Harmonized steps

Summary

This thesis addresses the challenges of data management in the healthcare sector, particularly the unstructured and dispersed nature of medical data, which hampers its efficient retrieval, analysis, and utilization. The research focusses on developing computer-assisted Natural Language Processing (NLP) systems to structure medical data in real-time, enhancing the reporting process, and data standardization with a focus on Tumour Node Metastases (TNM) staging and the International Classification of Diseases (ICD). The NLP systems prioritizes model explainability and seamless integration into medical workflows, adhering to the FAIR principles for data management.

Significant findings include the successful application of a rule-based expert system for classifying lung tumours in Dutch radiology reports, highlighting the feasibility of a specialized NLP in a less commonly supported language. Challenges arose from the scarcity of labelled data, necessitating a rule-based rather than machine learning approach. Institutional variations in medical documentation and coding practices demonstrated substantial differences, underscoring the importance of tailored methodologies.

The research identified several limitations, such as the restricted involvement of domain experts, which could introduce bias, the risk of selection bias in report selection, and the restricted number of datasets examined, which affects the generalizability and robustness of the approaches. This thesis also highlighted the constraints of radiology-based tumour staging, the limitations of description-based ICD coding, and the lack of real-world application and valorisation of the proposed assistants.

Recommendations for future research include a shift toward use-case-driven healthcare solutions, extending methodologies to other classification systems, utilizing multidisciplinary data for TNM staging, exploring large language models for classification, creating open-source datasets, leveraging active learning, and enhancing the valorisation of artificial intelligence (AI) co-pilots in clinical settings. Importantly, the research advocates for the indispensable role of medical professionals' expertise in the development and refinement of AI-assisted systems.

In conclusion, despite the remarkable potential of NLP and AI to improve healthcare data management and support professionals, human oversight and domain-specific expertise are critical for their success. The envisioned AI systems function as 'co-pilots', requiring a symbiotic relationship with healthcare practitioners to ensure optimal performance, safety, and compliance with emerging regulations, such as the EU AI Act. The thesis highlights the urgent need for innovations in NLP to help alleviate the burden on healthcare systems, allow medical professionals to focus on direct patient care, and harness the full capabilities of structured digital medical data.

Nederlandse samenvatting

Dit proefschrift richt zich op de uitdagingen van de gegevensverwerking in de gezondheidszorg, met name de vaak ongestructureerde aard van medische gegevens, die het efficiënt ophalen, analyseren en het gebruik ervan belemmert. Dit proefschrift onderzoek focust op het ontwikkelen van een computerondersteund Natural Language Processing (NLP) systemen voor het real-time structureren van medische gegevens. De systemen hebben als doel het proces van medische rapportage te optimaliseren en simultaan gegevens te standaardiseren, in het bijzonder voor TNM-stadiëring (Tumor Node Metastases) en de Internationale Classificatie van Ziekten (ICD). Het systeem prioriteert de interpreteerbaarheid van modellen en een vlekkeloze integratie in medische workflows, terwijl de FAIR-principes (Findable, Accessible, Interoperable, Reusable) voor gegevens worden nageleefd.

Een op regels gebaseerd expertsysteem is succesvol toegepast voor het classificeren van longtumoren in Nederlandse radiologieverslagen. Dit onderstreept de haalbaarheid van gespecialiseerde NLP in een historisch minder ondersteunde taal. Uitdagingen kwamen voort uit de schaarste van gelabelde gegevens, waardoor gekozen is voor een op regels gebaseerde benadering boven de tegenwoordig meer gebruikelijke machine learning methodes. Institutionele variaties in medische documentatie en coderingspraktijken lieten substantiële verschillen zien, wat het belang van op maat gemaakte methodologieën onderstreept.

Het onderzoek had verschillende beperkingen, zoals gelimiteerde aantal domeinexperts betrokken bij het onderzoek, waardoor vertekening zou kunnen optreden, het risico van selectiebias bij de selectie van rapporten, en het beperkte aantal onderzochte datasets, kan de generaliseerbaarheid en robuustheid van de benaderingen hebben beïnvloed. Het onderzoek wees ook op de beperkingen van op radiologie gebaseerde tumorstadiëring, de tekortkomingen van op beschrijving gebaseerde ICD-codering, en het gebrek aan valorisatie van de voorgestelde hulpmiddelen.

Aanbevelingen voor toekomstig onderzoek omvatten een verschuiving naar op gebruikssituaties gebaseerde (use-case-driven) gezondheidszorgtoepassingen, het uitbreiden van methoden naar andere classificatiesystemen, het inzetten van multidisciplinaire gegevens voor TNM-stadiëring, het verkennen van grote taalmodellen voor classificatie, het creëren van open-source datasets, het toepassen van actief leren (active-learning) en het verbeteren van de valorisatie van artificiële intelligentie (AI) 'co-piloten' in klinische omgevingen. Het onderzoek benadrukt de essentiële rol van medische professionals in de ontwikkeling en verfijning van AI-ondersteunde systemen.

Conclusie: Ondanks het aanzienlijke potentieel van NLP en AI om de gegevensverwerking in de gezondheidszorg te verbeteren en professionals te ondersteunen, zijn menselijk toezicht en domeinspecifieke expertise essentieel voor het succes van deze technolo-

gieën. De beoogde AI-systemen functioneren als een “co-piloot” en vereisen een symbiotische relatie met zorgverleners om optimale prestaties, veiligheid, en naleving van opkomende regelgeving, zoals de EU AI Act, te waarborgen. Het proefschrift onderstreept de urgente noodzaak van innovaties in NLP om de druk op gezondheidszorgsystemen te verminderen, medische professionals in staat te stellen zich te focussen op directe patiëntenzorg, en de volledige potentie van gestructureerde digitale medische gegevens te benutten.