Contents

List of Figures

List of Tables

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Chapter 1

Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Chapter 2

Linking Related Documents from Different Sources

2.1 Introduction

The goal of this chapter is to evaluate the feasibility of creating a dataset where the marketing authorization status of a certain medicine is automatically linked to the relevant clinical trials that were conducted to test this medicine and to the academic publications reporting the results of these trials. Currently, this information is dispersed in various 'data islands' that are not easily linked to each other: the marketing authorization status is found on the websites of (inter)national regulatory authorities (e.g. US Food and Drug Administration (FDA), European Medicines Agency (EMA)), information about clinical trials is found in designated registers (e.g. ClinicalTrials.gov, EU Register), and academic publications are found on the websites of various journals and in the PubMed database.

Linking these complementary sources of information to each other is of interest both to healthcare professionals and the general public. For example, the active ingredient *abaloparatide* (brand name *Eladynos*), which is intended as a treatment for *postmenopausal osteoporosis*, was refused marketing authorization in the EU in 2019. The decision was based on data from one pivotal Phase III clinical trial and supportive data from two Phase II studies and one extension Phase III study.[1] Interestingly, the same medicine (under

[1] https://www.ema.europa.eu/en/medicines/human/EPAR/eladynos (accessed 15-07-20)

the brand name *Tymlos*) was approved for marketing in the USA in 2017, based on the same four clinical trials.[2] In view of this, one could imagine that a doctor, a researcher or a patient would be interested in looking into the clinical trials that led to the discrepant regulatory decisions. However, neither EMA nor FDA provide links or identifiers that would allow to easily find the relevant trials in a register or the relevant scientific publications on PubMed.

This chapter explores three possible methods to address this issue, i.e. connect between drug authorizations (specifically - EMA), clinical trial registers (specifically - ClinicalTrials.gov, henceforth CTgov) and academic publications (PubMed). Each method approaches the task from a different route (see Figure 2.1):

- The *EPAR_CTgov* method extracts protocol numbers from EMA authorizations, and uses them to search the CTgov register. It then utilizes cross-references from CTgov to PubMed to identify related publications.
- The *PubMed_API* method extracts drug and disease names from EMA authorizations, and uses them to query PubMed. It then utilizes cross-references from PubMed to CTgov to identify related clinical trials.
- The *CUI_grouping* method extracts and normalizes drug and disease names from the free text of records from all three sources. It then groups together records that mention the same disease and drug names.

The experiments presented in this chapter reveal that each of these methods has its strengths and its limitations. None of the methods manages to consistently outperform the other two, and in many cases they seem to complement each other. While the task is far from being solved, the in-depth analysis provided in the chapter sheds light on each of the databases and on possible links between them. Hopefully, this information can be useful in future work on connecting the (partially) isolated biomedical databases, so that important insights can be easily and reliably accessed by all stakeholders.

The chapter is organized as follows. Section 2.2 summarizes existing work on the problem of linking biomedical data sources. Section 2.3 provides background information about the three databases: EMA drug register, CTgov

[2] https://www.accessdata.fda.gov/drugsatfda_docs/nda/2017/208743Orig1s000TOC.cfm (accessed 15-07-20)

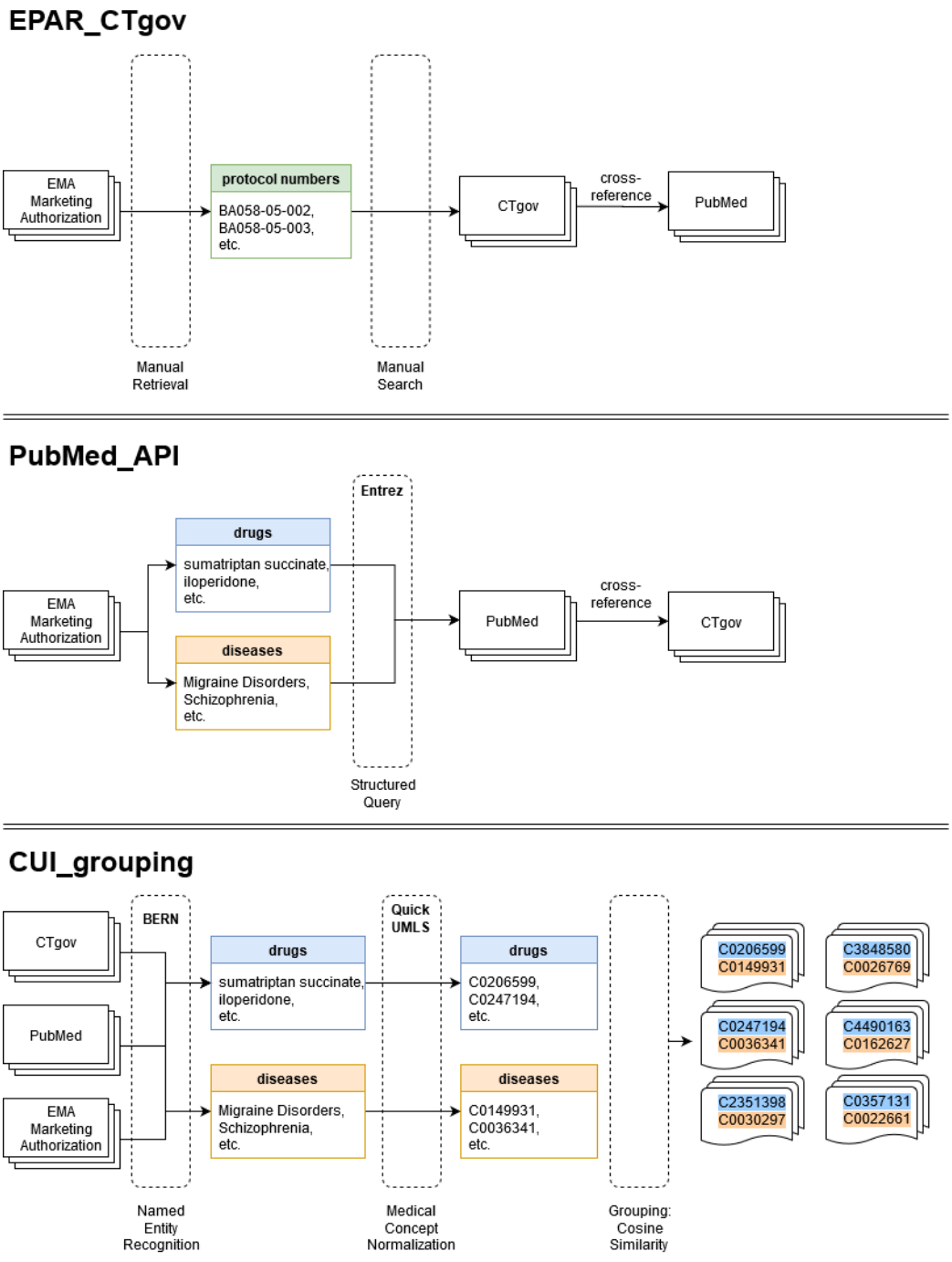

Figure 2.1: Overview of the three methods used for linking the data sources

and PubMed. Section 2.4 outlines the overall setup of the experiment. Section 2.5, Section 2.6 and Section 2.7 describe each of the three methods and the results obtained using it. Section 2.8 compares between the three methods, both quantitatively and qualitatively. Section 2.9 concludes the findings reported in the chapter.

2.2 Related Work

Biomedical research is characterized by a wealth of available data; however, discovering meaningful patterns and generating insights is extremely challenging, since the data are heterogeneous and dispersed over numerous isolated databases. The importance of linking these 'data islands' has been long recognized by the research community. Since the early 2000s, there have been significant efforts to connect biomedical data sources using the *Semantic Web* (Berners-Lee et al. 2001) and *Linked Data*[3] framework. The proposed solutions automatically download data from known biomedical databases (including DrugBank, DailyMed, PubMed, CTgov and others) and convert them from their original formats into RDF triples. These triples describe the relations and properties of each entity and allow building sophisticated data spaces of interlinked entities; the network can be then queried with the SPARQL query language.

A few systems have been implemented which focus on converting biomedical datasets into RDF; for example Bio2RDF (Belleau et al. 2008), which offers RDF versions of PubMed and CTgov, among others. LinkedCT (Hassanzadeh et al. 2009) is another example of a system that converts CTgov data to RDF and links it to various other data sources, including PubMed. TripleMap (Samwald et al. 2011) integrates LinkedCT (and other RDF datasets) into a web application that allows users to navigate, visualize and analyze the data. Unfortunately, both LinkedCT and TripleMap are currently not publicly available and the Linking Open Drug Data (LODD)[4] task force within the W3C's *Health Care and Life Sciences Interest Group*, which was behind these initiatives, also seems to be inactive in the last years.

The current research adds two main novelties to the existing research on linking biomedical data. First, it incorporates marketing authorization data, which to the best of my knowledge has not been attempted before.

[3] https://www.w3.org/wiki/LinkedData
[4] https://www.w3.org/wiki/HCLSIG/LODD

Second, while existing systems utilize the structured data available in the various databases, the current research examines the limitations of these structured links and explores the option to utilize information extracted from free text. Therefore, despite the limited scope of the experiment presented in this chapter, it offers a fresh point of view on this well-researched problem.

2.3 Data Sources

The methods described later in the chapter attempt to link information from three databases: EMA's medicine register, CTgov, and PubMed. In the following sections, I shortly present each of these sources and outline the considerations that led to choosing them over other comparable counterparts.

2.3.1 EMA Marketing Authorizations

For marketing authorization information, the two main candidates are the US FDA and EMA, since these are two major pharmaceutical markets for which data in English are available. The reason EMA was selected is that it publishes data both for medicines that were granted marketing authorization and for medicines that were refused it, while FDA only publishes data about medicines that were approved for marketing; for the second part of this research, I need examples of both approved and rejected medicines.

EMA (https://www.ema.europa.eu/en) provides a downloadable tabular summary of the European Public Assessment Reports (EPARs) of medicines submitted for authorization at a European Union level (see here). The tabular summary contains information such as the active substance, the disease for which the medicine is intended, the authorization status and the name of the sponsor. In addition, the full EPAR of each medicine is available on the medicine's web-page (e.g. *Eladynos*); this is usually a PDF document of 150-200 pages.

2.3.2 CTgov

CTgov (https://clinicaltrials.gov/) is a web database of clinical trials maintained by the US National Library of Medicine (NLM) at the National Institutes of Health (NIH). The information on CTgov is provided and updated by the sponsors or the principal investigators of clinical trials. CTgov offers

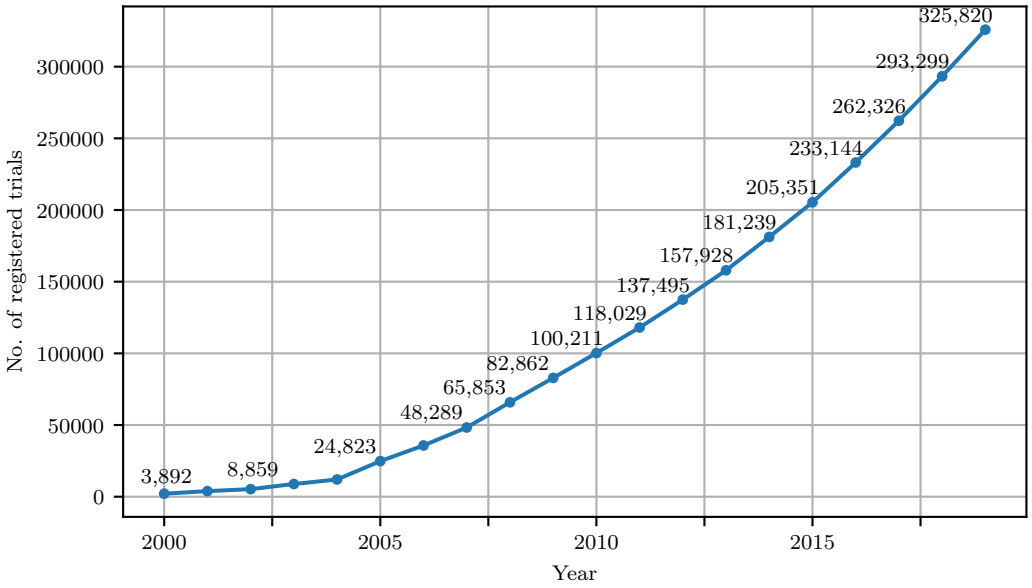

Figure 2.2: Number of registered trials over time (as of October 02, 2020); source: https://clinicaltrials.gov/ct2/resources/trends

an API which allows to display and download search results or the entire database in XML format (see here).

Figure 2.2 shows the total number of trials registered on CTgov since its establishment in 2000. The rate of study registration has increased over time as more policies and laws requiring registration have been enacted and as more sponsors and investigators have voluntarily registered their studies. As evident from Figure 2.2, there has been a big jump in registration numbers in 2005, when the International Committee of Medical Journal Editors (ICMJE) began requiring trial registration as a condition for publication. Another jump was around 2007-2008 after the US Congress passed the FDA Amendments Act, which requires more types of trials to be registered and additional trial information to be submitted.

Although the database is maintained by an American organization, it is a highly relevant source of information about trials worldwide, and not only in the US. Currently, the database lists over 350,000 trials conducted in 216 countries. In fact, as shown in Figure 2.3, half of the trials registered on CTgov are not taking place in the US at all.

Although EMA has its own clinical trials register, I decided to use CT-

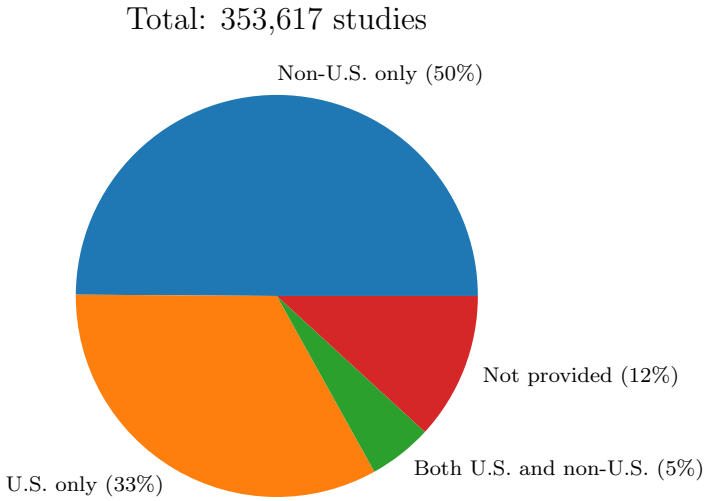

Figure 2.3: Percentage of registered trials by location (as of October 02, 2020); source: https://clinicaltrials.gov/ct2/resources/trends

gov since its records sometimes cross-reference to related PubMed articles, a feature that is utilized in one of the methods I experiment with.

2.3.3 PubMed

PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) is a search engine accessing the MEDLINE database, which contains academic publications from various biomedical disciplines including medicine, pharmacology, molecular biology and others. Like CTgov, PubMed is maintained by the US National Library of Medicine (NLM) at the NIH. PubMed contains about 30 million records from 1950 to the present; some records are citations only, some include abstracts and some include access to full-text articles.

PubMed data can be accessed and downloaded in several ways, including the E-utilities API (see here) and FTP servers (see here). The data is available in structured formats (e.g. xml) which contain various fields, such as publication type (e.g. 'Clinical Trial') and cross-references to other resources, including CTgov.

2.4 Experimental Setup

For the experiments described in this chapter, 72 EMA drug authorizations were selected; the three methods showed in Figure 2.1 were then used to link CTgov and PubMed records to these 72 authorizations. Since there is no gold standard of correct answers to evaluate against, the methods are evaluated comparatively to each other. For each method, the following quantitative metrics are calculated:

- The *linking rate* of the method: How many out of the 72 EMA authorizations were linked to at least one CTgov record? How many were linked to at least one PubMed record?
- The *total retrieval* of the method: How many unique CTgov / PubMed records were linked in total (to all authorizations together)?
- What is the median number of CTgov / PubMed records retrieved per EMA authorization?
- What is the maximum number of CTgov / PubMed records retrieved per EMA authorization?
- How much overlap (intersection) is there between the CTgov / PubMed records retrieved by the three methods?

In addition to the quantitative comparison, 4 out of the 72 authorizations are analyzed qualitatively. For these four examples, I manually check the CTgov and PubMed records linked by each method (if there are many records, a random sample of them is checked). The purpose of the manual analysis is to determine:

- Are the linked CTgov / PubMed records correct, i.e. related to the authorized drug?
- Are the correct records linked by all methods? If not - why does a certain method fail to link a certain correct record to the authorization?
- If there are incorrectly linked records - why does a certain method link a certain incorrect record to the authorization?

Based on the quantitative results and the in-depth analysis of the examples, I deduce: (a) the main strengths of each method, (b) the main weaknesses / limitations of each method, (c) the estimated precision of each method, i.e. what fraction of the retrieved CTgov / PubMed records is correct, and (d) the estimated recall of each method, i.e. what fraction of correct CTgov / PubMed records is retrieved.

The 72 EMA authorizations were selected out of the 1,410 available authorizations of human medicines (as of January 2020). From the total 1,410 records, I selected all records with a 'refused' status (N=46), as well as a random sample of comparable size and comparable time period of 'approved' records. It was then manually checked whether results of clinical trials were submitted as part of the authorization process (for biosimilar drugs or generic drugs, this is not the case); if not, the records were excluded. This procedure resulted in 72 authorizations, submitted between the years 2006 and 2019; 35 of these applications were approved and 37 of the applications were refused authorization. The full list of the 72 authorizations is found in Appendix A.

2.5 The *EPAR_CTgov* Method

This section describes the first method for linking clinical trials and publications to medicine authorizations. The method involves extracting protocol numbers from EMA authorizations, and using them to search the CTgov register. It then utilizes cross-references from CTgov to PubMed to identify related publications. The workflow is schematically summarized in Figure 2.1 above, under the name '*EPAR_CTgov*'.

The procedure was performed manually, as part of the initial data exploration for the project, but most of the steps can be automated, as described in Section 2.5.2 below.

2.5.1 Data

For this experiment, the full-length EPARs (PDF format) of the 72 EMA authorizations are used.

2.5.2 Method

The manual linking procedure is outlined below with an example:

- *Step 1: Find the full-length EPAR for the medicine.*
  The downloadable tabular summary of medicine authorizations contains a URL linking to each medicine's information page on the website; for example *Eladynos* info page. On this page, find the full EPAR related to the marketing authorization; this is a PDF document of 150-200 pages. Example: *Eladynos* EPAR.

- *Step 2: Find all protocol numbers in the EPAR.*
  In the report, find the section discussing the clinical data submitted with the application. In the *Eladynos* example, this is section *2.4 Clinical aspects*. In this section, there is usually a table or a paragraph mentioning the protocol numbers of the relevant clinical trials. In our example, three main studies are mentioned in the introduction paragraph ('BA058-05-002', 'BA058-05-003', 'BA058-05-005'); additional studies are mentioned in Table 3 and Table 4. In total 9 unique studies of different phases are mentioned for *Eladynos*; note that they are sometimes referred to by an abbreviated form, e.g. '003' instead of 'BA058-05-003'.

- *Step 3: Find the CTgov records corresponding to the protocol numbers.*
  On the CTgov website, search for the protocol numbers identified in *Step 2*. If successful, this leads to the trial's information page. For example, searching for 'BA058-05-002' leads to this CTgov record. The protocol number ('BA058-05-002') is listed under 'Other Study ID Numbers', since the primary identifier used in CTgov and other NLM resources (e.g. PubMed) is the NCT ID (in our case NCT00542425). To verify that you found the correct record, you can compare information about the study phase, the active ingredient, the disease, the sponsor, the number of participants, etc.

- *Step 4: Find the PubMed publications related to the CTgov records.*
  Some CTgov records contain links to related publications on PubMed. In our example, there is a link to publication PMID 25393645 (see under 'More Information' at the bottom of the page). The linked publications discuss either the results of the clinical trial (or multiple clinical trials) or other aspects of the trial, such as analyses of costs, motivation for the study design, etc. For this experiment, each publication was manually scanned and publications that do not discuss clinical results were excluded.

As mentioned above, the procedure was performed manually but most of it can be easily automated. To automate *Step 1*, the URL provided in the tabular data can be used to scrape the EPAR PDF from the webpage. To automate *Step 3*, the structured XML version of the CTgov record can be used; the protocol number is registered in the structured data within the `org_study_id` tag, see example here. The XML format can also be utilized to automate *Step 4*; cross-references to PubMed are registered in the structured data within the `PMID` tag, see example here. However, this will result in a lower retrieval rate than reported here, since the structured data contains less links to PubMed than the webpage. The structured data only contains the PubMed references that were manually entered by the person maintaining the trial record (i.e. the sponsor or the investigator), while the webpage automatically includes links to all PubMed records that mention the trial's NCT ID, in addition to the manually-entered links (Huser and Cimino 2013). For example, the web-page of NCT00542425 links to PubMed publication PMID 25393645, but the XML version of the same record does not; this is so because the cross-reference to PubMed was not manually entered into the structured data, but linked automatically due to the NCT ID being mentioned in the abstract.

The most problematic step in terms of automating is *Step 2*, i.e. extracting protocol numbers from the EPAR. This step is not easily automated because the location of the protocol numbers (section/table/etc.) is not uniform across EPARs; moreover, the tables in which the numbers are summarized are often low-quality scans.

2.5.3 Results

Table 2.1 summarizes the quantitative results obtained with the *EPAR_CTgov* method. As evident from the table, the method has a very good linking rate: it manages to link CTgov records to 90% of the 72 EMA authorizations and PubMed records to 88% of the authorizations. This results in a total of 451 unique CTgov records and 453 unique PubMed publications. The median number of CTgov records linked per authorization is 5 (excluding the cases when no records were linked), and the maximum is 42 CTgov records for one EMA authorization. The median number of PubMed records per authorization is 4, with a maximum of 59 publications linked to one EMA authorization.

In *Step 2* of the procedure described above, 844 protocol numbers were re-

	CTgov	PubMed
No. EMA authorizations for which records were found	65 (90%)	63 (88%)
Total no. of unique found records	451	453
Median no. of found records per EMA authorization	5	4
Max no. of found records per EMA authorization	42	59

Table 2.1: Summary of quantitative results: *EPAR_CTgov* method

trieved from the EPARs of the 72 EMA authorizations. In *Step 3*, I searched CTgov for these 844 protocol numbers; only for 451 of them a corresponding CTgov record was found. Failure to find a protocol number mentioned in an EPAR on CTgov can be related to either (a) a problem in the EPAR, i.e. incorrect or missing protocol number, (b) the trial being registered on CTgov without its protocol number, or (c) the trial not being registered on CTgov at all. While it is hard to evaluate the exact contribution of each of these factors to the reported retrieval rate, the general trend is that a bigger percentage of trials is identified on CTgov for more recent authorizations. This trend is not surprising in view of the data presented in Figure 2.2 in Section 2.3.2 above: CTgov exists since 2000 and the registration rate went up significantly over the years. The 72 authorizations took place between 2006 and 2019, meaning that the related studies were conducted a few year before that (Phase I trials might have taken place even up to 10-15 years before the authorization); therefore, for earlier authorizations, the chances of the trials being registered on CTgov are smaller.

About 45% of the linked CTgov records are Phase III studies, and the rest is Phase I studies (about 26%) and Phase II studies (about 27%). This distribution is directly related to the method itself: the trials that are always mentioned in the EPAR are the so-called 'pivotal trials' that are almost always Phase III; the non-pivotal Phase I and Phase II studies are only sometimes mentioned in the EPARs. Phase IV studies are conducted post-authorization and therefore are almost never mentioned in the EPAR.

As mentioned in Section 2.4, 37 out of the 72 EMA records are refused au-

thorizations and 35 are approved authorizations. Interestingly, 312 of the 453 linked PubMed records (69%) are related to approved authorizations. This might suggest that trials with non-significant or negative findings, which eventually led EMA to refuse the marketing authorization, tend to be published less, a phenomenon known as 'publication bias' (Simes 1986).

2.6 The *PubMed_API* Method

This section describes the second method for finding clinical trials and publications related to a specific medicine authorization. The method involves extracting drug and disease names from the authorizations, and using them to query the PubMed E-utilities API. It then utilizes cross-references from PubMed to CTgov to identify related clinical trials. The workflow is schematically summarized in in Figure 2.1 above, under the name '*PubMed_API*'.

2.6.1 Data

For this experiment, the 'active substance' and 'therapeutic area' fields of the 72 EMA authorizations are used. The fields are part of the structured tabular summary available on the EMA website.

2.6.2 Method

Entrez/E-utilities[5] is a data retrieval system of the NCBI, a division of the US National Library of Medicine (NLM). The system provides access to various biomedical databases, including PubMed. The data can be accessed manually with a URL syntax from a web browser, or programmaticly. For this experiment, I used the latter option, utilizing the Biopython[6] library.

For each of the 72 EMA authorizations, the names of the *drug* (active substance) and the *disease* (therapeutic area) were extracted from the tabular data and inserted into the following query:

```
drug [SUBS]
AND disease [MESH]
AND English [LANG]
```

[5] https://www.ncbi.nlm.nih.gov/books/NBK25497/
[6] https://biopython.org/

```
AND (
   Clinical Study[PTYP] OR
   Clinical Trial[PTYP] OR
   Randomized Controlled Trial[PTYP] OR
   Controlled Clinical Trial[PTYP] OR
   Clinical Trial, Phase I[PTYP] OR
   Clinical Trial, Phase II[PTYP] OR
   Clinical Trial, Phase III[PTYP] OR
   Clinical Trial, Phase IV[PTYP]
)
```

The square brackets indicate the structured fields where the term appearing before the brackets should be searched: `[SUBS]` refers to the 'substance name' field, `[MESH]` refers to the 'MeSH Terms' field[7], `[LANG]` refers to the language of the publication, and `[PTYP]` refers to the publication type[8]. In addition, the retrieved data was limited to papers published from 2000 onwards; this makes sense given the authorization dates (2006-2019) and the dates of the publications found by the previous method (2001-2020).

The system returns the PMIDs of the papers matching the query and allows to access further structured information about these PMIDs. One of the structured fields contains cross-references to CTgov identifiers (NCT ID's) of trials related to the publication. These CTgov identifiers are retrieved as well.

2.6.3 Results

Table 2.2 shows the quantitative results obtained through the procedure described above. For 46 out of the 72 EMA authorizations (64%), at least one PubMed record was found by running the above-mentioned query to the PubMed API, resulting in a total of 1,113 unique PubMed publications. For 41 out of the 72 EMA authorizations (57%), at least one CTgov record was

[7]This field contains the main topics discussed in the article, in the form of 'Medical Subject Headings', a thesaurus for life sciences terms.

[8]The categories I considered as related to clinical trials were chosen from the list of MeSH pubtypes (https://www.nlm.nih.gov/mesh/pubtypes.html) by searching for the keywords "trial" and "clinical" and excluding veterinary trials and other irrelevant types. Limiting the search to these publications types probably results in some data loss. However, since limiting the publication type was necessary in the *CUI_grouping* method (Section 2.7) to keep the data volume manageable, it was done here as well.

	CTgov	PubMed
No. EMA authorizations for which records were found	41 (57%)	46 (64%)
Total no. of unique found records	256	1,113
Median no. of found records per EMA authorization	3	12
Max no. of found records per EMA authorization	32	176

Table 2.2: Summary of quantitative results: *PubMed_API* method

identified through a structured link from a publication, resulting in a total of 256 unique CTgov records.

Compared to the quantitative results of the *EPAR_CTgov* method (Section 2.5.3), the *PubMed_API* method has a lower linking rate, i.e. there are more EMA authorizations to which no records are linked. Nonetheless, it finds twice as many PubMed records in total: 1,113 vs. 453 found by *EPAR_CTgov*. Similarly to the results of *EPAR_CTgov*, 69% (764 out of 1,113) of the PubMed records detected by *PubMed_API* are related to approved authorizations; this provides further support for a suspected publication bias.

For CTgov records, the opposite trend is observed: the *PubMed_API* method links half the number of CTgov records: 256 vs. 451 found by *EPAR_CTgov*. As explained in Section 2.6.2, the method retrieves CTgov NCT ID's from the structured data of PubMed; this structured field is automatically created when the authors of the publication mention an NCT ID in the original paper.[9] The fact that only 256 CTgov records are mentioned in 1,113 papers about clinical trial results suggests that the cross-references from PubMed to CTgov are very partial. This is further supported by the following little experiment: I took the 453 PubMed records found by the *EPAR_CTgov* method and retrieved the structured links to CTgov that they contain; this resulted in 273 unique CTgov records. Recall that the 453 PubMed records were found through links retrieved from 451 CTgov records; nonetheless, a reverse search retrieves only about 60% of these CTgov records, showing

[9] https://www.nlm.nih.gov/bsd/policy/clin_trials.html

that the cross-references are not symmetrical (i.e. PubMed records that are referenced on CTgov do not necessarily mention the relevant NCT ID).

2.7 The *CUI_grouping* Method

This section describes the third method used for linking EMA authorizations to CTgov and PubMed records. The method involves extracting and normalizing drug and disease names from the free text of records from all three sources. It then groups together records that mention the same disease and drug names. The workflow is schematically summarized in in Figure 2.1 above, under the name '*CUI_grouping*'.

2.7.1 Data

The data used for the *CUI_grouping* experiment include: (a) the downloadable tabular summary of the 72 EMA authorizations, specifically the 'active substance' and 'therapeutic area' fields of each authorization, (b) a subset (N=118,689) of the full CTgov data dump downloaded in October 2019, and (c) a subset (N=419,894) of the full PubMed data dump downloaded in August 2018.[10]

The workflow for creating the CTgov dataset is summarized in Figure 2.4. From the full CTgov data dump, downloaded from the website in October 2019, a subset was selected based on the following criteria: (a) the intervention type is 'drug' (as opposed to behavioral change, procedure, etc.), and (b) the end date of the trial is from 01-01-2000 onward. The titles and short summaries of the selected records were processed to detect biomedical named entities (i.e. names of diseases and drugs) and to convert the detected terms into normalized medical concepts (UMLS Concept Unique Identifiers - CUI's); the processing procedure is described in detail in Section 2.7.2; records that came out of the processing with no CUI's were excluded.

As shown in Figure 2.5, the resulting dataset of 118,689 CTgov records contains clinical trials from different phases of the drug development process, with Phase II studies being the largest group (N=32,080, 27%). In terms of the time period (Figure 2.6), the distribution is in accordance with what was discussed in Section 2.3.2: more and more studies over time. Specifically,

[10]The CTgov and PubMed data dumps used here were provided by myTomorrows in the course of an internship I followed there in September 2019 - January 2020.

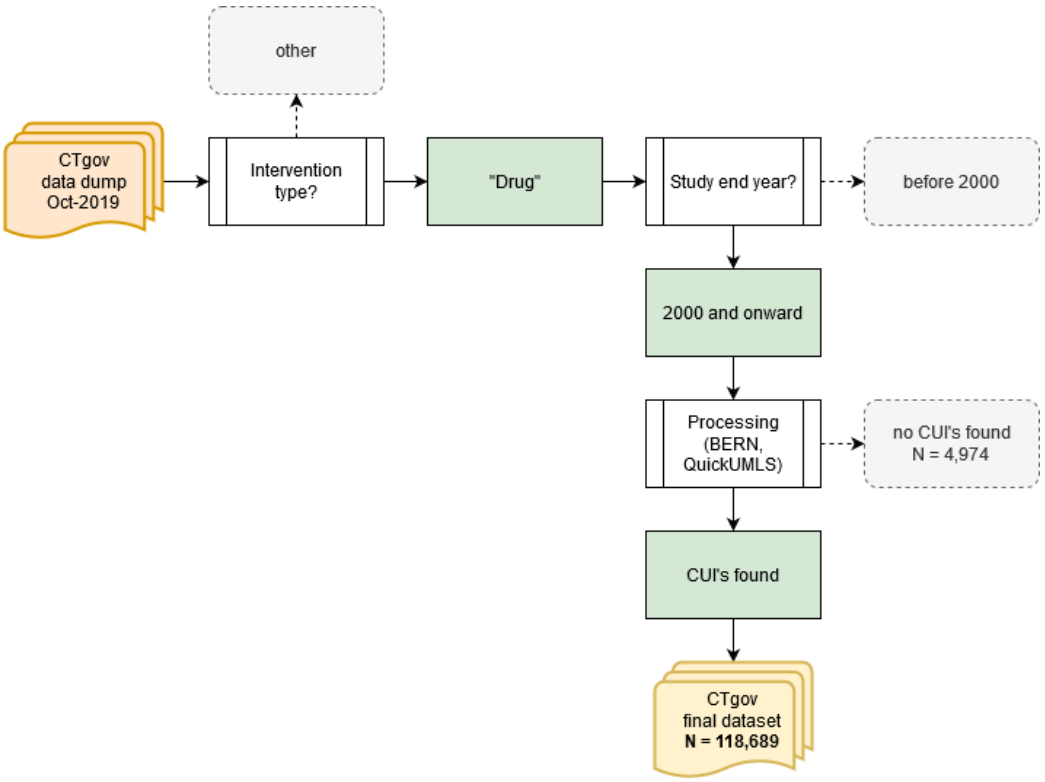

Figure 2.4: Creating the CTgov dataset for the *CUI_grouping* experiment

72% of the studies have their end date between 2007 and 2019, 19% of the studies are still ongoing (the end date of the study is 2020 or later), and only 9% of the studies ended between 2000 and 2007.

The workflow for creating the PubMed dataset is summarized in Figure 2.7. From the full PubMed data dump, downloaded from the website in August 2018, a subset was selected based on the following criteria: (a) the language of the publication is English, (b) the publication date is from 01-01-2000 onward, (c) the publication is tagged with at least one of the 'clinical trial' categories, i.e. 'Clinical Study', 'Clinical Trial', 'Randomized Controlled Trial', 'Controlled Clinical Trial', 'Clinical Trial, Phase I', 'Clinical Trial, Phase II', 'Clinical Trial, Phase III', 'Clinical Trial, Phase IV'. The titles and abstracts of the selected records were processed to detect biomedical named entities and to convert the detected terms into normalized medical concepts (CUI's), as described in detail in Section 2.7.2; records that came

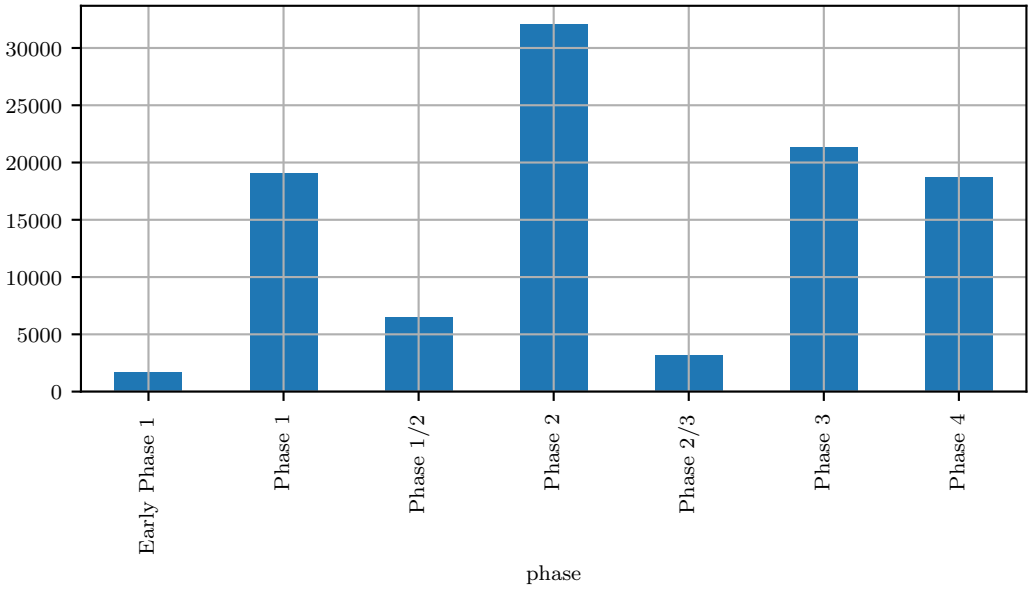

Figure 2.5: CTgov dataset: distribution by study phase

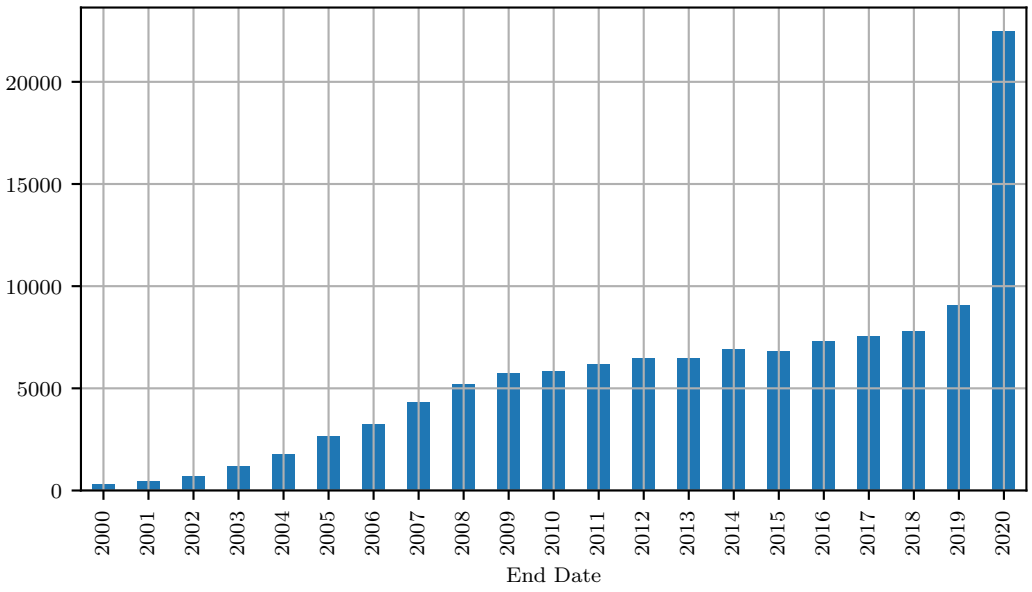

Figure 2.6: CTgov dataset: distribution by end date. Ongoing studies (end date in 2020 or later) are grouped under '2020'.

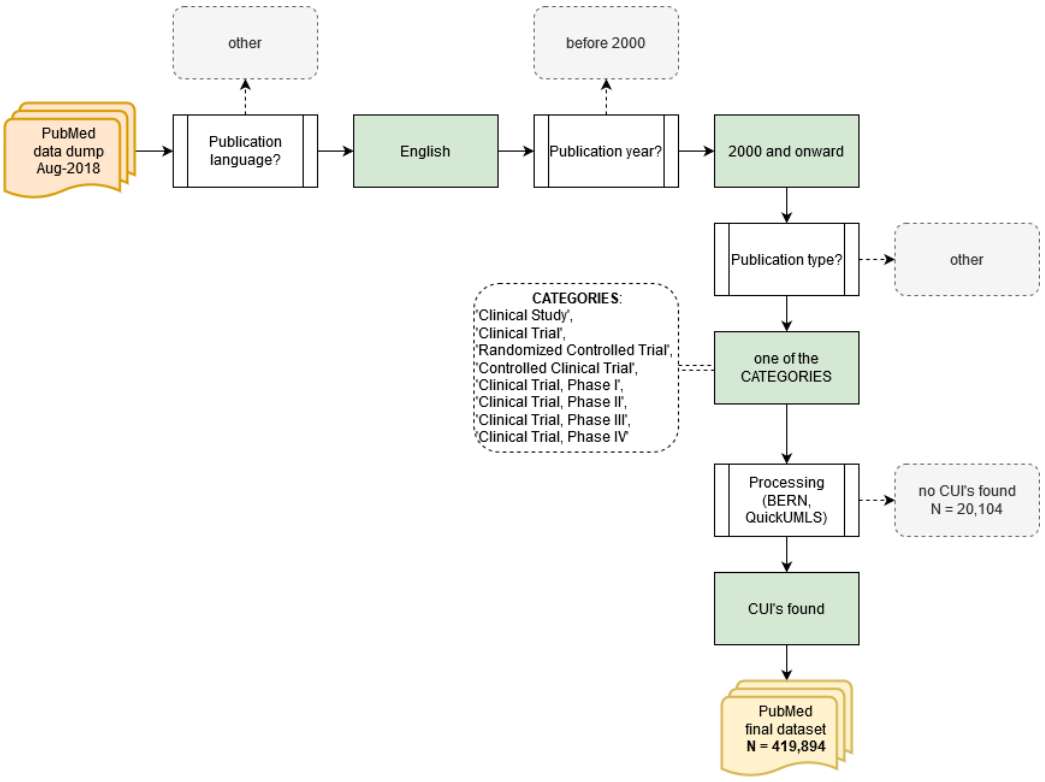

Figure 2.7: Creating the PubMed dataset for the *CUI_grouping* experiment

out of the processing with no CUI's were excluded.

The distribution of the resulting 419,894 PubMed records by publication year is shown in Figure 2.8. As evident, the data for 2018 is very partial, even though the data was downloaded in August 2018. This is probably due to the fact that I relied on tags (e.g. publication type) to select the data; some of these tags are added manually and therefore it might take a few months before they are present. The fact that the PubMed data from 2018 onward is partial / missing has consequences for the evaluation of the method, as discussed below.

2.7.2 Method

As mentioned above, the *CUI_grouping* method involves grouping records based on similarity between their drug and disease entities. These biomedi-

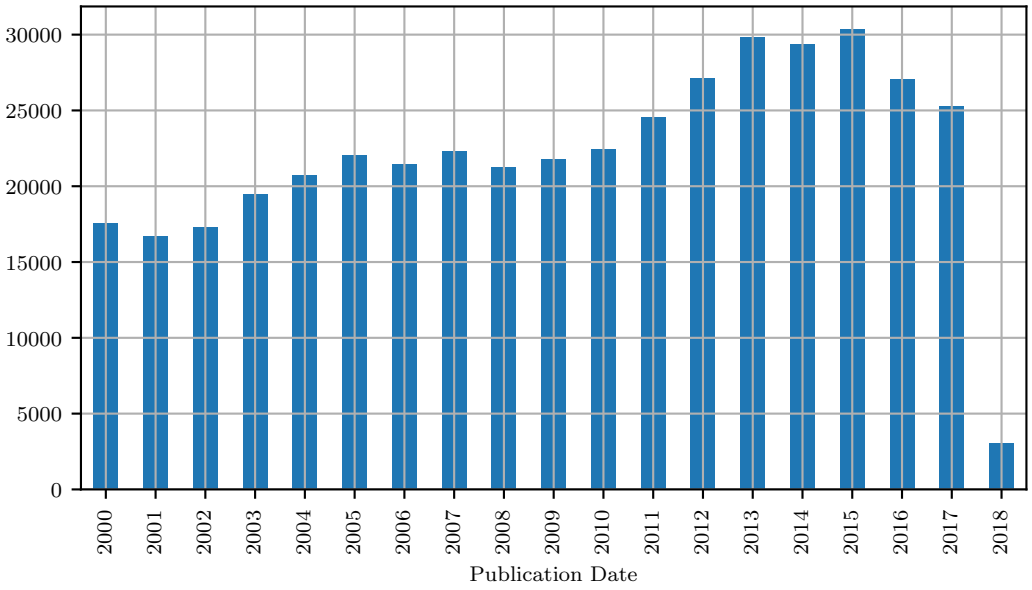

Figure 2.8: PubMed dataset: distribution by publication date

cal entities can be obtained from the data in a few ways; Appendix B reviews the various options in detail. The method chosen for the current experiment obtains the drug and disease entities from the free text of the records. Table 2.3 shows which text was utilized for each record type, including an example.

The first step in processing the text is to detect mentions that refer to drugs and diseases, a task known as biomedical named entity recognition (Bio-NER). For this task, I used the tool BERN (Kim et al. 2019), which utilizes the pre-trained biomedical language representation model BioBERT (Lee et al. 2020). BERN is the current state-of-the-art for Bio-NER, outperforming (in terms of F1-score) other tools in detection of both drug entities and disease entities. BERN was applied to the free text of CTgov and PubMed records to extract drug and disease entities; this step was not necessary for the EMA data since the entities were already given in a structured way.

However, detecting mentions of drugs and diseases is not enough for efficient linking. One of the challenges of medical language is that multiple terms can be used to refer to the same medical concept. For example, '*Hb-SS disease*', '*Herrick syndrome*' and '*sickle cell anemia*' all refer to the same

Dataset	No. records	Utilized textual data	Example
CTgov	118,689	Brief title	*Study to Evaluate the Safety and Efficacy of BA058 (Abaloparatide) for Prevention of Fracture in Postmenopausal Women*
		Official title	*A Randomized, Double-blind, Placebo-Controlled, Comparative Multicenter Phase 3 Study to Evaluate the Safety and Efficacy of BA058 (Abaloparatide) for Injection for Prevention of Fracture in Ambulatory Postmenopausal Women With Severe Osteoporosis and at Risk of Fracture*
		Brief summary	*The purpose of this study is to determine whether BA058 (abaloparatide), a parathyroid hormone-related peptide, is effective in preventing fractures in postmenopausal women with severe osteoporosis who are at risk of fractures.*
PubMed	419,894	Article title	*Effects of abaloparatide-SC (BA058) on bone histology and histomorphometry: the ACTIVE phase 3 trial.*
		Abstract	*There are a number of effective treatments for osteoporosis but most are in the antiresorptive class of compounds. Abaloparatide-SC is a new osteoanabolic agent, which increased bone mineral density and lowered the risk of osteoporosis-related fractures in the phase 3 ACTIVE trial. The objective of this report is to describe the effects of abaloparatide-SC (...)*
EMA	72	Therapeutic area	*Osteoporosis*
		Active substance	*abaloparatide*

Table 2.3: Experiment dataset: overview of textual data

disease. Therefore, efficient linking requires processing on a semantic, rather than lexical, level; namely, synonymous terms need to be normalized into one underlying medical concept. There are a few useful resources that offer a mapping of medical terms to concepts, most notably the MeSH thesaurus[11] and the UMLS metathesaurus[12]. For example, the UMLS metathesaurus maps all the disease terms mentioned above to one *concept unique identifier* (CUI) 'C0002895'. In the current experiment, the entities detected by BERN are mapped to their respective UMLS CUI's using the QuickUMLS tool (Soldaini and Goharian 2016).

After these two steps, the records in the experiment dataset are represented in terms of two sparse vectors: the set of their drug CUI's and the set of their disease CUI's. Cosine similarity is used to compare between the vectors: the drug vector of a record is (pairwise) compared to the drug vectors of all the other records in the dataset, the disease vector of a record is (pairwise) compared to the disease vectors of all the other records in the dataset. Cosine similarity is defined as the inner product of two vectors, while the length of both vectors is normalized to 1. This similarity metric is useful for our case, since only the non-zero dimensions of the vectors are considered, which is very suitable for dealing with sparse vectors (since the vectors are the size of the total number of unique CUI's, most of the elements in each vector are zero). If both the drug CUI's vector and the disease CUI's vector of two records are similar to each other (above a certain similarity threshold), these two records are grouped together into one group. At the end of the grouping process, all groups that are subsets of other groups were removed since for the current purposes we are interested in the largest possible group of related records. I experimented with three settings:

- *Metric*: cosine similarity; *Similarity threshold*: 0.8 (*name*: cos08)
- *Metric*: cosine similarity; *Similarity threshold*: 0.7 (*name*: cos07)
- *Metric*: cosine similarity; *Similarity threshold*: 0.6 (*name*: cos06)

[11] https://www.nlm.nih.gov/mesh/meshhome.html
[12] https://uts.nlm.nih.gov/uts/umls/home

	cos08	cos07	cos06
No. of output groups	28,234	74,452	100,048
No. of output groups containing an EMA authorization	28	183	200
Median no. of groups per EMA authorization	1	4	4
Max no. of groups per EMA authorization	5	34	35

Table 2.4: Overview of grouping results per experimental setup

2.7.3 Results

Overview of quantitative results

As mentioned above, the grouping method was applied to a big dataset containing 118,689 CTgov records, 419,894 PubMed records, and 72 EMA records. This resulted in about 28,000 groups with the 0.8 similarity threshold, about 74,000 groups with the 0.7 similarity threshold, and about 100,000 groups with the 0.6 similarity threshold; see Table 2.4. From these output groups, I selected the ones which contain an EMA record; there are 28 such groups for the cos08 setting, 183 groups for cos07, and 200 groups for cos06. One EMA record can be linked to multiple groups. With the high similarity threshold, this rarely happens; the median number of groups per EMA record is 1 and the maximum number is 5 (see Table 2.4). With the lower similarity thresholds (0.7 and 0.6), the median number of groups per EMA record is 4, with a maximum as high as 35 groups per one EMA record.

To illustrate the concept of multiple groups for one EMA record, consider the example of authorization # 1575: *Sovrima*. In the cos08 setting, this EMA record was linked to one group, in the cos07 setting to two groups, and in the cos06 setting to three groups. The results obtained with a 0.6 threshold are shown in Figure 2.9. As evident from the figure, the total 14 records linked to the EMA record are grouped into three groups, which partially overlap with each other. For our purposes, the differentiation into groups is not informative; therefore, from now on we focus on the total set of records that is associated with an authorization when using a certain similarity threshold. In other words, for this example we lump together the records in the three groups into one super-group.

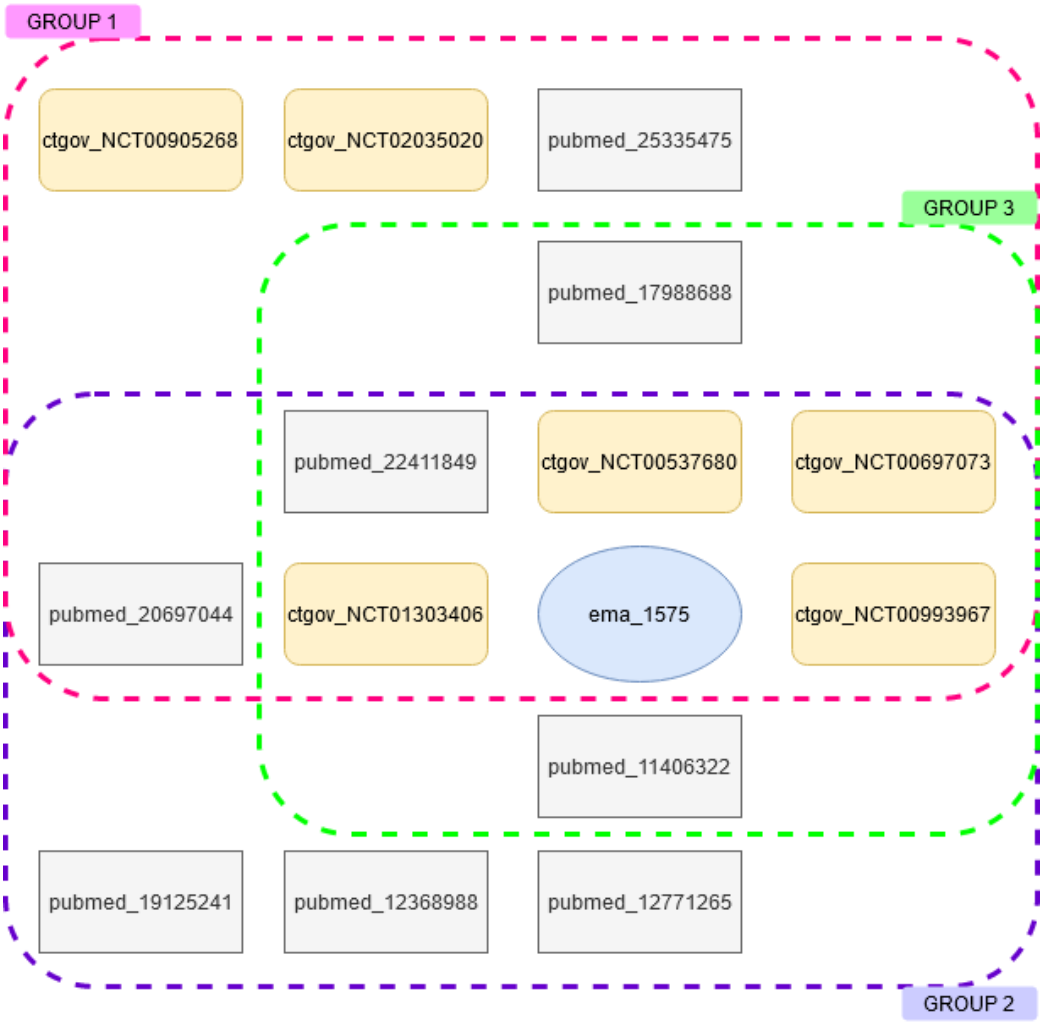

Figure 2.9: The groups linked to EMA record 1575 (*Sovrima*) (cos06)

	CTgov	PubMed
No. EMA authorizations which were grouped with CTgov/PubMed records	cos08: 19 (26%) cos07: 27 (37%) cos06: 27 (37%)	cos08: 13 (18%) cos07: 24 (33%) cos06: 25 (35%)
Total no. of unique CTgov/PubMed records grouped with EMA authorizations	cos08: 107 cos07: 1,920 cos06: 2,045	cos08: 20 cos07: 2,457 cos06: 2,770
Median no. of grouped records per EMA authorization	cos08: 6 cos07: 31 cos06: 34	cos08: 1 cos07: 22.5 cos06: 23
Max no. of grouped records per EMA authorization	cos08: 15 cos07: 703 cos06: 726	cos08: 5 cos07: 1,241 cos06: 1,382

Table 2.5: Summary of quantitative results for groups containing EMA authorizations: *CUI_grouping* method

Table 2.5 focuses on the super-groups that contain an EMA authorization and shows the quantitative results in terms of the CTgov and PubMed records associated with each EMA record. For all three similarity thresholds, the number of EMA authorizations that were grouped into at least one group is relatively low: 18-37% of the 72 EMA records in the dataset (a possible reason is discussed later in this section). On the other hand, the number of CTgov and PubMed records associated with those EMA records is very high, at least for the cos07 and cos06 setting, which associate more than 2,000 CTgov records and more that 2,000 PubMed records to about 25 EMA authorizations (see Table 2.5). At this point, we do not know whether these big numbers are due to the method being very "noisy", i.e. linking a lot of irrelevant records; this is explored via examples in Section 2.8.2.

Grouping Example

To better understand the mechanism of the grouping method, we can examine one example: the records that are linked to the EMA authorization of *Sovrima* (active substance: *idebenone*), a medicine intended to treat *Friedreich Ataxia*, which was refused marketing authorization by EMA in 2008. The results obtained with each experimental setting are as follows:

- *cos08*: 4 CTgov records and no PubMed records
- *cos07*: 6 CTgov records and 7 PubMed records
- *cos06*: 6 CTgov records and 8 PubMed records

With a high similarity threshold (0.8), only a few records that are extremely similar to the EMA authorization are linked to it; as the threshold is lowered, more and more records are considered "similar enough" to be linked to the authorization.

To make these various similarity levels more explicit, all the records linked to the *Sovrima* authorization are shown in Table 2.6. Each record appears with the biomedical entities detected in it by the bio-NER tool BERN and with the CUI's that the mapping tool QuickUMLS assigned to these entities. In the upper part of the table, we see the most similar records, namely the ones that were linked to the authorization even with a 0.8 similarity threshold. These records have exactly the same disease CUI (C0016719 *Friedreich Ataxia*) and the same drug CUI (C0123163 *idebenone*) as the EMA record. In the second part of the table, we see the records that were linked to the authorization in cos07 and cos06 settings, but not in the cos08 setting. These records are characterized by having one or both of the disease CUI's C0016719 (*Friedreich Ataxia*) / C0004134 (*ataxia*), and one or both of the drug CUI's C0123163 (*idebenone*) / C0387678 (*frataxin*). Some of these records serve as the connecting link, because they have both CUI's (e.g. 20697044 which has both C0016719 and C0004134, or 22411849 which has both C0123163 and C0387678) and therefore are similar to records that have either one of the two CUI's. In the lower part of the table, we see the record that was linked to the authorization only with a 0.6 similarity threshold. This record has 3 additional disease CUI's besides C0016719 and C0004134, and it has a drug CUI that none of the other records have (C0302583 *iron*) besides C0123163.

Number of CUI's

In Table 2.6, some records have only one disease CUI and one drug CUI, while others have multiple CUI's. This affects the grouping; a record is less similar to another record with multiple CUI's than to another record with fewer CUI's, even if the number of common CUI's is the same. For example, the cosine similarity between a record that has one CUI and a record that has 5 CUI's (one CUI in common) is 0.45, while the cosine similarity between a

Record ID	Disease entity	CUI	Drug entity	CUI
Grouped with all thresholds				
EMA_1575	Friedreich ataxia	C0016719	idebenone	C0123163
NCT00697073	Friedreich's ataxia	C0016719	idebenone	C0123163
NCT00993967	Friedreich's ataxia	C0016719	idebenone	C0123163
NCT00537680	Friedreich's ataxia	C0016719	idebenone	C0123163
NCT01303406	Friedreich's ataxia	C0016719	idebenone	C0123163
Grouped with 0.7 and 0.6 thresholds				
22411849	Friedreich ataxia	C0016719	frataxin, idebenone	C0123163, C0387678
20697044	Friedreich ataxia, neurological, philadelphia, ataxia	C0016719, C0004134	idebenone	C0123163
17988688	Friedreich ataxia, cardiomyopathy	C0016719, C0878544	idebenone	C0123163
25335475	Friedreich ataxia	C0016719	IFN, frataxin	C0387678
NCT02035020	Friedreich's ataxia, FRDA	C0016719	frataxin	C0387678
NCT00905268	Friedreich's ataxia, FRDA, neurological impairment and function	C0016719, C0521654	idebenone	C0123163
12368988	Friedreich s ataxia, ataxia	C0004134	idebenone	C0123163
19125241	Friedreich s ataxia	C0004134	idebenone, ubiquinone	C0123163, C0041536
12771265	Friedreich s ataxia, Friedreich ataxia, ataxia, cardiac hypertrophy, hypertrophy	C0016719, C0004134, C0020564, C1383860	idebenone, protoporphyrin ix	C0123163, C0033733
Grouped only with 0.6 threshold				
11406322	Friedreich ataxia, FA, degenerative ataxia, respiratory deficiency, FA-associated cardiomyopathy	C0016719, C0004134, C0878544, C0011164, C0020621	idebenone, iron, (31)P	C0123163, C0302583

Table 2.6: Biomedical entities and CUI's found in records related to *Sovrima*

	Disease CUI's				Drug CUI's			
	min	**max**	**mean (sd)**	**med**	**min**	**max**	**mean (sd)**	**med**
CTgov	0	29	2.2 (1.9)	2	0	23	1.5 (1.4)	1
PubMed	0	27	2.9 (2.4)	2	0	66	1.5 (1.8)	1
EMA	1	4	1.2 (0.5)	1	0	3	1.3 (0.6)	1

Table 2.7: Number of CUI's per record

record that has one CUI and a record that has 3 CUI's (one CUI in common) is 0.58.

This is especially crucial in view of the fact that EMA records tend to have only one disease CUI and one drug CUI, as shown in Table 2.7. The median number of both disease CUI's and drug CUI's in EMA records is 1, with a maximum of 3-4 CUI's per record. The other record types, however, tend to have multiple disease CUI's; the median number of disease CUI's in CTgov and PubMed records is 2, with outliers as high as 27-29 disease CUI's per record. This is probably the reason behind the low percentage of EMA authorizations to which records are linked (recall Table 2.5); since PubMed and CTgov records tend to have multiple CUI's, they are not very similar to the one-CUI EMA records.

The reason CTgov and PubMed records have multiple CUI's is because of the length and the nature of their processed text (recall Table 2.3 above). Especially the "brief summary" part of CTgov records and the "abstract" part of PubMed records contain a lot more information than just the investigated drug and the disease it is intended to treat. For example, a PubMed abstract might contain mentions of side effects, which also have disease CUI's, or comparator drugs, which have drug CUI's. One possible solution would be to use less text, for example only the titles; however, it should be taken into account that titles might not contain the name of the disease (e.g. the titles in Table 2.3 do not contain the term *osteoporosis*).

2.8 Evaluation of the Methods

2.8.1 Comparison of Quantitative Results

Table 2.8 provides an overview of the quantitative results obtained with the three methods (reported separately in Section 2.5.3, Section 2.6.3 and Sec-

tion 2.7.3 above). The first row shows how many EMA authorizations (out of the total 72) are linked to at least one CTgov / PubMed record by each method. In this regard, *EPAR_CTgov* is the best performing method, finding CTgov records for 90% of the authorizations and PubMed records for 88% of the authorizations. This means that if one has access to the protocol numbers related to an authorization, they can be effectively used to query the CTgov database and retrieve both clinical trial information and cross-references to related publications on PubMed. However, note that this performance is dependent on (parts of) the procedure being done manually.[13] In addition, while the method manages to find *at least one* related record for the vast majority of the authorizations, it is unclear whether there are also a lot of related records that it does not detect; this suspicion arises when comparing the total number of records found by each method, as discussed below. The other two methods, which are fully automated, are significantly less successful in terms of the percentage of authorizations they manage to link records to. The *PubMed_API* method finds PubMed records for 64% of the authorizations and CTgov records for 57%. The *CUI_grouping* method links records to between 18% and 37% of the authorizations, depending on the similarity threshold applied.

Even though they find records for fewer authorizations, the *PubMed_API* and *CUI_grouping* methods generally link more records to each authorization than *EPAR_CTgov*. This is reflected in the median number of records per authorization (third row in Table 2.8) and in the total numbers of unique records found by each method (second row in Table 2.8). While *EPAR_CTgov* finds about 450 CTgov records and 450 PubMed records in total, *PubMed_API* finds over 1,000 PubMed records and *CUI_grouping* finds over 2,000 PubMed records and over 2,000 CTgov records. The first exception to this trend is the low number of CTgov records found by *PubMed_API*: 256 only (median: 3 records per authorization); as discussed in Section 2.6.3, this has to do with the cross-references from PubMed to CTgov being very partial. The second exception is the very low numbers of records found by *CUI_grouping* with a 0.8 similarity threshold: 107 CTgov records (median: 6) and 20 PubMed records (median: 1). These results suggest that a 0.8 similarity threshold is

[13] As mentioned in Section 2.5.2, there are two obstacles to automating this procedure: (a) retrieving protocol numbers from the EPARs is not easily automated in their current format, and (b) not all cross-references to PubMed are found in the structured data of the CTgov records, meaning that automating will result in a certain drop in the number of the retrieved PubMed records.

	CTgov			PubMed		
	EPAR_CTgov	PubMed_API	CUI_grouping	EPAR_CTgov	PubMed_API	CUI_grouping
No. EMA authorizations for which records were found	65 (90%)	41 (57%)	cos08: 19 (26%) cos07: 27 (37%) cos06: 27 (37%)	63 (88%)	46 (64%)	cos08: 13 (18%) cos07: 24 (33%) cos06: 25 (35%)
Total no. of unique found records	451	256	cos08: 107 cos07: 1,920 cos06: 2,045	453	1,113	cos08: 20 cos07: 2,457 cos06: 2,770
Median no. of found records per EMA authorization	5	3	cos08: 6 cos07: 31 cos06: 34	4	12	cos08: 1 cos07: 22.5 cos06: 23
Max no. of found records per EMA authorization	42	32	cos08: 15 cos07: 703 cos06: 726	59	176	cos08: 5 cos07: 1,241 cos06: 1,382

Table 2.8: Summary of quantitative results: all methods

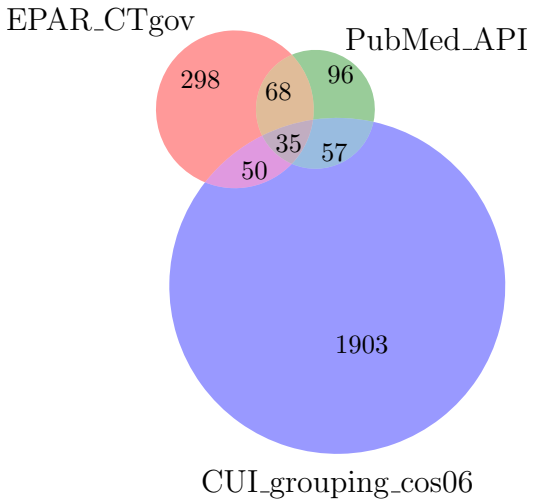

Figure 2.10: Overlap between the CTgov records identified by the three methods

too restrictive for the task at hand; possible reasons for this are discussed later in the section.

Figure 2.10 and Figure 2.11 show the relations between the records found by each method (for simplicity, only the 0.6 threshold is shown for *CUI_grouping*). As evident from the diagrams, the overlap between the different methods is relatively small, and the majority of the records detected by each method are not found by the other two. One important caveat to the numbers in the diagrams is that out of the records detected by *EPAR_CTgov* and *PubMed_API*, there are 62 CTgov and 420 PubMed records that are not part of the dataset used for the *CUI_grouping* experiment, meaning that they could not be linked by the *CUI_grouping* method. The reason they are missing from the dataset has to do with the selection criteria and the data dumps used for creating it; for example, as mentioned in Section 2.7.1, the PubMed data dump used for creating the *CUI_grouping* dataset is from 2018, therefore papers published after 2018 are not in the dataset at all, and the data for 2018 itself is partial. This caveat is especially relevant for the PubMed diagram (Figure 2.11), since 30% of the papers found by *EPAR_CTgov* and *PubMed_API* are not in the *CUI_grouping* dataset; if they were, the overlapping areas between the methods might have been bigger.

To sum up, comparison of the overall results obtained with the three

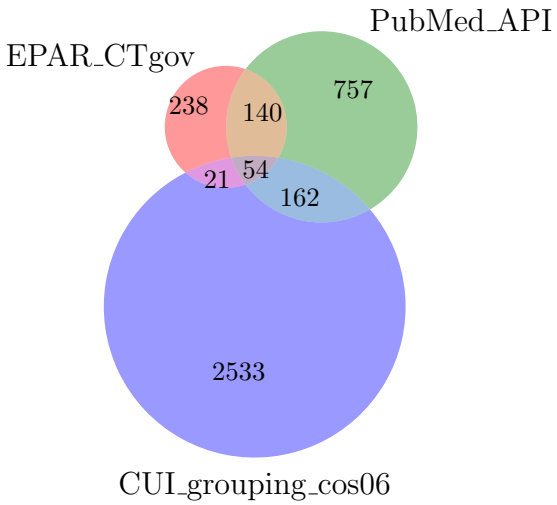

Figure 2.11: Overlap between the PubMed records identified by the three methods

methods suggests that each of them has its own strengths and its own limitations that lead to quantitatively and qualitatively different outcomes. To better understand these strengths and limitations, Section 2.8.2 below analyzes a few examples of EMA authorizations and the records that each method links to them.

2.8.2 In-depth Analysis of Examples

Since we do not have a gold standard of correct answers to evaluate against, it is necessary to examine individual examples in order to get an impression about how well each method performs. In this section, I analyze four EMA authorizations by manually checking the records that each method linked to it (when there are many records, a random sample of them is checked). The purpose of the analysis is to determine:

- Are the records linked by each method correct? Linked records are considered "correct" if they are directly related to the authorized drug, i.e. they discuss clinical trials that were conducted in order to investigate the safety and/or efficacy of a specific active substance as a treatment for a specific disease; this includes both trials that are mentioned in

the EPAR and trials that aren't, e.g. studies that were conducted after the authorization.

- Are the correct records linked by all methods? If not - why does a certain method fail to link a certain record to the authorization?
- If there are incorrectly linked records - why does a certain method link a certain incorrect record to the authorization?

The four examples were selected semi-randomly: the authorizations were randomly sampled and scanned until four examples were obtained that are different enough from each other in the phenomena they showcase:

- Refused EMA authorization #1094 (*Exondys*):
  - All three methods link records to this authorization. The results are balanced, i.e. the numbers of records linked by each method are comparable.
- Approved EMA authorization #57 (*Extavia*):
  - Only the *PubMed_API* method manages to link records to this authorization.
- Refused EMA authorization #1039 (*Eladynos*):
  - All three methods link records to this authorization. The results are unbalanced: the *CUI_grouping* method links significantly more records in comparison with the other methods.
- Approved EMA authorization #13 (*Mysimba*):
  - *EPAR_CTgov* links both PubMed and CTgov records to this authorization, *CUI_grouping* manages to link only CTgov records, and *PubMed_API* does not manage to link any records.

Below, these four examples are analyzed in detail; through this analysis the strengths and limitations of each method come forth.

EMA record 1094 (*Exondys*)

Exondys (active substance: *eteplirsen*), which is intended as a treatment for *Duchenne Muscular Dystrophy*, was refused marketing authorization by EMA in 2018. Its EPAR mentions 7 studies that were considered by the authorization committee. All three methods link records to this authorization:

- *EPAR_CTgov*: 7 CTgov records and 4 PubMed records;
- *PubMed_API*: 3 CTgov records and 7 PubMed records;
- *CUI_grouping*:
  - *cos08*: 6 CTgov records and no PubMed records;
  - *cos07*, *cos06*: 7 CTgov records and 2 PubMed records.

The relations between the records linked by the different methods are visualized in Figure 2.12 and Figure 2.13 (for simplicity, only the 0.6 threshold is included for the *CUI_grouping* method).

- CTgov (Figure 2.12):
  - There are 2 records (NCT01396239 and NCT01540409) that are detected by all three methods.
  - There are 3 records (NCT02255552, NCT02420379 and NCT02286947) that are identified by *EPAR_CTgov* and *CUI_grouping* but not by *PubMed_API*.
  - There is one record (NCT00844597) that is detected by *EPAR_CTgov* and *PubMed_API* but not by *CUI_grouping*.
  - There is one record (NCT00159250) that is only found by the *EPAR_CTgov* method (and neither of the other methods).
  - There are 2 records (NCT03218995 and NCT03992430) that are only found by the *CUI_grouping* method.
- PubMed (Figure 2.13):
  - There are 2 records (23907995 and 26573217) that are detected by all three methods.
  - There is one record (21784508) that is found by *EPAR_CTgov* and *PubMed_API* but not by *CUI_grouping*.

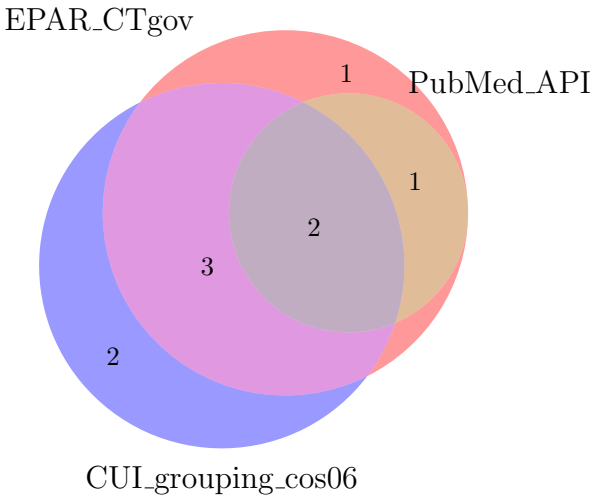

Figure 2.12: Overlap between the CTgov records associated by the three methods with *Exondys*

- There is one record (19713152) that is only found by the *EPAR_CTgov* method.
- There are 4 records (22676208, 29278896, 29752304 and 31261494) that are only found by the *PubMed_API* method. 3 out of them (29278896, 31261494 and 29752304) could not in principle be detected with the *CUI_grouping* method, since they are not part of the dataset used for the experiment (they were published in 2018 or later; see Section 2.7.1).

In total, the three methods link 9 CTgov records and 8 PubMed records to the authorization of *Exondys*. All the identified records are correct, i.e. they are clinical trials that investigate *eteplirsen* as a possible treatment for *Duchenne Muscular Dystrophy*, and publications reporting the results of these trials. 7 out of the 9 identified CTgov records are mentioned in the EPAR. The other two (NCT03218995 and NCT03992430) are currently ongoing clinical trials conducted by the same sponsor (Sarepta Therapeutics, previously known as AVI BioPharma) that further examine the safety and efficacy of different doses of *eteplirsen* as a treatment for *Duchenne Muscular Dystrophy*. All the identified PubMed records are related to trials mentioned in the EPAR.

In this example, the *EPAR_CTgov* and the *PubMed_API* methods are

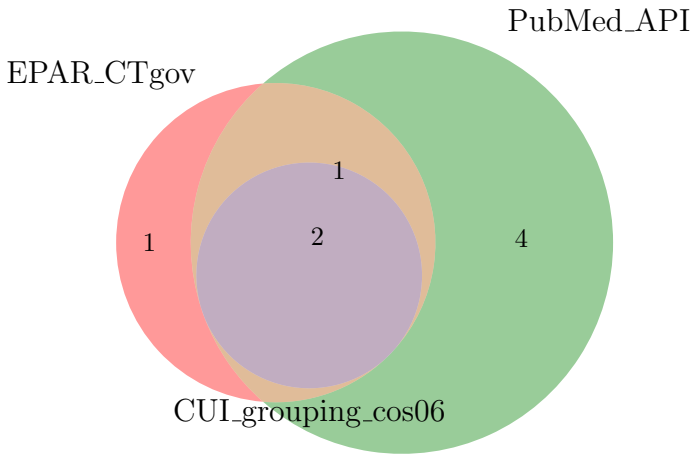

Figure 2.13: Overlap between the PubMed records associated by the three methods with *Exondys*

mirror images of each other; the former is successful in finding CTgov records (N=7) and less successful in finding PubMed records (N=4) and the latter shows the opposite performance (7 PubMed records, 3 CTgov records). Recall that *EPAR_CTgov* queries the CTgov database directly and relies on cross-references to retrieve PubMed records, while *PubMed_API* queries the PubMed database directly and relies on cross-references to retrieve CTgov records. As discussed in Section 2.5.2 and Section 2.6.3, the cross-references between the two databases are partial and not symmetrical; this is demonstrated again with the current example.

Despite their good performance in identifying the type of records that does not depend on cross-reference, *EPAR_CTgov* and *PubMed_API* are not perfect in detecting CTgov and PubMed records, respectively. There are 2 CTgov records that *EPAR_CTgov* fails to link because they are not mentioned in the EPAR (see above); this demonstrates an inherent limitation of this method. Similarly, there is one PubMed record that *PubMed_API* fails to detect. The reason is that this record (19713152) does not list *eteplirsen* as one of the substances; rather it has the terms *oligonucleotides* (a broader concept that includes *eteplirsen*) and *dystrophin* (this is not a drug, but the protein that muscular dystrophy patients lack). Since the *PubMed_API* query looks for exact matches of the active substance and disease names from

the EMA data, it cannot find records that do not list the exact same entity names.

The *CUI_grouping* method performs quite well in this example, especially in terms of CTgov records; it manages to link 7 out of the total 9 CTgov records, including 2 that are not retrieved by the other two methods. Out of the total 8 PubMed records, 5 records are part of the *CUI_grouping* dataset, and 2 of those are linked by the method. To understand why certain records are linked and others are not, their CUI's need to be examined. Table 2.9 and Table 2.10 show the linked and non-linked records, respectively; the non-linked records include the ones that are identified by the other methods and are part of the *CUI_grouping* dataset, i.e. the ones that could in principle be linked by the method. Each record appears with the biomedical entities detected in it by the bio-NER tool BERN and with the CUI's that the mapping tool QuickUMLS assigned to these entities.

The linked records in Table 2.9 are characterized by having the disease CUI C0013264 (*Muscular Dystrophy, Duchenne*) with occasionally one additional disease CUI, and the drug CUI C4283710 (*eteplirsen*) with occasionally one additional drug CUI. All the non-linked records in Table 2.10, on the other hand, do not have the correct drug CUI. Looking at the processed text of these records reveals that it does not use the term *eteplirsen* for the drug, but instead uses either the broader concepts *morpholino* and *antisense oligonucleotides*, or the narrower concept *AVI-4658*. These terms are, at least in some of the examples, identified as drug entities by BERN (it is unclear why this is not the case for NCT00159250 and 22676208); QuickUMLS maps *morpholino* to the CUI C0026560 (*morpholine*), which is a related but broader concept than *morpholino*, and it does not manage to map *AVI-4658*, even though this term has a CUI (hierarchically nested under *eteplirsen*).

In other words, the failure to link the records in Table 2.10 to the authorization is mainly due to the text itself, which uses broader / narrower concepts instead of the active substance *eteplirsen*. This issue could be solved (at least partially) by utilizing the hierarchical structure of the UMLS metathesaurus and including broader and/or narrower concepts; however, the usefulness of this approach should be tested empirically as it is expected to introduce a lot of noise to the results. The tools BERN and QuickUMLS also introduce some issues in this example, specifically - BERN fails to detect drug entities in 2 of the records (although they are present in the processed text) and QuickUMLS fails to map the concept *AVI-4658* (although it has a CUI in the metathesaurus).

Record ID	Disease entity (BERN)	CUI (QuickUMLS)	Drug entity (BERN)	CUI (QuickUMLS)
EMA_1094	'Muscular Dystrophy, Duchenne'	C0013264	eteplirsen	C4283710
NCT01396239	duchenne muscular dystrophy, DMD	C0013264	AVI-4658, eteplirsen	C4283710
NCT01540409	duchenne muscular dystrophy, DMD	C0013264, C0026850	eteplirsen	C4283710
NCT02255552	duchenne muscular dystrophy, DMD	C0013264	eteplirsen	C4283710
NCT02420379	duchenne muscular dystrophy, DMD	C0013264	eteplirsen	C4283710
NCT02286947	duchenne muscular dystrophy, advanced stage duchenne muscular dystrophy, DMD	C0013264	eteplirsen	C4283710
23907995	duchenne muscular dystrophy	C0013264	eteplirsen, phosphorodiamidate, nitric oxide	C4283710, C0028128
26573217	duchenne muscular dystrophy, loss of ambulation	C0013264, C2678024	eteplirsen	C4283710
NCT03992430	duchenne muscular dystrophy, DMD	C0013264	eteplirsen	C4283710
NCT03218995	duchenne muscular dystrophy, DMD	C0013264	eteplirsen	C4283710

Table 2.9: Biomedical entities and CUI's found in records related to *Exondys* (linked by cos06)

Record ID	Disease entity (BERN)	CUI (QuickUMLS)	Drug entity (BERN)	CUI (QuickUMLS)
NCT00159250	duchenne muscular dystrophy, DMD, muscle degenerative disorder	C0013264, C1285162	n/a	n/a
NCT00844597	duchenne muscular dystrophy, DMD	C0013264	AVI-4658	n/a
19713152	duchenne muscular dystrophy, x-linked disease duchenne muscular dystrophy	C0013264, C2748900, C0012634	AVI-4658, morpholino	C0026560
21784508	duchenne muscular dystrophy	C0013264, C0026850	phosphorodiamidate morpholino, AVI-4658 phospho-rodiamidate, AVI-4658, nitric oxide	C0028128, C0026560
22676208	paediatric disorders, duchenne muscular dystrophy, chronic paediatric disorders, progressive disorders	C0013264, C3839460, C0012634	n/a	n/a

Table 2.10: Biomedical entities and CUI's found in records related to *Exondys* (not linked by cos06)

All in all, this example shows good performance by all three methods; all the records linked to this authorization are correct. Moreover, the methods complement each other; each of them manages to find correct records that the other two do not detect.

EMA record 57 (*Extavia / Betaferon*)

Extavia (active substance: *interferon beta-1b*), a treatment for *Multiple Sclerosis*, was approved for marketing in the EU in 2008 based on the clinical data of another drug - *Betaferon* - through a procedure called 'informed consent'; *Betaferon* was approved by EMA in 1995. The information pages of both drugs mention 4 clinical trials considered for the authorization (see the section *'What benefits of Extavia have been shown in studies?'* here). However, the EPARs of both drugs do not mention any protocol numbers (see here and here); without protocol numbers, no CTgov records could be found by the *EPAR_CTgov* method, and consequently also no PubMed records.

With the *PubMed_API* method, 176 PubMed records and 15 CTgov records are linked to this authorization. A random sample of 5 CTgov records and 5 PubMed records was manually examined to check whether they are indeed related to *Extavia / Betaferon*:

- CTgov record NCT00099502 is a Phase III study of two different doses of *interferon beta-1b*; based on the number of patients, it is not one of the studies considered for the authorization, but it is **correct** under our definition.
- CTgov record NCT01766063 and PubMed record 30143019 are related to a post-authorization study conducted in 2012 that investigates sleep quality in MS patients treated with *interferon beta-1b*. → **correct**
- PubMed record 10636148 reports the results of a trial comparing the effect of a placebo treatment and two different doses of *interferon beta-1b* on lesion activity seen in MRI. → **correct**
- PubMed record 10809912 reports the results of a post-authorization study comparing between *interferon beta-1a* and *interferon beta-1b*. → **correct**
- CTgov record NCT01144052 is a post-authorization study comparing treatment with *interferon beta-1b* to treatment with natalizumab. →

correct

- PubMed record 24515732 is a post-authorization study that compares long term thyroid dysfunction in MS patients treated with *interferon beta-1b* or with glatiramer acetate. → **correct**

- PubMed record 28877664 and CTgov record NCT02121444 report the results of a study focused on an auto-injector used for the injection of *interferon beta-1b*. These records are considered **incorrect** by our definition, since they are not focused on investigating the effect of the drug itself but rather on the injector.

- CTgov record NCT00428584 is a study of a drug called Rebif (*interferon beta-1a*), where *interferon beta-1b* serves as an active comparator. This is considered **incorrect** by our definition, since it investigates the effects of Rebif and not of *interferon beta-1b*.

None of the sampled records is related to the studies considered for the authorization in 1995; however, most of the sampled records are correct under our definition, since they investigate the effects of treatment with *interferon beta-1b* in MS patients. Only 3 out of the sampled 10 records do not fit under our definition of correctly linked records.

The *CUI_grouping* method does not link any records to this authorization. 25 out of the 176 PubMed records detected by *PubMed_API* are not in the *CUI_grouping* dataset, but the other 151 PubMed records and 15 CTgov records are; to understand why those were not linked to the authorization by the *CUI_grouping* method, their CUI's need to be examined. Table 2.11 shows the EMA authorization and 5 related records that are identified by *PubMed_API* and discussed above. The correct CUI's are C0026769 (*Multiple Sclerosis*) and C0244713 (*interferon beta-1b*) but only 3 out of the 5 records have the correct disease CUI and none of them have the correct drug CUI. The main issue with the disease CUI is that QuickUMLS does not map *MS* to the *Multiple Sclerosis* CUI, even though it is listed as one of the synonyms in the metathesaurus; this is a problematic behavior of the tool since normalizing different synonymous terms to one CUI is the main goal of performing this step. Regarding the drug term, the problem can be traced back to the BERN tool, which does not detect the term *interferon beta-1b* in any of the records, even though it appears in 3 out of the 5.[14] In addition, for

[14]Recall that the text of the EMA record was not processed with BERN since the named

Record ID	Disease entity (BERN)	CUI (QuickUMLS)	Drug entity (BERN)	CUI (QuickUMLS)
EMA_57	Multiple Sclerosis	C0026769	interferon beta-1b	C0244713
NCT00099502	multiple sclerosis, MS	C0026769	betaseron, betaferon, copaxone, glatiramer acetate	C0284968, C0528175, C0289884, C0592527
NCT01766063	MS, fatigue, sclerosis	C0015672, C0036429	betaferon	C0592527
NCT01144052	multiple sclerosis, sclerosis, MS, neurological disorder, relapsing-remitting multiple sclerosis	C0026769, C0027765, C0751967, C0036429	n/a	n/a
10636148	MS, UBC	n/a	gadolinium	C0016911
10809912	multiple sclerosis, relapsing-remitting multiple sclerosis, MS	C0026769, C0751967	n/a	n/a

Table 2.11: Biomedical entities and CUI's found in records related to *Extavia* (not linked by *CUI_grouping*)

two records we see again the issue discussed in the previous example, namely that a narrower concept - *betaferon*, CUI C0592527 - is used instead of / in addition to the active substance *interferon beta-1b*.

To sum up, this example shows that the *PubMed_API* method succeeds where other methods fail by querying the PubMed dataset directly. It is not entirely clear how precise it is in finding relevant records; based on the sample, the precision is about 70%. The erroneously linked records are linked because the method detects all PubMed records that have the active substance in their 'substance name' field; this includes papers where the active substance is mentioned as e.g. a comparator, and not only as the investigated drug. The example also shows the shortcomings of the other two methods: *EPAR_CTgov* is fully reliant on the presence of protocol numbers in the EPAR, which apparently is not always the case; *CUI_grouping* is dependent on a pipeline of text processing tools, when the first step (i.e. bio-NER) fails, all relevant records are irreparably affected.

EMA record 1039 (*Eladynos*)

Eladynos (active substance: *abaloparatide*) is intended as a treatment for *osteoporosis*, and was refused marketing authorization by EMA in 2018. The EPAR mentions 9 studies considered during the authorization. All three methods link records to this authorization:

- *EPAR_CTgov*: 4 CTgov records and 6 PubMed records;
- *PubMed_API*: 3 CTgov records and 15 PubMed records;
- *CUI_grouping*:
  - *cos08*: 4 CTgov records and no PubMed records;
  - *cos07*: 26 CTgov records and 46 PubMed records;
  - *cos06*: 28 CTgov records and 48 PubMed records.

The relations between the records linked by the different methods are visualized in Figure 2.14 and Figure 2.15 (for simplicity, only the 0.6 threshold is included for the *CUI_grouping* method).

entities are part of the structured data; this is why the term is present and correctly mapped in the first row of the table.

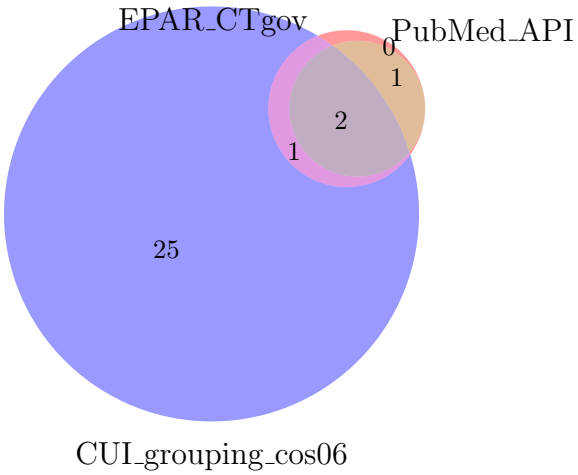

Figure 2.14: Overlap between the CTgov records associated by the three methods with *Eladynos*

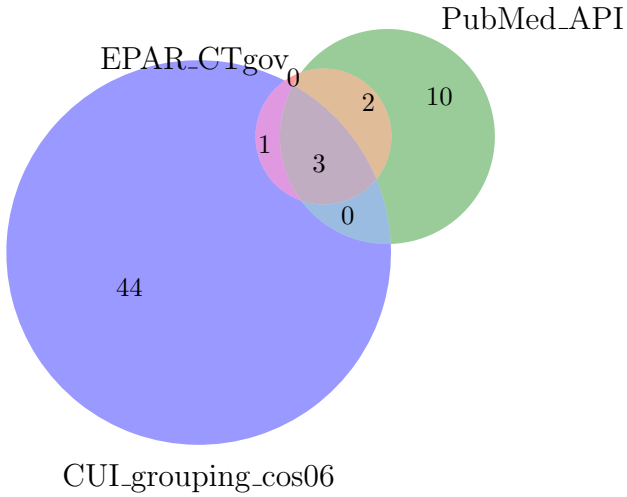

Figure 2.15: Overlap between the PubMed records associated by the three methods with *Eladynos*

- CTgov (Figure 2.14):
  - There are 2 records (NCT01343004 and NCT01657162) that are detected by all three methods.
  - There is one record (NCT00542425) that is found by *EPAR_CTgov* and *PubMed_API* but not by *CUI_grouping*.
  - There is one record (NCT01674621) that is found by *EPAR_CTgov* and *CUI_grouping* but not by *PubMed_API*.
  - There are 25 records that are only linked by the *CUI_grouping* and not the other methods.
- PubMed (Figure 2.15):
  - There are 3 records (25393645, 27612281 and 28160873) that are detected by all three methods.
  - There are 2 records (27533157 and 29800372) that are found by *EPAR_CTgov* and *PubMed_API* but not by *CUI_grouping*. One of these records (29800372) is not in the *CUI_grouping* dataset, since it was published in 2018 or later.
  - There is one record (27826127) that is found by *EPAR_CTgov* and *CUI_grouping* but not by *PubMed_API*.
  - There are 10 records found only by *PubMed_API* (28474780, 29167971, 29251010, 29285549, 29462094, 29951742, 30635696, 30719589, 30899994 and 31418585). 8 of these records are not in the *CUI_grouping* dataset, since they were published in 2018 or later.
  - There are 44 records found only by *CUI_grouping*.

The 4 CTgov records that are mentioned by their number above are correct: they are all trials that are mentioned in the EPAR. Out of the 16 PubMed records mentioned above by their number, 15 are correct (they are related to one of the trials mentioned in the EPAR); only record 29251010 (detected only by *PubMed_API*) is incorrect since it discusses the effect of a sequential treatment combining *abaloparatide* and *alendronate*, thus it is not directly related to the authorization of *abaloparatide* itself.

Out of the 25 CTgov records and 44 PubMed records that are linked only by *CUI_grouping*, 4 CTgov and 4 PubMed records were sampled and examined to see if they are correct. Two of the sampled CTgov records

(NCT03512262, NCT03710889) turned out to be correctly linked; these are recent studies (started after the authorization in 2018) conducted by the same sponsor (Radius), one of which examines the effects of *abaloparatide* in postmenopausal women with *osteoporosis* and another one which tests its effects in men with *osteoporosis*. The other 2 sampled CTgov records, as well as the 4 sampled PubMed records, are not related to *Eladynos*:

- NCT01760798 is a study comparing the effects of weekly vs. daily *teriparatide* treatment on postmenopausal women with *osteoporosis*.
- NCT00936260 is a study determining the duration of treatment with *alendronate* in postmenopausal women with *osteoporosis*.
- 24345886 is a paper that reports the results of a study evaluating the effects of *teriparatide* on postmenopausal women with *osteoporosis*.
- 10793867 is a paper that reports the results of a study evaluating the effects of *alendronate* on men with *osteoporosis* and postmenopausal women with *osteoporosis*.
- 15104559 is a paper that reports the results of a study evaluating the effects of recombinant human growth hormone, alone and combined with *alendronate*, in GH-deficient adults (half of them with *osteoporosis*).
- 24205135 is a paper that reports the results of a study comparing the tolerability and efficacy of generic versus brand *alendronate* in postmenopausal women with *osteoporosis*.

These examples share the same disease with *Eladynos*; however, they are not related to the drug *abaloparatide*, but rather to two different drugs: *teriparatide* and *alendronate*. To understand why these records are grouped with *Eladynos* at lower similarity thresholds, consider Table 2.12, which shows correct records that are identified by the other methods and are also linked by *CUI_grouping*. The correctly linked records are characterized by having the disease CUI C0029456 (*osteoporosis*), with optionally up to 3 additional disease CUI's, and the drug CUI C4042342 (*abaloparatide*), by itself or in combination with either C0070093 (*teriparatide*) or C0102118 (*alendronate*). These two drugs are mentioned in papers about *Eladynos*, since they often serve as active comparators or follow-up treatments in trials that investigate *abaloparatide*. In other words, it seems that with the lower similarity

thresholds, the *CUI_grouping* method links together all records that mention *osteoporosis* and one or more of these three drugs; this explains why this method links a lot more records to this authorization in comparison with the other two.

Table 2.13 shows correct records (identified by the other methods) that were not linked to *Eladynos* by the *CUI_grouping* method. The CTgov record NCT00542425, which is detected by both *EPAR_CTgov* and *PubMed_API*, is not linked by the *CUI_grouping* method because it does not have any drug CUI's. This happens since the text of this record does not contain the term *abaloparatide* but instead has *BA058* as the drug name. Similarly to the *Exondys* example above, while BERN correctly recognizes *BA058* as a drug entity, QuickUMLS does not map it to a CUI, even though *BA058* has a CUI of its own (C4548982), which is hierarchically a narrower concept under *abaloparatide*. The three non-linked PubMed records have the correct drug CUI, but the problem is with their disease CUI's: the first two (28474780, 29251010) do not have the common denominator CUI C0029456 (*osteoporosis*), the third one (27533157) has the *osteoporosis* CUI but also 7 additional disease CUI's. This prevents these records from being sufficiently similar to the ones in Table 2.12, even with the lowest similarity threshold.

As evident from both tables, the PubMed records in this example tend to have "extra" disease CUI's, i.e. various disease entities in addition to the main disease that the drug is intended to treat. These additional disease CUI's include both concepts that are related to the main disease, like C0016658 (fracture), and seemingly irrelevant concepts. Some of these "irrelevant" CUI's originate from the text itself. For example, the record 27826127 (Table 2.12) contains the CUI's C4540463 (bone marrow abnormalities) and C0016059 (fibrosis); looking at the text, it turns out that these disease terms appear under negation in the context of reporting absence of adverse reactions to the treatment ("There were no bone marrow abnormalities, marrow fibrosis..."). On the other hand, the record 27533157 (Table 2.13) contains CUI's that do not originate from the text (and thus are also not part of the BERN-identified entities): C1834129 (abnormal vertebral morphology), C3263723 (traumatic injury), and C0729233 (dissecting aneurysm of the thoracic aorta). These CUI's were erroneously introduced by the QuickUMLS tool. QuickUMLS is based on string matching with a default similarity threshold of 0.7 (Jaccard similarity); therefore, it might introduce incorrect mappings based on partial string overlap. For example, QuickUMLS introduced CUI C0729233 (dissecting aneurysm of the thoracic aorta) based on

Record ID	Disease entity (BERN)	CUI (QuickUMLS)	Drug entity (BERN)	CUI (QuickUMLS)
EMA_1039	osteoporosis	C0029456	abaloparatide	C4042342
NCT01657162	osteoporosis	C0029456	abaloparatide, BA058, BA058-05-003	C4042342
NCT01343004	osteoporosis, fracture, fractures	C0029456, C0016658	abaloparatide, BA058	C4042342
NCT01674621	osteoporosis	C0029456	abaloparatide, BA058	C4042342
27612281	osteoporosis, fractures, vertebral fractures, nonvertebral fractures, fracture, nonvertebral fracture, vertebral fracture	C0029456, C0080179, C0016658	abaloparatide, amino acid	C4042342, C0002520
25393645	osteoporosis, postmenopausal osteoporosis	C0029456, C0029458	abaloparatide, teriparatide	C4042342, C0070093
28160873	osteoporosis, vertebral fractures, nonvertebral fractures, osteoporotic, fractures, fracture	C0029456, C0016658	abaloparatide, alendronate, ALN	C4042342, C0102118
27826127	osteoporosis, fractures, bone marrow abnormalities, marrow fibrosis	C0029456, C4540463, C0016059, C0016658	abaloparatide, teriparatide	C4042342, C0070093

Table 2.12: Biomedical entities and CUI's found in records related to *Eladynos* (linked by cos06)

Record ID	Disease entity (BERN)	CUI (QuickUMLS)	Drug entity (BERN)	CUI (QuickUMLS)
NCT00542425	osteoporosis	C0029456	BA058	n/a
28474780	fracture, frax fracture, vertebral and nonvertebral fracture, rheumatoid arthritis, osteoporotic fractures, osteoporotic fracture	C0003873, C0521170, C1834129, C0016658	abaloparatide	C4042342
29251010	postmenopausal osteoporosis, postmenopausal osteoporotic, fractures, ALN	C0029458, C0232970, C0016658	abaloparatide, alendronate, ALN	C4042342, C0102118
27533157	osteoporosis, vertebral fractures, osteoporotic fractures, vertebral fracture, osteoporotic fracture, lumbar or thoracic vertebral fracture, low-trauma, nonvertebral fracture, fracture, hypercalcemia, vertebral and nonvertebral fractures	C0029456, C3263723, C0080179, C0020437, C0729233, C1834129, C0521170, C0016658	abaloparatide, teriparatide	C4042342, C0070093

Table 2.13: Biomedical entities and CUI's found in records related to *Eladynos* (not linked by cos06)

partial string overlap with the term 'lumbar or thoracic vertebral fracture' from the text (the overlapping part is "thoracic"). Another example is the term 'low-trauma nonvertebral fracture' in the text, which seems to have been mapped to C3263723 (traumatic injury) based on the "trauma" part in the strings. These types of mistakes can be minimized by raising the similarity threshold for the string matcher. For example, raising the threshold to 0.8 keeps all the correct mappings for record 27533157 and eliminates the incorrect mappings to 'dissecting aneurysm of the thoracic aorta' (C0729233) and 'abnormal vertebral morphology' (C1834129); the incorrect mapping to 'traumatic injury' (C3263723) remains, however (even at 0.9 threshold).

To sum up, the *EPAR_CTgov* and *PubMed_API* methods detect in total 4 CTgov records and 16 PubMed records for this authorization; all of those, except for one PubMed record, are correct, i.e. related to *Eladynos*. The *CUI_grouping* method fails to link some of these records to *Eladynos*, mainly because they contain multiple additional disease CUI's that the EMA record does not mention; this issue mainly occurs in PubMed records, which have a lot of text (abstracts). The additional disease CUI's either originate from the text itself (e.g. are mentioned as possible adverse reactions) or are erroneously introduced by the mapping tool QuickUMLS based on partial string overlap.

The *CUI_grouping* method links a lot of additional records to this authorization, that are not linked by the other methods. Based on a sample of these additional records, many of them seem to be unrelated to *Eladynos*; they are grouped together with *Eladynos*-related records because the latter mention not only the investigated drug *abaloparatide*, but also the drugs used as comparators or follow-up treatments, such as *teriparatide* and *alendronate*. This makes the records related to the authorizations of *teriparatide* and *alendronate* sufficiently similar to the records related to *abaloparatide*. This issue mainly applies to PubMed records, since the comparator drugs are likely to be mentioned in the abstracts.

EMA record 13 (*Mysimba*)

Mysimba (active substances: *bupropion* and *naltrexone*) is a treatment for obesity that was approved for marketing by EMA in 2015. The EPAR mentions 21 clinical trials that were submitted with the application. *PubMed_API* does not manage to link records to this authorization; the other two methods do:

- *EPAR_CTgov*: 5 CTgov records and 4 PubMed records;
- *PubMed_API*: no records;
- *CUI_grouping*:
  - *cos08*: no records;
  - *cos07*: 1 CTgov record and no PubMed records;
  - *cos06*: 4 CTgov records and no PubMed records.

Figure 2.16 shows the relation between the CTgov records identified with *EPAR_CTgov* and those identified with *CUI_grouping* (0.6 similarity threshold).

- There is one record (NCT00456521) that is found by both methods.
- There are 4 records (NCT00532779, NCT00567255, NCT00474630 and NCT00364871) that are detected only by *EPAR_CTgov*.
- There are 3 records (NCT00711477, NCT02638129 and NCT02616315) that are detected only by *CUI_grouping*.

In total, both methods detect 8 CTgov records. 7 of them are clearly related to *Mysimba*; the only record whose correctness is not entirely clear is NCT02616315, since it is conducted by a different sponsor. In addition, *EPAR_CTgov* finds 4 correct PubMed records for this authorization: 20673995 (results of NCT00532779), 20559296 (results of NCT00456521), 23408728 (results of NCT00567255), and 24144653 (results of NCT00474630). The 3 records that are identified by only by *CUI_grouping* are not detected by *EPAR_CTgov* because they are not mentioned in the EPAR (this is verified based on their protocol numbers).

PubMed_API, which queries the PubMed database using the active substance and the disease name from the EMA authorization, does not detect any records because of two issues:

- The names of the active substances in the EMA record are *bupropion hydrochloride* and *naltrexone hydrochloride*, while in the PubMed records they are registered as *bupropion* and *naltrexone*. Since the API query is looking for exact matches, this discrepancy results in failure to detect records.

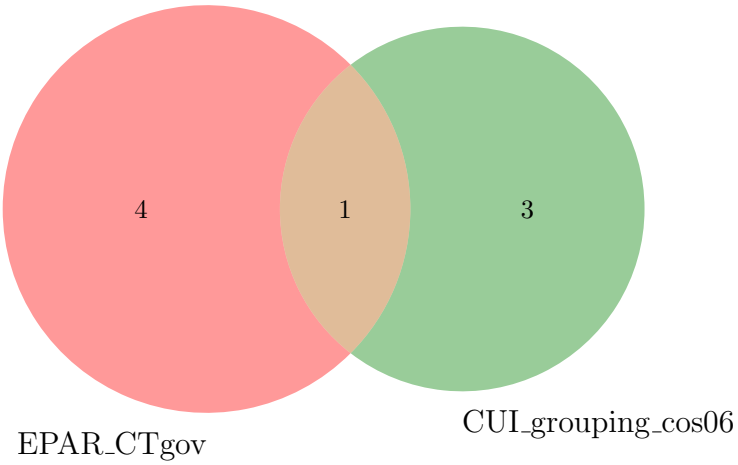

Figure 2.16: Overlap between the CTgov records associated by the three methods with *Mysimba*

- The EMA authorization lists two comma-separated active substances (*bupropion hydrochloride, naltrexone hydrochloride*) and two comma-separated diseases (*Obesity, Overweight*). The API query treats the whole string as one term and searches for a full match, which results in failure to detect records. Splitting the string on a comma would not be a good solution for this issue, since many MeSH disease terms contain a comma, e.g. 'Muscular Dystrophy, Duchenne', 'Diabetes Mellitus, Type 2'.

If these issues are manually solved (by separating the two drug terms and the two disease terms and removing 'hydrochloride' from the drug names), the query returns 12 PubMed records (the 4 identified by *EPAR_CTgov* and additional 8). However, this type of fix cannot be done automatically, and it therefore reveals a limitation of this method.

To understand the performance of the *CUI_grouping* method, we can compare the CUI's in the linked records (Table 2.14) to the CUI's in the records that were identified by *EPAR_CTgov* but not linked by *CUI_grouping* (Table 2.15).

Most of the linked records in Table 2.14 have two disease CUI's: C0028754 (*obesity*) and C0497406 (*overweight*); the exceptions are record NCT02616315, which has only the *obesity* CUI, and record NCT02638129, which has three

Record ID	Disease entity (BERN)	CUI (QuickUMLS)	Drug entity (BERN)	CUI (QuickUMLS)
EMA_13	obesity, overweight	C0028754, C0497406	bupropion hydrochloride, naltrexone hydrochloride	C0355800, C0700563
NCT02638129	obese, cardiovascular disease, cardiovascular events, mace, overweight and obese, cv disease	C0028754, C0497406, C1320716, C0012634, C0007222	naltrexone, bupropion, naltrexone hydrochloride, bupropion hydrochloride, HCl, bupropion HCl	C0355800, C0700563, C0027360, C1512523, C0085208
NCT02616315	weight loss, obese	C0028754	naltrexone HCl, bupropion HCl, naltrexone hydrochloride, HCl, bupropion hydrochloride	C0355800, C0700563
NCT00711477	overweight, obese	C0028754, C0497406	naltrexone SR, bupropion SR, naltrexone, bupropion	C0027360, C0085208
NCT00456521	overweight, obese, obesity	C0028754, C0497406	naltrexone SR, bupropion SR, naltrexone, bupropion	C0027360, C0085208

Table 2.14: Biomedical entities and CUI's found in records related to *Mysimba* (linked by cos06)

Record ID	Disease entity (BERN)	CUI (QuickUMLS)	Drug entity (BERN)	CUI (QuickUMLS)
NCT00364871	obesity, uncomplicated obesity	C0028754	naltrexone, bupropion SR	C0027360, C0085208
NCT00474630	obese, type 2 diabetes, type 2 diabetes mellitus, obesity	C0028754, C0011860	naltrexone, bupropion, naltrexone SR, bupropion SR	C0027360, C0085208
NCT00532779	obese, obesity	C0028754	naltrexone, bupropion, naltrexone SR, bupropion SR	C0027360, C0085208
NCT00567255	obese, obesity	C0028754	naltrexone, bupropion, naltrexone SR, bupropion SR	C0027360, C0085208
23408728	obesity, overweight, obese, dyslipidemia, hypertension, nausea, depression	C0028754, C0497406, C0242339, C0020538, C0027497, C0011570	naltrexone, bupropion	C0027360, C0085208
24144653	overweight and obese, type 2 diabetes, overweight, obese, 2 diabetes, antidiabetes, nausea, constipation, vomiting, depression, hypoglycemia, cardiovascular risk factors, diabetes	C0028754, C0497406, C0007220, C0042963, C0027497, C0011849, C0011860, C0011570, C0009806, C0035648, C0020615	naltrexone, bupropion, glucose, triglycerides, cholesterol	C0027360, C0085208, C0008377, C0041004, C0017725
20559296	cardiometabolic disease, nausea, obesity	C0028754, C0012634, C0027497	naltrexone, bupropion, BMOD	C0027360, C0085208
20673995	weight loss, overweight and obese, obesity, overweight, obese, dyslipidaemia, hypertension, nausea, headache, constipation, dizziness, vomiting, dry mouth, depression, suicidality	C0028754, C0497406, C0018681, C0242339, C0020538, C0042963, C0027497, C0012833, C0011570, C0043352, C0009806, C1262477	naltrexone, bupropion, orexigen	C0027360, C0085208

Table 2.15: Biomedical entities and CUI's found in records related to *Mysimba* (not linked by cos06)

additional CUI's besides *obesity* and *overweight*. In terms of drug CUI's, the linked records have either C0355800 (*naltrexone hydrochloride*) and C0700563 (*bupropion hydrochloride*), or C0027360 (*naltrexone*) and C0085208 (*bupropion*); record NCT02638129 has all four of them (C0355800, C0700563, C0027360, C0085208), and this is probably the link that makes all these records sufficiently similar to each other.

All the non-linked PubMed records in Table 2.15 also have the drug CUI's C0027360 (*naltrexone*) and C0085208 (*bupropion*); the reason they are not linked to the records in Table 2.14 is that they have either too few or too many disease CUI's. The CTgov records (first 4 records in the table) have the CUI C0028754 (*obesity*), but not C0497406 (*overweight*). The PubMed records, on the other hand, have multiple additional disease CUI's besides C0028754 and C0497406. These additional CUI's include the inclusion criteria for the trial (e.g. dyslipidaemia, hypertension), observed adverse events (e.g. headache, constipation, nausea), and ruled-out adverse events (e.g. depression, suicidality). This type of "noise" is characteristic of the PubMed records, since it is likely to be mentioned in the abstract, but not in the text of CTgov or EMA records.

To sum up, this example demonstrates a major limitation of the *PubMed_API* method, which cannot deal with strings containing more than one disease or drug entity. The other two methods both identify correct CTgov records; importantly, each of them identifies different records that the other one fails to detect. In addition, we encounter again a limitation of the *CUI_grouping* method: the free text, especially of PubMed records, contains additional biomedical entities besides the investigated drug and disease; this creates noise in the CUI's and prevents correct linking of PubMed records.

2.9 Conclusion

This chapter described three methods for linking EMA marketing authorizations to related CTgov and PubMed records:

- *EPAR_CTgov* extracts protocol numbers from EMA authorizations, and uses them to search the CTgov register. It then utilizes cross-references from CTgov to PubMed to identify related publications.
- *PubMed_API* extracts drug and disease names from EMA authorizations, and uses them to query PubMed. It then utilizes cross-references

	Linking rate	Precision CTgov	Precision PubMed	Recall CTgov	Recall PubMed
EPAR_CTgov	$\sim 90\%$	very high	very high	medium	medium
PubMed_API	$\sim 60\%$	medium-high	medium-high	low	high
CUI_grouping	$\sim 35\%$	*unknown*	*unknown*	medium	low

Table 2.16: Strengths and limitations of the methods

from PubMed to CTgov to identify related clinical trials.

- *CUI_grouping* extracts and normalizes drug and disease names from the free text of records from all three sources. It then groups together records that mention the same disease and drug names.

Table 2.16 summarizes the strengths and weaknesses of each method. The values in the table are approximated trends, based on the analyzed examples, rather than accurate quantitative results (except for the 'linking rate' column). 'Precision' estimates what fraction of the records retrieved by a method is correct (under our definition); 'recall' estimates what fraction of the correct records is retrieved by a method.

The *EPAR_CTgov* method manages to link records to about 90% of the EMA authorizations in the dataset, an impressive result which indicates that protocol numbers are often registered in CTgov as secondary identifiers and can be effectively used to find trials. Another strength of this method is its high precision: all the records it identifies are necessarily correct, since both the protocol numbers in the EPARs and the cross-references from CTgov to PubMed are manually entered by the responsible parties. However, the recall of the method is not very high. By definition, it can only retrieve CTgov records that are mentioned by their protocol numbers in the EPAR, which is a subset of all the correct CTgov records (e.g. post-authorization trials can never be retrieved by this method); furthermore, it can only retrieve PubMed records that are related to this subset and are referenced in CTgov. Finally, the biggest limitation this method suffers from (and is not indicated in the table) is that the extraction of protocol numbers from EMA's data was performed manually and is hard to automate. Future work that wishes to utilize this method should focus on solving this issue first and foremost.

The *PubMed_API* method manages to link records to about 60% of the authorizations. This limited success rate is related to the way the PubMed query is constructed: the string that appears in the 'therapeutic area' field in the EMA record is entered as a search term for the 'MeSH' field in PubMed, and the string that appears in the 'active substance' field in the EMA record is entered as a search term for the 'substance' field in PubMed. The query cannot deal with cases where the original string in the EMA record contains more than one entity (e.g. '*obesity, overweight*'); furthermore, if the term used in the PubMed record is not exactly the one used in EMA (e.g. narrower/broader concept, synonym, etc.), the query cannot find it. A second limitation of this method is the low recall of CTgov records; the cross-references from PubMed to CTgov are very partial, so this method gives very incomplete results in terms of CTgov records. On the other hand, the performance of this method in retrieving PubMed records is quite good; it manages to link twice as many PubMed records as *EPAR_CTgov*, and based on the analyzed examples it seems that most of them are correct. One thing that affects the precision is that sometimes the method links incorrect records which contain the right disease and drug terms, for example a record about a study in which the drug of interest serves as a comparator. This happens because the 'substance' field in PubMed lists all the important substances mentioned in the publication, not only the investigated drug.

The *CUI_grouping* method has the lowest linking rate: with a 0.6 similarity threshold, it manages to link records only to about 35% of the EMA authorizations; the percentage is even lower with higher similarity thresholds. There are three main reasons for the method's failure to link certain correct records. The first reason is that EMA records usually have one disease CUI and one drug CUI, while CTgov and PubMed records tend to have multiple CUI's (especially disease CUI's in PubMed records); as explained in Section 2.7.3, this affects the cosine similarity and makes EMA records not sufficiently similar to many CTgov and PubMed records (this also makes the 0.8 similarity threshold way too restrictive for the task). Another reason is that some records use broader or narrower concepts than the ones used in the EMA record (e.g. *ataxia* vs. *Friedreich ataxia*); the method does not take into account the hierarchical structure of UMLS concepts, so related CUI's are not considered similar. Finally, some of the linking errors are introduced by the NLP pipeline, namely the BERN and QuickUMLS tools. In the analyzed examples, we have seen three types of errors introduced by the tools: (a) an entity is not identified by BERN, (b) an entity is not mapped to a CUI

by QuickUMLS, (c) an entity is mapped to the wrong CUI by QuickUMLS because of a partial string match.

Despite the low linking rate, *CUI_grouping* links the total largest number of records: more than 2,000 CTgov records and more than 2,000 PubMed records (with the 0.7 and 0.6 thresholds). Surprisingly, this does not result in high recall, based on the analyzed examples; this means that although the method links many records, it does not necessarily manage to link the correct records. This is especially true for PubMed records. The reasons for this failure to link correct records are the same three reasons mentioned above, i.e. multiple CUI's, broader/narrower terms, and errors made by the tools. The precision of the method is hard to estimate based on the analyzed examples. In the two examples where *CUI_grouping* linked only a few records, the precision was good (7/7 CTgov and 2/2 PubMed records are correct for *Exondys*, 3/4 CTgov records are correct for *Mysimba*); however, in the *Eladynos* example *CUI_grouping* links 28 CTgov records and 48 PubMed records and from the analyzed sample of 8 records, only 2 are correct. The big number of total linked records suggests that the *Eladynos* example is more representative, i.e. that many of the linked records are actually not correct. Incorrect records end up being linked to the authorization as a side effect of the multiple CUI's issue: for example, if some drugs are frequently used as comparators to each other (i.e. their CUI's are frequently mentioned together), the *CUI_grouping* method might end up linking together records that are related to *any* of them, including records in which the drug of interest is not even mentioned.

Even though the performance of *CUI_grouping* is the lowest of the three, this method has the most potential for further experimentation and improvements. First, the discrepancy between the number of CUI's in EMA records vs. CTgov and PubMed records can be addressed in one of two ways: (a) reduce the number of CUI's in CTgov and PubMed by processing less text (e.g. only the title and not the abstract), or (b) increase the number of CUI's in EMA records by adding narrower and broader UMLS concepts. Second, the errors introduced by the tools might be reduced by using an alternative pipeline or fine-tuning the tools. As mentioned above, BERN is the current state-of-the-art in bio-NER and in our examples, it barely made any mistakes in entities detection; QuickUMLS, on the other hand, had a poorer performance. One alternative to experiment with is fine-tuning the QuickUMLS parameters, e.g. changing the string similarity threshold used for mapping. Another option is to use a different pipeline altogether, for example the

MetaMap tool[15] which identifies UMLS concepts in a text and maps them to CUI's (i.e. performs in one go the two steps that were done separately in our experiment). Whether these alternatives improve the linking results is an empirical question that I leave for future research.

One of the most important insights from the current experiments is that the three methods complement each other; each method manages to link correct records that the other ones do not identify. The fact that each method only obtains partial results illustrates the challenges that are inherent to the problem of linking clinical trial information from various sources. While the task is far from being solved, the analysis provided in this chapter sheds light on each of the databases and on possible links between them. Hopefully, this information can be useful in future work on connecting the (partially) isolated biomedical databases, so that important insights can be easily and reliably accessed by all stakeholders.

[15] https://metamap.nlm.nih.gov/

Chapter 3

Expressions of Uncertainty in Clinical Trial Publications

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Appendix A

List of EMA Authorizations

Below are the details of the 72 EMA authorizations used for the experiments in Chapter 2.

ID	58
Medicine Name	Vanflyta
Active Substance	quizartinib dihydrochloride
Disease	Leukemia, Myeloid, Acute
Status	Refused
Authorization Year	2019
Sponsor	Daiichi Sankyo Europe GmbH
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/vanflyta

ID	1039
Medicine Name	Eladynos
Active Substance	abaloparatide
Disease	Osteoporosis
Status	Refused
Authorization Year	2019
Sponsor	Radius International Ltd
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/eladynos

ID	1094
Medicine Name	Exondys
Active Substance	eteplirsen
Disease	Muscular Dystrophy, Duchenne
Status	Refused
Authorization Year	2018
Sponsor	AVI Biopharma International Ltd
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/exondys

ID	1164
Medicine Name	Dexxience
Active Substance	betrixaban maleate
Disease	Venous Thromboembolism
Status	Refused
Authorization Year	2018
Sponsor	Portola Pharma UK Limited
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/dexxience

ID	1216
Medicine Name	Aplidin
Active Substance	Plitidepsin
Disease	Multiple Myeloma
Status	Refused
Authorization Year	2018
Sponsor	Pharma Mar, S.A.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/aplidin

ID	1221
Medicine Name	Alsitek
Active Substance	masitinib mesylate
Disease	Amyotrophic Lateral Sclerosis
Status	Refused
Authorization Year	2018
Sponsor	AB Science
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/alsitek

ID	1225
Medicine Name	Xeljanz
Active Substance	Tofacitinib
Disease	Arthritis, Rheumatoid
Status	Refused
Authorization Year	2013
Sponsor	Pfizer Limited
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/xeljanz-0

ID	1291
Medicine Name	Masipro
Active Substance	masitinib mesylate
Disease	Mastocytosis
Status	Refused
Authorization Year	2017
Sponsor	AB Science
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/masipro

ID	1298
Medicine Name	Fanaptum
Active Substance	iloperidone
Disease	Schizophrenia
Status	Refused
Authorization Year	2018
Sponsor	Vanda Pharmaceuticals Ltd
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/fanaptum-0

ID	1311
Medicine Name	Onzeald
Active Substance	etirinotecan pegol
Disease	Breast Neoplasms
Status	Refused
Authorization Year	2018
Sponsor	Nektar Therapeutics UK Limited
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/onzeald

ID	1316
Medicine Name	EnCyzix
Active Substance	enclomifene citrate
Disease	Hypogonadism
Status	Refused
Authorization Year	2018
Sponsor	Renable Pharma Limited
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/encyzix

ID	1328
Medicine Name	Adlumiz
Active Substance	anamorelin hydrochloride
Disease	Cachexia, Anorexia, Carcinoma, Non-Small-Cell Lung
Status	Refused
Authorization Year	2017
Sponsor	-
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/adlumiz

ID	1336
Medicine Name	Human IGG1 monoclonal antibody specific for human interleukin-1 alpha XBiotech
Active Substance	human IgG1 monoclonal antibody specific for human interleukin-1 alpha
Disease	Colorectal Neoplasms
Status	Refused
Authorization Year	2017
Sponsor	XBiotech Germany GmbH
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/human-igg1-monoclonal-antibody-specific-human-interleukin-1-alpha-xbiotech

ID	1425
Medicine Name	Heparesc
Active Substance	Human heterologous liver cells
Disease	Urea Cycle Disorders, Inborn
Status	Refused
Authorization Year	2015
Sponsor	Cytonet GmbH KG
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/heparesc

ID	1439
Medicine Name	Lympreva
Active Substance	dasiprotimut-t
Disease	Lymphoma, Non-Hodgkin
Status	Refused
Authorization Year	2015
Sponsor	Biovest Europe Ltd
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/lympreva

ID	1458
Medicine Name	Masiviera
Active Substance	masitinib
Disease	Pancreatic Neoplasms
Status	Refused
Authorization Year	2014
Sponsor	AB Science
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/masiviera

ID	1470
Medicine Name	Reasanz
Active Substance	Serelaxin
Disease	Heart Failure
Status	Refused
Authorization Year	2014
Sponsor	Novartis Europharm Ltd
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/reasanz

ID	1471
Medicine Name	Nerventra
Active Substance	laquinimod
Disease	Multiple Sclerosis
Status	Refused
Authorization Year	2014
Sponsor	Teva Pharma GmbH
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/nerventra

ID	1490
Medicine Name	Masican
Active Substance	masitinib
Disease	Gastrointestinal Stromal Tumors
Status	Refused
Authorization Year	2014
Sponsor	-
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/masican

ID	1497
Medicine Name	Kynamro
Active Substance	mipomersen sodium
Disease	Hypercholesterolemia
Status	Refused
Authorization Year	2013
Sponsor	Genzyme Europe BV
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/kynamro

ID	1498
Medicine Name	Qsiva
Active Substance	phentermine, topiramate
Disease	Obesity
Status	Refused
Authorization Year	2013
Sponsor	Vivus BV
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/qsiva

ID	1508
Medicine Name	Labazenit
Active Substance	budesonide, salmeterol
Disease	Asthma
Status	Refused
Authorization Year	2013
Sponsor	Laboratoires SMB S.A.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/labazenit

ID	1509
Medicine Name	Istodax
Active Substance	romidepsin
Disease	Lymphoma, Non-Hodgkin
Status	Refused
Authorization Year	2013
Sponsor	Celgene Europe Ltd.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/istodax

ID	1514
Medicine Name	Elelyso
Active Substance	Taliglucerase alfa
Disease	Gaucher Disease
Status	Refused
Authorization Year	2012
Sponsor	Pfizer Ltd.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/elelyso

ID	1534
Medicine Name	Folotyn
Active Substance	Pralatrexate
Disease	Lymphoma, T-Cell
Status	Refused
Authorization Year	2012
Sponsor	Allos Therapeutics Ltd
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/folotyn

ID	1562
Medicine Name	Zeftera (previously Zevtera)
Active Substance	ceftobiprole medocaril
Disease	Skin Diseases, Infectious, Soft Tissue Infections
Status	Refused
Authorization Year	2010
Sponsor	Janssen-Cilag International NV
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/zeftera-previously-zevtera

ID	1567
Medicine Name	Impulsor
Active Substance	milnacipran
Disease	Fibromyalgia
Status	Refused
Authorization Year	2010
Sponsor	Pierre Fabre Medicament
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/impulsor

ID	1575
Medicine Name	Sovrima
Active Substance	idebenone
Disease	Friedreich Ataxia
Status	Refused
Authorization Year	2009
Sponsor	Centocor B.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/sovrima

ID	1579
Medicine Name	Rhucin
Active Substance	recombinant human C1 inhibitor
Disease	Angioedema
Status	Refused
Authorization Year	2008
Sponsor	Pharming Group N.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/rhucin

ID	1583
Medicine Name	Genasense
Active Substance	oblimersen
Disease	Melanoma
Status	Refused
Authorization Year	2007
Sponsor	Genta Development Ltd. c/o Ross Craig
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/genasense

ID	1585
Medicine Name	Zelnorm
Active Substance	tegaserod
Disease	Irritable Bowel Syndrome
Status	Refused
Authorization Year	2006
Sponsor	Novartis Europharm Limited
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/zelnorm

ID	1590
Medicine Name	Gemesis
Active Substance	becaplermin
Disease	Guided Tissue Regeneration, Periodontal
Status	Refused
Authorization Year	2010
Sponsor	BioMimetic Therapeutics Ltd
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/gemesis

ID	1611
Medicine Name	Cimzia
Active Substance	Certolizumab pegol
Disease	Crohn Disease
Status	Refused
Authorization Year	2008
Sponsor	UCB Pharma SA
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/cimzia

ID	1617
Medicine Name	Natalizumab Elan Pharma
Active Substance	natalizumab
Disease	Crohn Disease
Status	Refused
Authorization Year	2008
Sponsor	Elan Pharma International Ltd.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/natalizumab-elan-pharma

ID	1618
Medicine Name	Mylotarg
Active Substance	gemtuzumab ozogamicin
Disease	Leukemia, Myeloid, Acute
Status	Refused
Authorization Year	2008
Sponsor	Wyeth Europa Ltd
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/mylotarg

ID	1619
Medicine Name	Mycograb
Active Substance	recombinant human monoclonal antibody to hsp
Disease	Candidiasis
Status	Refused
Authorization Year	2007
Sponsor	NeuTec Pharma plc
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/mycograb

ID	1622
Medicine Name	Valdoxan
Active Substance	Agomelatine
Disease	Depressive Disorder, Major
Status	Refused
Authorization Year	2007
Sponsor	Les Laboratoires Servier
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/valdoxan-0

ID	3
Medicine Name	Zydelig
Active Substance	Idelalisib
Disease	Lymphoma, Non-Hodgkin, Leukemia, Lymphocytic, Chronic, B-Cell
Status	Authorised
Authorization Year	2014
Sponsor	Gilead Sciences Ireland UC
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/zydelig

ID	5
Medicine Name	Kuvan
Active Substance	Sapropterin dihydrochloride
Disease	Phenylketonurias
Status	Authorised
Authorization Year	2008
Sponsor	BioMarin International Limited
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/kuvan

ID	6
Medicine Name	Poteligeo
Active Substance	Mogamulizumab
Disease	Sezary Syndrome, Mycosis Fungoides
Status	Authorised
Authorization Year	2018
Sponsor	Kyowa Kirin Holdings B.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/poteligeo

ID	8
Medicine Name	Trumenba
Active Substance	Neisseria meningitidis serogroup B fHbp (recombinant lipidated fHbp (factor H binding protein)) subfamily A; Neisseria meningitidis serogroup B fHbp (recombinant lipidated fHbp (factor H binding protein)) subfamily B
Disease	Meningitis, Meningococcal
Status	Authorised
Authorization Year	2017
Sponsor	Pfizer Europe MA EEIG
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/trumenba

ID	9
Medicine Name	Xofigo
Active Substance	radium Ra223 dichloride
Disease	Prostatic Neoplasms
Status	Authorised
Authorization Year	2013
Sponsor	Bayer AG
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/xofigo

ID	11
Medicine Name	Firmagon
Active Substance	degarelix
Disease	Prostatic Neoplasms
Status	Authorised
Authorization Year	2009
Sponsor	Ferring Pharmaceuticals A/S
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/firmagon

ID	13
Medicine Name	Mysimba
Active Substance	bupropion hydrochloride, naltrexone hydrochloride
Disease	Obesity, Overweight
Status	Authorised
Authorization Year	2015
Sponsor	Orexigen Therapeutics Ireland Limited
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/mysimba

ID	14
Medicine Name	Invokana
Active Substance	canagliflozin
Disease	Diabetes Mellitus, Type 2
Status	Authorised
Authorization Year	2013
Sponsor	Janssen-Cilag International N.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/invokana

ID	16
Medicine Name	Besponsa
Active Substance	inotuzumab ozogamicin
Disease	Precursor Cell Lymphoblastic Leukemia-Lymphoma
Status	Authorised
Authorization Year	2017
Sponsor	Pfizer Europe MA EEIG
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/besponsa

ID	18
Medicine Name	Mepact
Active Substance	mifamurtide
Disease	Osteosarcoma
Status	Authorised
Authorization Year	2009
Sponsor	Takeda France SAS
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/mepact

ID	26
Medicine Name	Seebri Breezhaler
Active Substance	Glycopyrronium bromide
Disease	Pulmonary Disease, Chronic Obstructive
Status	Authorised
Authorization Year	2012
Sponsor	Novartis Europharm Limited
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/seebri-breezhaler

ID	28
Medicine Name	Omidria
Active Substance	ketorolac, phenylephrine
Disease	Lens Implantation, Intraocular, Pain, Postoperative
Status	Authorised
Authorization Year	2015
Sponsor	Omeros Ireland Limited
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/omidria

ID	31
Medicine Name	Xydalba
Active Substance	dalbavancin hcl
Disease	Soft Tissue Infections, Skin Diseases, Bacterial
Status	Authorised
Authorization Year	2015
Sponsor	Allergan Pharmaceuticals International Ltd
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/xydalba

ID	32
Medicine Name	Ebymect
Active Substance	dapagliflozin propanediol monohydrate, metformin hydrochloride
Disease	Diabetes Mellitus, Type 2
Status	Authorised
Authorization Year	2015
Sponsor	AstraZeneca AB
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/ebymect

ID	33
Medicine Name	Conbriza
Active Substance	bazedoxifene
Disease	Osteoporosis, Postmenopausal
Status	Authorised
Authorization Year	2009
Sponsor	Pfizer Europe MA EEIG
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/conbriza

ID	34
Medicine Name	Tygacil
Active Substance	Tigecycline
Disease	Bacterial Infections, Skin Diseases, Bacterial, Soft Tissue Infections
Status	Authorised
Authorization Year	2006
Sponsor	Pfizer Europe MA EEIG
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/tygacil

ID	37
Medicine Name	Saxenda
Active Substance	liraglutide
Disease	Obesity, Overweight
Status	Authorised
Authorization Year	2015
Sponsor	Novo Nordisk A/S
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/saxenda

ID	38
Medicine Name	Ristaben
Active Substance	sitagliptin
Disease	Diabetes Mellitus, Type 2
Status	Authorised
Authorization Year	2010
Sponsor	Merck Sharp & Dohme B.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/ristaben

ID	40
Medicine Name	Darzalex
Active Substance	Daratumumab
Disease	Multiple Myeloma
Status	Authorised
Authorization Year	2017
Sponsor	Janssen-Cilag International N.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/darzalex

ID	41
Medicine Name	Tremfya
Active Substance	guselkumab
Disease	Psoriasis
Status	Authorised
Authorization Year	2017
Sponsor	Janssen-Cilag International N.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/tremfya

ID	43
Medicine Name	Kyprolis
Active Substance	carfilzomib
Disease	Multiple Myeloma
Status	Authorised
Authorization Year	2015
Sponsor	Amgen Europe B.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/kyprolis

ID	44
Medicine Name	Edistride
Active Substance	dapagliflozin propanediol monohydrate
Disease	Diabetes Mellitus, Type 2, Diabetes Mellitus, Type 1
Status	Authorised
Authorization Year	2015
Sponsor	AstraZeneca AB
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/edistride

ID	46
Medicine Name	Tasigna
Active Substance	nilotinib
Disease	Leukemia, Myelogenous, Chronic, BCR-ABL Positive
Status	Authorised
Authorization Year	2007
Sponsor	Novartis Europharm Limited
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/tasigna

ID	48
Medicine Name	Qtern
Active Substance	Saxagliptin, dapagliflozin propanediol monohydrate
Disease	Diabetes Mellitus, Type 2
Status	Authorised
Authorization Year	2016
Sponsor	Astra Zeneca AB
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/qtern

ID	49
Medicine Name	Senshio
Active Substance	ospemifene
Disease	Postmenopause
Status	Authorised
Authorization Year	2015
Sponsor	Shionogi B.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/senshio

ID	51
Medicine Name	Cerdelga
Active Substance	eliglustat
Disease	Gaucher Disease
Status	Authorised
Authorization Year	2015
Sponsor	Genzyme Europe BV
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/cerdelga

ID	54
Medicine Name	NovoEight
Active Substance	turoctocog alfa
Disease	Hemophilia A
Status	Authorised
Authorization Year	2013
Sponsor	Novo Nordisk A/S
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/novoeight

ID	57
Medicine Name	Extavia
Active Substance	interferon beta-1b
Disease	Multiple Sclerosis
Status	Authorised
Authorization Year	2008
Sponsor	Novartis Europharm Ltd
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/extavia

ID	62
Medicine Name	Rxulti
Active Substance	brexpiprazole
Disease	Schizophrenia
Status	Authorised
Authorization Year	2018
Sponsor	Otsuka Pharmaceutical Netherlands B.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/rxulti

ID	63
Medicine Name	Tegsedi
Active Substance	inotersen sodium
Disease	Amyloidosis
Status	Authorised
Authorization Year	2018
Sponsor	Akcea Therapeutics Ireland Limited
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/tegsedi

ID	64
Medicine Name	Ocrevus
Active Substance	ocrelizumab
Disease	Multiple Sclerosis
Status	Authorised
Authorization Year	2018
Sponsor	Roche Registration GmbH
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/ocrevus

ID	66
Medicine Name	Crysvita
Active Substance	Burosumab
Disease	Hypophosphatemia, Familial, Hypophosphatemic Rickets, X-Linked Dominant
Status	Authorised
Authorization Year	2018
Sponsor	Kyowa Kirin Holdings B.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/crysvita

ID	72
Medicine Name	Mircera
Active Substance	Methoxy polyethylene glycol-epoetin beta
Disease	Anemia, Kidney Failure, Chronic
Status	Authorised
Authorization Year	2007
Sponsor	Roche Registration GmbH
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/mircera

ID	73
Medicine Name	Quofenix
Active Substance	delafloxacin meglumine
Disease	Skin Diseases, Bacterial
Status	Authorised
Authorization Year	2019
Sponsor	A. Menarini Industrie Farmaceutiche Riunite s.r.l.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/quofenix

ID	75
Medicine Name	Plegridy
Active Substance	peginterferon beta-1a
Disease	Multiple Sclerosis
Status	Authorised
Authorization Year	2014
Sponsor	Biogen Netherlands B.V.
URL	https://www.ema.europa.eu/en/medicines/human/EPAR/plegridy

Appendix B

Biomedical Entities in the Data Sources

In the *CUI_grouping* experiment described in Section 2.7, the three data sources – EMA marketing authorizations, CTgov, and PubMed – are linked based on similarity between the biomedical entities (drug terms and disease terms) that they contain. This appendix reviews in detail the formats in which biomedical entities are present in the three data sources and the various options for extracting and processing them.

Biomedical entities can be found in the data in the following formats:

- As part of the free text of the record, such as the title of a CTgov record; for example, *Efficacy Study of **AVI-4658** to Induce Dystrophin Expression in Selected **Duchenne Muscular Dystrophy** Patients.*
- As part of the structured data of the record, e.g. in the 'Condition' field of a CTgov record. In a structured field, biomedical entities can appear in the form of:
  - free text, e.g. *Duchenne Muscular Dystrophy*;
  - an official term from a controlled vocabulary, such as MeSH, e.g. *Muscular Dystrophy, Duchenne*;
  - an official ID, such as MeSH UI (e.g. D020388) or UMLS CUI (e.g. C0013264).

Table B.1, Table B.2, Table B.3 and Table B.4 show the disease and drug entities found in the structured and unstructured data of different record types.

Two EMA records (*Eladynos* and *Exondys*) were selected as examples, and for each of them one related CTgov record and one related PubMed record were sampled.[1] The first columns in the tables show the available structured fields and their contents, and the last two columns show the entities and CUI's resulting from the processing of the free text of these records. The processing, which is described in detail in Section 2.7.2, includes detection of disease and drug entities with a tool named BERN and mapping of the detected entities to CUI's with a tool named QuickUMLS.[2]

The information that is found in the structured data varies quite a bit between the different record types. For disease entities, EMA records have a field called 'Therapeutic area', which contains the MeSH term for the disease. CTgov has three structured fields: 'Condition', 'Condition MeSH Term' and 'Condition MeSH ID'. PubMed, on the other hand, does not have a special field for disease terms; rather, it has a field for the main topics discussed in the article (in the form of MeSH terms and ID's), and the disease term appears as part of this field. Examples can be seen in Table B.1, which shows three records related to the drug *Eladynos*, and in Table B.2, which shows records related to *Exondys*.

For drug entities, EMA records have two fields (besides the commercial name field): 'Active substance' and 'International non-proprietary name (INN) / common name'. CTgov has four structured fields (not all of them are always filled-in): 'Intervention Name', 'Intervention Other Name', 'Intervention MeSH Term' and 'Intervention MeSH ID', and PubMed has a 'Name of Substance' field, which contains the MeSH term and ID of the chemicals discussed in the article. Examples are shown in Table B.3 and Table B.4.

The *CUI_grouping* method involves grouping together records that share the same drug entities and disease entities. For the grouping, it makes sense to use ID's, rather than text strings, since this normalizes naming variations. The first option would be to utilize the MeSH ID's in the structured data of the records. It should be noted, however, that they cannot be used "as is", but require pre-processing:

- EMA records do not contain ID's; to obtain them, the textual data (disease and drug entities) needs to be automatically mapped either to

[1] The related CTgov and PubMed records were sampled from the records identified by the *EPAR_CTgov* method.

[2] For EMA records, which only have structured data, the first step is skipped and QuickUMLS is applied on the text found in the structured fields.

	Structured Data			Extracted from free text	
Record ID	Condition	MeSH Term	MeSH ID	Disease entity (BERN)	CUI (QuickUMLS)
ema_1039	n/a	Osteoporosis	n/a	n/a	C0029456 (obtained by running QuickUMLS on the MeSH term)
ctgov_ NCT01674621	'Post Menopausal Osteoporosis'	'Osteoporosis', 'Osteoporosis, Postmenopausal'	D000010024, D000015663	'Osteoporosis'	C0029456
pubmed_ 25393645	n/a	Aged, Aged, 80 and over, Bone Density, drug effects, Bone Density Conservation Agents, pharmacology, therapeutic use, Double-Blind Method, Female, Femur Neck, diagnostic imaging, drug effects, Humans, Lumbar Vertebrae, diagnostic imaging, drug effects, Middle Aged, **Osteoporosis, Postmenopausal**, drug therapy, Parathyroid Hormone-Related Protein, pharmacology, therapeutic use, Radiography, Teriparatide, pharmacology, therapeutic use, Treatment Outcome	D000368, D000369, D015519, Q000187, D050071, Q000494, Q000627, D004311, D005260, D005272, Q000000981, Q000187, D006801, D008159, Q000000981, Q000187, D008875, **D015663**, Q000188, D044162, Q000494, Q000627, D011859, D019379, Q000494, Q000627, D016896	'Osteoporosis', 'Post Menopausal Osteoporosis'	C0029456, C0029458

Table B.1: Disease entities in EMA, CTgov and PubMed records related to *Eladynos*

	Structured Data			Extracted from free text	
Record ID	Condition	MeSH Term	MeSH ID	Disease entity (BERN)	CUI (QuickUMLS)
ema_1094	n/a	Muscular Dystrophy, Duchenne	n/a	n/a	C0013264 (obtained by running QuickUMLS on the MeSH term)
ctgov_ NCT01396239	'Duchenne Muscular Dystrophy'	'Muscular Dystrophies', 'Muscular Dystrophy, Duchenne'	D000009136, D000020388	'Duchenne Muscular Dystrophy', 'DMD'	C0013264
pubmed_ 23907995	n/a	Adolescent, Child, Double-Blind Method, Dystrophin, genetics, Humans, Male, Morpholinos, Muscle, Skeletal, pathology, **Muscular Dystrophy, Duchenne**, drug therapy, genetics, pathology, Mutation, Oligonucleotides, therapeutic use, Treatment Outcome	D000293, D002648, D004311, D016189, Q000235, D006801, D008297, D060172, D018482, Q000473, **D020388**, Q000188, Q000235, Q000473, D009154, D009841, Q000627, D016896	'Duchenne Muscular Dystrophy'	C0013264

Table B.2: Disease entities in EMA, CTgov and PubMed records related to *Exondys*

	Structured Data				Extracted from free text	
Record ID	Intervention name / Active substance	Intervention other name / INN	MeSH Term	MeSH ID	Drug entity (BERN)	CUI (QuickUMLS)
ema_1039	abaloparatide	abaloparatide	n/a	n/a	n/a	C4042342 (obtained by running QuickUMLS on the active substance)
ctgov_NCT01674621	Abaloparatide Transdermal (50 mcg), Abaloparatide Transdermal (100 mcg), Abaloparatide Transdermal (150 mcg), Abaloparatide Injection (80 mcg), Abaloparatide Placebo	BA058 Transdermal (50 mcg), BA058 Transdermal (100 mcg), BA058 Transdermal (150 mcg), BA058 Injection (80 mcg), BA058 Placebo	Abaloparatide, Parathyroid Hormone-Related Protein	C000596789, D000044162	BA058, abaloparatide	C4042342
pubmed_25393645	n/a	n/a	Bone Density Conservation Agents, Parathyroid Hormone-Related Protein, Teriparatide, abaloparatide	D050071, D044162, D019379, C000596789	abaloparatide, teriparatide	C4042342, C0070093

Table B.3: Drug entities in EMA, CTgov and PubMed records related to *Eladynos*

	Structured Data				Extracted from free text	
Record ID	Intervention name / Active substance	Intervention other name / INN	MeSH Term	MeSH ID	Drug entity (BERN)	CUI (QuickUMLS)
ema_1094	eteplirsen	eteplirsen	n/a	n/a	n/a	C4283710 (obtained by running QuickUMLS on the active substance)
ctgov_NCT01396239	AVI-4658 (Eteplirsen), Placebo	Eteplirsen-Phosphorodiamidate Morphilino Oligomer, Phosphate buffered saline	n/a	n/a	AVI-4658, eteplirsen	C4283710
pubmed_23907995	n/a	n/a	Dystrophin, Morpholinos, Oligonucleotides, eteplirsen	D016189, D060172, D009841, C000611335	eteplirsen, phosphorodiamidate, nitric oxide	C4283710, C0028128

Table B.4: Drug entities in EMA, CTgov and PubMed records related to *Exondys*

MeSH ID's or CUI's.

- The MeSH ID's used in CTgov have a non-standard format, namely they contain additional zeroes. For example, the correct MeSH ID for the term '*osteoporosis, postmenopausal*' is D015663; however, in CTgov it is registered as D000015663 (see additional examples in the tables). To use these ID's for grouping, they would need to be processed and standardized (assuming that the format observed in the examples is consistent across the whole database).

- The disease MeSH ID's in PubMed do not have a dedicated field, but rather appear in the general field that describes the topics of the article. For the grouping experiment, the disease ID's would need to be isolated from the other MeSH ID's in the field; this could be done by e.g. utilizing the hierarchical tree structure of MeSH, which groups all diseases under one branch.

The second option would be to utilize the unstructured data of the records: detect the drug and disease terms in the free text and map them either to MeSH ID's or CUI's. In the examples shown in the tables, the results of this processing are quite accurate:

- In Table B.1, the disease terms used in the structured data are '*osteoporosis*' (used in the EMA and CTgov records) and the more specific term '*osteoporosis, postmenopausal*' (used in the CTgov and PubMed records). The processing of the free text results in the correct two CUI's: C0029456 (*osteoporosis*) and C0029458 (*osteoporosis, postmenopausal*).

- In Table B.2, the disease terms used in the structured data are '*Muscular Dystrophy, Duchenne*' (used in all three records) and the broader concept '*Muscular Dystrophies*' (used in CTgov). The processing of the free text results in the correct (more specific) CUI: C0013264 (*Muscular Dystrophy, Duchenne*).

- In Table B.3, the active substance in the authorization is '*abaloparatide*'; this term is present in all three records. The CTgov and PubMed records in the table contain additional terms: '*Parathyroid Hormone-Related Protein*' (*abaloparatide* is an analog of this protein), '*Bone Density Conservation Agents*' (a broader concept that *abaloparatide*

fits under), and '*teriparatide*' (another drug that is used as a comparator in one of the trials). The processing of the free text identifies the correct CUI in all three records: C4042342 (*abaloparatide*); in addition, the comparator drug *teriparatide* (C0070093) is identified in the PubMed record.

- In Table B.4, the active substance in the authorization is '*eteplirsen*'; this term is present in all three records. The PubMed record in the table contains additional terms: '*Dystrophin*' (the protein whose production the drug stimulates), '*Morpholinos*' (a broader concept that *eteplirsen* fits under), and '*Oligonucleotides*' (a broader concept that *eteplirsen* fits under). The processing of the free text identifies the correct CUI in all three records: C4283710 (*eteplirsen*). In addition, the CUI C0028128 (*nitric oxide*) is detected in the PubMed record; this substance is mentioned in the abstract but is not related to the investigated drug (i.e. can be considered as noise created by the processing method).

To sum up, based on the sampled examples, obtaining drug and disease ID's for the grouping experiment can be achieved either by processing the unstructured data and mapping the detected entities to CUI's, or by processing the available structured data. Which method would result in more accurate groupings is an empirical question, which was not investigated as part of this research. The option that was chosen in the current research is processing of the free text; the method and the results are described in Section 2.7.

Bibliography

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., & Morissette, J. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, *41*(5), 706–716.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, *284*(5), 34–43.

Hassanzadeh, O., Kementsietsidis, A., Lim, L., Miller, R. J., & Wang, M. (2009). LinkedCT: A linked data space for clinical trials. *arXiv preprint arXiv:0908.0567*.

Huser, V., & Cimino, J. J. (2013). Linking ClinicalTrials.gov and PubMed to track results of interventional human clinical trials. *PloS one*, *8*(7), e68409.

Kim, D., Lee, J., So, C. H., Jeon, H., Jeong, M., Choi, Y., Yoon, W., Sung, M., & Kang, J. (2019). A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, *7*, 73729–73740.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240.

Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C. S., Willighagen, E., Hajagos, J., Marshall, M. S., Prud'hommeaux, E., Hassanzadeh, O., Pichler, E. Et al. (2011). Linked open drug data for pharmaceutical research and development. *Journal of cheminformatics*, *3*(1), 19.

Simes, R. J. (1986). Publication bias: The case for an international registry of clinical trials. *Journal of clinical oncology*, *4*(10), 1529–1541.

Soldaini, L., & Goharian, N. (2016). Quickumls: A fast, unsupervised approach for medical concept extraction, In *Medir workshop, sigir*.