BMI 707 Project Check-In

HINTing at the Outcome: Predicting Clinical Trial Success using Deep Learning

Aishwarya Chander, Benedikt Geiger, Kezia Irene, Man Qing Liang, Thomas Smits

1. Primary Research Question:

Given a set of features on a clinical trial and drug data(e.g., drug, disease, protocol details, drug molecular properties), can we predict if a trial will succeed or not?

Our research question stems from reading about the HINT model, a hierarchical interaction graph for predicting clinical trial success by Fu et al.[1] We hope to combine data from the Trial Outcome Prediction (TOP) dataset (used by Fu et al. as a benchmark dataset) with drug feature data from the Therapeutic Data Commons (TDC) platform to note if the addition of molecular drug features can help improve performance scores as given by the HINT model.

Current model performance scores as reported on the HINT project's GitHub Page on data from various parts of the clinical trial pipeline:

Dataset	PR-AUC	F1	ROC-AUC
Phase I	0.745 (0.009)	0.820 (0.007)	0.726 (0.009)
Phase II	0.685 (0.011)	0.754 (0.010)	0.698 (0.008)
Phase III	0.709 (0.009)	0.757 (0.008)	0.784 (0.009)
Indication	0.702 (0.008)	0.776 (0.009)	0.786 (0.008)

Table 1: Model metrics from the Fu et al. study [1]

We're also adding trial characteristics from AACT, a publicly available relational database containing all information about every study registered in ClinicalTrials.gov to note how demographic labels can impact outcome prediction. Our data analysis shows a disproportionate distribution in Trial Data and will be discussed in further sections.

2. Summary of the datasets:

A summary of the datasets we propose to use and their features are as follows:

I. Trial Outcome Prediction Dataset

The dataset contains information on 17, 614 trials, 13, 880 small-molecule drugs and 5, 335 diseases, and is manually curated by Fu et al.

II. AACT Dataset

The AACT dataset contains 54 tables that provide information related to clinical trials from ClinicalTrials.gov. The main key used to link these tables is the NCT_ID which is common across HINT and AACT, allowing us to merge data from the two tables.

III. Therapeutic Data Commons Dataset

Contains a variety of datasets on drug pharmacokinetics, toxicity, drug targets and drug-drug interactions, which can be leveraged as additional features by matching the SMILES string in these datasets to the SMILES string in TOP.

3. Deep learning approaches:

Our goal is to evaluate the clinical trial success using a graph neural network. For this, the first task is to be able to represent each sample in the graph space as a node, and determine the space between different samples, such that trials that are similar have small distance in the graph space, whereas trials that are very different have a large distance.

For generating necessary embeddings, we split the data into various parts as described above to generate test, train and validation data across the 3 clinical trial phases. One component is the 'criteria' variable, which is in available free text form. To be able to extract meaningful embeddings, we will be applying a NLP approach with BERT and applying it to our model. We will be measuring model performance across different trial phases as executed by Fu et al., allowing us to compare their results to ours.

A summary of the approached that we will use are as follows:

1. **Classifiers:** Simple regression models to test associations between variables, and their relevance to the final model.
2. **Graph Neural Network:** A complex deep learning model that comprises several patient, drug, disease and demographic embeddings that will work to predict the final output, which is the outcome of the clinical trial, a label that says predictive trial success or failure.

   The structure of our GNN is as follows:

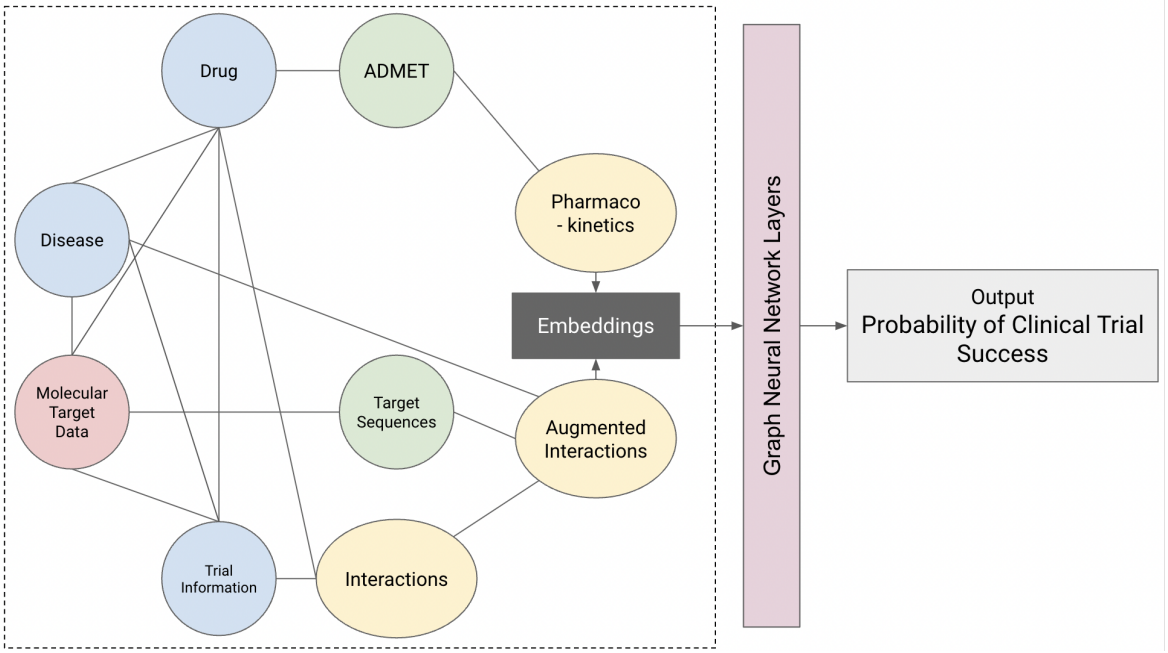

Fig 1. Tentative overview of our GNN model

4. Current results:

4.1. Generating the dataset

Currently, we've merged the AACT dataset with the TOP dataset using the NCT_ID column to generate a final dataset that has 13 columns. Each row is linked to a unique NCT_ID. A total of 9 datasets from TOP are combined with data from AACT; each file corresponding to training, testing and validation datasets for Phase I, II and III of clinical trials.

	file	nct_id	status	phase	diseases	icdcodes	drugs	smiless	criteria	study_date	country	participant_count	outcome
0	phase_I_train.csv	NCT01187615	terminated	phase 1	['small cell lung carcinoma']	['['D02.20', 'D02.21', 'D02.22']']	['regorafenib (bay73-4506) - sequential / cisp...	[[H][N]1([H])[C@@H]2CCCC[C@H]2[N]([H])([H])[P...	\n Inclusion Criteria:\n\n - ...	2010-07-16	United States	NaN	0
1	phase_I_train.csv	NCT01046487	completed	phase 1	['cancer']	['['C05.2', 'C10.0', 'C16.0', 'C16.4', 'C17.0'...	['imatinib mesylate, cyclophosphamide (dosing ...	['CC1=NC(NC2=NC=C(S2)C(=O)NC2=C(C)C=CC=C2Cl)=C...	\n Inclusion Criteria:\n\n - ...	2010-01-11	France	NaN	1
2	phase_I_train.csv	NCT01381887	completed	phase 1	['diabetes mellitus, type 2']	['['E11.65', 'E11.9', 'E11.21', 'E11.36', 'E11...	['placebo', 'canagliflozin 300mg/placebo', 'ca...	['CN1C(=O)C=C(N2CCC[C@@H](N)C2)N(CC2=C(C=CC=C2...	\n Inclusion Criteria:\n\n - ...	2011-06-09	United States	NaN	1
3	phase_I_train.csv	NCT02015676	completed	phase 1/phase 2	['breast cancer']	['['C79.81', 'D24.1', 'D24.2', 'D24.9', 'D49.3...	['trastuzumab', 'paclitaxel', 'myocet']	[[H][N]1([H])[C@@H]2CCCC[C@H]2[N]([H])([H])[P...	\n Inclusion Criteria:\n\n - ...	2013-12-03	Spain	54.0	1
4	phase_I_train.csv	NCT02015676	completed	phase 1/phase 2	['breast cancer']	['['C79.81', 'D24.1', 'D24.2', 'D24.9', 'D49.3...	['trastuzumab', 'paclitaxel', 'myocet']	[[H][N]1([H])[C@@H]2CCCC[C@H]2[N]([H])([H])[P...	\n Inclusion Criteria:\n\n - ...	2013-12-03	Spain	54.0	1

Fig 2. A poor view of our dataset

The columns are kept constant across the 9 datasets and are as follows:

From TOP:

1. **file**: Information about clinical trial phase, and training, testing or validation data.
2. **nct_id:** Unique identifier corresponding to each row of the dataset that is consistent across TOP and AACT. Represents one specific clinical trial study.
3. **status:** Status of the trial.
   Values include: 'terminated', 'completed', 'withdrawn', 'active, not recruiting', 'unknown status', 'suspended'
4. **phase:** Phase of the clinical trial in which the data has been collected.
5. **diseases:** Text representation of disease information that the patient has been exposed to.
6. **icdcodes:** International Classification of Disease 10 standard used to embed disease information as codes.
7. **drugs:** List of chemical drug names for each of the drugs that the patient is exposed to. Information includes granularity regarding placebos and which drug the placebo is being administered against.
8. **smiless:** String of drug information in IUPAC SMILES format. This column will be used to merge information from the TCD commons dataset as we build complex deep learning models and include information regarding drug makeup.
9. **criteria:** Eligibility criteria of the individual to participate in the trial.

From AACT:

10. **study_date:** Exact date of the clinical trial in YYYY-MM-DD format.
11. **country:** List of country names from which the data has been collected.
12. **participant_count:** Participant count from each clinical trial study.
13. **outcome:** Manually annotated label for whether a trial has passed or failed. Labels represent an overview of if the study has successfully passed onto the next stage of the trial.

Future additions from TDC linking on SMILES strings:

1. **Drug-Target Interaction:** Information about the activity of a small-molecule and its binding efficacy to its target.
2. **Drug-Drug Interaction:** Interaction between two or more drugs, and the implications of their interaction when administered at the same time.

4.2. Baseline:

So far, we've tested a simple logistic regression classifier to note the impact of participant count data on clinical trial outcome success. We were able to note that the model predicts with an AUC of near 0.5, or randomly assigns an outcome, indicating the necessity of the trial data embeddings.

Our team has also taken a look at the HINT model, its code and structure, the models they've built and tested to use as a baseline. Our initial model building has led us to believe that the performance metrics reported in their study are poor due to noisy data, which we will confirm as we work with the data ourselves.

4.3. Data Exploration post merge:

An important part of building any deep learning model is to understand the data we're working with to try and get ahead of any potential biases that we might see as we build our model. Some of the visualizations we generated to achieve this is as follows:

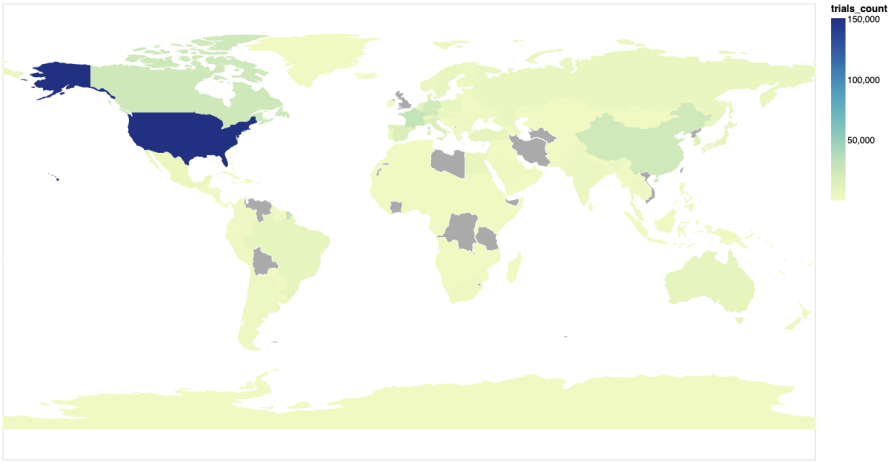

Fig 3. Number of clinical trials of all phase around the world

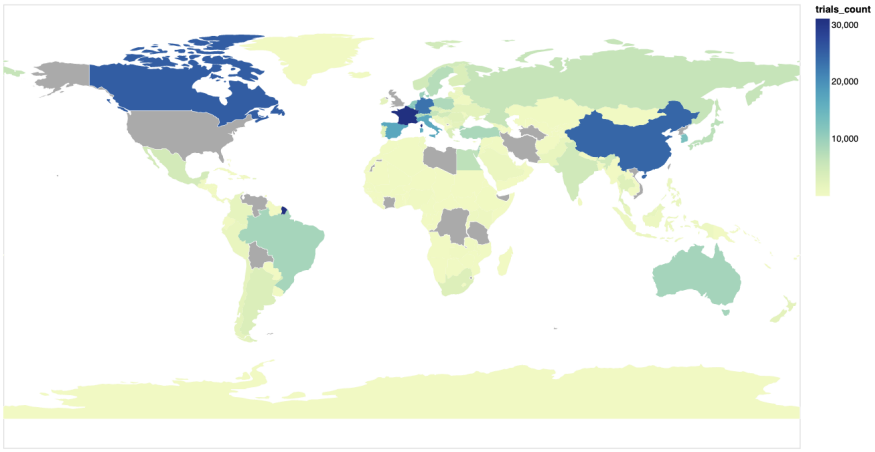

Fig 4. Number of clinical trials of all phase around the world, excluding USA

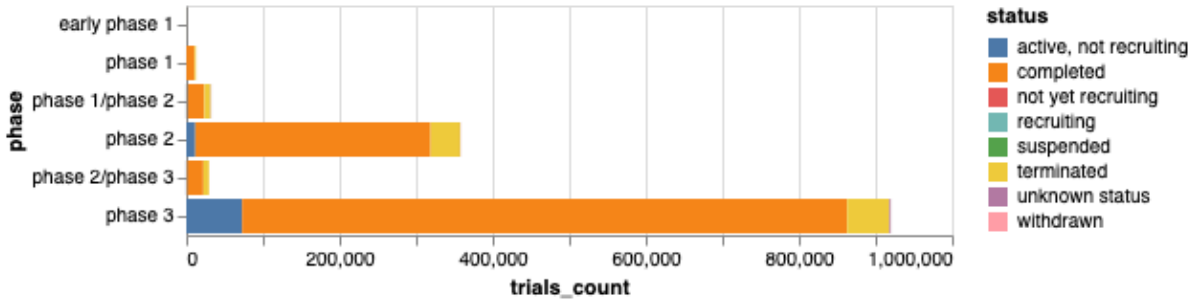

Fig 5. Number of trial counts over different phases and trial status

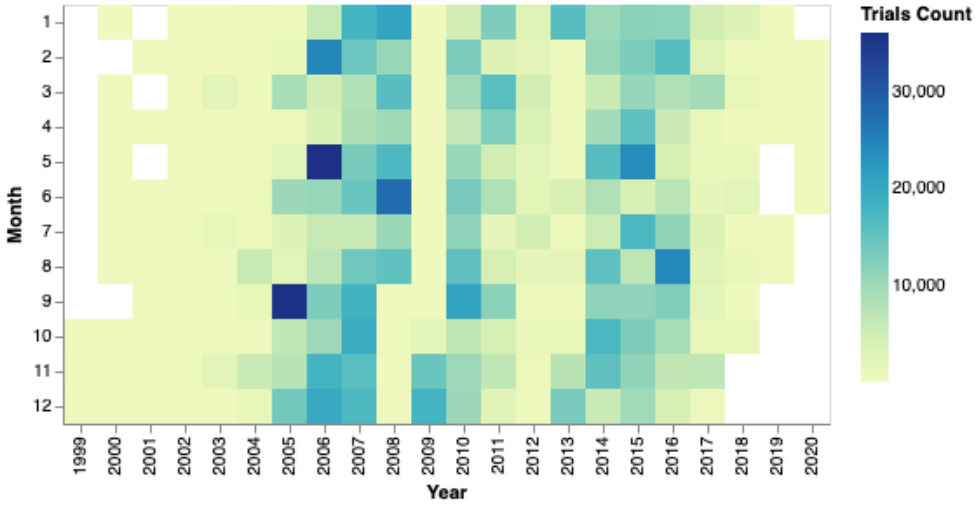

Fig 6. Number of clinical trials across the year

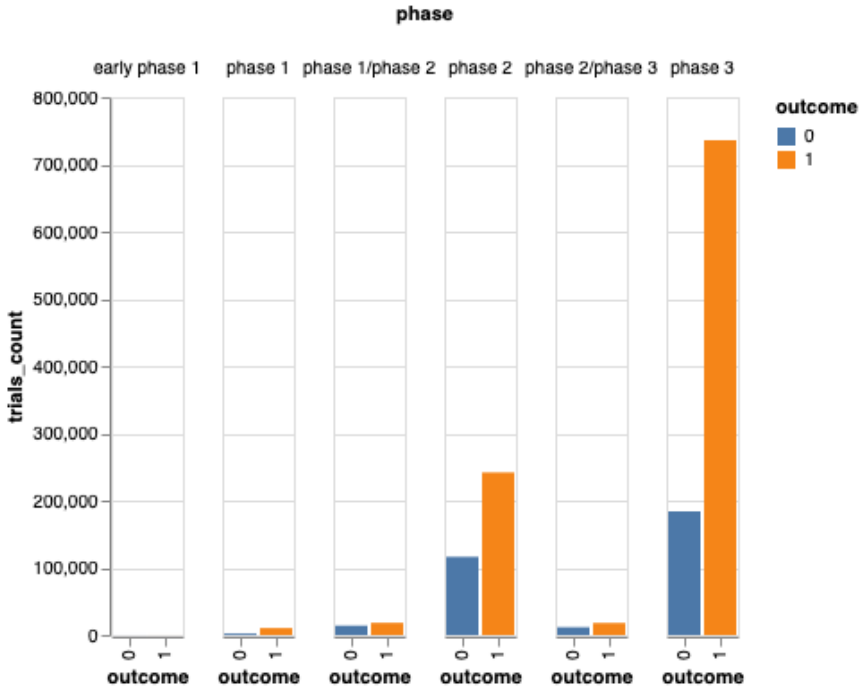

Fig 7. Trial Outcomes over Clinical Trial Phases

We can note a few demographic traits that might skew the results of our final model. Based on testing outcomes, we will look into potential data transformations to accommodate for skew-ness produced by these values. Furthermore, we note that the datasets for all trial phases contain class imbalance, which we plan to address with a weighted loss function and resampling (SMOTE).

We're currently in the process of exploring options for the best way to deal with these data issues.

5. Challenges

- While we initially thought that HINT had published their 'criteria' embeddings made with BioBERT, their .pkl file was empty. This has led us to set up our own NLP model to generate embeddings for the criteria, but BioBERT's documentation is sparse. We may decide to use the general BERT model.

- The HINT model relies on a series of dependencies that are not as well documented as initially interpreted by us. Understanding the code is a time consuming task which we're yet to fully conquer.

- By combining datasets and including information regarding small-molecules, we lose several data-points from our original dataset. We're yet to explore the implications of this and will have to test our model for overfitting. We're also yet to compute how much smaller our dataset will become with the addition of molecular data but are aware of this shortcoming. We continue to be motivated to include this information as the relevance of it to the success of clinical trials could be significant based on our preliminary literature review.

6. Ongoing work:

6.1. Generate embeddings: We are working on generation of embeddings for the eligibility criteria variable with BioBERT or BERT. We are also exploring the use of pre-trained embeddings (e.g., cui2vec[2]) to generate embeddings for drugs and diseases.

6.2. Balancing the dataset: During data analysis we've noted class-imbalances and skewness. We're currently in the process of exploring loss-function optimizations that can help us overcome these issues.

6.3. Building baseline model: Baseline models will be built using standard approaches as recorded by Fu et al. in the HINT study.

6.4. Addition of data from TDC commons: This goal requires a lot of data parsing and cleaning and is an ongoing effort by members of our team. We hope to use this as a 4th dataset and compare performance against the HINT baselines and our GNN models.

6.5. Describing layers for the GNN: We're currently exploring packages and models that utilize graph neural networks for clinical data prediction and are in the literature review stage.

7. Future Tasks:

7.1. Build and validate GNN using AACT+TOP data.
7.2. Build and validate GNN using AACT+TOP+TDC data.
7.3. Produce comparative metrics (F1/ Precision/Recall /AUC) between models.
7.4. Generate necessary reports.

References:

1. Fu, T., Huang, K., Xiao, C., Glass, L. M. & Sun, J. HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns* 100445 (2022) doi:10.1016/j.patter.2022.100445.
2. Beam, A. L. *et al.* Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 25, 295–306 (2020).