Utrecht University

Department of Computer Science

# Improving Health Policy Research through Automated Knowledge Extraction from Medicines Authorisation and Reimbursement Reports

*Author:*
Inge Gimbel *(4268733)*

*Supervisors:*
dr. M.W. Chekol
dr. R. Vreman
prof. dr. Y. Velegrakis

*A thesis submitted in fulfillment of the requirements for the degree of*
*Master of Science in Computer Science*

July 5, 2021

# Acknowledgement

I would like to thank my supervisors, Mel Chekol, Rick Vreman and Yannis Velegrakis, for their guidance and feedback throughout this project. They were always willing and enthusiastic to assist in any way they could throughout the thesis project.

# Abstract

The evaluation of new drugs from medicines authorisation and reimbursement reports is a time-consuming process. Efficient strategies for identification and extraction of information from these reports would benefit the evaluation significantly. Due to the unstructured nature of the reports, automatic extraction of information is challenging. In this thesis, we propose an information extraction method based on regular expressions to automatically extract information from the reimbursement reports. We showed that the developed information extraction method is able to achieve high performance, with an overall F-score of 0.87. Finally, we discuss how the performance of the information extraction method can be improved in future research.

# Contents

# 1 Introduction

Information extraction (IE) has been a continuously growing research field in recent decades, owing to its effectiveness in a wide range of application areas (1). One specific area in which automatic IE could improve research significantly, but is not yet applied, is the domain of health policy research, specifically the evaluation of new drugs from medicines authorisation and reimbursement reports.

The European Medicines Agency (EMA) is an agency of the European Union (EU) in charge of the evaluation and supervision of medicinal products (2). After approval of the EMA a drug may be administered to patients. Because new drugs are often expensive, national healthcare systems cannot afford to pay for all of them. National public organizations known as 'reimbursement organizations' analyze whether the benefits of medications outweigh their costs in order to determine which medications should be paid for. The usage of a drug will be determined by incremental clinically significant effects relative to other treatments presently available. The willingness or ability to pay for the drug, as well as insurance coverage, are also important factors determining the use of a drug. The process of clinical and economic value assessment is formalized in Europe through evaluations by Health Technology Assessment (HTA) bodies. The National Health Care Institute (Zorginstituut Nederland, ZIN) is the responsible body in the Netherlands (3). The decision-making process for the evaluation of drug applications is complex. Based on the assessment of non-clinical, clinical, and quality data submitted by the pharmaceutical industry, regulators have to make sure that only products with a positive benefit-risk balance are brought to the public (4). In some cases the EMA conditionally approves drugs based on a less comprehensive evidence package to allow timely access to novel drugs when the immediate availability of the drug outweighs the risks associated with the less complete evidence package. The benefit-risk balance stil needs to be positive, however given the drug's potential to treat unmet medical needs, additional uncertainty may be tolerated for conditionally approval (5). These drugs are of particular interest for health policy research because the greater uncertainty associated with these drugs means there is a bigger risk of making a wrong decision. Regulatory authorities and reimbursement organizations publish their assessment reports for each drug online. Researchers analyze these reports in a complex and inefficient manner. Data is extracted from each report manually, and the raw data is transformed to data that is suitable for analysis. This is a time-consuming process that is also prone to human error. As a result, a second researcher frequently duplicates the data extraction and modification, putting additional demand on resources. Individual researchers spend many hours personally locating, obtaining, and analyzing the data contained in these publicly available reports (6). Manual extraction is time consuming and expensive. Furthermore, readily adaptable information technologies exist to automate and reliably extract this type of information from free-text data.

To address the aforementioned concerns regarding the process of evaluating drug reports, the objective of this thesis is to automate knowledge extraction

from the reimbursement reports. More precisely, we developed an IE method based on a limited set of rules and regular expressions to extract information related to the drug. As a result health policy research will be less time-consuming and more accurate, which in turn will lead to more efficient evaluations and more appropriate decisions regarding the approval and reimbursement of new and expensive medicines.

# 2 Preliminaries

## 2.1 Regular expressions

A regular expression (shortened as regex) is a specific kind of text pattern that you can use with many modern applications and programming languages. You can use them to verify whether input fits into the text pattern, to find text that matches the pattern within a larger body of text, to replace text matching the pattern with other text or rearranged bits of the matched text and to split a block of text into a list of subtexts (7). A regular expression is a sequence of characters that defines a search pattern. The set of strings matched by the regular expression is a language. A regular expression $R_i$ represents a language $L(R)$ over an alphabet $\Sigma$, where $L(R)$ is a (possibly infinite) set of strings. For a given language, there are many regular expressions that can describe it (8).

An overview of common regexes is shown in Table 1. Next, we discuss some of these regexes.

Alternation in regexes enables matching with multiple sub-expressions. For example, [drug|medicine] will match an input string with drug or medicine or both. Quantifiers in regexes allow for a sub-expression to be used a varying number of times while matching. The most common quantifier is the zero or more quantifier denoted by a star [*]. This quantifier allows to match the sub-expression zero or more times in succession while matching. For example, for a regex [drug*] matches the string made up of any number of 'drug's. Other quantifiers are the question mark, a zero or once quantifier and plus, an once or more quantifier. Quantifiers can be either greedy or lazy. A greedy quantifier matches the longest possible string and a lazy quantifier the shortest. For example, the greedy expression [h.+l] matches 'hell' in 'hello', but the lazy [h.+?l] matches only 'hel'. The question mark is used as lazy quantifier. A character class can be negated so that it matches any single symbol not in the original character class. This is done by adding a caret symbol ˆ. For example, [ˆabc] will match any single symbol, except for a, b or c. Lookaround assertions allow for the placement of restrictions on prefixis or suffixes starting or ending at specific positions in the input string. There are four main types of lookaround assertions, positive or negative lookahead or lookbehind. For example, the positive lookahead ($[A(?=B)]$, find expression A where expression B follows). The positive lookbehind ($[(?<=B)A]$, find expression A where expression B precedes) syntaxes. The negative function similarly as the positive counterparts, but instead assert that the sub-expression should not match a prefix or suffix at some position (7; 9).

When combining different metacharachters and ordinary charachters complex regexes can be created. For example, `[a-zA-Z0-9]+`, matches one to infinete characters and numbers. Another example of a more complex regular expression is `(\W|^)[\w.\-]{0,25}@(yahoo|hotmail|gmail)\.com(\W|$)`. This regex

matches any email adress from the domains yahoo.com, hotmail.com, and gmail.com. A regular expression that can be used for checking if a given set of characters is any email address or not is `^([a-zA-Z0-9_\-\.]+)@([a-zA-Z0-9_\-\.]+)\.([a-zA-Z]{2,5})$`. These are some examples that illustrate the large amount of possibilities someone has with regular expressions.

Table 1: List of common regular expression

| Metacharacter | Description | Metacharacter | Description |
|---|---|---|---|
| . | Any character except newline | ^ | Start of string |
| *a* | The character a | $ | End of string |
| *ab* | The string ab | (?=...) | Positive lookahead |
| *a*\|*b* | a or b | (?!...) | Negative lookahead |
| *a** | 0 or more a's | (?<=...) | Positive lookbehind |
| \ | Escapes a special character | (?<!...) | Negative lookbehind |
| *[ab-d]* | One character of: a, b, c, d | * | 0 or more |
| *[âb-d]* | One character except: a, b, c, d | + | 1 or more |
| *[\b]* | Backspace character | ? | 0 or 1 |
| *d* | One digit | {2} | Exactly 2 |
| *D* | One non-digit | {2, 5} | Between 2 and 5 |
| *s* | One whitespace | {2,} | 2 or more |
| *S* | One non-whitespace | {,5} | Up to 5 |
| *w* | One word character | \b | Word boundary |
| *W* | One non-word character | \B | Non-word boundary |

## 2.2 Regular expression operation

Python has a built-in package called re[1], which can be used to work with Regular Expressions (Table 2). The re module offers a set of functions that allows to search a string for a match. One of the functions, named findall, returns all non-overlapping matches of pattern in a string, as a list of strings. The string is scanned left-to-right, and matches are returned in the order found. If one or more groups are present in the pattern, return a list of groups; this will be a list of tuples if the pattern has more than one group. Empty matches are included in the result. The function findall is used in our IE method, to capture all outcomes of the regex patterns.

[1]re - Regular expression operations. https://docs.python.org/3/library/re.html

```
re.findall(pattern, str)
```

| Sr. No. | Parameter and description |
|---|---|
| 1 | pattern: this is the regular expression to be matched. |
| 2 | str: this is the string, which would be searched to match the pattern at the beginning of string. |

Table 2: The Findall function.

## 2.3 Performance measurements

Performance of the information extraction algorithm is measured by precision, recall, accuracy and the F-score. Precision (positive predictive value) is the fraction of relevant information among the retrieved information, while recall (sensitivity) is the fraction of relevant information that were retrieved. Both precision and recall are therefore based on relevance. Precision is calculated by dividing the number of correct information element instances extracted by the algorithm (true positives), when compared to the reference standard, by the total number of information element instances extracted (true positives and false positives) [Eq. (1)]. Recall is calculated by dividing the number of correct information element instances extracted by the total number of information element instances existing. (true positives and false negatives) [Eq. (2)]. Accuracy is calculated by dividing the number of correct information element instances extracted by the total number of items reviewed [Eq. (3)]. The F-score is the harmonic mean of the precision and recall [Eq. (4)]. The highest possible value of an F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero.

$$P = \frac{number\ of\ correct\ information\ element\ instances\ extracted}{total\ number\ of\ information\ element\ instances\ extracted} \quad (1)$$

$$R = \frac{number\ of\ correct\ information\ element\ instances\ extracted}{total\ number\ of\ information\ element\ instances\ existing} \quad (2)$$

$$A = \frac{number\ of\ correct\ information\ element\ instances\ extracted}{total\ number\ of\ items\ reviewed} \quad (3)$$

$$F - score = \frac{2 * (precision * recall)}{(precision + recall)} \quad (4)$$

# 3 Literature overview

Information Extraction (IE) is the process of extracting specific (pre-specified) information from textual sources. The goal of IE is to offer a structured representation of the retrieved data from examined text (10). IE involves the ability to extract relevant information from unstructured data without having to manually search through a large volume of data for the exact information needed (11). Users define the information to be extracted, which includes predefined concepts of interest, associated entities, and relationships between entities and events (12).

Numerous research efforts have been conducted for IE in various domains (13; 14; 15). Named entity recognition (NER), attribute extraction, relation extraction and event extraction are four primary foci of IE research. Event extraction aims to extract instances of multiple concepts (16). The other primary foci aim at extracting instances of a single concept or of two related concepts (17).

*Natural language processing (NLP) approach*
Most of the IE methods make use of Natural language processing (NLP). NLP is a range of computational techniques for the automatic analysis and representation of human language (18). NLP entails obtaining information on how humans comprehend and utilize language. This is done in order to develop appropriate techniques that will allow computer systems to comprehend and modify natural languages in order to accomplish a variety of desired tasks. (19). NLP has been used for wide range of tasks, including information retrieval (IR) and information extraction (IE) (18). NLP employs a variety of knowledge representations, including a lexicon of words, their meanings, and grammatical characteristics, a collection of grammar rules, and, in certain cases, additional resources such as an ontology (20). Frequently used NLP techniques include tokenizing, sentence splitting, part of speech (POS) tagging, named entity recognition (NER) and entity linkage (21).

*Named-entity recognition and entity linkage*
Named entity recognition (NER), also known as named-entity indentification (NEI), addresses the problem of identification and classification of predefined concepts. NER is a frequently used technique in previously conducted medication extraction studies (22; 23; 24; 25). NER is an IE subtask that organizes and locates named entities referenced in unstructured text into pre-defined categories, such as dates, persons, organizations, places, etc. NER is followed by entity linkage. Entity linkage is the task of assigning a unique identity to entities mentioned in text. For example, given the sentence 'Amsterdam is the capital of The Netherlands', the expected output of an entity linkage system would be 'Amsterdam' and 'The Netherlands'. Linking of entities is normally performed through an ontology, which encompasses a representation, formal naming and definition of the relations between entities. After entity linkage, relation extraction can be performed to extract the required pre-defined information.

Tikk and Solt (22) applied a NER based approach a medication extraction challenge, in which medication names together with details of their administration were to be extracted from medical discharge summaries. They decomposed the medication extraction problem into three subtasks: NER, filtering and relation extraction. For this they developed a rule based NER based on dictionaries and manually created regular expressions. Filtering was performed by a context-aware rule based approach. They partitioned discharge summaries into zones and negative statements were removed. They achieved an entry level F-score of 80 % for exact and inexact matching for both approaches.

*Regular expressions*
Despite the fact that several machine learning techniques for IE have been proposed in recent years, manually generated regular expressions remain a commonly used practical option for IE (26). Regular expression is a common technique for providing an interpretable answer to text categorization among various rule-based approaches. Regular expressions are the key to powerful, flexible, and efficient text processing. By using regular expressions a large class of entity extraction tasks can be accomplished. The technique has been used in many Natural Language Processing (NLP) studies. Regular expressions are created manually by experts with domain knowledge in the majority of these studies (27). Manually created regular expressions can achieve high recall and precision. However, the manual construction regular expressions that guarantee a high performance on these measurements is a tedious manual task and requires expert knowledge. Therefore, multiple approaches have been proposed that automatically infers regular expressions from a set of sample entities (26; 28; 29). It learns effective regular expressions that can be easily interpreted and modified by a user.

In biomedical research regular expressions have been widely used for IE in recent years (30; 31; 32). These studies showed that regular expression based approaches can achieve high performance. A simple algorithm using regular expressions to detect systemic treatments administered to patients was created by Maguire and colleagues (31). As a result of the algorithm considerable time was saved compared to manual review. Garvin and colleagues (32) successfully extracted information on left ventricular ejection fraction from echo-cardiogram reports using a set of regular expressions, with a F-score higher than 99%.

*Rule-based approach vs machine-learning approach*
IE approaches can be classified as rule-based or machine-learning, depending on whether domain experts create the rules or a machine learns them automatically. The rule-based approach establishes rules for extracting specified information using domain knowledge and matching with extraction rules to retrieve relevant information. Because the extraction rules are developed manually, the rule-based technique has a relatively high accuracy in extracting text information, but it is time intensive and requires great effort to develop the rules. The machine-learning approach, on the other hand, is a method of automatically generating extraction rules by teaching a computer to recognize specific pat-

terns in text documents. However, if the amount of training data available in a given domain is insufficient, the accuracy of the machine-learning approach may be compromised. To put it another way, if the IE is done with machine learning without enough training data, the accuracy and dependability of the IE outcomes will be poorer than with the rule-based approach (21).

In this thesis, we used an unsupervised regular expression and rule-based IE approach. The main reason for choosing a rule-based approach is the difficulty of securing sufficient data for machine learning. Only a limited number of reimbursement reports were available as data for this thesis. Furthermore, rule-based approach can achieve high accuracy in extracting text information and are easy to adopt. Regular expressions can express a wide range of patterns. As a result regular expression based approaches are able to cover almost all instances of a specific concept to extract in free text. Due to the versatility of the regular expression it is widely used in text processing and parsing. Because of the heterogeneous nature of the reports used in this thesis, regular expressions can be of great value to extract the required information. Carefully constructed regular expressions can achieve high recall and precision. In our approach automatic-extraction rules and regular expression patterns were manually developed. Patterns of the occurrence of information in a sample of the available reports were analysed with the use of domain knowledge to create the regular expressions. Together with the rule based approach the required information was extracted. The developed method showed some promising results, with a high overall F-score of 0.87.

## 3.1 Research question

How to automatically extract knowledge from medicines authorization and reimbursement report to improve health policy research?

# 4 Methods

## 4.1 Data description

The data used in this thesis consists of a set of 22 Health Technology Assessment (HTA) reports from the Dutch National Health Care Institute (Zorginstituut Nederland, ZIN). The HTA reports were stored in Portable Document Format (PDF) files. Only reports regarding conditionally approved drugs were included. The reports contain the substantive assessment of the therapeutic value of the conditionally approved drugs. The length of the reports differed between 18 and 265 pages. ZIN systematically publishes full initial HTA reports and re-assessment reports on their websites. HTA reports were retrieved by searching agencies' websites for the drug generic and brand name. HTA reports were included from the first report until the most recent published reports, between 2006 and 2016. Vaccines were excluded, because HTA organizations assess vaccines differently from other drugs.
In this thesis different terms are used to refer to the used data: 'data', 'documents' and 'reports'. They are interchangeable.

## 4.2 Ethical and legal considerations of the data

Sources of the data used in this thesis include real-world data, clinical trials and adverse drug reaction reports. Data used in this thesis consists of aggregated information. No individual data is used. The data are freely available for downloading from the website of ZIN. Thus, all data is public and does not relate to individual patients.

## 4.3 Problem definition

We have a set of documents $D = \{D_1, ..., D_n\}$. A document $D_i$ is a sequence of syntactically correct sentences. An attribute (or feature) $A_i$ can be seen as a data field that represent the characteristics or features of a data object. An attribute $A_i$ comes with a corresponding value $v_i$. The values $v$ and corresponding attributes $A$ can be found in the documents $D$. Given a document $D_i$, the challenge is to extract the required value $v_i$, corresponding to the given attribute $A_i$. The required attributes $A$ to extract are pre-defined by a domain expert and are shown in Table 3. As the documents $D$ are heterogeneous of nature, the task of extracting the required values $v$ is difficult.
To address this challenge, we constructed an information extraction (IE) method based on regular expression patterns, and consequently extracted the attributes values $v$ from the documents $D$.

## 4.4 Data to extract

In total 7 attributes $A = \{A_1, ..., A_7\}$ were pre-defined by a domain expert to extract the corresponding values $v$ from the documents $D$. The attributes

consists of the generic drug name, $A_{gn}$, the brand drug name, $A_{dn}$, the date of the HTA report, $A_{date}$, the indication of the drug, $A_{ind}$, the orphan status of the drug, $A_{os}$, the trial name in which the drug is investigated, $A_{tn}$, and the therapeutic value outcome of the drug, $A_{tv}$. The attributes and a description of the attributes are listed in Table 3.

Table 3: Pre-defined attributes

| Attributes | Description | Values to extract |
|---|---|---|
| *Generic name* | Generic name of the drug. | Name |
| *Brand name* | Brand name of the drug | Name |
| *Date of HTA report* | Date the HTA report is publiced. | Date |
| *Indication* | The registered indication of the drug. | Sentence |
| *Orphan status* | The orphan status of the drug. | 'Yes' or 'No' |
| *Trial name* | Name of the pivotal trial. | Name |
| *Therapeutic value* | The therapeutic value of the drug. | Sentence. |

## 4.5 Objective F-score

The objective of the developed method is to reach a minimum F-score of 0.3.

## 4.6 Algorithm overview

In our IE method, we first preprocessed the documents by identifying tables and extracting text. Then we created a regular expression (regex) based pattern to identify the information from the text, finally we extracted the information using a rule based approach.

Figure 1 shows an overview of the IE method. Figure 2 illustrates an example of the information extraction method. The IE method is explained into more detail in the next sections.

### 4.6.1 Table identification

Data was available as a set of documents $D = \{D_1, ..., D_n\}$ in Portable Document Format (PDF) files. Tables were identified from the documents using tabula-py[2]. The tables were stored as CSV files. For each document $D_i$ multiple tables $\{T_{1Di}, T_{2Di}, ..., T_{nDi}\}$ were identified (Fig. 1(A))).

### 4.6.2 Text extraction

The text was extracted from the documents $D$ using PDFMiner[3]. The documents $D$ in PDF format were converted into a list of strings $S$. Each string $S_i$

[2]A. Ariga. tabula-py. Version 2.2.0. https://pypi.org/project/tabula-py/
[3]Y. Shinyama. PDFMiner. Version 20191125. https://pypi.org/project/pdfminer/

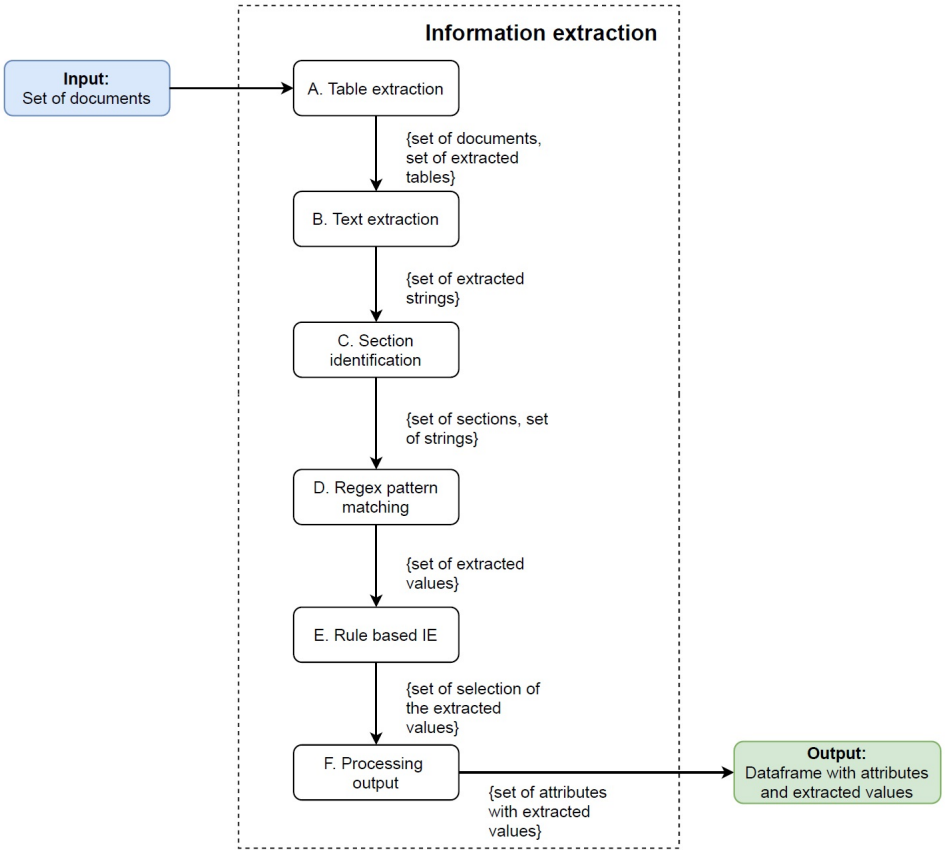


Figure 1: Overview of the information extraction method.

represented one document $D_i$ (Fig. 1(B) and 2(B)).

#### 4.6.3 Section extraction

Because of the general nature of some of the required values $v$ to extract, a particular section $P_i$ was extracted first and used as a filter (Figure 1, C). Section extraction was performed using regular expression patterns that identified the section by paragraph headers or specific words. After identification of the section header, a number of sentences after the matched pattern were extracted from the text with the syntax `([name of section header])([^.]*\.([^.]*\.){n})`. To be certain that the complete section is extracted, a large number of sentences after the header were extracted. Then, the extracted section $P_i$ was used to extract the required value $v_i$ (Fig. 1(C) and 2(C)).

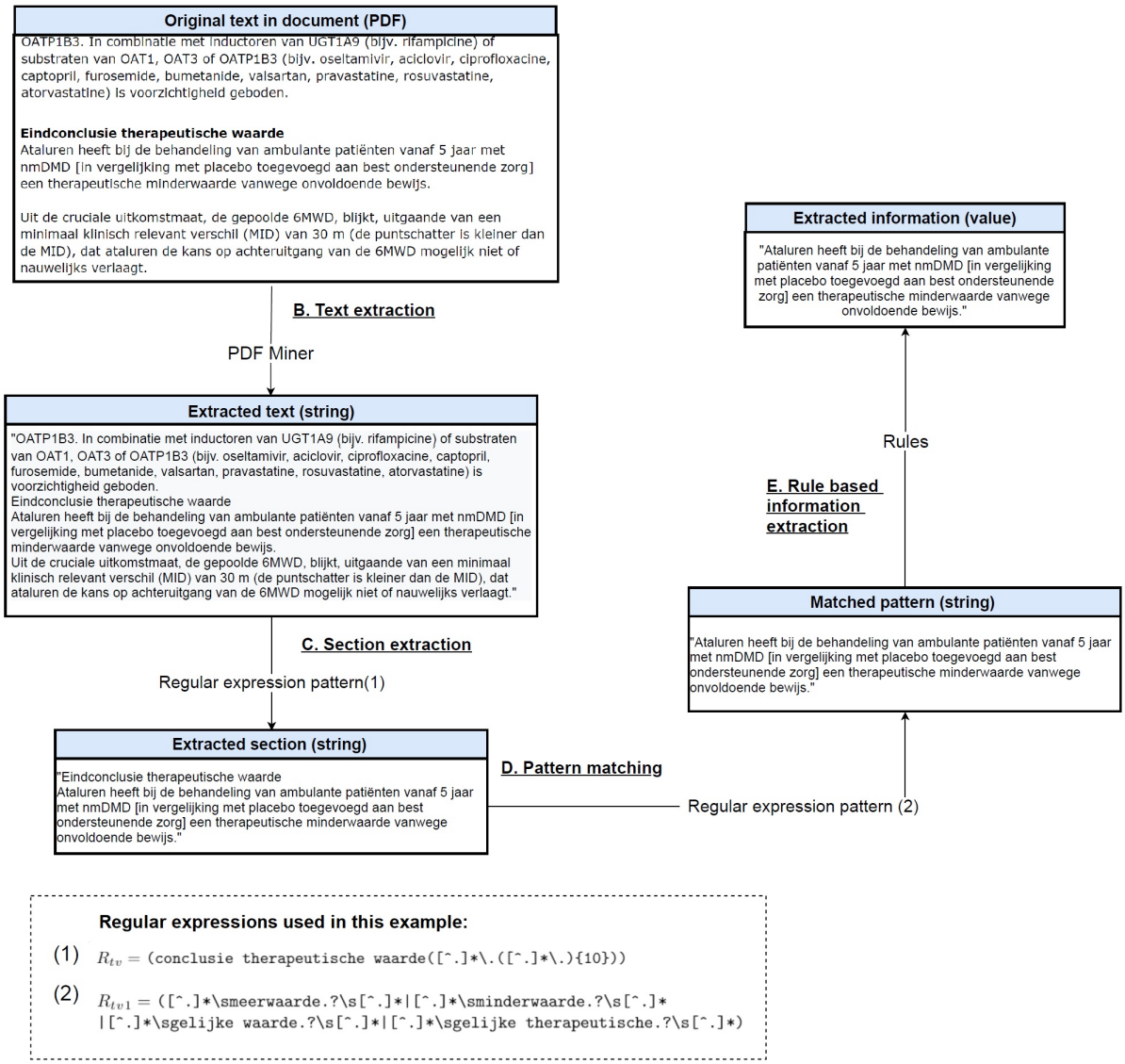


Figure 2: Illustrative example applying the proposed information extraction methodology

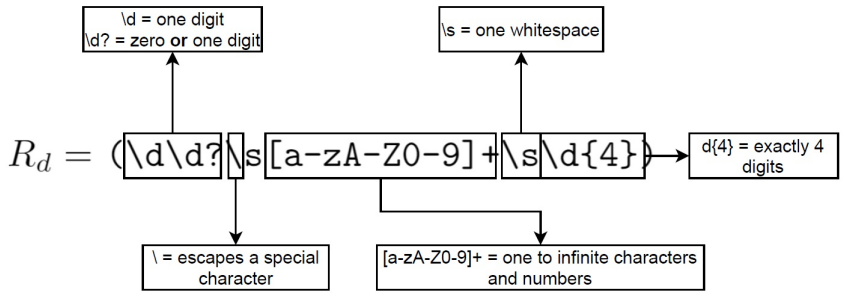


Figure 3: Regular expression of date parsed.

#### 4.6.4 Regular expression based patterns

Five documents $\{D_1, ..., D_5\}$ were randomly selected from the dataset $D$. These documents were manually analyzed to create the patterns $R$ in Python grammar, using regular expressions. Multiple patterns were used for the extraction of the required values $v_i$ (Fig. 1(D) and 2(D)). The patterns are hand-crafted and domain knowledge was used to develop them. Table 4 gives an overview of all the patterns used.
Patterns $R$ were based on the (1) structure, (2) context or (3) exact matching of the information to extract.

*(1) Structure based regular expressions*
The options to create patterns $R$ with regular expression are numerous. Regular expressions can capture a wide range of possible character and/or number combinations. Therefore, regular expressions can be used to extract values $v$ with a specific structure.

A structure based pattern was used to extract the value of the attribute date, $A_{date}$. A date has a specific structure, with one or two digits (day), followed by the name of the month or one or two digits that represent a month, and ending with four digits (year). A pattern was created that represented this particular date structure $[R_d]$. In figure 3 the different components of the pattern are showed and explained.

*(2) Context based regular expressions*
To correctly extract the required value $v_i$ from the documents, the context in which the value is located can be used. First the context of the value was indentified, then the value was extracted. Regular expressions offer multiple ways to extract values from surrounding areas, often using positive lookahead ($[A(? = B)]$, find expression A where expression B follows), and lookbehind

([(? <= $B$)$A$], find expression A where expression B precedes) syntaxes.

For example, the context was used to create a pattern for the attribute indication, $A_{ind}$. Based on a specific word combination that was always followed by the required value, $v_{ind}$, the value could be extracted easily. The sentence after the specific word combination was extracted by the following pattern: $R_{ind}$ = `Geregistreerde indicatie.?([^.]*\.[^.]*)`, in which `([^.]*\.[^.]*)` extracts the sentence.

A second example of a context based pattern is the value extraction $v_i$ for the attributes generic name $A_{gn}$ and brand name $A_{bn}$. Multiple patterns with different word combinations were used. The generic name was always followed by the brand name; this information was used to extract both names at the same time. The text between the multiple word combinations was extracted in order to succesfully extract the required values. One of the three used patterns is the following: $R_{dn1}$ = `(?<=Farmacotherapeutisch rapport\s).*(?=\sbij de indicatie|bij de behandeling|voor de behandeling)`, in which [*], means one or more and [|] represents a 'or' operator.

*(3) Exact matching based regular expressions*
Exact matching, in which the pattern is the exact match of the values $v_i$ to extract, can be used to identify and extract the required value.

An example of exact matching is the pattern created for the attribute therapeutic value outcome $A_{tv}$. To extract the therapeutic value outcome $v_{tv}$ a pattern was created based on synonyms of the possible outcomes: $R_{tv1}$ = `([^.]*\smeerwaarde.?\s[^.]*|[^.]*\sminderwaarde.?\s[^.]* |[^.]*\sgelijke waarde.?\s[^.]*|[^.]*\sgelijke therapeutische.?\s[^.]*)`. The sentence around the outcome was extracted with the syntax: `([^.]*)`.

Table 4 gives an overview of all used patterns, except for the patterns that were used for section extraction (Appendix C).

### 4.6.5 Rule development

To correctly extract the required values $v$, rules were created (Fig. 1(E) and 2(E)). The majority of the rules were based on 'If...Else' statements. For some attributes, multiple patterns were used to extract information. To use the multiple patterns in a specific order, 'If...Else' statements were used. The first pattern was identified first, if no result (no match) was found, then the second pattern was identified, and so on. As a result, multiple patterns could be used for one attribute $A_i$ to extract the corresponding $v_i$ in a document $D_i$.

*Indexing*
For some attributes $A_n$ the first or last match of the pattern in the documents had to be extracted. For this, 'If...Else' statement in combination with indexing

[n] was used to select the first or last match in the documents. If there was only one match, the information of that match was extracted.

*Negation*
To take into account negation, patterns were created to match negated information. For example, the attribute orphan status $A_{os}$ was extracted based on exact matching. To exclude matches with negating words, a negated pattern was used. Again, 'If...Else' statements were used to output the correct value for the orphan status based on the orphan status pattern and negated pattern. If the negated orphan patterns was matched in the text, the output was 'no'. If the orphan pattern is matched in the text, the output is 'yes'. And, if no matches are found, then the orphan status output is 'no'. Listing 1 shows a general example of rule based extraction in which a negated pattern is used.

Listing 1: Example of rule based extraction. In this example a pattern and a corresponding negated pattern are used to extract the correct values. 'If...Else' statement is used to output the correct value corresponding to the found pattern.

```
pattern1 = positive pattern
pattern2 = negated pattern
match1 = findall(pattern1, str)
match2 = findall(pattern2, str)
if len(match2) !=0:      #match with negated pattern
    status = 'no'
elif len(match1) !=0:    #match with positive pattern
    status = 'yes'
else:                    #no match at all
    status = 'no'
```

*Multiple outcomes therapeutic value*
For the therapeutic value one or more outcomes were possible to extract. For different indications of the drug, different outcomes were possible. Therefore, all possible outcomes were extracted, with the main outcome being extracted first.

## 4.7 Gold standard

The information extracted by the developed IE method was compared with manual review findings (gold standard) for each drug. The manual review was performed before the start of this thesis. This was done by researchers, which were domain experts, who evaluated the HTA reports by manually extracting the desired information from the reports. As manually extracting information is prone to error, the extraction was duplicated by a second researcher.

| Attribute | Dutch pattern | English pattern | Explanation of the components of the pattern | Explanation of the pattern |
|---|---|---|---|---|
| Generic name | $R_{dn1}$ = (?<=Farmacotherapeutisch rapport\s).* (?=\sbij de indicatie\|bij de behandeling\|voor de behandeling) | (?<=Farmacotherapeutic report\s).*(?=\sfor the indication\|for the treatment\|to treat) | [?<=B] = A find expression A where expression B precedes. [.*] = any character except newline 0 or more times. [A?=B] = find expression A where expression B follows. | Find 'generic name' where expression 'farmacotherapeutic report' precedes. Find 'generic name' where 'for the treatment' follows. Extract everything in between. |
| | $R_{dn2}$ = (?<=Budget impact analyse van\s).* (?=\sbij de indicatie\|bij de behandeling\|voor) | (?<=Budget impact analysis of\s).*(?=\sfor the indication\|for the treatment) | [?<=B] = A find expression A where expression B precedes. [.*] = any character except newline 0 or more times. [A?=B] = find expression A where expression B follows. | |
| | $R_{dn3}$ = rapport \d\d/?\d\d\s([a-zA-Z]+) | report\d\d/?\d\d\s([a-zA-Z]+) | [\d] = one digit. [/?]=0 or 1 backslash. [\s] = one whitespace. [a-zA-Z]+ = one to infinite characters. | Find 'report' followed by two digits a possible backslash and another two digits, followed by one whitespace. Extract on to infinite charachters except newline. |
| Brand name* | $R_{dn1,} R_{dn2,} R_{dn1}$ | | | |
| Date of HTA report | $R_{d}$ = (\d\d?\s[a-zA-Z0-9]+\s\d{4}) | (\d\d?\s[a-zA-Z0-9]+\s\d{4}) | [\d\d?]= one or two digits. [\s] = one whitespace. [a-zA-Z0-9]+ = one to infinite characters or digits. [\d{4}] = exactly 4 digits. | Find pattern with one or two digits followed by one or more characters or digits, and ending with 4 digits. Extract the complete pattern. |
| Indication | $R_{ind}$ = Geregistreerde indicatie.?([^.]*\.[^.]*) | Registered indication.?([^.]*\.[^.]*) | [.?] = zero or 1 dots. [^.]* = zero or more non-dots. [\.] = escapes dot. | Extract sentence after expression 'registered indication' followed by zero or one dot. |
| Orphan status | $R_{os1}$ = 'weesgeneesmiddel\|wees geneesmiddel' | orphandrug\|orphan drug" | [\|] = or. | Find and extract match of expression 'orphandrug' or 'orphan drug'. |
| | $R_{os2}$ = 'geen weesgeneesmiddel\|geen wees geneesmiddel' | no orphandrug\|no orphan drug' | [\|] = or. | Find and extract match of expression 'no orphandrug' or 'no orphan drug'. |
| Trial name | $R_{tn}$ = [A-Z]+[\d@]+[\w@]+[\-]+[A-Z]+[\d@]*\|[A-Z]+[\d@]+[\w@]+[\-]+[\d@]*\|[A-Z]+[\d@]+[\w@]* | [A-Z]+[\d@]+[\w@]+[\-]+[A-Z]+[\d@]*\|[A-Z]+[\d@]+[\w@]+[\-]+[\d@]*\|[A-Z]+[\d@]+[\w@]* | [A-Z]+ = one or more characters in capital. [\d]+ = one or more digits. [@] = delimiter. [\w]+ = one or more words. [\-]+ = escape the special character dash. [*] = 0 or more. | Find and extract every possible combination of capital letters and digits with possible dash. |
| Therapeutic value | $R_{tv1}$ = ([^.]*\smeerwaarde.?\s[^.]*\|[^.]*\sminderwaarde.?\s[^.]*\|[^.]*\sgelijke waarde.?\s[^.]*\|[^.]*\sgelijke therapeutische.?\s[^.]*) | ([^.]*\sadded value.?\s[^.]*\|[^.]*\sless value.?\s[^.]*\|[^.]*\sequal value.?\s[^.]*\|[^.]*\sequal therapeutic.?\s[^.]*) | [^.]* = zero or more non-dots. [\s] = one white space. [.?] = zero or 1 dots. [^.]* = zero or more non-dots. [\|] = or. | Find expression 'added value' with a white space preceeding and following and with or without a dot following. Extract sentence preceeding and extract sentence following the expression and extract expression itself. |

Table 4: Overview of used patterns per attribute.

## 4.8 Evaluation

For our evaluation, we compare the developed IE method to the manual report analysis by the researchers (the gold standard). The evaluation was performed manually. Extracted information by the IE method was compared to the gold standard. True positives, true negatives, false positives and false negatives were manually reviewed. True positives were defined as correctly extracted information, compared to the gold standard. Exact matching was required for considering true positives for the attributes generic name, $A_{gn}$, brand name, $A_{bn}$, date HTA report, $A_{date}$, orphan status, $A_{os}$ and trial name, $A_{tn}$. For the other two attributes, indication, $A_{ind}$, and therapeutic value, $A_{tv}$, inexact matching was required for true positives. The required matching for inexact matching was discussed with the domain expert. True negatives were defined as correct not extracted information, when compared to the gold standard. False positives were defined as incorrectly extracted information and false negatives were defined as incorrectly not extracted information.

# 5 Results

For the 22 HTA reports of conditionally approved drugs, the proposed method was able to automatically extract information for all 7 attributes. Our method extracted information about the generic and brand name of the drug, the indication of the drug, the date of the report, the orphan status of the drug, the trial name and the therapeutic value outcome(s).

Table 5 presents evaluation results of our IE method. Overall the F-score was high, 0.87, calculated over all attributes, which is far above the pre-defined required F-score of 0.3. The method achieved the highest overall score on recall, 0.90, and lowest score on accuracy, 0.81. The lowest overall score was produced for extracting the trial name, with a F-score of 0.63. For all other attributes the method produced a F-score of 0.87 and higher. The method performed best on extracting the orphan status of a drug, with a score of 0.95 or higher on all outcome measurements, and a F-score of 0.98. The method achieved high F-scores for extraction of the generic name and brand name as well, F-scores of 0.95 for both.

Table 5: Performance measurements of the information extraction methods for the list of attributes

| Attributes | Precision | Recall | Accuracy | F-score |
|---|---|---|---|---|
| *Generic name* | 0.95 | 0.95 | 0.91 | 0.95 |
| *Brand name* | 0.95 | 0.95 | 0.91 | 0.95 |
| *Date HTA report* | 0.76 | 1.00 | 0.75 | 0.87 |
| *Indication* | 0.81 | 0.94 | 0.77 | 0.87 |
| *Orphan status* | 0.95 | 1.00 | 0.95 | 0.98 |
| *Trial name* | 0.59 | 0.67 | 0.59 | 0.63 |
| *Therapeutic value* | 0.94 | 0.81 | 0.78 | 0.87 |
| *Overall* | 0.85 | 0.90 | 0.81 | 0.87 |

*Therapeutic value outcomes*

For the therapeutic value more than one outcome was possible, as the outcome of the therapeutic value could be different for different indications of the drug. For twelve reports, the majority, one outcome was extracted, for eight reports two outcomes were extracted, and for four reports three outcomes were extracted. The performance for the main outcome was good, with a F-score of 0.87. However, the performance dropped for the second and third outcome, with F-scores of 0.59 and 0.22 respectively (Table 6).

Table 6: Performance measurements of the information extraction methods for the list of attributes

| Therapeutic value | Precision | Recall | Accuracy | F-score |
|---|---|---|---|---|
| *Main outcome* | 0.94 | 0.81 | 0.78 | 0.87 |
| *Second outcome* | 0.63 | 0.55 | 0.68 | 0.59 |
| *Third outcome* | 0.25 | 0.20 | 0.68 | 0.22 |

# 6 Discussion

Identifying information in free-text is a challenging task. Scalability, dimensionality, and heterogeneity of the unstructured data appear as the main challenges for automatic information extraction (33). The objective of this thesis was to automatically extract information from medicine authorisation and reimbursement reports. Therefore, we developed an unsupervised IE method using regular expressions. The IE method was able to automatically extract information from the reports with good performance: an overall F-score of 0.87 was achieved (Table 5). The development of this method furthers the intention to automatically extract data from medicine authorisation reports to improve health policy research. This is of great value, because of the time consuming manner the reports are currently manually evaluated.

To perform IE, several methods are proposed in the literature. Many systems include NLP and other machine learning methods, developed through supervised learning techniques (19; 20; 21; 34; 35). For these machine learning approaches training is needed, which requires large amounts of training data. In this thesis, (training) data was not available to that extent. This data will also not be available in the (near) future, as there are only around 50 drug approvals a year. Moreover, the pattern based IE method used in this thesis showed some promising results; NLP methods may not necessarily perform better. NLP requires syntactic parsing, recognition of subject, verbs and objects, and named entity recognition. These are advanced techniques that require the use of an ontology. Furthermore, NLP methods have been successfully applied to the analysis of English texts, but advances in other languages have been limited by the lack or poor coverage of resources (36). Nonetheless, if more data come available, the use of NLP techniques should be considered as NLP is the current state-of-the-art method for accurate information extraction (34).

## 6.1 Limitations and recommendations for future research

Overall, our developed IE method showed promising results, with high overall F-scores (Table 5). Nonetheless, some limitations should be acknowledged. Next, we will address these limitations and make recommendations for further research.

The main limitation of this thesis is the generalizability of the IE method; when applied to another domain or reports that are drafted differently, the performance will probably drop. However, addressing the former, the objective of this thesis was to develop an IE method that is able to extract information for specific reimbursement reports, so portability to another domain is of secondary importance. Nevertheless, the portability to newly differently drafted reimburse reports is of key importance. For the majority of the patterns a context based approach and section extraction were used in order to identify the required in-

formation. In newly drafted reports this context and structure could change. As a result the performance of the IE method will probably drop. To mitigate this problem, multiple patterns were developed with a variety of word combinations to identify all required information. Part of the proposed plan for further development would be to extent the list of word combinations to capture the information in reports that are drafted differently.

*Use of a configuration file*
We developed a pattern based IE method using regular expressions. These patterns are manually crafted and required to be pre-defined with the the help of a domain expert, which is time consuming. Domain experts have knowledge about the relations and dependence's of the information to extract. However, it is challenging for them to express this in form of patterns and translate it to the algorithm. Using a configuration file could be used to address this problem. We did not use a configuration file in this thesis, as the algorithm used for the IE method was hard-coded. However, the use of a configuration file may improve the IE method considerably. To accomplish this, the rules created in the method should not be embedded, but should be outside of the program. A configuration file defines the parameters, options, settings and preferences applied to operating systems. The application or system opens and reads the configuration file at the start, when it parses and applies each setting. The configuration options may be changed while the algorithm stays the same. Because of this, a domain expert can change settings of the configuration file, without having to change the algorithm code itself. As the information to extract is required to be pre-defined by a domain expert, using a configuration file would be of great value.

*Lack of data*
Another limitation of this thesis is the lack of data. Only 22 reports were used for evaluation of the developed IE method, as a result the obtained results may be unreliable. Furthermore, only seven attributes were investigated during this thesis. For the complete evaluation of a report a much greater number of attributes need to be extracted. The number of attributes and amount of validation data should be extended in future research to improve the power of the performance of the IE method. Moreover, the IE method was developed based on 5 sample reports from the total set of 22. The IE method could therefore be overfitted on the small sample size. However, the evaluation results show us that the method performed well on new data.

*Information extraction from tables and position*
Information extraction based on specific position (e.g. page 3 line 10) of information could not be performed. The variable nature of the reports made this task difficult; the reports differed in headers, section, order of sections, content, length and table formats. Therefore, we were not able to extract information based on position. In the future the reports will probably remain unstructured. As a result, extracting information based on position will remain a difficult task.

Nevertheless, identifying specific sections and patterns in which information is mentioned could help making extracting information based on position in future research feasible. For example, the first pages of the reports include the letter to the Minister of Health of the Netherlands. This information may be used to extract text based on position.

Table identification was the first step of the developed IE method for information extraction. However, the desired result of table identification in order to extract information was not achieved. As a result information extraction from tables was not used. The different table layouts, the unstructured text (table titles, footnotes, etc.) and the nested structure of tables in the reports added to this challenging task. Alternative methods from the methods used in this thesis (tabula and camelot) for table extraction should be explored in order to use the tables for information extraction in further research. Rastan and collegues (37) propose a table extraction method system which takes a PDF file and automatically generates a set of XML and CSV files containing the extracted cells, rows and columns of tables, as well as a complete reading order analysis of the tables. This extraction method could be a sound method for table extraction from the data used in this thesis. Although PDF documents are overlooked in the table extraction field (38), table extraction as part of IE has been a continuously growing field, and novel methods will probably come available in the near future.

## 6.2 Discussing results

The developed IE method performed well overall, with a F-score of 0.87 (Table 5), which was far above the pre-defined objective F-score of 0.3. In retrospect, the standard was set too low with the objective F-score of 0.3. Comparing to F-scores reported in literature (35; 39; 40; 41), an overall objective F-score of 0.9 would have been more appropriate.

For all attributes a score of almost 0.9 or higher was achieved, except for the extraction of the trial name. The method under-performed for this attribute, with a F-score of 0.59 (Table 5). Thus, extracting trial names from the reports turned out to be a difficult task. The unstructured way trial names where mentioned in the reports, in combination with the heterogeneity in the trial names, made this a difficult task. Most trial names consisted of a combination of capital letters and numbers. Yet, dashes and lowercase letters were also used in some cases. For this reason, complex regular expressions were used to extract the trial names. For other attributes less complex regular expressions were used, as the corresponding values for these attributes were more straightforward to identify.
Because tables were not used for information extraction, the trial names listed in the tables could not be extracted. Therefore, the IE method on trial name extraction could be considerably improved if tables were used for information extraction.

The orphan status was mentioned in a consistent manner in the reports. Therefore, the orphan status could be identified easily with a simple regular expression pattern. This resulted in the highest F-score measured: 0.98 (Table 5). The orphan status was correctly extracted for all reimbursement reports, except one. For extracting the orphan status, a negation pattern was used to prevent false positives. Although the negation pattern performed well on the included reports, the list of words that negate 'orphan status' could be extended to capture all negations in future reports. Creating negating patterns requires iterative development of these for each attribute (that requires a negated pattern), as the negated patterns are attribute specific. Nonetheless, the time it requires to create a negated pattern is minimal and definitely less time-consuming in comparison with manually extraction.

Although document sections were extracted, complete sections could not be identified accurately. Sections were extracted by identifying the corresponding header. The start of a section could be identified accurately using header identification, however the end of a section could not. As a result, parts of other sections were extracted as well; which resulted in false positive results. This was one of the main reasons for the low performance of the IE method on extracting multiple outcome for the therapeutic value, with F-scores of 0.59 and 0.22 for multiple outcomes (Table 6). Although the results are far from perfect, false positives are easier to deal with than false negatives. False positives can be corrected by manually reviewing the results. On the contrary, false negatives, if unaware of missing data, cannot be corrected easily. We propose to examine the structure of sections carefully in the future to identify possible section limits. The end of the section may be identified by the start of a subsequent section. For example, numbers, bold or italic text, capital letters, a combination of numbers and letters could be identified as the start of a next section. An alternative method, would be to (partially) review the cases manually. Although this is time-consuming, overall the IE method still will save a great amount of time in comparison with the current manual review.

No ethical implications or considerations are in order for this research.

# 7 Conclusion

In this thesis we introduced an IE method for the automatic extraction of information from medicines authorisation and reimbursement reports. We showed that with the developed method automatic information extraction is achieved. The obtained results show that the developed system performs well, with overall high F-scores. As far as we know, this is the first time automatic information extraction on this particular domain is performed. We described how the IE method is applied on medicines authorisation and reimbursement reports, which are heterogeneous reports of textual nature. We evaluated the precision and recall of our approach using 22 reports. We conclude that with our IE method automatically extracting information from medicines authorisation and reimbursement reports can be achieved with good performance. However, using the developed IE method to replace manually evaluation would require several improvements, including a more structured approach of information extraction. Moreover, an approach where the created rules are not embedded inside the code would be advised. In the future, we hope improvement of the current IE method is achieved to ultimately replace the current time consuming manually evaluation of medicine authorisation and reimbursement reports. And as a result, improve health policy research.

# References


[1] J. Piskorski and R. Yangarber, "Information extraction: Past, present and future," in *Multi-source, multilingual information extraction and summarization*, pp. 23–49, Springer, 2013.

[2] "European medicines agency." https://www.ema.europa.eu/en. Accessed: 2021-06-03.

[3] R. A. Vreman, H. Naci, W. G. Goettsch, A. K. Mantel-Teeuwisse, S. G. Schneeweiss, H. G. Leufkens, and A. S. Kesselheim, "Decision making under uncertainty: comparing regulatory and health technology assessment reviews of medicines in the united states and europe," *Clinical Pharmacology & Therapeutics*, vol. 108, no. 2, pp. 350–357, 2020.

[4] G. Tafuri, P. Stolk, F. Trotta, M. Putzeist, H. Leufkens, R. Laing, and M. De Allegri, "How do the ema and fda decide which anticancer drugs make it to the market? a comparative qualitative study on decision makers' views," *Annals of oncology*, vol. 25, no. 1, pp. 265–269, 2014.

[5] R. A. Vreman, L. T. Bloem, S. Van Oirschot, J. Hoekman, M. E. Van Der Elst, H. G. Leufkens, O. H. Klungel, W. G. Goettsch, and A. K. Mantel-Teeuwisse, "The role of regulator-imposed post-approval studies in health technology assessments for conditionally approved drugs," *International Journal of Health Policy and Management*, 2020.

[6] W. H. Organization *et al.*, *Access to new medicines in Europe: technical review of policy initiatives and opportunities for collaboration and research*. WHO Regional Office for Europe, 2015.

[7] J. Goyvaerts and S. Levithan, *Regular expressions cookbook*. O'reilly, 2012.

[8] J. E. Friedl, *Mastering regular expressions*. " O'Reilly Media, Inc.", 2006.

[9] M. Fitzgerald, *Introducing Regular Expressions: Unraveling Regular Expressions, Step-by-Step*. O'Reilly Media, 2012.

[10] S. Sohn, C. Clark, S. R. Halgrim, S. P. Murphy, C. G. Chute, and H. Liu, "Medxn: an open source medication extraction and normalization tool for clinical text," *Journal of the American Medical Informatics Association*, vol. 21, no. 5, pp. 858–865, 2014.

[11] W. B. A. Karaa and N. Dey, *Mining multimedia documents*. CRC Press, 2017.

[12] G. Popovski, S. Kochev, B. Korousic-Seljak, and T. Eftimov, "Foodie: A rule-based named-entity recognition method for food information extraction.," in *ICPRAM*, pp. 915–922, 2019.

[13] E. Soysal, I. Cicekli, and N. Baykal, "Design and evaluation of an ontology based information extraction system for radiological reports," *Computers in Biology and Medicine*, vol. 40, no. 11-12, pp. 900–911, 2010.

[14] K. Sapkota, A. Aldea, M. Younas, D. A. Duce, and R. Banares-Alcantara, "Extracting meaningful entities from regulatory text: Towards automating regulatory compliance," in *2012 Fifth IEEE International Workshop on Requirements Engineering and Law (RELAW)*, pp. 29–32, IEEE, 2012.

[15] A. Hogenboom, F. Hogenboom, F. Frasincar, K. Schouten, and O. Van Der Meer, "Semantics-based information extraction for detecting economic events," *Multimedia Tools and Applications*, vol. 64, no. 1, pp. 27–52, 2013.

[16] S. Patwardhan, *Widening the field of view of information extraction through sentential event recognition*. PhD thesis, Citeseer, 2010.

[17] X. Ling and D. Weld, "Fine-grained entity recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, 2012.

[18] E. Cambria and B. White, "Jumping nlp curves: A review of natural language processing research," *IEEE Computational intelligence magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[19] S. R. Joseph, H. Hlomani, K. Letsholo, F. Kaniwa, and K. Sedimo, "Natural language processing: A review," *Natural Language Processing: A Review*, vol. 6, pp. 207–210, 2016.

[20] A. Kao and S. R. Poteet, *Natural language processing and text mining*. Springer Science & Business Media, 2007.

[21] J. Lee, J.-S. Yi, and J. Son, "Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based nlp," *Journal of Computing in Civil Engineering*, vol. 33, no. 3, p. 04019003, 2019.

[22] D. Tikk and I. Solt, "Improving textual medication extraction using combined conditional random fields and rule-based systems," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 540–544, 2010.

[23] M. Vazquez, M. Krallinger, F. Leitner, and A. Valencia, "Text mining for drugs and chemical compounds: methods, tools and applications," *Molecular Informatics*, vol. 30, no. 6-7, pp. 506–519, 2011.

[24] A. Gupta, I. Banerjee, and D. L. Rubin, "Automatic information extraction from unstructured mammography reports using distributed semantics," *Journal of biomedical informatics*, vol. 78, pp. 78–86, 2018.

[25] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain, "Named entity recognition and normalization applied to large-scale information extraction from the materials science literature," *Journal of chemical information and modeling*, vol. 59, no. 9, pp. 3692–3702, 2019.

[26] Y. Li, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. Jagadish, "Regular expression learning for information extraction," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 21–30, 2008.

[27] J. Liu, R. Bai, Z. Lu, P. Ge, U. Aickelin, and D. Liu, "Data-driven regular expressions evolution for medical text classification using genetic programming," in *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, IEEE, 2020.

[28] F. Brauer, R. Rieger, A. Mocan, and W. M. Barczynski, "Enabling information extraction by inference of regular expressions from sample entities," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1285–1294, 2011.

[29] K. Murthy, P. Deepak, and P. M. Deshpande, "Improving recall of regular expressions for information extraction," in *International Conference on Web Information Systems Engineering*, pp. 455–467, Springer, 2012.

[30] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries," *Journal of biomedical informatics*, vol. 34, no. 5, pp. 301–310, 2001.

[31] F. B. Maguire, C. R. Morris, A. Parikh-Patel, R. D. Cress, T. H. Keegan, C.-S. Li, P. S. Lin, and K. W. Kizer, "A text-mining approach to obtain detailed treatment information from free-text fields in population-based cancer registries: A study of non-small cell lung cancer in california," *PloS one*, vol. 14, no. 2, p. e0212454, 2019.

[32] J. H. Garvin, S. L. DuVall, B. R. South, B. E. Bray, D. Bolton, J. Heavirland, S. Pickard, P. Heidenreich, S. Shen, C. Weir, *et al.*, "Automated extraction of ejection fraction for quality measurement using regular expressions in unstructured information management architecture (uima) for heart failure," *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 859–866, 2012.

[33] K. Adnan and R. Akbar, "Limitations of information extraction methods and techniques for heterogeneous unstructured big data," *International Journal of Engineering Business Management*, vol. 11, p. 1847979019890771, 2019.

[34] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named entity recognition and relation extraction: State-of-the-art," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–39, 2021.

[35] J. Zhang and N. M. El-Gohary, "Semantic nlp-based information extraction from construction regulatory documents for automated compliance checking," *Journal of Computing in Civil Engineering*, vol. 30, no. 2, p. 04015014, 2016.

[36] N. Viani, C. Larizza, V. Tibollo, C. Napolitano, S. G. Priori, R. Bellazzi, and L. Sacchi, "Information extraction from italian medical reports: An ontology-driven approach," *International journal of medical informatics*, vol. 111, pp. 140–148, 2018.

[37] R. Rastan, H.-Y. Paik, J. Shepherd, S. H. Ryu, and A. Beheshti, "Texus: table extraction system for pdf documents," in *Australasian Database Conference*, pp. 345–349, Springer, 2018.

[38] Y. Liu, K. Bai, P. Mitra, and C. L. Giles, "Tableseer: automatic table metadata extraction and searching in digital libraries," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 91–100, 2007.

[39] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee, "Twiner: named entity recognition in targeted twitter stream," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 721–730, 2012.

[40] L. Bing, W. Lam, and T.-L. Wong, "Wikipedia entity expansion and attribute extraction from the web using semi-supervised learning," in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 567–576, 2013.

[41] A. Sun, R. Grishman, and S. Sekine, "Semi-supervised relation extraction with large-scale word clustering," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 521–529, 2011.

# Appendices

## A Python code

```
#!/usr/bin/env python
# coding: utf-8

# In[1]:


#import packages
import json
import pandas as pd
import regex as re
import os
import glob
import re


# In[2]:


#creat function for importing PDFs and convert to string
from pdfminer.pdfinterp import PDFResourceManager, PDFPageInterpreter
from pdfminer.converter import TextConverter
from pdfminer.layout import LAParams
from pdfminer.pdfpage import PDFPage
from io import StringIO

def extract_pdf_content(pdf):
    rsrcmgr = PDFResourceManager()
    codec = 'utf-8'
    outfp = StringIO()
    laparams = LAParams()
    device = TextConverter(rsrcmgr=rsrcmgr, outfp=outfp, codec=codec, laparams=la
    with open(pdf, 'rb') as fp:
        interpreter = PDFPageInterpreter(rsrcmgr, device)
        password = ""
        maxpages = 0
        caching = True
        pagenos=set()
        for page in PDFPage.get_pages(fp, pagenos, maxpages=maxpages, password=p
            interpreter.process_page(page)
    mystr = outfp.getvalue()
```

```
    device.close()
    outfp.close()
    return mystr


# In[3]:


#path of files
pdf_path = "C:/Users/ingeg/Documents/ADS_Thesis_Project/Dataset/ZIN/ZIN/PDF_file
pdfs = glob.glob("{}/*.pdf".format(pdf_path))


# In[4]:


#all files put together
zin = []
for i in range(22):
    content = extract_pdf_content(pdfs[i])
    zin.append(content)


# In[5]:


#check type
type(zin[1])


# In[6]:


# Function to convert list to String (needed further ahead)
def listToString(s):

    # initialize an empty string
    str1 = ""

    # traverse in the string
    for ele in s:
        str1 += ele

    # return string
    return str1
```

```
# In[7]:


def info_extraction(pdf):

    df = pd.DataFrame()

    Drugname = []
    Brandname = []
    HTA_date = []
    indication = []
    orphan = []
    REA_outcome = []
    trial = []

    for text in pdf:

        #Drugname
        medname_pattern = '(?<=Farmacotherapeutisch rapport\s).*(?=\sbij de indi
        medname_pattern2 = '(?<=Budget impact analyse van\s).*(?=\sbij de indica
        medname_pattern3 = 'rapport \d\d/?\d\d\s([a-zA-Z]+)'
        medname_pattern4 = 'rapport \d\d/?\d\d\s([^.]*)'
        medname_pattern5 = '(?<=\().*?(?=\))'
        medname1 = re.findall(medname_pattern, text)
        medname2 = re.findall(medname_pattern2, text)
        medname3 = re.findall(medname_pattern3, text)

        if len(medname1) != 0:
            medname_c = medname1[0]
            medname_c = medname_c.split(" ")
            drug_name = medname_c[0]
            brand_name = medname_c[1]
        elif len(medname2) !=0:
            medname_c = medname2[0]
            medname_c = medname_c.split(" ")
            drug_name = medname_c[0]
            brand_name = medname_c[1]
        elif  len(medname3) !=0:
            drug_name = medname3[0]
            medname4 = re.findall(medname_pattern4, text)[0]
            medname5 = re.findall(medname_pattern5, medname4)[0]
            brand_name = medname5
        else:
            drug_name = ''
            brand_name = ''
```

```
Drugname.append(drug_name)
Brandname.append(brand_name)

#Indication
ind_pattern = "Geregistreerde indicatie.?([^.]*\.[^.]*)"
text = text.replace('\n','')
ind = re.findall(ind_pattern, text)
if len(ind) != 0:
    ind2 = ind[-1]
else:
    ind2 = ind
indication.append(ind2)

#Date first HTA
date_pattern = "(\d\d?\s[a-zA-Z0-9]+\s\d{4})"
htadate = re.findall(date_pattern, text)[0]
HTA_date.append(htadate)

#Orphan status
orph_pattern = 'weesgeneesmiddel|wees geneesmiddel'
orph_not_pattern = 'geen weesgeneesmiddel|geen wees geneesmiddel'
orph_search = re.findall(orph_pattern, text)
orph_not_search = re.findall(orph_not_pattern, text)
if len(orph_not_search) !=0:
    orph = 'no'
elif len(orph_search) !=0:
    orph = 'yes'
else:
    orph = 'no'
orphan.append(orph)

#REA Outcome
concl_pattern = "(?i)conclusie therapeutische waarde([^.]*\.([^.]*\.){20
concl_pattern2 = "(?i)eindconclusie([^.]*\.([^.]*\.){20})"
conclusion = re.findall(concl_pattern, text)
conclusion_2 = re.findall(concl_pattern2, text)
if len(conclusion) != 0:
    conclusion2 = conclusion[-1]
elif len(conclusion_2) !=0:
    conclusion2 = conclusion_2[-1]
else:
    conclusion2 = conclusion
conclusion3 = listToString(conclusion2)
val_pattern = "([^.]*\smeerwaarde.?\s[^.]*|[^.]*\sminderwaarde.?\s[^.]*|
ther_val = re.findall(val_pattern, conclusion3)[:3]
REA_outcome.append(ther_val)
```

```
        #Trial name
        study_pattern = "(?i)(ge ncludeerde studies)([^.]*\.([^.]*\.){30})"
        study_pattern2 = "(?i)(werkzaamheid)([^.]*\.([^.]*\.){30})"
        study_pattern3 = "(?i)(evidentie)([^.]*\.([^.]*\.){30})"
        study_pattern4 = "(?i)(placebogecontroleerde studies)([^.]*\.([^.]*\.){30
        study = re.findall(study_pattern, text)
        study2 = re.findall(study_pattern2, text)
        study3 = re.findall(study_pattern3, text)
        study4 = re.findall(study_pattern3, text)
        if len(study) != 0:
            study2 = study[-1]
        elif len(study2) !=0:
            study2 = study2[-1]
        elif len(study3) !=0:
            study2 = study3[0]
        elif len(study4) !=0:
            study2 = study4[-1]
        else:
            study2 = study
        study3 = listToString(study2)
        study3 = study3.replace('CD4', '').replace('FEV 1', '').replace('FEV1',
        trial_pattern = "[A-Z]+[\d@]+[\w@]+[\-]+[A-Z]+[\d@]*|[A-Z]+[\d@]+[\w@]+[
        trials = re.findall(trial_pattern, study3)[:1]
        trial.append(trials)

    df['Drugname'] = Drugname
    df['Brandname'] = Brandname
    df['Date first HTA'] = HTA_date
    df['HTA indication assessed'] = indication
    df['Orphan status'] = orphan
    df['REA_Outcome'] = REA_outcome
    df['Trial names'] = trial


    return df


# In[8]:


#Extract the information for all documents
info_extraction(zin)


# In[9]:
```

```
#rename to df
df = info_extraction(zin)


# In[10]:


#export to csv and JSON file (df with one outcome for therapeutic value)
df.to_csv('df_ZIN_1.csv')
df.to_json('df_ZIN_1.json')


# In[11]:


#split outcomes into new columns to split the multiple therapeutic value outcome
df2 = pd.DataFrame([pd.Series(x) for x in df.REA_Outcome])
# print(df2)
df2.columns = ['REA_Outcome_{}'.format(x+1) for x in df2.columns]
print(df2.columns)


# In[12]:


#concat the two different dataframes
df_ZIN = pd.concat([df,df2], axis=1)
df_ZIN


# In[13]:


#drop Outcome column (=duplicate)
df_ZIN = df_ZIN.drop(['REA_Outcome'], axis=1)


# In[14]:


#show result
df_ZIN.head()
```

```
# In[15]:


#export to csv and JSON file (df with multiple outcomes for therapeutic value)
df_ZIN.to_csv('df_ZIN_2.csv')
df_ZIN.to_json('df_ZIN_2.JSON')
```

# B Extracted information

## B.1 Extracted information, output: JSON file

{"Drugname": {"0": "ataluren", "1": "aztreonam", "2": "", "3": "darunavir", "4": "etravirine", "5": "everolimus", "6": "fampridine", "7": "asfotase", "8": "", "9": "lapatinib", "10": "lomitapide", "11": "obeticholzuur", "12": "ofatumumab", "13": "axi-cel", "14": "tisagenlecleucel", "15": "tisagenlecleucel", "16": "panitumumab", "17": "pazopanib", "18": "raltegravir", "19": "Stiripentol", "20": "sunitinib", "21": "vandetanib"}, "Brandname": {"0": "(Translarna\u00ae)", "1": "(Cayston\u00ae)", "2": "", "3": "Prezista\u00ae", "4": "(Intelence\u00ae)", "5": "(Votubia\u00ae)", "6": "(Fampyra\u00ae)", "7": "alfa", "8": "", "9": "(Tyverb\u00ae)", "10": "Lojuxta", "11": "(Ocaliva\u00ae)", "12": "(Arzerra\u00ae)", "13": "(Yescarta\u00ae)", "14": "(Kymriah\u00ae)", "15": "(Kymriah\u00ae)", "16": "(Vectibix\u00ae)", "17": "(Votrient\u00ae)", "18": "(Isentress\u00ae)", "19": "(Diacomit\u00ae)", "20": "(Sutent\u00ae)", "21": "(Caprelsa\u00ae)"}, "Date first HTA": {"0": "23 november 2017", "1": "26 april 2011", "2": "28 november 2017", "3": "10 april 2007", "4": "1 december 2008", "5": "10 februari 2012", "6": "21 januari 2013", "7": "25 maart 2019", "8": "14 mei 2019", "9": "11 augustus 2008", "10": "22 april 2015", "11": "18 juli 2018", "12": "7 juni 2011", "13": "57 maart 2019", "14": "18 december 2018", "15": "07 maart 2019", "16": "11 februari 2007", "17": "14 januari 2011", "18": "5 februari 2008", "19": "3 maart 2008", "20": "26 oktober 2006", "21": "10 september 2012"}, "HTA indication assessed": {"0": "Ataluren is geregistreerd voor \u201cSpierdystrofie van Duchenne als gevolg van een nonsense mutatie in het dystrofine-gen, bij ambulante pati\u00ebnten van vijf jaar en ouder. De werkzaamheid is niet aangetoond bij niet-ambulante pati\u00ebnten", "1": " \u201cSuppressieve behandeling van chronische longinfecties veroorzaakt door Pseudomonas aeruginosa bij pati\u00ebnten vanaf 18 jaar met cystische fibrose.\u201d Dosering", "2": [], "3": "In combinatie met 100 mg ritonavir en andere antiretrovirale geneesmiddelen voor de behandeling van hiv-1 infectie bij sterk voorbehandelde volwassen pati\u00ebnten bij wie meer dan \u00e9\u00e9n antiretrovirale behandeling met een proteaseremmer heeft gefaald. Dosering Darunavir 600 mg, tweemaal daags, in te nemen samen met 100 mg ritonavir, tweemaal daags, en met voedsel", "4": "\u201cIntelence, in combinatie met een gebooste proteaseremmer en andere antiretrovirale geneesmiddelen, is aangewezen voor de behandeling van een infectie met het humaan immunodefici\u00ebntievirus type 1 (hiv-1) bij met antiretrovirale geneesmiddelen voorbehandelde volwassen pati\u00ebnten. Deze indicatie is gebaseerd op de analyses na 24 weken in 2 gerandomiseerde, dubbelblinde, placebogecontroleerde fase III-studies bij sterk voorbehandelde pati\u00ebnten met virusstammen waarin mutaties voorkwamen voor resistentie tegen niet-nucleoside reverse transcriptaseremmers en tegen proteaseremmers, waarin Intelence werd onderzocht in combinatie met een \u2018optimised background regimen\u2019 (OBR) met inbegrip van darunavir/ritonavir", "5": " \u201cVoor de behandeling van pati\u00ebnten van 3 jaar en ouder met subependymale reuscel

astrocytomen (SEGA), geassocieerd met Tubereuze Sclerose Complex (TSC), die een therapeutische interventie nodig hebben, maar niet ontvankelijk zijn voor een chirurgische ingreep.\u201d Dosering", "6": " \u201cverbetering van het lopen bij volwassen pati\u00ebnten met multiple sclerosis met beperkt loopvermogen (EDSS 4-7).\u201d Fampridine is door de Europese Registatie Autoriteit (EMA) \"voorwaardelijk\" geregistreerd", "7": "Asfotase alfa (Strensiq\u00ae) is geregistreerd voor 'langdurige enzymvervangingstherapie bij pati\u00ebnten met hypofosfatasia (HPP) bij wie de eerste symptomen voor de leeftijd van 18 jaar zijn opgetreden om de manifestaties van de ziekte met betrekking tot het bot te behandelen\u2018. 1", "8": "Myalepta\u00ae is ge\u00efndiceerd als aanvulling bij een dieet als vervangingstherapie om de complicaties van leptinedefici\u00ebntie te behandelen bij pati\u00ebnten met lipodystrofie: \uf0b7 met bevestigde aangeboren gegeneraliseerde lipodystrofie (Berardinelli-Seip-syndroom) of verworven gegeneraliseerde lipodystrofie (Lawrence-syndroom), bij volwassenen en kinderen van 2 jaar en ouder; \uf0b7 met bevestigde familiaire parti\u00eble lipodystrofie of verworven parti\u00eble lipodystrofie (Barraquer- Simons-syndroom), bij volwassenen en kinderen van 12 jaar en ouder bij wie met standaardbehandelingen geen adequate metabole controle werd bereikt. Bijzonderheid Registratie als weesgeneesmiddel", "9": "In combinatie met capecitabine bij gevorderde ofgemetastaseerde borstkanker bij tumoren met HER2-overexpressie bij progressieve ziekte na eerdere behandeling meteen antracyclinederivaat, een taxaan en trastuzumab.DoseringVolwassenen: 1", "10": "Lomitapide (Lojuxta\u00ae) is ge\u00efndiceerd \u2018als aanvulling bij een vetarm dieet en andere lipidenverlagende geneesmiddelen met of zonder low-density-lipoprote\u00efne-aferese (LDL-aferese) bij volwassen pati\u00ebnten met homozygote familiaire hypercholesterolemie (HoFH). Indien mogelijk, moet genetisch worden bevestigd dat er sprake is van HoFH", "11": "Obeticholzuur (Ocaliva\u00ae) is geregistreerd voor de behandeling van primaire biliaire cholangitis (ook primaire biliaire cirrose genaamd) in combinatie met ursodeoxycholzuur (UDCA) bij volwassenen met een ontoereikende respons op UDCA of als monotherapie bij volwassenen die UDCA niet kunnen verdragen. Primaire biliaire cholangitis (PBC) is een relatief zeldzame chronische leverziekte met kenmerken van een autoimmuun aandoening", "12": " \u201cOfatumumab is ge\u00efndiceerd voor de behandeling van chronische lymfatische leukemie (CLL) bij pati\u00ebnten die refractair zijn voor fludarabine en alemtuzumab.\u201d Dosering", "13": "De kosteneffectiviteitsanalyse moet plaatsvinden bij pati\u00ebnten met de geregistreerde indicatie voor axicabtagene ciloleucel. De geregistreerde indicatie luidt als volgt: \u201cbehandeling van volwassenen pati\u00ebnten met recidiverend of refractair diffuus grootcellig B-cellymfoom (DLBCL) en primair mediastinaal B-cellymfoom (PMBCL), na twee of meer lijnen systemische therapie\u201d", "14": "Tisagenlecleucel (Kymriah\u00ae) is geregistreerd voor \u2018pediatrische en jongvolwassen pati\u00ebnten tot de leeftijd van 25 jaar met refractaire B-cel acute lymfoblastaire leukemie (ALL), of met een recidief na transplantatie of met een tweede of later recidief van B-cel ALL.\u20191 Plaats in het behandelalgoritme ALL is een levensbedreigende ziekte", "15": "Tisagenlecleucel is geregistreerd voor twee therapeutische indicaties: \u2022

Volwassen pati\u00ebnten met een recidief of refractair (r/r) diffuus grootcellig B-cellymfoom na twee of meer lijnen systemische therapie (behandeld in dit rapport). Pediatrische en jongvolwassen pati\u00ebnten tot de leeftijd van 25 jaar met refractaire B-cel acute lymfoblastaire leukemie (ALL), of met een recidief na transplantatie of met een tweede of later recidief van B-cel ALL (niet behandeld in dit rapport)", "16": " De behandeling van pati\u00ebnten met gemetastaseerd colorectaal carcinoom met EGFR expressie die niet KRAS-gemuteerd (wild-type) is, na falen fluoropyrimidine, oxaliplatin- en irinotecan- bevattende chemotherapieregimes. Dosering 6 mg/kg lichaamsgewicht, \u00e9\u00e9nmaal per twee weken", "17": " \u201cVotrient is geregistreerd voor de eerstelijnsbehandeling van gevorderd niercelcarcinoom (RCC, Renal Cell Carcinoma) en voor pati\u00ebnten die eerder een cytokinebehandeling hebben ondergaan voor het gevorderde stadium van de ziekte.\u201d Dosering", "18": "in combinatie met andere antiretrovirale geneesmiddelen: behandeling van volwassenen met HIV1-infectie en aangetoonde HIV1-replicatie, ondanks eerdere en voortdurende antiretrovirale behandeling Dosering 400 mg 2 dd Werkingsmechanisme Remt het enzym integrase. Daardoor kan het virale DNA niet integreren in het DNA van de humane gastheercel en dus ook niet repliceren", "19": " Als adjuvans bij refractaire gegeneraliseerde tonisch-klonische aanvallen bij ernstige juveniele myoklonische epilepsie op zeer jonge leeftijd (syndroom van Dravet) in combinatie met clobazam en valpro\u00efnezuur. Dosering Kinderen: in een periode van 3 dagen dosis opbouwen tot de aanbevolen dosis van 50 mg/kg/dag verdeeld over 2-3 giften Werkingsmechanisme De belangrijkste werking wordt toegeschreven aan een Pagina 1 van 8 \f28002289 definitieve versie stiripentol (Diacomit\u00ae) versterking van de werking van andere anti-epileptica via farmacokinetische interacties die hoofdzakelijk zijn gebaseerd op de metabole remming van CYP3A4 en 2C191", "20": "\u2022 Behandeling van pati\u00ebnten met gevorderd en/of gemetastaseerd niercelcarcinoom na falen van behandeling met interferon-alfa of interleukine-2. \u2022 Behandeling van pati\u00ebnten met niet-operatief te verwijderen en/of gemetastaseerde gastro-intestinale stromaceltumoren (GIST) na falen van behandeling met imatinibmesylaat als gevolg van therapieresistentie of intolerantie", "21": " \u201cCaprelsa is ge\u00efndiceerd voor de behandeling van aggressieve en symptomatische medullaire schildklierkanker (MTC) bij pati\u00ebnten met niet-reseceerbare lokaal gevorderde of gemetastaseerde ziekte. Er moet rekening worden gehouden met een mogelijk kleiner voordeel bij pati\u00ebnten waarvan de RET-mutatie-status (Rearranged during Transfection) niet bekend of negatief is, voordat de beslissing over de individuele behandeling wordt genomen"}, "Orphan status": {"0": "yes", "1": "yes", "2": "no", "3": "no", "4": "no", "5": "yes", "6": "no", "7": "yes", "8": "yes", "9": "no", "10": "no", "11": "yes", "12": "yes", "13": "yes", "14": "yes", "15": "yes", "16": "no", "17": "no", "18": "no", "19": "yes", "20": "yes", "21": "no"}, "Trial names": {"0": ["PTC 1"], "1": [], "2": [], "3": ["POWER 1"], "4": ["V90I"], "5": ["CYP3A4"], "6": ["F202"], "7": ["ENB 0"], "8": [], "9": ["EGF10051"], "10": ["UP1002"], "11": ["CYP2C19-"], "12": [], "13": ["ZUMA-1"], "14": ["ECOG 0"], "15": ["C2201"], "16": [], "17": [], "18": [], "19": [], "20": [], "21": ["V1"]}, "REA_Outcome_1": {"0": "

Ataluren heeft bij de behandeling van ambulante pati\u00ebnten vanaf 5 jaar met nmDMD [in vergelijking met placebo toegevoegd aan best ondersteunende zorg] een therapeutische minderwaarde vanwege onvoldoende bewijs", "1": " aeruginosa-infectie heeft aztreonam-inhalatie een gelijke therapeutische waarde als inhalatie van tobramycine of colistine", "2": null, "3": " Bij de behandeling van een HIV-1 infectie in combinatie met andere anti-retrovirale middelen bij pati\u00ebnten die uitgebreid voorbehandeld zijn en resistentie hebben tegen meerdere proteaseremmers heeft darunavir (in combinatie met ritonavir) een meerwaarde. 2", "4": " Bij de behandeling van hiv-infectie heeft etravirine als toevoeging aan een optimale doch falende combinatie van antiretrovirale middelen een therapeutische meerwaarde boven placebo", "5": " Bij de behandeling van pati\u00ebnten van 3 jaar en ouder met subependymale reuscel astrocytomen (SEGA), geassocieerd met Tubereuze Sclerose Complex (TSC), die een therapeutische interventie nodig hebben, maar niet in aanmerking komen voor een chirurgische ingreep heeft everolimus een therapeutische meerwaarde ten opzichte van best ondersteunende zorg", "6": " Omdat hiervan uit onderzoek geen nadeel is aangetoond, is de conclusie therapeutisch gelijke waarde. Ook ten opzichte van oefentherapie/fysiotherapie, zijn er onvoldoende gegevens en is geen indirecte vergelijking in kwantitatieve zin mogelijk", "7": " Zorginstituut Nederland is tot de eindconclusie gekomen dat bij pati\u00ebnten met hypofosfatasia en een perinatale of een infantiele aanvang, de behandeling met asfotase alfa een therapeutische meerwaarde heeft ten opzichte van beste ondersteunende zorg alleen", "8": " Rekening houdend met de onzekerheden in de gunstige effecten en met de bezorgdheid over de ontwikkeling van neutraliserende antilichamen concluderen Zorginstituut Nederland en de CTG dat het niet mogelijk is om een therapeutische meerwaarde toe te kennen aan metreleptine (Myalepta\u00ae) bij pati\u00ebnten met gegeneraliseerde lipodystrofie en parti\u00eble lipodystrofie", "9": null, "10": " Vanwege de noodzaak de ziekte te behandelen luidt de conclusie, ondanks dbeperkte dataset en de suboptimale studie-opzet, dat lomitapide bij behandeling van pati\u00ebnten \u2265 18 jaar met bevestigde HoFH, toegevoegd aan optimale gebruikelijke lipidenverlagende behandelingen (combinatie van vetarm dieet, medicatie en LDL-aferese), een therapeutische meerwaarde heeft ten opzichte van dMede vanwege het ontbreken van gegevens over de lange termijn veiligheid en effectiviteit is het product door EMA geregistreerd onder \u2018Exceptional Circumstances\u2019", "11": " Er is geen meerwaarde aangetoond van obeticholzuur bij volwassenen PBC pati\u00ebnten met een ontoereikende respons op UDCA of als monotherapie bij volwassenen PBC pati\u00ebnten die UDCA niet kunnen verdragen, t", "12": " Op basis van een retrospectieve vergelijking en gezien de ernst van de ziekte en de afwezigheid van een alternatieve behandeling wordt, ondanks de beperkte gegevens, geconcludeerd dat ofatumumab bij de behandeling van chronische lymfatische leukemie (CLL) bij pati\u00ebnten die refractair zijn voor fludarabine en alemtuzumab een therapeutische meerwaarde heeft ten opzichte van geen ofatumumab behandeling", "13": null, "14": null, "15": " We concluderen dat tisagenlecleucel een therapeutische minderwaarde heeft op basis van onvoldoende bewijs bij volwassen pati\u00ebnten met een recidief

of refractair (r/r) diffuus grootcellig B-cellymfoom na twee of meer lijnen systemische therapie ten opzichte van salvage chemotherapie (+SCT)", "16": " Bij pati\u00ebnten met gemetastaseerde, chemotherapieresistente colorectale kanker bij wie geen mutatie van het KRAS gen kan worden aangetoond heeft panitumumab een therapeutische meerwaarde ten opzichte van de standaardbehandeling met beste ondersteunende zorg", "17": " Bij eerstelijnsbehandeling van pati\u00ebnten met lokaal gevorderd of gemetastaseerd niercelcarcinoom in de gunstige of intermediaire prognosegroep (volgens de indeling van Motzer) heeft pazopanib een gelijke waarde ten opzichte van sunitinib en een gelijke waarde ten opzichte van cytokinetherapie zoals interferon alfa al dan niet in combinatie met bevacizumab, of interleukine-2", "18": " Bij de behandeling van HIV-infectie heeft raltegravir als toevoeging aan een optimale doch falende combinatie van antiretrovirale middelen een therapeutische meerwaarde boven placebo", "19": " Bij de behandeling van refractaire gegeneraliseerde tonisch-klonische aanvallen bij ernstige juveniele myoklonische epilepsie op zeer jonge leeftijd (SMEI/syndroom van Dravet) als adjuvans in combinatie met clobazam en valpro\u00efnezuur heeft stiripentol een therapeutische meerwaarde. 2", "20": " De toepassing van sunitinib heeft een therapeutische meerwaarde bij pati\u00ebnten met gevorderd en/of gemetastaseerd niercelcarcinoom waarbij behandeling met interferon-alfa of interleukine-2 niet succesvol of niet ge\u00efndiceerd is", "21": " Bij de behandeling van agressieve en symptomatische medullaire schildklierkanker (MTC) bij pati\u00ebnten met niet-reseceerbare lokaal gevorderde of gemetastaseerde ziekte heeft vandetanib een therapeutische meerwaarde ten opzichte van placebo"}, "REA_Outcome_2": {"0": null, "1": " De deels gunstige effecten van aztreonam en de afwezigheid van ongunstige effecten zijn onvoldoende voor een therapeutische meerwaarde, omdat er slechts \u00e9\u00e9n, open effectiviteitsonderzoek is en omdat de ervaring met het middel beperkt is", "2": null, "3": null, "4": null, "5": " Claim van de fabrikant en oordeel van de CFH 4a Claim van de fabrikant \u201cBij pati\u00ebnten van 3 jaar en ouder met TSC-geassocieerde SEGA, die niet in aanmerking komen voor een chirurgische resectie, heeft de behandeling met everolimus een therapeutische meerwaarde boven geen behandeling", "6": " Hier is de situatie mede om onderstaande redenen echter geheel anders, waardoor de conclusie therapeutische minderwaarde is", "7": " Bij pati\u00ebnten met hypofosfatasia en een juveniele aanvang heeft de behandeling met asfotase alfa een therapeutische minderwaarde door onvoldoende gegevens", "8": null, "9": null, "10": null, "11": " Hierdoor is het thans te vroeg om te spreken van een aangetoond klinisch relevant voordeel op harde klinische eindpunten door de behandeling van volwassenen PBC pati\u00ebnten met een ontoereikende respons op UDCA of als monotherapie bij volwassenen PBC pati\u00ebnten die UDCA niet kunnen verdragen met obeticholzuur, met andere woorden, dit betekent dat het geneesmiddel een therapeutische minderwaarde heeft", "12": " Uit klinisch onderzoek blijkt dat behandeling met ofatumumab een duidelijke meerwaarde heeft in de behandeling van chronische lymfatische leukemie (CLL) bij pati\u00ebnten die refractair zijn voor: - - fludarabine en alemtuzumab fludarabine en voor wie een behandeling met alemtuzumab niet geschikt wordt

geacht", "13": null, "14": null, "15": null, "16": null, "17": " Na falen van voorafgaande behandeling gebaseerd op interferon-alfa (al dan niet in combinatie met bevacizumab) of interleukine-2 of wanneer deze niet werd verdragen, heeft tweedelijnsbehandeling met pazopanib bij deze pati\u00ebnten eveneens een gelijke waarde ten opzichte van sunitinib", "18": null, "19": null, "20": " Het gebruik van sunitinib heeft op basis van het te verwachten levensverlengende effect een therapeutische meerwaarde bij het merendeel van pati\u00ebnten met GIST die progressie vertonen tijdens behandeling met imatinibmesilaat of intolerant zijn voor deze behandeling", "21": " Bij de behandeling van agressieve en symptomatische medullaire schildklierkanker (MTC) bij pati\u00ebnten met niet-reseceerbare lokaal gevorderde of gemetastaseerde ziekte heeft vandetanib een therapeutische meerwaarde ten opzichte van placebo"}, "REA_Outcome_3": {"0": null, "1": " Aztreonam lysine heeft een therapeutische meerwaarde voor pati\u00ebnten met deze zeldzame en ernstige erfelijke aandoening waardoor opname op bijlage 1B van de Regeling zorgverzekering is aangewezen", "2": null, "3": null, "4": null, "5": "\u201d 4b Oordeel CFH over de claim van de fabrikant Bij de behandeling van pati\u00ebnten van 3 jaar en ouder met subependymale reuscel astrocytomen (SEGA), geassocieerd met Tubereuze Sclerose Complex (TSC), die een therapeutische interventie nodig hebben, maar niet in aanmerking komen voor een chirurgische ingreep heeft everolimus een therapeutische meerwaarde ten opzichte van best ondersteunende zorg", "6": " Eindconclusie: Ter verbetering van het lopen bij volwassen pati\u00ebnten met multiple sclerose (EDSS 4-7) heeft fampridine mga een gelijke therapeutische waarde als fampridine als apotheekbereiding (=4-AP)", "7": null, "8": null, "9": null, "10": null, "11": null, "12": " Om deze redenen wordt, ondanks de beperkte gegevens, toch geconcludeerd dat ofatumumab bij de behandeling van chronische lymfatische leukemie (CLL) bij pati\u00ebnten die refractair zijn voor fludarabine en alemtuzumab een therapeutische meerwaarde heeft ten opzichte van best ondersteunende zorg", "13": null, "14": null, "15": null, "16": null, "17": null, "18": null, "19": null, "20": null, "21": null}}

## B.2 Extracted information, output: dataframe

Table 7: Extracted information from the 22 reports by the developed IE method. Information of attributes 'Indication' and 'Therapeutic value' are not shown in order to create a readable dataframe

| Generic name | Brand name | Date first HTA | Orphan status | Trial name |
|---|---|---|---|---|
| *ataluren* | (Translarna®) | 23 november 2017 | yes | ['PTC 1'] |
| *aztreonam* | (Cayston®) | 26 april 2011 | yes | [] |
| | | 28 november 2017 | no | [] |
| *darunavir* | Prezista® | 10 april 2007 | no | ['POWER 1'] |
| *etravirine* | (Intelence®) | 1 december 2008 | no | ['DUET-1'] |
| *everolimus* | (Votubia®) | 10 februari 2012 | yes | ['C2485'] |
| *fampridine* | (Fampyra®) | 21 januari 2013 | no | ['F202'] |
| *asfotase* | alfa | 25 maart 2019 | yes | ['ENB 0'] |
| | | 14 mei 2019 | yes | [] |
| *lapatinib* | (Tyverb®) | 11 augustus 2008 | no | ['EGF10051'] |
| *lomitapide* | Lojuxta | 22 april 2015 | no | ['UP1002'] |
| *obeticholzuur* | (Ocaliva®) | 18 juli 2018 | yes | ['NCT02308111'] |
| *ofatumumab* | (Arzerra®) | 7 juni 2011 | yes | [] |
| *axi-cel* | (Yescarta®) | 57 maart 2019 | yes | ['ZUMA-1'] |
| *tisagenlecleucel* | (Kymriah®) | 18 december 2018 | yes | ['ECOG 0'] |
| *tisagenlecleucel* | (Kymriah®) | 07 maart 2019 | yes | ['C2201'] |
| *panitumumab* | (Vectibix®) | 11 februari 2007 | no | ['EPAR 1'] |
| *pazopanib* | (Votrient®) | 14 januari 2011 | no | ['VEG105192'] |
| *raltegravir* | (Isentress®) | 5 februari 2008 | no | ['BENCHMRK1'] |
| *Stiripentol* | (Diacomit®) | 3 maart 2008 | yes | [] |
| *sunitinib* | (Sutent®) | 26 oktober 2006 | yes | ['RTKC-0'] |
| *vandetanib* | (Caprelsa®) | 10 september 2012 | no | ['D4200C00058'] |

# C Patterns section identification

Table 8: Dutch patterns for section identification

| Attributes | Dutch pattern |
|---|---|
| *Therapeutic value* | (?i)conclusie therapeutische waarde([.]*\.([.]*\.){20}) |
| | (?i)eindconclusie([.]*\.([.]*\.){20}) |
| *Trial name* | (?i)(geïncludeerde studies)([.]*\.([.]*\.){30}) |
| | (?i)(werkzaamheid)([.]*\.([.]*\.){30}) |
| | (?i)(evidentie)([.]*\.([.]*\.){30}) |
| | (?i)(placebogecontroleerde studies)([.]*\.([.]*\.){30}) |

Table 9: English patterns for section identification

| Attributes | English pattern |
|---|---|
| *Therapeutic value* | (?i)conclusion therapeutic value([.]*\.([.]*\.){20}) |
| | (?i)final conclusion([.]*\.([.]*\.){20}) |
| *Trial name* | (?i)(included studies)([.]*\.([.]*\.){30}) |
| | (?i)(efficacy)([.]*\.([.]*\.){30}) |
| | (?i)(evidence)([.]*\.([.]*\.){30}) |
| | (?i)(placebo controlled studies)([.]*\.([.]*\.){30}) |