## КОМП'ЮТЕРНІ НАУКИ

*UDC* 004.8 *http://dx.doi.org/10.30970/vam.2024.34.00000*

# DETECTION OF DISCREPANCIES IN BILINGUAL CLINICAL TRIAL REGISTRIES USING RULE-BASED MAPPING AND NATURAL LANGUAGE PROCESSING

**V. Horlatch, V. Pasichnyk**

*Ivan Franko National University of Lviv,*
*1, Universytetska str., 79000, Lviv, Ukraine,*
*e-mail: vitaliy.horlatch@lnu.edu.ua, vasyl.pasichnyk.apmi@lnu.edu.ua*

Accurate and consistent clinical trial registries are important elements of clinical studies. However, errors and inconsistencies in trial documentation are common, particularly when support of multiple regions is required and local regulatory authorities maintain parallel registries using different languages and data standards. In this paper, an application designed to detect discrepancies between Ukrainian and English versions of clinical trial records is presented. The system combines structured field mapping and rule-based algorithms with modern natural language processing techniques to compare unstructured text. This combination allows the identification of semantic similarity in the fields which cannot be compared using standard methods. The proposed approach was evaluated on a set of records about clinical trials in Ukraine. The system successfully identified numerous discrepancies, including inaccuracies in patient eligibility criteria. This outcome demonstrates that a combination of deterministic algorithms with large language models (LLMs) can provide significant improvements to the quality of clinical trial documentation. Ultimately, the proposed approach contributes to the overall quality and integrity of medical research documentation and creates better grounds for further training of LLMs and developments in this field.

*Key words*: clinical trial registry; data quality; discrepancy detection; large language models; natural language processing.

## 1. Introduction

Sharing information about clinical trials is a crucial task, as it helps people identify potential novel treatments for severe diseases and enables clinical research organizations (CROs) and sponsors to accelerate patients enrollment and the development of new drugs. Different organizations are committed to making this information available for a broad audience. For example, the National Institutes of Health in the United States and the State Expert Center of the Ministry of Health of Ukraine provide comprehensive details on trials that are planned, completed, or ongoing within different regions.

At the same time, ensuring the consistency and accuracy of this information is not a trivial task. For example, a study found that 16.2% of clinical trials which are registered on both ClinicalTrials.gov and the European Union registry had conflicting information about their status [1]. The reasons for discrepancies are different: missing data, variations in names, human errors, translation issues, asynchronous updates, etc. [2].

Numerous studies have explored methods for extracting clinical trial data from clinical protocols or related documents, as well as how to measure the accuracy of these extraction results [3–4]. However, there is a notable gap in the literature when it comes to comprehensive discrepancy searches – particularly those that focus on comparing Ukrainian clinical trial data with their global counterparts.

In this paper, a system designed to identify potential inconsistencies between Ukrainian and English versions of clinical trial documentation is introduced. The proposed solution uses a variety of techniques to compare values from different sources in different languages and formats. Such automated discrepancy detection can assist regulators and researchers in ensuring that the data remains accurate across different registers and sources.

## 2. Materials and methods

First, the required datasets from the State Expert Center of the Ministry of Health in Ukraine (clinicaltrials.dec.gov.ua) and the corresponding records from the National Institutes of Health (clinicaltrials.gov) were collected and uploaded into a PostgreSQL database (a part of the dataset was also published on the Figshare repository for reference and further research [5]). Then, a Python-based system for discrepancy detection was created, employing several methods to compare the records.

For elements that could be directly compared, deterministic algorithms were utilized and mapping dictionaries were created to ensure high accuracy. This approach allowed to handle straightforward fields with the highest level of certainty. In particular, the trial phase field in Ukrainian may contain values like "I", "II", "III" (Roman numerals in Ukrainian script) which were mapped to English labels "PHASE1", "PHASE2", "PHASE3", etc. Similarly, one-to-one mappings were created for gender eligibility (male, female, all).

Fields such as age, enrollment, and completion date have different representations in the Ukrainian and English versions. Therefore, more sophisticated mapping strategies were required. For example, the eight age groups that are typical in Ukrainian records had to be mapped to three age groups from the English database. For completion dates, additional parsing routines were implemented to support multiple date formats and to allow comparisons between full and abbreviated notations. For enrollment values, a tolerance parameter was added to compare the values in a specific range.

For more complex unstructured text fields, such as titles and study sponsors, which are not easily comparable, Natural Language Processing and Large Language Models (LLM) techniques were applied to detect similarities. Specifically, an OpenAI-based transformer that provides high-dimensional vector embeddings for sentences was used. Such models are trained on a large amount of data and can project semantically similar sentences in different languages to nearby points in vector space. Recent work has shown that models like these are capable of aligning texts across different languages, including English -Ukrainian clinical data matching [6].

All mapping dictionaries, as well as LLM prompts, and other parameters were stored in a separate configurable and reusable YAML file. This approach enables easy adjustment and fine-tuning. The configuration file can also be reused for similar platforms and future research. A very high-level architecture of the system is presented in Figure 1.

The pipeline was executed on the entire set of 1657 records from the Ukrainian dataset and compared them with their English versions. The solution produced a list of inconsistencies for each trial, categorized by field type. In addition to manual evaluation, an error-injection method was used to test whether the system could identify all these manually introduced discrepancies. Specifically, 376 errors were introduced (43 in age, 88 in full name, 49 in phases, 15 in sex, 58 in enrollment, 39 in completion date, 84 in sponsor fields), randomly interchanging data between studies. Introducing those errors helped to create an experimental environment close to real-world conditions.

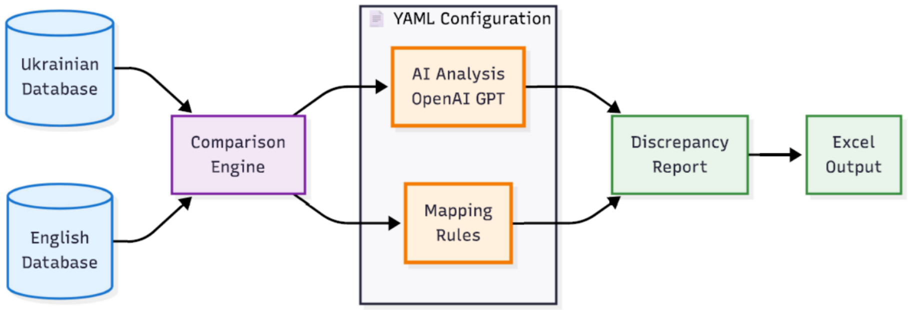

Fig. 1. High-level architecture diagram

## 3. Results

Out of 1657 records that were compared by the system, 1404 had at least one identified discrepancy. The detailed results are uploaded to the Figshare repository [7], and a summary is provided in Table 1.

*Table* 1

The summary of discrepancies

| **Field** | **Number of discrepancies (including empty fields)** | **Number of discrepancies (excluding empty fields)** |
|---|---|---|
| Age | 408 | 408 |
| Completion date | 814 | 256 |
| Enrollment (with 20% error range) | 704 | 584 |
| Phases | 119 | 109 |
| Sex | 31 | 18 |
| Sponsor | 295 | 295 |
| Title | 204 | 198 |

In some cases, the discrepancies might be reasonably justified. For example, a study may be designed for different age groups worldwide, while in Ukraine, only adults were permitted. Similarly, several phases may be listed in the English version but only one in the Ukrainian record. However, in many cases, the inconsistencies appear to be clear mistakes. Some examples of the discrepancies are presented in Table 2.

Next, 376 errors were introduced, and the script was rerun. As expected, the system detected 100% of mistakes in fields where deterministic algorithms were applied for the data comparison (age, phases, sex, enrollment, completion date). For fields where LLMs were utilized (full name and sponsor), the solution missed two errors and didn t highlight them as discrepancies. In those cases, the corresponding fields were very similar (the same therapeutic area, conditions, approach, etc.), and therefore, it proves the reliability of the approach rather than questioning it.

*Table* 2

Examples of discrepancies

| **NCT ID** | **Category** | **Original values from the Ukrainian dataset** | **Values from the international dataset** |
|---|---|---|---|
| NCT00141102 | Age | Неповнолітні (14-18 років), Літнього віку (старші 65 років), Інший віковий діапазон, Молодші 18 років | ADULT, OLDER_ADULT |
| NCT04345900 | Completion Date | 18.02.2021 | 4/20/2018 |
| NCT01307800 | Completion Date | 29.08.2012 | 2012-09 |
| NCT03558152 | Enrollment | 24.05.2018 Запланована – 270 | 195 |
| NCT01989676 | Enrollment | 29.10.2013 Запланована – 997 | 707 |
| NCT02401035 | Phases | II | PHASE4 |
| NCT01122927 | Phases | IV | PHASE3 |
| NCT02771795 | Sex | чоловіча, жіноча | FEMALE |
| NCT02448680 | Sex | чоловіча, жіноча | MALE |
| NCT01499290 | Sponsor | astrazeneca ab, sweden («астразенека аб», швеція) | pfizer |
| NCT02929329 | Sponsor | «амжен інк.» (amgen .), США | cytokinetics |

## 4. Conclusions

This paper demonstrates the importance and value of applying computer science algorithms to improve the quality of clinical trial data. The developed system identified 1404 discrepancies among 1657 clinical studies conducted in Ukraine. These findings have significant practical implications. The obtained results can be leveraged by state institutions and individual researchers to enhance data accuracy within the Ukrainian clinical trial domain. Another important outcome of this study is the configurable YAML file that was developed. This file can be utilized for further related research and integrated into other automated platforms that require mapping of Ukrainian and English clinical trial data.

This work has a high level of scientific novelty. Although some research has been conducted on identifying discrepancies between registries in the United States and the European Union, there remains a clear gap in similar studies focusing on the Ukrainian clinical data landscape. To the best of our knowledge, no similar research has been conducted previously.

In future work, the plan is to focus on the following directions. First, improve text comparison methods to achieve greater accuracy and precision when aligning unstructured clinical documentation across different sources and languages. Also, expand the number of extracted attributes and integrate additional datasets to enhance comparative analysis. Finally, the goal is to develop autonomous agentic systems based on the configurable framework introduced in this paper. These agents will be capable of continuously monitoring and improving the quality of clinical data in near real time.

## References

1. Fleminger J. Prevalence of clinical trial status discrepancies: A cross-sectional study of 10,492 trials registered on both ClinicalTrials.gov and the European Union Clinical Trials Register / J. Fleminger, B. Goldacre // PLOS ONE. – March 7, 2018. – URL: https://doi.org/10.1371/journal.pone.0193088.
2. Chaturvedi N. Some data quality issues at ClinicalTrials.gov / N. Chaturvedi, B. Mehrotra, S. Kumari, S. Gupta, H.S. Subramanya, G. Saberwal // Trials. – 2019. – Vol. 20. – URL: https://doi.org/10.1186/s13063-019-3408-2.
3. Kreimeyer K. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review / K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S.F. Jones, et al. // Journal of Biomedical Informatics. – 2017. – Vol. 73. – P. 14–29. – URL: https://doi.org/10.1016/j.jbi.2017.07.012.
4. Patricoski G. An Evaluation of Pretrained BERT Models for Comparing Semantic Similarity Across Unstructured Clinical Trial Texts / G. Patricoski, K. Kreimeyer, A. Balan, et al. // Stud. Health Technol. Inform. – 2022. Jan. – P. 18–21. – URL: https://doi.10.3233/SHTI210848.
5. Pasichnyk V. Datasets of clinical trials in Ukraine / V. Pasichnyk Figshare: Collection, 2025. – URL: https://doi.10. 6084/m9.figshare.c.7887785.v1.
6. Dyyak I. Assessing generative pre-trained transformer 4 in clinical trial inclusion criteria matching / I. Dyyak, V. Horlatch, T. Pasichnyk, V. Pasichnyk // Seventh International Workshop on Computer Modeling and Intelligent Systems. – May 3, 2024. – Zaporizhzhia. – Ukraine. – P. 305–316. – URL: https://doi.10.5281/zenodo.12636163.
7. Pasichnyk V. Discrepancies in clinical trial data / V. Pasichnyk. – Figshare, 2025. – URL: https://doi.10.6084/m9. figshare.29927594.

*Article: received* 03.09.2025
*revised* 22.09.2025
*printing adoption* 15.10.2025

## ЗАСТОСУВАННЯ ДЕТЕРМІНОВИХ АЛГОРИТМІВ ТА МЕТОДІВ ОБРОБКИ ПРИРОДНОЇ МОВИ ДЛЯ ВИЯВЛЕННЯ ПОМИЛОК У РЕЄСТРАХ КЛІНІЧНИХ ВИПРОБУВАНЬ

**В. Горлач, В. Пасічник**

*Львівський національний університет імені Івана Франка,
вул. Університетська 1, Львів, 79000, Україна
e-mail: vitaliy.horlatch@lnu.edu.ua, vasyl.pasichnyk.apmi@lnu.edu.ua*

Детальні та достовірні реєстри клінічних випробувань є надзвичайно важливими елементами у сфері клінічних досліджень. Однак помилки та розбіжності в такій документації є поширеним явищем, особливо якщо використовуються декілька мов. У цій роботі ми представляємо систему, яка виявляє розбіжності між українською та англійською версіями записів клінічних випробувань. Система поєднує алгоритми зі структурованими правилами та методи обробки природної мови для порівняння неструктурованого тексту. Ми оцінили цей підхід на наборі записів про поточні клінічні випробування в Україні. Система успішно виявила численні розбіжності, включаючи некоректні дані щодо відповідності пацієнтів критеріям дослідження. Наші результати демонструють, що поєднання детермінованих алгоритмів з великими мовними моделями може забезпечити суттєве покращення якості документації та процесу клінічних випробувань у цілому.

*Ключові слова*: реєстр клінічних випробувань; якість даних; виявлення розбіжностей; великі мовні моделі; обробка природної мови.