applied sciences

MDPI

Article

Information Extraction from German Clinical Care Documents in Context of Alzheimer's Disease

Lisa Langnickel [1,2,†], Kilian Krockauer [3], Mischa Uebachs [4,‡], Sebastian Schaaf [5,†], Sumit Madan [6,7], Thomas Klockgether [8,9] and Juliane Fluck [1,2,10,*,†]

1 Knowledge Management, ZB MED—Information Centre for Life Sciences, 50931 Cologne, Germany; langnickel@zbmed.de
2 Graduate School DILS, Bielefeld Institute for Bioinformatics Infrastructure (BIBI), Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany
3 IT Department, University Hospital Bonn, 53127 Bonn, Germany; kilian.krockauer@ukbonn.de
4 Department of Neurology, DRK Kamillus Klinik Asbach, 53567 Asbach, Germany; mischa.uebachs@kamillus-klinik.de
5 HPC and Scientific Computing, German Center for Neurodegenerative Diseases (DZNE) within the Helmholtz Association, 53127 Bonn, Germany; sebastian.schaaf@dzne.de
6 Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing SCAI, 53757 Sankt Augustin, Germany; sumit.madan@scai.fraunhofer.de
7 Institute of Computer Science, University of Bonn, 53115 Bonn, Germany
8 Department of Neurology, University Hospital Bonn, 53127 Bonn, Germany; thomas.klockgether@ukbonn.de
9 Clinical Research, German Center for Neurodegenerative Diseases, 53127 Bonn, Germany
10 The Agricultural Faculty, University of Bonn, 53115 Bonn, Germany
* Correspondence: fluck@zbmed.de or jfluck@uni-bonn.de
† These authors worked at *6* during conduction of the study.
‡ This author worked at *8* during conduction of the study.

Citation: Langnickel, L.; Krockauer, K.; Uebachs, M.; Schaaf, S.; Madan, S.; Klockgether, T.; Fluck, J. Information Extraction from German Clinical Care Documents in Context of Alzheimer's Disease. *Appl. Sci.* **2021**, *11*, 10717. https://doi.org/10.3390/app112210717

Academic Editor: Elena Cardillo

Received: 30 September 2021
Accepted: 9 November 2021
Published: 13 November 2021

Abstract: Dementia affects approximately 50 million people in the world today, the majority suffering from Alzheimer's disease (AD). The availability of long-term patient data is one of the most important prerequisites for a better understanding of diseases. Worldwide, many prospective, longitudinal cohort studies have been initiated to understand AD. However, this approach takes years to enroll and follow up with a substantial number of patients, resulting in a current lack of data. This raises the question of whether clinical routine datasets could be utilized to extend collected registry data. It is, therefore, necessary to assess what kind of information is available in memory clinic routine databases. We did exactly this based on the example of the University Hospital Bonn. Whereas a number of data items are available in machine readable formats, additional valuable information is stored in textual documents. The extraction of information from such documents is only applicable via text mining methods. Therefore, we set up modular, rule-based text mining workflows requiring minimal sets of training data. The system achieves F1-scores over 95% for the most relevant classes, i.e., memory disturbances from medical reports and quantitative scores from semi-structured neuropsychological test protocols. Thus, we created a machine-readable core dataset for over 8000 patient visits over a ten-year period.

Keywords: clinical text mining; data standardization; semantic interoperability

1. Introduction

For translational medicine, there is a tremendous need for broad, longitudinal health-related phenotype data. This holds true especially for Alzheimer's disease (AD), for which an etiology of decades is assumed, in major parts without clinical symptoms. In the field of AD, publicly available cohort data resources, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [1] are heavily used (as can be seen in the number of publications referring to ADNI). Other examples of cohort data in the field of neurodegeneration are the ADNeuroMed cohort [2] or The European Prevention of Alzheimer's

Dementia (EPAD) cohort [3]. In Germany, the German Center for Neurodegenerative Diseases (DZNE) has started large cohort studies, such as the DZNE Longitudinal Cognitive Impairment and Dementia Study (DELCODE) [4] and the DZNE Clinical Registry Study of Neurodegenerative Diseases (DESCRIBE) [5]. Despite all these efforts, it will take years to collect large subject groups and longitudinal data. Moreover, these cohorts follow their own, often unharmonized, study designs—hence, they might be biased toward their particular study designs.

In contrast, hospitals' clinical routine facilities are constantly visited by a broad range of patients. In memory clinics, cognitive deficits are routinely examined and documented together with the resulting diagnosis in medical reports. The most relevant cognitive deficits are memory, language, attention, and planning deficits. Further frequent deficit descriptions focus on general, temporal, and spatial orientation. In structured cohort data, these are core data items together with the estimated onset time. In addition, standardized assessments are highly relevant. In memory clinics, such assessments are often performed on a routine basis as well. One of the most prominent assessments—since it is quickly realizable and broadly covering—is the Mini-Mental State Examination (MMSE). The MMSE represents a short test consisting of just 11 questions that the patient has to answer concerning the cognitive aspects of mental functions [6].

At the University Hospital Bonn, in addition to MMSE, a more in-depth cognitive test, the CERAD (Consortium to Establish a Registry for Alzheimer's Disease; https://sites.duke.edu/centerforaging/cerad/, accessed on 25 September 2021) test battery is conducted. In order to quantify memory and language skills, the CERAD test battery offers a broad variety of well-established cognitive tests with a total of 18 test scores and as such, a deeper understanding of cognitive deficits. CERAD is subdivided into seven different subcategories, covering a broad range of commonly observed cognitive deficits. The subcategories comprise the MMSE, verbal fluency, the Boston Naming Test (BNT), construction ability, learning of word lists, recall, and recognition [7,8]. Because the full CERAD test battery is time consuming and quite exhausting for the patient, it is not measured on every visit, but mostly during the first, providing a detailed assessment right at the start. In follow-up visits, often the MMSE score is measured only (instead of the whole CERAD test battery), estimating any cognitive decline.

For each patient, the above-mentioned data points are captured in memory clinics on a routine basis in various documents. However, these clinical documents are not structured, as they consist of human-written texts. In this study, we therefore develop an information extraction workflow in order to automatically extract the relevant information from various clinical document types originated in the neurology department. We make use of text mining methods based on natural language processing (NLP). The application of text mining methods in hospital context is also known as clinical text mining and poses several additional challenges in comparison to general text mining.

Medical reports appear to not have a common and explicitly defined structure (as, for example, compared to a scientific paper). However, usual named form fields, paragraph headings, and topics/terms recurrently appear, especially when using a similar hospital information system or analyzing documents from one department. Furthermore, data privacy considerations limit the access to a feasible amount of training and test data in the local language. Publicly available datasets are often inaccessible or non-existing for languages other than English. Moreover, setting up a compliant software environment requires substantial efforts. Often, just small numbers of manually anonymized documents are available.

As the clinical records to be processed are written in German, commonly available medical text mining and NLP components for English language cannot be applied. Similarly, existing medical terminologies, such as ICD-10, are mainly focused on accounting processes, not matching the expressions used in medical reports [9]. Other important terminologies do not necessarily exist in the given local language. Furthermore, diagnosis criteria, such as assessment scales for neurodegenerative diseases, may differ between

countries and may even not be covered by any terminology yet. A well-known special characteristic of medical reports, especially in the context of anamnesis, is often the negation of mentioned diseases and symptoms for the sake of explicitly excluding them [10]. Therefore, the detection of negations needs to be considered well in clinical text mining. Similarly, family and disease history information need to be separated from current findings.

In this study, we implemented a text mining workflow to structure routine data hidden in German clinical care documents of a memory clinic. For this purpose, we created a training and a test dataset, which were both manually annotated by medical experts. The training dataset builds the foundation for the implementation of rule-based pipelines to extract information from medical and neuropsychological test reports of AD patients. In addition, we present a detailed evaluation of the workflow and, furthermore, show the valuable results collected by applying the workflow on a large memory clinic dataset. By structuring the large memory clinic dataset, we expect—despite moving toward digital transformation of the clinic itself—to extend cohort data with clinical routine data to support future research. We also expect to improve the case findings, due to the higher availability of clinical information that were hidden before in clinical care documents.

2. Related Work

In the past, the scientific community organized and published a series of shared tasks and datasets to promote exchange in the clinical NLP domain. A prominent example is the "Informatics for Integrating Biology and the Beside" (i2b2) initiative (https://www.i2b2.org/NLP/DataSets/Main.php, accessed on 25 September 2021), now known as the "National NLP Clinical Challenges" (n2c2 ; https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/, accessed on 25 September 2021) that aims to build a framework that provides clinical data to researchers. "SemEval Clinical TempEval challenges" and "TREC Precision Medicine/Clinical Decision Support Track" are further well-known communities that regularly campaign new tasks and datasets for clinical NLP. In general, these challenges focus on clinical NER tasks, such as the detection of diseases, symptoms, drugs/medication, or body parts [11] and on the temporal ordering of clinical events.

Clinical text mining systems are based either on rule-based approaches or machine learning (ML) techniques. In both cases, raw text is often preprocessed first by performing sentence detection, tokenization, stemming, part-of-speech tagging, or stopword removal. For the rule-based approach, a subsequent step involves the development of manually created rules that typically use regular expressions. With the availability of larger training datasets, the creation of ML-based applications using methods such as support vector machines (SVMs), conditional random fields (CRFs), or neural networks (NNs), have become popular.

For instance, the 2010 i2b2 challenge introduced a task on medical concept detection (such as symptoms and procedures), where most systems were based on CRFs. The performance of the best systems in terms of F1-score ranges from 0.788 to 0.852 [12]. Xu et al. developed a rule and dictionary–based system to extract medication findings (such as drug name, strength and frequency) from unstructured clinical narratives [13]. The authors achieved an F1-score of 93.2% for drug names. Fluck et al. [14] demonstrated two clinical applications focusing on the mining of unstructured German physician reports with rule-based approaches.

The first application dealt with information extraction from surgical reports of endoprosthesis implantations, e.g., anatomic region (knee/hip), reasons and type of surgery as well as surgical complications. The second application identified and interpreted TNM classifications of malignant tumors from pathology reports (where T, N, and M stand for tumor, lymph node, and metastasis, respectively). More recently, Becker et al. developed a pipeline to automatically annotate German clinical colorectal cancer notes. Using a dictionary-based approach, including pre-processing steps, they annotated UMLS concepts, achieving F1-scores in the range of 64.52% to 98.74% for the different classes [15].

As mentioned, negation detection plays an important role in clinical text mining. The well-known negation detector written for English medical reports is *NegEx*, which is a regular expression-based algorithm that works at the sentence level [10]. An extension of the NegEx algorithm, called *ConText*, focuses additionally on the recognition of hypothetical or historical statements as well as whether the clinical conditions are experienced by someone other than the patient [16]. Cotik et al. [17] adapted NegEx to German clinical reports by the creation of a new list of trigger terms, which is also publicly available.

Recent studies have shown the potential of using electronic health records as a source of valuable information for Alzheimer's disease research. Zhou et al. [18] proposed a rule-based NLP method to extract lifestyle exposures, thereby drawing conclusions about the risk factors. Similarly, Shen et al. [19] tested different deep learning-based models in order to also extract lifestyle factors for AD. As no training data were available, they first automatically developed a dataset using a rule-based approach and trained the models on the silver standard. However, for applying machine learning–based approaches, Kumar et al. [20] stated that the majority of research in this field is conducted using publicly available datasets.

To the best of our knowledge, nothing comparable has been yet achieved for German medical reports in the AD research domain. Moreover, the above-mentioned studies only focus on one data type each. In contrast, we focus on both unstructured and semi-structured data and developed dedicated text mining modules to extract a total of 26 attributes. Additionally, we designed the pipelines in such a way that structured data can be combined with clinical cohort data and can thus be used to generate more profound knowledge.

3. Materials and Methods

This study was conducted at the Memory Clinic of the University Hospital Bonn, Germany. It aims to automatically extract relevant patient features from both semi- and unstructured clinical care documents. In the following, we first provide an overview about the used datasets and then describe our two developed pipelines.

3.1. Data

The clinical care data comprise patient-related documents that can either be in a structured format (such as the date of birth or sex of the patient) or be stored as unstructured text in database fields or documents. The dataset studied was collected between 2002 and 2020 and includes raw data from 4834 patients with a mean age of 68 at visit. Detailed patient characteristics can be seen in Table 1.

Table 1. Patient characteristics of study data.

Total number of patients	4834
Gender	2291 males, 2285 females, 258 not specified
Mean age	68 ± 12 years
Time period	2002–2021
Amount of dementia-related ICD codes	1224

In this research, we aim to extract information from medical reports and neuropsychological tests. The following patient attributes are of interest: the date of visit, the AD-specific deficits (e.g., memory or language deficits) as well as their perceived onset, the corresponding diagnoses, the value of the MMSE score, and the cognitive test results. To extract these patients' attributes from the hospital information system (HIS), we prepared two different pipelines. The first, the medical report pipeline, extracts all patient attributes besides the cognitive test results. The second, the neuropsychological test (NPT) report pipeline, extracts the cognitive test results of the CERAD test battery, which is stored in a semi-structured format. All extracted attributes are mapped to the variables of the DESCRIBE

cohort study in order to enable semantic integration [5]. The variables of the DESCRIBE cohort study are also mapped to Unified Medical Language System (UMLS) concepts (see Medical Data Models Portal [21]; https://medical-data-models.org/details/13956, accessed on 25 September 2021). An overview of the developed pipelines is shown in Table 2.

Table 2. Overview of two developed pipelines with their data sources containing specific patient attributes. For both pipelines, the amount of training and test data is provided.

	Medical Report Pipeline	**NPT Report Pipeline**
Data source	Discharge letters, HIS	NPT reports
Document types	MS Word documents, HIS text fields	MS Word documents with well defined table
Patient attributes	Date of visit, diagnoses, AD-specific cognitive deficits, and MMSE score	Cognitive results of the CERAD test battery
Amount of training data	50 documents	20 documents
Amount of test data	100 documents	100 documents

3.2. Medical Report Pipeline

In order to develop and evaluate algorithms to build an information extraction system that can extract patients' features from unstructured medical reports, annotated training and test data are needed. Randomly chosen training and test data were manually annotated by two medical experts using the annotation tool *brat* [22]. The annotation was performed at the sentence level. In the following, we describe the annotation schema in detail. Afterwards, the inter-annotator agreement is introduced. We further explain the structure of the implemented medical report pipeline.

3.2.1. Annotation Schema

Firstly, we focused on the specific deficits that are often described in a medical report, triggered by typical questions asked by a physician to the patient. We annotated memory, language, attention, planning and orientation deficits, as well as their onset time. Typical indicators of the first category *memory deficit* are words such as "vergessen", "etwas verlegen", or "nachfragen" (engl. "to forget", "to mislay something", or "to ask again"). However, according to the medical expert, deficits concerning long-term memory do not indicate dementia of type Alzheimer's. The second category *language deficit* comprises language-related deficits and is often described by phrases such as "looking up for words" or "interchanging letters" (for instance "er suche oft nach Wörtern", engl. "he is often looking for words"). *Attention deficit* defines the third category and are for example simply stated as "Konzentration beeinträchtigt" (engl. "attention impaired"). The category *planning deficit* refers to deficits in complex human actions. An example sentence is "beispielsweise bei komplexeren Tätigkeiten wie dem Regeln von Finanzangelegenheiten gebe es zunehmend Unsicherheiten" (engl. "for example, in more complex activities such as the regulation of financial matters, there is increasing uncertainty"). The last category called *orientation deficit* mainly includes temporal or spatial orientation deficits (e.g., "räumliche Orientierungsstörungen") as well as general disorientation.

Furthermore, the recognition of the date of disease onset is tackled for both relative (e.g., three years ago) and absolute (e.g., since an accident in January 1998) time frames. In addition to this, a qualitative summary of the patient's anamnesis and examination, concomitant information on date of visit, diagnoses, and the MMSE score are subject to extraction from medical reports. In terms of diagnoses, the following 11 diagnoses of interest were specified: *Alzheimer's disease*, *subjective cognitive decline*, *mild cognitive impairment*, *Lewy-body dementia*, *posterior cortical atrophy*, *fronto-temporal dementia*, *primary progressive aphasia*, *limbic encephalitis*, *dementia syndrome with normal pressure hydrocephalus*, *vascular dementia* and *organic amnestic syndrome*. The MMSE score is usually described in

a numerical format (e.g., *X/30* as 30 is the maximum number of points to be achieved). An overview of the annotated classes with a corresponding example and the amount of occurrences in the test set can be seen in Table 3.

Table 3. Overview of all annotated classes subject to the extraction from medical reports. The amount column refers to the number of instances of the manually annotated test set (on sentence level). * The amounts are split into approved and negated (app./neg.) statements (on sentence basis).

Section	Entity Class	Example	No. of Instances
Personal info.	Date of visit	berichten über ... am DD.MM.YYYY	49
Diagnosis	Diagnosis	Alzheimer-Krankheit und subkortikale arterisklerotische Encephalopathie	84
Examination summary	MMSE score	Mini-Mental-Status v. 29.10.2007: 24 von 30 Punkten	100
Examination summary	MMSE date	Mini-Mental-Status v. 29.10.2007: 24 von 30 Punkten	25
Anamnesis	Onset time	und berichtete, seit ca. 6 Monaten vergesslich und unkonzentriert zu sein.	74
Anamnesis	Memory deficit	und berichtete, seit ca. 6 Monaten vergesslich und unkonzentriert zu sein.	168/07 *
Anamnesis	Language deficit	mache umständliche Beschreibungen wegen teils Wortfindungsstörungen	44/12 *
Anamnesis	Attention deficit	freundlich zugewandt, Konzentration beeinträchtigt, Auffassung und Merkfähigkeit waren eingeschränkt.	16/07 *
Anamnesis	Planning deficit	beispielsweise bei komplexeren Tätigkeiten wie dem Regeln von Finanzangelegenheiten	25/17 *
Anamnesis	Orientation deficit	gebe es zunehmend Unsicherheiten Keine räumlichen Orientierungsstörungen	34/22 *

3.2.2. Inter-Annotator Agreement

For the five mentioned deficits and their onset, we determined the inter-annotator agreement (IAA), as this helps to analyze the complexity of the wording that can be used by different physicians. For this purpose, two field experts annotated a total of 25 medical report documents. The F1-measure was used to determine the IAA, using the following formula where *TP*, *FP* and *FN* stand for *true positive*, *false positive* and *false negative*, respectively.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1-score} = 2 \times \left(\frac{precision \times recall}{precision + recall}\right) \tag{3}$$

However, before annotating the final IAA document set, three annotation rounds—each with ten documents—were performed. After each round, the annotation guidelines were adapted in a consecutive discussion. To determine the IAA, we used the annotations of one annotator as the gold standard and compared them to the results of the second annotator. The results of the IAA can be seen in Table 4. We assessed the results of cognitive deficits extraction at the sentence as well as the document level. The medical experts classified the deficits as being approved or negated. Thus, the absolute amount of negated as well as approved entities was calculated. However, at the document level, in some cases, contradictory information might be detected by the system. To cope with such cases, we performed a post-processing step in which the entity with the highest amount was taken into account. For example, if a memory deficit was detected twice as being approved and once as being negated, the approved entity was taken into account for the evaluation. If the occurrence of both cases was equally high, the approved one was taken into account.

These post-processing steps were applied on both the manually curated and automatically extracted data. We also chose to not evaluate the negation detection separately, but take only the final result into account, so that we were able to judge on the final performance of the developed systems.

Except for the attention deficit, the IAA of all classes is higher on document than on sentence level. The highest agreement is achieved for the memory deficit and amounts to an F1-score of 86% on sentence level and 97% on document level. For language and orientation deficits as well as the date of the disease onset, an agreement above 90% on document level is achieved. However, reaching an agreement on attention and planning deficits seems to be less trivial. Here, F1-scores of 75% and 62% are reached, respectively.

Table 4. Inter-annotator agreement on both sentence and document level.

Class	Level	Precision	Recall	F1-Score
Memory deficit	Sentence	0.9216	0.8103	0.8624
	Document	0.95	1.0	0.9744
Language deficit	Sentence	0.8667	0.619	0.7222
	Document	0.8333	1.0	0.9091
Attention deficit	Sentence	0.8333	0.7143	0.7692
	Document	0.75	0.75	0.75
Planning deficit	Sentence	0.5714	0.3636	0.4444
	Document	0.6667	0.5714	0.6154
Orientation deficit	Sentence	0.75	0.6	0.6667
	Document	1.0	0.8571	0.9231
Onset time	Sentence	0.9231	0.96	0.9412
	Document	0.9231	0.96	0.9412

3.2.3. Structure of the Medical Report Pipeline

An overview of the structure of the medical report pipeline can be seen in Figure 1 (on top). The pipeline consists of several steps that are performed in a consecutive manner. First, the textual contents from MS Word documents in DOC and DOCX format and text-based fields of the *Agfa ORBIS* hospital information system (HIS) were parsed to prepare the medical report documents for further analysis (e.g., cleaning, harmonizing encoding, converting to plain-text format). Furthermore, basic preprocessing steps, i.e., sentence detection, tokenization, and lemmatization are applied. Second, we apply a document decomposer to structure the text documents into different paragraphs. This is required because our information extraction components are highly modularized and are, therefore, specifically designed for different sections appearing in the documents. In this way, each of the IE components can easily be embedded in new workflows for different clinics. These IE components extract various annotations for the defined classes. Finally, we apply a post-processing step in order to detect negated statements. In the following, the various parts of the pipeline are described in detail.

Document decomposition: This module subdivides the medical reports into nine different sections, typical of medical reports created in neurology departments (e.g., anamnesis, and epicrisis). For their recognition, we created a terminology with the usual titles of the corresponding paragraphs. These titles were automatically extracted from the HIS and were further extended by a medical expert. The titles were embedded in a dictionary, which is used in a rule-based system to detect the begin and end of sections. Example rules with detailed explanations can be found in the Appendix A (see Table A4).

Cognitive Deficits Detection: To recognize AD-specific cognitive deficits, we developed four separate components that run on their associated sections (see Table 3) extracted from the medical reports. By analyzing the annotated training data, typical expressions for each of the entity classes could be identified. The general working mechanism of these components is similar to the document decomposer: we combine dedicated terminologies with regular expressions that use lemmas for detection.

Moreover, we normalize the detected entities to map them to the corresponding variables from the DZNE cohort study. The amount of rules needed for a single entity class varies strongly. The detection of the memory deficits is the most complex. Here, we first detect simple phrasings, such as "vergessen" or "häufig verlegen" (engl. "to forget", "to mislay often"). Second, we detect words and corresponding negation terms that indicate a memory deficit only if they occur together, e.g., "to not remember". An example can be found in the Appendix A. Third, we unmark detected annotations that contain the term "Langzeit" (engl. "long term").
We relate all the cognitive deficits to a detected onset time. Therefore, sophisticated rules were developed in order to detect the relative and the absolute onset (two years ago vs. since January 2019). If a single sentence contains both a cognitive deficit annotation and an onset, these are related to each other.

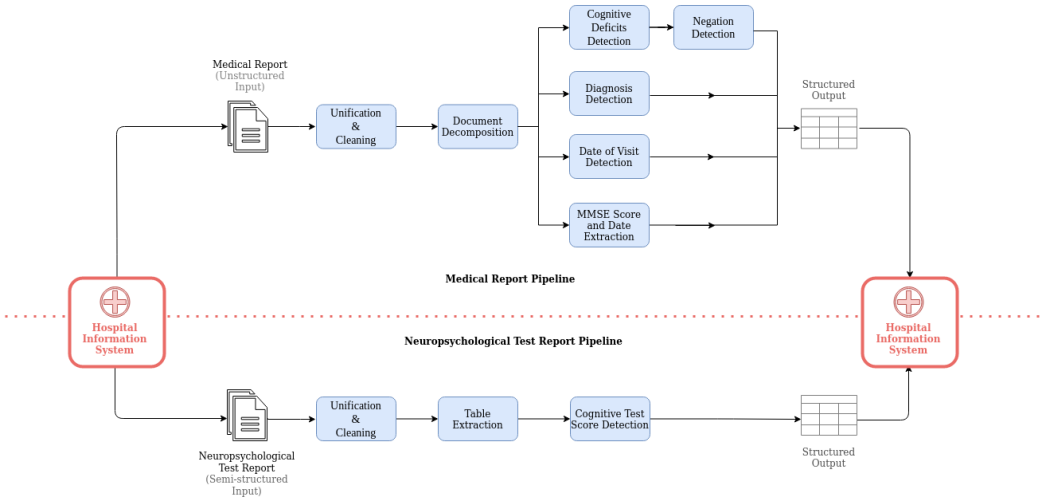


Figure 1. Patients' attributes extraction workflow integrated in hospital information system. The figure depicts the different text mining pipelines that are integrated into the hospital information system. The workflow of the medical report pipeline that is subdivided into four different components is illustrated at the top. The bottom half depicts the neuropsychological test report pipeline that extracts various scores from semi-structured test reports.

Negation detection: In order to detect negated annotations, we developed a German negation detector. The algorithm is based on two different kinds of terminologies that were assembled manually and supplemented by translations of NegEx [10]. The first terminology includes all terms that represent a negation itself, for example, "verneint" (engl. "negated") or "unproblematisch" (engl. "unproblematic"). The other contains terms that are only negated in the presence of four specific negation terms "nicht" (engl. "not"), "nie" (engl. "never"), "ohne" (engl. "without") and "kein" (engl. "no" or "none"). Two examples could be "nicht berichten" (engl. "not report") and "kein Hinweis" (engl. "no evidence"). Therefore, the terms of the second terminology are only annotated when they follow or precede one of the four mentioned negation terms.
The next step is to detect the concepts that are negated by the corresponding negation term. As German medical reports often contain long sentences, we first identify sub-clauses based on conjunctions. We then iterate over each sentence that contains both an annotated cognitive deficit and a negated term. If the sentence contains a sub-clause, we mark all cognitive deficits as negated that appear in the same part of the sentence as the negation term. If the sentence does not contain a sub-clause, all deficits within this sentence are marked as negated. To increase the precision, several rules are additionally implemented to catch exceptions that are found in the training set. Examples can be found in the Appendix A in Tables A5 and A6.

3.3. Neuropsychological Test (NPT) Report Pipeline

The second pipeline implements the extraction rules for semi-structured tabular CERAD test battery results that are included in NPT reports (see Figure 1, on bottom). The results of the CERAD test battery appear within a relatively standardized table format. Consequently, we ignore free text paragraphs outside tables. Similar to the medical report pre-processing step, the test reports are first pre-processed and converted with the document parser to plain text. Each paragraph is positively tested for being part of a table area that gets parsed. Consequently, extracted tables are collections of rows, which are assembled right after their respective ends.

The CERAD table consists of seven different categories. The amount of tests belonging to each category strongly varies. Whereas there is only one test for the verbal fluency, the category verbal memory consists of eight sub-tests. In total, the test battery comprises 18 tests. The scores of the CERAD test battery can be given in different units in NPT reports. The medical experts provide raw values for the specific evaluation metric, such as the absolute amount of words mentioned by the patient. Moreover, these raw values can be transformed into relative values, such as *Z-value*, which are standardized toward their associated feature, for example, age and gender. As not all reports are complete—some reports only contain raw values, whereas others contain only transformed values—we developed rules for the extraction of all different kinds of values. In Figure 2, an example of a complete CERAD test battery result table of a single patient can be seen that has two columns for both raw (in German abbreviated as RW) and Z-values. Example rules can be seen in Table A7.

1. Verb. Flüssigkeit kateg. (Sprachproduktion, kogn. Flexibilität)				
Genannte Worte (Kategorie „Tiere“, 1 min.)	RW:	17	Z:	**-1,1**
2. Boston Naming Test (Aphasie-Screening)				
Von 15 Bildern konnten richtig benannt werden:	RW:	13	Z:	**-1,0**
3. **Mini-Mental-State Examination** (Demenz-Screening)				
Gesamtpunkte (max. 30)	RW:	24	Z:	**-3,5**
4. Wortliste (Verbales Gedächtnis)				
Von 10 Wörtern konnten richtig wiedergegeben werden:				
Summe über 3 Lerndurchgänge (max. 30)	RW:	20	Z:	**-0,7**
Nach dem ersten Lerndurchgang	RW:	6	Z:	**0,1**
Nach dem zweiten Lerndurchgang	RW:	4	Z:	**-2,8**
Unmittelbar nach dem dritten Lerndurchgang	RW:	10	Z:	**1,1**
Abrufen nach Interferenzaufgabe	RW:	3	Z:	**-2,5**
Intrusionen	RW:	3	Z:	**-2,0**
Savings Wortliste (in Prozent)	RW:	30%	Z:	**-3,4**
Diskriminabilität (in Prozent)	RW:	85%	Z:	**-2,3**
5. Konstruktive Praxis (Visuokonstruktion, visuelles Gedächtnis)				
Von 11 Zeichnungselementen konnten richtig gezeichnet werden:				
Direktes Abzeichnen	RW:	11	Z:	**1,1**
Verzögerte Wiedergabe	RW:	5	Z:	**-2,0**
Savings	RW:	45%	Z:	**-1,8**
6. Verb. Flüssigkeit phon. (Sprachproduktion, kogn. Flexibilität)				
Genannte Worte (Anfangsbuchstabe „S“, 1 min.)	RW:	12	Z:	**-0,2**
7. TMT Pfadfindertest (visuomotor. Geschwindigkeit, Übersicht)				
Teil A:	Sek:	36	Z:	**0,7**
Teil B:	Sek:	117	Z:	**-0,5**
Teil B/A:	RW:	3,3	Z:	**-1,0**

Figure 2. Example of a CERAD test battery results table of a single patient.

3.4. Implementation

The document parsers for extracting information from Microsoft Word documents were created with the Apache POI library (v3.9; https://poi.apache.org/, accessed on 20 September 2021). For the various proprietary file formats (such as .doc, and .docx) of Microsoft Word, the POI library provides a Java API that allows iterating over potential paragraphs as well as tables via the so-called range iterator class that eases the programmatic accessibility of the included information. For both the sentence detector and the tokenizer, in-house built algorithms based on regular expressions were used. For detection of lemmas, the lemmatizer from *mate tools* (https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools/, accessed on 15 August 2021) was used.

The implementation of dedicated modules (such as document decomposition, negation detection, cognitive deficit entity recognition) was realized with the Unstructured Information Management System (UIMA) (https://uima.apache.org/, accessed on 20 September 2021) Rule-based Text Annotation (Ruta), which is an imperative rule language to define patterns of annotations, optionally with additional conditions [23]. The embedding and execution of the dedicated modules as a text mining workflow was performed with the popular UIMA framework. We provide our code under https://github.com/llangnickel/GermanClinicalTM (accessed on 20 September 2021) together with some sample data (see also the Data Availability Statement section).

4. Results

4.1. Medical Report Pipeline Evaluation

The medical report pipeline contains separate modules that are evaluated independently. First, the document decomposition is evaluated by means of 50 independent documents. The results can be seen in Figure 3. All sections could be identified with F1-scores above 93% with an equally high overall precision and recall. At the current stage, only the sections *Date*, *Anamnesis*, *Diagnosis* and *Examination Summary* are processed further. Those four sections are recognized with F1-scores of 99% (*Date*), 97% (*Anamnesis*) and 98% (both *Diagnosis* and *Examination Summary*) (see Figure 3). In order to have a reliable evaluation of the lower-level components, we feed them with the automatically extracted sections.

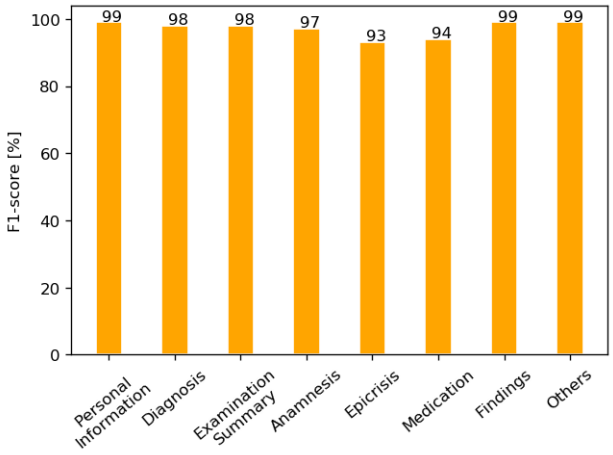
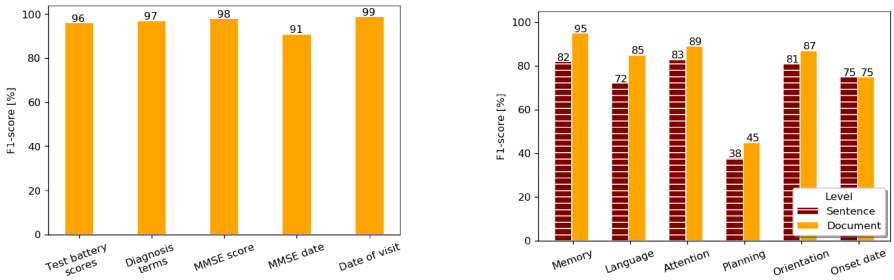


Figure 3. Results of extraction of sections of interest. The detailed results containing precision and recall are included in the Table A1. For further processing, currently only the first four sections are needed (compare Table 3).

From the *Diagnosis* sections, eleven different neurodegenerative diagnosis terms, defined by the medical expert, could be extracted with a recall of 94.05% and a precision of 100%, resulting in an F1-score of 96.93% (see Figure 4a). From the *Examination Summary* sections, we extracted the *MMSE* values. These are the most abundant follow-up cognitive scores we can use in data analyses to estimate changes in cognitive state over time. The evaluation of the *MMSE* score occurring in free text results in an F1-score of 98%. The extraction of the *Date of Visit* reaches a performance of 99% (see Figure 4a).

The extraction of AD-specific cognitive deficits (extracted from the *Anamnesis* section) is the most sophisticated and complex task. First, the *onset date* annotations were evaluated based on the gold standard. The onset date usually refers to the start/awareness of any cognitive deficits, for which the evaluation revealed a precision of 78.58%, a recall of 72.37% and an F1-score of 75.35% (see Figure 4b).



(a) Extraction results from test battery, diganosis, examination summary and personal information section

(b) Extraction results from anamnesis section

Figure 4. Summary of information extraction results. (**a**) Results of extraction test battery scores, diagnosis terms and date of visit. (**b**) Results of extraction of AD-specific cognitive deficits from anamnesis pipeline. The detailed results containing precision and recall can be found in Appendix A in Tables A2 and A3.

For both sentence and document levels, the extraction of cognitive deficits shows satisfactory results and are better on the document level (Figure 4b). This is expected, as a specific deficit is often described more than once in a single report. Thus, common wording is usually used in the beginning, such as "er leidet an einer Gedächtnisstörung" (engl. *"he suffers from memory loss"*); however, later in the report, wordings could become more specific, which makes it harder to detect at the sentence level. In brief, our system achieves an F1-score of 94.87% for the memory deficit on document level and an F1-score of 81.77% on sentence level. The second highest score on document level is achieved for the category *Attention*, amounting to 88.89% in terms of F1-score. Similar to the manual annotation measured with IAA, the extraction of deficits for the category *planning* are most difficult. The system achieves only an F1-score of 45%. This low F1-score is the result of a very low recall (31%), whereas the precision of 79% is acceptable.

4.2. NPT Report Pipeline Evaluation

The evaluation of the NPT report pipeline is averaged over the 18 different test score attributes that we extract. It shows, as expected, high results, due to the semi-structured format of the NPT reports. Precision and recall amount to 99.92% and 92.78%, respectively, leading to an F1-score of 96.22% (see Figure 4a). During a subsequent analysis, we found that most false negatives are caused by incomplete or falsely filled tables, containing either different wordings or a different tabular format. The exercises of the CERAD test battery are often only partly conducted with patients, and we, therefore, expect missing values. The *Boston Naming Test* is performed most often, whereas the *Trail Making Test* (category seven) is often not conducted anymore.

4.3. Application of the Extraction Workflow on Large Memory Clinic Dataset

The main motivation for the development of such a system was to apply it on a large memory clinic dataset to structure the patient information, which can be used for future research analyses. Hence, the implemented system was used to extract patient data from the memory clinic over an extended time period from 2002 to mid 2021. Overall, we gathered structured data for 4834 patients. A deeper analysis of the collected data, comparison to cohort data and data analysis scenarios are planned in future work. Nevertheless, an overview of the extracted data is summarized in Table 5.

As expected, on the first patient visit, the highest number of attributes could be collected from the clinical care documents. The implemented diagnosis detection focusing on dementia-related disorders identified 1664 such diagnoses. In comparison, only a total of 1224 dementia-related ICD-10 codes were registered (see Table 1), meaning that we can extract 36% more detailed diagnosis information from the medical reports. This emphasizes the benefit of the developed pipeline, even when structured information is available. The following items became only accessible through the extraction pipeline. The MMSE score,

which is used as a fast test to exclude cognitive deficits, could be extracted 1971 times. Furthermore, during the first visit of a patient, a high number of cognitive deficits, in total 2347, were recognized in medical reports. This covers all the five different cognitive deficits, such as *Memory* or *Language*. In follow-up visits, 828 further diagnosis terms and 849 MMSE scores were found.

Table 5. Summary of patient data collected from the memory clinic. It covers different classes and patient visits. For the patients, the first number indicates the number of patients having at least the indicated amount of visits, whereas the number in brackets indicates the number of patients with exactly the mentioned amount of visits. For example, 2052 patients have at least 4 visits and 495 from them have exactly 4 visits.

No. of Visits	No. of Patients	No. of Cognitive Deficits	No. of Test Batteries	No. of MMSE Scores	No. of Diagnosis Terms
First visit	4834 (1042)	2347	699	1971	1664
Second visit	3792 (965)	1134	906	849	828
Third visit	2827 (775)	898	353	532	676
Fourth visit	2052 (495)	648	166	354	497

5. Discussion

In this study, we have presented and evaluated a text mining workflow that extracts AD-specific patient attributes from clinical routine data of the neurology department with high extraction performance. Designing and implementing the workflow that can cope with a heterogeneous clinical data warehouse required identifying and solving many hurdles: first, we needed to identify what kind of information is stored in which resource; second, we found the database fields that include the relevant semi-structured data; third, for the textual documents (such as MS Word), we had to assess if the information can be found and extracted reliably; and fourth, we had to determine how the training and test datasets could be generated while at the same time complying with data privacy regulations. Especially for the last issue, it became clear that it will take a significant time investment of medical experts to annotate and anonymize a certain amount of data.

Due to the low number of training data and our preliminary experiments, we decided to focus on rule-based implementation, even though the current trends (such as using neural networks) show high potential for entity recognition tasks [24] that are also applied in the clinical domain [11]. The success of the state-of-the-art methods depends on the pre-training task, which is performed on huge amounts of both nonspecific and domain-specific data. The *Clinical-BERT* [25] model, for instance, is pre-trained on approximately two million clinical notes from the publicly available MIMIC-III corpus [26]. As there is no corresponding dataset available for the German language, a traditional rule-based approach was considered the best choice. It is expected that this situation will change in the future, as the need for appropriate training data is rising also in the German clinical domain. As more training data become available, we will extend our workflow with machine learning models to predict certain classes, which will certainly decrease the manual effort of adapting rules.

The developed text mining workflow consists of two pipelines that process two different types of documents. Whereas the neuropsychological test reports are already in a semi-structured (tabular) format, the medical reports contain free text. For the NPT reports, the extraction of the 18 cognitive test scores reaches an F1-score of 96%. This result shows that in some cases, it is easily—and with low technical effort—possible to make high-value information machine readable. We should not neglect these "low hanging fruits", but focus on making these data available for further research.

The medical report pipeline is more complex and consists of several components. The first step is the document decomposition that represents the basis of the extraction of different entity classes in a consecutive step. Therefore, we developed dedicated modules for these sections. The detection of the cognitive deficits in the *anamnesis* paragraph is quite sophisticated and needed careful considerations while defining the extraction rules.

As the written free text is based on individual patient narratives, both the content and wording can differ enormously. The determination of IAA represents a good indicator for the expected extraction quality. In our evaluation, the automatic system achieves quite comparable results to IAA; in some cases, the system even beats the IAA scores. Overall, the recognition of various sections achieves an F1-score of 97%.

For four of the detected sections, we developed further dedicated rules to extract diagnosis terms, the date of visit, the MMSE score with date and the cognitive deficits. Whereas we reached F1-scores above 90% for the first three entity classes, both the manual and automatic recognition of the six different cognitive deficits experience a wide range of F1-scores. The most often occurring deficit—the memory deficit—can be extracted reliably with an F1-score of 95% on document level.

One of the important steps is also the transfer of the extracted data back into the hospital information system. This is needed to ensure proper access to the extracted data, thereby enabling its re-use. In the present study, we used a specific data model tailored especially for the University Hospital Bonn to achieve the goal of extending cohort data with their clinical routine data. In future, it will be preferable to adapt to a more generic data model that can be applied across different hospitals and institutions. Such a schema could be modeled with Fast Healthcare Interoperability Resources (FHIR; https://hl7.org/FHIR, accessed on 25 September 2021), a standard for data exchange in health care.

In future, we aim to apply the developed pipelines in different hospitals and analyze how far the workflow needs to be adapted. In Germany, the CERAD test battery is commonly used to assess patients for AD. Even if different hospitals store these results differently, chances are quite high that our pipeline can be (at least partly) applied. The diversity of the clinical information systems and document formats in clinics can only be coped by enhancing standardization, which is yet hard to achieve. Therefore, we chose the best current option, i.e., to modularize the text mining pipeline, to allow for easy customization of various components to cope with the available diversity.

With the application of our developed workflow within the University Hospital, we could gain valuable, machine-readable data that can be used to enlarge cohort study data, such as the clinical study DESCRIBE [5] conducted currently in Germany. To achieve this semantic interoperability, we mapped our extracted attributes to the variables used by DESCRIBE (https://medical-data-models.org/details/13956, accessed on 25 September 2021) that can be found at Medical Data Models Portal [21]. Thus, the number of patient data can be increased about tenfold, which will allow for more significant analysis. Moreover, using the patient data from clinics could allow a realistic sample of the population, and a possible cohort selection bias could be significantly reduced. In the long run, the integration of such structured routine data with the cohort studies data and their combined analysis will enable new routes for Alzheimer's disease research.

6. Conclusions

For the purpose of extending cohort data with routine data in the context of Alzheimer's disease, this study processed, analyzed and structured German clinical care documents from the neurology department of the University Hospital Bonn. Two rule-based information extraction pipelines were built to extract and structure information from medical and neuropsychological test reports. The results of the workflow evaluation and its application on a large amount of patient records demonstrate the usefulness of the proposed workflow—this serves as a suitable method to gather valuable phenotype information from the hidden treasures in hospitals. An obvious extension of our work would include the adaptation and application of the developed tools to other hospitals to make more patient data accessible for Alzheimer's disease research.

Author Contributions: Conceptualization, S.S., S.M., T.K. and J.F.; data curation, M.U.; methodology, L.L., S.M. and J.F.; project administration, T.K.; resources, K.K.; software, L.L., K.K. and S.M.; supervision, S.S. and J.F.; validation, M.U.; visualization, L.L. and S.M.; writing—original draft, L.L.; writing—editing, L.L., S.M.; writing—review and editing, L.L., Kilian Krockauer, M.U., S.S., S.M., T.K. and J.F. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the German Federal Ministry of Education and Research (BMBF) as part of the program "i:DSem—Integrative Data Semantics in the Systems Medicine", project number 031L0029 [A-C].

Data Availability Statement: The clinical care documents of the patients cannot be published, due to data protection and privacy reasons. However, we provide all developed terminologies, the source code of the system, and few synthetic datasets that were generated by a physician. The system can be tested on synthetic data. All these resources can be found at https://github.com/llangnickel/GermanClinicalTM (accessed on 20 September 2021).

Acknowledgments: We would like to thank Marc Jacobs for the support in regular expression-based preprocessing of the textual documents.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Further Results

The appendix contains the detailed results and some code examples of the developed text mining components.

Appendix A.1. Detailed Results of the Text Mining Pipelines

In the following, detailed results from our evaluations are shown for all pipelines and all classes. In Table A1, precision, recall and F1-score are shown for the extracted sections of interest. Out of these sections, we currently further process four sections, namely *Personal Information*, *Anamnesis*, *Diagnosis* and *Examination Summary*. We reach F1-scores above or equal to 97% for these sections.

Furthermore, the information extraction results for the dedicated classes are shown in Table A2. For the test battery scores, the precision (99.92%) is higher than the recall (92.78%). As the test reports are semi-structured, the slightly lower recall can be explained by deviations from the standard structure, which leads to unrecognized values (false negatives). It is similarly true for the diagnosis terms, where unusual naming could prevent detection. The *MMSE date* is extracted less efficiently than the *MMSE score* itself, which is caused by a lower recall of 86.12%. This is due to the fact that the date is not always clearly assignable to the corresponding test score.

Table A1. Information extraction results for the sections of interest (see also Figure 3).

	Personal Information	Diagnosis	Examination Summary	Anamnesis	Epicrisis	Medication	Findings	Others
Precision	1.00	0.98	1.00	0.94	0.93	0.97	0.97	0.98
Recall	0.98	0.98	0.96	1.00	0.93	0.91	1.00	1.00
F1-score	0.99	0.98	0.98	0.97	0.93	0.94	0.99	0.99

Table A2. Results for the NPTs, the diagnosis terms and the MMSE from fulltext (see also Figure 4a).

	Precision	Recall	F1-Score
Test battery scores	0.9992	0.9278	0.9622
Diagnosis terms	1	0.9405	0.9693
MMSE score	0.96	1	0.9796
MMSE date	0.9741	0.8612	0.9141
Date of visit	0.99	1	0.99

Table A3 shows the automatic extraction results of the cognitive deficits detection pipeline at both the document and sentence levels. For the different categories, a wide range of F1-scores can be seen. At the document level, the lowest scores are reached for the *planning deficit* (44.4% for IAA and 44.9% for automatic extraction). In contrast, the best results are achieved for the memory category that occurs also the most often. Here, the system reaches an F1-score of 94.87%, which is close to the IAA of 97.44%.

Table A3. Information extraction results at both sentence and document levels (see also Figure 4b).

Category	Level	Precision	Recall	F1-Score
Memory	sentence	0.8093	0.8263	0.8177
	document	0.961	0.9367	0.9487
Language	sentence	0.7451	0.6786	0.7103
	document	0.9091	0.8	0.8511
Attention	sentence	0.7692	0.9091	0.8333
	document	0.8421	0.9412	0.8889
Planning	sentence	0.6875	0.2619	0.3793
	document	0.7857	0.3143	0.449
Orientation	sentence	0.8182	0.8036	0.8108
	document	0.8431	0.9149	0.8776
Onset date	sentence	0.7857	0.724	0.7534
	document	0.7857	0.724	0.7534

Appendix A.2. Exemplary Rules Implemented in the Text Mining Pipelines

In the following, we explain for each component a small excerpt and simplified version of the exemplary *UIMA Ruta* rules and regular expressions of our developed pipelines. We further refer the reader to the documentation of *UIMA Ruta* (https://uima.apache.org/d/ruta-current/tools.ruta.book.html#ugr.tools.ruta.language.syntax, accessed on 20 September 2021) to understand the language syntax.

Appendix A.2.1. Medical Report Pipeline

In this section, we provide examples for each parts of the medical report pipeline. The first step is the document decomposition. Afterward, we show some example rules for the cognitive deficits detection, and finally, we show some excerpts of the negation detector rules.

Document Decomposition: The developed rules for the detection of the Anamnesis Section are shown in Table A4. They consist of two different parts. First, we detect all anamnesis related words—based on lemmas—that are followed by a colon. Here, we also take care for often occurring misspellings. In the second rule, we perform a postprocessing step to remove wrongly classified terms, such as "family anamnesis".

Cognitive Deficits Detection: In Table A5, we show two simple, exemplary rules for the detection of memory deficits. In both cases, the value is set to *true* and the key corresponds to the study variable.

Negation Detection: The negation detector consists of several different parts. Some example rules can be seen in Table A6. In the first rule, we detect all terms that define a negation in combination with terms such as "not", "no", or "without". In order to find out which term is negated in the sentence, we need to know whether a sub-clause is available. Therefore, we detect conjunctions—shown in the second row. Based on this, we can define our logic: if a sentence contains a negation and a deficit but no conjunction, we can remove the annotated deficit. In contrast, if there is a conjunction in the sentence, we look for a comma in order to mark the whole sub-clause. Thus, we can then remove false positive annotations in the specific part of the sentence.

Appendix A.2.2. Neuropsychological Test Report Pipeline

Three rules from our NPT pipeline can be seen in Table A7. As the test reports are already in a semi-structured form, the basic principle can be applied for every test score. First, we detect the names of all the tests. Afterward, the units are detected, e.g., "Z". Finally, we detect the value that follows the unit (see Figure 2 for comparison).

Table A4. Exemplary rules for document decomposer.

Rule Description	Exemplary Rules for Document Decomposition (Anamnesis Section)

Recognition of anamnesis related words that precede a special character (colon)	`Lemma {`
	`  contains(Lemma.wordLemma,"anamnese")`
	`} COLON { ->`
	`  CREATE(Anamnese, 1, 2, "key"="Anamnese")`
	`};`
	`Lemma {`
	`  contains(Lemma.wordLemma,"zwischenbericht")`
	`} COLON { ->`
	`  CREATE(Anamnese, 1, 2, "key"="Anamnese")`
	`};`

Recognition and removal of wrongly classified terms (such as "family anamnesis") collected in a terminology	`WORDLIST Anamnese_FPList = 'config/anamnese_FP.txt';`
	`Document {->`
	`  MARKFAST(Others, Anamnese_FPList, true, 2)`
	`};`
	`Anamnese{`
	`  PARTOF(Others) -> UNMARK(Anamnese)`
	`};`
	`Anamnese{`
	`  CONTAINS(Others)-> UNMARK(Anamnese)`
	`};`

Table A5. Exemplary rules for cognitive deficit extraction.

Explanation	Code

Detect a memory deficit (such as "häufig fragen", engl. "ask frequently") and set a key that corresponds to the study variable	`Lemma {`
	`  contains(Lemma.wordLemma,"häufig")`
	`}`
	`ANY??`
	`Lemma{`
	`  contains(Lemma.wordLemma,"fragen") ->`
	`  CREATE(Gedaechtnis, 3,`
	`  "value"="True", "key"="ACXMEM")`
	`}`

Detect a memory deficit (such as "nicht erinnern", engl. "does not remember") and set a key that corresponds to the study variable	`Lemma{`
	`  REGEXP(Lemma.wordLemma,`
	`  "schlecht|nicht|problem|schwierigkeit")`
	`}`
	`W[0,6]?`
	`Lemma {`
	`  REGEXP(Lemma.wordLemma,"erinnern|merken") ->`
	`  CREATE(Gedaechtnis, 1, 3,`
	`  "value"="True", "key"="ACXMEM")`
	`}`

Table A6. Exemplary rules of the negation detector.

Rule Description	Exemplary Code

Detect all terms that define a negation in combination with terms such as “not”, “no”, or “without”	`Lemma{`
	`  REGEXP(Lemma.wordLemma, "nicht|nie|kein|ohne")`
	`}`
	`ANY??`
	`Lemma{`
	`  REGEXP(Lemma.wordLemma,`
	`  "berichten|...|problem|defizit") ->`
	`  MARK(Negation, 1, 3)`
	`}`

Detect conjunctions	`Lemma {`
	`  REGEXP(Lemma.wordLemma,`
	`  "jedoch|aber|trotzdem|dennoch|...") ->`
	`  MARK(Trennung_Unit)`
	`}`

If sentence does not contain a conjunction but a negation, remove deficit annotation	`Sentence {`
	`  AND(-CONTAINS(Trennung_Unit), CONTAINS(Negation))`
	`} -> {`
	`  Stoerung_Value { ->`
	`  UNMARK(Stoerung_Value)`
	`  };`
	`};`

Based on conjunction and comma, mark the sub-clause	`Document {`
	`  CONTAINS(Trennung_Unit)`
	`} -> {`
	`  COMMA ANY[0,5]? Trennung_Unit ANY*? PERIOD {->`
	`  MARK(Nebensatz_Unit, 2, 4)`
	`  };`
	`};`

Set all occurrences of a cognitive deficit in a denied sub-clause to *False*	`Nebensatz_Unit {`
	`  AND(CONTAINS(Negation), CONTAINS(Stoerung_Value))`
	`} -> {`
	`  Stoerung_Value {->`
	`  SETFEATURE("value", "False")`
	`  };`
	`};`

Table A7. Exemplary rules for the NPT report pipeline.

Rules Description	Exemplary Code

Detect verbal fluency ("verbale Flüssigkeit kategorisch")	`CW {`
	`  REGEXP("Verb|Verbale")`
	`} PERIOD? W SW {`
	`  REGEXP("kateg|kategorial")`
	`} PERIOD? { ->`
	`  MARK (VerbFluessigkeit, 1, 5)`
	`};`

Use predefined table to detect key Z	`Table {} -> {`
	`  SEMICOLON W{`
	`  REGEXP("Z|z") -> Z_Unit`
	`  }`
	`  COLON SEMICOLON;`
	`};`

Based on the two previous annotations, detect the corresponding value (NUM)	`Table {} -> {`
	`  VerbFluessigkeit ANY[1,30]? Z_Unit COLON SEMICOLON`
	`  SPECIAL.ct == "-"? NUM NUM? COMMA? PERIOD?`
	`  NUM? SEMICOLON {-> MARK(Z_VFATOT, 6, 11)`
	`  };`
	`};`

References

1. Alzheimer's Disease Neuroimaging Initiative. About ADNI. 2017. Available online: http://adni.loni.usc.edu/about/ (accessed on 30 July 2021).
2. Lovestone, S.; Francis, P.; Strandgaard, K. Biomarkers for disease modification trials–the innovative medicines initiative and AddNeuroMed. *J. Nutr. Health Aging* **2007**, *11*, 359–361. [PubMed]
3. Solomon, A.; Kivipelto, M.; Molinuevo, J.L.; Tom, B.; Ritchie, C.W. European Prevention of Alzheimer's Dementia Longitudinal Cohort Study (EPAD LCS): Study protocol. *BMJ Open* **2018**, *8*, e021017. [CrossRef] [PubMed]
4. German Center for Neurodegenerative Diseases. DELCODE: DZNE—Longitudinal Cognitive Impairment and Dementia Study. 2014. Available online: https://www.dzne.de/en/research/studies/clinical-studies/delcode/ (accessed on 30 June 2021).
5. German Center for Neurodegenerative Diseases. DESCRIBE: A DZNE Clinical Registry Study of Neurodegenerative Diseases—Longitudinal Cognitive Impairment and Dementia Study. 2015. Available online: https://www.dzne.de/en/research/studies/clinical-studies/describe/ (accessed on 8 May 2021).
6. Folstein, M.F.; Folstein, S.E.; McHugh, P.R. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **1975**, *12*, 189–198. [CrossRef]
7. Moms, J.C.; Heyman, A.; Mohs, R.C.; Hughes, J.P.; van Belle, G.; Fillenbaum, G.; Mellits, E.D.; Clark, C. The Consortium to Establish a Registry for Alzheimer's Disease CERAD. Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* **1998**, *39*, 1159–1165. [CrossRef] [PubMed]
8. Satzger, W.; Hampel, H.; Padberg, F.; Bürger, K.; Nolde, T.; Ingrassia, G.; Engel, R. Zur praktischen Anwendung der CERAD-Testbatterie als neuropsychologisches Demenzscreening. *Der Nervenarzt* **2001**, *72*, 196–203. [CrossRef] [PubMed]
9. World Health Organization *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*; World Health Organization: Geneva, Switzerland, 1993.
10. Chapman, W.W.; Bridewell, W.; Hanbury, P.; Cooper, G.F.; Buchanan, B.G. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *J. Biomed. Inform.* **2001**, *34*, 301–310. [CrossRef] [PubMed]
11. Dalianis, H. *Clinical Text Mining: Secondary Use of Electronic Patient Records*; Springer: Berlin/Heidelberg, Germany, 2018.
12. Uzuner, O.; South, B.R.; Shen, S.; DuVall, S.L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med Inform. Assoc.* **2011**, *18*, 552–556. [CrossRef] [PubMed]
13. Xu, H.; Stenner, S.P.; Doan, S.; Johnson, K.B.; Waitman, L.R.; Denny, J.C. MedEx: A medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 19–24. [CrossRef] [PubMed]
14. Fluck, J.; Senger, P.; Ziegler, W.; Claus, S.; Schwichtenberg, H. The cloud4health Project: Secondary Use of Clinical Data with Secure Cloud-Based Text Mining Services. In *Scientific Computing and Algorithms in Industrial Simulations: Projects and Products of Fraunhofer SCAI*; Griebel, M., Schüller, A., Schweitzer, M.A., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 285–315. [CrossRef]

15. Becker, M.; Kasper, S.; Böckmann, B.; Jöckel, K.H.; Virchow, I. Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. *Int. J. Med. Inform.* **2019**, *127*, 141–146. [CrossRef] [PubMed]
16. Harkema, H.; Dowling, J.N.; Thornblade, T.; Chapman, W.W. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J. Biomed. Inform.* **2009**, *42*, 839–851. [CrossRef] [PubMed]
17. Cotik, V.; Roller, R.; Xu, F.; Uszkoreit, H.; Budde, K.; Schmidt, D. Negation Detection in Clinical Reports Written in German. In Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), Osaka, Japan, 11–16 December 2016; The COLING 2016 Organizing Committee: Osaka, Japan, 2016; pp. 115–124.
18. Zhou, X.; Wang, Y.; Sohn, S.; Therneau, T.M.; Liu, H.; Knopman, D.S. Automatic extraction and assessment of lifestyle exposures for Alzheimer's disease using natural language processing. *Int. J. Med Inform.* **2019**, *130*, 103943. [CrossRef] [PubMed]
19. Shen, Z.; Yi, Y.; Bompelli, A.; Yu, F.; Wang, Y.; Zhang, R. Extracting Lifestyle Factors for Alzheimer's Disease from Clinical Notes Using Deep Learning with Weak Supervision. *arXiv* **2021**, arXiv:2101.09244.
20. Kumar, S.; Oh, I.; Schindler, S.; Lai, A.M.; Payne, P.R.O.; Gupta, A. Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: A systematic literature review. *J. Am. Med. Inform. Assoc. Open* **2021**, *4*, ooab052. [CrossRef] [PubMed]
21. Riepenhausen, S.; Varghese, J.; Neuhaus, P.; Storck, M.; Meidt, A.; Hegselmann, S.; Dugas, M. *Portal of Medical Data Models: Status 2018*; Studies in Health Technology and Informatics; IOS Press: Amsterdam, The Netherlands, 2019; pp. 239–240. [PubMed]
22. Stenetorp, P.; Pyysalo, S.; Topić, G.; Ohta, T.; Ananiadou, S.; Tsujii, J. brat: A Web-based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations Session at EACL 2012, Avignon, France, 23–27 April 2012; Association for Computational Linguistics: Avignon, France, 2012.
23. Klügl, P.; TOEPFER, M.; BECK, P.D.; Fette, G.; Puppe, F. UIMA Ruta: Rapid development of rule-based information extraction applications. *Nat. Lang. Eng.* **2014**, *22*, 1–40. [CrossRef]
24. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, 36, 1234–1240 [CrossRef] [PubMed]
25. Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.H.; Jindi, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 72–78. [CrossRef]
26. Johnson, A.E.; Pollard, T.J.; Shen, L.; wei H. Lehman, L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.A.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef] [PubMed]