How to Improve Information Extraction from German Medical Records

Johannes Starlinger, Madeleine Kittner, Oliver Blankenstein, Ulf Leser

Abstract: Vast amounts of medical information are still recorded as unstructured text. The knowledge contained in this textual data has a great potential to improve clinical routine care, to support clinical research, and to advance personalization of medicine. To access this knowledge, the underlying data has to be semantically integrated - an essential prerequisite to which is information extraction from clinical documents.
A body of work, and a good selection of openly available tools for information extraction and semantic integration in the medical domain exist, yet almost exclusively for English language documents. For German texts the situation is rather different: research work is sparse, tools are proprietary or unpublished, and rarely any freely available textual resources exist. In this survey, we (1) describe the challenges of information extraction from German medical documents and the hurdles posed to research in this area, (2) especially address the problems of missing German language resources and privacy implications, and (3) identify the steps necessary to overcome these hurdles and fuel research in semantic integration of textual clinical data.

ACM CCS: Applied computing → Life and medical sciences → Health care information systems; Information systems → Retrieval tasks and goals → Information extraction; Applied computing → Document management and text processing → Document preparation → Annotation

Keywords: medical text mining; information extraction; semantic information integration

1 Introduction

The phenotype describes a patient's observable physical characteristics and their pathological conditions. It is encoded in the patient's medical status and has been recorded by physicians in a wealth of clinical notes and data sheets for ages. Systematic use of such data may carry both economic impact through the healthcare system, and scientific impact by facilitating medical knowledge discovery from the wealth of available data and by easing the selection and recruitment of patients for clinical trials. It also promises to improve care options for each single patient in daily clinical routine [24]. More recently, phenotypic data is being shifted into a new focus with the trend towards personalized medicine [40], initiated by the development of high throughput next generation DNA-sequencing some 15 years ago. This trend is rapidly changing the way we perceive medical care: From a rather aggregate discipline where each patient receives a standard set of state-of-the-art treatments for their respective diagnosis, into a highly personalized discipline where each patient's individual (genetic) profile is potentially considered when choosing the best possible treatment [19]. With phenotypic data to complement the, so far, largely genotype centered view on personalized medicine, even more patients could benefit from this medical paradigm shift and new research opportunities could be unlocked. For instance, similarity search over large collections of patient cases could be used to infer options for further examinations or treatments [26] and the combination of phenotypic and genetic data could enable systematic research to reveal molecular mechanisms of rare diseases [9]. However, while genetic information is digitally available in structured databases, phenotypic information has traditionally been recorded in unstructured text notes, and predominantly still is today.

While phenotypic data thus has a great potential to improve patient care and pathomechanism research in various ways, several hurdles must be taken to allow caregivers to make full use of such data and the knowledge to be mined from it - the most important of which

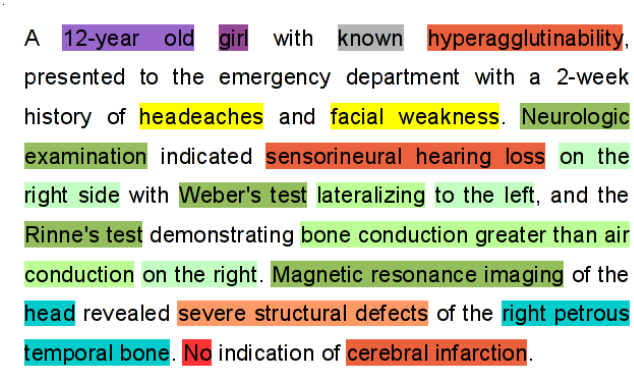

Figure 1: Sample medical text with various types of entities highlighted, including personal information (purple), past and current diagnoses (orange), symptoms (yellow), procedures (green), findings (light green), anatomical sites (blue), and negation (red). Relationships between entities are not shown.

is its semantic integration. In fact, one of the major obstacles to data-driven personalization of healthcare is the current state of how the overwhelming majority of patient data is stored in clinical information systems: as free form text notes [25]. To a certain degree, singular data points of care events are recorded in structured form, such as billing information or blood sampling data. Input of staccato exclamation-style phrases using non-uniform abbreviations of medical terminology, however, is often the preferred, because fastest way for clinicians to record their findings [3]. As such, a prerequisite to semantic integration of clinical data and documents are semantic annotation of, and information extraction from these free form text notes via natural language processing (NLP) techniques. Figure 1 shows a sample clinical note with various types of entities highlighted.

Much research has been invested in NLP of biomedical and clinical documents in English, e.g., [14, 20, 38, 43, 48]. Precise NLP solutions are highly language and domain dependent, and it is often not trivial to modify solutions that exist for one language to be used for another. This makes it inevitable to re-evaluate, to adjust, and often to (re)create approaches specifically for a given language of application. Yet, for other languages, such as Danish, Finnish, French, Dutch, Swedish, and German, much less published research work exists, e.g., [1, 8, 27, 30, 45]. Research and development of medical information extraction solutions in Germany is currently driven almost exclusively by commercial efforts, such as ProMiner [21] or RadMiner [17], or highly local applications at singular clinics, e.g., [18, 28, 49], which are seldom published about in reproducible detail. An independent investigation, development, and evaluation of different approaches by the research community is still largely missing, and much redundant effort is being made due to a lack of shared resources.

The objectives of this review are (1) to describe the challenges of clinical NLP and the hurdles posed to research in this area, (2) to address the problems of missing German language resources and privacy implications, and (3) to identify the steps necessary to overcome these hurdles and fuel research in semantic integration of textual clinical data. Following these objectives, Sections 2, 3, and 4 survey characteristics of clinical text, point to existing solutions and resources, and enumerate essential cornerstones for clinical NLP research. We conclude in Section 5.

2 Research Challenges

Information extraction from any kind of document requires a number of processing steps, as sketched in Figure 2. These typically include general linguistic operations like detection of sentence boundaries, tagging of terms or phrases by their part of speech (e.g., verbs or noun phrases), and stemming, i.e., reducing inflected words to their word stem. Following these preparational steps, single terms or term sequences have to first be recognized as relevant entities (named entity recognition, NER), and then normalized to a unique identifier (named entity normalization, NEN). This semantically annotates the entity and the document with the respective concept linked to the chosen identifier. Algorithms for relation extraction can then be applied to assess how the extracted entities are semantically connected amongst each other.

This task of information extraction poses a number of challenges. From clinical documents, several types of entities with different characteristics have to be properly recognized, normalized, and put into context, such as diagnoses, symptoms, physiological conditions, medication, procedures, anatomical sites, and genomic mutations. Some of these include value-and-unit expressions (e.g., measurable physiological conditions like blood pressure), while others may carry a modal adjunct to qualify the respective finding (e.g., diagnoses marked as being finite, suspected, or possible alternatives). Entities such as tumor classifications (TNM), on the other hand, can often only be extracted from multiple documents [53], which, even more, may need to be temporally aligned by the extraction of date and time [10]. Additionally, up to approximately half of all clinical conditions mentioned in medical documents are negated due to the explicit documentation of excluded diagnoses or treatment options [6]. Furthermore, clinical texts are often not well formed, neither in terms of grammar, nor in terms of vocabulary; instead, non-uniform abbreviations are concatenated to staccato exclamation-style phrases, as exemplified in Figure 3 [46]. Sometimes, these texts are not even recorded digitally in the first place, but handwritten and scanned with often poor OCR results. Standard NLP tools like sentence splitters or part-of-speech taggers can not be used on such texts, but solutions have to be developed which are tailored to the specificities of the domain, possibly taking inspiration from the aspects it shares with other domains,

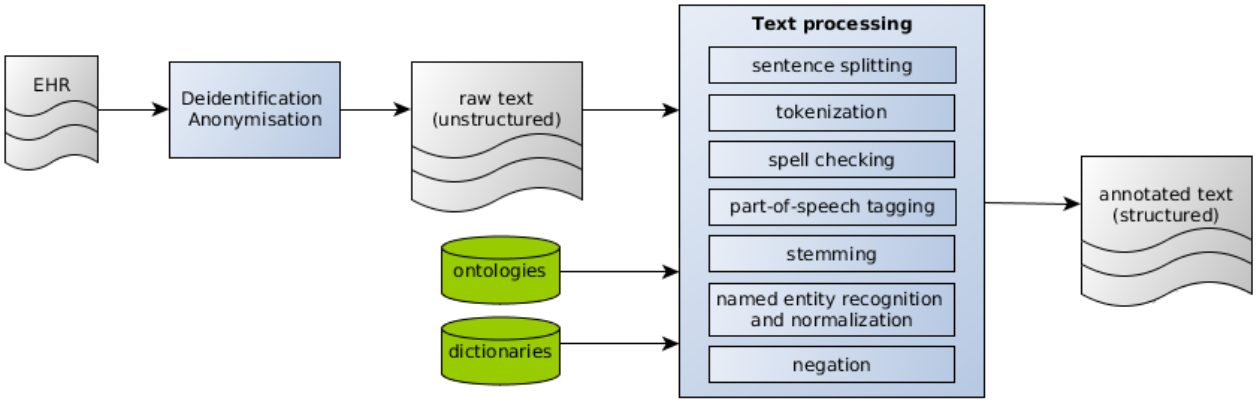

Figure 2: Common setup of NLP pipeline: Electronic health records (EHR) are de-identified and input to the text processing module performing several subtasks including linguistic preprocessing, named entity recognition and normalization, and negation detection to produce annotated text. Ontologies and dictionaries are used as resources for entity recognition and normalization.

such as social media messages from Twitter [33].

In other languages, especially English, many of the issues enumerated above have been targeted and high annotation quality for selected corpora of medical texts has been achieved, e.g., [7, 14, 39, 43, 48]. The respective systems often use dictionary- and rule-based approaches, heavily relying on the availability of comprehensive lexica and ontologies for their target entities. Important terminologies to be used for entity normalization, however, have been originally developed in English and their translations to other languages often have a much lower coverage. Also, such ontologies are usually proprietary. SNOMED-CT, for instance, as the most widely used terminology incurs a license fee for use in Germany[1] - with no official German translation available. For entity recognition, such broad classifications are also not sufficient because only subsets of the terminology used in clinical documents are covered, especially regarding acronyms and abbreviations and their disambiguation [36]. For instance, the German issue of the ICD-10 classification (which is used to code diagnoses for billing purposes and of which, again, only last year's issue is available for research for free[2]) contains no mention of the acronym *PCP*, frequently used in clinical documentation to refer to a chronic disease of the joints. Thus, existing sources for dictionary-based annotation have to be largely enriched and complemented with synonym and abbreviation data.

The acronym *PCP* also highlights another difficulty with clinical text, as the semantics of a given term often depend on the organ context it is used in, i.e., the medical specialty the respective text stems from: When used in the context of a pulmonary report, *PCP* instead refers to a fungal infection of the lung. Similarly, the term *lymphocytosis* may refer to a diagnosis of an increased count of white blood cells, or merely describe a property of the microscopy specimen retrieved in an examination. In this case, proper disambiguation again depends on the organ context, but also on the type of document the term is encountered in, e.g., either a full discharge summary of a patients stay or a short text examination report. Which of these specific contexts each potential annotation applies to is highly important and must be supplied to information extraction algorithms.

The clinical jargon used in a medical note not only depends on the respective medical specialty and document type, but also on the concrete individual institution: The selection of synonyms and abbreviations used, and grammatical style are strongly influenced by local institutional conventions [49]. While this observation is not specific to the medical domain, it has to be considered when evaluating reports of existing approaches: Local solutions within singular institutions have been reported to achieve highly accurate results [11, 28, 49] for a set of target entities on a (typically small) local evaluation corpus. Although these results are encouraging, they are hard to extrapolate and compare to each other across institutions, as, often enough, different types of evaluation metrics are used, neither the tools created, nor the corpus trained or evaluated on are publicly available, and - due to said institutional specificities - the generalizability of the respective approaches and of their reported performance is unclear.

3 Resources and Privacy Regulations

The most striking obstacle for German clinical NLP is the absence of shared resources, both in terms of clinical data, and of tools. While German issues of some

[1] even for research
[2] https://www.dimdi.de/static/de/klassi/icd-10-gm/formate/index.htm

Marginal lymphocytosis with increased CD4/CD8 ratio, compatible with sarcoidosis, DD: tuberculosis, collagenosis.

Figure 3: Sample examination summary text including modal *DD* to qualify differential diagnoses; translated from German.

important classifications of medical terminology, such as ICD-10 for diagnoses, LOINC for tests, measurements, and observations [13], or the NegEX negation lexicon [7], can be obtained online (yet often with reduced coverage), actual text data and domain-tailored applications are not shared or even publicly announced. In fact, the only openly available such resource we are aware of are text analysis models for sentence splitting, tokenization, and part-of-speech tagging[3] trained on the FraMed corpus [11].

The FraMed corpus itself is not available, however, as it contains sensitive clinical data. This showcases a fundamental hurdle of data provision in the medical domain: One of the distinct problems with handling clinical data is the inherent danger of inflicting on the privacy of patients. Both legal and ethical standards pose strong requirements to patient data privacy and safety. Access to clinical data is highly regulated and restricted, also for research [44]. Data can often only be accessed on hospital premises after lengthy negotiations and consultations with data protection officers. Only data which is fully de-identified or anonymized could be considered for publication. The unavailability of tools for automatic de-identification of German language texts and the prohibitive cost of manual de-identification, in conjunction with strong German data protection laws result in a near complete lack of public data. As a consequence, an independent, comparative evaluation of different approaches on comprehensive test data is currently impossible and the initial hurdle for starting research in German clinical NLP and semantic integration is high: It requires to not only get access to a sufficiently large data set of clinical documents (either de-identified or confidential), but to also collect manual evaluation data on these documents with the help of medical professionals. The absence of a sufficiently large and representative, publicly available corpus of clinical documents and gold-standard annotations for even a single type of entity is the most pressing problem to be solved.

As for tool resources, on the other hand, neither the applications for medical information extraction created by in-house efforts at specific clinics, e.g., [18, 28, 49], and even less commercial applications, e.g., [21, 17], have been released to the public. Correspondence with respective authors of tools from the former class revealed that, generally, sharing of these tools is possible and under consideration. Yet, so far, we are not aware of any such effort actually being made.

[3] http://www.julielab.de/Resources.html

4 Research Roadmap

To best allow the (German) research community to engage in the field of semantic integration of clinical documents, a number of steps are required which we enumerate in the following. We draw from successful solutions in other disciplines, and practical considerations and experience regarding the specifics of the medical domain.

Corpus Creation. Only with a sufficiently representative corpus available to the scientific community, different approaches and tools become comparable. As such, one of the foremost necessities is to provide both textual data and its semantic annotations. In fact, we believe that whenever research uses clinical data, every reasonable measure should be taken to pursue contribution of such data to the research community. Due to the aforementioned intra-specialty and intra-clinical idiosyncrasies regarding textual characteristics, a corpus of medical documents would ideally be comprised of documents a) from a variety of medical sub-disciplines to represent different specialties, and b) from a number of different institutions to incorporate different in-house colloquials. When each such partition of documents is tagged with its respective source, the influence of local particularities can be pinpointed, and generalizable observations can be derived. In the end, a German language clinical data set comparable to the MIMIC II database [42] - covering more than 32,000 patients with over 700,000 clinical notes - would be an invaluable asset for driving clinical data research. Only with such large corpora, modern machine learning methods become applicable, for example, the use of word embeddings to solve the problem of abbreviation sense disambiguation in clinical notes [52], or to perform semantic annotation itself [32].

De-identification. Before clinical documents can be publicly shared, they have to be carefully de-identified. For sufficiently large quantities of documents, manual de-identification is too costly. Algorithmic methods for automatically performing such anonymizing operations have to be established and evaluated for German language texts. Again, numerous possible approaches have been reported for English documents [29], typically using rule-based and pattern matching approaches, machine learning approaches when large annotated corpora are available, or even combined approaches [34]. A systematic evaluation of 5 de-identification tools available for English tested on the i2b2 [51] de-identification reference corpus showed, however, that a fair amount of adaptation is required for any de-identification tool to obtain acceptable results with new, unseen documents [12], even within the same language. For evaluation, training and adaption of such methods on German language text, an initial, manually de-identified corpus will be inevitable. To mitigate the cost of a fully manual effort, solutions

inspired by approaches in other languages could provide a starting ground to semi-automatic approaches with iterative refinement of computational results by human-provided feedback on algorithmic results, as sketched in Figure 4.

Human Provided Semantic Annotations. For specific types of applications, such as abbreviation sense disambiguation through word embeddings as mentioned above, it is not even necessary (and not feasible) for the corpus to carry gold-standard semantic annotations. For the development and evaluation of semantic integration, on the other hand, ground truth semantic annotations have to be provided by human experts for each entity type of interest. Several tools exist for this task [37], which not only allow to identify singular entities, but are also able to capture complex relationships as necessary for aggregate annotations like tumor classifications. Despite these positive premises, our experience has shown that the time medical research collaborators are able to invest into providing a sufficient amount of annotations is highly limited. Clinical duties often leave little room for assignments not directly related to patient care. Catering to this dilemma, tools for annotating clinical documentation would thus ideally be tightly integrated into the clinical workflow itself: The physician would record their clinical notes as usual, while an iteratively trainable system for semantic annotation of these notes would automatically highlight detected entities in the background. The resulting annotations could then be instantly reviewed by the physician. Of course, such a system would have to be unobtrusive and intuitive, and not interfere with the regular workflow of patient care.

Common Annotation Data Model. For representation and storage of annotations for a given text, a data model is required. Such a data model typically encompasses information about the annotated mentions as found in the text, the normalized entity identifiers these mentions are annotated with, as well as extracted relations between different entities. In a clinical medical setting, the data model should also capture which algorithm or human an annotation was created by, and the institution, subdiscipline, and procedure the underlying text stems from. Which data is captured, and how this data is formatted and represented (e.g., tabular or as XML, standoff or within the annotated text) depends on the concrete data model used. Currently, virtually every annotation tool employs its own data model [35]. This heterogeneity greatly hinders interoperability between different tools, comparison of annotation results, and community creation of annotated corpora. Some suggestions have been made towards common standards for data models in clinical NLP, e.g., [15, 35], none of which have found widespread adoption, yet. A satisfying solution to this challenge will only be found through broad

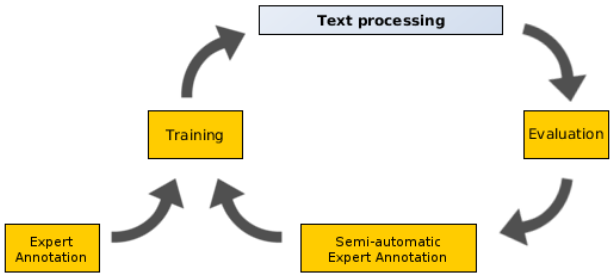

Figure 4: Annotation/Training-Cycle: Provision of annotations by human experts which are used to train models and resources used for text processing (see also Figure 2). Processed text and automatic annotations are iteratively evaluated and submitted to semi-automatic re-annotation by human experts to improve annotation performance after re-training.

community discourse.

Shared Task Challenges. A major driver in various areas of computer science research have been shared task challenges. They not only raise awareness for a specific problem, but also concisely define such problems and provide evaluation data for the participating teams. Such shared tasks allow comparison of different approaches in an objective manner, and could also serve as a platform for discussing, evaluating, and establishing community standards, e.g., for annotation data models. In the related domain of bioinformatics, and in English language medical NLP, repeated community challenges like BioCreative [23], i2b2 [51], or Clinical TempEval [4] have been established. Taking from these highly successful initiatives, German medical text mining could greatly benefit from such a shared task competition. As a requirement to such an initiative, of course, suitable corpora for given target entities or tasks would have to be created - yet with the immediate bonus of their direct use by the community. The annual focus on a particular NLP target would, in turn, lead to a gradual extension of the overall corpus and annotation data available. First studies even indicate that high quality annotations may be crowdsourced from the community of participants itself, helping to overcome the 'ground truth development bottleneck', that may hamper such shared efforts [50].

Dictionary Enrichment. Successful existing tools rely on dictionaries to identify medical entities. As single existing terminologies are often insufficient in their coverage of relevant entities and of terms used in clinical jargon, these terminologies have to be enriched with synonyms, abbreviations, and entities themselves. Options for this task include the expansion of the available vocabulary by integration of several ontologies [31], or the use of statistical machine learning methods on raw data [52]. Inline with [49], our experience has shown that especially expert annotations provided for even small

collections of sample documents from specific examinations can greatly improve annotation quality regarding terms and wording not represented in standard terminologies - for the respective source of documents [46]. As stated in Section 2, this observation makes it highly desirable to capture which medical specialty and type of documentation annotations within the (enriched) dictionaries relate to. Of course, the lack of freely available comprehensive ontologies, like SNOMED-CT, for German (see Section 2), is a major obstacle to any annotational effort. This current shortcoming may even make it necessary to include English versions of these ontologies for initial, manual annotation. Associating the German terms thus annotated with the identifiers linked to their English language counterparts in medical terminology, extensions to currently insufficient existing resources could be generated. Ideally, of course, the resulting lexica should be shared with the community.

Shared Resource Library. In the related area of bioinformatics, the open culture of sharing both data and tools has fueled computer aided analysis which in turn has come to speed biological discovery. Next to aforementioned shared task initiatives, an online repository of reusable applications, data sets, and data resources will be necessary. Such a platform could take inspiration from similar efforts such as GEO [2] or BioConductor [16] in the area of bioinformatics, or the myExperiment repository of scientific workflows [41]. In myExperiment, social features like shared tagging and rating, or giving credit to original authors in derivative work are coupled with the central aspects of versioned publication and retrieval of resources. Such sharing could also include models and tools derived from, and trained and tested on undisclosable corpora [22]. As stated above, however, we do believe that only the availability of sufficient original (de-identified) textual resources allows for both fully independent development of approaches, and fully independent (re)evaluation.

Automatic Data Privacy Classification and Declaration. Analogous to methods of de-identification, methods to automatically infer the privacy level of a given clinical data set, e.g., with respect to its level of k-anonymity [47], could help judge the constraints to be imposed on sharing and processing of the respective data set. This could especially allow to transfer some of the data to cloud infrastructures for complex processing - automatically evaluating policies for doing so w.r.t. the privacy level of the data. Complementary declarations of the maximum level of privacy at which a given algorithm needs data for proper operation (i.e., to deliver meaningful results) could help to evaluate a priori, whether, how, and under which circumstances a given, possibly sensitive clinical data set can be analyzed. With increasing size of digital clinical data to be processed, and increasing complexity of available algorithms, such mechanisms will become highly relevant in the foreseeable future.

5 Conclusion

In this survey, we have described the current obstacles inherent to the area of information extraction from clinical documents in German and have identified the major steps we deem necessary to overcome them. We argue that only a shared effort can generate results fast enough and of sufficient quality to bring the value of semantic integration to best use in the medical domain. The advent of publicly funded projects targeted at semantic integration of clinical documents in Germany, such as the BMBF funded i:DSem [5], is a good start, but for sustained benefit to clinical NLP, an effort has to be made to make both data and tools openly available to the public.

In the short term, clinical decision support systems based on semantically integrated information extracted from electronic health records could provide instant benefit for each single patient. Prospectively, in conjunction with genomic data, the phenotypic information mined could lead to completely new insights about how genetic regions or singular mutations map to physical representations [24]. To bring such benefits to patients in German speaking countries, the research community has to start making the necessary information, data, and knowledge accessible now.

Acknowledgement

This work was partly funded by BMBF grants PERSONS (031L0030B) and PREDICT (031L0023A), and by DFG grant SOAMED (GRK1651).

Literature

[1] Z. Afzal, E. Pons, N. Kang, M. C. Sturkenboom, M. J. Schuemie, and J. A. Kors. Contextd: an algorithm to identify contextual properties of medical terms in a dutch clinical corpus. *BMC bioinformatics*, 15(1):373, 2014.

[2] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, et al. NCBI GEO: archive for functional genomics data sets - 10 years on. *Nucleic acids research*, 39(suppl 1):D1005–D1010, 2011.

[3] R. C. Barrows Jr, M. Busuioc, and C. Friedman. Limited parsing of notational text visit notes: ad-hoc vs. nlp approaches. In *Proceedings of the AMIA Symposium*, page 51. American Medical Informatics Association, 2000.

[4] S. Bethard, L. Derczynski, G. Savova, G. Savova, J. Pustejovsky, and M. Verhagen. SemEval-2015 task 6: Clinical TempEval. *Proc. SemEval*, 2015.

[5] BMBF. Bekanntmachung des Bundesministeriums für Bildung und Forschung von Richtlinien zur Förderung von Projektideen im Rahmen der Massnahme "i:DSem – Integrative Datensemantik in der Systemmedizin". *https://www.bmbf.de/foerderungen/bekanntmachung-920.html*, 2014.

[6] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. Evaluation of negation phrases in narrative clinical reports. In *Proceedings of the AMIA Symposium*, page 105. American Medical Informatics Association, 2001.

[7] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[8] L. Deléger, C. Grouin, and P. Zweigenbaum. Extracting medication information from french clinical texts. *Stud Health Technol Inform*, 160(Pt 2):949–53, 2010.

[9] J. C. Denny, L. Bastarache, M. D. Ritchie, R. J. Carroll, R. Zink, J. D. Mosley, J. R. Field, J. M. Pulley, A. H. Ramirez, E. Bowton, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*, 31(12):1102–1111, 2013.

[10] L. Döhling and U. Leser. Extracting and aggregating temporal events from text. *4th Temporal Web Analytics Workshop*, pages 839–844, 2014.

[11] E. Faessler, J. Hellrich, and U. Hahn. Disclose models, hide the data – how to make use of confidential corpora without seeing sensitive raw data. *9th International Conference on Language Resources and Evaluation*, 2014.

[12] O. Ferrández, B. R. South, S. Shen, F. J. Friedlin, M. H. Samore, and S. M. Meystre. Evaluating current automatic de-identification methods with veteran's health administration clinical documents. *BMC medical research methodology*, 12(1):1, 2012.

[13] A. Forrey, C. McDonald, G. DeMoor, S. Huff, D. Leavelle, D. Leland, T. Fiers, L. Charles, B. Griffin, F. Stalling, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical Chemistry*, 42(1):81–90, 1996.

[14] C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

[15] C. Friedman, G. Hripcsak, L. Shagina, and H. Liu. Representing information in patient reports using natural language processing and the extensible markup language. *Journal of the American Medical Informatics Association*, 6(1):76–87, 1999.

[16] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[17] A. Gerstmair, P. Daumke, K. Simon, M. Langer, and E. Kotter. Intelligent image retrieval based on radiology reports. *European radiology*, 22(12):2750–2758, 2012.

[18] U. Hahn, M. Romacker, and S. Schulz. MedSynDIKATe - a natural language system for the extraction of medical information from findings reports. *International journal of medical informatics*, 67(1):63–74, 2002.

[19] M. A. Hamburg and F. S. Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

[20] D. A. Hanauer, Q. Mei, J. Law, R. Khanna, and K. Zheng. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the electronic medical record search engine (EMERSE). *Journal of biomedical informatics*, 55:290–300, 2015.

[21] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck. ProMiner: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6(1):1, 2005.

[22] J. Hellrich, F. Matthies, E. Faessler, and U. Hahn. Sharing models and tools for processing german clinical texts. *Studies in health technology and informatics*, 210:734, 2015.

[23] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of BioCreative: critical assessment of information extraction for biology. *BMC bioinformatics*, 6(Suppl 1):S1, 2005.

[24] P. B. Jensen, L. J. Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[25] D. Kalra and D. Ingram. Electronic health records. *Information technology solutions for healthcare*, pages 135–181, 2006.

[26] S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464, 2009.

[27] M. Kreuzthaler and S. Schulz. Disambiguation of period characters in clinical narratives. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL*, volume 100, 2014.

[28] H.-U. Krieger, C. Spurk, H. Uszkoreit, F. Xu, Y. Zhang, F. Müller, and T. Tolxdorff. Information extraction from german patient records via hybrid parsing and relation extraction strategies. In *LREC*, pages 2043–2048, 2014.

[29] C. A. Kushida, D. A. Nichols, R. Jadrnicek, R. Miller, J. K. Walsh, and K. Griffin. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, 50:S82–S101, 2012.

[30] V. Laippala, T. Viljanen, A. Airola, J. Kanerva, S. Salanterä, T. Salakoski, and F. Ginter. Statistical parsing of varieties of clinical finnish. *Artificial intelligence in medicine*, 61(3):131–136, 2014.

[31] Y. Lee, K. Supekar, and J. Geller. Ontology integration: experience with medical terminologies. *Computers in Biology and Medicine*, 36(7):893–919, 2006.

[32] Y. Li and H. Liu. Learning semantic tags from big data for clinical text representation. *AMIA Summits on Translational Science Proceedings*, 2015:461, 2015.

[33] D. Maynard, K. Bontcheva, and D. Rout. Challenges in developing opinion mining tools for social media. *Proceedings of the @NLP can u tag #user_generated_content*, pages 15–22, 2012.

[34] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):1, 2010.

[35] S. M. Meystre, S. Lee, C. Y. Jung, and R. D. Chevrier. Common data model for natural language processing based on two existing standard information models: CDA+GrAF. *Journal of biomedical informatics*, 45(4):703–710, 2012.

[36] S. Moon, S. Pakhomov, and G. B. Melton. Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1310. American Medical Informatics Association, 2012.

[37] M. Neves and U. Leser. A survey on annotation tools for the biomedical literature. *Briefings in bioinformatics*, page bbs084, 2012.

[38] Y. Ou and J. Patrick. Automatic negation detection in narrative pathology reports. *Artificial intelligence in medicine*, 64(1):41–50, 2015.

[39] J. P. Pestian, C. Brew, P. Matykiewicz, D. Hovermale, N. Johnson, K. B. Cohen, and W. Duch. A shared task involving multi-label classification of clinical free text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104, 2007.

[40] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615, 2008.

[41] D. Roure, C. Goble, and R. Stevens. The design and realisation of the myExperiment virtual research environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5):561–567, 2009.

[42] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

[43] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[44] I. Schlünder. Datenschutzkonforme Lösungen für die Versorgungsforschung. *14. Deutscher Kongress für Versorgungsforschung*, 2015.

[45] M. Skeppstedt, M. Kvist, G. H. Nilsson, and H. Dalianis. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. *Journal of biomedical informatics*, 49:148–158, 2014.

[46] J. Starlinger, B. T. Schmeck, and U. Leser. Challenges in automatic diagnosis extraction from medical examination summaries. *CIKM Workshop on Web Science and Information Exchange in the Medical Web*, 2011.

[47] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[48] D. Tikk and I. Solt. Improving textual medication extraction using combined conditional random fields and rule-based systems. *Journal of the American Medical Informatics Association*, 17(5):540–544, 2010.

[49] M. Toepfer, H. Corovic, G. Fette, P. Klügl, S. Störk, and F. Puppe. Fine-grained information extraction from german transthoracic echocardiography reports. *BMC medical informatics and decision making*, 15(1):1, 2015.

[50] Ö. Uzuner, I. Solti, F. Xia, and E. Cadag. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523, 2010.

[51] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[52] Y. Wu, J. Xu, Y. Zhang, and H. Xu. Clinical abbreviation disambiguation using neural word embeddings. *ACL-IJCNLP 2015*, page 171, 2015.

[53] W.-W. Yim, S. Kwan, and M. Yetisgen. In-depth annotation for patient level liver cancer staging. In *Sixth International Workshop on Health Text Mining and Information Analysis (LOUHI)*, page 1, 2015.

Dr.Ing. Dr.med. Johannes Starlinger is a Research Associate at the Department of Computer Science at Humboldt-Universität zu Berlin. After studying medicine at Medical University of Vienna, and Computer Science at HU-Berlin, he joined the DFG-funded SOAMED graduate program in 2010 to research service-oriented architectures in a medical area of application. He received his PhD from HU-Berlin in 2015. Johannes' current research focus is on similarity search over data relevant to the biomedical domain, including scientific workflows, genomic, and medical data.

Address: Humboldt-Universität zu Berlin, Institut für Informatik, Unter den Linden 6, 10099 Berlin, Germany, E-Mail: starling@informatik.hu-berlin.de

Dr.rer.nat. Madeleine Kittner studied chemistry at TU Berlin and University of Strathclyde Glasgow. In 2011, she received a PhD in biochemistry from Universität Potsdam, Germany. She has experience in analyzing transcriptomics data, signaling pathways and text mining of Dutch medical records. Currently, she is a research associate at the Department of Computer Science at Humboldt-Universität zu Berlin focusing on text mining of biomedical documents.

Address: Humboldt-Universität zu Berlin, Institut für Informatik, Unter den Linden 6, 10099 Berlin, Germany, E-Mail: kittnema@informatik.hu-berlin.de

Dr.med. Oliver Blankenstein is a pediatrician at the Department of Pediatric Endocrinology and Diabetology and the head of the Newborn Screening Laboratory at Charité Universitätsmedizin Berlin. He also heads the department of endocrinology and metabolism at Labor Berlin. He has served as PI in a number of randomized controlled clinical trials and, for several years, has been advising computer science researchers in the DFG RTG SOAMED.

Address: Charité Universitätsmedizin Berlin, Pädiatrische Endokrinologie und Diabetologie,Augustenburger Platz 1, 13353 Berlin, Germany, E-Mail: Oliver.Blankenstein@charite.de

Prof. Dr.Ing. Ulf Leser studied informatics at TU München and obtained his PhD from TU Berlin. In 2002 be became professor for Knowledge Management in Bioinformatics at HU Berlin. His highly interdisciplinary research focuses on scientific data management, statistical Bioinformatics, biomedical text mining, and scientific workflows. He is speaker of the DFG-RTG SOAMED and a board member of the DFG-excellence RTG BSIO.

Address: Humboldt-Universität zu Berlin, Institut für Informatik, Unter den Linden 6, 10099 Berlin, Germany, E-Mail: leser@informatik.hu-berlin.de