Automatic extraction of clinically relevant parameters for tumor board decision support

Automatische Extraktion klinisch relevanter Parameter zur Unterstützung der Entscheidungsfindung im Tumorboard

Imge Yüzüncüoglu[1]
Yuxuan Chen[1]
A. Altar Lüser[1]
Nikitha Shruthi Ramasetti[2]
Robert Oehring[2]
Felix Krenzien[2]
Philippe Thomas[1]
Sebastian Möller[1,3]
Roland Roller[1]

1 German Research Centre for Artificial Intelligence (DFKI) Berlin, Germany
2 Charité – Universitätsmedizin Berlin, Germany
3 Technische Universität Berlin, Germany

Abstract

Introduction: Clinical decision-making relies on information documented in unstructured reports, which are often non-standardized and difficult to process efficiently. This challenge is particularly critical in multidisciplinary tumor board meetings, where decisions must be made under time constraints and based on incomplete information. Automated information extraction (IE) as presented in **ADBoard** supports structuring relevant clinical data for decision-making.
Methods: We developed and evaluated an IE pipeline for liver tumor cases, focusing on hepatocellular carcinoma, cholangiocarcinoma, and colorectal liver metastases. A dataset of 743 clinical documents from 134 patients was manually annotated to create ground truth labels for clinically relevant parameters. The task was formulated as named entity recognition (NER). We compared rule-based methods (regular expressions), transformer-based models, and an exploratory large language model (LLM) under realistic clinical constraints.
Results: Transformer-based models achieved the strongest overall performance, particularly for frequent parameters. Regular expressions performed competitively for well-defined patterns and proved robust in low-resource settings. The LLM showed inconsistent performance and did not outperform the other methods. Across all approaches, performance was also affected by data imbalance and annotation variability.
Discussion: Our findings indicate that, under the evaluated realistic, local deployment constraints, the exploratory LLaMA 3.1 8B setup did not outperform task-specific transformer or rule-based approaches for specialized clinical IE. Instead, a hybrid approach combining transformer-based and rule-based methods might be most effective. Furthermore, evaluation methodology substantially influences performance interpretation, as *strict* span-based metrics can underestimate clinically relevant results. These results highlight that improving annotation quality and evaluation design is as important as advances in model architecture for clinical IE.

Keywords: natural language processing, information extraction, hepatocellular carcinoma, cholangiocarcinoma, colorectal liver metastases, sequence evaluation, tumor board, clinical decision support

Zusammenfassung

Einleitung: Die klinische Entscheidungsfindung stützt sich auf Informationen, die in unstrukturierten Patientenberichten dokumentiert sind, die oft nicht standardisiert und nur schwer effizient zu verarbeiten sind. Diese Herausforderung ist besonders kritisch bei multidisziplinären Tumorboard-Sitzungen, bei denen Entscheidungen unter Zeitdruck und auf Grundlage unvollständiger Informationen getroffen werden müssen. Die automatisierte Informationsextraktion (IE), wie sie in **ADBoard** vorgestellt wird, unterstützt die Strukturierung relevanter klinischer Daten für die Entscheidungsfindung.

Methoden: Wir haben eine IE-Pipeline für Lebertumorfälle entwickelt und evaluiert, wobei der Schwerpunkt auf dem hepatozellulären Karzinom, dem Cholangiokarzinom und kolorektalen Lebermetastasen lag. Ein Datensatz mit 743 klinischen Dokumenten von 134 Patienten wurde manuell annotiert, um Ground-Truth-Labels für klinisch relevante Parameter zu erstellen. Die Aufgabe wurde als Named-Entity-Recognition (NER) formuliert. Wir verglichen regelbasierte Methoden (reguläre Ausdrücke), Transformer-basierte Modelle und ein exploratives großes Sprachmodell (LLM) unter realistischen klinischen Rahmenbedingungen.
Ergebnisse: Transformer-basierte Modelle erzielten die beste Gesamtleistung, insbesondere bei häufig vorkommenden Parametern. Reguläre Ausdrücke schnitten bei klar definierten Mustern gut ab und erwiesen sich in Umgang mit begrenzten Ressourcen als robust. Das LLM zeigte eine uneinheitliche Leistung und konnte die anderen Methoden nicht übertreffen. Bei allen Ansätzen wurde die Leistung zudem durch Datenungleichgewichte und Schwankungen bei den Annotationen beeinträchtigt.
Diskussion: Unsere Ergebnisse deuten darauf hin, dass unter den untersuchten realistischen, lokalen Einsatzbedingungen die explorative LLaMA-3.1-8B-Konfiguration bei der spezialisierten klinischen IE keine bessere Leistung erzielte als aufgabenspezifische Transformer- oder regelbasierte Ansätze. Stattdessen könnte ein hybrider Ansatz, der Transformer-basierte und regelbasierte Methoden kombiniert, am effektivsten sein. Darüber hinaus hat die Bewertungsmethodik einen erheblichen Einfluss auf die Interpretation der Leistung, da *strenge*, auf Sequenzen basierende Metriken klinisch relevante Ergebnisse unterbewerten können. Diese Ergebnisse unterstreichen, dass die Verbesserung der Annotationsqualität und des Bewertungsdesigns für die klinische IE ebenso wichtig ist wie Fortschritte in der Modellarchitektur.

Schlüsselwörter: natürliche Sprachverarbeitung, Informationsextraktion, hepatozelluläres Karzinom, Cholangiokarzinom, kolorektale Lebermetastasen, Sequenzauswertung, Tumorboard, Unterstützung bei der klinischen Entscheidungsfindung

1 Introduction

Within the medical domain, patient information, such as clinical condition, treatments, and medication, is often documented in non-standardized, unstructured reports. Although these reports contain essential information for clinical decision-making, extracting relevant data remains time-consuming and error-prone. This challenge is particularly critical in multidisciplinary tumor board meetings, where treatment decisions must be made under time constraints and often with incomplete information due to administrative or procedural reasons. To support physicians in this setting, we developed **ADBoard**, a clinical decision support system for liver tumor treatment recommendations [1]. The system targets hepatocellular carcinoma, cholangiocarcinoma, and colorectal liver metastases. It extracts clinically relevant parameters from unstructured reports and transforms them into a structured and consolidated tumor board protocol, complemented by treatment recommendations. Some of these parameters are tumor-specific findings (e.g., *tumor size*, *tumor count*, and *TNM staging*) as well as clinically relevant factors such as *ascites*, and *vascular invasion*, which are routinely considered during tumor board decision-making. The extracted information is presented in form of a normalized, interpretable tumor board protocol to support decision-making during tumor board discussions [2].

In this work, we focus on the information extraction (IE) component, which aims to derive structured information from unstructured clinical text. As we extract predefined clinical parameters, we formulate the task as named entity recognition (NER), where entities of interest are identified and extracted [3], [4]. We compare rule-based, transformer-based, and an exploratory large language model (LLM)-based approaches under realistic clinical constraints. Our contributions are as follows:

1. a comparative evaluation of rule-based, transformer-based, and LLM-based approaches for German clinical IE;
2. an analysis of the role of annotation quality in clinical extraction tasks; and
3. a discussion of practical limitations of current approaches in real-world settings.

2 Tumor board data

Data description

Treatment decisions for liver tumor patients are based on a set of clinically relevant parameters, including tumor-independent factors (e.g., *cirrhosis*) and tumor-specific characteristics that vary across mentioned tumor types. Identifying these parameters is essential for guideline-compliant decision-making in tumor board settings. To extract such information from unstructured clinical reports, we iteratively defined a set of relevant parameters motivated by the guideline program [5] together with clinical experts. The parameters include numerical values (e.g., *tumor size*, *number of tumors*), binary indicators (e.g., presence of *cirrhosis*), and categorical or free-text information (e.g., *TNM staging*). In total, 31 parameters, represented in Table 1, were defined. Due to class imbalance, varying clinical relevance, and limited representation, we mainly focus on a subset of 14 key parameters (highlighted in Table 1) that are consistently relevant across the three mentioned tumor types. These parameters were selected based on two criteria:

1. their relevance for treatment decisions according to clinical guidelines and tumor board workflows, and
2. expert assessment by participating physicians regarding their practical importance and informational value in routine clinical decision-making.

Data generation

We constructed a pseudonymized German clinical dataset, approved for scientific use by the relevant ethics committee (EA4/169/22). Development started with hepatocellular carcinoma due to higher data availability, starting with an initial corpus of 144 unstructured clinical documents, including discharge letters, radiology reports, pathology reports, and surgical reports. For experiments, we focused on radiology and pathology reports, as these are available in the targeted real-time clinical decision support setting. To enable training and evaluation of extraction methods, the documents were manually annotated to create a reference dataset, i.e. clinically relevant parameters were explicitly marked in the text, defining the expected correct information ("ground truth"). Annotation was performed by a single annotator using INCEpTION [6]. Hence, annotation omissions, boundary inconsistencies, and differences in annotation interpretation may have influenced both model training and evaluation. Due to class imbalance, with some parameters occurring fewer than five times, additional radiology reports were incorporated iteratively as they became available and were annotated using the same annotation scheme and guidelines. The final dataset comprises 743 documents from 134 patients (581 radiology, 162 pathology reports). As some parameters were introduced at later stages, previously annotated documents were not retrospectively re-annotated for these parameters. Consequently, not all documents contain annotations for all parameters, which is considered in the error analysis.

3 Methodology

This section provides an overview of the experimental setup and the evaluated IE approaches. After generating the dataset and performing some preprocessing, we compared three methodological paradigms for clinical IE: regular expression (RegEx) patterns, transformer-based sequence labelling models, and an exploratory prompt-based large language model (LLM) baseline. All approaches are trained and evaluated using the same data splits and evaluation framework, enabling a consistent comparison of symbolic, statistical, and generative methods for extracting clinical IE.

3.1 Preprocessing

To enable consistent processing and comparison across different extraction methods, the dataset was preprocessed prior to experimentation. Sentence segmentation by INCEpTION was refined using RegEx postprocessing to account for frequent clinical abbreviations (e.g., *Pat.*, *klin.*), which can otherwise lead to incorrect sentence boundaries. All annotations were converted into the BIO tagging scheme to support token-level sequence labeling and ensure comparability across approaches [4]. Finally, the dataset was split at document level into training (476 documents), development (119 documents), and test (148 documents) sets. Rather than using a purely proportional split, we aimed to ensure sufficient representation of relevant parameters in each subset. A patient-level split was not applied because, given the limited cohort size and the low prevalence of several parameters, this would have resulted in some extraction targets being insufficiently represented or absent from individual subsets, limiting their suitability for model development and evaluation. Consequently, 59 of 134 patients (44.0%) occur in more than one of the training, development, and test subsets. This patient overlap may lead to optimistic performance estimates and limits conclusions regarding generalization to previously unseen patients. Despite efforts to balance the parameters across the subsets, the data remains naturally imbalanced, with some parameters (e.g. portal hypertension) being underrepresented (see Table 1).

3.2 Symbolic, statistical, and generative IE approaches

Regular expressions

RegEx are a well-established method for IE [3]. Patterns were iteratively developed exclusively on the training set using domain knowledge from clinical experts and inspection of annotated data. To improve precision, document structure was incorporated by restricting extraction to

Table 1: Overview of the F1 score evaluation results for each parameter, based on the XLM-RoBERTa, RegEx and LLaMA approaches.

The following are depicted: The support of the parameter within the test set (#), as well as the *lenient* (L), *strict* (S) and *manual* (M) evaluation results. Grey highlighted rows mark the 14 parameters we focused on. Although no evaluation results were yet available for Child-Pugh and LiMAx, these were included to provide a comprehensive overview of the parameters.

		F1 scores								
		XLM-RoBERTa			RegEx			LLaMA		
Parameters	#	L	S	M	L	S	M	L	S	M
Ascites	26	**0.71**	**0.37**	0.83	0.56	0.21	0.75	0.67	0.34	**0.86**
Benign lesion	2	0	0	–	**0.09**	**0.06**		–	–	–
Child-pugh	–	–	–	–	–	–	–	–	–	–
Cholestasis	9	**0.50**	**0.50**	–	0.12	0.04	–	–	–	–
Cirrhosis	11	**0.78**	**0.67**	0.83	0.71	0.59	0.74	0.67	0.57	**0.84**
Collateral circulation	5	0.20	0	0.33	**0.33**	**0.21**	**0.53**	0.05	0	0.04
Clinical diagnosis	24	**0.42**	**0.40**	–	0	0	–	–	–	–
Differentiation	10	**0.67**	**0.67**	–	0.46	0.26	–	–	–	–
Distant metastasis	8	0.17	**0.17**	0.17	**0.22**	0.10	**0.38**	0.14	0.09	0.31
Hepatitis	3	0.16	0	–	**0.41**	0.29	–	–	–	–
Imaging modality	105	**0.92**	**0.91**	–	0.13	0	–	–	–	–
Limax	–	–	–	–	–	–	–	–	–	–
Lymph nodes	41	**0.54**	**0.32**	–	0.20	0.03	–	–	–	–
Negation	43	**0.19**	**0.16**	–	0	0	–	–	–	–
Organ measurement	33	**0.78**	**0.47**	–	0	0	–	–	–	–
Origin specimen	56	**0.63**	**0.58**	–	0.13	0.03	–	–	–	–
Pe	15	0.35	0.29	–	**0.60**	**0.57**	–	–	–	–
Portal hypertension	2	0.67	0.67	0.67	**0.80**	**0.80**	**0.80**	0.17	0	0.50
Portal vein thrombosis	21	**0.60**	**0.22**	**0.76**	0.25	0.06	0.55	0.24	0.02	0.38
Resection	44	**0.59**	**0.42**	**0.70**	0.13	0	0.31	0.15	0.10	0.35
Scanned bodypart	92	**0.92**	**0.92**	–	0	0	–	–	–	–
Staining	23	**0.72**	**0.54**	–	0.18	0.11	–	–	–	–
Tumor count	14	**0.64**	**0.64**	**0.64**	0.14	0.14	0.17	0.04	0.04	0.18
Tumor description	87	**0.42**	**0.16**	–	0	0	–	–	–	–
Tumorhood evidence	10	**0.61**	**0.61**	–	0.22	0.10	–	–	–	–
Tumor location	89	**0.46**	**0.37**	**0.55**	0.41	0.27	0.47	0.23	0.17	0.54
Tnm	7	**0.92**	**0.80**	**0.93**	0.56	0.12	0.88	0.33	0	0.75
Tumor reference	61	**0.53**	**0.52**	–	0	0	–	–	–	–
Tumor measurement	84	**0.73**	**0.73**	**0.73**	0.26	0.22	0.59	0.51	0.51	0.68
Tumor entity	9	**0.16**	**0.13**	–	0.11	0.10	–	–	–	–
Vascular invasion	10	**0.56**	**0.27**	**0.87**	0.13	0.05	0.09	0.04	0.01	0.09

relevant sections (e.g., *diagnosis, microscopy, macroscopy*). In total, 38 patterns were defined. For negation detection, used to determine the existence or non-existence of binary parameters such as *ascites* and *vascular invasion*, we implemented **pynegex** (https://pypi.org/project/pynegex/). Documents were processed at the sentence level, matching each sentence against the defined patterns.

Transformer models

Transformer-based models, particularly BERT variants, are widely used for NER tasks and are often state-of-the-art methods [7], [8], [9]. We evaluate three models:

1. google-bert/bert-base-german-cased [10] as a German general-domain baseline,
2. GerMedBERT/medbert-512 [11] as a German domain-specific medical BERT variant, and
3. FacebookAI/xlm-roberta-base [12] as a multilingual transformer model supporting German,

hereafter referred to as BERT, MedBERT, and XLM-RoBERTa, respectively. All models were trained on the

training set and tuned on a development set using AdamW optimization, cross-entropy loss, and early stopping based on F1 score (patience of six epochs). Input sequences were truncated or padded to the maximum model length. Hyperparameter tuning considered learning rates of 3e-5 and 5e-5, batch sizes of 16 and 32, and maximum sequence lengths of 256 and 512 tokens. To assess the impact of input granularity, experiments were conducted at both sentence and document level. Sentence-level processing treats each sentence independently, whereas document-level processing uses entire reports to capture broader contextual dependencies. The best-performing configuration for each transformer model was selected based on development-set performance and subsequently evaluated on the test set.

Exploratory LLM baseline

Model selection was constrained by two requirements:

1. All processing had to be performed locally due to sensitive patient data, and
2. models had to be deployable in real-world clinical settings, limiting the use of large models.

We therefore evaluated LLaMA 3.1 8B (LLaMA) [13], a lightweight open-source model suitable for local deployment. Documents were provided as full input to enable the model to leverage document-level context. Prompts were manually designed as German **zero-shot prompts** for each extraction parameter. They instruct the model to act as an experienced German physician reviewing clinical reports in a tumor board context. For each parameter, a separate inference call was performed on the full document, and the model was asked to extract the relevant information and return only JSON output. Inference was performed locally with a temperature of 0 and a maximum generation length of 512 tokens. Other decoding parameters were left at the backend defaults. For example, for tumor size extraction, the model was instructed to return a list under the key TSIZE, with each entry containing the extracted size including its unit and the original sentence from which it was extracted. If no tumor size was found, the model was instructed to return an empty list. Generated outputs were subsequently parsed, validated for the expected parameter format, and mapped back to BIO spans for evaluation.

3.3 Evaluation metrics

Evaluation of NER in clinical text is challenging due to inconsistencies, particularly in span boundaries (e.g., missing units such as mm or cm). To address this, we employ three evaluation strategies: *strict*, *lenient*, and *manual*, reporting performance using the F1 score, as it provides a balanced measure of precision and recall, both important in the clinical domain. The *strict* evaluation requires exact matches between prediction and ground truth and is highly sensitive to boundary deviations. In contrast, *lenient* evaluation measures character-level overlap using the *Jaccard Index* [14]. For the 14 key parameters, we additionally conducted a *manual* evaluation based on clinically validated reference values provided by the participating physicians. Predictions were primarily reviewed by the computer scientists and considered correct if the relevant clinical information matched these reference values, irrespective of exact span boundaries. Ambiguous cases were discussed with the clinical experts, whose assessment was considered decisive. These complementary evaluation strategies account for annotation variability and provide a more realistic assessment of IE performance in clinical settings.

4 Results discussion and error analysis

4.1 Transformer model comparison

Figure 1 summarizes the test set performance of the evaluated transformer models. MedBERT achieved the highest aggregate F1 scores (51% at sentence level and 47% at document level). All models achieved comparable performance, with the largest difference between MedBERT and XLM-RoBERTa at sentence level being 5 percentage points only. Across all transformer variants, sentence-level IE slightly outperformed document-level IE, suggesting that shorter and more focused contexts may facilitate parameter extraction compared to longer and potentially noisier reports. However, the results should be interpreted considering the dataset characteristics: Several parameters were underrepresented, making performance estimates sensitive to individual prediction errors and class imbalance.

4.2 Per parameter evaluation

The detailed per-parameter results are shown in Table 1 for XLM-RoBERTa, RegEx, and LLaMA. While Figure 1 shows that MedBERT achieves the strongest aggregate transformer performance, the differences between the transformer models are relatively small. XLM-RoBERTa is used as the representative transformer model in Table 1 because detailed manually reviewed parameter-level results were available for this model. Thus, Figure 1 provides the overall transformer comparison, whereas Table 1 enables a detailed parameter-level comparison with RegEx and LLaMA. The LLM evaluation is restricted to the subset of 14 focused parameters.

Comparing evaluation methods

We observe substantial discrepancies between *strict*, *lenient*, and *manual* evaluation, indicating that the choice of metric strongly affects performance interpretation. For example, XLM-RoBERTa, RegEx, and LLaMA achieve F1 scores of 37%, 21%, and 34% under *strict* evaluation for IE of *ascites*, compared to 71%, 56%, and 67% under *lenient* and 83%, 75%, and 86% under *manual* evaluation.

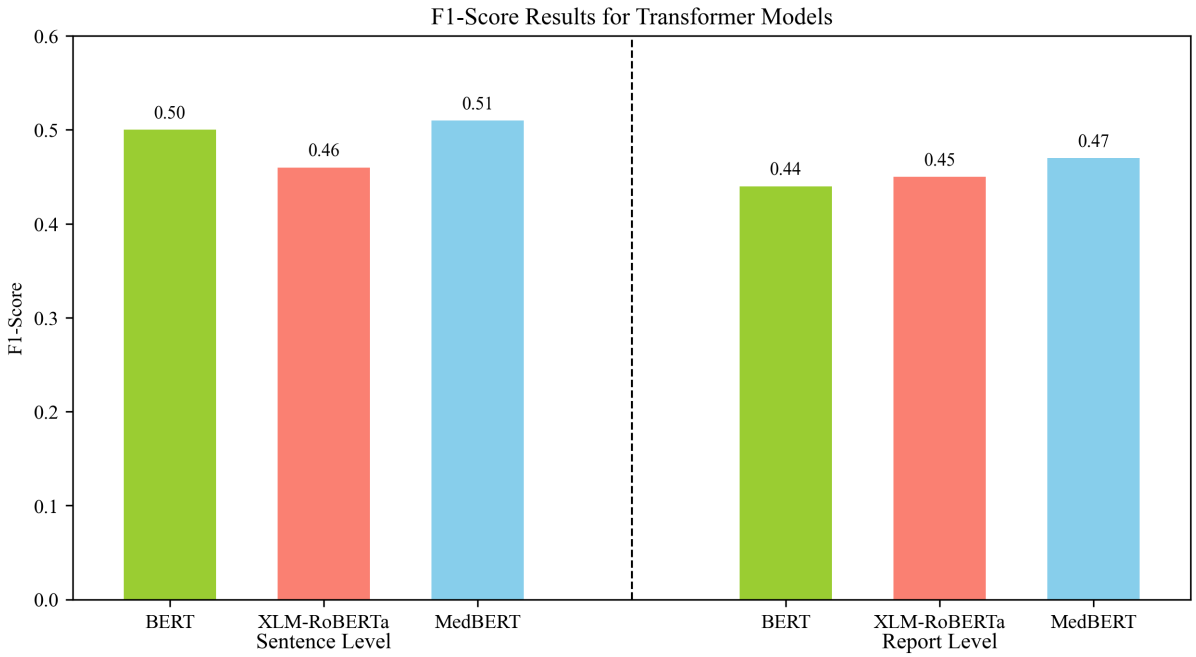

Figure 1: The aggregated micro-F1 scores of BERT, XLM-RoBERTa and MedBERT on the test set are compared across sentence- and document-level extraction using *lenient* evaluation.

Across the 14 parameters, differences of at least ten percentage points between *strict* and *manual* evaluation are observed for most parameters. *Strict* span-based evaluation tends to underestimate clinically relevant performance, while *manual* evaluation better reflects practical utility. *Lenient* evaluation provides an intermediate perspective by partially rewarding overlapping spans as they frequently lie between *strict* and *manual* results and do not consistently approximate either. These results highlight that evaluation protocols are a critical factor in assessing clinical IE systems. In settings with free-text data and non-standardized annotations, *strict* span-based metrics may not adequately reflect clinically meaningful performance. This points to an open challenge in clinical NLP: designing evaluation schemes that capture relevant information beyond exact span boundaries.

Comparing F1 scores

Several parameters are underrepresented in the dataset, making evaluation sensitive to individual prediction errors. Based on manual evaluation, LLaMA performs slightly better than XLM-RoBERTa and RegEx for only two parameters *(ascites* and *cirrhosis*), but the differences are marginal (3 and 1 percentage points) and do not indicate a meaningful advantage. Overall, XLM-RoBERTa achieves the best performance across most parameters, outperforming RegEx on nine out of 14 key parameters. However, results vary considerably depending on data availability. For example, RegEx outperforms XLM-RoBERTa for *portal hypertension*, but this parameter occurs only twice, limiting the reliability of this observation. This highlights the strong dependence of transformer-based models on data quantity and quality. Frequently occurring parameters are learned effectively, whereas rare parameters remain challenging. In our exploratory LLM setup, zero-shot prompt-based IE did not outperform the transformer or rule-based approaches. Additionally, results were sensitive to prompt formulation, limiting reproducibility, and the model's opacity made error analysis and controlled improvements difficult. These results should be interpreted as evidence of limitations of the local deployment scenario, rather than as a general conclusion about the reliability of LLMs for clinical IE.

4.3 Error analysis

To better understand the limitations of the evaluated approaches, we conducted a comprehensive error analysis during manual evaluation. The error categories identified were found to be consistent across transformer-based approaches and are therefore discussed at the methodological level rather than for a single transformer variant. The analysis has shown that model limitations, parameter-specific errors, and annotation errors contribute to incorrect predictions.

Model limitations

The approaches exhibit different limitations. RegEx performs well for structured patterns but lacks contextual understanding, resulting in false positives for ambiguous terms (e.g., *Flüssigkeit* (fluids) but not related to *ascites)*. Transformer models, in contrast, struggle with rare parameters and complex clinical language, particularly in cases requiring contextual interpretation such as negation or implicit references.

Representative parameter-level errors

Certain parameters are particularly prone to errors. *Tumor size* is often over-predicted, as models incorrectly label related measurements (e.g., *organ size* or *resection size*) as *tumor size*. Binary parameters like *ascites* and *cirrhosis* are frequently affected by negation errors, leading to false positives despite explicit negation. *Distant metastasis* remains challenging due to ambiguous, context-dependent expressions requiring additional contextual information beyond annotated spans.

Annotation-related errors

Annotation quality may have influenced both model training and evaluation. During manual inspection of prediction errors, we observed cases in which model outputs appeared clinically plausible despite the absence of corresponding annotations. For example, the phrase *Keine freie Flüssigkeit im kleinen Becken* (No free fluid in the pelvis) was predicted to be relevant to *ascites*, although no corresponding annotation was present in the reference dataset. This particularly affected parameters introduced later in the annotation process, as earlier documents had not been retrospectively re-annotated. Consequently, missing annotations could not always be distinguished from true negative labels during automated training and evaluation. While such observations do not allow conclusions about the overall annotation quality, they suggest that annotation omissions may contribute to some apparent false positives. We further observed inconsistencies related to span boundaries and negation handling. In some cases, negated expressions were annotated without including the negation cue itself, which may lead to mismatches under *strict* span-level evaluation. Consequently, part of the observed error rate may reflect differences between model predictions and annotation conventions rather than purely incorrect IE. All in all, the observed errors likely reflect a combination of model limitations, annotation related factors, data imbalance, and the inherent complexity of clinical language.

5 Related work

Clinical IE has traditionally been addressed using rule-based and machine learning based approaches. Earlier studies show that rule-based systems dominated the field [15], while more recent work highlights persistent challenges such as limited annotated data, restricted access to clinical corpora, and difficulties in model sharing [16]. Even with domain-adapted models, improvements remain moderate as there are reports of only around 6% improvement for medical decision support tasks [8]. A fundamental limitation of many IE approaches is their focus on sentence-level extraction, neglecting relations across sentences and documents. However, clinical narratives often distribute relevant information across multiple sentences or reports, requiring contextual aggregation. Prior work has addressed this using methods such as reference resolution to link related mentions across text [17]. This improves downstream extraction of tumor characteristics but remains challenging and achieves only moderate performance. Similarly, transformer-based approaches using question-answering formulations have been applied for IE [18]. They demonstrate strong performance for explicit information (e.g., *age*, *gender*, *histology*), but limitations for more complex attributes such as *tumor location*.

Recent work with LLMs investigates their ability to capture contextual and temporal relations. While they perform well on benchmark datasets [19], their effectiveness in real-world clinical settings remains limited, particularly for long and fragmented patient records. Therefore, LLMs are being combined with Retrieval-Augmented Generation (RAG) and structured representations [20]. For instance, CliCARE transforms longitudinal electronic health records into temporal knowledge graphs to model temporal dependencies and align them with clinical guidelines, significantly outperforming standard RAG approaches. However, RAG approaches remain limited in reliability with reports that 17% of generated references are hallucinated, indicating that such systems are not yet sufficiently robust for clinical deployment [21].

All in all, prior work shows that clinical IE requires not only accurate NER but also modeling of contextual and temporal relations. Despite recent advances, these challenges remain partially solved, particularly under realistic constraints such as limited data and domain-specific settings. Our findings are consistent with this: in a German clinical IE setting under realistic constraints, the LLM approach does not outperform transformer or rule-based methods, highlighting the need for more robust approaches to clinical IE.

6 Limitations

This study has several limitations:

1. We had only one medical specialist as an annotator. Therefore, no inter-annotator agreement could be assessed. Annotation noise, missing labels, and boundary inconsistencies may affect both model training and evaluation and limit the reliability of automatic metrics.
2. The dataset is relatively small and imbalanced, with some parameters occurring rarely, restricting the ability of data-driven models to generalize and making performance comparisons for rare parameters less reliable. Furthermore, the document-level splitting strategy resulted in 59 of 134 patients occurring in more than one subset, which may lead to optimistic performance estimates and limits conclusions regarding generalization to unseen patients.
3. The study is conducted on German clinical text, where publicly available resources are limited due to data protection constraints. Although this increases practical relevance, it reduces comparability with prior

work, which is predominantly based on English datasets.
4. RegEx patterns and LLM prompts were manually developed and are therefore sensitive to design choices, limiting reproducibility and generalizability to other datasets and clinical settings.
5. Due to resource constraints, detailed parameter-level evaluation was conducted only for a subset of models and parameters.
6. The requirement for local deployability in a realistic setting constrained the evaluation to a single exemplary LLM-based approach. Therefore, the findings cannot be generalized to LLMs in general, particularly to larger models that could not be considered under these requirements.
7. Finally, this work evaluates the IE component of ADBoard retrospectively and does not include prospective validation within actual tumor board meetings. Such workflow-level validation is outside the scope of this present NER-focused work and is being investigated separately.

7 Conclusion

In this work, we investigate IE for clinical decision support as part of the **ADBoard** pipeline. Using a real-world dataset, we compared rule-based, transformer-based, and an exploratory LLM approach under real-world clinical conditions. Our results show that transformer-based models achieve the strongest overall performance, while RegEx approaches can provide a resource-efficient complement for well-structured parameters. In contrast, the LLM-based approach does not outperform existing methods in this setting. We further demonstrate that evaluation methodology strongly influences performance interpretation, as strict span-based metrics can underestimate clinically relevant results. In practice, a hybrid approach combining transformer-based and rule-based methods could be most effective, leveraging the strengths of both paradigms depending on the parameter and data availability. Furthermore, our findings suggest that annotation quality and evaluation design are as important as advances in model architecture for clinical IE. Future work should focus on improving annotation, a better distribution of the parameters, exploring more advanced LLM-based approaches, and developing evaluation schemes that better reflect clinically meaningful extraction performance.

Notes

Author contributions

RR, SM, PT, FK: Conception and design of the study; NSR, RO: Collecting, preparing and annotating data; IY, YC, AAL: Implementation and conduction of experiments; YC, IY, AAL, RO, FK: Evaluation; IY: Writing of manuscript; RR, AAL, YC, PT, SM, RO, FK, NSR: Review of manuscript. All authors have approved the manuscript as submitted and assume responsibility for the scientific integrity of the work. The authors declare that there is no conflict of interest.

Acknowledgements

This work was carried out as part of the **ADBoard** project (01VSF21047) funded by the Joint Federal Committee and supported by the Federal Ministry of Research, Technology and Space (BMFTR) through the project **Veranda** (16KIS2048).

Competing interests

The authors declare that they have no competing interests.

References

1. Ng SST, Oehring R, Ramasetti N, Roller R, Thomas P, Chen Y, Moosburner S, Winter A, Maurer MM, Auer TA, Kamali C, Pratschke J, Benzing C, Krenzien F. Concordance of a decision algorithm and multidisciplinary team meetings for patients with liver cancer-a study protocol for a randomized controlled trial. Trials. 2023 Sep 9;24(1):577. DOI: 10.1186/s13063-023-07610-8
2. Yüzüncüoglu Y, Chen Y, Lüser AA, Roller R, Thomas P, Möller S, Ramasetti NS, Ng SST, Oehring R, Krenzien F. Entwicklung einer automatisierten Informationsextraktion und Entscheidungsunterstützung für die Tumorkonferenz hepatozellulärer Karzinome. KI in der Arzt-Patienten-Kommunikation. Forthcoming.
3. Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, with language models. 3rd ed. draft. 2026 Jan 6. Available from: https://web.stanford.edu/~jurafsky/slp3/
4. Alshammari N, Alanazi S. The impact of using different annotation schemes on named entity recognition. Egyptian Informatics Journal. 2021;22(3):295-302. DOI: 10.1016/j.eij.2020.10.004
5. Leitlinienprogramm Onkologie. S3-Leitlinie Diagnostik und Therapie des hepatozellulären Karzinoms und biliärer Karzinome. AWMF-Registernummer 032-0530L. Version 5.2. Langfassung. 2025 Jun [accessed 2026 Mar 31]. Available from: https://www.leitlinienprogramm-onkologie.de/leitlinien/hcc-und-biliaere-karzinome
6. Klie JC, Bugert M, Boullosa B, de Castilho RE, Gurevych I. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations; 2018 Aug 20-26; Santa Fe, New Mexico. Association for Computational Linguistics; 2018. p. 5-9.
7. Gardazi NM, Daud A, Malik MK, Bukhari A, Alsahfi T, Alshemaimri B. BERT applications in natural language processing: a review. Artif Intell Rev. 2025;58(6):166. DOI: 10.1007/s10462-025-11162-5
8. Saxena A, Santhanavijayan A. A layer-wise survey on internal modifications in BERT and its variants: techniques, applications, and performance trade-offs. International Journal of Data Science and Analytics. 2026;22(1):1.

9. Joloudari JH, Hussain S, Nematollahi MA, Bagheri R, Fazl F, Alizadehsani R, Lashgari R, Talukder A. BERT-deep CNN: State of the art for sentiment analysis of COVID-19 tweets. Social Network Analysis and Mining. 2023;13(1):99. DOI: 10.1007/s13278-023-01102-y

10. Chan B, Schweter S, Möller T. German's next language model. In: Scott D, Bel N, Zong C, editors. Proceedings of the 28th International Conference on Computational Linguistics; 2020 Dec 8-13; Barcelona, Spain (online). International Committee on Computational Linguistics; 2020. p. 6788-6796.

11. Bressem KK, Papaioannou JM, Grundmann P, Borchert F, Adams LC, Liu L, Busch F, Xu L, Loyen JP, Niehues SM, Augustin M, Grosser L, Makowski MR, Aerts HJWL, Löser A. MedBERT.de: A comprehensive German BERT model for the medical domain [preprint]. arXiv. 2023. DOI: 10.48550/arXiv.2303.08179

12. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. Unsupervised cross-lingual representation learning at scale. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5-10; online. Association for Computational Linguistics; 2020. p. 8440-8451.

13. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al Dahle A, Letman A, Mathur A, Schelten A, Vaughan A, Yang A, Fan A, Goyal A, Hartshorn A, Yang A, Mitra A, Sravankumar A, Korenev A, Hinsvark A, Rao A, Zhang A, Rodriguez A, Gregerson A, Spataru A, Roziere B, Biron b, Tang B, Chern B, Caucheteux C, Nayak C, Bi C, Marra C, McConnell C, Keller C, Touret C, Wu C, Wong C, Ferrer CC, Nikolaidis C, Allonsius D, Song D, Pintz D, Livshits D, Wyatt D, Esiobu D, Choudhary D, Mahajan D, Garcia-Olano D, Perino D, Hupkes D, Lakomkin E, AlBadawy E, Lobanova E, Dinan E, Smith EM, Radenovic F, Guzmán F, Zhang F, Synnaeve G, Lee G, Anderson GL, Thattai G, Nail G, Mialon G, Pang G, Cucurell G, Nguyen H, Korevaar H, Xu H, Touvron H, Zarov I, Ibarra IA, Kloumann I, Misra I, Evtimov I, Zhang J, Copet J, Lee J, GeffertJ, Vranes J, Park J, Mahadeokar J, Shah J, van der Linde J, Billock J, Hong J, Lee J, Fu J, Chi J, Huang J, Liu J, Wang J, Yu J, Bitton J, Spisak J, Park J, Rocca J, Johnstun J, Saxe J, Jia J, et al. The llama 3 herd of models. arXiv. 2024. DOI: 10.48550/arXiv.2407.21783

14. Bossy R, Golik W, Ratkovic Z, Bessieres P, Nédellec C. BioNLP shared task 2013 – an overview of the bacteria biotope task. In: Proceedings of the BioNLP shared task 2013 workshop; 2013 Aug 9; Sofia, Bulgaria. Association for Computational Linguistics; 2013. p. 161-169.

15. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: A literature review. J Biomed Inform. 2018 Jan;77:34-49. DOI: 10.1016/j.jbi.2017.11.011

16. Wu H, Wang M, Wu J, Francis F, Chang YH, Shavick A, Dong H, Poon MTC, Fitzpatrick N, Levine AP, Slater KT, Handy A, Karwath A, Gkoutos GV, Chelala C, Shah AD, Stewart R, Collier N, Alex B, Whiteley W, Sudlow C, Roberts A, Dobson RJB. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. NPJ Digit Med. 2022 Dec 21;5(1):186. DOI: 10.1038/s41746-022-00730-6

17. Yim WW, Kwan SW, Yetisgen M. Tumor reference resolution and characteristic extraction in radiology reports for liver cancer stage prediction. J Biomed Inform. 2016 Dec;64:179-191. DOI: 10.1016/j.jbi.2016.10.005

18. Zhu S, Gilbert M, Ghanem AI, Siddiqui F, Thind K. Feasibility of using zero-shot learning in transformer-based natural language processing algorithm for key information extraction from head and neck tumor board notes [ASTRO 2023 abstract]. Int J Radiat Oncol Biol Phys. 2023;117(2):e500.

19. Hu Y, Zuo X, Zhou Y, Peng X, Huang J, Keloth VK, Zhang VJ, Weng RL, Shyr C, Chen Q, Jiang X, Roberts KE, Xu H. Information extraction from clinical notes: are we ready to switch to large language models? J Am Med Inform Assoc. 2026 Mar 1;33(3):553-562. DOI: 10.1093/jamia/ocaf213

20. Li D, Liang J, Li W, Wang X, Cao L, Yu K. CliCARE: Grounding large language models in clinical guidelines for decision support over longitudinal cancer electronic health records. In: Proceedings of the 40th AAAI Conference on Artificial Intelligence; 2026 Jan 20-27; Singapore. Washington, DC, USA: AAAI Press; 2026. p. 31554-31562.

21. Berman E, Sundberg Malek H, Bitzer M, Malek N, Eickhoff C. Retrieval Augmented Therapy Suggestion for Molecular Tumor Boards: Algorithmic Development and Validation Study. J Med Internet Res. 2025 Mar 5;27:e64364. DOI: 10.2196/64364

Erratum

The spelling of the author Krenzien has been corrected.

Corresponding author:
Imge Yüzüncüoglu
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Salzufer 15/16, 10587 Berlin, Germany
imge.yuezuencueoglu@dfki.de

Please cite as
Yüzüncüoglu I, Chen Y, Lüser AA, Ramasetti NS, Oehring R, Krenzien F, Thomas P, Möller S, Roller R. Automatic extraction of clinically relevant parameters for tumor board decision support. GMS Med Inform Biom Epidemiol. 2026;22:Doc13.
DOI: 10.3205/mibe000311, URN: urn:nbn:de:0183-mibe0003116

This article is freely available from
https://doi.org/10.3205/mibe000311

Published: *2026-09-22*
Published with erratum: *2026-09-24*