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

doi:10.3233/SHTI240744

Towards a Reporting Guideline for Studies on Information Extraction from Clinical Texts

Daniel REICHENPFADER[a,1] and Kerstin DENECKE[a]
[a] *Bern University of Applied Sciences, Institute for Patient-centered Digital Health, Biel/Bienne, Switzerland*
ORCiD ID: Daniel Reichenpfader https://orcid.org/0000-0002-8052-3359,
Kerstin Denecke https://orcid.org/0000-0001-6691-396X

Abstract. Background: The rapid technical progress in the domain of clinical Natural Language Processing and information extraction (IE) has resulted in challenges concerning the comparability and replicability of studies. Aim: This paper proposes a reporting guideline to standardize the description of methodologies and outcomes for studies involving IE from clinical texts. Methods: The guideline is developed based on the experiences gained from data extraction for a previously conducted scoping review on IE from free-text radiology reports including 34 studies. Results: The guideline comprises the five top-level categories information model, architecture, data, annotation, and outcomes. In total, we define 28 aspects to be reported on in IE studies related to these categories. Conclusions: The proposed guideline is expected to set a standard for reporting in studies describing IE from clinical text and promote uniformity across the research field. Expected future technological advancements may make regular updates of the guideline necessary. In future research, we plan to develop a taxonomy that clearly defines corresponding value sets as well as integrating both this guideline and the taxonomy by following a consensus-based methodology.

Keywords. Reporting guideline, Standardization, Information Extraction, Natural Language Processing

1. Introduction

The current era is marked by the advent of language-centric machine learning that is influencing various domains, including healthcare. Although still lacking behind industry, the medical sector starts leveraging the capabilities of Natural Language Processing (NLP), for example to extract information from unstructured clinical texts, a task commonly referred to as information extraction (IE). IE methods extract instances of specific, pre-defined generic information types from unstructured text [1]. However, due to the fast technological progress, integrating latest approaches, like utilizing large language models (LLMs), the scope of NLP keeps expanding rapidly. While this progress introduces potential for innovation, it also causes challenges in ensuring

[1] Corresponding Author: Daniel Reichenpfader, Institute for Patient-centered Digital Health, Bern University of Applied Sciences, Biel/Bienne, Switzerland; E-Mail: daniel.reichenpfader@bfh.ch.

comparability and standardization across studies. A comparison among IE approaches allows for a more in-depth analysis of model performance with respect to influencing factors like model size, computational resources, and data set size. This could lead to aggregated knowledge on quality-related factors in IE systems, eliminating the need to repeat experiments among research groups. Unfortunately, existing practices often hinder comparability due to two key reasons, being rarity of datasets and methodological inconsistencies. The former arises because of high standards to be cohered to regarding data protection and security, due to the sensitive nature of healthcare data according to regulations such as the General Data Protection Regulations (GDPR). Furthermore, most of the few available datasets contain English data, resulting in an ‘anglophone bias’. Regarding the latter, methodological inconsistencies in different studies emerge due to the broadness of the NLP domain and fast, recent technological advancements. These inconsistencies regard e.g. the calculation of performance measures, splitting data, design of annotation processes, etc. The introduction of model cards, which became usual in the context of transfer learning, is a first step towards increased comparability [2]. However, these model cards are practically oriented and might not provide all necessary details, leaving room for ambiguity.

In the broad context of healthcare and artificial intelligence, several reporting guidelines are available: For example, Liu et al. introduced CONSORT-AI, a “guideline for clinical trials evaluating interventions with an AI component” [3]. MI-CLAIM is a similar, generic guideline for reporting AI algorithms in medicine [4]. Other guidelines focus on specific use cases, e.g., prediction and prognosis [5,6], or a certain clinical domain, e.g., urology [7] or radiology [8].

Currently, there is no reporting guideline available specifically targeting clinical IE studies. This study aims to deliver such a guideline that can support standardized reporting of methodology and outcomes in studies describing IE from clinical text. In the following sections, we describe the development process and the guideline itself, providing details on its categories and aspects.

2. Methodology

We base the content of the reporting guideline on a previously carried out scoping review on LLM-based IE from free-text radiology reports [9]. According to the JBI Manual for Evidence Synthesis, a data extraction table was created based on the defined research questions. This table contains all aspects to be extracted from each included source of evidence. The data extraction table was updated after a pilot test of analyzing two sources of evidence, adding additional aspects. The finalized data extraction table was converted to a list of aspects to be included in the reporting guideline. Last, these aspects were grouped and generalized where applicable to not only include LLM-based approaches and radiology reports, but to generalize to any NLP method and unstructured clinical text source.

3. Results

The developed guideline consists of five categories of aspects to be reported, as depicted in Fig. 1.

3.1. Information Model

The information model serves as a framework for the pre-defined generic information types according to the definition of IE. The number of information types to be extracted together with a short description should be described. If applicable, the underlying theoretical model should be mentioned (e.g., clinically validated scores). Furthermore, it should be described whether the extracted information types are structured and/or normalized after extraction.

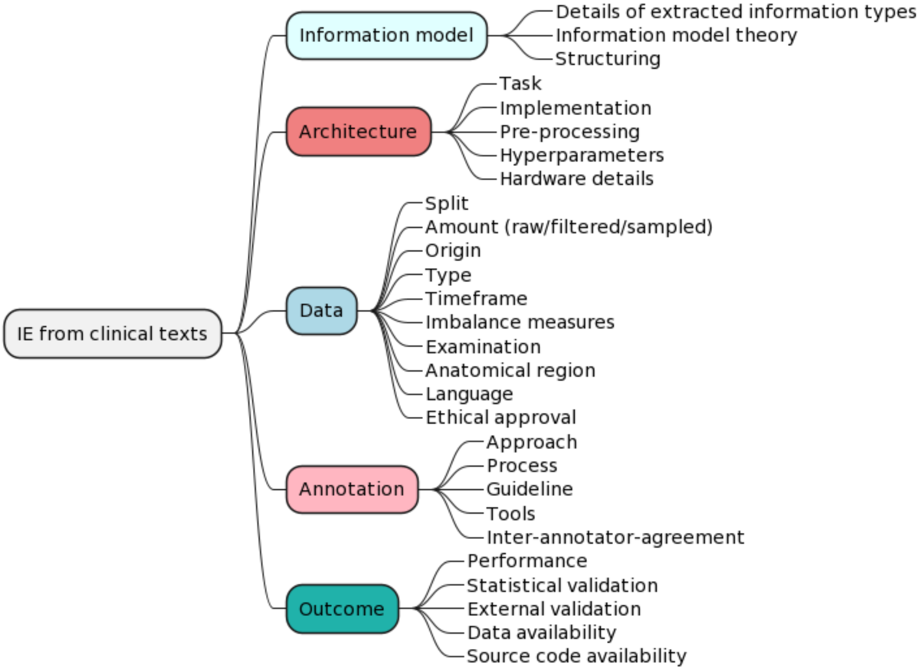

Figure 1. Reporting guideline for information extraction (IE) from clinical text.

3.2. Architecture

IE can be distinguished between document-level and entity-level extraction: The task of document-level extraction can be regarded as a multi-class classification task of the whole text. Entity-level extraction comprises named entity recognition as well as relation extraction. Classification tasks in general can be separated into binary classification, multi-class, and multi-label classification. With the advent of LLMs, the above-mentioned traditional tasks were augmented by approaches based on extractive question answering as well as generative approaches based on autoregressive sampling [10]. The NLP implementation might be heuristic-, machine-learning-, deep-learning- or LLM-based, a hybrid or comparative approach. If applicable, base model, pre-training, further pre-training as well as fine-tuning should be reported in detail, as well as any pre-processing steps (e.g., tokenization, stemming, lemmatization, tagging, de-identification). Depending on the chosen implementation, hyperparameters must be documented to facilitate replicability of the described approach. Furthermore, the hardware specification of the training and inference environment as well as GPU hours should be included in the report.

3.3. Data

When it comes to training data, it is important to describe related aspects in detail, as data is often not available, making replication of results impossible. A data flow diagram demonstrates the applied data splitting method (e.g., train, test, and validation splits), as well as number of documents, sentences, and tokens of each split. Furthermore, the diagram shows the process of filtering and sampling data from the originally available pool of documents. A separate diagram should be included in case the model was validated on external data. Any imbalance measures (e.g. stratification) should also be reported. Other aspects include the name and country of the originating institution, the clinical text source (e.g. reports, clinical notes, discharge letters), timeframe of dataset, and, in case of reports, which examination and anatomical region the document corresponds to. Moreover, document language should be explicitly stated and whether ethical approval was granted or waived.

3.4. Annotation

The source of evidence should clarify if and how data was annotated. It should be mentioned whether a manual, automated or hybrid approach was applied to label data. The annotation process should be thoroughly described, including number of documents and information types, annotators, as well as test rounds, details regarding annotation guideline development and tools used. The final annotation guideline should be made available. Furthermore, it should be explained how inter-annotator-agreement was calculated, including the corresponding scores. The background of annotators should be briefly explained including their experience, role and level of domain expertise.

3.5. Outcome

Model performance should be reported separately for each information type as well as averaged over all extracted information types. We do not recommend specific performance measures but highlight the importance of including the formulas of how the measures are calculated. Any statistical tests or cross-validation should be described including confidence intervals. All performance measures should be reported separately on internal validation data as well as external data, if applicable. Last, we emphasize the importance of making datasets and source code available via open research data repositories, e.g. Zenodo (https://zenodo.org) or OSF (https://osf.io).

4. Discussion and Conclusions

With this paper, we suggest a reporting guideline to be adhered to by researchers reporting on experiments and studies with IE systems in the clinical domain. In general, if an aspect of the guideline is not applicable, a short rationale should be given why corresponding information is not described (e.g., application of a pre-trained LLM, so no annotated data was used). This guidance ensures that all essential information is described in the publication and therefore fosters replicability and comparability of studies. The need of improving comparability of studies was already shown by Davidson et al., who conducted a review on the quality of NLP studies [11]. The 15 criteria the authors used to assess quality are all included in our results.

Our paper shows the following limitations: First, the guidelines are based on our experiences and have not yet been validated by an expert panel. There might be dependencies between aspects, e.g. the description of hyperparameters is not applicable to heuristic-based approaches, which remains currently unconsidered. We furthermore assume that the guideline cannot be considered final, as future technological advancements might impact IE methodologies and therefore items to be reported.

As a future research direction, we plan to develop a taxonomy including the definition of value sets. Such taxonomy allows to investigate the occurrence of values as well as interactions between aspects as done by Hupkes et al. [12]. Next, the guideline proposed in this paper might be finalized by integrating the taxonomy and by validation according to a consensus-based methodology, e.g. conducting a Delphi study [13].

References

[1] Okurowski ME. Information Extraction Overview. In: TIPSTER TEXT PROGRAM: PHASE I: Proceedings of a Workshop held at Fredricksburg, Virginia, September 19-23, 1993 [Internet]. Fredericksburg, Virginia, USA: Association for Computational Linguistics; 1993 [cited 2024 Mar 15]. p. 117–21. Available from: https://aclanthology.org/X93-1012

[2] Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency [Internet]. New York, NY, USA: Association for Computing Machinery; 2019 [cited 2024 Mar 18]. p. 220–9. (FAT* '19). Available from: https://dl.acm.org/doi/10.1145/3287560.3287596

[3] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020 Sep;26(9):1364–74.

[4] Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020 Sep;26(9):1320–4.

[5] Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016 Dec 16;18(12):e323.

[6] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015 Jan 6;162(1):W1-73.

[7] Kwong JCC, McLoughlin LC, Haider M, Goldenberg MG, Erdman L, Rickard M, et al. Standardized Reporting of Machine Learning Applications in Urology: The STREAM-URO Framework. Eur Urol Focus. 2021 Jul;7(4):672–82.

[8] Klontzas ME, Gatti AA, Tejani AS, Kahn CE. AI Reporting Guidelines: How to Select the Best One for Your Research. Radiology: Artificial Intelligence. 2023 May;5(3):e230055.

[9] Reichenpfader D, Müller H, Denecke K. Large language model-based information extraction from free-text radiology reports: a scoping review protocol. BMJ Open. 2023 Dec 1;13(12):e076865.

[10] Shanahan M, McDonell K, Reynolds L. Role play with large language models. Nature. 2023 Nov;623(7987):493–8.

[11] Davidson EM, Poon MTC, Casey A, Grivas A, Duma D, Dong H, et al. The reporting quality of natural language processing studies: systematic review of studies of radiology reports. BMC Medical Imaging. 2021 Oct 2;21(1):142.

[12] Hupkes D, Giulianelli M, Dankers V, Artetxe M, Elazar Y, Pimentel T, et al. A taxonomy and review of generalization research in NLP. Nat Mach Intell. 2023 Oct;5(10):1161–74.

[13] Spranger J, Homberg A, Sonnberger M, Niederberger M. Reporting guidelines for Delphi techniques in health sciences: A methodological review. Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen. 2022 Aug 1;172:1–11.