Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240507

Improving the Quality of Unstructured Cancer Data Using Large Language Models: A German Oncological Case Study

Yongli MOU[a,1] , Jonathan LEHMKUHL[a,b], Nicolas SAUERBRUNN[c], Anja KÖCHEL[c], Jens PANSE[c,d], Daniel TRUH[e], Sulayman SOWE[a,b], Tim BRÜMMENDORF[d], and Stefan DECKER[a,b]
[a] *Chair of Computer Science 5, RWTH Aachen University, Germany*
[b] *Fraunhofer FIT, Germany*
[c] *Center for Integrated Oncology, University Hospital Aachen, Germany*
[d] *Department of Hematology, Oncology, Hemostaseology and Stem Cell Transplantation, University Hospital Aachen, Germany*
[e] *Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Germany*

Abstract. With cancer being a leading cause of death globally, epidemiological and clinical cancer registration is paramount for enhancing oncological care and facilitating scientific research. However, the heterogeneous landscape of medical data presents significant challenges to the current manual process of tumor documentation. This paper explores the potential of Large Language Models (LLMs) for transforming unstructured medical reports into the structured format mandated by the German Basic Oncology Dataset. Our findings indicate that integrating LLMs into existing hospital data management systems or cancer registries can significantly enhance the quality and completeness of cancer data collection - a vital component for diagnosing and treating cancer and improving the effectiveness and benefits of therapies. This work contributes to the broader discussion on the potential of artificial intelligence or LLMs to revolutionize medical data processing and reporting in general and cancer care in particular.

Keywords. Large Language Models, Information Extraction, Prompt Engineering, Cancer Registry, Data Quality

1. Introduction

According to the World Health Organization[2], cancer is a leading cause of death worldwide with nearly 10 million deaths in 2020, while the Robert Koch Institute (RKI)[3] reports more than half a million people are diagnosed with cancer in Germany every year resulting in approximately 230,000 deaths annually with the number steadily rising in recent decades [1]. Cancer registries are crucial for monitoring and improving the quality of oncological care, increasing transparency, and contributing to

[1] Corresponding Author: Yongli Mou; E-mail: mou@dbis.rwth-aachen.de.

[2] WHO, Cancer Fact Sheets: https://www.who.int/news-room/fact-sheets/detail/cancer

[3] RKI, Cancer in Germany for 2019/2020

new findings in scientific research. In Germany, each federal state has its independent cancer registration structure that collects epidemiological and clinical data for people living in the state. All physicians and health care providers involved in the diagnosis or treatment of cancer are required to notify cancer cases. Large hospitals have an internal institution that does the documentation for all cancer treatment done in the hospital and submits the reports to the corresponding federal cancer registry on the state level. Nationally, the German Centre for Cancer Registry Data (ZfKD) at the RKI systematically aggregates the data from the federal registries to create a comprehensive national cancer database [2].

In the modern hospital systems, diverse and integrated oncology departments generate vast amounts of structured and unstructured patient data, including physician notes, tumor board protocols and pathology reports, to mention a few, which results in a heterogeneous landscape. On the other hand, in Germany, the data reported to the federal cancer registry must adhere to a structured data schema known as the German Basic Oncology Dataset enforced by the Federal Cancer Registry Data Act 2021 [2], which poses a significant challenge in harmonizing and transforming those data into the appropriate format for reporting to the federal cancer registry. Currently, this data transformation process heavily relies on manual tumor documentation, wherein unstructured reports are converted into structured formats. However, manually extracting information from the documents is not only time-consuming and labor-intensive but also prone to error. With budget constraints preventing the recruitment of additional cancer registrars, the registry faces challenges in meeting regulatory reporting requirements.

The recent emergence of Large Language Models (LLMs) has opened unprecedented opportunities to effectively extract information from medical reports and convert unstructured cancer data into the structured format [3]. In this work, we aim to explore the feasibility of enhancing tumor documentation in the local cancer registry by integrating LLMs into the current process. As a case study, we first look at the operations of the cancer registry of the Center for Integrated Oncology (CIO) Aachen. Then we build an LLMs-based Proof-of-Concept tool, apply it to a subset of medical reports of pseudonymized patients and evaluate the performance. The results demonstrate the potential of seamlessly integrating LLMs into the existing workflow at the CIOs to efficiently convert diverse and unstructured data into standardized formats. This could facilitate accurate and timely reporting to the federal registry, significantly improve cost-effectiveness, and simultaneously enhance the quality and completeness of the reported cancer data.

2. Methods

2.1. Documentation Workflow in the CIO Aachen

In the CIO Aachen, tumor documentation happens separately for every tumor instance of a patient, i.e., a case. In the case of CIO Aachen, it uses the proprietary software Onkostar for tumor documentation, which defines structured forms to be filled out by the registrars. All relevant information is stored in the Hospital Information System (HIS). The main task of the cancer registrars is data transformation between these two systems.

Free-text reports contain a vast amount of information to be reported in Onkostar. Various types of reports such as pathology reports, referral letters and tumor board protocols exist in the HIS, while in Onkostar there are various types of forms, for example for diagnosis, pathology findings, surgery and tumor boards. By human documentation, the information from these reports is extracted and transformed into completed structured forms in Onkostar, as shown on the left-hand side in Figure 1. Unfortunately, there is not a one-to-one correspondence between the reports in the HIS and the forms in Onkostar. To improve cost-efficiency and data quality within CIO Aachen, we propose a semi-automated approach where the data transformation is done by an LLM-based tool with subsequent human review, shown on the right-hand side in Figure 1.

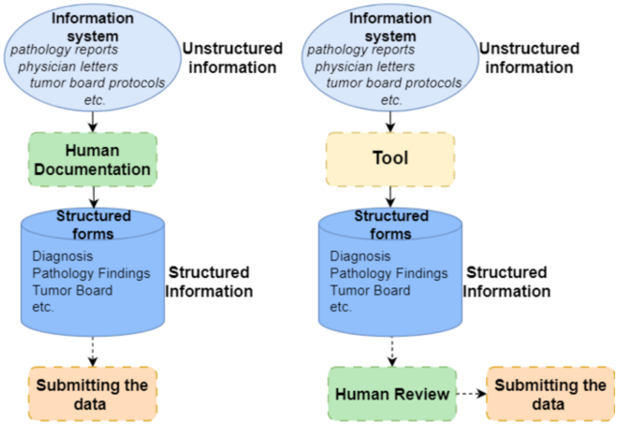

Figure 1. Comparison of two approaches for data transformation in the CIO Aachen. Current data flow (left) and semi-automated data flow with LLMs-based tool (right).

2.2. Data Transformation of Pathology Reports using Large Language Models

Since every type of report has to be processed differently, we focus on pathology reports in this work. For one pathology examination, we feed all associated pathology reports to the LLM, together with a structured data model and a surrounding prompt that instructs the LLM to determine the correct value for each feature of the data model based on the provided pathology reports. The structured data model models the important features of pathology reports and, for each feature, contains a description that explains to the LLM how to determine the correct value. To improve the performance, we experimented a lot with different prompting strategies such as zero-shot, few-shot and chain-of-thought prompting [4]. After the features were extracted by the LLM, they are inserted into the forms "Pathology Findings" and "Diagnosis" in Onkostar. As a Proof-of-Concept we built an LLM-based tool that accomplishes the information extraction and import steps and is openly accessible[4]. The resulting JSON outputs of the LLM are joined together and then parsed against the data model. As LLMs to power our tool we tried both open-source LLMs such as Mixtral-8×7B [5] and closed-source LLMs such as GPT-4 by OpenAI [6].

[4] Code available: https://github.com/MouYongli/MIE2024

2.3. Dataset

To evaluate the performance of the developed tool, we created a dataset of biopsy reports from breast cancer patients at the University Hospital Aachen. We randomly selected 50 patients and divided them into groups of 25 training samples and 25 testing samples. We pseudonymized all reports, which involved replacing all values that could be considered personally identifiable information with dummy values, such as record number, name, date of birth, admission number, inbound date, and outbound date.

3. Results

Two essential data quality dimensions of cancer registry data were examined: completeness and correctness [7,8]. Table 1. shows the results of human registrars against GPT-4 and Mixtral-8×7B, based on average scores across 27 examinations. The findings reveal that GPT-4 already achieves near-human-level accuracy in the correctness, while Mixtral-8×7B exhibits notably lower performance, particularly in identifying Localization and ICD-10 Diagnosis features. When evaluating completeness, both LLMs yield significantly higher completeness scores than human registrars in our example.

Table 1. Comparison of correctness and completeness scores between the human registrars (H), GPT-4 (G), and Mixtral-8×7B (M)

Feature	**Correctness**			**Completeness**		
	H	**G**	**M**	**H**	**G**	**M**
Examination Date	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Submission Number	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Examined Preparation	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Type of Biopsy	**0.96**	0.92	0.93	**1.0**	0.96	**1.0**
Biopsy Sampling Site	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Localization	**0.96**	0.78	0.41	**1.0**	**1.0**	**1.0**
Tumor Proof	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
ICD-O3 History	**1.0**	0.96	**1.0**	**1.0**	**1.0**	**1.0**
Grading	0.91	**1.0**	0.90	0.95	**1.0**	0.95
Estrogen Positive Cells	**1.0**	**1.0**	0.96	0.81	**0.93**	0.85
Estrogen Intensity	**1.0**	0.96	0.89	0.04	**1.0**	1.0
Estrogen Score	**1.0**	0.96	**1.0**	0.22	**0.93**	0.78
Progesterone Positive Cells	**1.0**	0.91	0.96	0.81	0.81	**0.85**
Progesterone Intensity	**1.0**	0.96	0.93	0.04	**1.0**	**1.0**
Progesterone Score	**1.0**	0.91	**1.0**	0.22	**0.81**	0.74
HER2	**1.0**	**1.0**	0.90	**1.0**	**1.0**	0.95
Ki-67	**1.0**	0.96	0.79	**1.0**	**1.0**	**1.0**
ICD-10 Diagnosis	**0.93**	0.81	0.48	**1.0**	**1.0**	**1.0**
Side	**1.0**	0.96	0.96	**1.0**	**1.0**	**1.0**
Mean	**0.99**	**0.95**	**0.90**	**0.79**	**0.97**	**0.95**

4. Discussion

For application within the cancer registry in Aachen, our tool could already enhance the efficiency of tumor documentation when being used in a semi-automated manner with subsequent human review. Since correctness scores of both models are below the human benchmark, humans sometimes would have to correct a false value. However, our assumption is that this would still increase time- and cost-efficiency. Moreover, our results indicate that using our tool would significantly improve completeness of documentation. Currently, our Proof-of-Concept tool can only handle biopsy reports of breast cancer patients. To effectively use such a system for comprehensive tumor documentation, other types of reports as well as other cancer types need to be considered. Another challenge is that closed-source LLMs performed better than open-source LLMs in our evaluation, but they raise privacy concerns in the clinical context.

5. Conclusions

We presented a case study that shows the promise of utilizing LLMs to improve both the quality and cost-effectiveness of cancer data documentation. Closed-source LLMs offer superior performance, but they raise significant privacy concerns. Conversely, open-source LLMs present a compelling advantage by allowing for local deployment and fine-tuning, which ensures compliance with the existing privacy regulations at the healthcare facilities, while applying them to real patient data. In future work, we will extend our tool to more types of reports, use fine-tuning of models to improve the performance and integrate it end-to-end into the workflow in the cancer registry.

Acknowledgements

This work was supported by the German Ministry for Research and Education (BMBF) projects, the FAIR Data Spaces project under the grant number FAIRDS05 and the WestAI project under the grant number 01IS22094D.

References

[1] Ronckers C, Spix C, Trübenbach C, Katalinic A, Christ M, Cicero A, Folkerts J, Hansmann J, Kranzhöfer K, Kunz B, Manegold K. Krebs in Deutschland für 2019/2020.

[2] Katalinic A, Halber M, Meyer M, Pflüger M, Eberle A, Nennecke A, Kim-Wanner SZ, Hartz T, Weitmann K, Stang A, Justenhoven C. Population-based clinical cancer registration in Germany. Cancers. 2023 Aug 2;15(15):3934.

[3] Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, Bressem KK. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology. 2023 Apr 4;307(4):e230725.

[4] Yang J, Jin H, Tang R, Han X, Feng Q, Jiang H, Zhong S, Yin B, Hu X. Harnessing the power of llms in practice: A survey on chatgpt and beyond. ACM Transactions on Knowledge Discovery from Data. 2024 Apr 27;18(6):1-32.

[5] Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, Chaplot DS, Casas DD, Hanna EB, Bressand F, Lengyel G. Mixtral of experts. arXiv preprint arXiv:2401.04088. 2024 Jan 8.

[6] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R. Gpt-4 technical report. arXiv preprint arXiv:2303.08774. 2023 Mar 15.

[7] Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. European journal of cancer. 2009 Mar 1;45(5):747-55.

[8] Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods Part II. Completeness. European journal of cancer. 2009 Mar 1;45(5):756-64.