# $CT_{CL}$: A Cross-Language Benchmark for Matching Patients to Clinical Trials


Maciej Rybinski
Universidad de Málaga/University of New South Wales
Spain/Australia, Málaga/Sydney
maciek.rybinski@uma.es

Wojciech Kusa
NASK National Research Institute
Warsaw, Poland
wojciech.kusa@nask.pl

Necva Bölücü
CSIRO
Sydney, Australia
necva.bolucu@csiro.au

Georgios Peikos
University of Milano-Bicocca
Milan, Italy
georgios.peikos@unimib.it

Aditya Joshi
University of New South Wales
Sydney, Australia
aditya.joshi@unsw.edu.au

Sarvnaz Karimi
CSIRO/Monash University
Sydney/Melbourne, Australia
sarvnaz.karimi@csiro.au

Aitziber Atutxa
University of the Basque Country (UPV/EHU)
Bilbao, Spain
aitziber.atucha@ehu.eus

Javier Del Ser
TECNALIA/University of the Basque Country (UPV/EHU)
Bilbao, Spain
javier.delser@tecnalia.com

Ahmet Bölücü
Ministry of Health of Turkey
Ankara, Turkey
ahmet.bolucu@saglik.gov.tr

Monica Chierichetti
Independent Researcher
Milan, Italy
monica.chierichetti@gmail.com

Pritam Dasgupta
St John of God Health Care
Perth, Australia
pritam.dasgupta@sjog.org.au

Nicolás Jiménez García
Hospital Universitario Costa del Sol
Marbella, Spain
nicolas.jimenez.g.sspa@juntadeandalucia.es

Borja Pedruzo
Biobizkaia Health Research Institute
Bilbao, Spain
borja.pedruzobagazgoitia@osakidetza.eus

Angelika Romańska
Independent Researcher
Warsaw, Poland
romanskaangelika@gmail.com

Ioulia Symeonidou
University of Nicosia Medical School
Nicosia, Cyprus
symejul@gmail.com



## Abstract

The success of clinical trials depends on the recruitment of patients who match strict inclusion criteria. The development of effective patient to clinical trial matching systems depends on benchmarking datasets that support systematic evaluation. Apart from resources that are created in English by the TREC Clinical trials tracks (2021-23), very limited corpora exist for other languages and cross-language settings, despite the need of automatic support for clinical trial recruitment being global. To address this gap, we combine machine translation with medical expert annotation to construct $CT_{CL}$ (Clinical Trials Cross Lingual retrieval), a cross-lingual evaluation benchmark for patient-clinical trial retrieval in seven languages. We benchmark the cross-lingual retrieval task using 14 large language (embedding) models. We showcase how our dataset can be used to evaluate the cross-lingual capability of the models for languages with varying resource availability.




*SIGIR '26, Melbourne, VIC, Australia*



ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808620



## CCS Concepts

• **Applied computing** → **Health informatics**; • **Information systems** → **Evaluation of retrieval results**.

## Keywords

Language models, Clinical trials search, Medical IR, Evaluation, Cross-lingual IR

**ACM Reference Format:**
Maciej Rybinski, Wojciech Kusa, Necva Bölücü, Georgios Peikos, Aditya Joshi, Sarvnaz Karimi, Aitziber Atutxa, Javier Del Ser, Ahmet Bölücü, Monica Chierichetti, Pritam Dasgupta, Nicolás Jiménez García, Borja Pedruzo, Angelika Romańska, and Ioulia Symeonidou. 2026. $CT_{CL}$: A Cross-Language Benchmark for Matching Patients to Clinical Trials. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3805712.3808620


## 1 Introduction

Advances in medicine are largely dependent on knowledge derived from clinical trials (CTs) [21]. However, a significant impediment to the success of these trials is the enrollment of patients, which often fails to meet required recruitment targets [12, 15, 21]. The recruitment phase is laborious and involves the meticulous screening of candidates for eligibility [4]. The (semi) automation of this

process has the dual advantage of expediting translational science (i.e., shortening the path for new drugs to reach the market) [12] and facilitating a broader patient access to novel treatments and procedures via CTs appropriate for individual patients [8].

Automated systems are proposed to streamline the screening process by matching eligible patients with suitable CTs. In the literature, the problem is often modeled as a retrieval (ranking) task [8, 23, 25, 31]. The matching itself involves analysing patient data (often in the form of unstructured patient notes, i.e., medical narratives) with the multifaceted clinical trial documents [9, 18, 29, 30] to assess patient eligibility.

While searching electronic health records (EHRs) is already used in practice to bolster CT recruitment, experts responsible for patient screening point to several barriers reducing the effectiveness of the semi-automated approach [15]. Some of the barriers include data governance fragmentation, regulatory issues, and technical shortcomings. It is plausible that many regulatory and governance obstacles can be overcome through initiatives such as the implementation of the European Health Data Space (EHDS). EHDS is expected to provide patients with a regulatory framework and infrastructure to govern and share their healthcare data, including EHRs.

While providing a regulatory and technical framework for patient data sharing is an important step towards more effective secondary use of patient records, its implementation in a *federated* healthcare system inevitably brings new technical challenges. One of such challenges relates to cross-language settings implied by such a federated model. We note that EHDS and other similar initiatives will provide access to patient records in multiple languages. For the task we focus on in this paper—matching patients to clinical trials—the patient records in multiple languages would have to be matched to a clinical trial specification in another language (e.g., English). However, there are no cross-language benchmarks for clinical trials matching currently available to the IR community, and all the existing collections cover a setup where patient notes in English are matched to clinical trials in English. Evaluating the effectiveness of systems for CT recruitment in cross-language settings is crucial in allowing the IR community to carry out research towards equitable solutions for healthcare use.

In this paper we address this gap by introducing $\text{CT}_{CL}$, a benchmark collection for matching patients to clinical trials, where the task is to match patient records in languages other than English (i.e., Basque, Bengali, Greek, Italian, Polish, Spanish, and Turkish) to clinical trials specifications (i.e, their descriptions and inclusion criteria)[1]. Our collection builds on an established English-only patient–trials matching benchmark, *TREC Clinical Trials 2021* (henceforth, TREC CT 2021)[2]. While these languages are described as mid- to high-resource in NLP [35], we hypothesise that they benefit from varying degrees of biomedical resource availability and should therefore yield more nuanced results.

We have compiled our collection by translating the patient notes from TREC CT 2021 from English into the respective target languages, and validating and correcting these translations with the help of health professionals proficient in these languages. These translated topics can be used alongside the (English-language) document corpus and human judgments sourced from TREC CT 2021 (thus producing cross-lingual settings). We illustrate how our dataset can be used to carry out cross-language IR (CLIR) experiments by evaluating a selection of state-of-the-art multi-lingual dense retrieval models and comparing them against a strong (but computationally expensive) translate-and-match baseline. Code of all our experiments is available at https://github.com/maciekrybinski/CL_clinical_trials.

[1]Our dataset is available at https://zenodo.org/records/18503691
[2]https://www.trec-cds.org/2021.html

## 2 Related Work

*Clinical Trial Search.* The success of clinical trials depends on matching complex inclusion and exclusion criteria against potential participants [5, 6, 10, 13, 19]. From 2017 to 2022, Precision Medicine and Clinical Trials tracks in TREC [23, 24, 26, 27, 33] focused on this matching problem. Different methods resulted from these tracks, such as a multi-stage neural ranking system [20], Trial2Vec [37], BERT-based reranking methods [30] and search prototypes such as Science2Cure [28] or TrialGPT [7]. One major problem is that, apart from the data provided by the TREC and a couple of other English variants (e.g., [8], or Leaf Clinical Trials [2]), there are no major benchmarks and datasets that could be leveraged for machine-learning-based ranking, especially for non-English languages. We develop a resource that addresses this gap.

*Cross-Language Information Retrieval (CLIR).* It has been studied within the IR and NLP communities for decades. CLIR was introduced in the 90s as a retrieval task where the query is in one language, and the retrieved documents are in another, with machine translation playing a key role [1, 14]. In 2012, Zhou et al. [39] surveyed and categorised how translation models can be integrated into a CLIR system. While the fundamental ideas remain relevant to date, the advances in machine translation have made it much more usable. Most recently, in 2025 Wang et al. [36] proposed a novel framework that focused on low-performing languages by aligning their internal representations with those of high-performing languages during inference.

*Machine Translation of Medical Text.* In high-stakes applications such as health, machine translation quality determines the usability and deployment of technology. Pecina et al. [17] investigate query translation quality for medical queries in a CLIR system for three languages, German, Czech and French, into English using statistical machine translation methods. More recent studies investigate neural machine translation. In a low-resource language such as Vietnamese, Toan et al. [34] show that small pretrained language models, such as Qwen, can provide accurate and efficient domain-specific translations. In our study, we also consider a similar approach in translation for clinical text.

## 3 $\text{CT}_{CL}$ dataset creation

We create $\text{CT}_{CL}$ by translating and validating patient topics from TREC CT 2021 into multiple languages. We outline the dataset (Subsection 3.1), our translation and validation process (Subsection 3.4), and the rationale for choosing the 2021 collection over later TREC CT releases (Subsection 3.2).

**Topic 49:** A 12 year old girl came to the clinic with her mother, complaining of short stature, delayed in puberty and developmental delay. Her karyotype study revealed 45X and confirmed the diagnosis of Turner syndrome. She is treating with GH since 6 months ago without estrogen therapy to avoid menarche and reach the ideal height. She is an obese, mentally retarded girl in the physical exam. Her breast bulb were in stage 1 with no course hair in the pubic or axillary. Her TSH was 3 and FBS was 75 in the latest lab study.

**Figure 1: Example topic from the TREC CT 2021.**

## 3.1 TREC CT 2021

The source dataset we use to create $CT_{CL}$, TREC CT 2021, consists of 75 patient topics with 35,832 relevance judgments and uses as its corpus a 2020 snapshot of the ClinicalTrials.gov registry[3] as its corpus, containing over 375K clinical trials.

Clinical trials are *(semi) structured documents* with the following fields: brief summary, brief title, identifier, detailed description, drug name, drug keywords, inclusion/exclusion criteria, gender, general keywords, intervention type, maximum age, minimum age, official title, and primary outcome. Intervention type, gender, and primary outcome refer to controlled vocabularies; age-related fields are numeric. All other fields except the clinical trial ID are textual.

Each topic simulates a patient admission note, i.e., lengthy topics spanning several sentences (e.g., Figure 1). For each evaluated topic–document pair, a relevance judgment assigns a score of 0 for *not relevant*, 1 for *excluded* (the patient meets the inclusion criteria but is ruled out by one or more exclusion criteria), and 2 for *eligible* (the patient meets the inclusion criteria and none of the exclusion criteria apply).

## 3.2 On the choice of source dataset

The TREC CT track provides three topic formats across its 2021 to 2023 editions. The 2021 and 2022 editions formulate topics as free text medical narratives (admission notes), while the 2023 edition formulates topics as semi structured questionnaires. We restrict our experiments to the 2021 and 2022 editions because retrieval from free text clinical narratives is typically harder for cross language methods due to abbreviations, shorthand, and specialized terminology, and it better matches the unstructured text encountered in real clinical documentation. We opt to use the 2021 dataset because our preliminary experiments with English-language retrieval showed that evaluation scores are less affected by unjudged documents in 2021 than in 2022. This indicates that the 2021 relevance judgments are more complete, likely because its judgment pool was built from a larger and more diverse set of participating systems. As a result, 2021 provides a more reliable basis for cross-lingual experiments.

## 3.3 Data

Our dataset includes the seven topic files containing translated and validated patient notes (one file for each of the target languages).

[3]http://clinicaltrials.gov

INSTRUCTIONS
Given a patient note automatically translated into [TARGET LANGUAGE] and its original version (in English), the next step is to correct the translated version (or "validate" it, confirming that the machine translation is correct). The goal of the corrections is to replace untranslated (or poorly translated) terms with terms used in [TARGET LANGUAGE] medical practice and to ensure that the translated notes are grammatically correct.
Examples of terms that often result in correctable translations include: abbreviations (e.g., RLE), names of procedures, diseases, and symptoms, body parts, shorthand (e.g., s/p), and drug names (e.g., use of the popular brand name in the US instead of the name used in [TARGET LANGUAGE]).

**Figure 2: Instructions for correcting automatically generated translations.**

Since documents and topic-document relevance judgments are *inherited* directly from the original TREC CT 2021, these can be accessed via `ir_datasets` python library, which supports the original dataset. We provide translated topic files in the dataset repository (https://zenodo.org/records/18503691). The Github repository of our dataset (https://github.com/maciekrybinski/CL_clinical_trials) additionally contains code examples, a quickstart guide, and code to replicate the experiments presented below.

## 3.4 Translation and correction

Each patient note (corresponding to a single topic of TREC CT 2021) was translated automatically to the target language. It was then reviewed and corrected by medical practitioners who are native speakers of the target language. All our annotators are practising clinicians, either having obtained their postgraduate specialist degree or being close to obtaining one. In particular, we prompted the reviewers to pay attention to terms that were not translated by the automatic translation or appear to be translated incorrectly. We further highlighted that automatic translation errors, apart from grammatical errors, often originate from the use of shorthand notation, abbreviations, and names of drugs, diseases, procedures, and symptoms. In the content shown to reviewers, together with the instruction and the automatically translated note, we included an original (English) note for reference. Instructions for reviewers (in English) are presented in Figure 2.

The format of the content displayed to medical professionals correcting the notes was agreed on individually. For example, for Spanish, we used separate text files repeating the instructions, while for other languages (e.g., Polish, Greek, Bengali), we used a shared spreadsheet with all the notes and instructions shown on a single screen.

Initial automatic translations were sourced from Google Translate[4] for all languages except for Basque. For Basque, we obtained the automatic translations by using es2eu [3] [5] based on MarianMT

[4]https://translate.google.com/; last accessed in December 2025
[5]https://huggingface.co/HiTZ/medical_es-eu

applied to corrected Spanish versions to ensure a better quality of the automatic translation process [3], and to consequently present our annotator with a simpler task[6].

We report the quality of initial automatic translations with respect to reference (i.e., corrected) translations across languages, using standard machine translation metrics. We report BLEU [16], TER [32], and COMET[7] [22] scores in Table 1. Spanish, Italian, Greek, and Turkish achieve high BLEU and COMET with low TER, indicating strong translation quality. Basque translation also scored highly (meaning it required few corrections), but it is worth noting that it was carried out using a specialised medical translation model. Polish shows moderate BLEU and higher TER, reflecting morphological complexity, while Bengali scores lowest, highlighting challenges in lower-resource clinical translation. Overall, automatic translations are reliable for well-resourced languages, with caution needed for Bengali.

**Table 1: BLEU, TER, and COMET scores calculated for automatically translated patient notes and their corrected versions for each of the languages.**

| Language | BLEU (↑) | TER (↓) | COMET (↑) |
|---|---|---|---|
| Spanish (ES) | 0.93 | 6.15 | 0.87 |
| Italian (IT) | 0.92 | 6.34 | 0.89 |
| Polish (PL) | 0.75 | 22.84 | 0.90 |
| Turkish (TR) | 0.89 | 8.07 | 0.89 |
| Bengali (BN) | 0.14 | 74.01 | 0.78 |
| Greek (EL) | 0.93 | 6.49 | 0.91 |
| Basque (EU) | 0.96 | 3.14 | 0.90 |

Figure 3 shows an initial automatic Spanish translation (top) and its corrected version (bottom; corrected terms were highlighted) for topic 49 (original version of which is presented in Figure 1).

Self-reported time spent on correction and verification ranged from 8 hours to around 30 hours, with a median reported time close to 12 hours (for all 75 notes).

**Topic 49, automatic translation:** Una niña de 12 años acudió a la clínica con su madre quejándose de baja estatura, retraso en la pubertad y retraso en el desarrollo. Su estudio de cariotipo reveló 45X y confirmó el diagnóstico de síndrome de Turner. Está en tratamiento con GH desde hace 6 meses sin terapia estrogénica para evitar la menarquia y alcanzar la altura ideal. En el examen físico se observa que es una niña obesa y con retraso mental. Su bulbo mamario estaba en etapa 1 sin vello grueso en el pubis ni en la axila. En el último estudio de laboratorio su TSH era 3 y su FBS era 75.

**Topic 49, corrected:** Una niña de 12 años acudió a la clínica con su madre quejándose de baja estatura, retraso en la pubertad y retraso en el desarrollo. Su estudio de cariotipo reveló 45X y confirmó el diagnóstico de síndrome de Turner. Está en tratamiento con GH desde hace 6 meses sin terapia estrogénica para evitar la menarquia y alcanzar la altura ideal. En el examen físico se observa que es una niña obesa y con retraso mental. Su botón mamario estaba en etapa 1 sin vello grueso en el pubis ni en la axila. En el último estudio de laboratorio su TSH era 3 y su glucemia basal en ayunas era 75.

**Figure 3: An example topic (topic 49) from the TREC CT 2021 translated automatically to Spanish and corrected by a medical professional.**

## 4 Experiments

As we mentioned, the main focus of experiments presented here is dense retrieval using LLM-based embedding models. When applied directly in cross-language settings we vectorise the queries (in one of the languages of $CT_{CL}$) and evaluate and evaluate their similarity to the embeddings of the CT corpus (in English) to create document rankings for each topic. In translate-and-retrieve experiments the patient note is translated to English prior to retrieval. Details of the setup are presented further in this section.

The experiment with multilingual dense retrievers is motivated by the fact that these models represent an efficient solution to matching patients to clinical trials in cross-lingual settings. In such a setup, each text (i.e., a patient note, or a clinical trial description) is *vectorised* only once. We, therefore, investigate dense retrievers as a straightforward baseline for the cross-lingual CT retrieval task we introduce with $CT_{CL}$.

We note, however, that our dataset follows the setup of TREC CT 2021, where a CT database is searched using a patient note (i.e., the topic). This (patient searching for trials) setup could represent a scenario of automatically finding suitable trials for a given patient in point-of-care settings. Of note, a clinical trial recruitment process could well consist of searching a patient database with a specific clinical trial description (so, a trial searching for patients). This two-way nature of the core problem (where, in practice, both trial descriptions and patient notes could be expressed in different languages) highlights the appeal of multilingual embedding models for the patient-trial matching task.

### 4.1 Experimental Setup

*Models.* We benchmark 14 instruction-tuned dense retrieval models. Our primary evaluation includes the top-8 open dense retrieval models from MTEB [11] multilingual leaderboard[8]: Qwen3-Embedding-8B, Qwen3-Embedding-4B, inf-retriever-v1, jina-embeddings-v4, Qwen3-Embedding-0.6B, inf-retriever-v1-1.5b, gte-Qwen2-7B-instruct, and gte-Qwen2-1.5B-instruct. These models account for all top-10 scores on the MTEB multilingual leaderboard attained by openly available models on retrieval tasks.

We further extend the experiments with six models that account for the remaining top-10 scores on the multilingual medical retrieval

[6]We confirmed the effectiveness boost by manual evaluation on a handful of randomly sampled instances

[7]Via evaluate library.

[8]As of August 2025; https://huggingface.co/spaces/mteb/leaderboard

subset of MTEB: NV-Embed-v2, NV-Embed-v1, SFR-Embedding-Mistral, SFR-Embedding-2_R, Linq-Embed-Mistral, and e5-mistal-7b-instruct.

All models were evaluated using the prompt: *'Given a patient note, find clinical trials the patient is eligible for.'*. For some models, we appended additional tokens, as per respective Huggingface model cards (e.g., end-of-sequence tokens, instruction prefix, etc.).

*Indexing and Retrieval Pipeline.* Our experiments were implemented in Python, with indexing performed using Sentence Transformers[9] library. We accessed document content for indexing using ir_datasets[10] package and indexed a concatenation of title, summary, and eligibility (criteria) fields. Dense vectors were indexed using a flat FAISS[11] index, which was also used for similarity search.

All models were evaluated using 4-bit quantisation via BitsAndBytes to enable reproducibility of our experiments on commodity GPUs. The only exceptions are NVIDIA's NV-Embed models, which do not appear to support quantisation and were therefore run in half precision (FP16) mode. All GPU experiments ran on NVIDIA A100 GPUs. Full implementation details are available in our GitHub repository.

*Evaluation Metrics.* In our experiments, for each set of topics (i.e., for each language), we follow the evaluation procedure outlined by the TREC CT 2021 organisers. Each system is evaluated based on a ranking of 1000 clinical trials per topic. We report the main evaluation measures used in the official TREC CT 2021 evaluations: P@10, RR, nDCG@10.

*Other Baselines.* We compare dense retrievers in cross-lingual setting against two other baselines: (i) BM25 and (ii) Qwen3-Embedding-4B (the best performing model among the evaluated group), both evaluated in a translate-and-retrieve setting.

For translate-and-retrieve baselines, patient notes are translated into English (from one of the seven languages of $CT_{CL}$) using NLLB-200-3.3B. Topics translated to English are then used directly as queries (so, either for the BM25 retrieval, or for the dense retrieval). We chose NLLB-200-3.3B, as it covers all of our languages and is reported to provide reasonably strong results in domain-specific evaluations [38]. Translation is performed sentence-by-sentence in FP16; sentence segmentation uses NLTK for European languages, while '|' character is used as a sentence delimiter for Bengali. This pipeline effectively corresponds to a back-translation procedure with an intermediate manual medical correction step, whose implications we discuss in Section 5. For BM25, we use the CT 2021 index and BM25 ranking implemented in Pyserini[12], for easy reproducibility of our results.

*Statistical significance.* We evaluate the statistical significance of differences between target language and English results yielded by respective models using a paired t-test with Bonferroni correction, wherever comparisons are made.

**Table 2: Results reported for all languages aggregated across all 14 dense retrieval models. These runs implement direct ranking with cross-lingual embeddings. 'Best' denotes results obtained by Qwen3-Embedding-4B. $^{\dagger}$ indicates a significant difference of the best model effectiveness vs. its English results (evaluated with a paired t-test, $p < 0.0029$, Bonferroni-corrected).**

| Lang | Agg. | nDCG@10 (↑) | P@10 (↑) | RR (↑) |
|---|---|---|---|---|
| EN | avg | 0.382 ± 0.06 | 0.257 ± 0.045 | 0.475 ± 0.083 |
| | max | 0.473 | 0.319 | 0.629 |
| | min | 0.227 | 0.148 | 0.296 |
| | best | 0.472 | 0.318 | 0.545 |
| ES | avg | 0.328 ± 0.061 | 0.22 ± 0.045 | 0.43 ± 0.086 |
| | max | 0.443 | 0.305 | 0.592 |
| | min | 0.236 | 0.153 | 0.31 |
| | best | 0.443 | 0.307 | 0.589 |
| IT | avg | 0.319 ± 0.057 | 0.213 ± 0.048 | 0.4 ± 0.069 |
| | max | 0.427 | 0.300 | 0.542 |
| | min | 0.242 | 0.141 | 0.298 |
| | best | 0.425$^{\dagger}$ | 0.297 | 0.538 |
| PL | avg | 0.272 ± 0.078 | 0.182 ± 0.06 | 0.356 ± 0.097 |
| | max | 0.396 | 0.276 | 0.495 |
| | min | 0.145 | 0.071 | 0.202 |
| | best | 0.398 | 0.279 | 0.477 |
| TR | avg | 0.263 ± 0.089 | 0.178 ± 0.062 | 0.338 ± 0.112 |
| | max | 0.43 | 0.281 | 0.555 |
| | min | 0.127 | 0.069 | 0.169 |
| | best | 0.429 | 0.281 | 0.552 |
| BN | avg | 0.284 ± 0.088 | 0.184 ± 0.058 | 0.365 ± 0.099 |
| | max | 0.428 | 0.279 | 0.531 |
| | min | 0.155 | 0.107 | 0.234 |
| | best | 0.418$^{\dagger}$ | 0.277$^{\dagger}$ | 0.515 |
| EL | avg | 0.168 ± 0.095 | 0.112 ± 0.066 | 0.231 ± 0.114 |
| | max | 0.362 | 0.259 | 0.469 |
| | min | 0.049 | 0.028 | 0.09 |
| | best | 0.360$^{\dagger}$ | 0.235$^{\dagger}$ | 0.471 |
| EU | avg | 0.184 ± 0.08 | 0.117 ± 0.059 | 0.243 ± 0.097 |
| | max | 0.353 | 0.251 | 0.415 |
| | min | 0.078 | 0.045 | 0.078 |
| | best | 0.322$^{\dagger}$ | 0.211$^{\dagger}$ | 0.405 |

## 5 Results and Discussion

Table 2 shows the aggregated results[13] for each language in our dataset, alongside English–English retrieval results for reference. The best retriever in our evaluation is Qwen3-Embedding-4B (best NDCG@10 for 4 languages, second-best for the remaining 3), reported as *best* in the same table.

[9] https://pypi.org/project/sentence-transformers
[10] https://ir-datasets.com
[11] https://faiss.ai
[12] https://github.com/castorini/pyserini
[13] Due to the scale of our evaluation (14 models, 7 languages), we report only aggregated results per language (i.e., averages, max, and min values across all models, averaged over all topics, for each model). We also report detailed results for the best-performing model.

**Table 3: Translate-and-retrieve baseline with BM25 and Qwen3-Embedding-4B ranking.**

| | BM25 | | | Qwen3-Embedding-4B | | |
|---|---|---|---|---|---|---|
| **Lang** | **NDCG@10** | **P@10** | **RR** | **NDCG@10** | **P@10** | **RR** |
| EN | 0.29 | 0.16 | 0.32 | 0.47 | 0.32 | 0.54 |
| ES | 0.28 | 0.15 | 0.34 | 0.45 | 0.32 | 0.49 |
| IT | 0.28 | 0.15 | 0.34 | 0.46 | 0.31 | 0.53 |
| PL | 0.24 | 0.13 | 0.28 | 0.42 | 0.28 | 0.50 |
| TR | 0.28 | 0.15 | 0.30 | 0.45 | 0.31 | 0.53 |
| BN | 0.25 | 0.13 | 0.30 | 0.46 | 0.30 | 0.53 |
| EL | 0.24 | 0.13 | 0.29 | 0.42 | 0.29 | 0.50 |
| EU | 0.26 | 0.14 | 0.31 | 0.42 | 0.30 | 0.50 |

Cross-lingual retrieval effectiveness is reasonably good for Spanish and Italian, comparable to English-only retrieval. For Turkish, the strongest models also yield comparable results; however, lower average and minimum values indicate that fewer models handle Turkish–English retrieval well. Polish and Bengali follow a similar , with competitive performance among the best-performing models but reduced effectiveness on average. In contrast, effectiveness is consistently lower across all models for Greek and Basque.

Table 3 reports results of the translate-and-retrieve baselines using BM25 and Qwen3-Embedding-4B. BM25 (at a comparable computational cost, since machine translations cost factors in against query embedding cost) performance is generally comparable to the average dense retriever performance for most languages, though Greek and Basque perform worse than BM25 on average. In contrast, Qwen3-Embedding-4B consistently outperforms BM25 across all languages (see 'best' in Table 2).

Comparing direct cross-lingual retrieval ('best' in Table 2) to the translate-and-retrieve approach with Qwen3-Embedding-4B shows similar effectiveness for most languages, indicating that direct retrieval achieves comparable results at roughly half the computational cost. The exceptions are Greek and Basque, where translate-and-retrieve yields higher effectiveness, reflecting the challenge of lower-resource cross-lingual setting.

## 6 Conclusions

In summary, our experiments provide strong baselines for the introduced dataset (and viable first-step retrieval strategies for languages with different resource availability). Benchmarking covers best results across all experiments for particular metrics, as well as strong single model retrievers (i.e., Qwen3-Embedding-4B in direct and translate-and-retrieve setups). Notably, results for Qwen3-Embedding-4B for Spanish and Italian are comparable to, or surpass, previously described first-step retrieval methods for the original CT 2021 task (namely, for monolingual English retrieval). Our results also point to several promising avenues for future work, e.g., closing the effectiveness gap between direct cross-lingual retrieval and translate-and-retrieve for low resource languages, or investigating the impact of either strategy on systems with multiple ranking steps (i.e., re-ranking, in cross-lingual and translate-and-retrieve settings). The GitHub repository for $CT_{CL}$ features code to reproduce our experiments, which can be a good starting point for further follow-up research. While the basics of working with the dataset are simple (topics in new languages share an identical XML format with the original ones; corpus and human judgements are available through both TREC and ir_datasets), we also supply simple *how-to* code examples in the GitHub repository.

## 7 Limitations and Future Work

One limitation of our methodology is the use of a single annotator per language for correcting automatically translated notes. However, we note that since we opted for highly specialised and experienced annotators, the resulting resources can be viewed as medical notes that pass as plausible in the target languages. Using a larger annotator cohort could possibly help in mitigating possible biases of using an automated translation as the preliminary step.

Another limitation stems from the aforementioned use of automatic translation and also relates to the best use of highly qualified annotators. Translating notes from scratch would require a heavier workload than correction and validation, and its time-cost could prove prohibitive. As a result, we believe that our dataset is better suited to experiments with cross-lingual models than to experiments with translate-and-retrieve approaches, where the impact of double automatic translation (with a manual correction step in-between translations) cannot be ruled out.

A natural extension of this work would be addition of more languages. Another interesting avenue for further IR research in the cross–lingual CT matching could be experiments with neural re-ranking with $CT_{CL}$.

## Acknowledgements

MR's contribution was supported by EMERGIA research grant (Emergia23-00364), awarded by *Consejería de Universidad, Investigación e Innovación* of *Junta de Andalucía (Spain).*

## References

[1] Mark Davis. 1996. Advances in Multilingual Text Retrieval. In *TIPSTER TEXT PROGRAM PHASE II Proceedings*. Association for Computational Linguistics, Vienna, Virginia, USA, 185–194. doi:10.3115/1119018.1119057

[2] Nicholas Dobbins, Tony Mullen, Özlem Uzuner, and Meliha Yetisgen. 2022. The Leaf Clinical Trials Corpus: a new resource for query generation from clinical trial eligibility criteria. *Sci Data* 9, 490 (2022). doi:10.1038/s41597-022-01521-0

[3] Ane García Domingo-Aldama, Irune Palacios, Maitane Urruela, Iker De la Iglesia, Ander Barrena, and Josu Goikoetxea. 2025. EuMediCS - Euskarazko Medikuntzaren Domeinuko Corpus Sintetikoa, Itzultzaile Automatikoen Ekarpena. In *G. Domingo-Aldama, A., Palacios, I., Urruela, M., De la Iglesia, I., Barrena, A., & Goikoetxea, J. (2025). EuMediCS - Euskarazko Medikuntzaren Domeinuko Corpus Sintetikoa, Itzultzaile Automatikoen Ekarpena. IkerGazte. Nazioarteko Ikerketa Euskaraz, 3, 173–180.*

[4] Peter J Embi, Anil Jain, and C Martin Harris. 2008. Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey. *BMC medical informatics and decision making* 8, 1 (2008), 1–8.

[5] David B Fogel. 2018. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemporary clinical trials communications* (2018), 156–164.

[6] Hamed Hassanzadeh, Sarvnaz Karimi, and Anthony Nguyen. 2020. Matching patients to clinical trials using semantically enriched document representation. *Journal of Biomedical Informatics* 105 (2020), 103406.

[7] Qiao Jin, Zifeng Wang, Charalampos Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. 2024. Matching patients to clinical trials with large language models. *Nature Communications* 15, 1 (2024), 9074. doi:10.1038/s41467-024-53081-z

[8] Bevan Koopman and Guido Zuccon. 2016. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 669–672.

[9] Wojciech Kusa, Óscar E. Mendoza, Petr Knoth, Gabriella Pasi, and Allan Hanbury. 2023. Effective matching of patients to clinical trials using entity extraction and neural re-ranking. *Journal of biomedical informatics* 144 (2023), 104444.

[10] Wojciech Kusa, Patrick Styll, Maximilian Seeliger, Oscar E. Mendoza, and Allan Hanbury. 2023. DoSSIER at TREC 2023 Clinical Trials Track. In *Proceedings of the Thirty-Second Text REtrieval Conference (TREC 2023)*.

[11] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. doi:10.18653/v1/2023.eacl-main.148

[12] Yizhao Ni, Stephanie Kennebeck, Judith W Dexheimer, Constance M McAneney, Huaxiu Tang, Todd Lingren, Qi Li, Haijun Zhai, and Imre Solti. 2015. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. *Journal of the American Medical Informatics Association* 22, 1 (2015), 166–178.

[13] Jill M Novitzke. 2008. The significance of clinical trials. *J Vasc Interv Neurol* 1, 1 (Jan. 2008), 31.

[14] Douglas Oard and Bonnie Dorr. 1996. *A survey of multilingual text retrieval*. Technical Report. USA.

[15] Emily O'Brien, Sudha Raman, Alicia Ellis, Bradley Hammill, Lisa Berdan, Tyrus Rorick, Salim Janmohamed, Zachary Lampron, Adrian Hernandez, and Lesley Curtis. 2021. The use of electronic health records for recruitment in clinical trials: a mixed methods analysis of the Harmony Outcomes Electronic Health Record Ancillary Study. *Trials* 1, 22 (2021), 465. doi:10.1186/s13063-021-05397-0

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. doi:10.3115/1073083.1073135

[17] Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth J. F. Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, Martin Popel, Rudolf Rosa, Aleš Tamchyna, and Zdeňka Urešová. 2014. Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial Intelligence in Medicine* 61, 3 (2014), 165––185. doi:10.1016/j.artmed.2014.01.004

[18] Georgios Peikos, Pranav Kasela, and Gabriella Pasi. 2024. Leveraging Large Language Models for Medical Information Extraction and Query Generation. In *2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. 367–372. doi:10.1109/WI-IAT62293.2024.00058

[19] Georgios Peikos, Gabriella Pasi, et al. 2025. A Decision-Theoretic Framework to Multidimensional Relevance Estimation. *International Journal of Information Technology & Decision Making* (2025), 1–43.

[20] Ronak Pradeep, Yilin Li, Yuetong Wang, and Jimmy Lin. 2022. Neural Query Synthesis and Domain-Specific Ranking Templates for Multi-Stage Clinical Trial Matching. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Madrid, Spain, 2325–2330. doi:10.1145/3477495.3531853

[21] Taylor Pressler, Po-Yin Yen, Jing Ding, Jianhua Liu, Peter Embi, and Philip Payne. 2012. Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-screening tools. *BMC Medical Informatics and Decision Making* 12 (2012), 47.

[22] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. doi:10.18653/v1/2020.emnlp-main.213

[23] Kirk Roberts, Dina Demner-Fushman, Ellen Voorhees, William R. Hersh, Steven Bedrick, Alexander Lazar, and Shubham Pant. 2017. Overview of the TREC 2017 Precision Medicine Track. In *TREC*. Gaithersburg, MD.

[24] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, Steven Bedrick, and William R Hersh. 2020. Overview of the TREC 2020 Precision Medicine Track. In *TREC*.

[25] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, Steven Bedrick, and Willian R Hersh. 2021. Overview of the TREC 2021 Clinical Trials Track. In *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021)*.

[26] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, and Alexander J. Lazar. 2018. Overview of the TREC 2018 Precision Medicine Track. In *TREC*. Gaithersburg, MD.

[27] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, Shubham Pant, and Funda Meric-Bernstam. 2019. Overview of the TREC 2019 Precision Medicine Track. In *TREC*. Gaithersburg, MD.

[28] Maciej Rybinski, Sarvnaz Karimi, and Aleney Khoo. 2021. Science2Cure: A Clinical Trial Search Prototype. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2620–2624.

[29] Maciej Rybinski, Wojciech Kusa, Sarvnaz Karimi, and Allan Hanbury. 2024. Learning to match patients to clinical trials using large language models. *Journal of Biomedical Informatics* 159 (2024), 104734.

[30] Maciej Rybinski, Jerry Xu, and Sarvnaz Karimi. 2020. Clinical trial search: Using biomedical language understanding models for re-ranking. *Journal of Biomedical Informatics* 109 (2020), 103530.

[31] Chaitanya Shivade, Courtney Hebert, Marcelo Lopetegui, Marie-Catherine De Marneffe, Eric Fosler-Lussier, and Albert M Lai. 2015. Textual inference for eligibility criteria resolution in clinical trials. *Journal of biomedical informatics* 58 (2015), S211–S218.

[32] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2026. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. https://aclanthology.org/2006.amta-papers.25/

[33] Ian Soboroff. 2021. Overview of TREC 2021. In *TREC*.

[34] Phan Minh Toan, Nguyen Xuan Phi, Nguyen Van Tai, Trang Minh Quang, and Dang Van Thin. 2025. Bosch@AI_Team at MMT 2025: Medical Machine Translation by Bidirectional Training with Small Language Models. In *Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing*, Luong Chi Mai, Nguyen Thi Minh Huyen, and Nguyen Thi Thu Trang (Eds.). Hanoi, Vietnam, 393–400. https://aclanthology.org/2025.vlsp-1.47/

[35] Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 15894–15939. doi:10.18653/v1/2024.acl-long.845

[36] Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025. Bridging the Language Gaps in Large Language Models with Inference-Time Cross-Lingual Intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Vienna, Austria, 5418–5433. doi:10.18653/v1/2025.acl-long.270

[37] Zifeng Wang and Jimeng Sun. 2022. Trial2Vec: Zero-Shot Clinical Trial Document Similarity Search using Self-Supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates, 6377–6390. https://aclanthology.org/2022.findings-emnlp.476

[38] Aman Kassahun Wassie, Mahdi Molaei, and Yasmin Moslem. 2024. Domain-specific translation with open-source large language models: Resource-oriented analysis. *arXiv preprint arXiv:2412.05862* (2024).

[39] Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. 2012. Translation techniques in cross-language information retrieval. *ACM Computung Surveys* 45, 1, Article 1 (2012), 44 pages. doi:10.1145/2379776.2379777