Paper ML11

Leveraging open data and machine learning enhanced data extraction for clinical trials insight

Hugo Bottois, Sanofi, Paris, France

Paulin Charliquart, Sanofi, Stockholm, Sweden

Andre Couturier, Sanofi, New Jersey, USA

Edward Gillian, Cytel, Gorzów Wielkopolski, Poland

Mukul Mittal, Sanofi, New Jersey, USA

ABSTRACT

Open databases such as ClinicalTrials.gov provide valuable information about clinical trials for specific diseases or drugs, including data from protocol and study results.

However, some information is not always directly provided in a structured format in protocol descriptions (dosage, frequency, and route of drugs) or in results (type of group).

These limitations raise challenges to extract information and fully leverage open data for clinical projects. Machine learning tools like Med7 and BioBERT allow extraction and mapping of this information, making exploitation of data easier over a larger number of clinical trials and reducing human labor.

These processes allow multiple use cases such as extracting adverse event data for meta-analytic risk modeling, gathering data to produce tables, or providing a landscape of clinical trials over time with a data visualization dashboard.

Thus, making data extraction and monitoring easier unlocks numerous use cases for clinical trial analysis, automation, and exploration.

INTRODUCTION

Efforts for sharing clinical data become increasingly important to leverage machine learning and data analysis and foster biomedical discovery during the drug development process. The scientific and clinical trial community has increasingly recognized the role and utility of open data and the transparency of clinical trial protocols and results.

Among the benefits, those data can generate new research hypotheses or help to think about an analysis strategy for clinical programmers and statisticians. This also provides a database prior to project in signal detection or machine learning/Artificial Intelligence (AI) applications.

Open clinical databases such as ClinicalTrials.gov have become an important source of data for reviews and meta-analyses.

This data contains a myriad of information including the description of the experimental protocol, cohort characteristics, objective, and endpoint over clinical trial sponsor. For some study, it also contains the results, including safety (adverse event distribution) and efficiency (endpoint result) data.

This project aims to monitor and analyze the landscape of clinical trials for specific diseases and drugs.

It primarily leverages data from open public databases (ClinicalTrials.gov and PubMed) and machine learning models/AI to extract additional information from non-structural data, enabling more comprehensive analysis and insights.

These data allow for multiple use cases we have in our company such as:

- Providing a landscape of clinical trials over time
- Gathering data to produce custom tables and data visualizations
- Extracting adverse event safety data for meta-analytic risk modeling

PROVIDING A LANDSCAPE OF CLINICAL TRIALS

CLINICALTRIALS.GOV DATA EXTRACTION AND VISUALIZATION

The first step of the project was to gather data about the clinical trials on ClinicalTrials.gov (maintained by the National Center for Biotechnology Information). Initially, this platform was only a web-based application with raw XLM format data. Fortunately, the release 2.0 of an API (Application Programming Interface) in 2024 has significantly simplified data extraction from trial.gov [1].

We first used the API [2] to fetch data about the protocol section related data for different disease or drugs (**Figure 1**).

We extracted information such as study type (Interventional), phase (Phase 3), intervention model (Parallel), primary purpose (Treatment), masking methods (Double-blind), study status (Completed), enrollment count, enrollment type (Actual), condition (Asthma), and drugs used (Aspirin). Additionally, it included details about eligibility criteria like age and sex, and outcomes measured.

All this data is provided across time and sponsor (academic or industrial) and provides a comprehensive overview of the clinical trial design actors and can be used to answer various questions:

How many clinical trials and at which phase / status? The drug is involved in which condition? by which pharmaceutical company?

What is the demographic distribution (age, sex/gender) of the study participants and how many? What is the study design of these trials? What is the duration of these trials?

Which big pharmas have the most clinical trials and patients recruited for a disease or a drug?

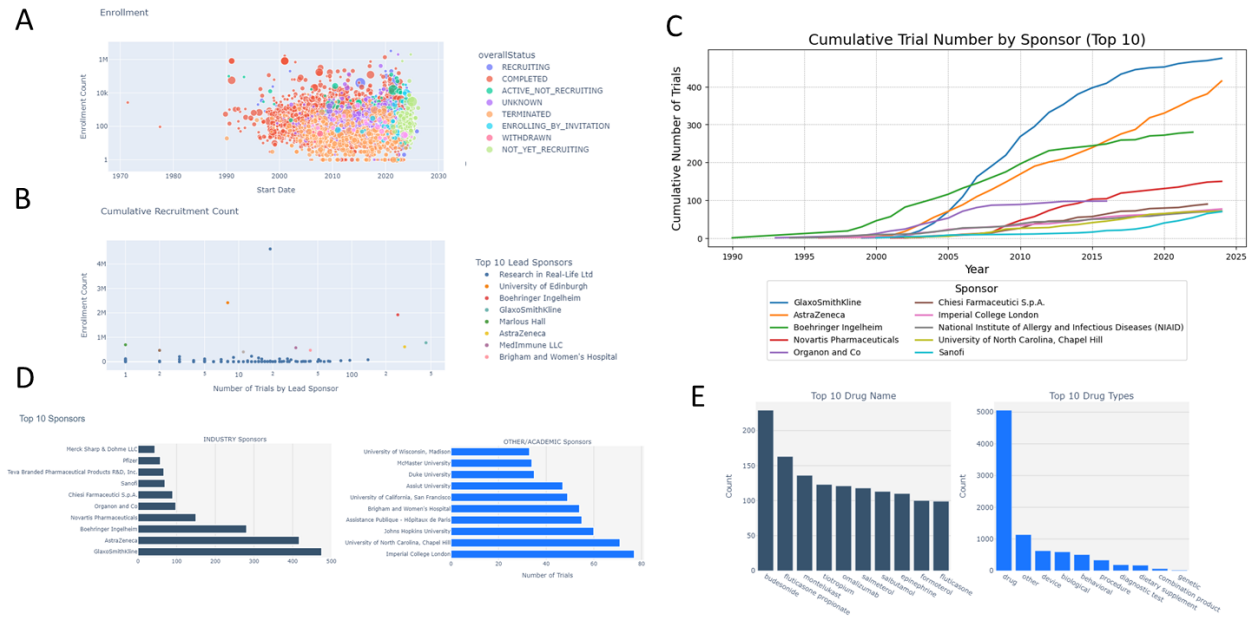

Figure 1: Landscape of clinical trials for Asthma disease with trial.gov data

(A) Relationship between the start date of clinical trials and the enrollment count (actual or planned) with the number of study arms per trial (point size) and overall status (complete, terminated). This data helps identify trends in enrollment size and trial complexity over time.

(B) Relationship between the total enrollment count and the number of clinical trials conducted by each industrial sponsor (top 10 sponsors displayed). It shows the cumulative enrollment of patients for sponsors, providing an idea of the scale and reach of their clinical trial activities.

(C) Number of cumulative clinical trials conducted by sponsors over time. It tracks the accumulation of trials over the years, providing insights into the leaders in clinical trial volume.

(D) Top 10 industrial and academic sponsors by total clinical trial number.

(E) Top 10 most frequently mentioned drug names and types (drug, device, procedure). This provides a clear overview of which drugs and types are most commonly referenced in clinical trials.

CLINICAL TRIAL PUBLICATIONS VIA PUBMED API

Interestingly, the list of publications associated with a clinical trial is also indicated as “PMID” references (pubmed identifier) as well as the title and date of publication. It also specifies whether the publication is categorized as background (context information, research), result (reporting the study's result), or derived (containing secondary analyses or data derived from the study).

Using a PMID, additional information can be retrieved from another well-known database: Pubmed [3]. This platform (also maintained by the National Center for Biotechnology Information and the U.S. National Library of Medicine) is dedicated to search and retrieve biomedical and life sciences literature. PubMed offers an API that enables the extraction of various information, such as the publication URL, journal, type of publication (e.g., article or book), and associated keywords. This functionality is useful for determining which clinical trials have more publications and for prioritizing research efforts (**Figure 2**).

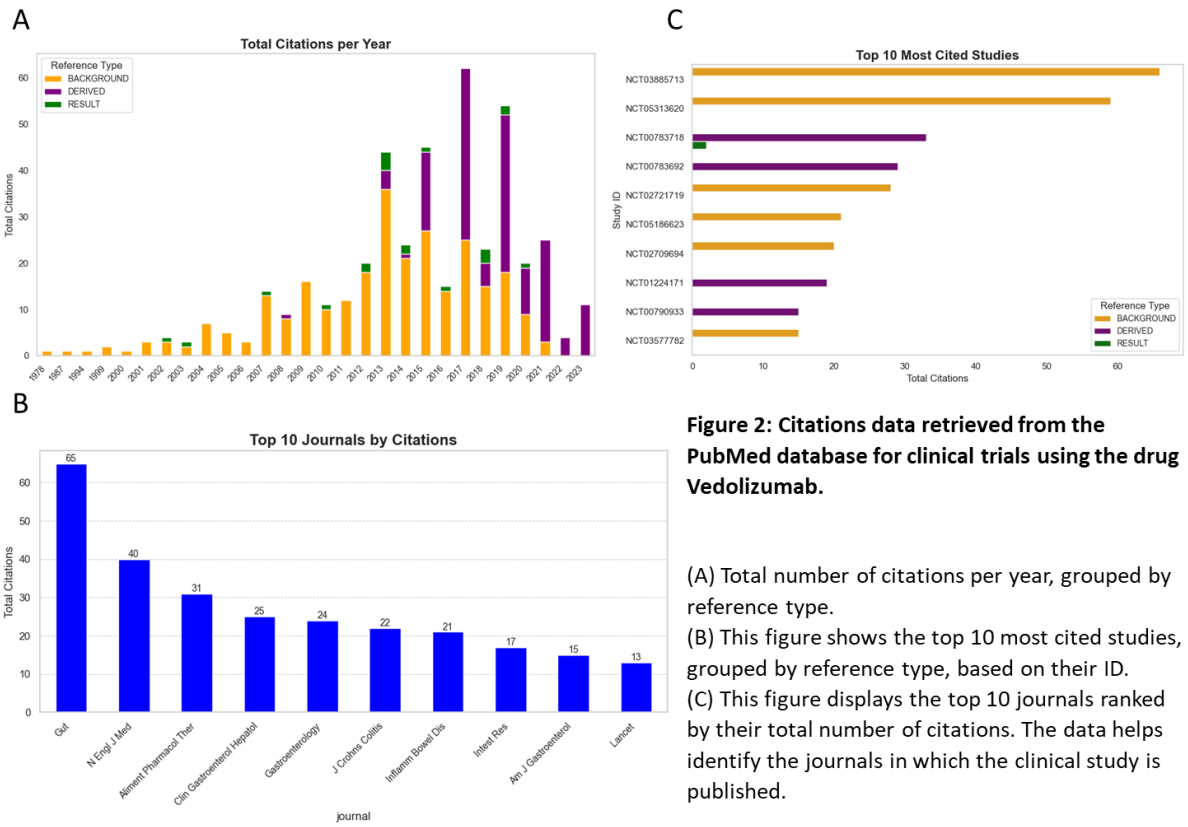

Figure 2: Citations data retrieved from the PubMed database for clinical trials using the drug Vedolizumab.

(A) Total number of citations per year, grouped by reference type.
(B) This figure shows the top 10 most cited studies, grouped by reference type, based on their ID.
(C) This figure displays the top 10 journals ranked by their total number of citations. The data helps identify the journals in which the clinical study is published.

Overall, we can provide a comprehensive overview of the current clinical trial landscape. This enables users to identify which companies or institutions are conducting the most trials on specific diseases or drugs, as well as which ones are generating the most publications. Additionally, it offers detailed data on protocol design and inclusion/exclusion criteria, providing further insights and focus of these trials.

Biopharma can use these API to get information about clinical trials for competing disease or drugs to prioritize R&D, analysis plans or marketing strategy.

RESULT DATA SECTION AND MACHINE LEARNING DATA EXTRACTION

DATA MAPPING PROTOCOL TO RESULT DATA

ClinicalTrials.gov also provides data about the results of a clinical trial for efficacy and safety including adverse events. An adverse event (AE) is defined by “an untoward medical occurrence after exposure to a medicine, which is not necessarily caused by that medicine” [4]. It could be events such as bacterial infection, abdominal pain, nausea.

We started acquiring AE from results data occurring to a patient over clinical trial arm for a drug or a disease like we did for the protocol description data. Each AE is associated with an organSystem (Infections and infestations, Nervous system disorders), eventType (Serious or Others) and numAtRisk (Total number of patients per group) and numAffected (number of patients experiencing at least once an AE), numEvents (number of total occurrences of an AE) for each group **(Table 1).**

A

Participant Group/Arm	Intervention/Treatment
Experimental: Alirocumab SC Q2W Alirocumab SC every 2 weeks (Q2W) from baseline (day 1) through week 10 during the double-blind treatment period Starting at week 12, and continuing through week 22, participants will receive open-label alirocumab SC Q2W	Drug: Alirocumab • Alirocumab SC Q2W • Other Names: ◦ PRALUENT® ◦ REGN727 ◦ SAR236553
Experimental: Placebo SC Q2W Matching placebo SC Q2W from baseline through week 10 during the double-blind treatment period Starting at week 12, and continuing through week 22, participants will receive open-label alirocumab SC Q2W	Drug: Alirocumab • Alirocumab SC Q2W • Other Names: ◦ PRALUENT® ◦ REGN727 ◦ SAR236553 Drug: Placebo • Matching placebo SC Q2W

B

Arm/Group Title	DB Placebo (DBTP)	DB Alirocumab 150 Q2W (DBTP)	DB Placebo (OLTP)	DB Alirocumab 150 Q2W (OLTP)
Arm/Group Description	Participants received matching placebo subcutaneously (SC) every 2 weeks (Q2W) from baseline (Day 1) through Week 10 during the double-blind treatment period.	Participants in this arm received alirocumab 150 milligrams (mg) SC Q2W from baseline (Day 1) through Week 10 during the double-blind treatment period.	Participants received matching placebo subcutaneously (SC) every 2 weeks (Q2W) from baseline (Day 1) through Week 10 during the double-blind treatment period. Starting at Week 12, and continuing through Week 22, all participants received open-label alirocumab SC Q2W	Participants in this arm received alirocumab 150 milligrams (mg) SC Q2W from baseline (Day 1) through Week 10 during the double-blind treatment period. Starting at Week 12, and continuing through Week 22, all participants received open-label alirocumab SC Q2W

C

Other (Not Including Serious) Adverse Events

Frequency Threshold for Reporting Other Adverse Events	5%							
Arm/Group Title	DB Placebo (DBTP)		DB Alirocumab 150 Q2W (DBTP)		DB Placebo (OLTP)		DB Alirocumab 150 Q2W (OLTP)	
	Affected / at Risk (%)	# Events	Affected / at Risk (%)	# Events	Affected / at Risk (%)	# Events	Affected / at Risk (%)	# Events
Total	4/24 (16.67%)		8/45 (17.78%)		2/24 (8.33%)		4/45 (8.89%)	
Gastrointestinal disorders								
Diarrhoea †1	0/24 (0.00%)	0	3/45 (6.67%)	3	0/24 (0.00%)	0	1/45 (2.22%)	1
Infections and infestations								
Nasopharyngitis †1	0/24 (0.00%)	0	2/45 (4.44%)	3	1/24 (4.17%)	1	3/45 (6.67%)	4
Upper respiratory tract infection †1	2/24 (8.33%)	3	2/45 (4.44%)	2	0/24 (0.00%)	0	0/45 (0.00%)	0
Nervous system disorders								
Headache †1	2/24 (8.33%)	8	2/45 (4.44%)	2	1/24 (4.17%)	2	1/45 (2.22%)	1

† Indicates events were collected by systematic assessment

1 Term from vocabulary, MedDRA (22.0)

Table 1: Arm description and Adverse events in the result section for study NCT03156621

(A) Title and description of arm in protocol section and (B) result group in result section
(C) Adverse events data with numatrisk (total number of patient per group) and numaffected (number of patient experiencing at least once the corresponding AE)

However, the result groups are not directly distinguishable with the titles used in the protocol, making it difficult to map the result and protocol data. Here, an arm refers to a group of participants receiving a specific type of intervention, which could be a comparator (Placebo comparator, Sham comparator, Active comparator) or the experimental treatment. Additionally, the group names are sometimes uninformative ("Group A" or "Group B"), requiring reading through the group descriptions to understand them.

As a result, while AE data may be presented for each group, it is unclear which groups are experimental or comparator without cross-referencing the protocol section, where the intervention details are listed.

One of our challenges was to map the arm types (experimental, placebo comparator, active comparator) from the protocol section to the groups presented in the results section.

This seems tricky, but we came across an article [5, 6] that outlined the same issue and found an elegant solution. The author used the BioBERT [7, 8] model for sentence embedding to map the arm type to the corresponding group. BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a pre-trained language model built for biomedical text mining tasks. We used dmis-lab/biobert-v1.1 model.

It can be used for sentence embedding, converting sentences into a vector representation to capture their semantic meanings. It allows comparison based on a contextual meaning rather than superficial features such as word order or exact matching. Each sentence is tokenized into tokens (word or sub word), then is fed to the Biobert model to be embedded into the vector representation.

These vectors can be mapped into a dimensional space, where the similarity between two sentences can be computed (cosine similarity) and indicate a semantic closeness.

In our case, we map the arm type (as described in the protocol) to its corresponding result group (as described in the result) by comparing their semantic similarities. This approach allows us to match the arm type with the greatest similarity along result group description/title (**Table 2**).

	Arm description	Type	Intervention Names	group	Result group description	Similarity	Best match
0	Cohort 1 Placebo Drug: Placebo	PLACEBO_COMPARATOR	[Drug: Placebo]	EG000	Main Period: Cohort 1 Placebo Placebo: Placebo	0.954078	True
5	Cohort 2 Minimum Dose IFX-1 (400 mg Q4W) Drug: IFX-1	EXPERIMENTAL	[Drug: IFX-1]	EG001	Main Period: Cohort 2 Minimum Dose IFX-1 (400 mg Q4W) IFX-1: Single IV infusions of IFX-1 diluted in sodium chloride.	0.970261	True
10	Cohort 3 Low dose IFX-1 (800 mg Q4W) Drug: IFX-1	EXPERIMENTAL	[Drug: IFX-1]	EG002	Main Period: Cohort 3 Low dose IFX-1 (800 mg Q4W) IFX-1: Single IV infusions of IFX-1 diluted in sodium chloride.	0.969034	True
15	Cohort 4 Medium Dose IFX-1 (800 mg Q2W) Drug: IFX-1	EXPERIMENTAL	[Drug: IFX-1]	EG003	Main Period: Cohort 4 Medium Dose IFX-1 (800 mg Q2W) IFX-1: Single IV infusions of IFX-1 diluted in sodium chloride.	0.969153	True
20	Cohort 4 Medium Dose IFX-1 (800 mg Q2W) Drug: IFX-1	EXPERIMENTAL	[Drug: IFX-1]	EG004	Main Period: Cohort 5 High Dose IFX-1 (1200 mg Q2W) IFX-1: Single IV infusions of IFX-1 diluted in sodium chloride.	0.968896	True
25	Cohort 3 Low dose IFX-1 (800 mg Q4W) Drug: IFX-1	EXPERIMENTAL	[Drug: IFX-1]	EG005	Extension Period: Cohort 3 Low dose IFX-1 (800 mg Q4W) IFX-1: Single IV infusions of IFX-1 diluted in sodium chloride.	0.971898	True
30	Cohort 4 Medium Dose IFX-1 (800 mg Q2W) Drug: IFX-1	EXPERIMENTAL	[Drug: IFX-1]	EG006	Extension Period: Cohort 4 Medium Dose IFX-1 (800 mg Q2W) IFX-1: Single IV infusions of IFX-1 diluted in sodium chloride.	0.971861	True

Table 2: Example of mapping using BioBERT embeddings applied to the arm description and group description based on the higher cosine similarity (example study NCT03487276).

We just had to adapt this approach as the ClinicalTrials.gov was not available at the time of the publication, and the authors worked with xml formatted data.

The authors also validated this approach by manually annotating 100 study records and evaluating the mapping. They achieved an accuracy of 84% for correctly mapping the intervention group and 80% for the comparator group over 100 result records which were manually annotated and randomly selected [5]. From our experimentation, we observed that studies with more complex protocols (dose escalation, groups receiving placebo followed by active drugs during randomization then open label) are more challenging to map.

While a manual review may still be necessary, these machine learning-based approaches, combined with trial data from ClinicalTrials.gov, can serve as a valuable data source for monitoring adverse event occurrences across treatment arms, drugs, or diseases. In the paper [5], the authors construct a knowledge graph to represent efficacy and safety results over arm, but various other use cases are possible as well.

We initially fetch those data to be used for meta-analytic risk modeling by using AE occurrence over comparator and experimental group.

This project's ambition is to estimate, based on the incidence of adverse effects observed in the blinded data from ongoing studies, the probability of an imbalance between the treatment groups, assuming that the baseline incidence of adverse effects follows the distribution we previously estimated.

EXTRACTING DRUG ADMINISTRATION DATA IN TRIALS FROM GROUP DESCRIPTION

Mapping the arm protocol to the result group is useful for identifying which group is the comparator or experimental. However, this approach has certain limitations. We encountered mismatches where some arms describe groups receiving both placebo and experimental drugs, yet they are still labeled as experimental. Additionally, this information may sometimes be incomplete or missing.

Moreover, certain information is still not easily accessible on ClinicalTrials.gov. For example, details such as drug dosage and frequency, route of administration, treatment dosage, and treatment duration are often more challenging to extract.

This information is typically only found within the text of the arm or group descriptions, such as:

“Participants received placebo to adalimumab subcutaneous injection once every other week and methotrexate orally once a week at a starting dose of 7.5 mg/week (could be escalated up to 20 mg/week) during the 2-year double-blind treatment phase.”

(Example from clinical Study NCT04115748)

To achieve this, we used a Named Entity Recognition (NER) open-source model called Med7 [9, 10]. This model extracts relevant variables related to drug administration and can identify 7 key entities.

- **Drug**: Generic or brand name of the medication (adalimumab).
- **Dosage**: The total amount of a drug administered.
- **Strength**: The amount of drugs in a given dosage (200 mg)
- **Route**: The path by which the drug is taken into the body (oral, subcutaneous).
- **Frequency**: The dosage regimen at which the medication should be administered (once daily, every two weeks)
- **Duration**: The length of time that the drugs was prescribed for (for 16 weeks).
- **Form**: A particular configuration of the drugs intended for marketed use (tablet, injection)

We used this model to extract those entities to have more usable information. This extraction process enables us to gather more comprehensive and structured information about how the drug is used, particularly in terms of dosage, administration, and route. More importantly, it could help us understand the actual procedures followed in each result group (**Table 3**).

	nctId	groupTitle	groupDescription	group_info_arm_type	DRUG	DOSAGE	STRENGTH	ROUTE	FREQUENCY	DURATION	FORM
0	NCT04115748	Filgotinib 200 mg (Main Study)	Filgotinib 200 mg tablet orally once daily + PTM filgotinib 100 mg tablet orally once daily + PTM adalimumab SC injection every two weeks for 16 weeks.	EXPERIMENTAL	['PTM filgotinib', 'adalimumab']	[]	['200 mg', '100 mg']	['orally', 'orally', 'SC']	['once daily', 'once daily', 'every two weeks']	['for 16 weeks']	['tablet', 'tablet', 'injection']
1	NCT04115748	Filgotinib 100 mg (Main Study)	Filgotinib 100 mg tablet orally once daily + PTM filgotinib 200 mg tablet orally once daily + PTM Adalimumab SC injection every two weeks for 16 weeks.	EXPERIMENTAL	['PTM filgotinib', 'Adalimumab']	[]	['100 mg', '200 mg']	['orally', 'orally', 'SC']	['once daily', 'once daily', 'every two weeks']	['for 16 weeks']	['tablet', 'tablet', 'injection']
2	NCT04115748	Adalimumab 40 mg (Main Study)	PTM filgotinib 200 mg tablet orally once daily + PTM filgotinib 100 mg tablet orally once daily + Adalimumab 40 mg SC injection every two weeks for 16 weeks.	EXPERIMENTAL	['PTM filgotinib', 'Adalimumab']	[]	['200 mg', '100 mg', '40 mg']	['orally', 'orally', 'SC']	['once daily', 'once daily', 'every two weeks']	['for 16 weeks']	['tablet', 'tablet', 'injection']
3	NCT04115748	Placebo (Main Study)	PTM filgotinib 200 mg tablet orally once daily + PTM filgotinib 100 mg tablet orally once daily + PTM adalimumab SC injection every two weeks for 16 weeks.	PLACEBO_COMPARATOR	['PTM filgotinib', 'adalimumab']	[]	['200 mg', '100 mg']	['orally', 'orally', 'SC']	['once daily', 'once daily', 'every two weeks']	['for 16 weeks']	['tablet', 'tablet', 'injection']
4	NCT04115748	Filgotinib 200 mg From Filgotinib 200 mg (LTE)	Filgotinib 200 mg tablet orally once daily + PTM filgotinib 100 mg tablet orally once daily for up to 34 weeks. Participants received filgotinib 200 mg in the Main Study.	EXPERIMENTAL	['PTM filgotinib', 'filgotinib']	[]	['200 mg', '100 mg', '200 mg']	['orally', 'orally']	['once daily', 'once daily']	[]	['tablet', 'tablet']

Table 3: Example of data extraction from group description text with med7 NER model to obtain drug dosage, strength, route of administration, frequency, and duration of treatment (example from NCT04115748 study).

However, one significant limitation is to capture the relationships between detected entities. This is particularly challenging when a group is receiving multiple medications over different period (randomization, open label phase) and dosage.

Nonetheless, this open-source model allows us to fast development of a first pipeline for clinical language processing.

CONCLUSION

Open-source databases like ClinicalTrials.gov and PubMed are valuable resources to monitor protocol practices and results. However, these data are not always structured or consistent over studies and require further data extraction, standardization, cleaning, and formatting.

This emphasizes the importance of data standardization across the study and the adoption of open data practices to improve the consistency and quality of clinical trial data.

That is why leveraging machine learning models can help to extract semantic meaning and relevant information over different study notably on complex protocols.

This also underscores the importance and interest of promoting open data in clinical trials, as this could promote a more systematic meta-analysis practice and provide valuable insights for future clinical trials.

REFERENCES

[1] "ClinicalTrials.gov API Version 2.0 Now Available," [Online]. Available: https://www.nlm.nih.gov/pubs/techbull/ma24/ma24_clinicaltrials_api.html.

[2] "clinicaltrialsAPI," [Online]. Available: https://clinicaltrials.gov/data-api/api.

[3] "NCBI," [Online]. Available: https://www.ncbi.nlm.nih.gov/.

[4] "European Medecines Agency," [Online]. Available: https://www.ema.europa.eu/en/glossary-terms/adverse-event.

[5] X. D. J. Shi, "Constructing a finer-grained representation of clinical trial results from ClinicalTrials.gov," *Sci Data,* vol. 41, p. 41, 2024.

[6] "GitHub - xuanyshi/Finer-Grained-Clinical-Trial-Results," [Online]. Available: https://github.com/xuanyshi/Finer-Grained-Clinical-Trial-Results.

[7] "Github Biobert," [Online]. Available: https://github.com/dmis-lab/biobert.

[8] K. S. e. a. Yoon W, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics,* vol. 36, pp. 1234-1240, 2020.

[9] V. N. L. Q. N.-H. A. Kormilitzin A, "Med7: A transferable clinical natural language processing model for electronic health records," *Artificial Intelligence in Medicine (Artif Intell Med),* vol. 118, 2021.

[10] "Github med7," [Online]. Available: https://github.com/kormilitzin/med7.

ACKNOWLEDGMENTS

Hugo Bottois and Paulin Charliquart participated in providing and implementing the technical solutions. Andre Couturier handled project management and guided the project direction. Edward Gillian and Mukul Mittal conducted the project and documentation review.

The authors would also like to acknowledge the contributions of the Sanofi safety team members.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Hugo Bottois

Company: Sanofi

Email: hugo.bottois@sanofi.com

Location: Paris, France