*Health Informatics Meets eHealth*
*D. Hayn and G. Schreier (Eds.)*

*doi:10.3233/978-1-61499-759-7-24*

# Semantic Technologies for Re-Use of Clinical Routine Data

Markus KREUZTHALER[a,b,1], Catalina MARTÍNEZ-COSTA[b], Peter KAISER[c] and Stefan SCHULZ[b]
[a]*CBmed GmbH - Center for Biomarker Research in Medicine*
[b]*Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*
[c]*SAP SE, SAP Connected Health Platform*

**Abstract.** Routine patient data in electronic patient records are only partly structured, and an even smaller segment is coded, mainly for administrative purposes. Large parts are only available as free text. Transforming this content into a structured and semantically explicit form is a prerequisite for querying and information extraction. The core of the system architecture presented in this paper is based on SAP HANA in-memory database technology using the SAP Connected Health platform for data integration as well as for clinical data warehousing. A natural language processing pipeline analyses unstructured content and maps it to a standardized vocabulary within a well-defined information model. The resulting semantically standardized patient profiles are used for a broad range of clinical and research application scenarios.

**Keywords.** Electronic Health Records, Natural Language Processing, Semantics

## 1. Introduction

The totality of electronic health records (EHRs) in health care institutions of all levels constitutes a highly interesting "data treasure" for primary and secondary usage scenarios [1]. Innovative re-use of this wealth of information about millions of patients, available as structured and unstructured data from heterogeneous sources, requires a combined effort capitalising on data semantics approaches, biomedical terminologies, natural language processing, big data management and predictive content analytics.

Natural language processing (NLP) is a fundamental technology, which enables access to relevant information within patient narratives, and it can be used to get semantically enriched patient profiles [2]. This is especially important for secondary use case scenarios and retrospective cohort building within a clinical environment. Very often, attributes required for addressing a specific information need, or attributes that characterise a patient cohort are present in EHRs as parts of daily routine documentation within a certain document type produced in a specific unit by a certain type of health professional. Using this raw data without further processing, even simple queries may require time-consuming efforts, involving combined expertise in information extraction and information management in the medical domain. Therefore, further processing is

[1] Corresponding Author: Markus Kreuzthaler, CBmed GmbH - Center for Biomarker Research in Medicine, Stiftingtalstrasse 5, 8010 Graz, Austria; Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2/V, 8036 Graz, Austria, E-Mail: markus.kreuzthaler@medunigraz.at

necessary in order to ensure that the information needs are met, required attributes and their sources are identified, a workflow is devised and, finally, the information need is correctly formulated with the query syntax supported by the information retrieval tool in use. This paper will demonstrate how such a retrieval scenario can benefit from appropriate natural language processing methods together with a well-defined information model.

## 2. Background

### *2.1. Medical language*

The complexity of the understanding of written language depends on the domain of discourse as well as on characteristics of domain-specific sublanguages, which often heavily deviate from syntactic and stylistic conventions of standard language. This is necessary for understanding the characteristics of clinical texts, especially in contrast to scientific papers or textbooks, as the following text snippet demonstrates [3]:

```
   Ca. 2 x 1 cm, große ovaläre Verschattung im UF li. Rez. re.
lat. frei, li.teilhärent
```

The translation is “An approximately 2 by 1 cm sized, oval-shaped opacity (in the chest X-ray) over (- laying) the left lower lobe; pleural recess (on the right side of the lobe) free of fluids, on the left side partly adhesive (after non-recent inflammation)”. Such highly condensed text is understood by clinicians (at least by those of the same area), but it poses challenges to computer-based morphological, syntactic and semantic processing. Numerous language idiosyncrasies that are typical for clinical texts have to be considered when tailoring an NLP pipeline and supporting lexical resources to clinical narratives: acronyms, abbreviations, ambiguous terms, synonyms, derivations, single-word compounds, uncorrected spelling, spelling variants, typing and punctuation errors, jargon expressions, telegram style, non-standardized numeric expressions, and non-standard variations of negations [1,4].

### *2.2. Clinical natural language processing*

Clinical NLP is a relatively small sub-area, compared to e.g. bioNLP, which targets scientific texts. A reason for this is the difficulty of accessing clinical data due to privacy concerns. This results in a lack of shared data and gold standards. Nevertheless, scientific challenges with special tracks have been established to foster *clinical* NLP research (i2b2 NLP research data sets, ShARe/CLEF eHealth, SemEval, TREC Medical Tracks, MedNLPDoc). Research institutions that have considerable merits in clinical NLP are the Mayo Clinic [5] (cTakes), the Veterans Affairs network of hospitals [1,6] (The Leo framework - The VINCI-developed NLP infrastructure using UIMA [17], the Apache Unstructured Information Management Architecture) and the Columbia-Presbyterian Medical Center [7,8] (MedLEE). MetaMap [9] uses concepts of the UMLS Metathesaurus for semantic annotations and is used as a core engine for SemRep [10], a system that generates subject-predicate-object triples out of MetaMap annotations. Usually applied in the biomedical domain, its applicability to clinical narratives had been

investigated [11]. HITEx [12] is based on GATE [13], an NLP engine. The DKPro Core collection also has to be mentioned [14] in the context of NLP functionalities.

In most of them, the scope is limited to English text, whereas the International Workshop on Health Text Mining and Information Analysis (Louhi) is worth mentioning due to its focus on European languages. Nevertheless, the situation for the German language in this research area is not satisfying. The lack of openly available gold standards makes comparison between competing approaches almost impossible. Notable exceptions are the JULIE Labs in Germany, which made their clinical language models and UIMA based NLP framework openly accessible.

Pipeline components are mostly a combination of various methods exploiting rule-based engines (e.g. Apache UIMA Ruta [18] or regular expressions), distributional semantics, or unsupervised / supervised machine learning methods, with a special focus on (i) not violating real time constraints and (ii) reaching a certain level of annotation quality. Especially terminology / ontology management in the background and its mapping to narrative content is a main building block in a pipeline. Recently, deep learning methods, especially bi-directional long short term memories (BI-LSTM), one type of recurrent neural network, have attracted attention for processing clinical narratives e.g. for the de-identification task [15,16].

In the use case presented here we extended the core NLP engine provided by the company Averbis [25] with six rule-based extraction approaches based on regular expressions. We built a custom medication terminology used by the parameter-tuned concept matcher within the pipeline and also mapped to ICD-10 codes. The results of the extraction engine are fed into the generic clinical data model of the SAP Connected Health platform [19], which is described in the following section.

## 3. Methods

### *3.1. Data source*

The data set for first initial experiments on clinical information extraction, using the in-memory SAP HANA technology as data sink, was a sample from in- and outpatient discharge summaries from a dermatology department, filtered by ICD codes and admission dates. Using an Extract Transform Load (ETL) process with Talend Open Studio, 1,696 summaries were extracted. After manual de-identification this corpus could then be used various text mining experiments.

**Table 1.** Entity types to be extracted from the narratives together with examples of their representation.

| Entity type | Representation examples |
|---|---|
| Diagnosis (ICD-10 code) | “malignant melanoma”, “C43.9” |
| Medication information | “Norvasc 5 mg 1-0-1”, ”Atarax 25 mg 0-0-1” |
| Tumor staging (pTNM) | “pT1a N3 M1c”, “pT3aN0M1c”, “pT-2b” |
| Breslow level | “TD <0,5mm”, “TD unter 0,5 mm”, “Tumordicke 0.9 mm” |
| Clark level | “Invasionslevel III”, “Level II”, “Clark-Level III” |
| Mitosis index | “Mitosen < 1/mm²” |
| S100 biomarker | “S100 0.058 (Normbereich)” |
| Ulceration | “ulceriertes Ca in situ”, “exulzeriertes MM” |

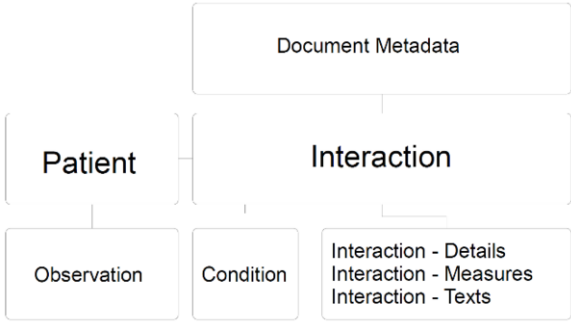


**Figure 1.** Logical clinical data model in the SAP Connected Health platform.

### *3.2. Use case: malignant melanoma*

Based on the use case "malignant melanoma" formulated by the administrators of a large biobank, we analysed the above data set in order to gather additional phenotype information for defining retrospective patient cohorts. Table 1 shows the types of entities to be extracted in a first stage of the project. Specific annotators were implemented for the information extraction process and combined with the NLP core in use.

### *3.3. Modelling within the SAP Connected Health platform*

A core entity of the clinical data model within the SAP Connected Health platform is the *Patient* and the corresponding *Interactions* with healthcare providers (see Figure 1). An *Interaction* is an event, which may occur at a specific time or time interval. Examples are diagnostic procedures, chemotherapy treatments, genomic analyses or hospital check-ups. Each *Interaction* is uniquely identified by an InteractionID and classified by an InteractionType. The latter can be coded according to standardized terminologies such as SNOMED CT. *Patients* usually participate in several *Interactions*.

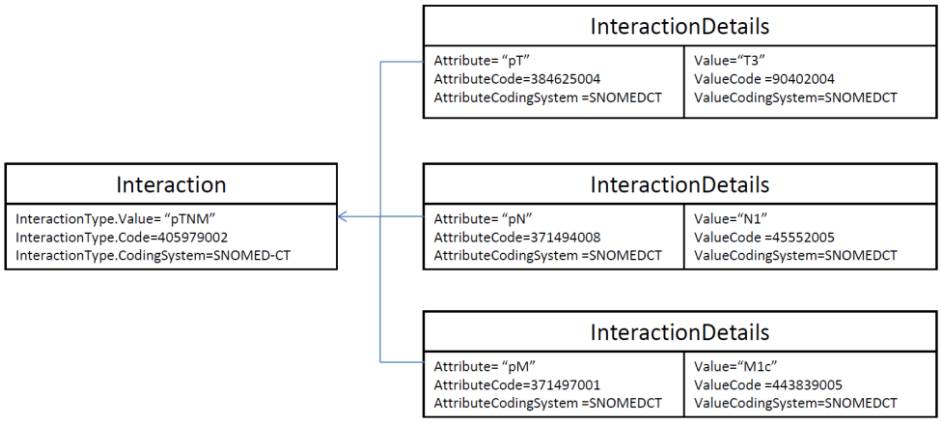


**Figure 2.** pTNM = T3N1M1c representation according to the SAP Clinical Data Model. For simplicity reason the version of the SNOMED coding system is not shown (Jan 2017 release).

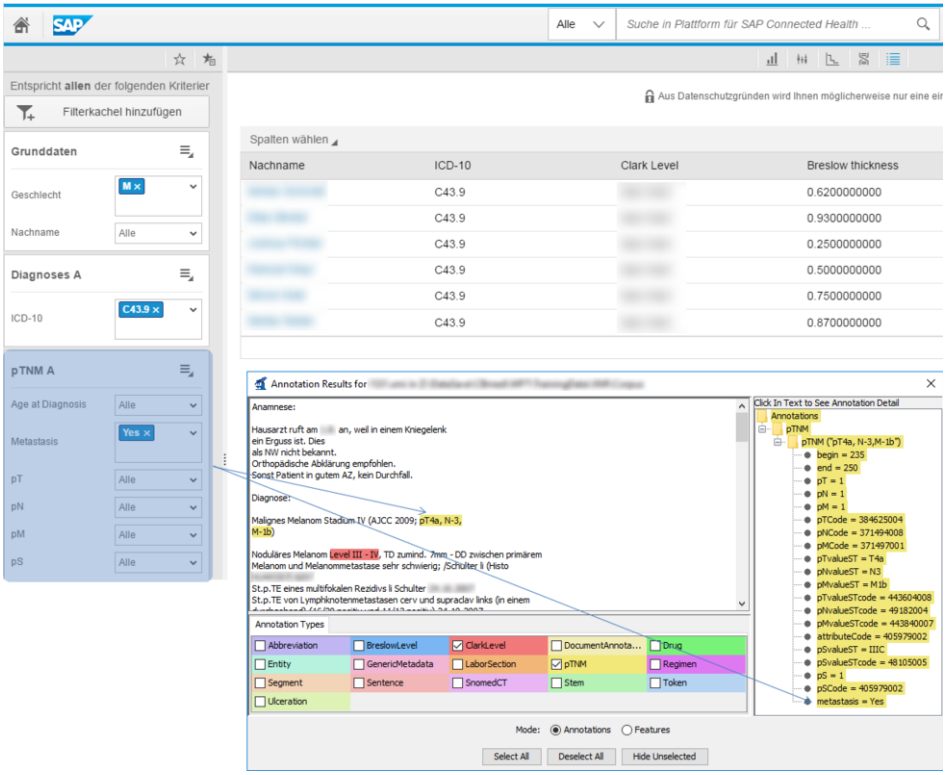


**Figure 3.** Customized filter cards for the secondary use of the structured and standardized clinical routine data within SAP Medical Research Insights. The annotation results for the entities Clark Level and pTNM within the corresponding clinical narrative of the NLP pipeline are depicted before data harmonization within the SAP Connected Health platform.

Besides, an *Interaction* can be further described by *Interaction Details* / *Measures* / *Texts* to represent coded (i.e. code and code system), quantitative (i.e. numeric value and unit) or free-text data respectively. Another relevant entity is the *Document,* which is constituted by free-text and directly relates with *Interactions* (e.g. a discharge summary, blood test report, etc.). *Observations* record a time-specific observation or measurement for a patient (e.g. blood measurement or tumour sizing). Finally a *Condition* is a specific instance of a medical condition in the patient (e.g. a disease) and is used to filter entities associated with the same condition.

## 4. Results

### *4.1. Integrating data within the SAP Connected Health platform - Modelling pTNM*

Structured data extracted from the clinical narratives can be integrated in the Clinical Data Warehouse (CDW), by using custom or standard plug-ins and adapters provided by SAP or, alternatively, by using an ETL tool. At an early stage of the project, we imported structured data stored as CSV files, the result file format from the applied NLP pipeline to the narratives. However, in a later stage, structured and unstructured data from a source

system will be directly imported into the SAP Connected Health platform by using an ETL tool like Talend Open Studio. Such a tool will communicate directly with the NLP engine via a RESTful service which will return a set of annotations using e.g. JSON, also avoiding intermediate result files.

Once in the system, the data is mapped to the clinical data model of the SAP Connected Health platform, described in Section 3.3. SAP Medical Research Insights will then query this model to retrieve clinical data for the corresponding use cases. Figure 2 shows how pTNM is represented according to the clinical data model. It corresponds to an *Interaction* of the type "pTNM" described by three coded attributes with SNOMED CT (i.e. InteractionDetails), each one describing the primary tumour (pT), regional lymph nodes (pN) and distant metastasis (pM) according to the 7th edition (2009) of the AJCC Cancer Staging Manual.

### *4.2. Querying clinical information: SAP Medical Research Insights*

SAP Medical Research Insights allows data access from heterogeneous sources such as clinical information systems, tumour registries, biobank systems etc. It allows filtering and grouping patients according to different attributes based on the SAP Connected Health clinical data model. Figure 3 depicts its main interface. On the left side, filter cards can be used to formulate queries. Each filter card corresponds to an interaction type. One of the filter cards is directly related to the pTNM classification and its corresponding values are now searchable within a defined structured and standardized information model.

In this case, we would like to retrieve all male patients with the diagnosis "C43.9" (ICD-10 code for "malignant melanoma of skin") and metastasis. More advanced queries including disjunction and temporal information are also supported.

## 5. Discussion and Conclusion

Much relevant information in electronic medical records is still only available as unstructured text. NLP methods need to account for specific morphological, lexical and syntactic features of clinical language. The extended natural language processing pipeline analyses unstructured text and maps it to a standardized vocabulary within a well-defined data model. This model is part of the SAP Connected Health platform, the core of which is a high performance in-memory database. Structured information within SAP HANA can be exploited for a broad range of clinical and research application scenarios e.g. using the SAP Medical Research Insights.

We are currently processing a corpus of de-identified German outpatient discharge summaries from dermatology, together with test queries to retrieve documents on melanoma according to criteria formulated by biobank experts. Based on them, eight entity types were identified and annotated within the clinical narratives, using standardized terminologies such as ICD-10 and SNOMED CT.

The extracted information has been fed into the clinical data model within the SAP Connected Health platform, which has as focus patients' *Interactions* with healthcare providers. *Interactions* are events such as a diagnostic procedure or a medication administration. As a disadvantage we mention that this clinical model is proprietary of the SAP Connected Health platform, which complicates the interoperability of data stored within the SAP HANA environment with other information systems. This could

be solved by transforming data in SAP Connected Health platform into a standardized representation according to existing international EHR standards and specifications such as HL7 CDA/FHIR [20-21], ISO 13606 [22] or openEHR [23].

Future work has to present detailed evaluation measures (precision, recall, F-measure with respect to exact / inexact matching and micro / macro averaging) regarding (i) the *clinical information extraction* task of clinical entities and (ii) for the *clinical information retrieval* use cases for which an adopted TREC-based evaluation scenario would give the most appropriate measurements. For both scenarios a human revised gold standard is needed. For the clinical information extraction case e.g. annotation tools like BRAT can be used. For information retrieval evaluation already performed and evaluated information needs of retrospective scientific inquiries for cohort building could be exploited. Nevertheless, a rigorous analysis of needed attributes fur fulfilling the information retrieval task has to be done. This directly influences the information extraction task as often cohort specific attributes reside just within clinical narratives [24]. These considerations directly lead to a minimal data set of patient-based attributes needed for most of the scientific queries. Defining such a set to exploit the SAP Medical Research Insights we see as a future task. In addition, usability aspects of the SAP Medical Research Insights and the ability to express complex information needs for cohort building using the SAP Medical Research Insights filter cards have to be investigated in detail.

## Acknowledgements

This work is part of the IICCAB project (Innovative Use of Information for Clinical Care and Biomarker Research) within the K1 COMET Competence Center CBmed (http://cbmed.at), funded by the Federal Ministry of Transport, Innovation and Technology (BMVIT); the Federal Ministry of Science, Research and Economy (BMWFW); Land Steiermark (Department 12, Business and Innovation); the Styrian Business Promotion Agency (SFG); and the Vienna Business Agency. The COMET program is executed by the FFG. We also thank KAGes (Styrian hospital company) and SAP SE to provide significant resources, manpower and data as basis for research and innovation, Averbis GmbH for providing the Information Discovery platform, Biobank Graz for the use case descriptions and finally Werner Aberer, director of the Department of Dermatology, Medical University of Graz, for the provision of sample data (anonymised texts).

## References

[1] Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform, 35(128), 44.

[2] Friedman, C., & Elhadad, N. (2014). Natural language processing in health care and biomedicine. In Biomedical Informatics (pp. 255-284). Springer London.

[3] Kreuzthaler, M., Daumke, P., & Schulz, S. (2015). Semantic retrieval and navigation in clinical document collections. Stud Health Technol Inform, 212, 9-14.

[4] Patterson, O., Igo, S., & Hurdle, J. F. (2010, November). Automatic acquisition of sublanguage semantic schema: towards the word sense disambiguation of clinical narratives. AMIA Annu Symp Proc. 2010, 612-616.

[5] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010, 507-513

[6] Patterson, O. V., Forbush, T. B., Saini, S. D., Moser, S. E., & Duvall, S. L. (2015). Classifying the Indication for Colonoscopy Procedures. Stud Health Technol Inform 216, 614-618.

[7] Friedman, C., Johnson, S. B., Forman, B., & Starren, J. (1995). Architectural requirements for a multipurpose natural language processor in the clinical environment. Proc Annu Symp Comput Appl Med Care, 1995, 347-351.

[8] Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S. B., & Clayton, P. D. (1995). Natural language processing in an operational clinical information system. Natural Language Engineering, 1(01), 83-108.

[9] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001, 17-21.

[10] Rindflesch, T. C., & Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics, 36(6), 462-477.

[11] Liu, Y., Bill, R., Fiszman, M., Rindflesch, T. C., Pedersen, T., Melton, G. B., & Pakhomov, S. V. (2012). Using SemRep to label semantic relations extracted from clinical text. In AMIA Annu Symp Proc. 2012; 587-95

[12] Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N., & Lazarus, R. (2006). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Medical Informatics and Decision Making, 6(1), 30.

[13] Cunningham, H., Wilks, Y., & Gaizauskas, R. J. (1996, August). GATE - A General Architecture for Text Engineering. In Proceedings of the 16th conference on Computational linguistics-Volume 2 (pp. 1057-1060). Association for Computational Linguistics.

[14] Bär, D., Zesch, T., & Gurevych, I. (2013, August). DKPro Similarity: An Open Source Framework for Text Similarity. In ACL (Conference System Demonstrations) (pp. 121-126).

[15] Shweta, A. E., Saha, S., & Bhattacharyya, P. (2016). Deep Learning Architecture for Patient Data De-identification in Clinical Records. ClinicalNLP 2016, 32.

[16] Lee, J. Y., Dernoncourt, F., Uzuner, O., & Szolovits, P. (2016). Feature-augmented neural networks for patient note de-identification. arXiv preprint arXiv:1610.09704.

[17] Savova, G., Kipper-Schuler, K., Buntrock, J., & Chute, C. (2008). UIMA-based clinical information extraction system. Towards enhanced interoperability for large HLT systems: UIMA for NLP, 39.

[18] Kluegl, P., Toepfer, M., Beck, P. D., Fette, G., & Puppe, F. (2016). UIMA Ruta: Rapid development of rule-based information extraction applications. Natural Language Engineering, 22(01), 1-40.

[19] SAP Connected Health platform, https://help.sap.com/platform_health, last access: 15.3.2017.

[20] HL7 Implementation Guide for CDA® Release 2: IHE Health Story Consolidation, Release 1.1 - US Realm, http://www.hl7.org/implement/standards/product_brief.cfm?product_id=258 last access: 13.3.2017.

[21] HL7 FHIR. https://www.hl7.org/fhir/ last access: 13.3.2017.

[22] ISO/TC 215, Health informatics (2008). Electronic health record communication Part 2: Archetype interchange specification, (ISO 13606-2:2008)

[23] T. Beale & S. Heard: Archetype Definitions and Principles version 1.0.2. http://www.openehr.org/releases/1.0.2/architecture/am/archetype_principles.pdf last access: 13.3.2017.

[24] Kreuzthaler, M., Schulz, S., & Berghold, A. (2015). Secondary use of electronic health records for building cohort studies through top-down information extraction. Journal of biomedical informatics, 53, 188-195.

[25] Averbis text analytics - Healthcare: https://averbis.com/en/industries/healthcare/ last access: 13.3.2017.