Article published online: 2020-10-14

Designing an openEHR-Based Pipeline for Extracting and Standardizing Unstructured Clinical Data Using Natural Language Processing

Antje Wulff[1] Marcel Mast[1] Marcus Hassler[2] Sara Montag[1] Michael Marschollek[1] Thomas Jack[3]

[1] Peter L. Reichertz Institute for Medical Informatics, TU Braunschweig and Hannover Medical School, Hannover, Germany
[2] Econob, Informationsdienstleistungs GmbH, Klagenfurt am Wörthersee, Austria
[3] Department of Pediatric Cardiology and Intensive Care Medicine, Hannover Medical School, Hannover, Germany

Address for correspondence Antje Wulff, MSc, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Karl-Wiechert-Allee 3, 30625 Hannover, Germany (e-mail: antje.wulff@plri.de).

Methods Inf Med 2020;59:e64–e78.

Abstract

Background Merging disparate and heterogeneous datasets from clinical routine in a standardized and semantically enriched format to enable a multiple use of data also means incorporating unstructured data such as medical free texts. Although the extraction of structured data from texts, known as natural language processing (NLP), has been researched at least for the English language extensively, it is not enough to get a structured output in any format. NLP techniques need to be used together with clinical information standards such as openEHR to be able to reuse and exchange still unstructured data sensibly.

Objectives The aim of the study is to automatically extract crucial information from medical free texts and to transform this unstructured clinical data into a standardized and structured representation by designing and implementing an exemplary pipeline for the processing of pediatric medical histories.

Methods We constructed a pipeline that allows reusing medical free texts such as pediatric medical histories in a structured and standardized way by (1) selecting and modeling appropriate openEHR archetypes as standard clinical information models, (2) defining a German dictionary with crucial text markers serving as expert knowledge base for a NLP pipeline, and (3) creating mapping rules between the NLP output and the archetypes. The approach was evaluated in a first pilot study by using 50 manually annotated medical histories from the pediatric intensive care unit of the Hannover Medical School.

Results We successfully reused 24 existing international archetypes to represent the most crucial elements of unstructured pediatric medical histories in a standardized form. The self-developed NLP pipeline was constructed by defining 3.055 text marker entries, 132 text events, 66 regular expressions, and a text corpus consisting of 776 entries for automatic correction of spelling mistakes. A total of 123 mapping rules were implemented to transform the extracted snippets to an openEHR-based representation to be able to store them together with other structured data in an existing openEHR-based data repository. In the first evaluation, the NLP pipeline yielded 97% precision and 94% recall.

Keywords

- natural language processing
- clinical decision support systems
- openEHR
- pediatric intensive care
- medical history taking

received
May 12, 2020
accepted after revision
July 18, 2020

DOI https://doi.org/10.1055/s-0040-1716403.
ISSN 0026-1270.

Conclusion The use of NLP and openEHR archetypes was demonstrated as a viable approach for extracting and representing important information from pediatric medical histories in a structured and semantically enriched format. We designed a promising approach with potential to be generalized, and implemented a prototype that is extensible and reusable for other use cases concerning German medical free texts. In a long term, this will harness unstructured clinical data for further research purposes such as the design of clinical decision support systems. Together with structured data already integrated in openEHR-based representations, we aim at developing an interoperable openEHR-based application that is capable of automatically assessing a patient's risk status based on the patient's medical history at time of admission.

Introduction

Rationale and Background

Digitalization in medicine comes along with an increasing interest in the reuse of existing data sets for other purposes than originally intended. Today, the importance of reusing clinical data for improved health care is widely recognized.[1] However, not only the interest has risen but also the technical possibilities for integrating heterogeneous datasets have been expanded. While bringing data together in a *syntactical interoperable* way is one important building block of enabling enhanced reuse and exchange, in recent years, the awareness also rose toward forming a shared meaning of data across institutions and countries (*semantic interoperability*[2]). Nowadays, researchers work on the integration of data originating from various sources by using different clinical information standards such as openEHR,[3] HL7 FHIR,[4] HL7 V3 RIM,[5] HL7 CDA/CCR,[6] or HL7 VMR.[7] It can be observed that the primary goal of these research projects is often to harmonize datasets that are already available in a (semi)structured format but completely disparate. However, although this already is well-known as a challenging task, the next step must be the incorporation of unstructured data such as medical documents as these texts also carry crucial information for clinical care and research. Along with the increasing digitalization in medicine, these free texts are now electronically available and accessible. Although this is an improvement, it does not seem enough because the sole electronic availability is not necessarily associated with faster readability and information processing.[8] Clinicians and researchers "(…) spend considerable time reading free texts (…)"[9] which potentially hinders the everyday routine, moreover, the free text format is also not appropriate for a *multiple use* or an exchange of data. Consequently, there is a clear need of an approach for (1) extracting crucial information from such texts, and (2) representing the extracted data in a structured, semantically enriched way. Here, the use of natural language processing (NLP) techniques together with clinical information modeling standards might be appropriate. NLP can help to "(…) bridge the gap between textual and structured data, allowing humans to interact using familiar natural language while enabling computer applications to process data effectively."[8]

In the context of bringing NLP techniques together with clinical information standards to reach a structured representation of the NLP output, most recently Hong et al[10,11] presented an FHIR-based approach to standardize and structure texts from electronic health records (EHRs) by using existing NLP tools for the English language. For German, a related but not yet clinically evaluated attempt using FHIR is available.[12] Some older publications dealing with HL7 CDA for structuring texts such as discharge letters are available, too.[13,14] In terms of openEHR, Kropf et al[15] presented a way to structure a pathology report into sections represented by openEHR archetypes by a regular expression-based approach to enable section-sensitive queries on these texts. The work successfully shows the feasibility of transforming the general structure of a document into an openEHR-based representation and formulating semantic queries on previously unstructured pathology reports. However, the work is limited on only finding sections and is not underpinned by a full-pipe NLP approach possible of retrieving key items and storing them on entry-level in an openEHR template. Hence, to the best of our knowledge, recent publications have not developed an openEHR-based pipeline for extracting and standardizing unstructured clinical data to the extent as we intend to do. We aim at designing a new approach of seamlessly integrating NLP and openEHR for transferring unstructured documentation into standardized and semantically enriched data items using openEHR.

The Importance of Medical Histories

The feasibility of an openEHR-based pipeline for transformation of unstructured clinical data into standardized representations is tested on examples of pediatric medical histories as these texts bear an immense meaning in everyday routine of clinicians.

Medical practice in critical care is characterized by solving complex decision-making problems under challenging conditions of routine care such as critical situations, time pressure, and work interruptions.[16,17] The need for timely decision-making on diagnoses and early therapies especially gain in importance when critically ill patients are admitted. For an immediate impression of the patient's condition, medical interviews are performed and medical histories are composed. Back in 1975, Hampton et al already reported that in more than 82%

of cases the medical history provided sufficient information for an exact initial diagnosis.[18,19] Later, Peterson et al supported these findings by describing that 76% of medical histories contain crucial information that led to the final diagnosis.[20] Similar early findings on medical history research were presented by Keifenheim et al.[21] Today, the significance of this rather time consuming approach for diagnostics is being discussed as new innovative diagnostic technologies such as imaging methods or laboratory analyses are fast and accurate. However, a medical history contains a great diversity of heterogeneous information at an aggregated level, therefore, they are still recognized as highly valuable. Different researchers in several scenarios report on the significant meaning of medical histories, e.g., in geriatrics,[22] in ophthalmology,[23] in pediatrics,[24,25] and in the diagnosis of pneumonia.[26] Along with the increased digitalization and availability of patient's data in EHRs or patient data management systems (PDMS) in intensive care units, medical histories became available electronically. Although these reports are now easily accessible, there is no further support for faster clinical care as the health care professionals still need to review the entire report. There is a clear need for NLP-based solutions that are able to extract important information from unstructured medical histories. This alone already enables clinicians to assess the patient's situation more quickly at the time of admission. However, bringing structured and unstructured data together in a semantically enriched and unambiguous manner, thus sensibly brings the chance to reuse heterogeneous data for further purposes in research and patient care. In the context of medical histories, this would open up the possibility of developing helpful risk scoring applications (comparable to the widely used pediatric mortality and morbidity scores such as PIM II [pediatric index of mortality] and PRISM III [pediatric risk of mortality]). Automatic generating of a reliable morbidity and mortality score based on medical history analysis could be an innovative and valuable tool for clinicians in their daily routine.

Objectives

We aim at developing an approach to automatically extract crucial information from medical free texts and to transform this unstructured clinical data by using NLP into a standardized and structured openEHR-based representation. Therefore, we designed and implemented an exemplary pipeline for the processing of pediatric medical histories.

Methods

openEHR

For structured representation of extracted information, we adopted the openEHR approach as semantic modeling methodology and interoperability standard. In openEHR, a clear separation of technical and domain content is realized by following a multilevel modeling approach. The underlying reference model provides the basis for any software implementation of openEHR by describing standardized definitions of structures, data types, and functions (first level of modeling). The further levels consider the formal definition of clinical concepts and use cases as data models, regardless of the technical implementation. By applying constraints on the openEHR reference model, clinical concepts such as a *diagnosis* or a *laboratory result* are modeled as machine-readable and computable but predominantly domain-level concept definitions called *archetypes* [3]. Consequently, archetypes are often developed in close cooperation with medical domain experts. All attributes, characteristics, data structures, and internal or external terminologies relevant for the clinical concept are defined and bound within archetypes by using the Archetype Definition Language. Archetypes are then reused and nested in so-called *templates*[3,27] to represent specific use cases. Typically, templates express entire clinical documents containing different information modeled as several archetypes such as *discharge letters*, *result reports,* or *medical histories*. The multilevel modeling approach allows for exchanging archetypes between all institutions implementing the openEHR reference model and reusing archetypes without in-depth technical understanding of the underlying persistence structure of the data repository implemented. Different implementations of the openEHR reference model that can used as data repository are available.[28–31] To retrieve data from an openEHR-based data repository, a semantically enriched query language called Archetype Query Language (AQL)[a] is provided. As long as the same archetypes are used to represent the same clinical concepts, these queries will work in any openEHR implementation.

To allow the reusability of our data models, applications and results, we strive for using existing archetypes as much as possible. Hence, when designing archetypes for representing a patient's medical history, we first reviewed existing archetypes from a global and freely accessible archetype repository (*Clinical Knowledge Manager*, CKM[b]). Since not all contents have already been modeled, we also might need new archetypes. Of course, we aim at providing our new models to the international CKM to contribute to the global openEHR activities. The archetypes are selected and designed in close cooperation with domain experts such as the clinicians from our pediatric intensive care unit. To structure and monitor our modeling processes, we take advantage of an existing clinical knowledge governance framework that we designed for the purpose of openEHR modeling in a nationwide data infrastructure project. All other openEHR related projects in our department are aligned to this governance process. To learn more about the details of our modeling activities, including IT tools used and modeler roles defined, we refer to Wulff et al.[32]

Natural Language Processing

Free text documentation seems to be very common in clinical practice. The use of natural language is not only more convenient for clinicians, but it also includes various means of expressions that could reflect the complexity and diversity of clinical cases.[33] However, it is a well-known bottleneck for computer-aided processing and utilization of free texts due to the crucial point that equivalent information can be

[a] http://www.openehr.org/releases/QUERY/latest/docs/AQL/AQL.html.

[b] https://www.openehr.org/ckm/.

represented by a large variety of words and grammatical structures.[8] Tackling this challenge is one of the main tasks of NLP. From our perspective, Dubitzky et al provides a complex, but accurate definition of NLP that we bear in mind during our work: "NLP is the analysis of linguistic data, most commonly in the form of textual data such as documents or publications, using computational methods. The goal of NLP is generally to build a representation of the text that adds structure to the unstructured natural language, by taking advantage of insights from linguistics. This structure can be syntactic in nature, capturing the grammatical relationships among constituents of the text, or more semantic, capturing the meaning conveyed by the text."[34]

Knowledge Acquisition for Pipeline Construction

As suggested by Friedman et al,[35] the development of NLP systems requires corpora for training, a domain model, and a domain as well as a linguistic knowledge. Hence, we decided to work closely together with experienced clinicians and researchers from the Department of Pediatric Cardiology and Intensive Care Medicine from the Hannover Medical School. By regularly meeting and interviewing these experts, we were able to define the most important information from medical histories. With this knowledge, we were able to construct a dictionary that summarized various clinical markers and events. In addition, operational aspects such as the selection of methods, tools, and systems play a major role in the design of NLP applications.[35]

NLP Pipeline Components

For our work, we have built an NLP pipeline of well-known components such as morphological analysis, part-of-speech tagging, syntactic, semantic and pragmatic analysis. Instead of developing new procedures, we decided to reuse and apply existing methods and algorithms such as statistical methods, linguistic rules, and regular expressions.

For extracting crucial information from pediatric medical histories, an NLP process consisting of five successive tasks was developed. The first step describes the segmentation of the medical history into various morphemes such as roots, prefixes, and suffixes (morphological analysis). Thereby, the words included in the text are analyzed by having a look at their generic structure. We implemented the morpheme segmentation by using finite-state machines.[8] In a second step, the segmented morphemes need to be tagged by a so-called part-of-speech tagging (POS tagging) task. Here, the recognized words were marked and identified as belonging to a specific category of words (part of speech) such as preposition or noun. Moreover, we performed an additional step to the classical POS tagging by adding or removing spaces to gain a standardized punctuation within the output, improving the quality of the resulting tags and the following steps. In a third step, the syntactical structure of the tagged words included in the phrase must be analyzed (syntactic analysis). We implemented a backtracking parser to extract the syntactic structure of the input and to represent it by using parsing trees.[8] By this task, the component is capable of understanding the location and relationship of the words included in the recognized sentence. After performing the syntactic analysis, the fourth step comprises the task of semantic analysis to be able to understand the meaning of the sentence. Here, well-known semantic patterns of the language are bailed-in for better understanding the combination of words to find out the semantic meaning of the whole sentence. We based our semantic analysis on the so-called Montague Semantics.[36] The fifth step represents the task of pragmatic analysis in which not only the plain lexical meaning is considered but also the discursive meaning of the statement. To be able to extract crucial information, the clinically relevant artifacts have to be defined. In our context, these artifacts were determined through an enhanced requirement analysis, expert interviews, and a literature review. Here, we implemented the idea of marker concepts. A marker concept consists of various collections of entries, called marker, that represent clinically relevant artifacts to be extracted during the NLP process. The occurrence of at least one but also multiple marker entries predefine events. An occurrence can either be a single entry from one marker concept or a combination of different entries originating from other marker concepts.

Data, Materials, and Tools

OpenEHR Modeling Tools

For modeling openEHR archetypes and templates, we used the Archetype Editor 2.8 and the Template Designer 2.8 from Ocean Informatics.[c] For retrieving existing archetypes from the international openEHR community, we accessed the international Clinical Knowledge Manager (CKM)[d]. Furthermore, for building our local and project-specific set of reused and newly created archetypes and starting specific review rounds with our experts, we reused a national version of the CKM[e] that was implemented previously for a nationwide data infrastructure project in Germany.[37] This instance is linked with the international CKM so that all existing archetypes are directly referenced. All archetypes and templates used for this project are available in the CKM.

OpenEHR Data Repository

In our work, we use an existing openEHR-based data repository, which has been used for related research projects before.[37–40] Currently, the platform which is separated in two instances (research and patient care) is continuously filled with data needed in the context of a nationwide data infrastructure project called HiGHmed.[37] It builds the technical basis of the so called medical data integration center of the Hannover Medical School[f]. The repository is based on the *better platform* by Marand[g] and is used together with various commercial but also self-developed mapping and integration

[c] https://www.openehr.org/downloads/modellingtools/.
[d] https://www.openehr.org/ckm/.
[e] https://ckm.highmed.org/ckm/.
[f] https://www.mhh.de/forschungseinrichtungen/medic/.
[g] http://www.better.care/.

tools for transferring primary source data to this openEHR-based data repository. Currently, these tools are only able to integrate structured data from primary source systems. Hence, because no unstructured, free text can be treated as input source, medical histories could not be integrated up to now.

Data Source and Access

The platform already stores some datasets from different local primary source systems, e.g., the electronic medical record (i. s. h. med), which are available in a structured format. In a previous project, we already tested the integration of structured intensive care data from the PDMS of the pediatric intensive care unit of the Hannover Medical School (m.life and the legacy system COPRA).[38–40] The medical histories used within this project originate from the same PDMS. For data safety concerns, the medical histories are used in an anonymized form by removing or modifying sensible data manually.

NLP Tools

For our work, we used *LingRep*, provided by econob,[h] as exemplary NLP application because it offers a sample pipeline of different well-known methods and components required for our application as well as a high flexibility in the individual adaptation and extension of the pipeline. *LingRep* has not been used in medical contexts yet.

Workflow Design

The workflow is realized in a Java-based application that consists of an *input module* to load all relevant settings, dictionaries as well as the medical histories as free text in a text format. Before starting the pipeline, the *spelling correction module* is passed. By implementing a REST client, the LingRep configuration can be accessed and the NLP pipeline configured by our previously designed *marker dictionary* can be started. The output format of the NLP pipeline from LingRep is a JSON file that is transferred to the *mapping module* of our application. The mapping module performs the interpretation of the extracted NLP snippet and the assignment to the items of the openEHR archetypes. By using the REST interface of our data repository, the *integration module* loads the datasets into our platform. A *querying module* can be used to access the integrated datasets by using AQL.

Evaluation

To evaluate the feasibility of the NLP pipeline, a proof-of-concept evaluation was conducted. The prototype was evaluated by retrieving 50 anonymized randomly chosen medical histories from the pediatric intensive care unit (anonymization was performed by modifying sensible data manually). These medical histories were transferred to a structured openEHR-based representation by running through the designed pipeline to get finally stored in the openEHR-based data repository. According to the defined dictionaries, two independent reviewers with a medical informatics background extracted information related to the defined marker concepts from these medical histories. In case of disagreement, a third reviewer was involved to reach a final set of extracted events. The manually extracted information snippets were compared with the results of the automated extraction process by the NLP pipeline to determine precision and recall. To evaluate the viability of the prototypical workflow implementation for transforming data into an openEHR-based representation, we queried all data elements available in the openEHR data repository after executing the entire workflow. By using the querying module of our prototypical application, we evaluated the existence of all extracted information snippets and their assignment to a suitable archetype.

Ethical Considerations

This manuscript does not contain research involving human subjects.

Results

Archetypes for Information Representation

For representing the extracted information in a structured and semantically enriched format, we constructed an openEHR template nesting all relevant marker concepts as archetypes. As shown in ►**Table 1**, we were able to reuse 23 archetypes from the international CKM. One archetype defining the admission details of the patient was designed from the ground (see ►**Supplementary Appendix A.2**, available in the online version). The process of selecting or newly creating archetypes is crucial to be able to transform the information extracted from the unstructured text by the NLP components into a harmonized and standardized data representation. Only if appropriate archetypes are available, it is possible to start the process of mapping the extracted information snippets to the final representation. A brief overview of the developed template is given in ►**Fig. 1**.

Marker Dictionary

Currently, our dictionary contains 19 marker concepts, 60 markers, 3.055 marker entries, 132 , and 66 regular expressions.

Marker Concepts

In cooperation with experienced pediatricians, 19 different concepts, each representing highly relevant aspects occurring in medical histories, were created (a schematic representation is given in ►**Fig. 2**).[2] These include nonclinical marker concepts as unit-, negation or date-concepts, patient-specific marker concepts as medication-, diagnosis-, allergy-, or general patient's condition-concepts, and systemic marker concepts as skin-, body temperature-, respiration, or heart-concepts. Each of the concepts are further described by markers and their attributes, e.g., the skin concept contains entries describing the coloring of the skin ("blass" [pale skin], "rosig" [rosy skin]) or the patient's condition concept comprises items characterizing the patient's state as

[h] http://www.econob.com.

Table 1 Overview of the openEHR archetypes used for representing medical history data

Concept name	Archetype ID	Internationally available?
Adverse reaction risk	EVALUATION.adverse_reaction_risk.v1[1]	Yes – published
Age	OBSERVATION.age.v0[2]	Yes—Draft
Blood pressure	OBSERVATION.blood_pressure.v2[3]	Yes—published
Body temperature	OBSERVATION.body_temperature.v2[4]	Yes—published
Capillary refill	CLUSTER.capillary_refill_time.v0[5]	Yes—draft
Dosage	CLUSTER.dosage.v1[6]	Yes—published
Examination of abdomen	CLUSTER.exam_abdomen.v0[7]	Yes—draft
Examination of a pupil	CLUSTER.exam-pupil.v0[8]	Yes—draft
Examination of skin	CLUSTER.exam_skin.v0[9]	Yes—draft
Family history	EVALUATION.family_history.v2[10]	Yes—published
Food and nutrition summary	EVALUATION.nutrition_summary.v0[11]	Yes—draft
Gender	EVALUATION.gender.v1[12]	Yes—Published
Laboratory test result	OBSERVATION.laboratory_test_result.v1[13]	Yes—Published
Medication management	ACTION.medication.v1[14]	Yes—Draft
Pediatric Glasgow Coma Scale (pGCS)	OBSERVATION.glasgow_coma_scale_pediatric.v0[15]	Yes—Draft
Physical examination findings	OBSERVATION.exam.v1[16]	Yes—Published
Problem/Diagnosis	EVALUATION.problem_diagnosis.v1[17]	Yes—Published
Pulse/Heart beat	OBSERVATION.pulse.v2[18]	Yes—Published
Pulse oximetry	OBSERVATION.pulse_oximetry.v1[19]	Yes—Published
Report	COMPOSITION.report.v1[20]	Yes—Published
Respiration	OBSERVATION.respiration.v2[21]	Yes—Published
Story/History	OBSERVATION.story.v1[22]	Yes—Published
Symptom/Sign	CLUSTER.symptom_sign.v1[23]	Yes—Published
Patient admission	ADMIN_ENTRY.admission.v0	No

[1]https://ckm.openehr.org/ckm/#showArchetype_1013.1.1713; [2]https://ckm.openehr.org/ckm/#showArchetype_1013.1.3361; [3]https://ckm.openehr.org/ckm/#showArchetype_1013.1.3574; [4]https://ckm.openehr.org/ckm/#showArchetype_1013.1.2796; [5]https://ckm.openehr.org/ckm/#showArchetype_1013.1.3319; [6]https://ckm.openehr.org/ckm/#showArchetype_1013.1.2751; [7]https://ckm.openehr.org/ckm/#showArchetype_1013.1.219; [8]https://ckm.openehr.org/ckm/#showArchetype_1013.1.3882; [9]https://ckm.openehr.org/ckm/#showArchetype_1013.1.3933; [10]https://ckm.openehr.org/ckm/#showArchetype_1013.1.2469; [11]https://ckm.openehr.org/ckm/#showArchetype_1013.1.2755; [12]https://ckm.openehr.org/ckm/#showArchetype_1013.1.3715; [13]https://ckm.openehr.org/ckm/#showArchetype_1013.1.2191; [14]https://ckm.openehr.org/ckm/#showArchetype_1013.1.123; [15]https://ckm.openehr.org/ckm/#showArchetype_1013.1.4188; [16]https://ckm.openehr.org/ckm/#showArchetype_1013.1.271; [17]https://ckm.openehr.org/ckm/#showArchetype_1013.1.169; [18]https://ckm.openehr.org/ckm/#showArchetype_1013.1.4295; [19]https://ckm.openehr.org/ckm/#showArchetype_1013.1.3084; [20]https://ckm.openehr.org/ckm/#showArchetype_1013.1.677; [21]https://ckm.openehr.org/ckm/#showArchetype_1013.1.4218; [22]https://ckm.openehr.org/ckm/#showArchetype_1013.1.68; [23]https://ckm.openehr.org/ckm/#showArchetype_1013.1.195.

"kompensiert" [patient is hemodynamically compensated] or "schläfrig" [patient is somnolent].

Marker Events

The occurrence of at least one but also multiple marker entries predefine events. An occurrence can either be a single entry from one marker concept such as "Tachykardie"[tachycardia] or a combination of different entries originating from other marker concepts (►**Fig. 2**). One common example is the connection of one marker entry from the systemic marker concepts as "Herzfrequenz" [heart rate] with another marker entry as "hoch" [high]. The latter is related to another marker concept called adjective concept. Consequently, it is possible to combine different marker concepts to define events.

Regular Expressions

For numeric values such as in any prescription of medications (e.g., "50" in "50 mg") or dates, we designed regular expressions.

Spelling Correction Module

The developed spelling correction module was constructed by using the developed marker concepts, a list of approximately 300,000 German words and our available medical histories. The final module consists of approximately 776 entries relevant for our use case. For each entry, a list of spelling mistakes occurred in the medical histories is stored. To consider a yet unknown word as a potential misspelling of a relevant marker, the word is checked against the list of all known German words. In case of mismatching against this list, the word will be added as misspelling to our 776 entries. To assign this word as a

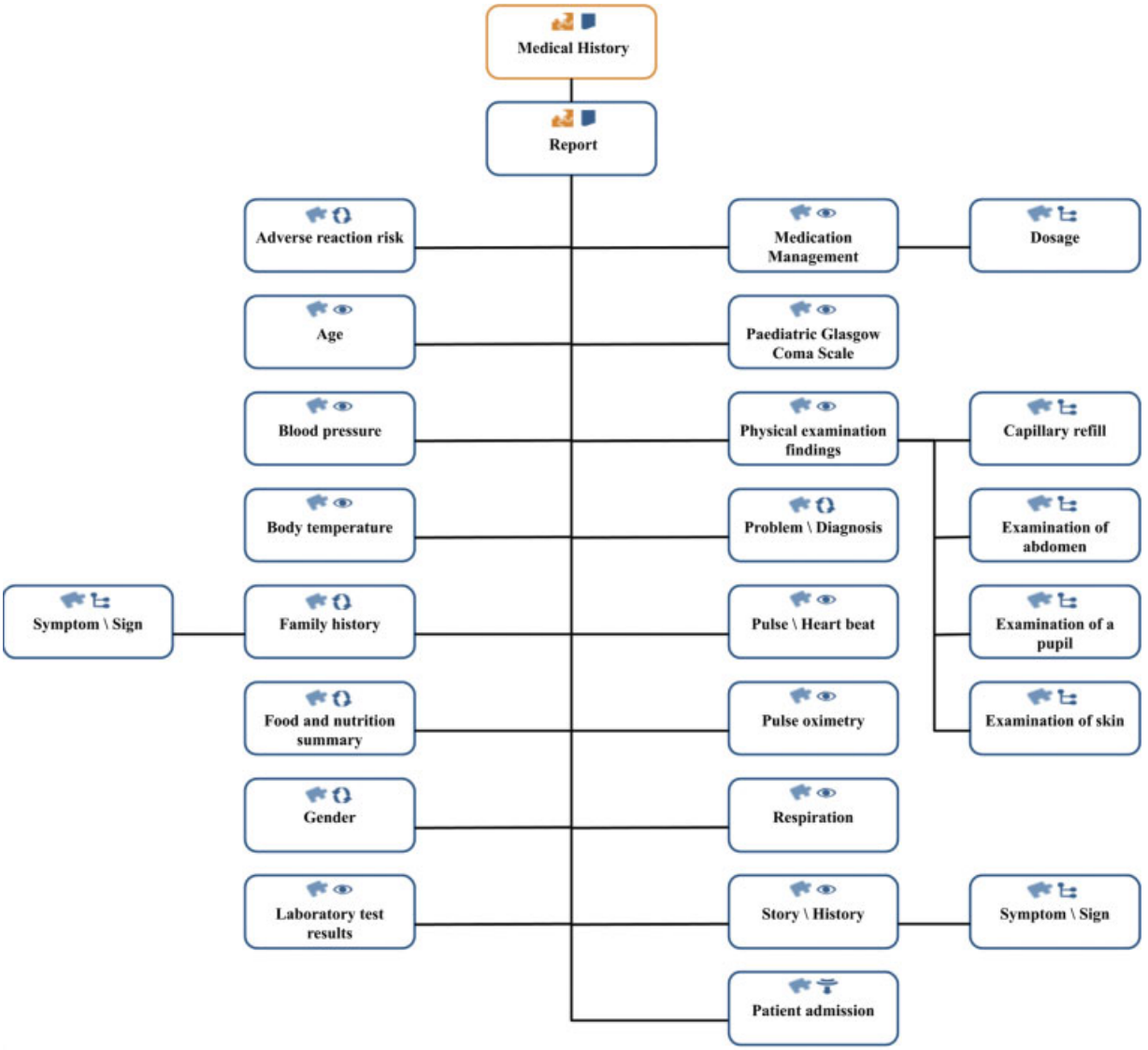

Fig. 1 OpenEHR template for representing a pediatric medical history.

misspelling to an existing entry, different similarity measures, including the Damerau–Levenshtein distance,[41,42] the Jaccard similarity coefficient, and the Soundex algorithm[43] are calculated. Depending on the word length and each calculated similarity measure, words can be matched. To reach a match, the similarity values calculated need to be higher than the values listed in ►**Supplementary Appendix A.1** (available in the online version). Based on this module, known misspelling words can be corrected before passing to the NLP pipeline and unknown misspelling words can either be handled as not relevant for our use case or added as another misspelling to our list.

Mapping and Integration Module

By connecting the NLP pipeline with the openEHR template, it is possible to extract crucial information from an unstructured medical history and integrate the extracted data into an openEHR-based data repository. Therefore, we defined a prototypical workflow and designed a Java-based application. Depending on its content and a unique event identifier, the extracted information is mapped to the item of the corresponding openEHR archetype (►**Fig. 2**).[4] ►**Figure 3** presents the mapping process within the Java code on the example of the age event. The age event is provided as an output from the NLP pipeline together with a unique identifier "2106." All possible events were converted to 123 mapping rules defined in a switch-case method. The methods called within this rules enable the generation of instances of the corresponding archetype. To be able to create a new archetype object and setting its values, the overall medical history template was imported and generated as Java class before. The eventObject carrying the extracted information snippet is processed within the called method by setting its content as value of the corresponding archetype attribute. For each unique archetype path, a specific setter method can be used.

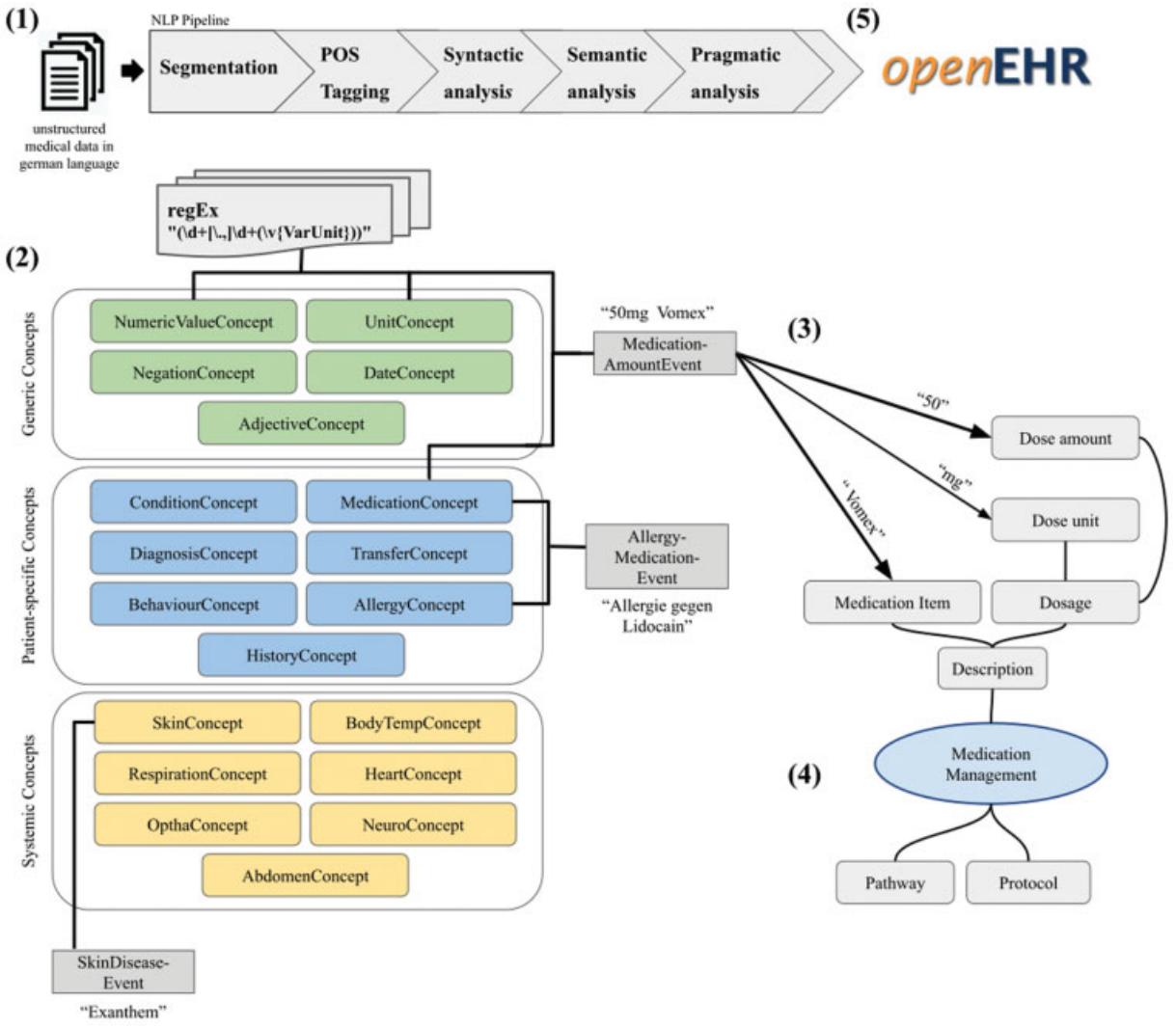

Fig. 2 Schematic representation of the developed workflow, including (1) the input module, (2) the marker concepts and regular expressions realized in the NLP pipeline module, (3) the process of mapping to the (4) an archetype nested in the openEHR medical history template stored in the (5) openEHR-based data repository. NLP, natural language processing.

Example Workflow

To demonstrate our workflow, we use the following fictional medical history.

"Die Patientin, 10 Jahre alt, wurde aus Klinikum Musterstadt verlegt. Patient blass, klagt seit 5 Tagen über Erbrechen und Kopfschmerzen; 39.7°C Körpertemperatur, Herzfrequenz bei 130. Pupillen eng, Abdomen weich. Vorherig bestand Lungenentzündung, Sauerstoffsättigung bei 82%, Rekapillarisierungszeit <2 Sekunden. Allergie gegen Latex. Familiär bekannter Immundefekt. Familiär D84. Nach Gabe von 50mg Vomex kein Erbrechen mehr."

[The patient, 10 years old, was transferred from another hospital. Patient pale, complaining of vomiting and headache for 5 days; 39.7°C body temperature, heart rate at 130. Pupils are narrow, abdomen soft. Previously there was pneumonia, oxygen saturation at 82%, capillary refill time <2 seconds. Allergy to latex. Familially known immunodeficiency. Familial D84. No more vomiting after administration of 50-mg Vomex.]

In a first step, the medical history was loaded into the NLP pipeline. Then, the text passed the NLP pipeline. During that process, all relevant information were extracted. For the aforementioned exemplary medical history, the pipeline extracted 32 events (e.g. "10 Jahre alt" [*10 years old*]). The third step of the workflow comprises the mapping of the extracted components to the archetypes by using the unique paths and so called *at-codes* that identify the items of an archetype. Depending on internal identifiers for every defined event within the NLP pipeline, extracted information can uniquely be categorized and mapped onto the archetype. For example, events with the identifier "2106" will always contain information related to the patient's age and, thus, will always be mapped onto the corresponding *age* archetype path. During this process, some contradictory or overlaying information was detected. In that case, we decided to integrate the component carrying the most detailed information. For example, a component describing "body temperature" with a specific value as "39.7°C" would be preferred over a more unspecific component consisting of the snippet "high body temperature."

A special case is the extraction of negated information such as "no headache." Here, the pipeline would extract both

Fig. 3 Snippet from the Java code for mapping the extracted information snippet on unique archetype paths (mapping and integration module), including (1) running through all defined rules and the firing of a suitable rule which then (2) enables the instantiation of a new age observation by filling the associated archetype paths with the extracted information delivered in the eventObject.

"headache" and "no headache" because the two words are handled as both two separate markers and one event. To prevent the integration of contradictory information, in this case, the negated information will be preferred. Because of the described contradictory or overlapping components, 18 of 32 extracted snippets were mapped onto archetypes and, in a fourth step, integrated into an openEHR-based data repository.

As a result, all information extracted from the pipeline should be available and, hence, queryable. Therefore, in the last step, we successfully retrieved the integrated datasets by using AQL. An exemplary query used to access the datasets stored in a specific composition is constructed as follows:

SELECT a
FROM EHR e
CONTAINS COMPOSITION a
WHERE a/uid/value = "986f1cc6–0709–47e6-b6e8–-6a065263c8fd::NLP::1."

The last line of the query contains the identifier of the chosen medical history report.

As a result, the text snippets representing the most important information of the medical history were successfully retrieved (►**Table 2**).

Evaluation

The proof-of-concept evaluation resulted in 529 manually extracted events, which were compared with the results of the automated extraction process by the NLP pipeline. The pipeline correctly extracted 499 concepts (true positives), wrongly identified 16 concepts (false positives), and missed 30 concepts (false negatives) (►**Table 3**). This yielded to a precision of 96.89% and a recall of 94.32%.

The 529 extracted ground truth events contain 81 events which were clearly understandable but misspelled in the raw input. In a first evaluation approach, none of these events were extracted. After implementation of the spelling correction module, 69 of the 81 misspelled events were successfully extracted. Without the spelling correction component, the misspelled events would have been treated as false negatives (recall of 81.29%).

Discussion

We designed an approach to extract important information from German medical free texts and to transform it into a structured openEHR representation on the example of pediatric medical histories.

Design and Evaluation of a Prototypical openEHR-Based Pipeline

By following the openEHR approach, we were able to represent the extracted information in a structured, semantically enriched and computable format. We have successfully represented all marker concepts as 24 archetypes, and the entire medical history as one template that contains all archetypes. We strived for reusing as many archetypes from the international CKM as possible. This resulted in just one newly created admission archetype which has been designed in close cooperation with clinical, technical, and international modeling experts (see ►**Supplementary Appendix A.2** , available in the online version). However, since we focused on the technical feasibility of the overall approach, some archetype selections should be reconsidered from a semantic point of view which might include a

Table 2 Results of the AQL query to retrieve extracted information snippets

Event ID	Snippet, extracted from pipeline	Archetype	Archetype path and archetype term code (at-code)	
2107	Patientin [patient, female]	Gender	Administrative gender *at0022*	Patientin [patient, female]
2106	10 Jahre alt [10 years old]	Age	Chronological age *at0004*	P10Y
			Comment *at0006*	10 Jahre alt [10 y old]
2104	Klinikum Musterstadt [Hospital Musterstadt]	Patient admission	Type of admission *at0049*	Klinikum Musterstadt [Hospital Musterstadt]
3103	Patient blass [pale patient]	Physical examination findings	Clinical description *at0015*	Patient blass [pale patient]
2101	Erbrechen [vomiting]	Problem/Diagnosis	Problem/Diagnosis name *at0002*	Erbrechen [vomiting]
2101	Kopfschmerzen [headache]	Problem/Diagnosis	Problem/Diagnosis name *at0002*	Kopfschmerzen [headache]
3206	39.7°C Körpertemperatur [39.7°C body temperature]	Body temperature	Temperature *at0004*	39.7 Cel
3411	Herzfrequenz bei 130 [heart rate at 130]	Pulse/Heart beat	Pulse rate *at0004*	130 bpm
3503	Pupillen eng [pupils are narrow]	Physical examination findings	Clinical description *at0003*	Pupillen eng [pupils are narrow]
3701	Abdomen weich [soft abdomen]	Physical examination findings	Clinical description *at0003*	Abdomen weich [soft abdomen]
2710	Vorherig bestand Lungenentzündung [previously existing pneumonia]	Story/History	Story *at0004*	Vorherig
			Symptom/Sign name *at0001*	Lungenentzündung [pneumonia]
3303	Sauerstoffsättigung bei 82% [oxygen saturation at 82%]	Pulse oximetry	SpO_2 *at0006*	82.0
3404	Rekapillarisierungszeit < 2 Sekunden [capillary refill time <2 seconds]	Capillary refill	Capillary refill time *at0026*	Less than 2 s
2502	Allergie gegen Latex [allergy to latex]	Adverse reaction risk	Category *at0120*	Allergie [allergy]
			Substance *at0002*	Latex
2705	Familiär bekannter Immundefekt [family history: immune deficiency]	Family history	Symptom/Sign name at *at0001*	Immundefekt [immun deficiency]
2707	Familiär D84 [familial D84]	Family history	Symptom/Sign name at *at0001*	D84
2202	50 mg Vomex	Medication management	Medication item *at0020*	Vomex
			Dose amount *at0144*	50.0
			Dose unit *at0145*	mg
2101	Kein Erbrechen [no more vomiting]	Problem/Diagnosis	Problem/Diagnosis name *at0002*	Kein Erbrechen [no more vomiting]

Abbreviation: AQL, Archetype Query Language.

conduction of cross-institutional and international expert review rounds. For example, the representation of medication use has always been a highly discussed concept. In our template, we only retrieve the medication a patient is taking at the time of admission or shortly before, e.g., a medication directly administered at admission. However, medical histories often also contain information about former medications which then should be transferred into a different archetype, e.g., openEHR-EHR-EVALUATION. medication_summary.v0. The same case might occur when looking into problems and diagnoses: there also might be current diagnoses and former diagnoses that already have been resolved. For representing all diagnoses a patient suffered during his life, an additional problem list

Table 3 Overview of the types of marker concepts identified within the manual annotation (ground truth) and the distribution of true positives, false negatives, and false positives

Type of marker concept	Number of events extracted (ground truth)	True positives	False negatives	False positives
Summary	*529*	*499*	*30*	*16*
Vital signs	190	168	22	7
Diagnosis	107	103	4	4
General condition and behavior	90	87	3	2
Skin characteristics	50	50	0	0
Abdomen characteristics	25	25	0	0
Medication	22	22	0	0
Special situations (e.g., transfer, emergency)	19	18	1	1
Ophthalmology	13	13	0	0
Neurology	8	8	0	0
Allergies	5	5	0	2

(openEHR-EHR-COMPOSITION.problem_list.v1) would be a good choice.

Furthermore, some of our defined markers might be already available in a structured and higher quality form, e.g., in an EHR. In some cases, it might be useful to rather use this structured data than extracting this from a medical history. Examples are birth data, gender, laboratory results, or standardized scores such as the Glasgow Coma Scale (GCS). However, when accessing this information from structured elements of the EHR, we still need to design or choose appropriate archetypes for them. Consequently, only the primary source will change and we still can use our openEHR template for representing the pediatric medical history.

With our exemplary integration into an openEHR-based data repository, we have successfully demonstrated the technical viability of transforming unstructured, free text into an interoperable openEHR format. Although the focus was on medical histories from the pediatric intensive care unit, we are confident that our workflow will be more generic and applicable in other contexts as the choice of archetypes and mapping rules does not strongly affect the overall methodological pipeline approach. With regard to the implemented assignments, some extensions are conceivable such as the consideration of times of measurements or the storage of the corresponding original phrase from which the concept was extracted (e.g., within the openEHR feeder audit[i]). The latter could improve transparency and understanding of the extraction process. Furthermore, there is the possibility that *different entries are extracted for the same marker or archetype*. If there is a clinical relevance, the template should allow multiple instances of one archetype to be stored. It may also be worth considering an integration of plausibility checks to decide which fact is the most important (e.g., in case of a co-occurrence of a normal and an abnormal temperature, the latter is used). A similar approach has already been considered in the treatment of *negations* and *overlaying information*. As explained above, if the same marker occurs without and with a negation, we will prefer to integrate the negation. For any case in which information snippets from different marker concepts are contradictory from a clinical perspective some expert rules will be needed to make an adequate decision. This would be a future development step since this case is not covered currently.

Evaluation Results

In the context of the conducted evaluation, 16 events were marked as *false positives*. These events contain a combination of multiple markers. All 16 false positives occurred due to a mix-up of the markers as seen in the following example: "[...] 70% FiO2 [...]. Later, 30% FiO2 [...]." The numerical values closest to the respective "FiO2" should be matched together to form an event. However, currently, the extracted events were built by cross-matching the numerical values and markers. Although the overall interpretation is not wrong, because the same marker is used, the matching process is not correct, leading to both, two false positives and false negatives. Hence, 16 of the total 30 *false negatives* resulted indirectly from the extraction of false positives, leaving 14 to be considered as new errors. Of these 14 false negatives, 12 resulted due to not corrected misspellings in the spelling correction step as mentioned above. However, although the spelling correction module was not capable of correcting these 12 events, it is again worth mentioning that the implementation of the spelling correction module clearly optimized the previous results by correcting 69 out of 81 misspelled events. This led to an improvement in the recall from 81.3 to 94.3%. The remaining two false negatives are due to insufficient built regular expressions during the dictionary construction step. Consequently, the spelling correction module and the regular expressions need to be optimized. For the false positives, it seems like the applied

[i] https://specifications.openehr.org/releases/RM/latest/common.html#_feeder_audit_class.

distance-based strategy explained above is not adequate since all false positives occurred due to a mix-up within the event construction step of the NLP component. It might be a promising approach to take even more the syntactic structure of the sentence into consideration (syntactical analysis step).

The overall performance of the pipeline in terms of the processing speed at runtime was satisfying (<1 minute for processing of all 50 medical histories). Furthermore there were no technical performance issues that can be inferred to the amount of marker and event concepts. In future work, standardized performance and speed tests at runtime should be performed.

Related Work

Research on the use of NLP techniques in health-related contexts has increased significantly in recent years. Many literature reviews, each focusing a slightly different topic, have been published in the last 2 years, such as a summary of current approaches to identify sections within clinical narratives from EHRs (published by Pomares-Quimbaya et al in 2019,[44] a review of recent publications on clinical information extraction applications (published by Wang et al in 2018),[45] an overview of published articles discussing the application of NLP techniques for mining health-related information not only from EHRs but also from social media (published by Gonzalez-Hernandez et al in 2017),[46] and a presentation of opportunities and challenges for clinical NLP in languages other than English (published by Névéol et al in 2018).[47]

Of course, also some commercial and noncommercial NLP tools exist that enable either the construction of a complete pipeline, or the completion of some specific tasks. For the former, and with a focus on the German language, mEX as an information extraction platform for German medical texts[48] as well as the well-known Mayo clinical text analysis and knowledge extraction system Apache cTAKEs[49] are worth mentioning. Furthermore, Averbis Health Discovery as a commercial product for analyzing medical texts has gained attention in the last years.[50] OpenNLP[51] or LingRep[52] are other examples for such full pipeline-oriented tools.

For the latter, MedXN is an open source tool for extracting and normalizing medication snippets from clinical texts,[53] MedTime for the extraction of temporal information[54] and POS taggers such as the Stuttgart-Tübingen-Tagset are available (also for the German language) for supporting specific NLP tasks. Tools for detecting abbreviations (Schwartz Hearst algorithm[55]) and negations (e.g., NegEx[56]) also fall into this category. However, the majority of the existing approaches focus on the English language as for example MedLEE as a natural language text extraction system for the medical domain, MetaMap as a tool to map biomedical text to the unified medical language system (UMLS), and caTIES as an application for extracting cancer information from clinical reports. While the research in the English-speaking world is ongoing in this field,[9] there is a lack of related work in German. However, the work presented by Becker and Böckmann[57] is notable, because the authors used a customized NLP pipeline with the help of cTAKES for German Language to extract UMLS concepts from clinical notes and to map these with SNOMED-CT codes. Although they only reached a moderate F1 measure, the results are promising because they reached these results without implementing German stemming. They even were able to further optimize this approach and evaluate it again in a clinical-driven use case of colorectal cancer with an improved F1 score of 81%.[58] A second notable approach for extracting information from German medical free text documents is provided by König et al.[59] The authors used NLP methods for the detection of clinical events with a precision of 95.6% and a recall of 96.7%. Within this work, the focus was mainly on two single concepts and could therefore be a promising approach to be integrated into a more holistic work. A third publication for extraction with NLP methods from a German source was published by Löpprich et al.[60]

Regardless of the tool used, it seems to be necessary to customize the NLP pipeline in terms of the concrete use case to reach satisfying results in clinical-driven evaluations. Existing NLP tools and already implemented NLP techniques and tasks (e.g., POS tagging) are very helpful but they always need a customization to reach the desired output in the specific medical use case. The modification and development process of German dictionaries and corpora are very time consuming and experts need to be involved. If done precisely, the resulting German markers carry great potential to be reused in other tools or other settings. Hence, in our work, we put a lot of effort into developing a specialized German dictionary (including markers, events, and regular expressions) for pediatric medical histories.

Some related work is available for using semantic interoperability standards for capturing former unstructured information from medical free texts in a structured format. Hong et al[11] present the development of an FHIR-based clinical data normalization pipeline for standardization and integration of unstructured and structured EHR data. For evaluation, they used gold standard annotation corpora converted in an FHIR-based schema.[61] Their first evaluation was not based on a specific clinical use case but on core clinical resources for which NLP tools and dictionaries already exist. In addition to the more general first evaluation, in a recent publication, the authors applied the developed pipeline to textual discharge summaries for reaching the further goal of using machine learning modules on the FHIR resource instances.[10] Altogether, the authors present a great approach by reaching satisfying, albeit widely ranging F-scores from 0.69 to 0.99 for various FHIR elements. In our work, we also needed to define mapping and normalization rules, but additionally, we had to define our very clinical-driven use case of pediatric medical histories and construct a new German NLP dictionary for this reason. Using FHIR in clinical text mining also has been discussed by the German working group of Daumke et al. In this study,[12] they presented the harmonization of an existing commercial text-mining tool, called Averbis Health Discovery, with FHIR. It is a very interesting, but methodological-driven paper, demonstrating mappings between the output formats of the tool and the FHIR resources. The feasibility of this approach in a

clinical context has not been shown yet. Some older publications concentrate on using HL7 CDA as interoperability standard. In 2014, Lin et al combined NLP with a semi-automatic annotation approach to generate entry-level CDA documents.[13] Before, in 2012, Meystre et al combined HL7 CDA with the ISO Graph Annotation Format to develop a new standard-based data model out of unstructured clinical data, tested on discharge summaries and progress notes.[14] As already denoted in the introduction, for openEHR, we only identified one other article in this context, published by Kropf et al.[15] Their work from 2017 shows initial successful attempts to use openEHR archetypes as final structured representation of a German pathology report. In our work, we contribute to this research by using regular expressions for information extraction and enriching it with a dictionary-based approach. Furthermore, since Kropf et al. demonstrated the feasibility of representing sections of unstructured texts by openEHR, we focused on storing retrieved facts at an entry-level to load a filled medical history as openEHR template presentation into an openEHR-based data repository.

Limitations and Future Work

Currently, our pipeline is not able to take retrospective points into account, such as the description of the patient's status from last month, last week or yesterday. We plan to integrate a combination of marker concepts and regular expressions to be able to assign each marker entry to a specific time or period and thus to visualize the timeline of a patient. Additionally, the pipeline can be further enriched by including further strategies for treating contradictory information as explained in the section above. We are aware that our workflow can be optimized by broadening our marker and event dictionary and conducting an enhanced clinical study. The first evaluation yielded promising results. However, it is limited due to a small sample size and focused on testing the technical feasibility. Further evaluations will be conducted in short term.

In the long-term, our goal is to prioritize markers and assign weightings to the archetype instances for developing a scoring application able to evaluate the condition of the patient at the time of admission. Additionally, we will access further structured information such as vital signs measurement, since this data can also be integrated into the same data repository (as presented by Haarbrandt et al[38,40] and Wulff et al[39]). With this approach, we can merge unstructured and structured information into an interoperable format. As such application will be built on top of the openEHR platform, it is potentially implementable in a "plug-and-play"-fashion at other institutions that follow the same interoperability approach and reuse the same archetypes. Alongside, it also would be a great future research question to find out whether our pipeline might be able to transform data not only into openEHR-based formats but also other various EHR standard representations. As presented in our work, this would require the design of appropriate data models represented with the specific standard format and the development and evaluation of the mapping rules and processes. For that, our work delivered all methods and knowledge assets, including a definition of relevant markers for medical histories, a summary of important items needed in the standard data models, a German dictionary for medical histories, and a definition of the required mapping rules. Together with the approaches presented in the related work section, it would be a good starting point to examine the possibilities of reaching a full pipeline based on various EHR standards. This would make the pipeline even more usable for designing interoperable applications. Hence, for future work, we recognize the efforts presented as a foundation for the development of "(…) clinically striking NLP applications that can be widely used."[35]

Conclusion

The use of an NLP-based solution to extract important information from medical histories in conjunction with a semantically enriched and structured openEHR representation is a promising approach. We successfully implemented a workflow that allows transforming medical histories as free text into a structured representation format. Based on these efforts, the long-term goal of developing interoperable application that rely on both, structured and unstructured data, e.g., to assess the condition of a patient at admission, becomes tangible. Health care professionals will benefit from such applications because they consolidate unstructured and structured information, analyze a large amount of heterogeneous data, and present the most important pieces of information. These applications will have the potentials to enable accurate, fast, and informed decision-making even in time-critical and high-risk situations. A workflow such as the one presented in this work allows the use of the full depth and width of natural language to express an observed clinical situation without obstructing the ability to reuse this valuable routine data in a structured form.

Authors' Contributions

A.W. was responsible for drafting the methodological approach, managed the overall project work, led the proof-of-concept evaluation, and has authored the manuscript. M. M. developed the described NLP pipeline, designed the openEHR archetypes and template, and co-authored the manuscript. T. J. and S. M. provided clinical expertise for requirement analysis and dictionary construction. M. H. gave subject-specific advices on the design of NLP pipelines and provided the NLP software. M. M. provided further technical and medical expertise and, together with all authors, co-authored and proofread the manuscript. All authors read and approved the final manuscript.

Funding

None.

Conflict of Interest

None declared.

References

1 Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform 2017;26(01):38–52
2 Martínez-Costa C, Cornet R, Karlsson D, Schulz S, Kalra D. Semantic enrichment of clinical models towards semantic interoperability. The heart failure summary use case. J Am Med Inform Assoc 2015;22(03):565–576
3 Beale T. Archetypes: constraint-based domain models for future-proof information systems. In: Eleventh OOPSLA Workshop on Behavioral Semantics: Serving the Customer. Seattle, Washington, BostonNortheastern University2002:16–32
4 HL7. FHIR v1.0.2. Available at: http://hl7.org/fhir/index.html. Accessed June 12, 2020
5 HL7. HL7 RIM—das Referenzinformationsmodell. Available at: http://hl7.de/themen/hl7-v3-rim-das-referenzinformationsmodell/. Accessed June 12, 2020
6 HL7. Clinical Document Architecture Release 2.0 (CDA R2). Available at: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7. Accessed June 12, 2020
7 HL7. HL7 Version 3 Standard: clinical decision support; Virtual Medical Record (vMR) Logical Model, Release 2. Available at: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=338. Accessed June 12, 2020
8 Friedman C, Johnson SB. Natural language and text processing in biomedicine. In: Shortliffe EH, Cimino JJ, eds. . Biomedical Informatics. New York, NY: Springer New York; 2006:312–343. Health Informatics
9 Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform 2017;73:14–29
10 Hong N, Wen A, Stone DJ, et al. Developing a FHIR-based EHR phenotyping framework: a case study for identification of patients with obesity and multiple comorbidities from discharge summaries. J Biomed Inform 2019;99:103310
11 Hong N, Wen A, Shen F, et al. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. JAMIA Open 2019;2(04):570–579
12 Daumke P, Heitmann KU, Heckmann S, Martínez-Costa C, Schulz S. Clinical text mining on FHIR. Stud Health Technol Inform 2019;264:83–87
13 Lin C-H, Wu N-Y, Lai W-S, Liou D-M. Comparison of a semi-automatic annotation tool and a natural language processing application for the generation of clinical statement entries. J Am Med Inform Assoc 2015;22(01):132–142
14 Meystre SM, Lee S, Jung CY, Chevrier RD. Common data model for natural language processing based on two existing standard information models: CDA+GrAF. J Biomed Inform 2012;45(04):703–710
15 Kropf S, Krücken P, Mueller W, Denecke K. Structuring legacy pathology reports by openEHR archetypes to enable semantic querying. Methods Inf Med 2017;56(03):230–237
16 Williams CN, Bratton SL, Hirshberg EL. Computerized decision support in adult and pediatric critical care. World J Crit Care Med 2013;2(04):21–28
17 Lighthall GK, Vazquez-Guillamet C. Understanding decision making in critical care. Clin Med Res 2015;13(3-4):156–168
18 Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. BMJ 1975;2(5969):486–489
19 Summerton N. The medical history as a diagnostic technology. Br J Gen Pract 2008;58(549):273–276
20 Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med 1992;156(02):163–165
21 Keifenheim KE, Teufel M, Ip J, et al. Teaching history taking to medical students: a systematic review. BMC Med Educ 2015;15:159
22 Ghosh D, Karunaratne P. The importance of good history taking: a case report. J Med Case Reports 2015;9:97
23 Wang MY, Asanad S, Asanad K, Karanjia R, Sadun AA. Value of medical history in ophthalmology: a study of diagnostic accuracy. J Curr Ophthalmol 2018;30(04):359–364
24 Masic I, Begic Z, Naser N, Begic E. Pediatric cardiac anamnesis: prevention of additional diagnostic tests. Int J Prev Med 2018;9:5
25 Ikiz MA, Cetin II, Ekici F, Güven A, Değerliyurt A, Köse G. Pediatric syncope: is detailed medical history the key point for differential diagnosis? Pediatr Emerg Care 2014;30(05):331–334
26 Brander P, Garin N. Utilité de l'anamnèse et de l'examen clinique dans le diagnostic de la pneumoniae. Rev Med Suisse 2011;7(313):2026–2029
27 Garde S, Knaup P, Hovenga E, Heard S. Towards semantic interoperability for electronic health records. Methods Inf Med 2007;46(03):332–343
28 vitasystems GmbH. EHRbase: Open Electronic Health Record Platform. Available at: https://ehrbase.org/. Accessed March 11, 2020
29 DIPS AS. DIPS Electronic Patient Record. Available at: https://www.dips.com/uk/dips-electronic-patient-record. Accessed March 11, 2020
30 Ripple Foundation C.I.C. Ltd. EtherCIS: Enterprise Clinical Data Repository. Available at: http://ethercis.org/. Accessed March 11, 2020
31 CaboLabs. CloudEHRServer: Clinical Data Management and Sharing Platform. Available at: https://cloudehrserver.com/. Accessed March 11, 2020
32 Wulff A, Haarbrandt B, Marschollek M. Clinical knowledge governance framework for nationwide data infrastructure projects. Stud Health Technol Inform 2018;248:196–203
33 Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. Yearb Med Inform 2015;10(01):183–193
34 Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H, eds. . Encyclopedia of Systems Biology. New York, NY: Springer New York; 2013
35 Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. J Biomed Inform 2013;46(05):765–773
36 Montague R. Universal grammar. Theoria 1970;36(03):373–398
37 Haarbrandt B, Schreiweis B, Rey S, et al. HiGHmed - An open platform approach to enhance care and research across institutional boundaries. Methods Inf Med 2018;57(S01):e66–e81
38 Haarbrandt B, Jack T, Marschollek M. Automated transformation of openEHR data instances to OWL. Stud Health Technol Inform 2016;223:63–70
39 Wulff A, Haarbrandt B, Tute E, Marschollek M, Beerbaum P, Jack T. An interoperable clinical decision-support system for early detection of SIRS in pediatric intensive care using openEHR. Artif Intell Med 2018;89:10–23
40 Haarbrandt B, Tute E, Marschollek M. Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. J Biomed Inform 2016;63:277–294
41 Damerau FJ. A technique for computer detection and correction of spelling errors. Commun ACM 1964;7(03):171–176
42 Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Dokl Akad Nauk SSSR 1965;163(04):845–848
43 Knuth DE. The Art of Computer Programming: Sorting and Searching. 2nd ed. Boston: Addison-Wesley; 2017
44 Pomares-Quimbaya A, Kreuzthaler M, Schulz S. Current approaches to identify sections within clinical narratives from

electronic health records: a systematic review. BMC Med Res Methodol 2019;19(01):155
45 Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018;77:34–49
46 Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. Yearb Med Inform 2017; 26(01):214–227
47 Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: opportunities and challenges. J Biomed Semantics 2018;9(01):12
48 DFKI—German Research Center for Artificial Intelligence. mEx—Medical Information Extraction. Available at: http://biomedical.dfki.de/mEx. Accessed April 19, 2020
49 Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010; 17(05):507–513
50 Averbis. Health Discovery. Available at: https://averbis.com/health-discovery. Accessed March 11, 2020
51 OpenNLP. OpenNLP. Available at: https://opennlp.apache.org/. Accessed April 19, 2020
52 LingRep. LingRep. Available at: https://www.econob.com/de/demos/. Accessed April 19, 2020
53 Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inform Assoc 2014;21(05):858–865
54 Lin Y-K, Chen H, Brown RA. MedTime: a temporal information extraction system for clinical narratives. J Biomed Inform 2013; 46:S20–S28
55 Schwartz AS, Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text. Pac Symp Biocomput 2003;8:451–462
56 Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(05): 301–310
57 Becker M, Böckmann B. Extraction of UMLS® Concepts Using Apache cTAKES™ for German Language. Stud Health Technol Inform 2016;223:71–76
58 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
59 König M, Sander A, Demuth I, Diekmann D, Steinhagen-Thiessen E. Knowledge-based best of breed approach for automated detection of clinical events based on German free text digital hospital discharge letters. PLoS One 2019;14(11):e0224916
60 Löpprich M, Krauss F, Ganzinger M, Senghas K, Riezler S, Knaup P. Automated classification of selected data elements from free-text diagnostic reports for clinical research. Methods Inf Med 2016;55 (04):373–380
61 Hong N, Wen A, Mojarad MR, Sohn S, Liu H, Jiang G. Standardizing heterogeneous annotation corpora using HL7 FHIR for facilitating their reuse and integration in clinical NLP. AMIA Annu Symp Proc 2018;2018:574–583