Prior to each implementation phase, as required by regulations, such as the FDAAA 801[7] and Regulation (EU) no. 536/2014,[8] CTs are carefully planned and many of the execution details are precisely described in what are known as CT protocols.[9] In fact, these regulations are becoming stricter; for example, with requirements of risk assessment, and CTs may have to adapt to situations like the COVID-19 pandemic, where vaccine development timelines needed to be significantly shortened.[10] Despite careful planning, less than 14% of trials succeed in getting from phase I to final market approval by regulatory agencies.[11] Direct consequences of these failures include increased prices of medications when reaching patients, prolonged drug-to-market time, as well as other financial and ethical burdens associated with project failures incurred for pharma companies and research partners (see, e.g., an industry report[12] showing a decrease in R&D return for pharma companies).

Using retrospective analyses of CT databases, various reasons behind CT failures have been identified. While some reasons are associated with the clinical intervention under study, e.g., drug safety issues or lack of efficacy, a significant portion relates to CT logistics and execution details, such as insufficient participant enrollment, ineffective site selection, business or funding decisions, protocol issues, or lack of drug supply.[13–15] Specifically for some therapeutic groups, research shows that interventions in certain areas, such as oncology, are riskier than others.[16–18]

Thanks to the availability of large CT protocol collections, in the past years there have been several works in the literature reporting data-driven methods to assess risk factors based on CT protocols[19–21] and estimate risks prior to CT execution. Instead of resorting to manual inspection techniques, which often do not scale favorably with the complexity of the variables involved in CT risk analyses, these works propose automatic and reliable machine learning methods trained on historical CT registry data to predict whether a CT would complete its phase or terminate before achieving its objectives. Some of these works[19,22] use classic text mining approaches to identify keywords associated with CT termination, as well as random forests and latent Dirichlet allocation to perform risk classification. Other recent studies[20,21] pose the problem as a binary classification of CT protocols into completed and terminated categories and create a set of hand-crafted features that are fed to different off-the-shelf classifiers to predict the risk of phase success. They furthermore perform traditional feature selection and ranking strategies to identify top factors associated with CT termination. Similar methodologies have been used to carry out risk assessment for COVID-related CTs.[23]

Methods based on hand-crafted feature engineering often do not generalize well to textual data. As an alternative, deep learning,[24] that is, artificial neural network architectures with several learning layers, has emerged as a paradigm based on automatic extraction of useful representations from data through multiple stages of processing, and with successful examples across digital medicine.[25,26] As for the use of deep learning for CT risk assessment, recent studies[27–29] have explored the use geometric deep learning, a branch of deep learning used for graph-based data, to predict whether a CT would successfully complete a particular phase. By exploiting the hierarchical nature of the CT protocol document, graph-based models provide significant performance improvements compared with models encoding the protocol using non-hierarchical representations.

In this work, rather than solely relying on the reported final status of the CTs, that is, completed (low) or terminated (high), we incorporate historical CT statistics gathered from protocol updates, such as patient enrollment drop rate, study duration, and number of protocol amendments, to characterize the notion of risk. These statistics are calculated by mining a large collection containing historical evolution of CT protocols and use domain knowledge to propose a ternary risk-assignment approach: low, medium, and high risk. In contrast to previous works, this strategy is more aligned with established ways in the literature for risk assessment through manual methods.[11] Subsequently, we hypothesize whether these retrospectively assigned ternary risk labels can be automatically learned and thus show correspondence to their protocol risk. Benefitting from recent advances in deep learning architectures, we investigated various transformers-based language models[30,31] and graph neural network (GNN)[32] models trained to predict the proposed ternary risk categories. These models are combined into an ensemble using the probabilities of the individual models to obtain the final predictions. Interestingly, the ternary risk hypothesis is confirmed in our experiments with a robust performance of area under the receiving operator characteristic curve (AUROC) of 0.8453 (95% CI: 0.8409–0.8495).

Our main contributions can be summarized as follows:

- We perform a large-scale analysis of historical changes in CT protocols, focusing on enrollment, duration, outcome, and amendment aspects, and show their relations to CT risk levels.
- We propose a new risk assignment methodology using termination status as well as key protocol design features stratified by phase and condition.
- We design a new machine learning model for CT risk prediction, combining state-of-the-art transformer and GNN architectures to leverage both contextual word representation and hierarchical features of semi-structured documents.
- We evaluate our model against state-of-the-art baselines and show that it achieves superior performance on both binary and ternary risk prediction tasks. We also conduct an explainability study to demonstrate the capacity of the model to automatically identify protocol design aspects associated with risk factors.
- We create a new benchmark that goes beyond the classic binary classification, enabling the design and evaluation of finer-grained CT risk assessment methodologies.

RESULTS

Protocol versioning analyses

The divergence in months between the planned and actual CT duration obtained from the historical CT database is

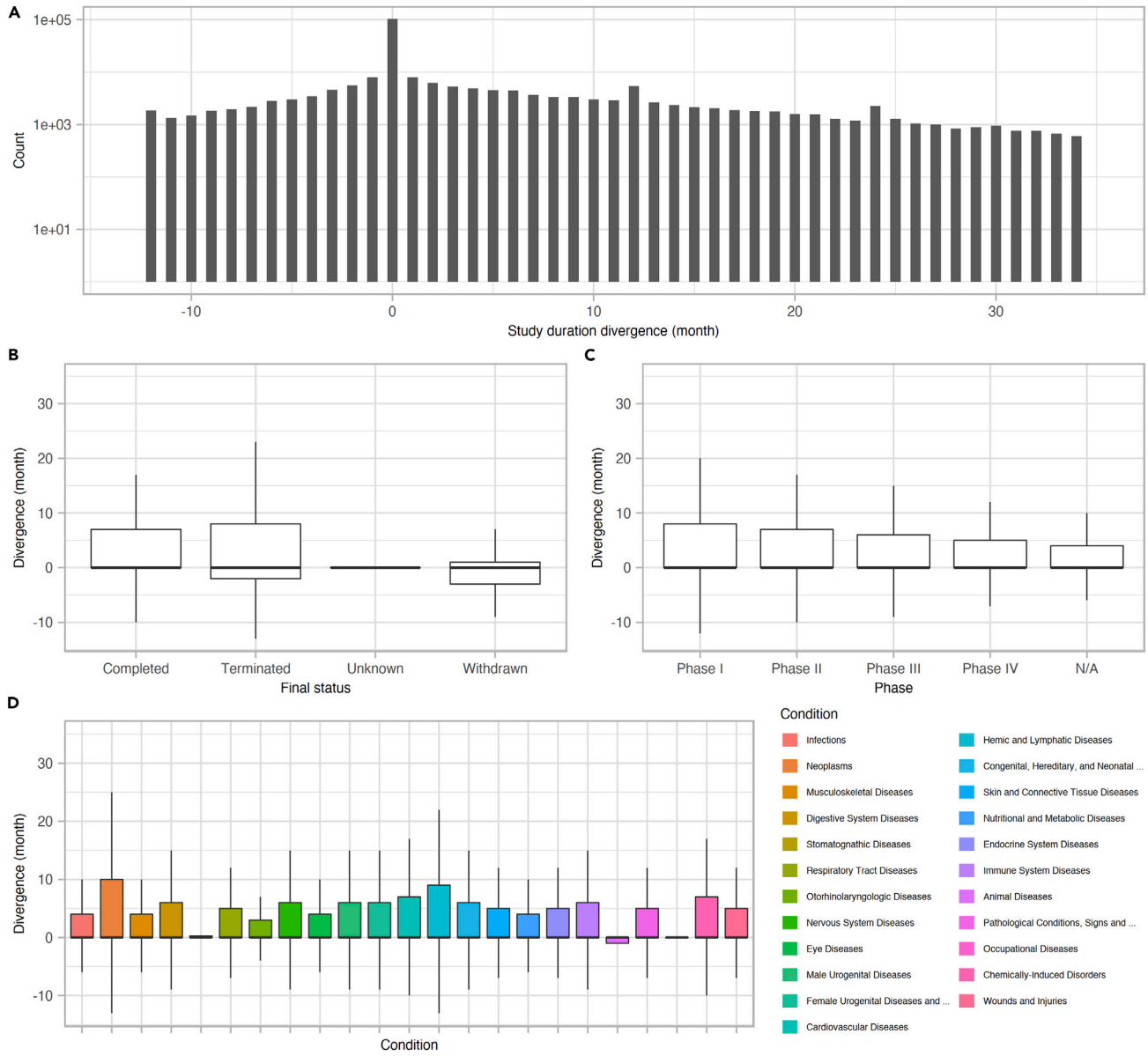


Figure 1. Divergence between planned and actual clinical trial duration

(A–D) Histogram of the divergence between planned and actual clinical trial duration (A), and stratification per overall status (B), phase (C), and clinical conditions (D).

shown in Figure 1. Results indicate that most CTs underestimate the study duration (Figure 1A), with a higher divergence spread in terminated CTs (SD = 20.9 and SD = 16.4 for non-completed and completed CTs, respectively, $p = 0.007$) (Figure 1B), and a more accurate estimation is seen as they progress across phases (phase I and II: SD = 19.7; phase III: SD = 18.7; and phase IV: SD = 14.6, $p < 0.001$) (Figure 1C). If we consider conditions composing at least 1% of the dataset, *Stomatognathic Diseases* CTs have the lowest divergence (SD = 11.5), while *Hemic and Lymphatic Diseases* as well as *Neoplasms* CTs underestimate the study duration by 6 months on average, taking 44.4 months (SD = 32.7 and SD = 31.8, respectively) on average to complete a phase (Figure 1D).

Figure 2 describes the number of major protocol amendments for different statuses, phases, and conditions. To provide a robust measure, we only consider changes to key protocol sections, namely *Arms and Interventions*, *Conditions*, *Groups and Interventions*, *Eligibility Criteria*, *Interventions*, *Outcome Measures*, *Sponsor and Collaborators*, *Study Design*, and *Study Status*. As expected, the terminated trials have the highest number of substantial protocol changes (Figure 2B) ($p < 0.001$). Premarket CT phases (i.e., phases I to III) present a similar distribution of major changes ($\mu = 3.3$, $p = 0.01$) (Figure 2C), with a

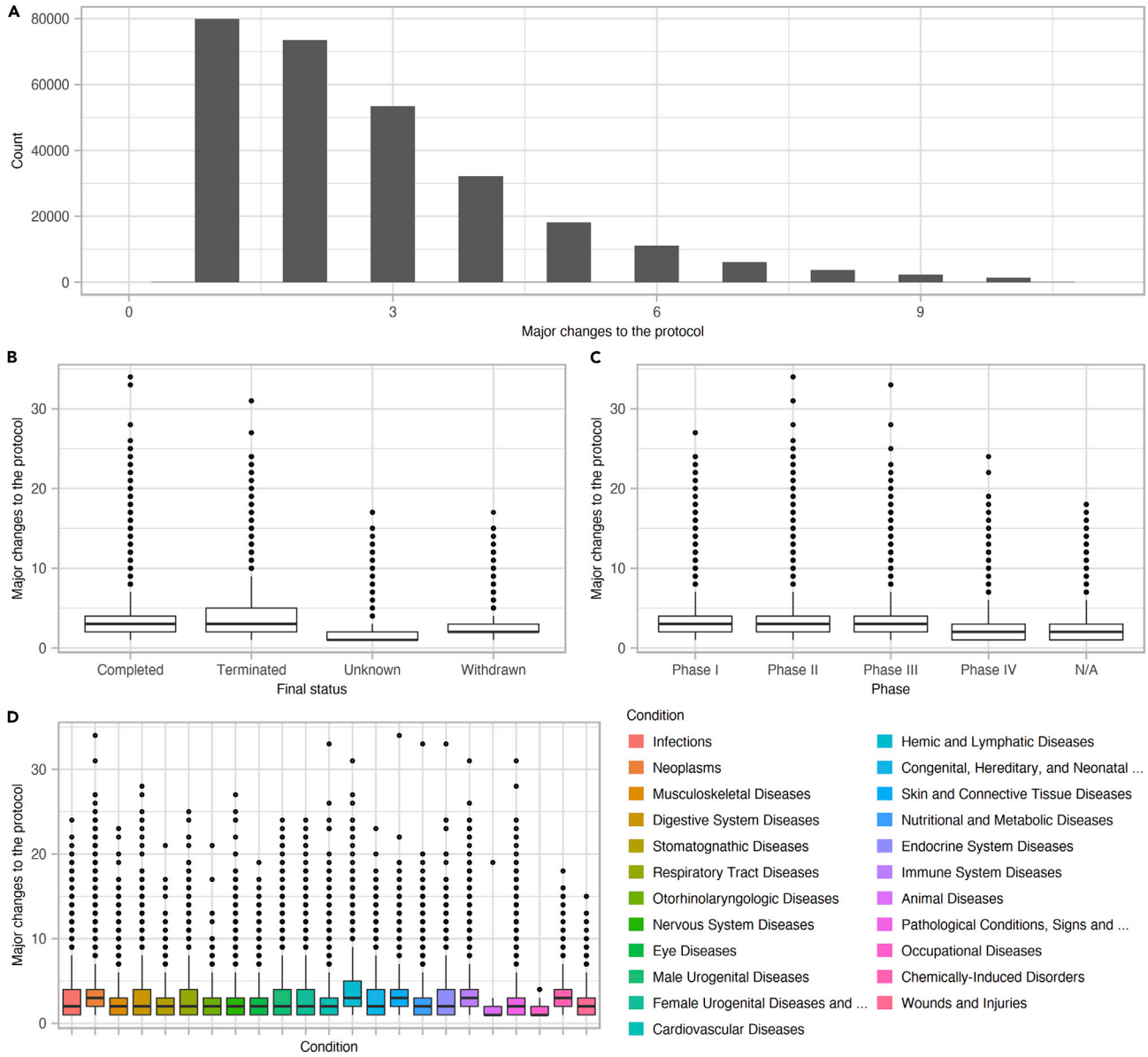


Figure 2. Major changes to the protocol

(A–D) Histogram of the number of major changes to the protocol (A), and stratification per overall status (B), phase (C), and clinical condition (D).

decrease in the number of changes for the post-market phase (phase IV) ($\mu = 2.5$, $p < 0.001$). Within the clinical condition categories (Figure 2D), the *Hemic and Lymphatic Diseases* have the highest frequency of major changes.

Results shows that terminated CTs enroll 52% of the planned number of subjects compared with 88% for completed CTs ($p < 0.001$) (Figure 3A). For some CTs, the enrollment is higher in the last version compared with the first. This does not correspond to risky behavior; therefore, in order not to bias the statistics, we upper-bound the ratio to one and consider only the dropping in enrollment. Similarly, as shown in Figure 3B, the number of outcome changes is relatively higher in terminated CTs (30%) when compared with the completed CTs (27%) ($p < 0.001$).

Retrospective multi-label risk assignment

After removing CT protocols for which the risk metrics could not be applied, e.g., due to missing or inconsistent values while computing the historical statistics, the number of unique CT protocols reduced to 135,940. Since the statistics were computed for unique phase-condition group pairs to better reflect the specificity of the trial, and to account for multiple phases (e.g., phases I and II) and condition groups (e.g., *Neoplasms* and *Hemic and Lymphatic Diseases*) that a unique CT protocol may refer to, the multi-label risks were assigned to CT-phase-condition triplets. This resulted in a dataset containing 283,776 unique CT-phase-condition triplets, for which the ternary risk model was applied.

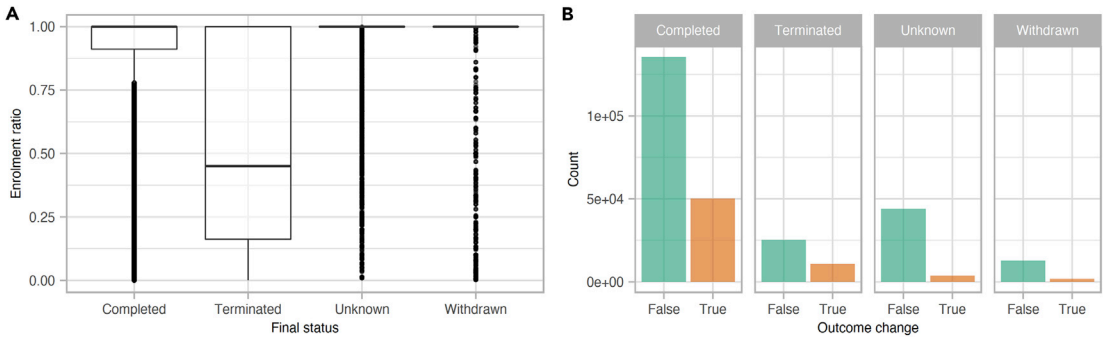


Figure 3. Enrollment ratio drop (A) and changes in outcome (B) per overall status

Figure 4 sketches the distribution of the computed ternary labels stratified by final status, CT phases, and clinical condition categories, which are distributed among 85,820 (30%) low-risk, 42,694 (15%) medium-risk, and 155,262 (55%) high-risk categories. As shown in Figure 4A, from the set of CTs that have the final status as completed, 43k triplets were labeled medium risk, while 57k triplets were labeled high risk (the remaining high risk comes from terminated, unknown, and withdraw statuses). While on average 30% of CT protocols are low risk, this ratio is the lowest among trials of phase II (around 25%), with trials on *Infections* (C01) and *Musculoskeletal Diseases* (C05) having the highest ratio of low-risk CTs (around 35%).

As an example of the historical changes that a CT protocol can go through during its execution, Figure 4D shows the timeline of changes for the ClinicalTrials: NCT01432886 study, from its creation and submission to the ClinicalTrials.gov repository to the study completion. The CT completed the phase (i.e., completed overall status); however, the protocol had seven major changes during its execution (10 in total). The study design section was modified twice, changing the enrollment number from 24 planned participants to 12 enrolled (50% drop) and the allocation type from "non-randomized" to "N/A." Moreover, the primary outcome changed from one primary outcome at the beginning of the study (evaluation of dose limiting toxicity [DLT]) to two outcome measurements in the final version (number of participants with DLT and number of participants with adverse events). Finally, the study was expected to complete in 12 months, but it lasted for 27 months (2.25-fold increase in duration). Thus, instead of considering this CT protocol as low risk based only on its final status, the study was labeled high risk according to the ternary risk model.

Analyses of the distribution of risk categories according to study duration, major changes, and enrollment ratio risk factors (Figure 5) show that there are no clear boundaries between risk categories. Medium-risk protocols are particularly scattered around low- and high-risk protocols, suggesting a coherent risk continuum across the three categories. Nevertheless, some clusters can still be identified around the median values of the risk factor metrics analyzed (low-risk CT protocols) or far from them (high-risk CT protocols).

CT risk prediction performance

We investigated three deep learning models to predict the risk of a CT based on its protocol: a graph model—Graph-BOW and Graph-LM; a transformer-based language model—Sampled-LM; and an ensemble combining results of the former two models. For the graph-based model, two strategies were used to encode the protocol leaves: bag-of-words (Graph-BOW) and word embeddings provided by the PubMedBERT language model[33] (Graph-LM). The proposed models were compared with a multi-layer perceptron baseline using bag-of-words (MLP-BOW) or pre-trained language model embeddings (MLP-LM) to encode features of the protocol sections (design, criteria, condition, intervention, etc.). In this setting, the representation of a CT protocol is obtained by the feature concatenation of the k sections (k = 15 in our settings). In our experiments, the models were trained using the dataset with ternary codes assigned to the triplets CT-phase-condition.

The prediction performance for the multi-label risk approach is presented Table 1. As we can see, both graph and language models outperform the MLP baseline, achieving AUROC of nearly 84% for both Graph-LM (0.8395 [95% CI: 0.8352–0.8439]) and Sampled-LM (0.8363 [95% CI: 0.8318–0.8407]). The ensemble model further improves the individual models in terms of AUROC, achieving a performance of around 85% (0.8453 [95% CI: 0.8409–0.8495]); however, this improvement is not statistically significant compared with the Graph-LM model. In terms of accuracy, the best results are provided by the Sampled-LM model (around 68%), which can correctly predict the CT protocol risk in more than two out of three attempts on average, while recall is maximum (around 62%) for the Graph-LM model. For the macro F1 score, which reflects how well the models can predict the individual risk classes, the ensemble achieves the highest results, with a performance of around 61%.

To have a comparative reference with state-of-the-art approaches, which uses only the completed and terminated status of a CT to assess risk, we also experimented with a binary labeling strategy. In our case, the CTs within the low-risk class had a completed status, while the high-risk class was composed of any of the terminated, withdrawn, or unknown statuses. The risk prediction performance was obtained in an out-of-sample dataset composed of 31,684 CT protocols (15%), and the remaining 178,813 protocols (85%) were used to train our models (70% train and 15% develop). These results are reported in Table 2, where we also show the performance of Elkin and Zhu[21] and Fu et al.,[29] which were carried out within a similar phase success setup but using different training and evaluation CT protocol datasets extracted from ClinicalTrials.gov. It is

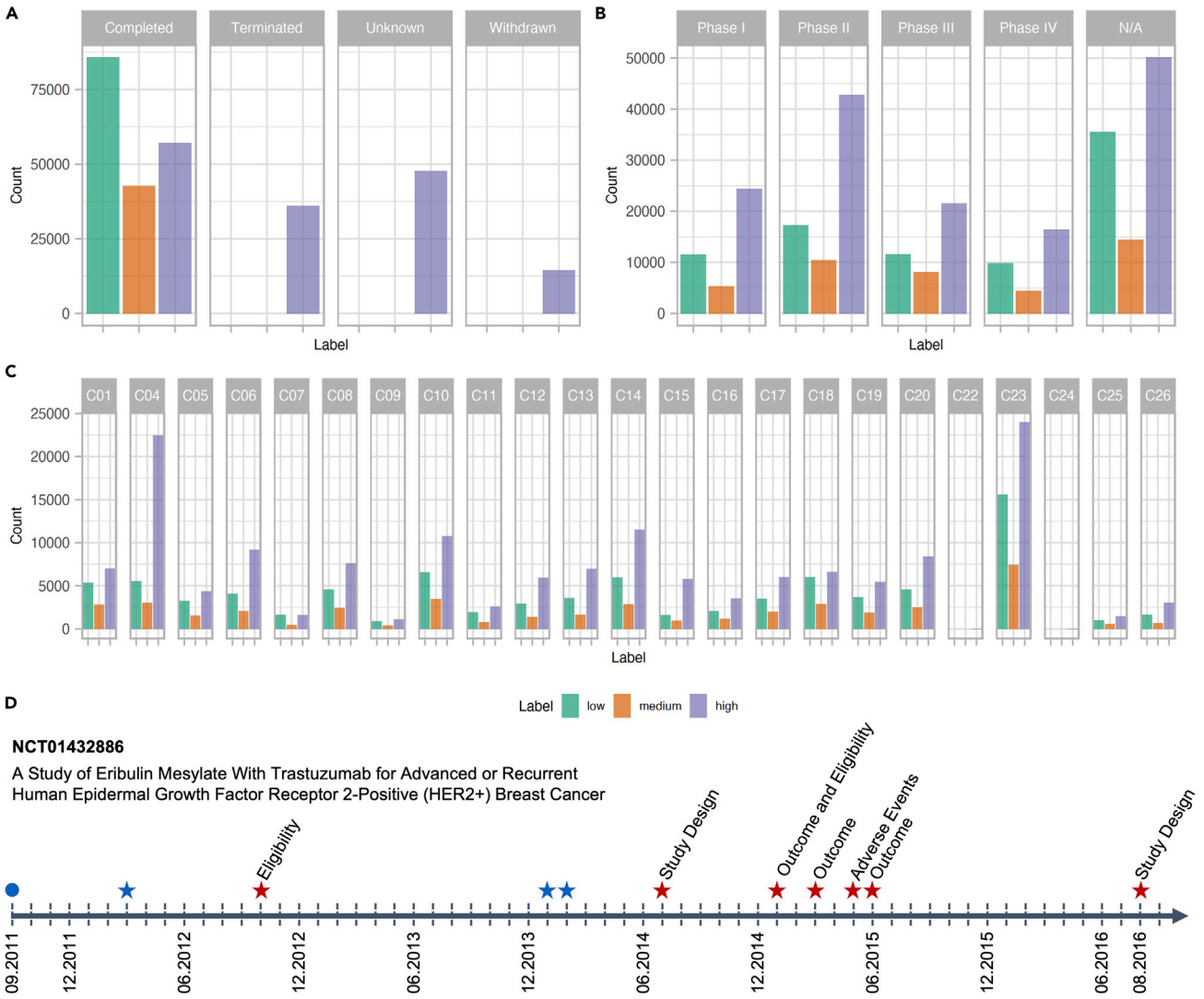


Figure 4. Distribution of ternary risk labels retrospectively assigned to CTs

(A–D) Distribution per overall status (A), phase (B), and clinical condition (C), and an example of historical changes for a CT protocol (D).

important to note that, in addition to a smaller training set, these previous experiments also used a smaller evaluation set compared with our setting. Therefore, they are not directly comparable.

Results show that the ensemble model achieves excellent class discrimination performance in the binary labeling strategy with an AUROC of 92% (0.9234 95% [CI: 0.9193–0.9274]) and an accuracy of almost 90%, outperforming our individual models and baseline. Comparing the individual models, the Sampled-LM achieves the highest performance of the aggregated metrics with a decision cutoff (84% macro F1 score and 87% accuracy).

Performance analyses

We further analyze the ternary predictive performance per risk type, phase, and condition categories using the results of best individual model according to the AUROC metric (i.e., Graph-LM). Overall, the model provides robust performance across risk type, phase, and condition (Figure 6), being above an AUROC of 78% for all strata, despite CT dynamics being significantly distinct for phases and conditions.[16–18] Nonetheless, we see a 9% drop in performance for the medium-risk class (0.7769 [95% CI: 0.7716–0.7824]) compared with the high-risk class (0.8691 [95% CI: 0.8658–0.8724]) (Figure 6A). Note that correctly predicting high-risk CTs is the most important for risk mitigation purposes. The performance is consistent phase-wise (Figure 6B), with an AUROC difference of only 2% between the best performing phase—phase I (0.8493 [95% CI: 0.8408–0.8573])—and the worst performing phase category—"N/A" (0.8327 [95% CI: 0.8274–0.8379]). Similarly, among the conditions with at least 1,000 samples in our test set (Figure 6C) we see an AUROC difference of only 5% between the best performing—*Hemic and Lymphatic Diseases* (0.8647 [95% CI: 0.8452–0.8841])—and the worst performing—*Nutritional and Metabolic Diseases* (0.8146 [95% CI: 0.8009–0.8277]). The results show that these

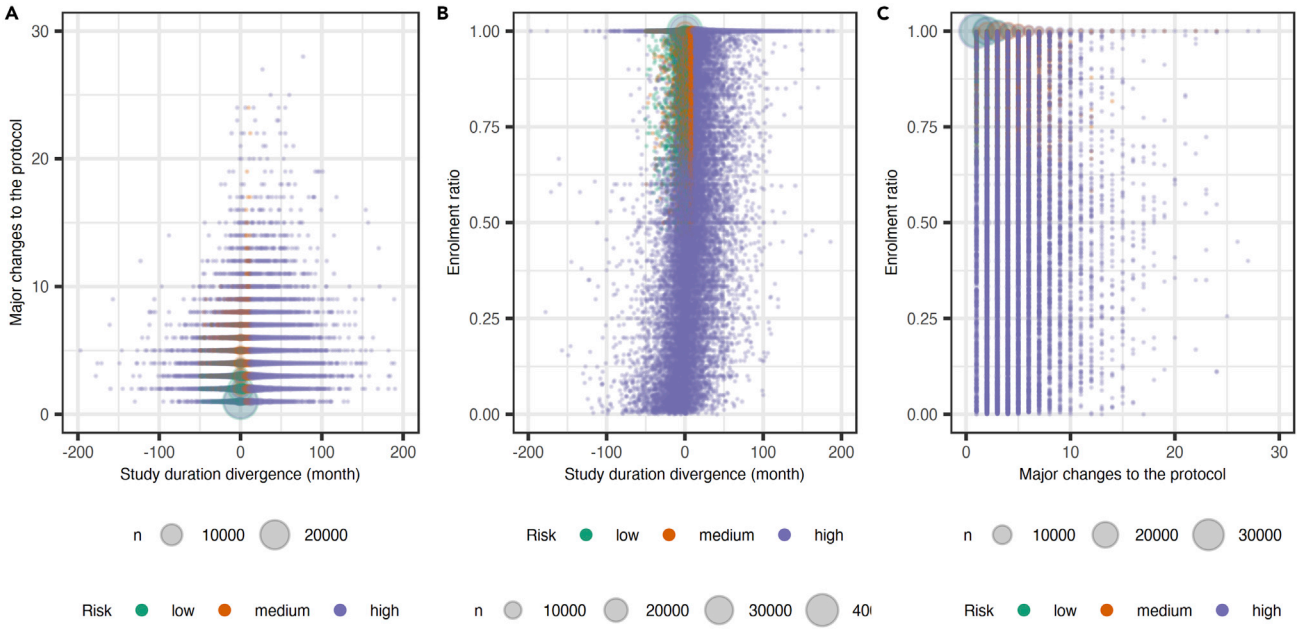


Figure 5. Distribution of risk according to the study duration, major changes, and enrollment ratio risk factors

models can be used consistently to assess CT for different trial phases and conditions. Finally, looking at the confusion matrix (Figure 6D), we notice that the model tends to confuse medium-risk with low- and high-risk instances. This is expected due to the continuity of the risk factors and the fact that there are no clear boundaries between risk categories (see Figure 5).

Model explainability and risk factor identification

To identify the parts of the CT protocols that the models associate with the high-risk label in their prediction, we use the integrated gradients[34] method, which has been successfully applied to explain factors associated with inference of deep learning models. For CTs classified by our graph-based method as high risk, we use this method to calculate the output-input activities for all CT nodes and consider those with the highest norm of activity across dimensions as possible risky fields. For better explainability, we focus on our graph-based model, which explicitly preserves the initial trial protocol sections. While we initially calculate the association of each single node from the input CT to the final predicted risk label one-by-one, and use the integrated gradients method to make results more robust and improve explainability, we pool the individual nodes to the main parent nodes (or sections) in the CT protocol hierarchy, i.e., any of *Arms and Interventions*, *Conditions*, *Contacts and Locations*, *Eligibility Criteria*, *Outcome Measures*, *Oversight*, *Sponsor and Collaborators*, or *Study Design*.

To validate this idea, we took a random set of correctly predicted high-risk CTs using our graph-based model for *Breast Cancer* (n = 28) and *Cardiovascular* (n = 36) conditions, and averaged the aggregated activities calculated from the integrated gradients method for the CT sections. The results are shown in Figure 7, where we note that the top 2 risky CT protocol sections for these analyzed conditions are *Study Design* (32% for breast cancer and 43% for cardiovascular diseases) and *Contacts and Locations* (31% for breast cancer and 29% for cardiovascular diseases). Note that these two fields are where the information regarding patient enrollment phase as well as site selection are detailed, which are indeed identified by the model as part of the top 5 nodes contributing most to the *Study Design* and *Contacts and Locations* risks. For example, according to the model *Enrollment*, *Count* contributes to 35% and 36% of the *Study Design* risk for the breast cancer and cardiovascular conditions, respectively (Figure 7B), while the top 5 location-related nodes, such as facility, institution

Table 1. Performance of the CT risk prediction models using the ternary risk labeling methodology

Model	Precision	Recall	F1 score	Accuracy	AUROC (95% CI)
MLP-BOW	0.5064	0.5068	0.5060	0.5869	0.7548 (0.7493–0.7603)
MLP-LM	0.5804	0.5934	0.5723	0.6173	0.8216 (0.8171–0.8262)
Graph-BOW	0.5501	0.5603	0.5403	0.5834	0.7936 (0.7886–0.7988)
Graph-LM	0.6039	**0.6215**	0.5913	0.6283	0.8395 (0.8352–0.8439)
Sampled-LM	0.6017	0.5855	0.5858	**0.6759**	0.8363 (0.8318–0.8407)
Ensemble	**0.6047**	0.6152	**0.6065**	0.6677	**0.8453** (0.8409-0.8495)

Precision, recall, and F1 score are reported using macro-average statistics.

Table 2. Performance of the CT risk prediction models using the binary risk labeling methodology

Model	Precision	Recall	F1 score	Accuracy	AUROC (95% CI)
Elkin and Zhu[21]	–	–	–	–	0.7281
Fu et al.[29]	–	–	–	0.837	0.817 (0.802–0.832)
MLP-BOW	0.8056	0.8148	0.8099	0.8449	0.8792 (0.8735–0.8846)
MLP-LM	0.8143	0.8292	0.8211	0.8527	0.9063 (0.9016–0.9109)
Graph-BOW	0.7907	0.8082	0.7983	0.8328	0.8865 (0.8813–0.8914)
Graph-LM	0.8400	**0.8410**	0.8405	0.8716	0.9161 (0.9118–0.9203)
Sampled-LM	**0.8956**	0.8203	0.8476	0.8883	0.9120 (0.9077–0.9163)
Ensemble	0.8913	0.8359	**0.8577**	**0.8933**	**0.9234** (0.9193–0.9274)

Precision, recall, and F1 score are reported using macro-average statistics.

(affiliation), and zip code, account together for 70% and 71% of the risk in the *Contacts and Locations* section for the breast cancer and cardiovascular conditions, respectively. This is in line with manual CT risk analysis literature, in which patient attrition and poor selection of study sites are identified as the main reasons behind trial failure.[13–15]

DISCUSSION

CT literature identifies various reasons behind CT failure, including issues related to the safety or efficacy of the medical interventions under study, but also related to logistics and protocol design. The intricate combination of numerous factors potentially prone to failure, such as clinical conditions, study sites, and enrollment criteria, and their analyses using large datasets of historical CT protocols, makes manual inspection infeasible. As an alternative, utilization of systematic machine learning approaches that benefit from large databases of past records have been proposed. Our study benefit data from historical CT evolution to retrospectively derive risk-related measures that go beyond the binary approach and to combine transformer- and graph-based methodologies to predict risk labels from CT protocols. Our models are stratified based on clinical conditions and trial phases and show promising risk predictive performances for CTs.

Our approach carries some similarities to previous studies since it uses neural networks for the CT protocol classification

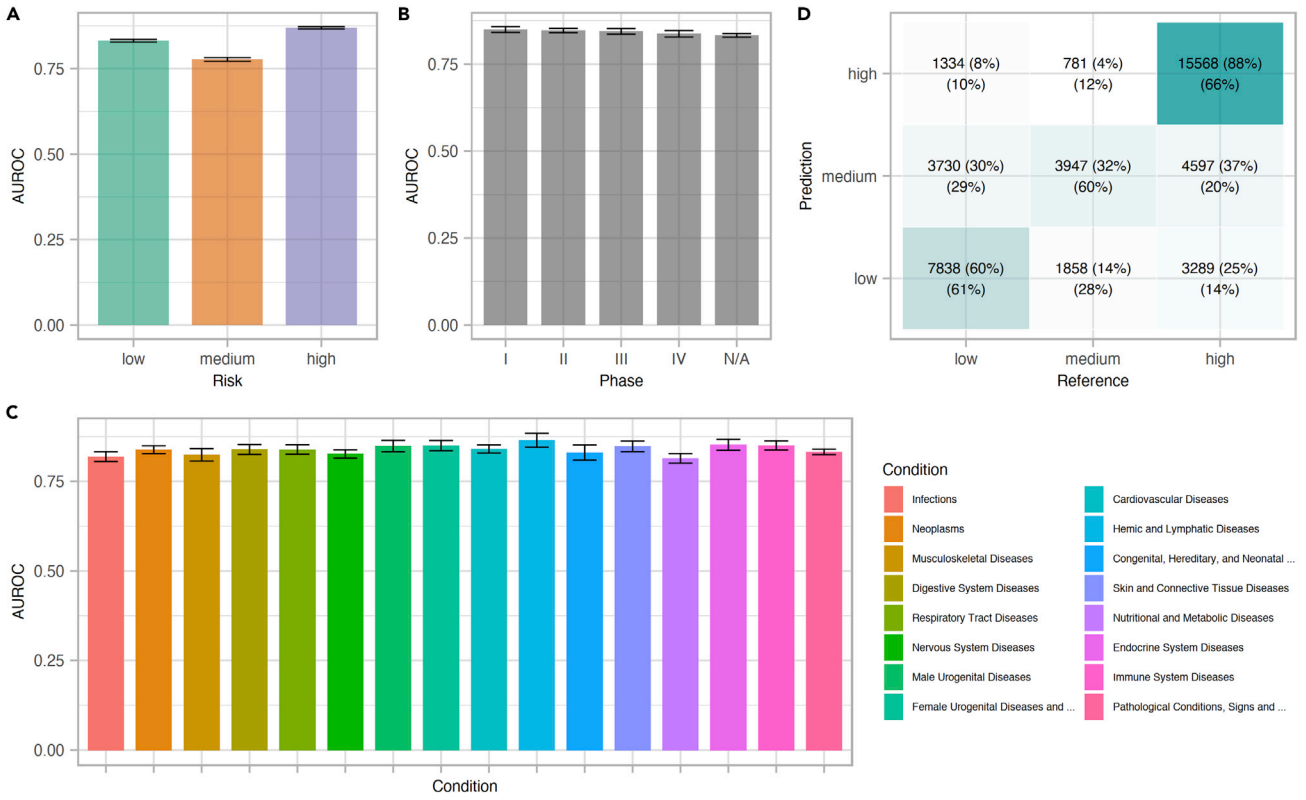


Figure 6. Classification results using the Graph-LM model

(A–D) Classification results per risk type (A), phase (B), and condition (C), and the confusion matrix (D).

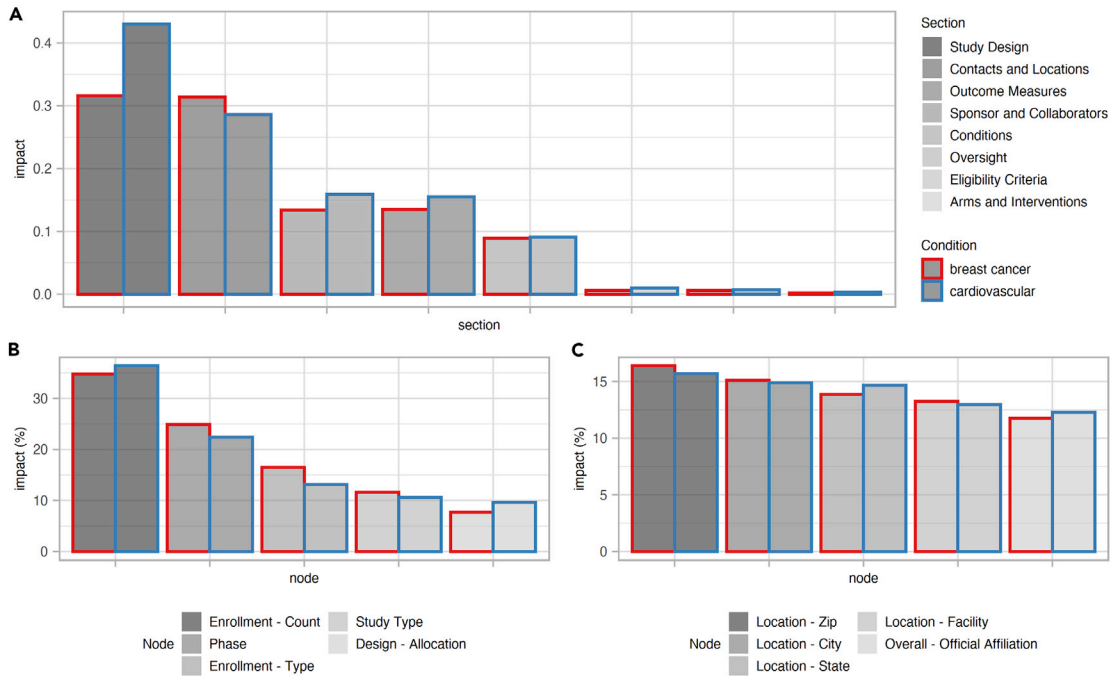


Figure 7. Results interpretability using integrated gradientes and the Graph-LM model

(A–C) Risk factors for high-risk CT protocols for breast cancer and cardiovascular conditions: impact of protocol sections on the risk prediction (A), top 5 nodes within the study design section (B), and top 5 nodes within the contacts and locations section (C).

task in addition to the use of embeddings to represent textual features as well as by including the disease hierarchy in the feature set. Elkin and Zhu[21] used feature engineering combined with word embeddings and assessed their impact using different off-the-shelf classifiers (logistic regression, random forest, etc.). Differently, Fu et al.[29] use multimodal data features, integrating CT protocols with a drug knowledge base, which are fed to an extended graph convolutional network architecture for phase success prediction. Our CT classification architecture differentiates mostly from previous attempts by the fact that it treats the protocol itself as graph, and features are extracted from all leaves independently, both by the graph- and transformer-based approaches. Thus, the model is less prone to information collapse (notice how the graph- and transformed-based models improve upon the MLP model). Moreover, by combining the graph- and transformer-based models in the ensemble, both the hierarchical protocol information and the word contextual representations are captured, which could lead to improved performance (as shown by the ensemble results).

Identifying the common causes behind CT failures is a first step toward mitigating their risk and increasing the odds of their success, hence reducing the expected time of drug-to-market, and potentially reducing medication prices. Our retrospective analyses studied common trial failure reasons similarly to what was previously highlighted by case-based studies in the CT literature. However, we go a step further by using a big data approach under various clinical condition categories and trial phases, and by benefiting from historical versions of CTs available from the ClinicalTrials.gov registry. We show that terminated trials tend to misestimate their duration, enroll fewer subjects than planned, but also to have more major and outcome changes compared with completed trials. These results enable the creation of fine-grained risk models as proposed here, which allows for more detailed analyses of past trials and for targeted predictive risk models.

Another important step toward optimization of trials is to be able to predict the outcome of a given trial study directly from its design protocol. This per-case prediction can potentially be more useful than solely relying on average-behavior studies, since customized optimization would be possible before the CT execution. Relying on databases of protocols along with their retrospective outcomes, machine learning can provide an appropriate framework to capture structures within protocols and map between protocol-outcome pairs in a way that can generalize to newly designed CT studies. Among the various machine learning paradigms, deep learning techniques are specifically suitable for large-scale databases and have recently shown significant performance improvements across many domains including healthcare and medicine. We utilize transformer-based language models and GNNs to adapt to the hierarchical structure of trial protocols with free text and contextualized biomedical concepts. With careful study design, statistical analyses, and performance measurements, we demonstrate that deep learning provides a robust framework to successfully predict trial risks in different stratification strategies.

Trial designers can benefit from deep learning-based risk prediction models based on the protocol in various ways. One such way is to keep changing the protocol components (e.g., the eligibility criteria, the study locations, etc.) of a predicted high-risk trial until the model predicts a low-risk label. Another way would be to perform nearest neighbor search, not on the protocol text, but rather on the latent space learned by the model.[35–37] The designers can then find similar past trials with their known

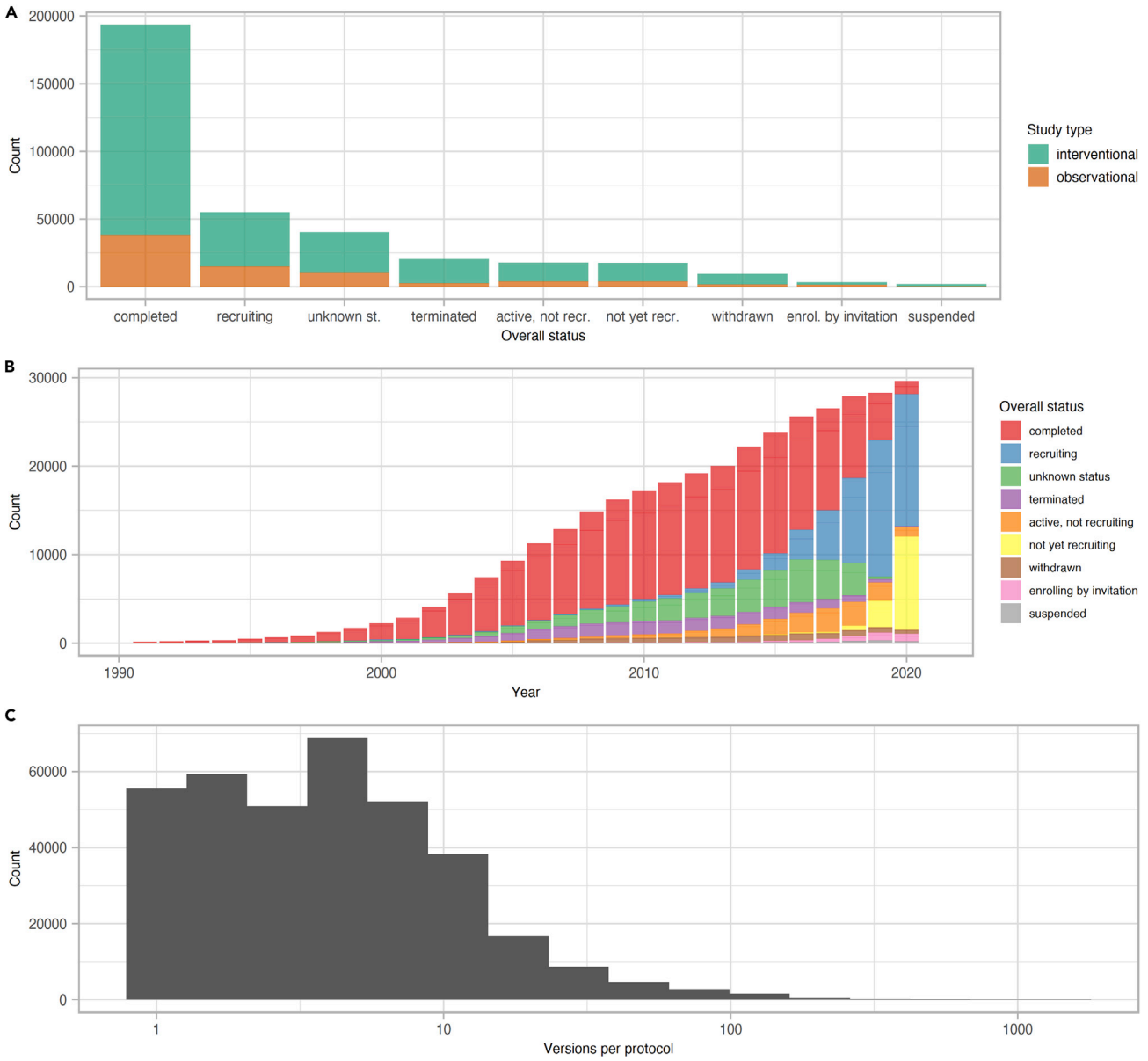


Figure 8. Dataset description statistics

(A–C) Overall status distribution for interventional and observational CTs of the ClinicalTrials.gov registry (A), starting year of interventional CTs from 1990 to 2020 stratified by overall status (B), and histogram of the number of versions per protocol (C).

outcomes based on some criteria that the model has learned. This is more in-depth than keyword-based matching, since the latent space of the model is implicitly enriched by the patterns of all the examples it had been trained on. We leave this direction of research for a future investigation.

To further enhance trial optimization, as a complementary step it would be desirable to explicitly find certain patterns within a given trial protocol that could potentially lead to an eventual risky behavior. While deep learning models are highly successful in incorporating large-scale data and automatically finding relevant patterns, their architectures are essentially backbox functions with powerful input-output mapping capabilities. In other words, while they can map their inputs to their outputs very successfully, it is usually not clear how they reach their decisions. Significant progress, however, has been made in this regard thanks to the areas of explainable machine learning.[38,39] In our studies, we used the popular integrated gradients approach with our graph-based model, where we assign an importance weight coefficient to nodes (or sections) of the protocol graph, i.e., the individual components of the trial protocol. As a result, apart from being able to predict the outcome of a study, the trial designers can have an estimation of the individual sections within their designed protocol that are more likely to be risk-inducing. While it is hard to quantitatively evaluate explainability results in healthcare,[40]

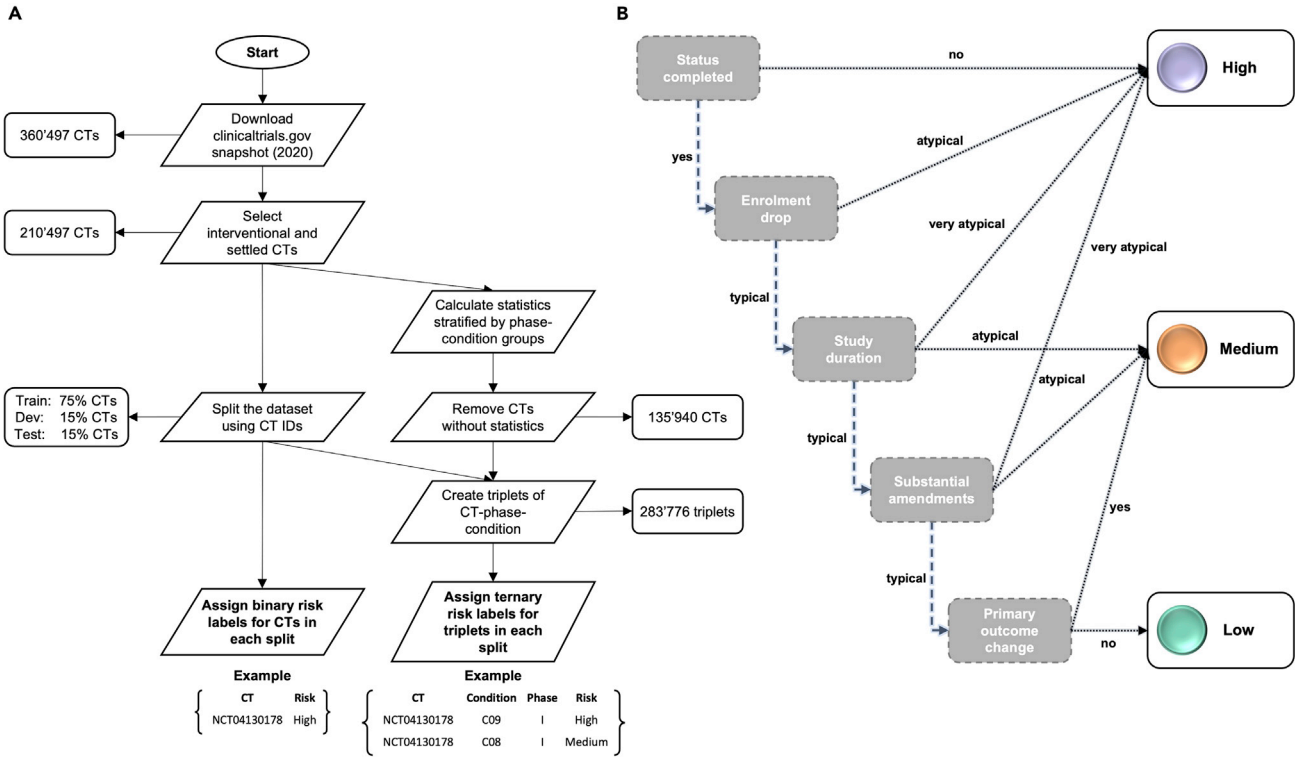


Figure 9. Dataset creation and risk labeling flowcharts

(A and B) Flow chart for the benchmark dataset creation (A) and the procedural approach to assign ternary risk labels to CT protocols based on statistics from the history-set as well as the final status reported in the protocols-set (B). Atypical values refer to statistics that are more than 1 standard deviation SD above the average within the respective phase and condition strata. Very atypical values refer to those that are beyond 2 standard deviations from the average.

preliminary qualitative analyses show that the risk factors identified by our models are supported by previous findings in the CT risk analysis literature.

While we made important strides in incorporating the history of CTs in our analyses, an important limitation of our study is that we do not consider patient-based or drug-related information. While, in general, it would be very difficult to find records of patients recruited and potentially dropped from a sufficiently large number of CT studies, several drug-related databases are publicly available that could be incorporated into our study to enrich the results. As an example, reports of adverse events for some medications are available that can be used along with the clinical conditions section of protocols to make them more specific. Moreover, the chemical structure itself could be incorporated as another node of the intervention section of the CT graph and be exploited by the model for safety and efficacy related risks. Indeed, our graph-based paradigm can be an ideal approach to incorporate external metadata to the protocols. We leave this direction for our future studies.

Table 3. Stratification of the of the CT protocols by phase

Phase	No.	%
I	41,175	20
II	70,496	34
III	41,228	20
IV	30,709	15
N/A	100,167	48

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

The lead contact for this work is Douglas Teodoro (douglas.teodoro@unige.ch).

Materials availability

This study did not generate any physical materials.

Data availability

The datasets and code used in this work are available at https://doi.org/10.5281/zenodo.7509571.

CT protocol dataset

The US-based ClinicalTrials.gov registry (publicly available at https://clinicaltrials.gov/) is a major source of trials information with daily updates and currently contains more than 390k registered CT items. In this work, we use a snapshot collection of ClinicalTrials.gov protocols with trials registered up to 2020. Figure 8 sketches the different status types (Figure 8A) of these CTs for both interventional and observational studies, as well as their starting year (Figure 8B), where an important increase in the number of launched CTs per year is observed for the past 20 years. We limit our study to interventional CTs, and consider only those with settled status, i.e., either completed, terminated, unknown, or withdrawn, resulting in a subset of 210,497 CTs out of the 360,497 downloaded. In addition to the latest version of the CT protocol, ClinicalTrials.gov also stores the historical protocol evolution from the day they were first submitted until their last update. Since these data are not readily available for download and there does not exist an API to access them, after the authorization of the registry administrator, we use web scraping

Table 4. Stratification of the of the CT protocols by condition

MeSH code	Condition group	No.	%
C01	infections	15,132	7.20
C04	neoplasms	30,992	14.75
C05	musculoskeletal diseases	9,111	4.34
C06	digestive system diseases	15,246	7.26
C07	stomatognathic disease	3,681	1.75
C08	respiratory tract diseases	14,584	6.94
C09	otorhinolaryngologic diseases	2,392	1.14
C10	nervous system diseases	20,799	9.90
C11	eye diseases	5,301	2.52
C12	male urogenital diseases	10,196	4.85
C13	female urogenital diseases and pregnancy complications	12,155	5.78
C14	cardiovascular diseases	20,283	9.65
C15	hemic and lymphatic diseases	8,316	3.96
C16	congenital, hereditary, and neonatal diseases and abnormalities	6,731	3.20
C17	skin and connective tissue diseases	11,493	5.47
C18	nutritional and metabolic diseases	15,515	7.38
C19	endocrine system diseases	11,034	5.25
C20	immune system diseases	15,467	7.36
C22	animal diseases	25	0.01
C23	pathological conditions, signs and symptoms	46,974	22.36
C24	occupational diseases	24	0.01
C25	chemically induced disorders	3,008	1.43
C26	wounds and injuries	5,316	2.53

Trials with unknown phase, or those without a particular phase assigned to them, are represented as N/A.

techniques to retrieve all protocol versions and engineered them to a format suitable for our studies. We call this new dataset the history-set to distinguish it from the main protocols-set, which contains only the latest protocol version and is readily available for download. The number of versions in the history-set vary significantly among the different CT protocols, most of them having between 1 and 10 versions, while a still important number has between 10 and 100 versions (Figure 8C). The flow chart for the benchmark dataset creation is depicted in Figure 9A.

Multi-label risk assignment strategy

For CTs that have finished their execution, the main protocols-set available from ClinicalTrials.gov provides only the final status (completed, terminated, withdrawn, etc.). This status is what previous risk prediction studies use[20,21,27,29] based on which CTs they label into two classes—completed or terminated—or minor variations therein, e.g., low or high risk. While we also experiment with this labeling procedure for comparison purposes, to better reflect realistic risk scenarios, we propose a more detailed procedure based on the history-set to retrospectively assign labels to CTs into different risk categories.

Supported by retrospective risk analysis studies,[13–18] we identified five factors that could be used to fine-grain risk levels and at the same time be derived from the protocol and history sets: (1) the final status of the CT protocol, (2) the attrition rate of recruited patients, (3) the divergence between the planned and actual study durations, (4) the number of major protocol amendments, and (5) major changes in the primary outcomes. We calculate the values for items (2) to (5) using the history-set and item (1) is provided by the respective protocol status in the protocols-set. To make our risk analyses more specific, we stratify the CT protocols according to their study phases (consisting of four phases plus one extra N/A category to account for trials without phase) as well as their condition groups, and we compute their stratified statistics. Within the ClinicalTrials.gov registry, a CT protocol is associated with one or multiple medical conditions and is annotated with Medical Subject Headings concepts for the respective study conditions. Using the MeSH ontology tree, we identify the condition group for each investigated disease, which is used during statistics computation. For each CT, we compare its values to these statistics to decide whether it deviates substantially (1 or 2 standard deviations) from other CTs in the same phase-condition category. Tables 3 and 4 show the distribution of CT protocols according to their phase and condition group for the study dataset, respectively.

As shown in the procedural approach of Figure 9B, we finally use these five factors together with the computed statistics and a set of rules to assign either low-, medium-, or high-risk labels to each CT. Overall, if trials fail to complete, fail to enroll enough participants, take much longer than planned, or have too many amendments, they are considered as high risk. Otherwise, if they take relatively longer than expected, have many amendments, or changes in the primary outcome, they are considered as medium risk; otherwise, as low risk.

CT risk prediction models

Figure 10 sketches the general schema of the predictive risk assignment model. Given a CT protocol, a phase, and a condition (group), after a basic pre-processing step, these features are fed to two machine learning models based on the transformer[30,31] (Figure 10B) and GNN[27,28,32] (Figure 10C) architectures. During the training phase, these models learn the CT protocol representation conditioned to the phase and clinical condition, and at the inference phase they predict the CT risk. The results of the transformer- and GNN-based models are merged using an ensemble model (Figure 10A), which computes the average of the probabilities provided by the individual models to obtain the final risk label. In the following sections we describe the transformer- and GNN-based models.

Sampled language model for CT risk prediction

While traditional approaches for text classification represent the text in vectorial representations (e.g., term frequency-inverse document frequency, word2vec,[41] etc.) and classify them based on off-the-shelf classifiers (e.g., logistic regression, support vector machine, etc.), the state-of-the-art approach usually with large performance gaps is the family of transformer-based[30] pre-trained language models.[31] The idea behind such models is to pretrain them on massive amounts of text, so that they learn contextual word vector representations but also syntactic and semantic relations[42] thanks to their highly expressive architecture. Such models can then be further trained (or fine-tuned) on a downstream task, e.g., text classification as in our case, while benefiting from all they had already learnt in their pre-training phase.

In our algorithm, we pre-trained a BERT-like[31] language model on the PubMed corpus and fine-tuned the resulting model to classify CT protocols according to the risk classes. The language model architecture has 6 transformer layers with 8 attention heads each and, like other BERT-based models, an embedding dimension of 512 tokens. To overcome the sequence length limitation of 512 tokens, which is usually surpassed by the length of the CT protocols, as shown in Figure 10B the pre-trained language model computes vectorial representations for n different text inputs sampled from the CT protocol leaves (n = 8 in our experiments). The n samples are drawn considering the leaf token size to reduce sampling bias, in which the larger the field length, the higher its sampling probability, and the higher its sampling frequency, the lower its probability of being redrawn. Subsequently, the sampled representations are passed through a mean pooling layer to obtain a CT-vector representation, which is then fed to a fully connected layer that predicts the risk label probabilities. This process is parallelly computed N times (N = 100 in our experiments) and the final CT risk label is computed by taking the label with the highest mean probability across the N predictions.

Graph neural network model for CT risk prediction

As an alternative method to the sampled-based language model, we use a variant of deep learning suitable to handle graph-based data, notably GNNs. In fact, the structure of a CT protocol, unlike traditional textual documents, is semi-structured with many hierarchical dependencies between different

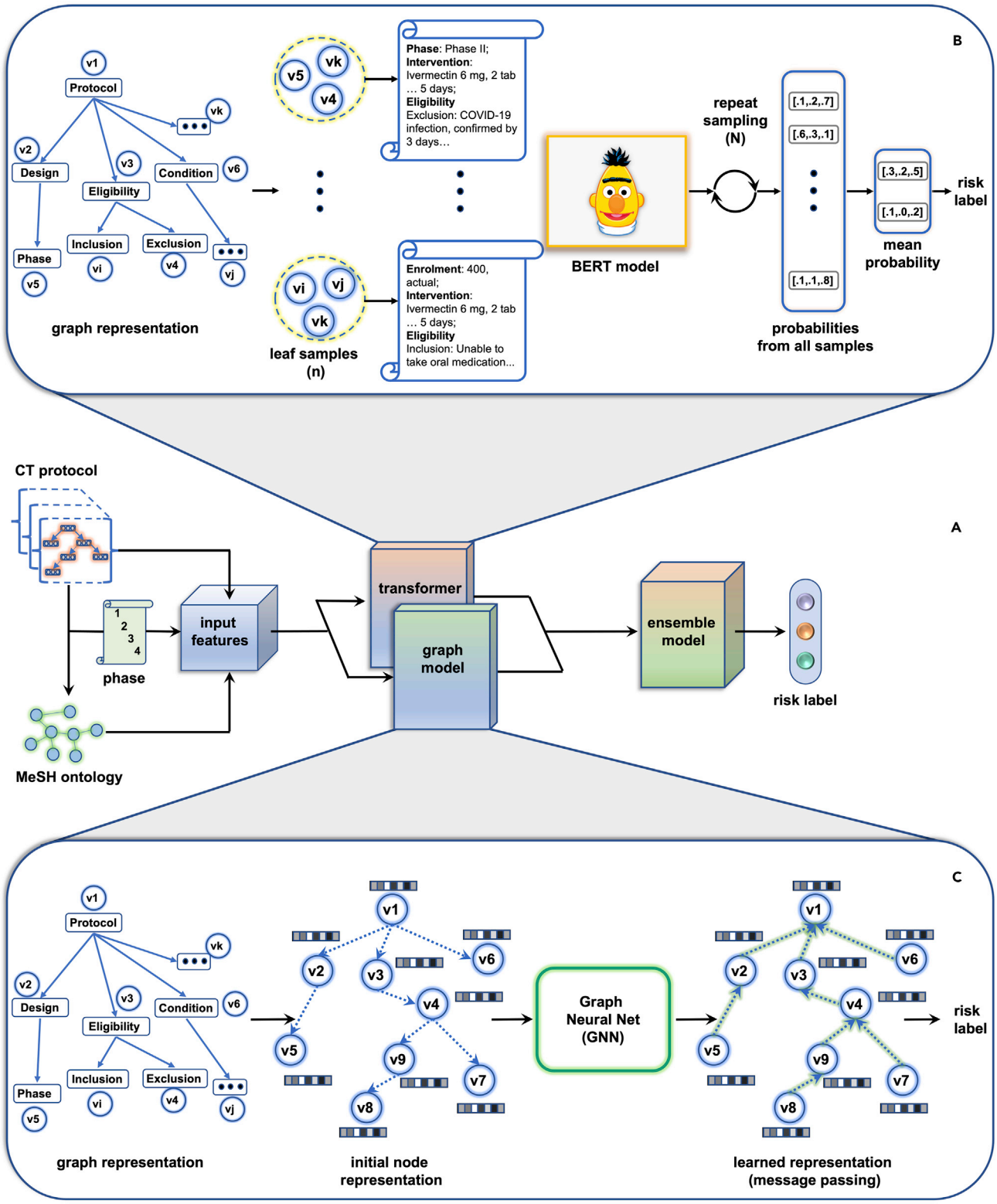


Figure 10. CT risk classification diagram

The schematic description of the predictive models: CT protocols, their associated phase, and condition categories are input as features to our two machine learning models, which are ensembled to predict the CT risk class.

protocol sections. This choice is furthermore driven by many successful applications of GNNs in various fields, e.g., the works of Stokes et al.[43] and Jumper et al.,[44] as well as our recent work addressing a simplified scenario of CT classification.[27,28]

The GNN-based CT risk prediction algorithm, as shown in Figure 10C, works by extracting raw vectorial features from individual textual pieces of the CT protocol and initializes the graph structure defined by a standard CT template with the vectorial features on their corresponding nodes.[27] The nodes of the graph will then propagate information between each other using L layers ($L = 4$ in our experiments) of the message passing algorithm.[45] This is to ensure that the network, along with the textual content, also incorporates the hierarchical structure of the CTs into account. This propagated information is then pooled together in a final vectorial representation, on which a fully connected neural network classifier is applied to predict the risk label. All these operations are performed with respect to a weighted cross-entropy classification loss to account for class imbalances and are optimized using the Adam algorithm with standard parameters. To extract vectorial features from text, we use a BOW representation as a baseline as well as the state-of-the-art transformer-based models described above. Unlike the above case and to minimize computational complexities, however, instead of fine-tuning the transformer weights we use pre-trained networks on large medical corpora and with sentence embedding capabilities.[33]

Statistical analyses

The selected subset of CTs were split into train, validation, and test sets with proportions of 70% (95,197), 15% (20,242), and 15% (20,501) unique CTs for the ternary risk model, respectively. The model parameters were trained on the training set, the hyper-parameters were tuned on the validation set, and model's performance results are reported on the test set using standard classification metrics: precision, recall, and F1 score (macro), accuracy and AUROC.

ACKNOWLEDGMENTS

This work is funded by the Swiss Innovation Agency – Innosuisse under the project with funding number 41013.1 IP-ICT.

AUTHOR CONTRIBUTIONS

S.F. and J.K. designed and implemented the models and ran the experiments and analyses. S.F. and D.T. wrote the manuscript draft. S.F. and N.B. created the benchmark dataset. D.T. and P.A. designed the experiments. All authors reviewed and approved the manuscript.

DECLARATION OF INTERESTS

P.A. and N.B. are employees and shareholders of Risklick AG.

Received: August 29, 2022
Revised: November 7, 2022
Accepted: January 16, 2023
Published: February 10, 2023

REFERENCES

1. Plenge, R.M. (2016). Disciplined approach to drug discovery and early development. Sci. Transl. Med. *8*, 349ps15. https://doi.org/10.1126/scitranslmed.aaf2608.
2. Friedman, L.M., Furberg, C.D., DeMets, D.L., Reboussin, D.M., and Granger, C.B. (2015). Fundamentals of Clinical Trials (Springer International Publishing).
3. Martin, L., Hutchens, M., and Hawkins, C. (2017). Trial watch: clinical trial cycle times continue to increase despite industry efforts. Nat. Rev. Drug Discov. *16*, 157. https://doi.org/10.1038/nrd.2017.21.
4. DiMasi, J.A., Grabowski, H.G., and Hansen, R.W. (2016). Innovation in the pharmaceutical industry: new estimates of R& D costs. J. Health Econ. *47*, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012.
5. Wouters, O.J., McKee, M., and Luyten, J. (2020). Estimated research and development investment needed to bring a new medicine to market, 2009-2018. JAMA *323*, 844–853. https://doi.org/10.1001/jama.2020.1166.
6. Sertkaya, A., Wong, H.-H., Jessup, A., and Beleche, T. (2016). Key cost drivers of pharmaceutical clinical trials in the United States. Clin. Trials *13*, 117–126. https://doi.org/10.1177/1740774515625964.
7. FDAAA 801 and the Final Rule. ClinicalTrials.gov. https://clinicaltrials.gov/ct2/manage-recs/fdaaa.
8. European Medicines Agency. Clinical trials regulation. https://www.ema.europa.eu/en/human-regulatory/research-development/clinical-trials/clinical-trials-regulation.
9. Cipriani, A., and Barbui, C. (2010). What is a clinical trial protocol? Epidemiol. Psichiatr. Soc. *19*, 116–117.
10. Turner, J.R. (2020). New FDA guidance on general clinical trial conduct in the era of COVID-19. Ther. Innov. Regul. Sci. *54*, 723–724. https://doi.org/10.1007/s43441-020-00160-0.
11. Wong, C.H., Siah, K.W., and Lo, A.W. (2019). Estimation of clinical trial success rates and related parameters. Biostatistics *20*, 273–286. https://doi.org/10.1093/biostatistics/kxx069.
12. Terry, C., and Lesser, N. (2018). Unlocking R&D Productivity: Measuring the Return from Pharmaceutical Innovation 2018. Deloitte.
13. Williams, R.J., Tse, T., DiPiazza, K., and Zarin, D.A. (2015). Terminated trials in the ClinicalTrials.gov results database: evaluation of availability of primary outcome data and reasons for termination. PLoS One *10*, e0127242. https://doi.org/10.1371/journal.pone.0127242.
14. Fogel, D.B. (2018). Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp. Clin. trials Commun. *11*, 156–164. https://doi.org/10.1016/j.conctc.2018.08.001.
15. Harrison, R.K. (2016). Phase II and phase III failures: 2013-2015. Nat. Rev. Drug Discov. *15*, 817–818. https://doi.org/10.1038/nrd.2016.184.
16. Hui, D., Glitza, I., Chisholm, G., Yennu, S., and Bruera, E. (2013). Attrition rates, reasons, and predictive factors in supportive care and palliative oncology clinical trials. Cancer *119*, 1098–1105. https://doi.org/10.1002/cncr.27854.
17. Bernardez-Pereira, S., Lopes, R.D., Carrion, M.J.M., Santucci, E.V., Soares, R.M., de Oliveira Abreu, M., Laranjeira, L.N., Ikeoka, D.T., Zazula, A.D., Moreira, F.R., et al. (2014). Prevalence, characteristics, and predictors of early termination of cardiovascular clinical trials due to low recruitment: insights from the ClinicalTrials.gov registry. Am. Heart J. *168*, 213–219.e1. https://doi.org/10.1016/j.ahj.2014.04.013.
18. DiMasi, J.A., Florez, M.I., Stergiopoulos, S., Peña, Y., Smith, Z., Wilkinson, M., and Getz, K.A. (2020). Development times and approval success rates for drugs to treat infectious diseases. Clin. Pharmacol. Ther. *107*, 324–332. https://doi.org/10.1002/cpt.1627.
19. Follett, L., Geletta, S., and Laugerman, M. (2019). Quantifying risk associated with clinical trial termination: a text mining approach. Inf. Process. Manag. *56*, 516–525. https://doi.org/10.1016/j.ipm.2018.11.009.
20. Feijoo, F., Palopoli, M., Bernstein, J., Siddiqui, S., and Albright, T.E. (2020). Key indicators of phase transition for clinical trials through machine learning. Drug Discov. Today *25*, 414–421. https://doi.org/10.1016/j.drudis.2019.12.014.
21. Elkin, M.E., and Zhu, X. (2021). Predictive modeling of clinical trial terminations using feature engineering and embedding learning. Sci. Rep. *11*, 3446. https://doi.org/10.1038/s41598-021-82840-x.
22. Geletta, S., Follett, L., and Laugerman, M. (2019). Latent Dirichlet Allocation in predicting clinical trial terminations. BMC Med. Inform. Decis. Mak. *19*, 242. https://doi.org/10.1186/s12911-019-0973-y.
23. Elkin, M.E., and Zhu, X. (2021). Understanding and predicting COVID-19 clinical trial completion vs. cessation. PLoS One *16*, e0253789. https://doi.org/10.1371/journal.pone.0253789.
24. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.

25. Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. N. Engl. J. Med. *380*, 1347–1358. https://doi.org/10.1056/NEJMra1814259.

26. Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. *25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7.

27. Ferdowsi, S., Borissov, N., Knafou, J., Amini, P., and Teodoro, D. (2021). Classification of hierarchical text using geometric deep learning: the case of clinical trials corpus. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).

28. Ferdowsi, S., Copara, J., Gouareb, R., Borissov, N., Jaume-Santero, F., Amini, P., and Teodoro, D. (2022). On graph construction for classification of clinical trials protocols using Graph Neural Networks. In 20th International Conference on Artificial Intelligence in Medicine, AIME 2022, M. Michalowski, S.S.R. Abidi, and S. Abidi, eds. (Springer Cham).

29. Fu, T., Huang, K., Xiao, C., Glass, L.M., and Sun, J. (2022). HINT: hierarchical interaction network for clinical-trial-outcome predictions. Patterns *3*, 100445. https://doi.org/10.1016/j.patter.2022.100445.

30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems, pp. 5998–6008.

31. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North (Association for Computational Linguistics), pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

32. Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. IEEE Signal Process. Mag. *34*, 18–42. https://doi.org/10.1109/MSP.2017.2693418.

33. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2022). Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthc. *3*, 1–23. https://doi.org/10.1145/3458754.

34. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, D. Precup and Y.W. Teh, eds. (PMLR), pp. 3319–3328.

35. Gysel, C.V., de Rijke, M., and Kanoulas, E. (2018). Neural vector spaces for unsupervised information retrieval. ACM Trans. Inf. Syst. *36*, 1–25. https://doi.org/10.1145/3196826.

36. Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., and Ward, R. (2016). Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. IEEE/ACM Trans. Audio Speech Lang. Process. *24*, 694–707. https://doi.org/10.1109/TASLP.2016.2520371.

37. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550.

38. Tjoa, E., and Guan, C. (2021). A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans. Neural Netw. Learn. Syst. *32*, 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314.

39. Ghassemi, M., Oakden-Rayner, L., and Beam, A.L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. Lancet. Digit. Health *3*, e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9.

40. Babic, B., Gerke, S., Evgeniou, T., and Cohen, I.G. (2021). Beware explanations from AI in health care. Science *373*, 284–286. https://doi.org/10.1126/science.abg1834.

41. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings.

42. Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 3651–3657. https://doi.org/10.18653/v1/P19-1356.

43. Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. Cell *180*, 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021.

44. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

45. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., and Dahl, G.E. (2020). Message passing neural networks. In Machine learning meets quantum physics (Springer), pp. 199–214. https://doi.org/10.1007/978-3-030-40245-7_10.