German Medical Natural Language Processing – A Data-centric Survey

Torsten Zesch[1] and Jeanette Bewersdorff[2]

[1] Computational Linguistics, CATALPA – Center for Advanced Technology-Assisted Learning and Predictive Analytics, FernUniversität in Hagen, Germany
torsten.zesch@fernuni-hagen.de
[2] RTG WisPerMed, University of Duisburg-Essen, Germany
jeanette.bewersdorff@uni-due.de

Abstract. Even though AI in general, and NLP in particular, has made a lot of progress in recent years, the impact on the processing of medical written data has so far been limited. We argue that this is mainly because publicly available data is scarce in the medical domain and thus provide an overview of available data sources as well as strategies to overcome data scarcity. We also discuss de-identification approaches and possible challenges when working with de-identified data. Finally, we give an overview of available German NLP models for the medical domain and discuss domain adaptation as a way to transfer models from a specific application area to another.

Keywords: language technology; medical NLP; German; datasets; domain adaptation

1 Introduction

Making sense of written medical data (e.g. from electronic patient records or laboratory analyses) is still a major challenge that gets even more difficult if we want to tackle languages other than English [35]. Medical NLP gained a lot of attention in recent years, e.g. in the form of re-occurring challenges like the National NLP Clinical Challenges (n2c2).[1] However, only very few medical datasets get released after being created, mainly because medical data contains sensitive personal information. When data is not available or cannot be shared, it has been proposed to instead share the resulting models [9, 19], however still have to care about model inversion attacks [9, 11], especially for overparametrized recent neural models. As a result, only very few datasets or models are available for public use. We thus give an comprehensive overview of the (very few) data sources in German medical NLP and discuss strategies to overcome data scarcity.[2]

2 Data Acquisition in Medical NLP

Even though AI in general, and NLP in particular, has made a lot of progress recently, the impact on the processing of medical written data has so far been limited. We argue that this is mainly because (as has been noted before for example by Lohr et al. [30]) publicly available data is scarce in the medical domain. To understand why, it is helpful to have a closer look at the data collection process.

Figure 1 presents an overview of the process of collecting or creating medical data. The origin of data can be divided into two sources: real data and synthetic data. **Real data** is collected as part of the patient treatment process in form of clinical notes and referral letters. In most cases, **de-identification** [24, 31, 37, 38] is required, as all personal information that may be contained in real data has to be removed. It is crucial to ensure that at no point a link between the data and the respective patient it originates from can be made. An alternative (that does not require de-identification) is using **synthetic data**. It can either be manually created (writing medical documents imagining patients and cases) or even automatically generated [29]. A natural source of manually created synthetic data are made-up reports and case studies written by medical professionals for educational purposes and published in medical textbooks [30], but also fictitious data written for the purpose of training a specific model [21]. The advantage of synthetic data, that it can be freely distributed without data protection issues, is countered by the looming question whether it closely enough resembles real data.

[1] https://n2c2.dbmi.hms.harvard.edu/

[2] Our focus on German is based on research within the WisPerMed DFG research training group ('Knowledge- and data-based personalization of medicine at the point of care', https://www.uni-due.de/grk_wispermed/grk_wispermed.php). The RTG combines the research expertise of Dortmund University of Applied Sciences and Arts, University of Duisburg-Essen, University Medical Center Essen and FernUniversität in Hagen. The overarching goal of the RTG is to make the knowledge contained in various data formats available and usable at the point of treatment for concrete individual therapy decisions. As a prototypical use case, the RTG is focusing on the treatment of malignant melanoma. All usable patient documentation in form of clinical notes is exclusively available in German.

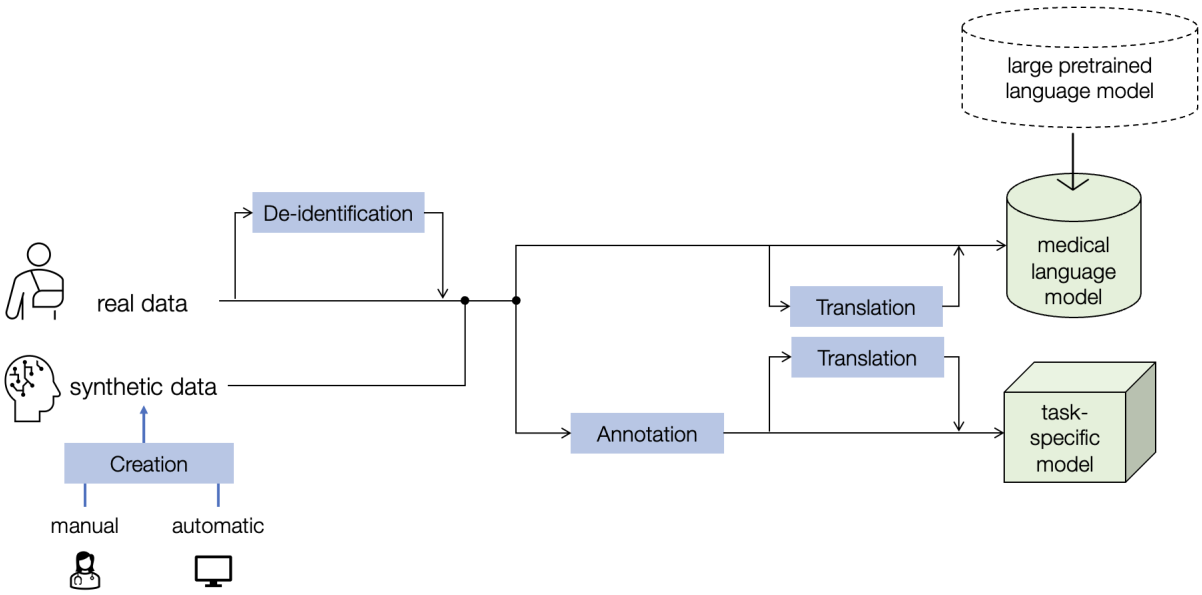

Fig. 1: Overview of collecting and creating data for medical NLP

As there is very little data in any given language, one might use **translation** to convert either real or synthetic data into the target language [13]. Finally, the raw data can be used to train a general purpose medical language model or the dataset needs additional **annotation** so that one can train a task-specific model.

We now give a more detailed overview of the individual steps involved in collecting and creating data for medical NLP.

2.1 De-identification

When working with *real data*, de-identification (also called anonymization) needs to be performed, before the data can be used. De-identification removes or replaces sensitive information such as the patient's name, their phone number, or address. Sensitive information items are collectively called *protected health information* (PHI) in the literature. For an overview of PHI types see Dernoncourt et al. [7]

For English, several de-identification approaches have been evaluated, for example [34, 45], especially as part of the i2b2 de-identification challenges [43].[3] Similar approaches have also been adapted to German [24, 38].

Table 1: Examples of different de-identification approaches, adapted from Berg et al. [2]

ORIGINAL	Mr. John Berry was prescribed 400mg of Ibuprofen today.
PSEUDO	Mr. Harry Miller was prescribed 400mg of Ibuprofen yesterday.
CLASS	Mr. {FirstName} {LastName} was prescribed 400mg of Ibuprofen {Date}.
MASK	Mr. XXX XXX was prescribed 400mg of Ibuprofen XXX.
REMOVE	-

Removing PHI from documents might however decrease the performance of NLP models. Berg et al. [2] found that the impact is significant and strongly depends on the de-identification strategy. Table 1 shows examples for the different methods. ORIGINAL shows the text as it was found in real medical data. The PSEUDO strategy replaces PHI with surrogates (i.e. another name/date/number etc.) that are as close to the intended meaning as possible depending on the context. For example, when replacing the age of a patient, granularity might be in decades for adults but that would not work well for young children where it makes quite a difference if they are 1 or 9 years old. CLASS replaces the PHI with a class marker, e.g. {FirstName} for given names like John or Mary. MASK hides the PHI by replacing it with some masking character, e.g. *X* or *#*. Finally, REMOVE just deletes any sentence

[3] The de-identification pipeline used for preparation of the i2b2 challenge is described in Stubbs and Özlem Uzuner [44].

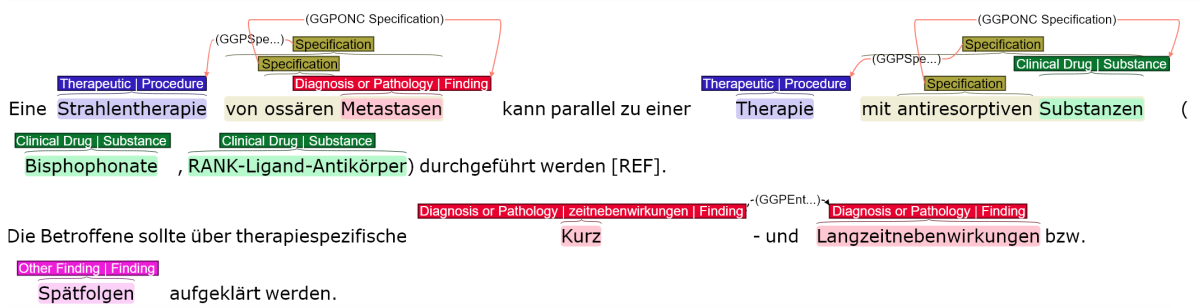

Fig. 2: Example of an annotated text from the GGPONC corpus created using the INCEpTION annotation platform (Source: `https://inception-project.github.io/use-cases/ggponc/`).

containing a single PHI from the corpus, which might be an unacceptable strategy if the the rate of PHIs in a text is high.

The study by Berg et al. [2] found that PSEUDO has the least impact on downstream task performance, while (unsurprisingly) REMOVE dramatically reduces performance (with the other two strategies in-between). However, at the same time it can be argued that PSEUDO has the weakest protection level and that surrogates have to be carefully designed to ensure de-identification.

So in de-identification, we not only have to find PHI, but also make an informed decision on how to retain as much information as possible, without compromising the de-identification itself.

2.2 Creating Synthetic Data

Synthetically created datasets have mainly been used in the domain of structured data [47]. The main goal is to transfer statistical properties (i.e. dependencies and distributions within the real data) to the synthetic data. While these methods cannot be used to create synthetic *texts*, other methods have been proposed for this purpose. Libbi et al. [29] compare an LSTM and a transformer-based generative model for creating synthetic medical care reports in Dutch. Guan et al. [15] propose to use generative adversarial networks (GANs) to generate Chinese electronic medical records.

As those methods have to be trained, at least some non-synthetic data is always required which could lead to a cold-start problem. It is also unclear, whether the automatically generated synthetic data is of high enough quality to be used as a replacement for manual synthetic data [15]. Another issue is that the generative model might leak PHI from the training data into the synthetic data [29].

2.3 Annotation

Real or synthetic data in its raw form is in most cases not enough to train a task-specific NLP model. What is also needed is *annotations* on the data, i.e. some kind of markup or codes that provide additional information that cannot be derived in some obvious fashion directly from the text. Usually this annotation process is thus performed manually. Figure 2 gives an example, where data from the GGPONC corpus [3] was annotated with SNOMED-CT classes classes using the INCEpTION platform [23].[4]

Some annotations are related to the linguistic properties, e.g. token, sentence, POS, or negation scope [48]). They do not directly correspond to a use case, but support downstream medical NLP tasks like concept extraction (e.g. findings, symptoms, drugs, diseases), relation extraction (e.g. adverse effects), or multi-label document classification (e.g. according to a scheme like ICD-10).[5] For a more comprehensive overview of different NLP tasks see for example Névéol et al. [35].

2.4 Translation

As so few datasets are available, it is an interesting strategy to translate data from other languages. If done manually, a high quality corpus can be created at very high costs, but it remains an open question whether the time-consuming effort to translate a dataset manually actually pays off or whether the synthetic creation or collection of new

[4] `https://inception-project.github.io/`
[5] `https://www.who.int/standards/classifications/classification-of-diseases`

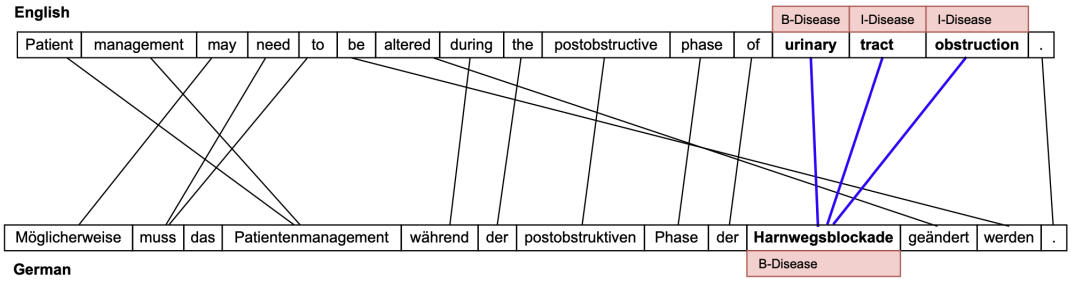

Fig. 3: Example of aligning original and translated sentence for annotation transfer (Figure by Schäfer et al. [41]).

data in the target language would be more efficient. Automatic machine translation can be used instead, which is much cheaper. If a remote translation service is used, only already de-identified data can be translated; otherwise, sensitive information could be disclosed. A work-around is to use a locally installed translation service, but the translation quality is usually lower. In any case, the automatic translation step is likely to introduce mistakes that will lower the quality of downstream task trained on that data [18].

When already annotated data is translated, a major challenge is to correctly transfer the annotations. Figure 3 shows an example provided by Schäfer et al. [41]. As the tokens do not align one-to-one between languages, it is challenging to correctly transfer the original annotation.

3 German Medical Corpora and Datasets

Table 2 gives an overview of the current situation for German medical corpora. A major distinction is whether a corpus is publicly available (upper part of the table) or not (lower part). The table is sorted by publication year, meaning that all more recently published corpora are available, while all older ones are not. Hopefully this represents a trend towards Open Science. As can be seen in the 'Origin' column, two of the available corpora are synthetic (GGPONC and JSYNCC), one is translated from English (GERNERMED) and only one contains real data (BRONCO150). The latter one is also the smallest, as it was manually de-identified. As medicine is a very broad field, the corpora cover a lot of domains (like radiology, surgery, or dermatology) and document types.

We now describe the four publicly available corpora in more detail:

GGPONC (German Guideline Program in Oncology NLP Corpus) is a corpus[6] consisting of 30 medical guidelines related to oncology. The corpus was created by Borchert et al. [3] and contained in its first release 25 guidelines. Version 2.0 added 5 more guidelines in 2022. The guidelines were manually annotated with SNOMED-CT classes (*finding, substance, procedure*). The corpus contains over 1.8 million tokens. More than 200,000 entities were annotated. As no sensitive data is contained, the corpus is publicly available for research purposes upon request.

BRONCO150 the Berlin-Tuebingen-Oncology Corpus[7] was released in 2021 by Kittner et al. [22]. It contains a selection of manually de-identified sentences from 150 discharge summaries, collected from melanoma or hepatocelluar carcinoma patients at Universitätsklinikum Tübingen and Charité Berlin. The full corpus consists of 200 documents, but only 150 were released (thus the name BRONCO150). The remaining 50 documents are disclosed for unbiased evaluation of future models.

To further increase the data protection, the order of all sentences from the 150 documents is randomized, so that a reconstruction of documents is nearly impossible. The resulting corpus contains 67,456 tokens. Annotations were made for diagnoses (following ICD10), treatments (following OPS) and medications (following ATC). To access the corpora, a user-agreement has to be signed.

GERNERMED is a translated corpus created by Frei and Kramer [13]. The English source data was taken from the 2018 ADE and medication extraction challenge (n2c2, Track 2) [20]. It contains 303 annotated discharge summaries with overall 172,695 tokens. The English source text was translated using the pretrained *transformer.wmt19.en-de* model from the Facebook fairseq model architecture [36]. Named entity annotations (drug, route, strength, frequency, duration, form, and dosage) were mapped to the respective positions in the translated documents using *FastAlign* [8], an unsupervised method for the word alignment between two

[6] https://www.leitlinienprogramm-onkologie.de/projekte/ggponc-deutsch/
[7] https://www2.informatik.hu-berlin.de/~leser/bronco/index.html

Table 2: Overview of German medical corpora and datasets.

Corpus	Year	Domain	Document Type	Origin	Available	# tokens [10^3]	# docs
GGPONC [3]	2022	oncology	guidelines	synthetic	yes	1800	-
BRONCO150 [22]	2021	oncology	discharge summaries	real	yes	70	150
GERNERMED [13]	2021	-	discharge summaries	real (translated)	yes	173	303
JSYNCC [30]	2018	surgery	mixed educational texts	synthetic	yes	313	867
3000PA [17]	2018	various	various	real	no	-	3000
Krebs et al. [25]	2017	radiology	reports	real	no	-	3000
Roller et al. [40]	2016	nephrology	clinical notes/ discharge summaries	real	no	158	1725
Cotik et al. [6]	2016	-	clinical notes/ discharge summaries	real	no	13	183
Lohr and Herms [32]	2016	surgery	intervention reports	real	no	266	450
Kreuzthaler et al. [26], Kreuzthaler and Schulz [27]	2016	dermatology	discharge summaries	real	no	-	1696
Toepfer et al. [46]	2015	cardiology	reports	real	no	-	140
Bretschneider et al. [4]	2013	radiology	reports	real	no	28	2713
Fette et al. [10]	2012	-	various	real	no	-	544
FraMed [49]	2004	-	various	real	no	100	-

languages. Additionally, as the original English data used masking for de-identification, pseudo names were introduced in the German translation to give it a more natural appearance. The corpus is available via Github.[8]

JSYNCC [30] is a synthetic German corpus.[9] It contains 867 documents with overall 312,784 tokens. It is based on ten medical e-books from different domains. The corpus itself consists of the various synthetic reports and discharge summaries contained in the e-books.[10] The conversion from the e-books to the final corpus is performed by a fully automated script, ensuring that everyone can recreate an exact copy of the corpus. The only prerequisite is that one needs to obtain the e-books.

4 Model and Tools for German

The list of publicly available German medical NLP tools and models is rather short:

German-MedBERT [42] is a finetuned version of the German BERT model and is publicly available on Huggingface.[11] For fine-tuning, medical reports and articles related to symptoms, diseases, and diagnoses were collected.

German NegEx by Cotik et al. [6] adapts the original NegEx [5], a regular expression algorithm developed to identify negations in English discharge summaries, to work on German clinical data by developing a publicly available German trigger set.[12] The trigger set is adapted to medical text by including also lexical items such as *gram-negativ* that are not found in standard trigger sets.

[8] https://github.com/frankkramer-lab/GERNERMED
[9] https://github.com/JULIELab/jsyncc
[10] We consider them as synthetic data, as they were written as examples for the text books.
[11] https://huggingface.co/smanjil/German-MedBERT
[12] http://macss.dfki.de/german_trigger_set.html

GERNERMED++ [12] is a named entity recognition model based on the dataset with the same name and the successor of the GERNERMED model [13]. It is available via Github and HuggingFace.[13]

JCoRe 2.0 the Julie Lab Component Repository by Hahn et al. [16] is a Java-based framework for creating NLP pipelines.[14] It also contains a German medical tokenizer, sentence splitter, and POS tagger [19].[15] The POS model was trained and evaluated on the non-publicly available FraMed corpus (see Table 2).

mEx [39] provides models for German POS tagging, NE recognition, and relation extraction.[16] It is based on a non-publicly available dataset (see Table 2 Roller et al.), so these models are good examples of the 'share the model, if you cannot share the data' paradigm.

Beyond that, there are quite some scientific papers on processing German medical text, e.g. on sentence boundary and abbreviation detection [26, 27], negation detection [6] and negation scope detection [14], or grammars and parsing [4, 21, 28], but to the best of our knowledge none of those are currently available for public use.

To improve this situation, a possible short-term strategy could be to rely more on translation from other languages or on synthetic training data.

4.1 Domain Adaptation

The German-MedBERT [42] model is already an example of domain adaption, where a general language model is fine-tuned to medical language. Another example is by Kara et al. [21], who adapt a standard German dependency parser (Stanford Core NLP) with relatively little in-domain training data. They show that the transfer learning model outperforms both the standard model as well as a model directly trained on the in-domain data. Apart from this domain shift from the non-medical into the medical domain, there is also an intra-medical domain shift when a model that is developed in one domain (e.g. multiple myeloma [33], colorectal cancer [1], nephrology [40], or radiology [4]) is applied in another. Frei et al. [12] find that their GERNERMED++ model performs much worse on out-of-distribution samples. Recognition performance drops considerably from 0.95 to 0.87 F_1 when applied to another domain.

In summary, it remains unclear if annotation efforts in one domain can be leveraged in another, or if each domain needs to train there one models (with the corresponding effort and data availability problems). With the very few available datasets, it is also challenging to even test the domain transfer capabilities of existing models.

5 Summary

Availability of suitable data is currently the major bottleneck for German medical NLP.

Even if more datasets have been made publicly available in the last few years, the amount and diversity of public data is still not sufficient. Releasing more real data depends on reliable de-identification and bears legal and ethical risks. Thus, using synthetic data and translating data from other languages are increasingly used (but still under-explored) strategies.

Besides the data scarcity issue, there is also a lack of high-quality models being made public for research purposes. In theory, models can be more freely distributed, even if the underlying training data cannot. However, in a practical setting there is always the fear that sensitive data could be exposed through the model. High-quality medical NLP models also have potentially high commercial value which also impedes open distribution.

When no big and comprehensive medical corpus is available, resulting models are necessarily domain-specific. Thus, domain adaptation is another under-explored area with some promising results, but it still requires at least some in-domain data.

Availability of suitable data is going to remain the major bottleneck for German medical NLP.

Acknowledgements

This work was funded by a PhD grant from the DFG Research Training Group 2535 'Knowledge- and data-based personalization of medicine at the point of care (WisPerMed)'. This work was partially conducted in the framework of CATALPA - Center for Advanced Technology-Assisted Learning and Predictive Analytics of the FernUniversität in Hagen, Germany. We thank Andrea Horbach and Christin Seifert for their insightful comments and suggestions.

[13] https://github.com/frankkramer-lab/GERNERMED-pp
[14] https://julielab.de/Resources/JCoRe.html
[15] https://github.com/JULIELab/jcore-pipelines/tree/master/jcore-medical-pos-pipeline
[16] https://github.com/DFKI-NLP/mEx-Docker-Deployment

Bibliography

1. 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. International Journal of Medical Informatics **127**, 141–146 (2019), ISSN 1386-5056
2. Berg, H., Henriksson, A., Dalianis, H.: The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pp. 1–11, Association for Computational Linguistics, Online (Nov 2020)
3. Borchert, F., Lohr, C., Modersohn, L., Langer, T., Follmann, M., Sachs, J.P., Hahn, U., Schapranow, M.P.: GGPONC: A Corpus of German Medical Text with Rich Metadata Based on Clinical Practice Guidelines. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pp. 38–48, Association for Computational Linguistics, Online (Nov 2020)
4. Bretschneider, C., Zillner, S., Hammon, M.: Identifying Pathological Findings in German Radiology Reports Using a Syntacto-semantic Parsing Approach. In: Proceedings of the 2013 Workshop on Biomedical Natural Language Processing, pp. 27–35, Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013)
5. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics **34**(5), 301–310 (2001), ISSN 1532-0464
6. Cotik, V., Roller, R., Xu, F., Uszkoreit, H., Budde, K., Schmidt, D.: Negation Detection in Clinical Reports Written in German. In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), pp. 115–124, The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016)
7. Dernoncourt, F., Lee, J.Y., Uzuner, O., Szolovits, P.: De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association **24**(3), 596–606 (12 2016), ISSN 1067-5027, doi:10.1093/jamia/ocw156
8. Dyer, C., Chahuneau, V., Smith, N.A.: A simple, fast, and effective reparameterization of ibm model 2. In: NAACL (2013)
9. Faessler, E., Hellrich, J., Hahn, U.: Disclose Models, Hide the Data - How to Make Use of Confidential Corpora without Seeing Sensitive Raw Data. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014)
10. Fette, G., Ertl, M., Wörner, A., Klügl, P., Stoerk, S., Puppe, F.: Information Extraction from Unstructured Electronic Health Records and Integration into a Data Warehouse (01 2012)
11. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. Proceedings of the ... USENIX Security Symposium. UNIX Security Symposium **2014**, 17–32 (aug 2014)
12. Frei, J., Frei-Stuber, L., Kramer, F.: GERNERMED++: Transfer Learning in German Medical NLP (2022)
13. Frei, J., Kramer, F.: GERNERMED – An Open German Medical NER Model (2021)
14. Gros, O., Stede, M.: Determining Negation Scope in German and English Medical Diagnoses, pp. 113–126. Brill, Leiden, The Netherlands (2014), ISBN 9789004258174
15. Guan, J., Li, R., Yu, S., Zhang, X.: A Method for Generating Synthetic Electronic Medical Record Text. IEEE/ACM Transactions on Computational Biology and Bioinformatics **18**, 173–182 (2021)
16. Hahn, U., Matthies, F., Faessler, E., Hellrich, J.: UIMA-Based JCoRe 2.0 Goes GitHub and Maven Central — State-of-the-Art Software Resource Engineering and Distribution of NLP Pipelines. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 2502–2509, European Language Resources Association (ELRA), Portorož, Slovenia (May 2016)
17. Hahn, U., Matthies, F., Lohr, C., Löffler, M.: 3000PA-Towards a National Reference Corpus of German Clinical Language. Studies in health technology and informatics **247**, 26–30 (01 2018)
18. Hayakawa, T., Arase, Y.: Fine-grained error analysis on English-to-Japanese machine translation in the medical domain. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pp. 155–164, European Association for Machine Translation, Lisboa, Portugal (Nov 2020)
19. Hellrich, J., Matthies, F., Faessler, E., Hahn, U.: Sharing models and tools for processing German clinical texts. Studies in health technology and informatics **210**, 734–738 (2015), ISSN 1879-8365 (Electronic)
20. Henry, S., Buchan, K., Filannino, M., Stubbs, A., Uzuner, Ö.: 2018 N2c2 Shared Task on Adverse Drug Events and Medication Extraction in Electronic Health Records. Journal of the American Medical Informatics Association : JAMIA (2020)

21. Kara, E., Zeen, T., Gabryszak, A., Budde, K., Schmidt, D., Roller, R.: A Domain-adapted Dependency Parser for German Clinical Text. In: KONVENS (2018)
22. Kittner, M., Lamping, M., Rieke, D.T., Götze, J., Bajwa, B., Jelas, I., Rüter, G., Hautow, H., Sänger, M., Habibi, M., Zettwitz, M., Bortoli, T.d., Ostermann, L., Ševa, J., Starlinger, J., Kohlbacher, O., Malek, N.P., Keilholz, U., Leser, U.: Annotation and initial evaluation of a large annotated German oncological corpus. JAMIA Open **4**(2) (04 2021), ISSN 2574-2531, ooab025
23. Klie, J.C., Bugert, M., Boullosa, B., de Castilho, R.E., Gurevych, I.: The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pp. 5–9, Association for Computational Linguistics (Juni 2018), veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018)
24. Kolditz, T., Lohr, C., Hellrich, J., Modersohn, L., Betz, B., Kiehntopf, M., Hahn, U.: Annotating German Clinical Documents for De-Identification. Studies in health technology and informatics **264**, 203–207 (aug 2019), ISSN 1879-8365 (Electronic)
25. Krebs, J., Corovic, H., Dietrich, G., Ertl, M., Fette, G., Kaspar, M., Krug, M., Stoerk, S., Puppe, F.: Semi-Automatic Terminology Generation for Information Extraction from German Chest X-Ray Reports. Studies in health technology and informatics **243**, 80–84 (01 2017)
26. Kreuzthaler, M., Oleynik, M., Avian, A., Schulz, S.: Unsupervised Abbreviation Detection in Clinical Narratives. In: Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), pp. 91–98, The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016)
27. Kreuzthaler, M., Schulz, S.: Detection of sentence boundaries and abbreviations in clinical narratives. BMC Medical Informatics and Decision Making **15**(2), S4 (2015), ISSN 1472-6947
28. Krieger, H.U., Spurk, C., Uszkoreit, H., Xu, F., Zhang, Y., Müller, F., Tolxdorff, T.: Information Extraction from German Patient Records via Hybrid Parsing and Relation Extraction Strategies. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 2043–2048, European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014)
29. Libbi, C.A., Trienes, J., Trieschnigg, D., Seifert, C.: Generating synthetic training data for supervised de-identification of electronic health records. Future Internet **13**(5) (2021), ISSN 1999-5903, doi:10.3390/fi13050136
30. Lohr, C., Buechel, S., Hahn, U.: Sharing Copies of Synthetic Clinical Corpora without Physical Distribution — A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSYNCC Corpus. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan (May 2018)
31. Lohr, C., Eder, E., Hahn, U.: Pseudonymization of PHI Items in German Clinical Reports. Studies in health technology and informatics **281**, 273–277 (may 2021), ISSN 1879-8365 (Electronic)
32. Lohr, C., Herms, R.: A Corpus of German Clinical Reports for ICD and OPS-based Language Modeling (05 2016)
33. 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 of information in medicine **55**(4), 373–380 (aug 2016), ISSN 2511-705X (Electronic)
34. Neamatullah, I., Douglass, M., Lehman, L.: Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 8, 32 (2008)
35. Névéol, A., Dalianis, H., Velupillai, S., Savova, G., Zweigenbaum, P.: Clinical Natural Language Processing in languages other than English: opportunities and challenges. Journal of Biomedical Semantics **9**(1) (2018)
36. Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M.: fairseq: A fast, extensible toolkit for sequence modeling (2019)
37. Richter-Pechanski, P., Amr, A., Katus, H.A., Dieterich, C.: Deep Learning Approaches Outperform Conventional Strategies in De-Identification of German Medical Reports. Studies in health technology and informatics **267**, 101–109 (sep 2019), ISSN 1879-8365 (Electronic)
38. Richter-Pechanski, P., Riezler, S., Dieterich, C.: De-Identification of German Medical Admission Notes. Studies in health technology and informatics **253**, 165–169 (2018), ISSN 1879-8365 (Electronic)
39. Roller, R., Alt, C., Seiffe, L., Wang, H.: mEx - An Information Extraction Platform for German Medical Text. In: Proceedings of the 11th International Conference on Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4HCLS'2018), Antwerp, Belgium (Dec 2018)
40. Roller, R., Uszkoreit, H., Xu, F., Seiffe, L., Mikhailov, M., Staeck, O., Budde, K., Halleck, F., Schmidt, D.: A fine-grained corpus annotation schema of German nephrology records. In: Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), pp. 69–77, The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016)
41. Schäfer, H., Idrissi-Yaghir, A., Horn, P., Friedrich, C.: Cross-Language Transfer of High-Quality Annotations: Combining Neural Machine Translation with Cross-Linguistic Span Alignment to Apply NER to Clinical

Texts in a Low-Resource Language. In: Proceedings of the 4th Clinical Natural Language Processing Workshop, pp. 53–62, Association for Computational Linguistics, Seattle, WA (Jul 2022)
42. Shrestha, M.: Development of a Language Model for Medical Domain. masterthesis, Hochschule Rhein-Waal (2021)
43. Stubbs, A., Kotfila, C., Uzuner, Ö.: Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform. 2015;58 Suppl(Suppl):S11-S19 (2015)
44. Stubbs, A., Özlem Uzuner: Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. Journal of Biomedical Informatics **58**, S20–S29 (2015), ISSN 1532-0464, doi:https://doi.org/10.1016/j.jbi.2015.07.020, supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data
45. Szarvas, G., Farkas, R., Busa-Fekete, R.: State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. Journal of the American Medical Informatics Association **14**(5), 574–580 (09 2007), ISSN 1067-5027
46. Toepfer, M., Corovic, H., Fette, G., Klügl, P., Störk, S., Puppe, F.: Fine-grained information extraction from German transthoracic echocardiography reports. BMC Medical Informatics Decis. Mak. **15**, 91 (2015)
47. Tucker, A., Wang, Z., Rotalinti, Y., Myles, P.: Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. Digit. Med. **3**(147) (2020)
48. Vincze, V., Szarvas, G., Farkas, R., Móra, G., Csirik, J.: The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics **9**(11), S9 (2008), ISSN 1471-2105, doi:10.1186/1471-2105-9-S11-S9
49. Wermter, J., Hahn, U.: An Annotated German-Language Medical Text Corpus as Language Resource (01 2004)