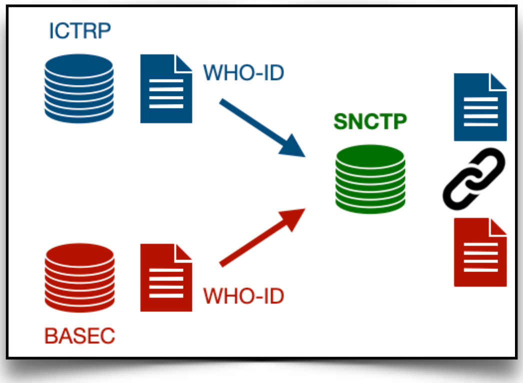
Matthias Rinderknecht
matthiasrinderknecht@gmx.ch

The missing link: text-based probabilistic record linkage of two clinical trial databases

Final Project Report



Final Project for the Certificate of Advanced Studies Course in Applied Data Science
2023 / 2024 Cycle
University of Bern

24 June 2024

Abstract

Clinical trial registries were introduced at the beginning of the new millennium to foster transparency about the existence, the purpose and the results of clinical trials. They have become an important tool for health care professionals, the medical research community and the general public to obtain information about completed and ongoing clinical trials.

In Switzerland, medical researchers are required by law to register any clinical trial they conduct. They must register their trial in an international clinical trial registry (usually in English language) and provide additional local details in a national language in a Swiss database. The two records from the two databases are then linked to each other using a common unique identifier entered by the researcher. The linking is performed in the Swiss National Clinical Trials Portal (SNCTP) based on a researcher-provided unique identifier.

In this project, an alternative, cross-language semantic similarity based linking method using Large Language Models and the python libraries „linktransformer“ and „Facebook AI similarity search“ was developed. The linking of the trials was based on the semantic similarity across languages of highly trial-specific features available in both databases, e.g. the title of the trial or the intervention studied.

During the course of the project, the combination of 3 different features (title of the trial, intervention and disease studied) resulted in a correct re-linking of up to 96.5% of over 3’000 trial pairs separated from both databases. In a real-world application and using the optimal matching confidence score threshold, the method is expected to allow linking of 1’535 yet unlinked trials with a sensitivity of 74%, a specificity of 89% and a positive predictive value of 92%, based on an extrapolation of 100 manually adjudicated links.

Using a vector index to store the precomputed LLM-derived text-embedings of the trial corpus to which new trials shall be linked massively reduced the required computing time by a factor of at least 8x compared to on-the-fly linking. This allowed searching for the best match for a single trial in a corpus of more than 12’000 trials in less than 0.5 seconds on a local CPU.

In the ongoing technical redesign of the SNCTP, the linking method developed in this project may be used as a backup method for cases where the trials could not be linked using the common unique trial identifier.

Table of Contents

1. Introduction

Public clinical trial registries were introduced in the 2000s to combat several common problems in medical research[1]. The main purposes are:

1. **During the design phase of clinical trials:**
   - Researchers can inform themselves about similar or identical research that is or was already conducted: this helps to use common end-points across different clinical trials, but also avoids unnecessary repetition of research

2. **Before starting an approved clinical trial**
   - Researchers must provide certain details about the approved clinical trial, e.g. the primary end-point of the study (the main goal), the inclusion and exclusion criteria, the responsible person etc: this serves as an a-priori statement about the goals of the study and make it more difficult to adjust the goals of the study according to the results of the study after it has been conducted

3. **During the conduct of an approved clinical**
   - Patients or their treating doctors can search for clinical trials that might be suitable for them: this offers the patients possible treatment option that would otherwise not be available
   - Researchers should update the recruitment status of the trial regularly: this helps to direct interested patients to open trials and can lead to increased recruiting into the trial

4. **After the conclusion of a trial**
   - Researchers must publish the main results of the concluded trials: this serves to make public whether the trial reached its main goal and whether the goal actually aligned to the a-priori stated hypothesis of the trial. Patients, doctors, other researchers and the general public can inform themselves about the outcomes of the trial.

Clinical trials are a subfield of research involving humans. In Switzerland, medical research involving humans is regulated on the federal level in the Federal Act on Research involving Human Beings (the „Human Research Act“, HRA)[2]. As stated in article 1 of the HRA, its goal is multifold:

1. The **primary goal** is the **protection of the research participants.**
2. The **secondary goals** are providing a legal framework for **fostering human research**, ensuring the **quality of the research** and increasing the **transparency of the research** carried out.

To increase the transparency of clinical trials, the HRA states that approved clinical trials must be registered in a clinical trial registry before they are begun. Further details how this must be done are given in the Federal Ordinance on Clinical Trials (the „Clinical Trials Ordinance“, CTO)[3]. Article 65 of the CTO states which clinical trial registries must be used for this.

In order to keep the burden on researchers as minimal as possible, a hybrid registration procedure is used:

1. **Registration in a primary registry:**
   The primary registration is done in one of the clinical trial registries acknowledged by the International Clinical Trial Registry Platform (ICTRP)[4]. This results in a unique primary registry trial ID for the clinical trial. The ICTRP is an initiative by the World Health Organization (WHO), which defines a certain dataset that registrations in all primary registries must fulfil (the WHO minimal registration set)[5]. In order to publish their research in one of the medical journals, practically all medical journals have introduced policies that make the *a priori* registration in an acknowledged clinical trial registry mandatory. Meaning these registrations are anyway done by the researchers and do not result in additional burden. The most frequently used registry is the registry of the United States National Institutes of Health, called „clinicaltrials.gov"[6]. The language in which registrations in these primary registries are carried out is mostly English and thus the information is not easily understood by all persons in Switzerland.

2. **Complementation with local information**
   To complement the registration in a primary registry, the researcher needs to provide some additional, Swiss-specific information about the clinical trial during the electronic application process

(in the BASEC system)[7] for a clinical trial at the ethics committee. This information entails, amongst others, a summary of the clinical trial in German, French or Italian written for lay-persons, a list of the Swiss study sites or the provision of a local contact in Switzerland (further information see chapter 3, data). To link the local information provided in BASEC with the information provided in the primary registry, the researcher needs to enter the unique trial ID from the primary registry.

To inform the public about clinical trials carried out in Switzerland, the CTO specifies in article 67, that the Federal Office of Public Health (FOPH) provides a public portal to search and display clinical trials.

This portal is called the Swiss National Clinical Trials Portal (SNCTP)[8]. The SNCTP displays information for every clinical trial conducted in Switzerland. It combines the information from the hybrid registration procedure described above, by downloading information from the ICTRP and BASEC for every clinical trial and linking the two records using the primary registry trial ID.

The process is described in the schema below (figure 1):

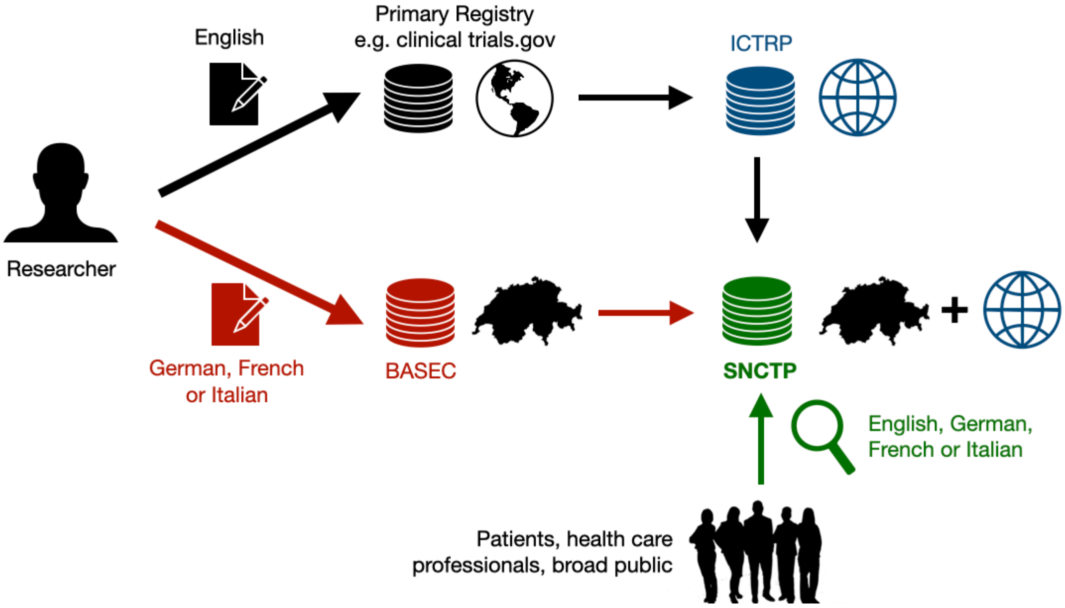


Figure 1: Flow of information from the researcher to the SNCTP and the public. *The researcher registers information about his trial in the local BASEC database and an international primary registry, which forwards to the ICTRP. The SNCTP combines the information from BASEC and ICTRP and enables searching by the public.*

The linking of the two records sometimes fails, because the primary registry trial ID is not always available in the Swiss part of the information or not always correct. This can happen, because the primary registry trial ID is not necessarily already available at the time when the request for authorisation is submitted to the regional Swiss ethics committee.

2. Project Objectives

As discussed above, the SNCTP sometimes cannot link the two corresponding records from ICTRP and BASEC, because there is not always a common identifier (the unique primary registry trial ID, or WHO-ID) available to link the records.

The current linking process is described in figure 2 below:

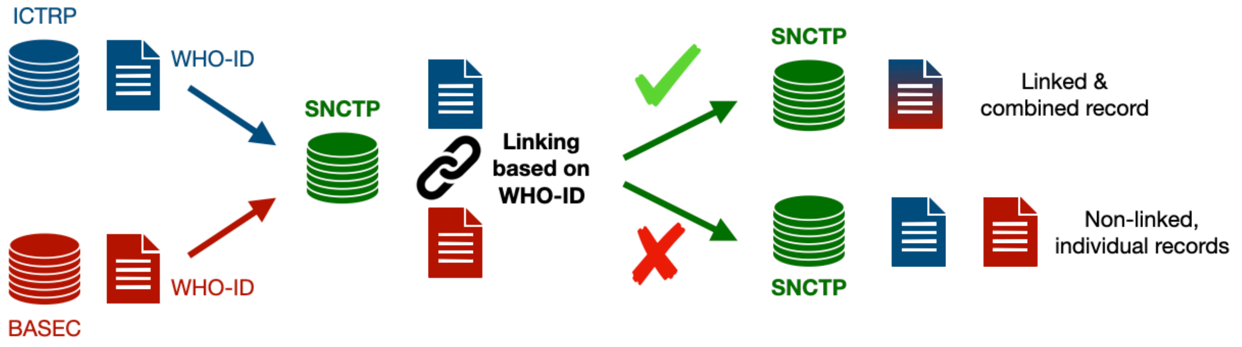


Figure 2: Record linkage performed by the SNCTP and possible outcomes. *The linking is attempted using the unique primary registry trial ID, here called the WHO-ID. If successful, the corresponding records from the ICTRP and BASEC are linked and displayed as a combined record. If linking was not successful, both records are displayed individually.*

The objective of this project is therefore to develop an alternative record linking method that could be used as a backup, whenever the direct linking via a common unique identifier is not possible.

Since there will probably not be another common unique identifier, the alternative record linking method needs to employ a probabilistic record linking method, most probably relying mainly on semantic text similarities in features present both in the BASEC and the ICTRP data.

3. Data

3.1. Preparation of datasets

Data Rows

A complete download of all data from the SNCTP database (including data sourced from BASEC and ICTRP) was provided on September 14 2023 by moxi ltd, the company who maintains the SNCTP on behalf of the FOPH.

This download included a total of 64’209 rows (trials), trials either being approved in Switzerland (imported from BASEC) or being marked to be conducted in Switzerland or one of the neighbouring countries (imported from ICTRP). The rows may contain information from both BASEC and ICTRP (linked trials), or information from BASEC only or from ICTRP only (unlinked trials).

All data obtained is from public sources. There is no personal data with the exception of the names, addresses and contact details from the study contacts; however this information has been provided deliberately for the purpose of contact and is already publicly available on the primary data sources. It is therefore concluded that there are no special requirements needed for the protection of the data.

To train the linking model, test it and use it to link yet unlinked trials, we need the following datasets:

1. The BASEC + ICTRP dataset:
This is the training dataset, containing already linked trials from BASEC and ICTRP

2. The BASEC-only dataset:
This is the dataset to match, containing only information from BASEC, with trials yet unlinked to trials from ICTRP

3. The ICTRP-only dataset:
This is the „corpus“, containing trials sourced from ICTRP, to which the trials from the BASEC-only datasets should be linked

Based on the source download, the different datasets were prepared as follows (figure 3):

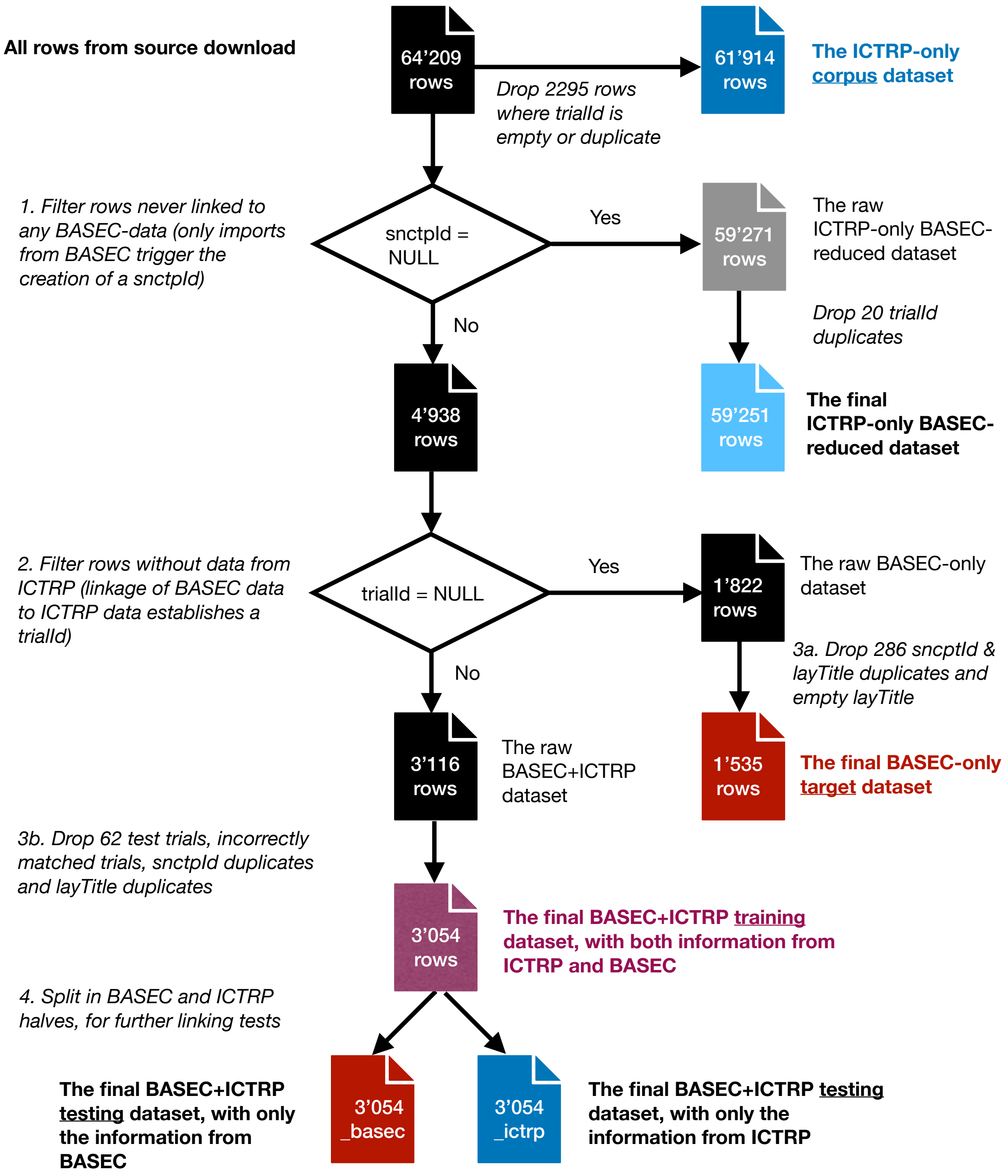

All rows from source download
64’209 rows
Drop 2295 rows where trialId is empty or duplicate
61’914 rows
The ICTRP-only corpus dataset
1. Filter rows never linked to any BASEC-data (only imports from BASEC trigger the creation of a snctpId)
snctpId = NULL
Yes
No
59’271 rows
The raw ICTRP-only BASEC-reduced dataset
Drop 20 trialId duplicates
59’251 rows
The final ICTRP-only BASEC-reduced dataset
4’938 rows
2. Filter rows without data from ICTRP (linkage of BASEC data to ICTRP data establishes a trialId)
trialId = NULL
Yes
No
1’822 rows
The raw BASEC-only dataset
3a. Drop 286 sncptId & layTitle duplicates and empty layTitle
1’535 rows
The final BASEC-only target dataset
3’116 rows
The raw BASEC+ICTRP dataset
3b. Drop 62 test trials, incorrectly matched trials, snctpId duplicates and layTitle duplicates
3’054 rows
The final BASEC+ICTRP training dataset, with both information from ICTRP and BASEC
4. Split in BASEC and ICTRP halves, for further linking tests
3’054 _basec
3’054 _ictrp
The final BASEC+ICTRP testing dataset, with only the information from BASEC
The final BASEC+ICTRP testing dataset, with only the information from ICTRP

Figure 3: Preparation of different datasets. *The different datasets described above are prepared from the source download using filtering, removing duplicates and removing certain trials such as database test trials or trials with empty key fields. Colors indicate the source of the information in the dataset, red = BASEC, blue = ICTRP, purple = BASEC + ICTRP.*

Data Columns

In order to reduce the dataset size, certain columns that were non-informative for the training process, the linking process or the manual checking of the linkage where dropped from the individual datasets (table 1).

Table 1: Overview of columns retained in the different datasets. *Colors indicate the retained information in the dataset, red = BASEC, blue = ICTRP, purple = BASEC + ICTRP.*

Dataset	Original number of columns	Number and names of retained columns
BASEC + ICTRP_full	64 (40 ICTRP + 24 BASEC)	11 ICTRP (trialId, publicTitle, secondaryId, scientificTitle, inclusionCriteria, exclusionCriteria, interventions, primarySponsor, healthConditions, publicContactAffiliation, scientificContactAffiliation) 12 BASEC (basecId, snctpId, whoId, layTitle, laySummary, disease, intervention, inclusionCriteria, exclusionCriteria, stuysites, studySitesOther, tags)
BASEC + ICTRP_ictrp	64 (40 ICTRP + 24 BASEC)	11 ICTRP (same as above 2 ICTRP (trialId/publicTitle)
BASEC + ICTRP_basec	64 (40 ICTRP + 24 BASEC)	12 BASEC (same as above) 2 BASEC (whoId/layTitle)
BASEC-only	24 BASEC	12 BASEC (same as above)
ICTRP-only	40 ICTRP	11 ICTRP (same as above) 23 ICTRP (same as above + countries, secondarySponsors, alternativeNames, publicContactFirstname, publicContactLastname, publicContactAddress, publicContactEmail, publicContactTel, scientificContactFirstname, scientificContactLastname, scientificContactAddress, scientificContactEmail)

Dataset terminology

The datasets used in this work follow this terminology:

- *„Description“_„number of rows“x„number of columns“_„optionally: part of dataset (basec or ictrp or full)“,* e.g. **BASEC-only_1540x11** or **BASEC+ICTRP_3114x2_basec**

3.2. Data quality

The characteristics of the raw dataset with 64’209 rows are shown in table 2. For the purpose of this project, the number of missing values and unique values for the features potentially useful for the record linking are especially interesting. We are looking for features that are present in both the BASEC and ICTRP subsets and have few missing values and a high proportion of unique values. In the BASEC subset, the features *layTitle, intervention* and *disease* seem promising, while in the ICTRP subset, the corresponding features *publicTitle, scientificTitle, interventions* and *healthConditions* look suitable.

Table 2: Overview of features potentially useful for the linking of the two datasets. *Potentially useful features are highlighted in bold. The complete set with all 64 features is shown in annex 1.*

Column name	BASEC Rows	Missing values	Remaining values	Duplicate values	Unique values	% unique values	Comment
snctpId	4938	0	4938	284	4654	94.2%	4938 trials of the 64209 trials have an snctpId, meaning they were imported from BASEC
whoId	4938	826	4112	403	3709	75.1%	Of the 4938 trials imported from BASEC, 3709 have a unique whoId => not all can be linked to an ICTRP-trial

Column name	BASEC Rows	Missing values	Remaining values	Duplicate values	Unique values	% unique values	Comment
contactName	4938	23	4915	1426	3489	70.7%	Not enough unique values for matching
contactMail	4938	51	4887	1988	2899	58.7%	Not enough unique values for matching
contactPhone	4938	28	4910	1491	3419	69.2%	Not enough unique values for matching
layTitle	4938	**2**	**4936**	**317**	**4619**	**93.5%**	**93.5% unique values => suitable for matching**
laySummary	4938	3	4935	307	4628	93.7%	Long text => not suitable for matching
disease	4938	**11**	**4927**	**688**	**4239**	**85.8%**	**85.8% unique values => suitable for matching**
intervention	4938	**14**	**4924**	**363**	**4561**	**92.4%**	**92.4% unique values => suitable for matching**
inclusionCriteria	4938	3	4935	332	4603	93.2%	Long text => not suitable for matching
exclusionCriteria	4938	3	4935	363	4572	92.6%	Long text => not suitable for matching
studySitesOther	4938	4285	653	276	377	7.6%	Not enough unique values for matching
studysites	4938	281	4657	3874	783	15.9%	Not enough unique values for matching
tags	4938	80	4858	4298	560	11.3%	Not enough unique values for matching

Column name	ICTRP rows	Missing values	Remaining values	Duplicate values	Unique values	% unique values	Comment
trialId	62387	0	62387	248	62139	99.6%	1822 trials of the 64209 trials from the raw download do not have a trialId, the remaining 62387 trials do have information from ICTRP
countries	62387	86	62301	41565	20736	33.2%	Not enough unique values for matching
publicTitle	62387	**86**	**62301**	**1287**	**61014**	**97.8%**	**97.8% unique values => suitable for matching**
secondaryId	62387	1734	60653	2379	58274	93.4%	High proportion of unique values, but corresponding value on BASEC side not clear
scientificTitle	62387	**855**	**61532**	**758**	**60774**	**97.4%**	**97.4% unique values => suitable for matching**
inclusionCriteria.1	62387	700	61687	672	61015	97.8%	Long text => not suitable for matching
exclusionCriteria.1	62387	29997	32390	1414	30976	49.7%	Not enough unique values for matching
interventions	62387	**3343**	**59044**	**3367**	**55677**	**89.2%**	**89.2% unique values => suitable for matching**
primarySponsor	62387	42	62345	44920	17425	27.9%	Not enough unique values for matching
secondarySponsors	62387	53745	8642	3861	4781	7.7%	Not enough unique values for matching
healthConditions	62387	**888**	**61499**	**21569**	**39930**	**64.0%**	**Rather high proportion of duplicate values, but 85.8% unique values on BASEC side => may be still suitable for matching**
alternativeNames	62387	48202	14185	49	14136	22.7%	Not enough unique values for matching
publicContactFirstname	62387	40510	21877	16125	5752	9.2%	Not enough unique values for matching
publicContactLastname	62387	23326	39061	16476	22585	36.2%	Not enough unique values for matching
publicContactAddress	62387	40340	22047	14446	7601	12.2%	Not enough unique values for matching
publicContactEmail	62387	33760	28627	11442	17185	27.5%	Not enough unique values for matching

Column name	BASEC Rows	Missing values	Remaining values	Duplicate values	Unique values	% unique values	Comment
publicContactTel	62387	35873	26514	7415	19099	30.6%	Not enough unique values for matching
publicContactAffiliation	62387	15795	46592	24336	22256	35.7%	Not enough unique values for matching
scientificContactFirstname	62387	39005	23382	17396	5986	9.6%	Not enough unique values for matching
scientificContactLastname	62387	21884	40503	17915	22588	36.2%	Not enough unique values for matching
scientificContactAddress	62387	38968	23419	14796	8623	13.8%	Not enough unique values for matching
scientificContactEmail	62387	33398	28989	12069	16920	27.1%	Not enough unique values for matching
scientificContactTel	62387	35575	26812	7607	19205	30.8%	Not enough unique values for matching
scientificContactAffiliation	62387	15328	47059	24712	22347	35.8%	Not enough unique values for matching

3.3. Example of a typical data row

An extract of a typical data row showing the features that will be used for the matching is shown below in table 3. Note that the title of the trial is present as two features in the ICTRP data: as *publicTitle* and the usually longer *scientificTitle*.

Table 3: Example of a potentially useful columns in a typical data row. *The complete data row with all 64 features is shown in annex 2.*

Feature	Feature name and data from BASEC	Feature name and data from ICTRP
Title of trial	**layTitle** Studie zur Beurteilung der Wirkung von Vancomycin im Vergleich zur verlängerten Fidaxomicin-Therapie bei der nachhaltigen klinischen Heilung einer Clostridium difficile-Infektion bei älteren Patienten	**publicTitle** A Phase IIIB/IV Study to Compare the Efficacy of Vancomycin Therapy to Extended Duration of Fidaxomicin Therapy in the Clinical Cure of Clostridium Difficile Infection (CDI) in an Older Population **scientificTitle** A Phase IIIB/IV Randomized, Controlled, Open-label, Parallel Group Study to Compare the Efficacy of Vancomycin Therapy to Extended Duration Fidaxomicin Therapy in the Sustained Clinical Cure of Clostridium Difficile Infection in an Older Population
Disease studied	**disease** Darminfektion mit Bakterium Clostridium Difficile	**healthConditions** Clostridium Difficile
Intervention studied	**intervention** Einnahme von Tabletten: Fidaxomicin für 25 Tage oder Vancomycin für 10 Tage	**interventions** Drug: Fidaxomicin;Drug: Vancomycin

3.4. Data cleaning

The datasets were cleaned as follows:

1. convert all text into strings
2. Replace all <br> with a whitespace
3. Remove all alphanumerical disease codes (such as „M50.1“) as well as all characters other than letters, numbers or whitespaces from the *healthConditions* column
4. copy *layTitle* into new *layTitleCorr* column, *scientificTitle* into new *scientificTitleCorr* column, *publicTitle* into new *publicTitleCorr*, if *scientificTitle* is empty, copy *publicTitle* into *scientificTitleCorr* instead
5. convert text in *layTitleCorr*, *scientificTitleCorr, publicTitleCorr* columns from UPPERCASE to lowercase

4. Methods

4.1. General introduction

Probabilistic record linking, also known as entity resolution or deduplication, is the process of identifying and matching records that refer to the same entity across two or more data sources.

Various techniques can be used to link two data frames with similar fields but no unique common identifier.

In this project, I focussed on the „text embeddings“ method.

Text embedding (also called „text vectorisation“) is a way to translate a text (i.e. multiple words) into a machine readable form, a single vector. Depending on the text embedding method, not only the frequency of the words itself, but also the significance of the word in the context of the whole text and the meaning of the text can be encoded into this vector up to a certain level. When text embeddings of two texts are performed, their resulting vectors can be compared and the nearness of the vectors can be computed, which is a measure of the similarity of the embedded text[9,10].

4.2. Jupyter notebooks

The python codes used in this work are organised in five notebooks available on the GitHub account of the author[11]:

- Notebook 1: Preparing the datasets and cleaning the data
- Notebook 2: Linking the datasets using the *linktransformer* library
- Notebook 3: Fine-tuning the LLM using the *linktransformer* library
- Notebook 4: Linking the datasets using the *Facebook AI similarity search* library
- Notebook 5: Linkage Analysis

4.3. The *linktransformer* library

Linktransformer[12] is a python library that is dedicated to perform record linkage tasks. In contrast to other libraries for this task, *linktransformer* uses large language model (LLM) embeddings to match records based on the semantic similarity of the meaning of the text of chosen data fields across the two datasets to match. *Linktransformer* can use a variety of LLMs, such as OpenAI models or popular open-source models from HuggingFace[13].

The basic usage is as follows[14]:

```
merged_df = linktransformer.merge(df1, df2, on='key_column', model=‚your-pretrained-model-from-
huggingface‘)
```

where

- *merged_df* is the output merged data frame
- *df1* and *df2* are the two dataframes to be merged
- *on=„key_column“* specifies the column name in both data frames that contains the text to be used for determining the similarity of two records
- *model=„your-pretrained-model-from-huggingface“* specifies the LLM to be used to create the text embeddings and determine similarity.

For every row in the left dataframe (df1), the method compares the indicated matching columns and computes a „score“ value between 0 and 1, which indicates the semantic similarity of the row in df1 and every row in df2. It then attributes the one row from the right data frame (df2) with the highest score (the „best-match“) to the current row in df1. Instead of linking just the best-matching row in df2, we can also specify that it should link *k* number of best matching rows in df2 to every row in df1.

Because we need to compare German, French or Italian text from BASEC to English text from ICTRP, we need to use a multi-lingual LLM that directly compares the texts in their source language, without the need to translate first into English.

The open-source multi-lingual LLMs used in this work are:

- paraphrase-multilingual-mpnet-base-v2[15]
- distiluse-base-multilingual-cased-v1[16]

Both of these models are pre-trained for the sentence similarity task.

Linktransformer also has the *train_model* method, which is used to finetune pretrained LLMs. The usage is as follows:

```
best_model_path=lt.train_model(
        model_path=„Pretrained-base-model“,
        data=your-training-dataframe,
        left_col_names=[„Left-training-column-1“, „Left-training-column-2“],
        right_col_names=[„Right-training-column-1“, „Right-training-column-2“],
        left_id_name=[„Left-common-ID“],
        right_id_name=[„Right-common-ID“],
        log_wandb=False
        training_args={„num_epochs“: 3})
```

where

- *model_path* specifies the pretrained base LLM to be fine-tuned, in our case *paraphrase-multilingual-mpnet-base-v2* (see above)
- *data* is the dataframe containing the training data
- *left_col_names* and *right_col_names* specify the column names in the dataframe that contain the text to be used for determining the similarity of two records. The left columns will be matched to the right columns during the matching process.
- *left_id_name* and *right_id_name* specify the column names in the dataframe that contain the common unique identifiers for the left and the right columns, the ground truth used for matching.
- l*og_wandb* specifies whether the training should be logged or not
- *training_args* specifies for how many epochs the model is trained.

4.4. The *Facebook artificial intelligence similarity search (Faiss)* library

Faiss[17],[18] is a python library built around vector similarity search. It offers tools for computing a vector index of a database (or „corpus“) by embedding their text content, searching the index for semantically similar texts (the „query“) and retrieving the search results from the corpus.

By precomputing the vector index of a corpus and store it, instead of computing it on-the-fly for every new search, one can re-use the index for multiple searches and massively reduce the search time. This especially matters if you need to perform many repeated single queries against a big corpus (as opposed to multiple parallel queries, as in this case, the computing time for the multiple parallel queries will become more constraining).

Basic usage[19]:

```
# Build index:
corpus_embeddings = model.encode(corpus) # embed the corpus into vectors using an LLM
faiss.normalize_L2(corpus_embeddings) # normalize the embeddings for cosine similarity
index = faiss.IndexFlatIP(d) # create an index of dim. d (the shape of the embeddings)
index.add(corpus_embeddings) # add embedding vectors to the index

# Build query:
query_embedding = model.encode([query]) # embed the query using the same LLM as above

# Search in corpus for similar embeddings to the query
k = 1  # Retrieve only the top match
```

```
query_distance, query_index = index.search(query_embeddings, k)

# Retrieve most similar match from corpus
match = corpus[query_index]
```

4.5. Infrastructure

Since the *linktransformer* library and the LLMs used are quite big in data size and to be able to benefit from GPUs, Google colab with the T4 GPU runtime was used for all record linking tasks using the *linktransformer* library.

In a later step, a local python installation (Anaconda Distribution for Python) running on a private machine (Apple MacBook Air M1, 8GB RAM) was used for performing the one-off matching for single new trials using the *Faiss* library and precomputed embeddings.

4.6. Explanation of code used for record linking using *linktransformer* and evaluating the matching performance (pseudo-code)

For full python code see Jupyter Notebook 2.

```
1. The basec and ictrp csv files are loaded into df1 and df2 pandas data frames:
   df1 = BASEC+ICTRP_3054x12_basec.csv
   df2 = BASEC+ICTRP_3054x11_ictrp.csv

2. The linkage is performed, matching on one or multiple columns of df1 vs. one or multiple  columns
   of df2:
   merged_df = lt.merge(df1, df2, on=None, model="sentence-transformers/paraphrase-multilingual-
   mpnet-base-v2", left_on=„columns in df1", right_on="columns in df2“)

3. The matching accuracy is assessed by counting the number of true matches of the unique common
   identifier in the „whoId“ column (from df1) and „trialId“ column (from df2)
```

4.7 Explanation of code used for fine-tuning the LLM used in *linktransformer* (pseudo-code)

For full python code see Jupyter Notebook 3.

```
1. The BASEC+ICTRP training csv file is loaded into a pandas data frame:
   dfTrain = pd.read_csv("BASEC_with_ICTRP_3054x64.csv")

2. The training is performed, using the pretrained LLM as a basis:
   best_model_path=lt.train_model(
           model_path="sentence-transformers/paraphrase-multilingual-mpnet-base-v2",
           data=dfTrain,
           left_col_names=["layTitle", "layTitle", "disease", "intervention"],
           right_col_names=['scientificTitle', "publicTitle", „healthConditions",
               "interventions"],
           left_id_name=['whoId'],
           right_id_name=['trialId'],
           log_wandb=False,
           training_args={"num_epochs": 3})

3. The fine-tuned model is then zipped and written to the content folder in Google Colab and
   subsequently downloaded to disk for local use:
   !zip -r /content/model5.zip /content/models
```

4.8 Explanation of code used for establishing and searching using a precomputed corpus (pseudo-code)

For full python code see Jupyter Notebook 4.

Create the precomputed corpus:

```
1. Load the fine-tuned LLM  and the corpus to be encoded from disk:
   model = SentenceTransformer(„Model_7/linkage“)
   corpus_df = pd.read_csv(‚ICTRP_only(CH=true)_12820x23.csv')

2. Specify the columns to be embedded for index search and concatenate them:
   text_columns = ["scientificTitle", "publicTitle", "interventions", „healthConditions"]
   corpus = corpus_df[text_columns].apply(lambda row: ' '.join(row.values.astype(str)),
   axis=1).tolist()

3. Embed the corpus, normalize embeddings and initialise index:
   corpus_embeddings = model.encode(corpus)
   faiss.normalize_L2(corpus_embeddings)
   d = corpus_embeddings.shape[1]
   index = faiss.IndexFlatIP(d)  # Use IndexFlatIP to search with inner product

4. Add normalized corpus embeddings to the index and save the index:
   index.add(corpus_embeddings)
   faiss.write_index(index, NAME.index')
```

Search the precomputed index against a new query:

```
1. Load the index (=precomputed corpus), the corpus and the query:
   index = faiss.read_index(NAME.index)
   corpus_df = pd.read_csv(CORPUS.csv)
   query_df = pd.read_csv(SEARCH.csv) #query_df may contain several queries

2. Specify the columns to be embedded for querying and concatenate them:
   query_columns = ["layTitle", "layTitle", "intervention", "disease"]
   queries = query_df[query_columns].apply(lambda row: ' '.join(row.values.astype(str)),
   axis=1).tolist()

3. Embed the query and normalise the embeddings:
   query_embeddings = model.encode(queries)
   faiss.normalize_L2(query_embeddings)

4. Search the index for similar embeddings to the query:
   k = 1  # Retrieve the top match
   distances, indices = index.search(query_embeddings, k)

5. Initialize an empty DataFrame to store the results and the query distance („score“):
   result_df = pd.DataFrame(columns=list(new_row.columns) + list(corpus_df.columns) + [„score“])

6. Process the search results:
   for i, row in query_df.iterrows():
    query_embedding = query_embeddings[i]
    query_distance, query_index = distances[i][0], indices[i][0]

    # Create a new row by combining the query row and the matched corpus row and the query distance
    result_row_values = list(query_df.iloc[i]) + list(corpus_df.iloc[query_index]) +[query_distance]
    result_row_df = pd.DataFrame([result_row_values], columns=result_df.columns)

    # Append the result_row_df to result_df
    result_df = pd.concat([result_df, result_row_df], ignore_index=True)

7. Save the result_df to a CSV file:
   result_df.to_csv(FILE_NAME, index=False)
```

5. Results

5.1. Establishing the linking strategy using *linktransformer*

Using the BASEC+ICTRP dataset with known linkages, the linking strategy was set up.
The ictrp and basec parts of the full BASEC+ICTRP dataset were read into separate data frames (see methods chapter 4.4) and linking was performed using the merge method of the *linktransformer* library.

To test the procedure, linking was performed on the whoId column of the basec part vs. the trialId column of the ictrp part. These columns contain the unique common identifier and linking on these columns should result in 100% accuracy, which was achieved (see match #1 in table 4).

After validating the correct functioning of the library, the linking using text embeddings was approached. In a first attempt the *layTitle* column of df1 was matched against the *publicTitle* column of df2. The matching accuracy was 69.2% (match #2). Then, the *layTitle* column was matched against the *scientificTitle* column, which improved the accuracy to 77.5% (match #3). Other combinations, also with multiple columns were tested in the following. The combination of 4 columns (*layTitle, layTitle, disease, intervention* on the BASEC side vs. *scientificTitle, publicTitle, healthcondition, interventions* on the ICTRP side) proved to be the best tested combination, with an accuracy of 81.8% (match #4).

Effect of using another LLM
Instead of the *paraphrase-multilingual-mpnet-base-v2 (pmmbv2)* model, the *distiluse-base-multilingual-cased-v1 (dbmcv1)* model was tested which resulted in a less good linkage accuracy (72.7% vs 81.8%, matches #4 and #5). For the rest of the project, the *paraphrase-multilingual-mpnet-base-v2* model was therefore used.
During the course of this work, a newer LLM became available, the *multilingual-e5* in the sizes large/base/small. *Multilingual-e5-large (me5l)* reached 89.6% accuracy, but with a more than 6x longer computing time (match #6).

Effect of using cleaned vs. raw data
After the first matches based on *layTitle* vs. *scientificTitle*, the incorrectly matched rows where inspected in detail. The following three patterns were discovered:

1. *scientificTitle* was empty in about 4% of incorrectly matched rows.
2. *layTitle* as well as *scientificTitle* was written in uppercase letters in 6% and 2.5% of rows respectively. It was reported in literature that using uppercase letters could impair embedding and retrieval quality in LLMs[20].
3. *healthConditions* and *interventions* columns sometimes contain formatting codes such as <br>

The datasets were therefore cleaned as described in chapter 3.3 „Data Cleaning“. Using cleaned data increased the matching accuracy from 81.8% to 83.6% (match #7).

Effect of k=3 vs. k=1 matching
Instead of matching only the best match from the _ictrp dataset to a given trial from the _basec dataset (k=1 matching), matching the best three trials from the _ictrp dataset to a given trial from the _basec dataset (k=3 matching) was also tried out. Compared to the k=1 matching, the second and third best matches contributed additional accuracies of 6.9% and 2.1%, which resulted in a total accuracy of 92.6% (match #9).

Table 4: Overview of linkages performed, their parameters and linkage accuracy achieved. *LLMs used: paraphrase-multilingual-mpnet-base-v2 (pmmbv2), distiluse-base-multilingual-cased-v1 (dbmcv1), multilingual-e5-large (me5l)*

Match #	df1 (left data frame)	df2 (right data frame)	Match on left columns	Match on right columns	LMM used	Correct linkages (out of a total of 3054 rows	Accuracy (% correct linkages)	Computing time on Google Colab T4
1	BASEC+ICTRP_3054x12_basec	BASEC+ICTRP_3054x11_ictrp	whoID	trialID	pmmbv2	3054	100%	18s
2	BASEC+ICTRP_3054x12_basec	BASEC+ICTRP_3054x11_ictrp	layTitle	publicTitle	pmmbv2	2113	69.2%	20s
3	BASEC+ICTRP_3054x12_basec	BASEC+ICTRP_3054x11_ictrp	layTitle	scientificTitle	pmmbv2	2368	77.5%	25s
4	BASEC+ICTRP_3054x12_basec	BASEC+ICTRP_3054x11_ictrp	layTitle, layTitle, disease, intervention	scientificTitle, publicTitle, healthConditions, interventions	pmmbv2	2497	81.8%	47s
5	BASEC+ICTRP_3054x12_basec	BASEC+ICTRP_3054x11_ictrp	layTitle, layTitle, disease, intervention	scientificTitle, publicTitle, healthConditions, interventions	dbmcv1	2220	72.7%	32s
6	BASEC+ICTRP_3054x12_basec	BASEC+ICTRP_3054x11_ictrp	layTitle, layTitle, disease, intervention	scientificTitle, publicTitle, healthConditions, interventions	me5l	2735	89.6%	5 min
7	BASEC+ICTRP_3054x12_basec_cleaned	BASEC+ICTRP_3054x11_ictrp_cleaned	layTitleCorr, layTitle, disease, intervention	scientificTitleCorr publicTitle, healthConditions, interventions	pmmbv2	2552	83.6%	51s
8	BASEC+ICTRP_3054x12_basec_cleaned	BASEC+ICTRP_3054x11_ictrp_cleaned	layTitleCorr, layTitle, disease, intervention	scientificTitleCorr publicTitle, healthConditions, interventions	pmmbv2 k=3	2827 (2552 +210 +65)	92.6% (83.6% +6.9% +2.1%)	2 min
9	BASEC+ICTRP_3054x12_basec	BASEC+ICTRP_3054x11_ictrp	layTitle, layTitle, disease, intervention	scientificTitle, publicTitle, healthConditions, interventions	me5l k=3	2946 (2735 +172 +39)	96.5% (89.6% +5.6% +1.3%)	6 min

Evolution of linkage accuracy and settling on a linking strategy

The evolution of the linkage accuracy is plotted in figure 4. Using more and more refined strategies, the linkage accuracy could be increased from initially 69.2% (matching *layTitle* vs. *publicTitle* only) to 96.5% (matching on 4 columns, using cleaned data, 3 top hits and an especially large LLM).

Taking these preliminary results in account and in order to settle on a strategy that provides a good trade-off between accuracy, data cleaning burden and computing time, I decided to use the following linking strategy for the remainder of the work,:

- Use the *paraphrase-multilingual-mpnet-base-v2* LLM or a fine-tuned model based on this LLM
- Match of 4 columns: l*ayTitle/layTitle/intervention/disease* on the BASEC side vs. *scientificTitle/publicTitle/interventions/healthConditions* on the ICTRP side
- No data cleaning applied
- Use k=1 matching (only the top match)

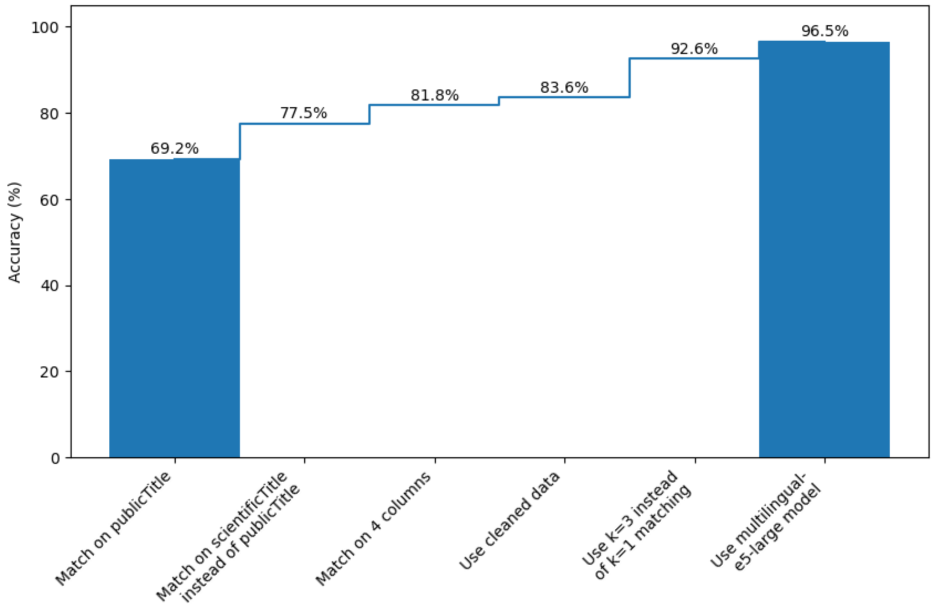


Figure 4: Step-plot of the evolution of the linking accuracy using different strategies.

5.4. Fine-tuning the used LLM

After the linking strategy was established, the pretrained LLM used for text embedding was fine-tuned using the *train_model* method of the *linktransformer* library (see chapter 4.4).

The BASEC + ICTRP_3054x64 dataset was used for the training. The *train_model* method splits the training dataset in a train, test and validation datasets. The training was run for 3 epochs. Training took about 8 minutes on Google Colab T4 runtime.

Compared to the accuracy using the pre-trained LLM (81.8%), linking using the fine-tuned LLM resulted in a gain of 10.0% accuracy resulting in a total of 91.8%. For the remainder of the work, the fine-tuned LLM („model 7“) was used.

Table 5: Parameters used for the fine-tuning of the LLM and comparison of the linkage accuracies achieved using the pre-trained vs. the fine-tuned model

Source Data used	Training parameters	Accuracy in the validation set according to train_model method	Accuracy on the full source dataset using the fine-tuned model	Accuracy on the full source dataset with the pre-trained model
BASEC+ICTRP_3054x64	left_col_names=[„layTitle“, "layTitle", "disease", "intervention"], right_col_names=['scientificTitle', "publicTitle", "healthConditions", "interventions"]	96.1%	91.8%	81.8%

5.5. Applying the linking strategy on yet unlinked data

With the established linking strategy and the fine-tuned LLM, the core task of matching yet unlinked BASEC trials from the BASEC-only dataset (1535 rows) with the ICTRP-trials from the ICTRP-only corpus (61’914 rows) dataset was tackled.

Since in the BASEC-only dataset, no linking with an ICTRP-trial based on the whoId/trialId could be established yet, no automated checking of correct linking is possible. The matches must be inspected manually and matching must be adjudicated based on all available data for the matched trials.

The first 100 matched trials were inspected manually. 56 of the 100 inspected matched trials (56%) were correct matches.

The *linktransformer* library adds a „score“ value to every merged data row, which indicates the similarity score of the two merged data rows from df1 and df2. In order to determine the significance and reliability of the score value, the matching accuracies for every 0.1 bin of the score value is calculated (figure 5). The matching accuracies in the 0.8 - 0.9 and the 0.9 - 1.0 bins of the score values amounts to 90% and 100% respectively and is therefore pretty reliable, while the reliability in the 0.7 - 0.8 bin drops to 56% and decreases further for bins below. In order to reduce the number of incorrect linkages, one could therefore set a threshold at e.g. score-value of 0.8, under which no linkages are accepted and no result is returned.

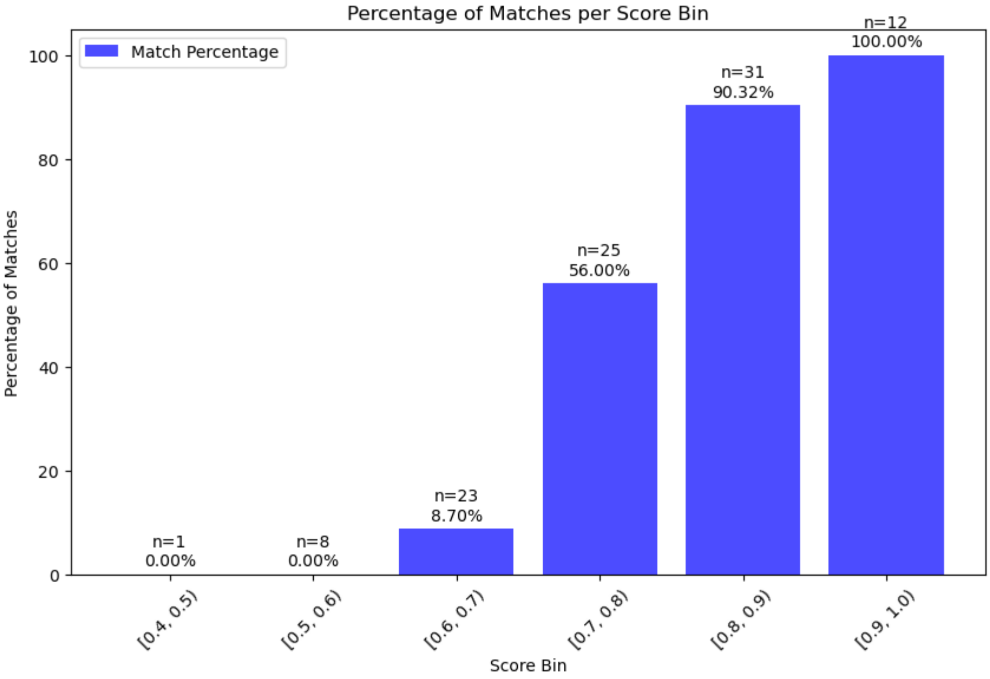


Figure 5: Linkage accuracy per 0.1 confidence score bin

The matching of the 1’535 row BASEC-only dataset against the 61’914 row ICTRP-only dataset took about 8 minutes on Google Colab T4 runtime. The computing time needed depends on the size of the right dataframe, since also the matching of a 20 row sample from the BASEC-only dataset against the 61’914 row ICTRP-only dataset took 8 minutes. This severely hampers the use of the matching process for on-the-fly matching for single new trials. Therefore, an alternative, faster method for searching matches for single new trials should be developed.

5.6. Finding an alternative, more efficient linking strategy using precomputed embeddings

The reason why the *merge* method of the *linktransformer* library takes a lot of time even when matching just one trial with a corpus dataset is because it recomputes text embeddings every time anew for both datasets. This is because the method is designed to be used for one-off merging of datasets and not for repeated searches for the best match of a single trial/row within a corpus of rows. In order to do this efficiently, the text embeddings of the corpus should be computed only once („precomputed“) and stored in an index (a so called "vector store“). With this strategy, the text embeddings for a new search need to be computed only for the trial to be matched, and can then be compared to the precomputed embeddings in the index. The most similar vector/row in the index is then computed using cosine vector similarity and the optimal match is retrieved from the corpus dataset.
The embeddings were precomputed using fine-tuned model 7 and stored as an index (vector store) of the different corpus datasets as described in chapter 4.6 (table 6):

Table 6: Parameters, computing time and size of precomputed indexes

Corpus dataset	Embedded columns	Computing time	Size of index on disk
BASEC+ICTRP_3054x11_ictrp	scientificTitle, publicTitle, interventions, healthConditions	1m 50s on local CPU	9.4 MB
ALL_64209x19_ictrp	scientificTitle, publicTitle, interventions, healthConditions	23min 54s on local CPU	197.3 MB
ICTRP_only(CH=true)_12820x23.csv	scientificTitle, publicTitle, interventions, healthConditions	8min 23s on local CPU	39.4 MB
ICTRP_only_61914x23_ictrp.csv	scientificTitle, publicTitle, interventions, healthConditions	25min 44s on local CPU	190 MB

Matching was then performed using the precomputed indexes. Computing times needed to precompute the indexes and for searching the precomputed indexes as compared to using the on-the-fly embedding using *linktransformer* is shown below (table 7):

Table 7: Parameters, linkage accuracies and computing time for precomputed vs. on-the-fly linking. *Linkages were computed using layTitle, layTitle, disease, intervention on the BASEC side vs. scientificTitle, publicTitle, healthConditions, interventions on the ICTRP side.*

Match #	df1 (left data frame)	df2 (right data frame)	Total rows	Correct linkages (%), precomputed	Correct linkages (%), on-the-fly	Computing time (precomputed)	Computing time (on-the-fly)
1	BASEC+ICTRP_3054x12_basec	BASEC+ICTRP_3054x11_ictrp	3054	2804 (91.8%)	2822 (92.4%)	1min 10s on local CPU	54 s on Colab T4
2	BASEC+ICTRP_3054x12_basec	ALL_64209x19_ictrp	3054	2093 (68.5%)	2176 (71.3%)	1 min 8s on local CPU	8 min on Colab T4

Matching accuracy was almost retained while computing time for big corpuses was reduced significantly.

However, an interesting effect was observed: when matching the BASEC+ICTRP_3054x12_basec dataset vs. the whole ICTRP corpus (ALL_64209x19_ictrp) (match #2), linking accuracy was reduced significantly compared to when linking vs. the ICTRP-part of the BASEC+ICTRP_3054x12 dataset (BASEC+ICTRP_3054x12_ictrp) (match #1). This probably is due to a „contamination" or „dilution“ effect, because in the ALL_64209x19_ictrp dataset, the possibilities to match to are more than 20-fold greater than in the BASEC+ICTRP_3054x12_ictrp dataset. Figuratively spoken, the haystack to search got bigger.

The performance of the established linking strategy as described in chapter 5.4 is therefore expected to be lower when matching against a very big corpus.

5.7. Applying the linking strategy on real-world data, using precomputed embeddings

As a final step, the established linking strategy and the precomputed embeddings were used to perform the core task again.

Using the complete corpus of 64’209 rows, 41 of 100 matches were assessed to be correct (see table 8, match #1).

In order to counter the haystack-effect observed in the previous section, I shrunk the corpus by removing all trials in the corpus that did not contain „Switzerland“ in the countries column. This resulted in a reduced corpus of 12’820 rows and an increased percentage of correct matches of 64% (match #2).

Table 8: Overview of linkage results and computing time using precomputed embeddings

Match #	df1 (left data frame) = query	df2 (right data frame) = corpus	Matches inspected	Correct linkages (%)	Computing time (on local CPU)
1	BASEC_withouth_ICTRP_1535x12	ALL_64209x19_ictrp	100	41 (41%)	1min 15s
2	BASEC_withouth_ICTRP_1535x12	ICTRP_only(CH=true)_12820x23	100	64 (64%)	39s
3	BASEC_withouth_ICTRP_1535x12	ICTRP_only_61914x23_ictrp	100	59 (59%)	55s
4	Single random row from BASEC_withouth_ICTRP_1535x12	ICTRP_only(CH=true)_12820x23	Not inspected	Not inspected	**0.18s to 0.5s**

To reduce the number of reported false positives, the score value (indicating matching confidence) will be used to define a threshold, below which no linkages are accepted. In order to find the optimal threshold of the score value, a Receiver-Operating Curve (ROC) showing sensitivity vs. 1-specificity for several different threshold values for match #2 from above was plotted (see figure 6). The optimal point of the ROC lies in the top left corner, indicating 100% sensitivity and 100% specificity.

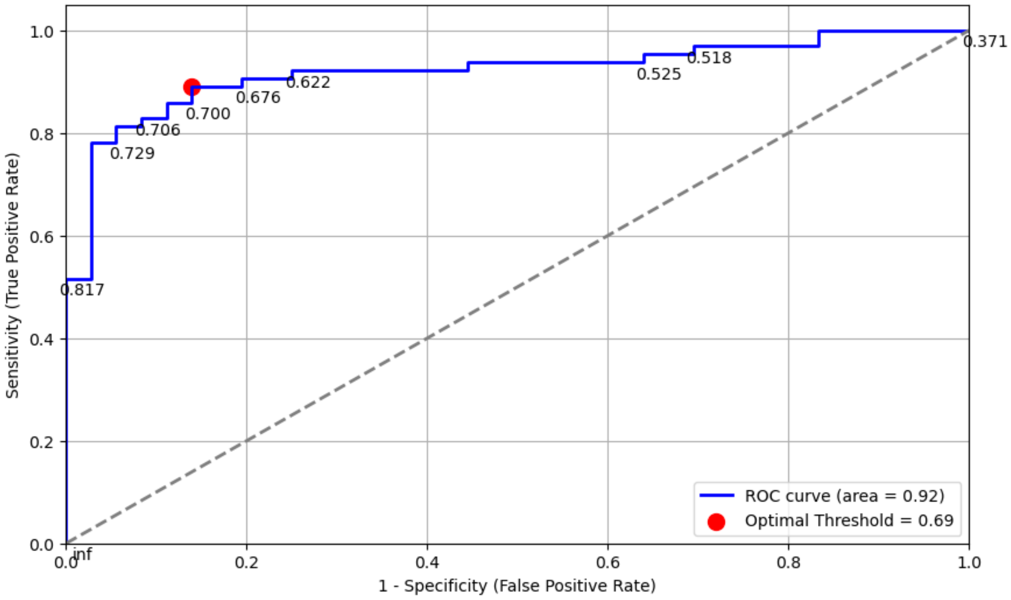


Figure 6: Receiver Operating Characteristics (ROC) curve depicting sensitivity and specificity at 10 different matching confidence score thresholds from match #2.

When applying the optimal score threshold (0.69) for accepting matches derived from the ROC curve for match #2 (see red dot in figure 6), the metrics at this threshold are as in table 9. The observed metrics from the 100 query adjudicated set (an achievable percentage of 64% correct matches and a positive predictive value of 92% at a threshold of 0.69 for the confidence score) were extrapolated to the full 1535 queries set. This resulted in a predicted number of 722 correct matches within 785 retained matches above the threshold and a sensitivity of 74%.

Table 9: Metrics for a manually adjudicated and the full target dataset using the chosen linkage strategy and matching confidence score threshold

Set	% and number of correct matches	Score threshold applied	Total matches above threshold	Positive predictive value (precision)	False discovery rate	Sensitivity (true positive rate)	Specificity (true negative rate)
100 randomly adjudicated queries from the 1535 queries	64% (64 correct matches)	0.69	62 of 100 (62%)	92% (57 of 62 are correct matches)	8% (5 of 62 are incorrect matches)	89% (57 out of 64 correct matches retained above threshold)	86% (31 of 36 incorrect matches removed by threshold)
All 1535 queries	Assuming same percentage (64%) as above: 982 correct matches	0.69	785 of 1535 (51%)	Predicted correct matches above threshold at 92% positive predictive value: 722	Predicted false matches above threshold at 8% false discovery rate: 63	74% (722 of 982 extrapolated correct matches retained above threshold)	89% (490 of 553 extrapolated incorrect matches removed by threshold)

Example of a correct and an incorrect match

Table 10: Example of correctly linked trial *(matching score 0.84)*

Feature	Feature name and data from BASEC	Feature name and data from ICTRP
Title of trial	**layTitle** Impact respiratoire des agents de courtes demies vies utilisés en anesthésie générale chez les patients souffrants ou suspects de syndrome d'apnée du sommeil (SAOS)	**publicTitle** Short Life Agents in Balanced Anesthesia on Obstructive Sleep Apnea Syndrome **scientificTitle** Respiratory Impact of Short Life Agents Used in Balanced Anesthesia on Patients Suffering or Suspected of Obstructive Sleep Apnea (OSA) Syndrome
Disease studied	**disease** Patient souffrant du syndrome d'apnée du sommeil obstructif non traité par CPAP (pression positive continue) ou patient suspecté de souffrir du syndrome d'apnée du sommeil avec une réponse au questionnaire STOP BANG de detection du SAOS supérieure ou égal à 3.	**healthConditions** Sleep Apnea Syndromes;Sleep Apnea, Obstructive
Intervention studied	**intervention** L'étude comparera un groupe "intervention" à un groupe "contrôle" ; le groupe intervention bénéficiera des médicament et agents inhalés anesthésiant à courtes durées d'actions alors que le groupe contrôle sera au bénéfice de médicaments et d'agents inhalés anesthésiant à moyenne durée d'action. Chez des patients souffrants du syndrome d'apnée du sommeil non traité par CPAP ou suspect de SAOS nous comparerons lors d'une intervention chirurgicale de type orthopédique interessant les membres inférieurs, l'utilisation concomitante de desflurane et de rémifentanil pour le groupe "intervention" à un groupe contrôle recevant du sevoflurane et du fentanyl. (…)	**interventions** Drug: Fentanyl and sevoflurane;Drug: Remifentanil and desflurane

Table 11: Example of incorrectly linked trial *(matching score 0.46)*

Feature	Feature name and data from BASEC	Feature name and data from ICTRP
Title of trial	**layTitle** Vergleich zweier Wirkstoffe für eine frühe Verlängerung der Behandlungsintervalle bei feuchter altersbedingter Makuladegeneration	**publicTitle** Comparison of Treatment rOutine Using afLibERcept: Strict vs relAxed retreatmeNT Regimen **scientificTitle** Comparison of Treatment rOutine Using afLibERcept: Strict vs relAxed retreatmeNT Regimen (TOLERANT Study)
Disease studied	**disease** feuchte altersbedingte Makuladegeneration	**healthConditions** Age Related Macular Degeneration
Intervention studied	**intervention** Die zugelassenen Medikamente Aflibercept (Eylea®) und Brolucizumab (Beovu®) wirken sehr lange. Wir gehen davon aus, dass in den meisten Fällen am Anfang Spritzen ins Auge alle sechs Wochen ausreichen zur Stabilisierung und ohne die Sehschärfe zu beeinträchtigen. Je nach Krankheitsverlauf können die Zeitabstände weiterhin verlängert oder verkürzt werden. Die Studie vergleicht die minimal nötige Anzahl Injektionen und Therapie-Intervalle der beiden Medikamente.	**interventions** Drug: Aflibercept

6. Discussion

The results of this study indicate that it is possible to develop a linking method based on the semantic similarity of multiple text fields in the BASEC and ICTRP subsets. The matching accuracy depends on several factors that need to be weighted against each other: corpus size, corpus targetedness, matching features, LLM size, computing time at disposal, data pre-processing and matching thresholds.

Using a precomputed index massively reduced the computing time needed for the linking of the subsets, especially when attempting to link single new trials.

When looking at the final aim of this study, the linking of yet unlinked studies from the BASEC_only subset to the ICTRP subset, the results indicate that from the 1'535 unlinked trials in BASEC, 722 could potentially be linked correctly to a trial from the ICTRP subset, with a high positive predictive value of 92% (meaning only 63 false positives) when applying the optimal score threshold and using a precomputed search index. This accuracy is acceptable for the purpose required, especially given the alternative of no link at all. However, this result is based on the extrapolation of a manually checked subset of 100 linked trials.

In terms of choosing the optimal corpus size and targetedness (making the „haystack" smaller without loosing any „needles" in the discarded „hay"), there are some trade-offs that applied here. By using the Swiss-targeted corpus, the accuracy improved, because the corpus was about 4x smaller than the full corpus, but some potential matches will be lost for good because some matching trials do not have „Switzerland" listed as a country of conduct, although the trial is conducted in Switzerland.

7. Conclusion & Outlook

It was shown in this work, that the developed method using a precomputed index and a specific matching threshold can be used to search for every new incoming trial from BASEC, not able to be linked to its ICTRP counterpart by the unique identifier, for a potential match in the corpus in less than 0.5 seconds.

If the potential match exceeds the confidence threshold, the potential match could be displayed to the user in BASEC or on SNCTP, if it does not exceed the threshold, it will not be displayed. By indicating to the user that the link was generated automatically and probability-based and also give the confidence score of the linking, the user will be informed transparently that the linked trial is not guaranteed to be the correct one. The displayed trial could even be accompanied by two feedback buttons on the webpage, where the user could give feedback whether the linking was correct or not.

The *linktransformer* library might in the future even receive the option of using a precomputed index instead of relying on on-the-fly computing of the corpus embeddings. This would make the process easier, because *linktransformer* integrates many of the steps that need to be executed separately using the *Faiss* library.

8. Acknowledgements

I would like to thank Dr. Sigve Haug and all the teachers of this CAS for sharing their enthusiasm and expertise in data science. I am especially grateful to Dr. Aris Marcolongo and Dr. Mykhailo Vladymyrov for making me aware of the text embeddings method for semantic similarity search during the discussions held at the CAS module 3 retreat.

I would also like to thank Abhishek Arora[21] from the *linktransformer* team at Harvard University who kindly answered a question from me concerning the use of precomputed indexes with *linktransformer* and pointed me to the possibility of using the *Faiss* library to do the trick.

Statement

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: 24.06.2024 Signature(s): Sig. M. Rinderknecht

Annex

1. Description of raw dataset (64’209 rows) with BASEC-derived (red) and ICTRP-derived (blue) features

Column name	BASEC-Rows	Missing values	Remaining values	Duplicate values	Unique values	% unique values
basecId	4938	891	4047	282	3765	76.2%
snctpId	4938	0	4938	284	4654	94.2%
whoId	4938	826	4112	403	3709	75.1%
ECName	4938	123	4815	4804	11	0.2%
ecFinalDecisionDate	4938	898	4040	2448	1592	32.2%
WHO register	4938	944	3994	3985	9	0.2%
flagRareDisease	4938	0	4938	4935	3	0.1%
flagForChildren	4938	0	4938	4935	3	0.1%
flagForAdolescents	4938	0	4938	4935	3	0.1%
flagForHealthy	4938	0	4938	4935	3	0.1%
published	4938	0	4938	4935	3	0.1%
contactName	4938	23	4915	1426	3489	70.7%
contactMail	4938	51	4887	1988	2899	58.7%
contactPhone	4938	28	4910	1491	3419	69.2%
lang	4938	0	4938	4933	5	0.1%
layTitle	4938	2	4936	317	4619	93.5%
laySummary	4938	3	4935	307	4628	93.7%
disease	4938	11	4927	688	4239	85.8%
intervention	4938	14	4924	363	4561	92.4%
inclusionCriteria	4938	3	4935	332	4603	93.2%
exclusionCriteria	4938	3	4935	363	4572	92.6%
studySitesOther	4938	4285	653	276	377	7.6%
studysites	4938	281	4657	3874	783	15.9%
tags	4938	80	4858	4298	560	11.3%

Column name	ICTRP-Rows	Missing values	Remaining values	Duplicate values	Unique values	% unique values
trialId	62387	0	62387	248	62139	99.6%
countries	62387	86	62301	41565	20736	33.2%
publicTitle	62387	86	62301	1287	61014	97.8%
secondaryId	62387	1734	60653	2379	58274	93.4%
scientificTitle	62387	855	61532	758	60774	97.4%
inclusionCriteria.1	62387	700	61687	672	61015	97.8%
exclusionCriteria.1	62387	29997	32390	1414	30976	49.7%
interventions	62387	3343	59044	3367	55677	89.2%

Column name	BASEC-Rows	Missing values	Remaining values	Duplicate values	Unique values	% unique values
primarySponsor	62387	42	62345	44920	17425	27.9%
secondarySponsors	62387	53745	8642	3861	4781	7.7%
healthConditions	62387	888	61499	21569	39930	64.0%
alternativeNames	62387	48202	14185	49	14136	22.7%
publicContactFirstname	62387	40510	21877	16125	5752	9.2%
publicContactLastname	62387	23326	39061	16476	22585	36.2%
publicContactAddress	62387	40340	22047	14446	7601	12.2%
publicContactEmail	62387	33760	28627	11442	17185	27.5%
publicContactTel	62387	35873	26514	7415	19099	30.6%
publicContactAffiliation	62387	15795	46592	24336	22256	35.7%
scientificContactFirstname	62387	39005	23382	17396	5986	9.6%
scientificContactLastname	62387	21884	40503	17915	22588	36.2%
scientificContactAddress	62387	38968	23419	14796	8623	13.8%
scientificContactEmail	62387	33398	28989	12069	16920	27.1%
scientificContactTel	62387	35575	26812	7607	19205	30.8%
scientificContactAffiliation	62387	15328	47059	24712	22347	35.8%
url	62387	1	62386	341	62045	99.5%
dateEnrollement	62387	2101	60286	54193	6093	9.8%
dateRegistration	62387	1173	61214	55406	5808	9.3%
studyType	62387	20	62367	62347	20	0.0%
studyDesign	62387	5856	56531	46415	10116	16.2%
phase	62387	27304	35083	34988	95	0.2%
primaryOutcome	62387	1819	60568	2851	57717	92.5%
secondaryOutcomes	62387	11966	50421	1376	49045	78.6%
resultsSummary	62387	56088	6299	112	6187	9.9%
resultsUrlLink	62387	56937	5450	30	5420	8.7%
resultsIpdPlan	62387	51285	11102	11074	28	0.0%
resultsIpdDescription	62387	57778	4609	1451	3158	5.1%
sourceSupport	62387	6199	56188	42819	13369	21.4%
dateCompletion	62387	54368	8019	4786	3233	5.2%
recruitmentStatus	62387	385	62002	61955	47	0.1%

2. Example of an SNCTP entry with BASEC (red) and ICTRP (blue) features

Feature / column	Data
date	2020-12-20 01:04:34
basecId	nan
snctpId	SNCTP000001408
whoId	NCT02254967
ECName	EC_TI
ecFinalDecisionDate	nan
WHO register	NCT
flagRareDisease	0
flagForChildren	0
flagForAdolescents	0
flagForHealthy	0
published	1
contactName	Prof. Dr. med. A B
contactMail	A.B@eoc.ch
contactPhone	+41 91 811 XXXX
lang	de
layTitle	Studie zur Beurteilung der Wirkung von Vancomycin im Vergleich zur verlängerten Fidaxomicin-Therapie bei der nachhaltigen klinischen Heilung einer Clostridium difficile-Infektion bei älteren Patienten
laySummary	Ihre Teilnahme an der Studie wird etwa 3 Monate dauern. Sie werden gebeten, 6 Besuchstermine beim Studienpersonal wahrzunehmen. Beim ersten Besuchstermin wird sich das Studienpersonal vergewissern, dass Sie an einer Clostridium difficile Infektion leiden. Hierfür werden Sie eine Stuhlprobe für Untersuchungen abgeben. Die Patienten werden nach dem Zufallsprinzip (etwa wie beim Werfen einer Münze) entweder einer Behandlung mit Fidaxomicin oder mit Vancomycin zugeteilt. Wenn Sie in die Behandlungsgruppe mit Fidaxomicin gekommen sind, erhalten Sie Fidaxomicin-Tabletten für 25 Tage. Wenn Sie in die Behandlungsgruppe mit Vancomycin gekommen sind, erhalten Sie Vancomycin-Tabletten für 10 Tage. Sie werden gebeten, während der Behandlungsphase ein Studientagebuch für Patienten auszufüllen und jeden Tag Informationen über Anzahl und Menge der ungeformten Stühle sowie über die Menge der jeden Tag eingenommenen Studienmedikation einzutragen. Beim Besuchstermin am Prüfzentrum wird Ihr Prüfarzt überprüfen, ob Sie immer noch Krankheitszeichen haben. Anschliessend werden Sie bis Tag 90 der Studie nachbeobachtet, um abzuklären, dass Sie kein CDI mehr haben, und um Ihren allgemeinen Gesundheitszustand zu kontrollieren.
disease	Darminfektion mit Bakterium Clostridium Difficile
intervention	Einnahme von Tabletten: Fidaxomicin für 25 Tage oder Vancomycin für 10 Tage
inclusionCriteria	Patient/In muss: mindestens 60 Jahre alt sein, an Diarrhoe leiden mit dem Nachweis einer Infektion mit dem Bakterium Clostridium Difficile, darf nicht an einer anderen klinischen Studie teilnehmen
exclusionCriteria	mehr als 2 Dosen eines Durchfallmittels innerhalb der letzten 24 Stunden, Pat. kann keine Tabletten schlucken
studySitesOther	nan
studysites	Lugano, St Gallen, Zürich
tags	Erkrankungen des Verdauungssystems (nicht Krebs), Infektionen und Parasitenbefall

Feature / column	Data
trialId	NCT02254967
url	https://clinicaltrials.gov/show/NCT02254967
countries	Switzerland, Germany, Turkey, France, Greece, United Kingdom, Austria, Belgium, Ireland, Italy, Portugal, Sweden, Russian Federation, Czech Republic, Finland, Hungary, Norway, Poland, Spain, Romania, Croatia, Denmark, Slovenia, Czechia
publicTitle	A Phase IIIB/IV Study to Compare the Efficacy of Vancomycin Therapy to Extended Duration of Fidaxomicin Therapy in the Clinical Cure of Clostridium Difficile Infection (CDI) in an Older Population
recruitmentStatus	Completed
secondaryId	2013-004619-31;2819-MA-1002
scientificTitle	A Phase IIIB/IV Randomized, Controlled, Open-label, Parallel Group Study to Compare the Efficacy of Vancomycin Therapy to Extended Duration Fidaxomicin Therapy in the Sustained Clinical Cure of Clostridium Difficile Infection in an Older Population
inclusionCriteria.1	 Inclusion Criteria: - CDI is confirmed by clinical symptoms (either > 3 unformed bowel movements or = 200ml of unformed stool (for subjects having rectal collection devices)) in the 24 hours prior to randomization and CDI test confirmed positive for presence of C. difficile toxin A or B in stool within 48 hr prior to randomization. - Subject agrees not to participate in another interventional study whilst participating in this study. Exclusion Criteria: - Subject is taking or requiring to be treated with prohibited medications - Subject has received more than one day of dosing of any therapy for CDI within the last 48 hours - Subject has experienced more than 2 previous episodes of CDI in the 3 months prior to study enrolment - Subject is unable to swallow oral study medication. - Subject has a current diagnosis of toxic megacolon. - Subject is not willing to adhere to the provisions of treatment and observation specified in the protocol. - Subject has been randomized into this study previously, has taken any investigational drug within 28 days or 5 half lives, whichever is longer, prior to enrollment, or is currently participating in another clinical study which may influence the assessment of efficacy and/or safety endpoints of this study, in the opinion of the Sponsor. - Subject has previously participated in a CDI vaccine study - Subject has hypersensitivity to fidaxomicin, vancomycin or any of its components.
exclusionCriteria.1	nan
interventions	Drug: Fidaxomicin;Drug: Vancomycin
dateEnrollement	2014-11-06
dateRegistration	2014-09-25
studyType	Interventional
studyDesign	Allocation: Randomized. Intervention model: Parallel Assignment. Primary purpose: Treatment. Masking: None (Open Label).
primarySponsor	Astellas Pharma Europe Ltd.
secondarySponsors	Merck Sharp & Dohme Corp.
phase	Phase 4
healthConditions	Clostridium Difficile
primaryOutcome	Percentage of Participants with a Sustained Clinical Cure of CDI at 30 Days after End of Treatment
secondaryOutcomes	Disease-free Survival After Day 10;Time to Recurrence of CDI after End of Active Treatment;Percentage of Participants with a Recurrence of CDI at Day 40, Day 55 and Day 90;Time to Resolution of Diarrhea (TTROD);Number of Participants with a Relapse on Day 90 as Determined by Whole Genome Sequencing of C. Difficile Isolates;Percentage of Participants with a Clinical Response of CDI at Day 12;Percentage of Participants with a Clinical Response of CDI at 2 Days after End of Treatment;Percentage of Participants with a Sustained Clinical Cure of CDI at Day 40, Day 55 and Day 90
resultsSummary	nan
resultsUrlLink	nan

Feature / column	Data
resultsIpdPlan	Yes
resultsIpdDescription	Access to anonymized individual participant level data collected during the trial, in addition to study-related supporting documentation, is planned for trials conducted with approved product indications and formulations, as well as compounds terminated during development. Conditions and exceptions are described under the Sponsor Specific Details for Astellas on www.clinicalstudydatarequest.com.
sourceSupport	Please refer to primary and secondary sponsors
alternativeNames	nan
dateCompletion	nan
publicContactFirstname	nan
publicContactLastname	Medical Director
publicContactAddress	nan
publicContactEmail	nan
publicContactTel	nan
publicContactAffiliation	Astellas Pharma Europe Ltd.
scientificContactFirstname	nan
scientificContactLastname	Medical Director
scientificContactAddress	nan
scientificContactEmail	nan
scientificContactTel	nan
scientificContactAffiliation	Astellas Pharma Europe Ltd.

References and Bibliography

[1] Fernández-González, L. Registering transparency: the making of the international clinical trial registry platform by the world health organization (2004–2006). *Global Health* **19**, 71 (2023). https://doi.org/10.1186/s12992-023-00970-5

[2] https://www.fedlex.admin.ch/eli/cc/2013/617/en

[3] https://www.fedlex.admin.ch/eli/cc/2013/643/en

[4] https://www.who.int/clinical-trials-registry-platform

[5] https://www.who.int/clinical-trials-registry-platform/network/who-data-set

[6] https://www.clinicaltrials.gov/

[7] https://swissethics.ch/en/basec

[8] https://kofam.ch/en/snctp-portal/searching-for-a-clinical-trial

[9] https://scholar.harvard.edu/sites/scholar.harvard.edu/files/dell/files/linktransformer.pdf

[10] https://en.wikipedia.org/wiki/Word_embedding

[11] https://github.com/Rinderkm/CAS-Github-Project/tree/main/Final_Project

[12] Arora, Abhishek and Dell, Melissa. (2023). LinkTransformer: A Unified Package for Record Linkage with Transformer Language Models, arXiv eprint, https://doi.org/10.48550/arXiv.2309.00789; https://linktransformer.github.io/

[13] https://huggingface.co/models

[14] https://github.com/dell-research-harvard/linktransformer?tab=readme-ov-file#getting-started

[15] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[16] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

[17] https://github.com/facebookresearch/faiss

[18] Douze, M. et al. (2024). The Faiss library. arXiv eprint, https://doi.org/10.48550/arXiv.2401.08281

[19] https://huggingface.co/dell-research-harvard/lt-wikidata-comp-multi/discussions/1#662263d82e46887f72c2f08e

[20] Cabrera-Diego, Luis Adrián & Moreno, Jose & Doucet, Antoine. (2021). Simple Ways to Improve NER in Every Language using Markup. 10.5281/zenodo.4680998. https://ceur-ws.org/Vol-2829/paper2.pdf

[21] https://econabhishek.github.io/