Automatically Labeling Clinical Trial Outcomes: A Large-Scale Benchmark for Drug Development

Chufan Gao*, Jathurshan Pradeepkumar*, Trisha Das*, Shivashankar Thati, and Jimeng Sun

Project Page

Paper

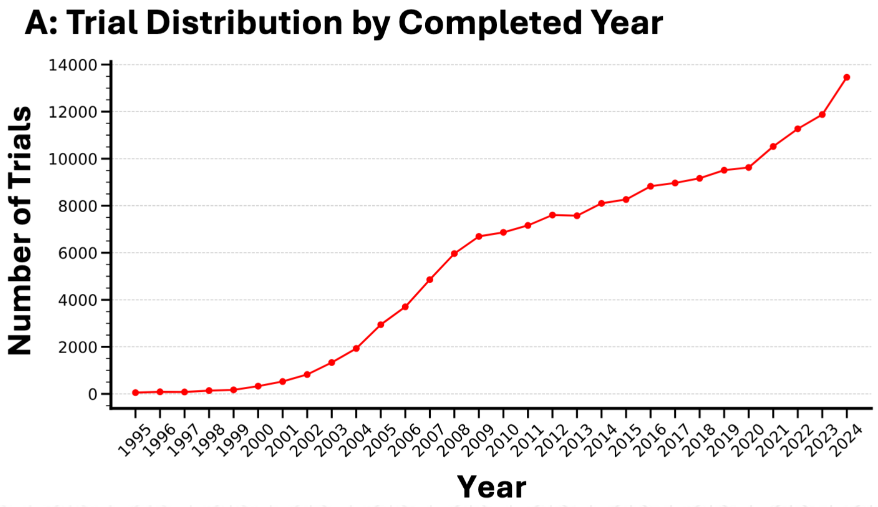
Code

Background

- Drug discovery and development are expensive, with clinical trial results vital for regulatory approval and patient care.

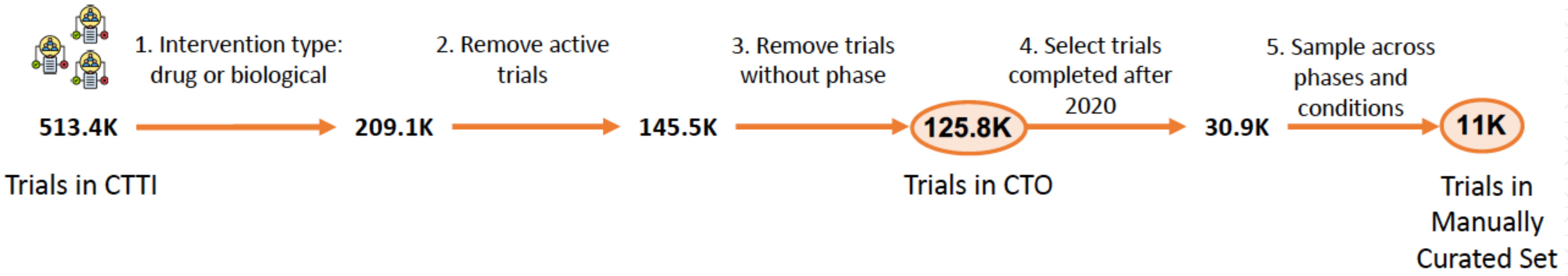
- Large-scale, high-quality clinical trial outcome data remains limited.
    - Hindering the development of predictive models

- Dynamic & Rapidly Growing Data
    - Clinical trial data grows rapidly and is affected by diverse external factors (e.g., COVID-19, regulatory changes).
    - Frequent label updates are needed, but **manual labeling** is **impractical** at scale.



Clinical Trial Outcome Benchmark (CTO)

- Clinical Trial Outcome (CTO) benchmark, a fully reproducible, regularly updated, large-scale repository encompassing approximately ~125K drug and biologics trials.
  - A comprehensive trial knowledge base
  - Automated labeling framework based on aggregation of weak labeling function
  - Manually curated around 11K trials.



1. Intervention type: drug or biological

2. Remove active trials

3. Remove trials without phase

4. Select trials completed after 2020

5. Sample across phases and conditions

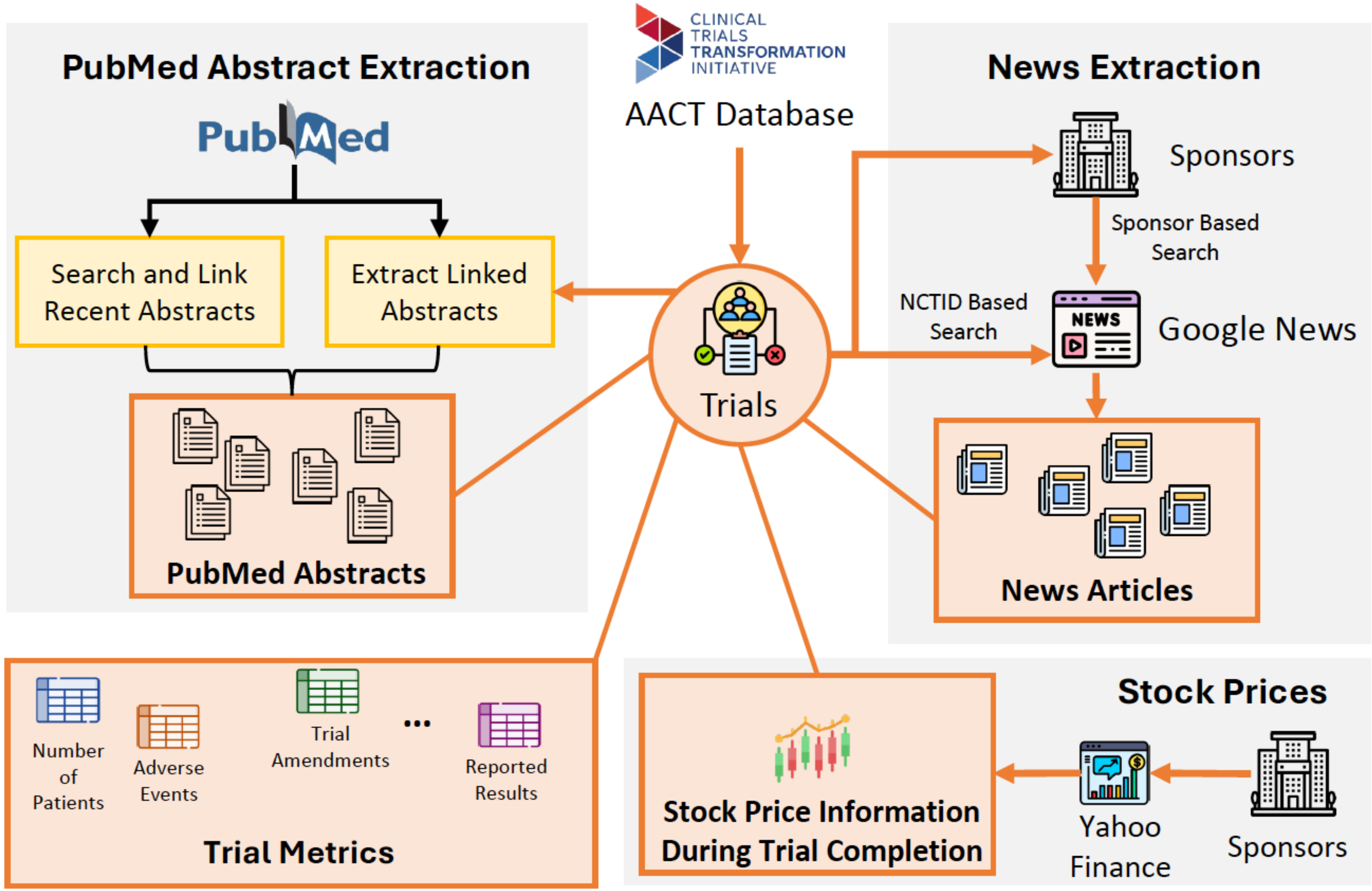
513.4K → 209.1K → 145.5K → 125.8K → 30.9K → 11K

Trials in CTTI

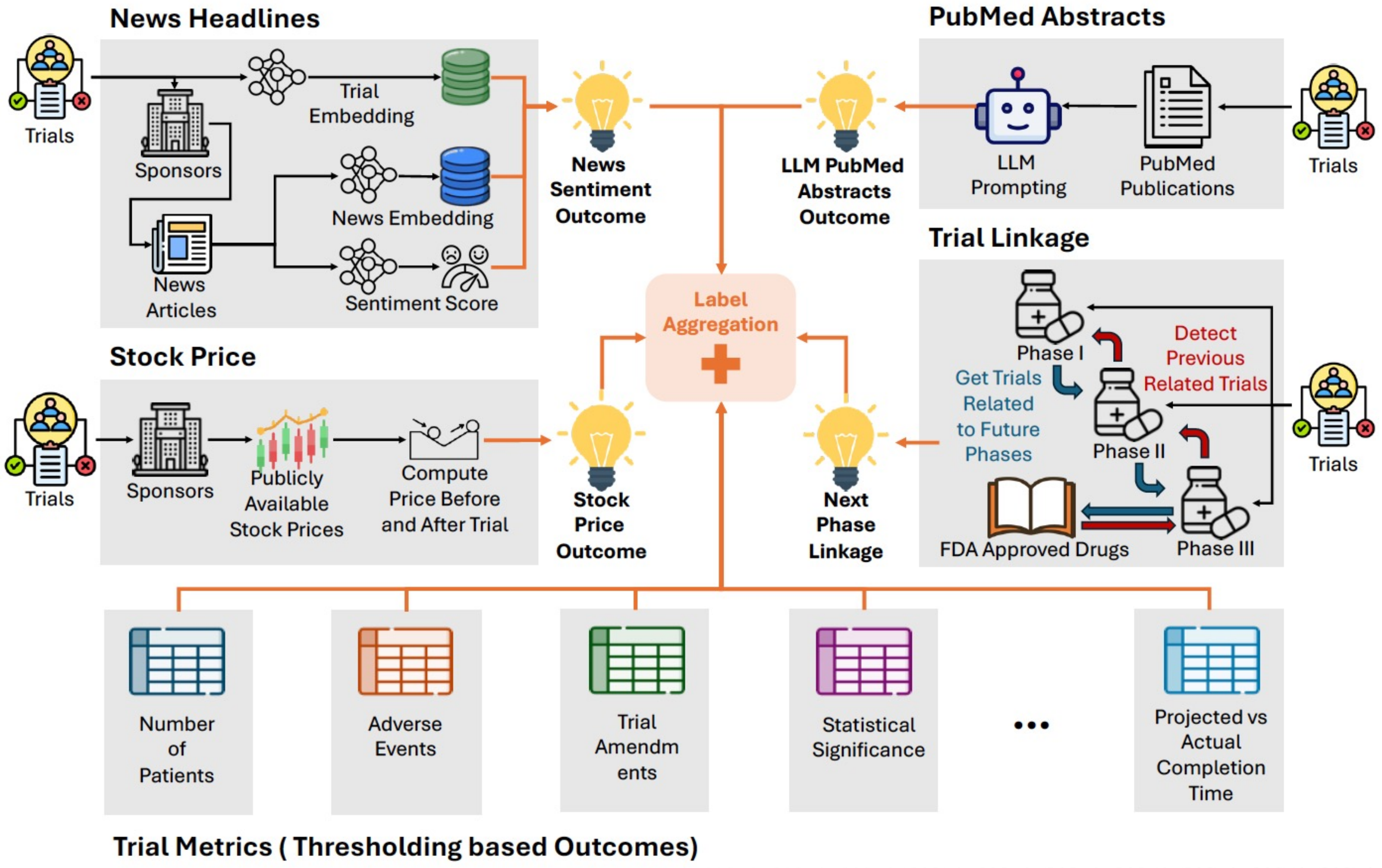
Trials in CTO

Trials in Manually Curated Set

Trial Knowledge Base



CTO Automated Labeling Framework



News Headlines

Trials

Sponsors

Trial Embedding

News Articles

News Embedding

Sentiment Score

News Sentiment Outcome

Stock Price

Trials

Sponsors

Publicly Available Stock Prices

Compute Price Before and After Trial

Stock Price Outcome

Label Aggregation

PubMed Abstracts

LLM Prompting

PubMed Publications

Trials

LLM PubMed Abstracts Outcome

Trial Linkage

Phase I

Get Trials Related to Future Phases

Phase II

Detect Previous Related Trials

Trials

FDA Approved Drugs

Phase III

Next Phase Linkage

Number of Patients

Adverse Events

Trial Amendm ents

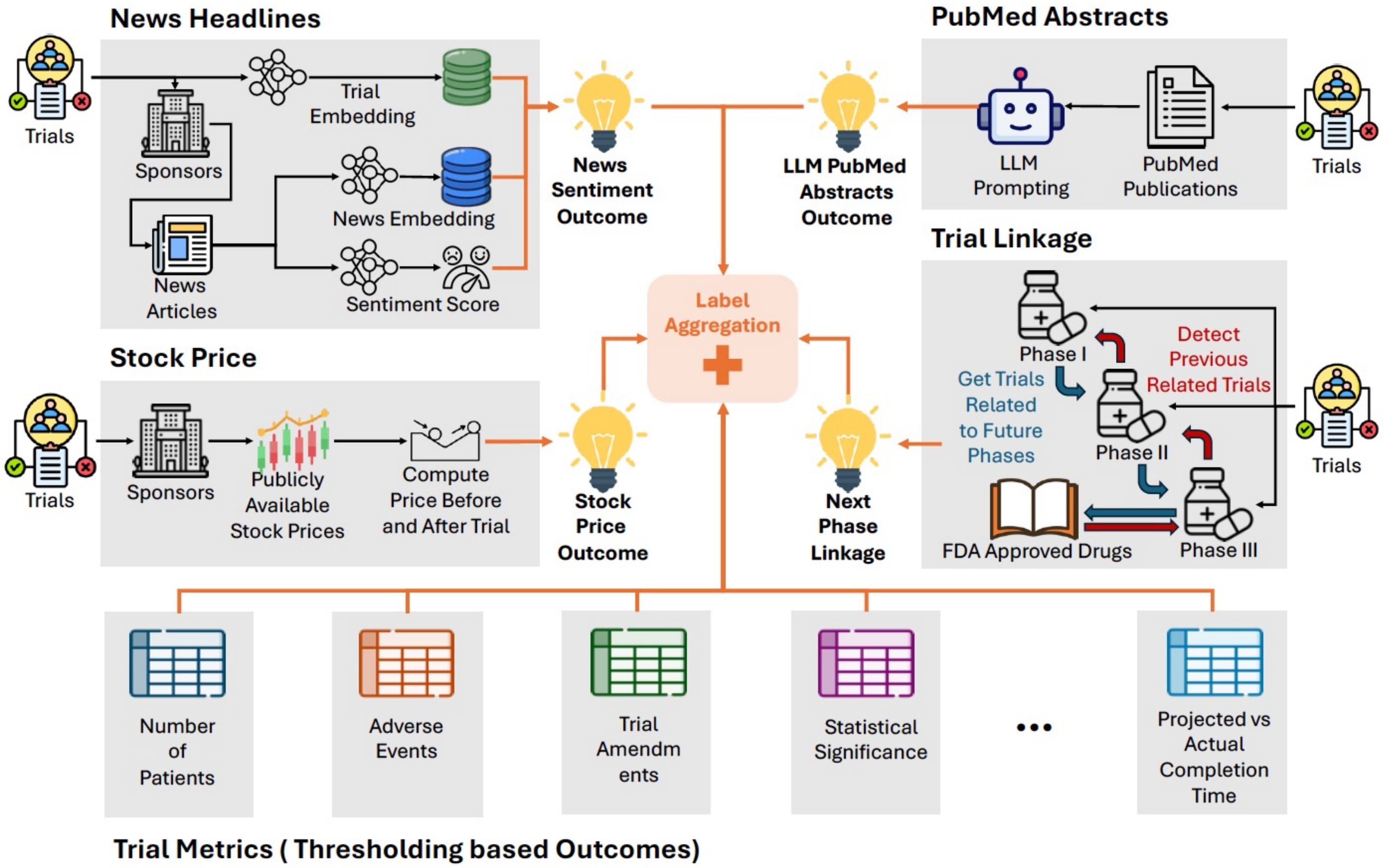
Statistical Significance

...

Projected vs Actual Completion Time

Trial Metrics (Thresholding based Outcomes)

CTO Automated Labeling Framework

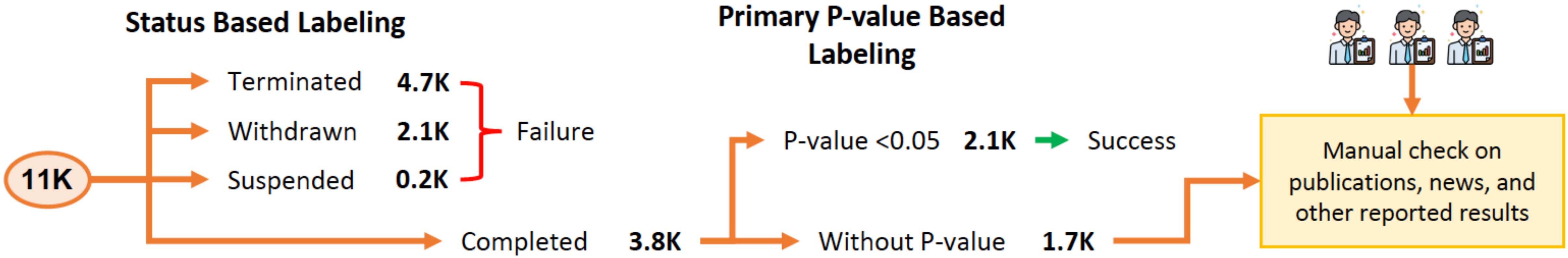


Phase	Aggregation Method	F1	κ
I	CTO_{MV}	0.726	0.490
	CTO_{DP}	0.870	0.700
	CTO_{RF}	**0.913**	**0.790**
II	CTO_{MV}	0.689	0.430
	CTO_{DP}	0.856	0.623
	CTO_{RF}	**0.878**	**0.693**
III	CTO_{MV}	0.904	0.606
	CTO_{DP}	0.921	0.582
	CTO_{RF}	**0.941**	**0.710**
All	CTO_{MV}	0.793	0.529
	CTO_{DP}	0.884	0.646
	CTO_{RF}	**0.909**	**0.729**

Agreement of CTO automated labels with human-labeled TOP[1] dataset

[1] Fu, Tianfan, et al. "Hint: Hierarchical interaction network for clinical-trial-outcome predictions." *Patterns* 3.4 (2022).

Manual Curation Process



Status Based Labeling

11K

Terminated 4.7K

Withdrawn 2.1K

Suspended 0.2K

Failure

Completed 3.8K

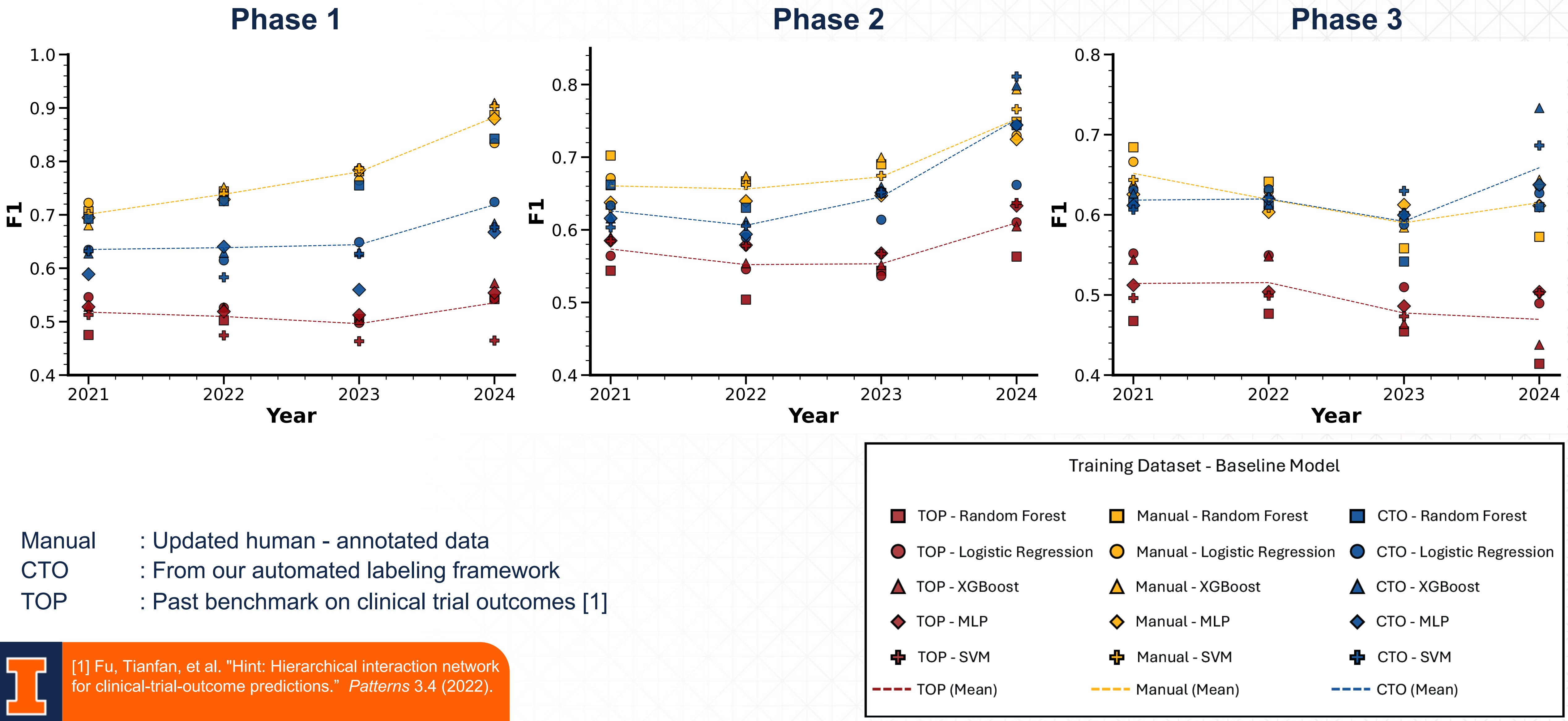
Primary P-value Based Labeling

P-value <0.05 2.1K → Success

Without P-value 1.7K

Manual check on publications, news, and other reported results

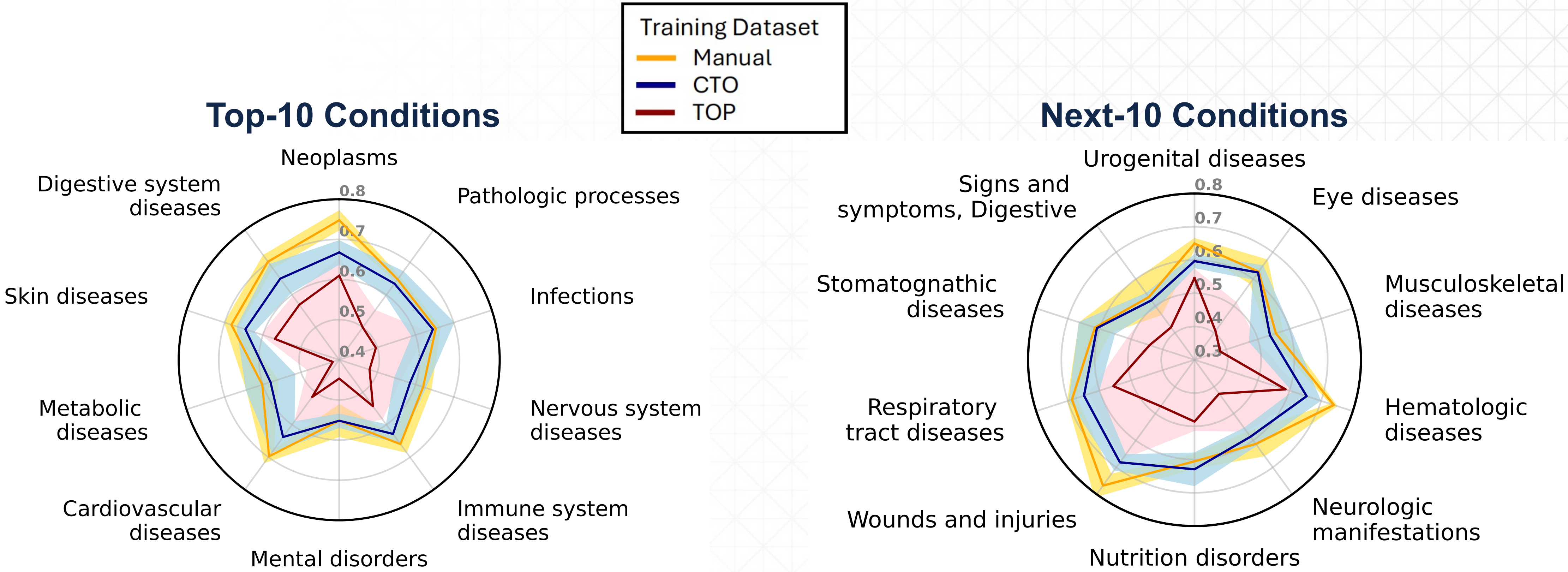
Which labels are better?



Manual : Updated human - annotated data
CTO : From our automated labeling framework
TOP : Past benchmark on clinical trial outcomes [1]

[1] Fu, Tianfan, et al. "Hint: Hierarchical interaction network for clinical-trial-outcome predictions." *Patterns* 3.4 (2022).

Which labels are better?



Top-10 Conditions

Next-10 Conditions

[1] Fu, Tianfan, et al. "Hint: Hierarchical interaction network for clinical-trial-outcome predictions." *Patterns* 3.4 (2022).

Manual	: Updated human - annotated data
CTO	: From our automated labeling framework
TOP	: Past benchmark on clinical trial outcomes [1]

Automatically Labeling Clinical Trial Outcomes: A Large-Scale Benchmark for Drug Development

Chufan Gao*, Jathurshan Pradeepkumar*, Trisha Das*, Shivashankar Thati, and Jimeng Sun

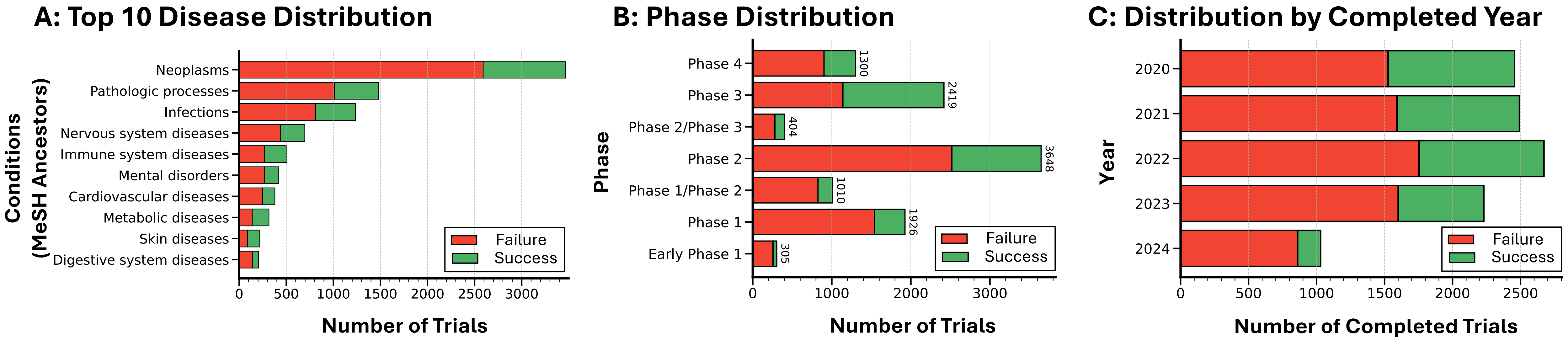
ILLINOIS

Project Page

Paper

Code

Manual Curated Trial Outcome Benchmark

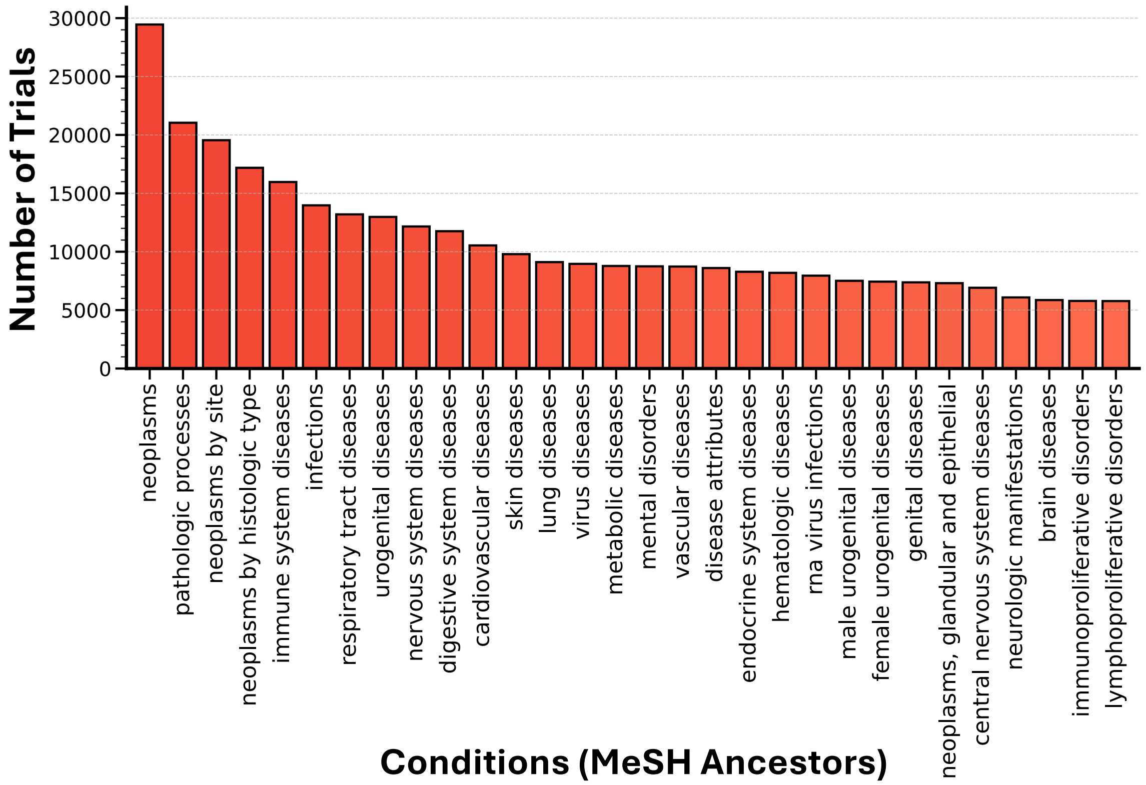


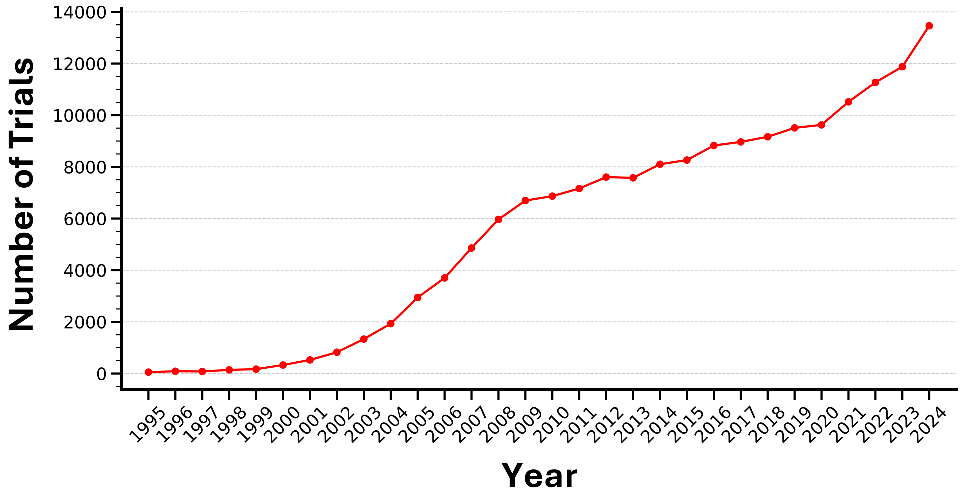
A: Top 10 Disease Distribution

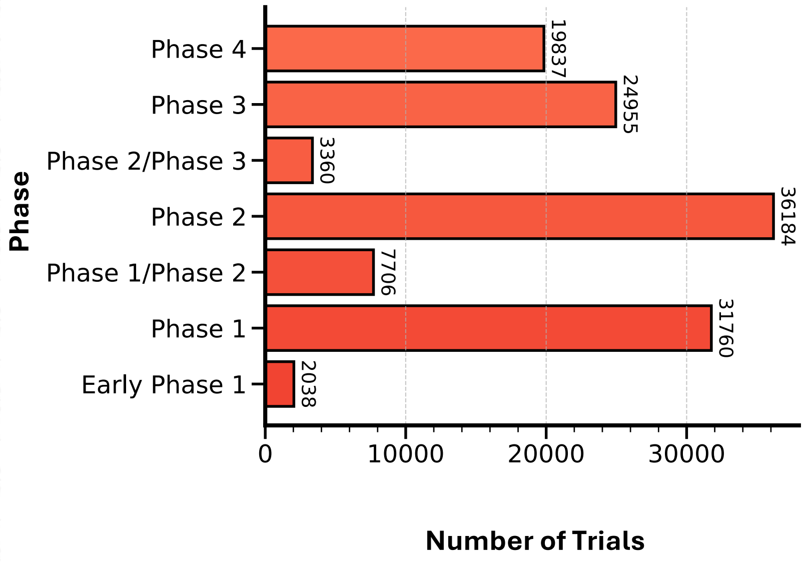
B: Phase Distribution

C: Distribution by Completed Year

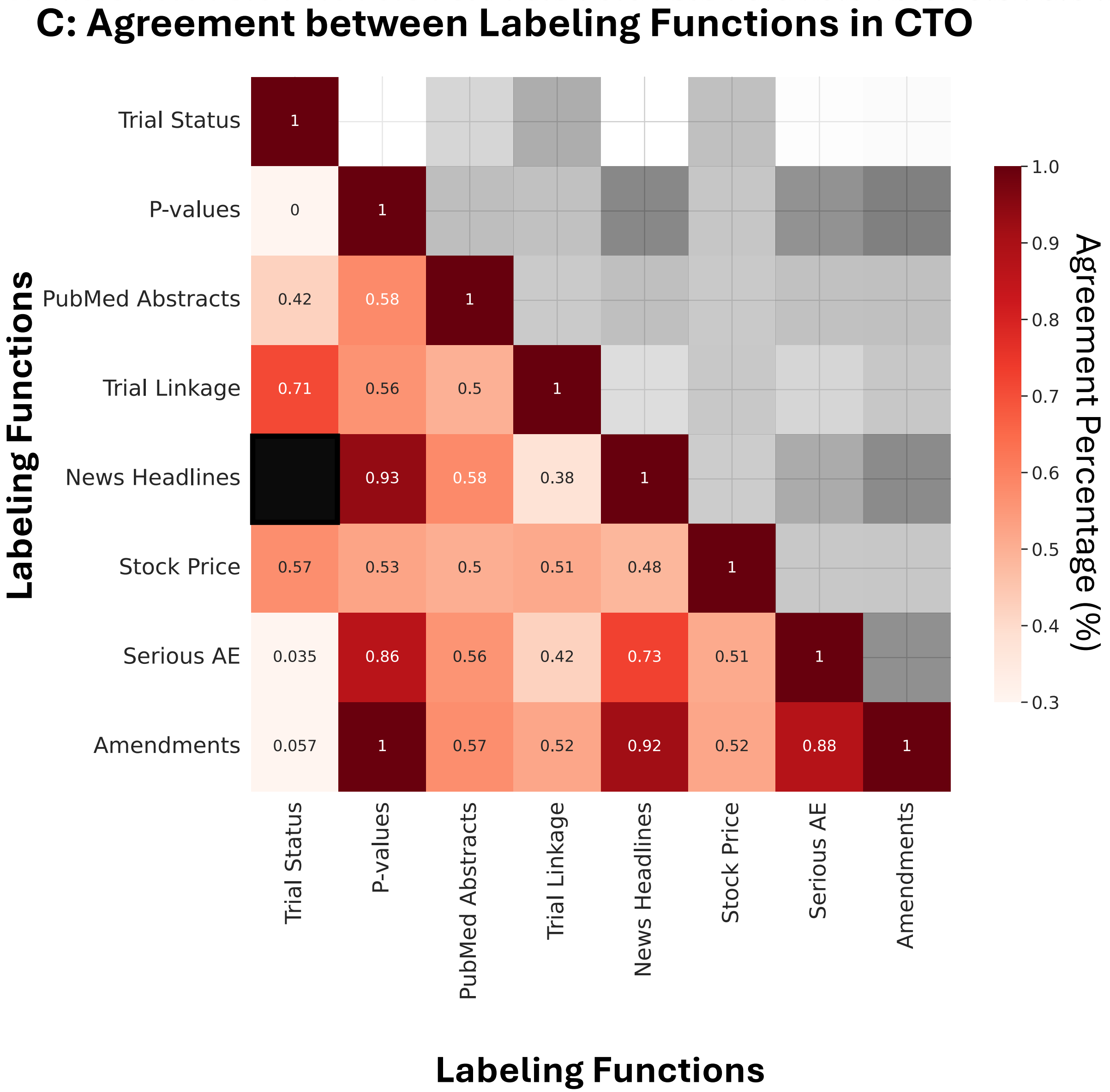
CTO Statistics


B: Frequency of Top 30


A: Trial Distribution by Completed Year

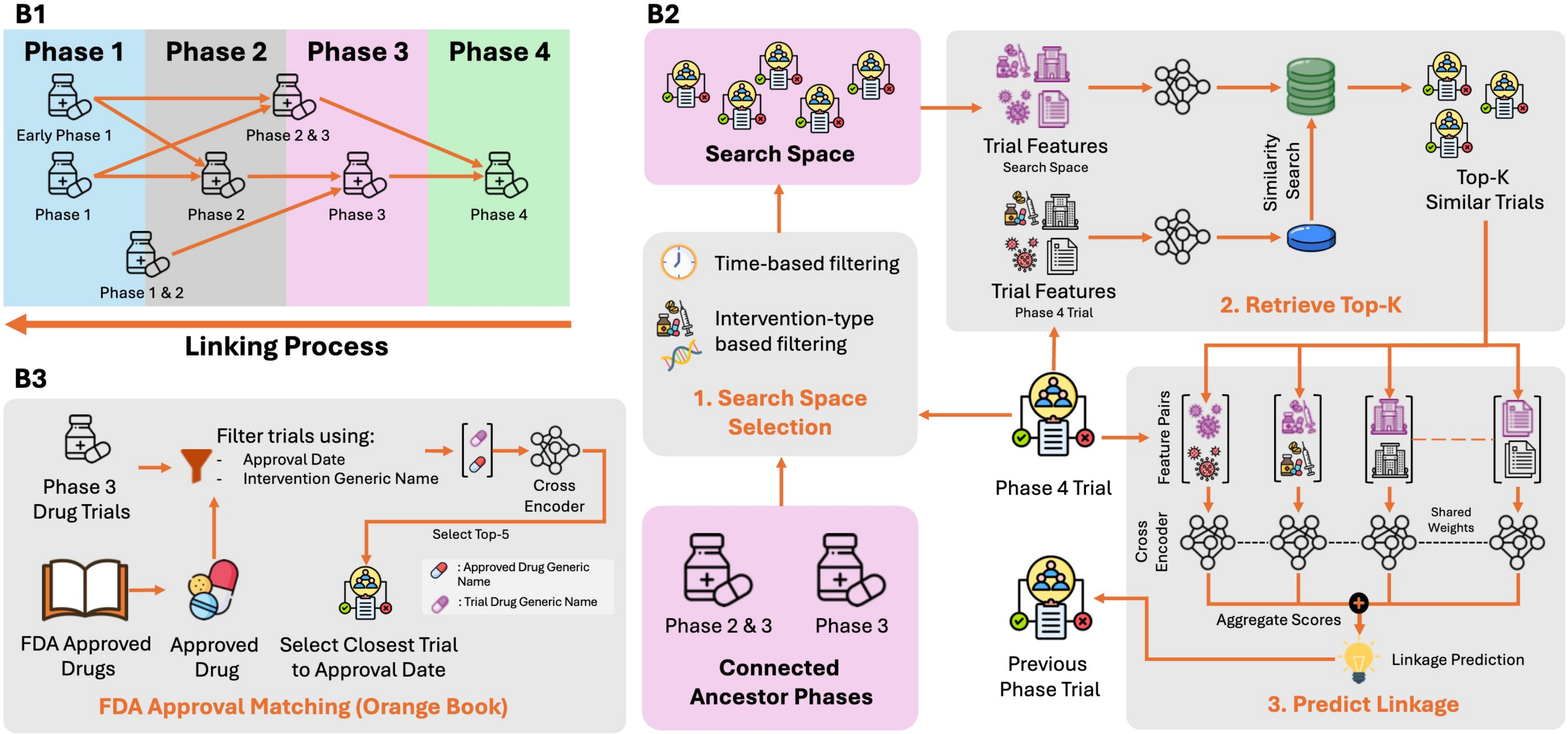

D: Phase Distribution

Agreement between Labeling Functions in CTO



Additional Method / Weak Labeling Slides

B. Trial Linkage Algorithm



B1

Phase 1 | Phase 2 | Phase 3 | Phase 4

Early Phase 1

Phase 1

Phase 1 & 2

Phase 2 & 3

Phase 2

Phase 3

Phase 4

Linking Process

B3

Phase 3 Drug Trials

Filter trials using:

- Approval Date
- Intervention Generic Name

Cross Encoder

Select Top-5

: Approved Drug Generic Name

: Trial Drug Generic Name

FDA Approved Drugs

Approved Drug

Select Closest Trial to Approval Date

FDA Approval Matching (Orange Book)

B2

Search Space

Time-based filtering

Intervention-type based filtering

1. Search Space Selection

Phase 2 & 3

Phase 3

Connected Ancestor Phases

Trial Features

Search Space

Trial Features

Phase 4 Trial

Similarity Search

Top-K Similar Trials

2. Retrieve Top-K

Phase 4 Trial

Feature Pairs

Cross Encoder

Shared Weights

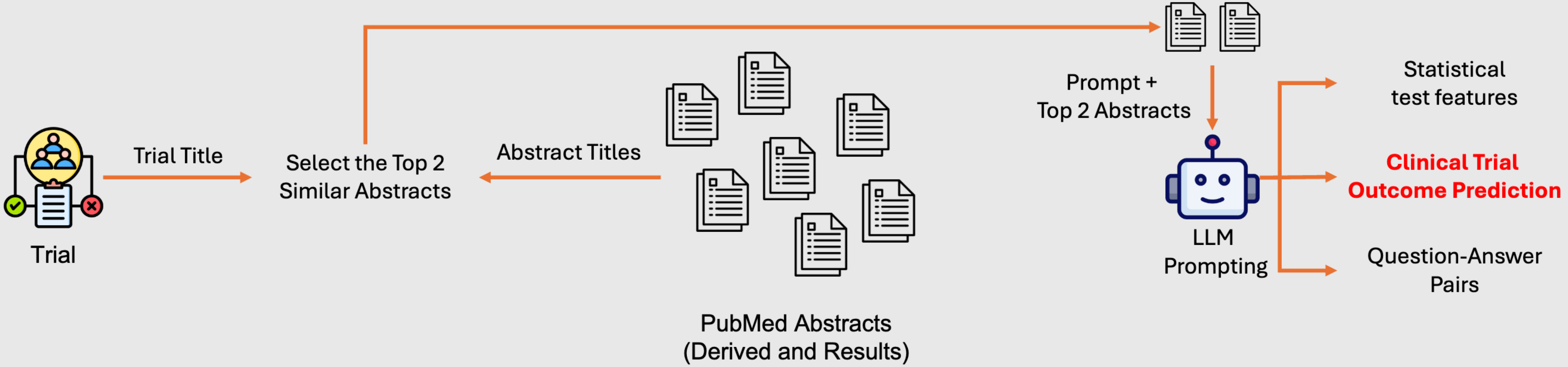
Aggregate Scores

Linkage Prediction

Previous Phase Trial

3. Predict Linkage

Additional Method / Weak Labeling Slides


Trial
Trial Title
Select the Top 2 Similar Abstracts
Abstract Titles
PubMed Abstracts (Derived and Results)
Prompt + Top 2 Abstracts
LLM Prompting
Statistical test features
Clinical Trial Outcome Prediction
Question-Answer Pairs