npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01219-0

A scoping review of large language model based approaches for information extraction from radiology reports

Daniel Reichenpfader [1,2] ✉, Henning Müller [3,4] & Kerstin Denecke [1]

Radiological imaging is a globally prevalent diagnostic method, yet the free text contained in radiology reports is not frequently used for secondary purposes. Natural Language Processing can provide structured data retrieved from these reports. This paper provides a summary of the current state of research on Large Language Model (LLM) based approaches for information extraction (IE) from radiology reports. We conduct a scoping review that follows the PRISMA-ScR guideline. Queries of five databases were conducted on August 1st 2023. Among the 34 studies that met inclusion criteria, only pre-transformer and encoder-based models are described. External validation shows a general performance decrease, although LLMs might improve generalizability of IE approaches. Reports related to CT and MRI examinations, as well as thoracic reports, prevail. Most common challenges reported are missing validation on external data and augmentation of the described methods. Different reporting granularities affect the comparability and transparency of approaches.

In contemporary medicine, diagnostic tests, particularly various forms of radiological imaging, are vital for informed decision-making[1]. Radiologists create for image examinations semi-structured free-text radiology reports by dictation, sticking to a personal or institutional schema to organize the information contained. Structured reporting that is only used in few institutions and for specific cases on the other hand offers a possibility to enhance automatic analysis of reports by defining standardized report layouts and contents.

Despite the potential benefits of structured reporting in radiology, its implementation often encounters resistance due to the possible temporary increase in radiologists' workload, rendering the integration into clinical practice challenging[2]. Natural language processing (NLP) can provide the means to make structured information available by maintaining existing documentation procedures. NLP is defined as "tract of artificial intelligence and linguistics, devoted to making computers understand the statements or words written in human languages"[3]. Applied on radiology reports, methods related to NLP can extract clinically relevant information. Specifically, information extraction (IE) provides techniques to use this clinical information for secondary purposes, such as prediction, quality assurance or research.

IE, a subfield within NLP, involves extracting pertinent information from free-text. Subtasks include named entity recognition (NER), relation extraction (RE), and template filling. These subtasks are realized using heuristic-based methods, machine learning-based techniques (e.g., support vector machines or Naıve Bayes), and deep learning-based methods[4]. Within the field of deep learning, a new architecture of models has recently emerged - namely large language models (LLMs).

LLMs are "deep learning models with a huge number of parameters trained in an unsupervised way on large volumes of text"[5]. These models typically exceed one million parameters and have proven highly effective in information extraction tasks. The transformer architecture, introduced in 2017, serves as the foundation for most contemporary LLMs, comprising two distinct architectural blocks; the encoder and the decoder. Both blocks apply an innovative approach of creating contextualized word embeddings called attention[6]. Prior to the "age of transformers" still present today, recurrent neural network (RNN)-based LLMs were regarded as state-of-the-art for creating contextualized word embeddings. ELMo, a language model based on a bidirectional Long Short Term Memory (BiLSTM) network[7], is an example thereof. Noteworthy transformer-based LLMs include encoder-based models like BERT (2018)[8], decoder-based models like GPT-3 (2020)[9] and GPT-4 (2023)[10], as well as models applying both encoder and decocoder blocks, e.g., Megatron-ML (2019)[11]. Models continue to evolve, being trained on expanding datasets and consistently surpassing the performance benchmarks established by previous state-of-the-art models. The question arises how these new models shape IE applied to radiology reports.

[1]Institute for Patient-Centered Digital Health, Bern University of Applied Sciences, Biel/Bienne, Switzerland. [2]Faculty of Medicine, University of Geneva, Geneva, Switzerland. [3]Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland. [4]Informatics Institute, HES-SO Valais-Wallis, Sierre, Switzerland. ✉e-mail: daniel.reichenpfader@bfh.ch

Table 1 | Overview of existing literature

Authors	Year	Scope
Pons et al.[12]	2016	SR: NLP in radiology
Casey et al.[13]	2021	SR: NLP applied to radiology reports
Davidson et al.[14]	2021	SR: Quality of NLP studies applied to radiology reports
Saha et al.[15]	2023	ScR: NLP applied to breast cancer reports
Gholipour et al.[16]	2023	SR: NLP-based extraction of cancer concepts from clinical notes
Gorenstein et al.[17]	2024	SR: BERT-based NLP in radiology

SR systematic review, *ScR* scoping review.

Regarding existing literature concerning IE from radiology reports, several reviews are available, although these sources either miss current developments or only focus on a specific aspect or clinical domain, see Table 1. The application of NLP to radiology reports for IE has already been subject to two systematic reviews in 2016[12] and 2021[13]. While the former is not freely available, the latter searches only Google Scholar and includes only one study based on LLMs. Davidson et al. focused on comparing the quality of studies applying NLP-related methods to radiology reports[14]. More recent reviews include a specific scoping review on the application of NLP to reports specifically related to breast cancer[15], the extraction of cancer concepts from clinical notes[16], and a systematic review on BERT-based NLP applications in radiology without a specific focus on information extraction[17].

As LLMs have only recently gained a strong momentum, a research gap exists as there is no overview of LLM-based approaches for IE from radiology reports available. With this scoping review, we therefore intend to answer the following research question:

What is the state of research regarding information extraction from free-text radiology reports based on LLMs?

Specifically, we are interested in the subquestions that arise from the posed research question:

- RQ.01 - Performance: What is the performance of LLMs for information extraction from radiology reports?
- RQ.02 - Training and Modeling: Which models are used and how is the pre-training and fine-tuning process designed?
- RQ.03 - Use cases: Which modalities and anatomical regions do the analyzed reports correspond to?
- RQ.04 - Data and annotation: How much data was used to train the model, how was the annotation process designed and is the data publicly available?
- RQ.05 - Challenges: What are open challenges and common limitations of existing approaches?

The objective of this scoping review is to answer the above-mentioned questions, provide an overview of recent developments, identify key trends and highlight future research by identifying outstanding challenges and limitations of current approaches.

Results

Study selection

As shown in Fig. 1, the systematic search yielded 1,237 records, retrieved from five databases. After removing duplicate records and records published before 2018, 374 records (title, abstract) were screened for eligibility. The screening process resulted in the exclusion of 302 records. The remaining 72 records were sought for full-text-retrieval, of which 68 could be retrieved. During data extraction, 43 papers were excluded due to not fulfilling inclusion criteria, which was not apparent based on information provided in the abstract.

Within the cited references of included papers, nine additional papers fulfilling all inclusion criteria were identified. Therefore, following the above-mentioned methodology, 34 records in total were included in this review.

Study characteristics

In the following, we organize the extracted information according to the structure of the extraction table, which in turn reflects the defined research questions. This review covers studies that were published between 01/01/2018 and 01/08/2023. The earliest study included was published in 2019. After eight included studies published in 2020, the topic reaches its peak with eleven studies published in 2021. Eight studies of 2022 were included. Six included studies were published in the first half of 2023.

Based on corresponding author address, 15 out of 35 papers are located in the USA, followed by six in China and three each in the UK and Germany. Other countries include Austria ($n = 1$), Canada ($n = 2$), Japan ($n = 2$), Spain ($n = 1$) and The Netherlands ($n = 1$) (Table 2).

Extracted information

This chapter describes the NLP task, the extracted entities, the information model development process and data normalization strategies of the included studies.

Extracted concepts encompass various entities, attributes, and relations. These concepts relate to abnormalities[18–20], anatomical information[21], breast-cancer related concepts[22], clinical findings[23–25], devices[26], diagnoses[27,28], observations[29], pathological concepts[30], protected health information (PHI)[31], recommendations[32], scores (TI-RADS[33], tumor response categories[34]), spatial expressions[35–37], staging-related information[38,39], and stroke phenotypes[40]. Several papers extract various concepts, e.g., ref. 41.

Studies solely describing document-level single-label classification were excluded from this review. Two studies apply document-level multi-class classification. Document-level multi-label classification is described in nine studies (26%), whereof three only classify more than two classes for each entity. The majority of the included studies ($n = 21$, 62%) describes NER methods, ten studies additionally apply RE methods. These studies encompass sequence-labeling and span-labeling approaches. Question answering (QA)-based methods are described in two studies, see Fig. 2.

The number of extracted concepts (including entities, attributes, and relations) ranges from one entity in both papers describing multi-class classification[33,34] up to 64 entities described in a NER-based study[30].

Three studies base their information model on clinical guidelines, namely the *Response evaluation criteria in solid tumors*[42] and the *TNM Classification of Malignant Tumors* (TNM) staging system[43]. Development by domain experts ($n = 2$), references to previous studies ($n = 3$), regulations of the Health Insurance Portability and Accountability Act[44] ($n = 1$), the Stanza radiology model[45] ($n = 1$) and references to previously developed schemes ($n = 2$) are other foundations for information model development. One study provides detailed information about the development process of the information model as supplementary information[19]. One study reports development of their information model based on the RadLex terminology[46], another based on the National Cancer Institute Thesaurus[47]. 21 studies (62%) do not report any details regarding the development of the information model.

Out of the 34 included studies, only three describe methods to structure and/or normalize extracted information. While Torres-Lopez et al. apply rule-based methods to structure extracted data based on entity positions and combinations[30], Sugimoto et al. additionally apply rule-based normalization based on a concept table[24]. Datta et al. describe a hybrid approach to normalize extracted entities by first generating concept candidates with BM25, a ranking algorithm, and then choosing the best equivalent with a BERT-based classifier[48].

Regarding the distribution of annotated entities within the datasets, only one study reports on having conducted measures to counteract class imbalance[19]. Another study reports on not having used F1 score as a performance measure, as the F1 score is not suited when class imbalances are present[27]. Four studies (12%) report coarse entity distributions and seven studies (21%) describe granular entity distributions.

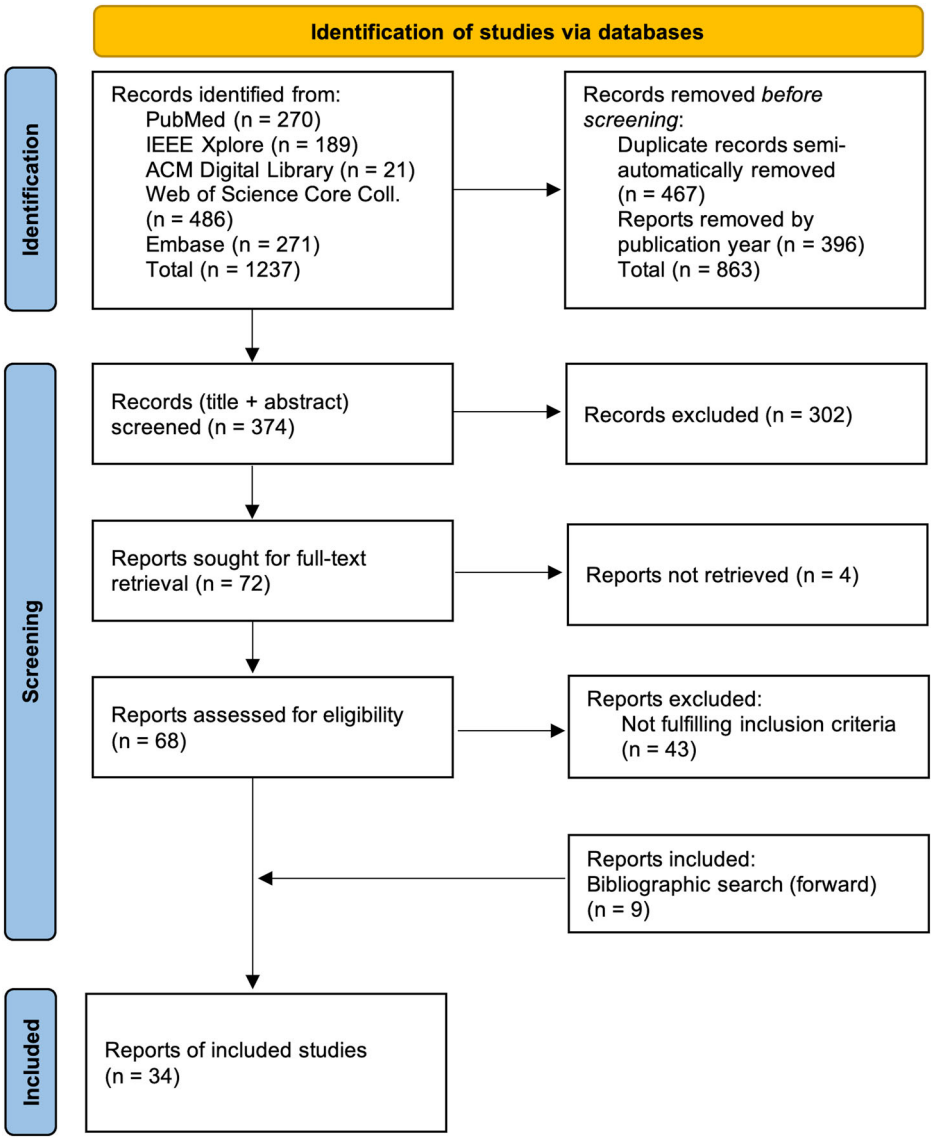

Fig. 1 | PRISMA flowchart describing the source of evidence retrieval and selection process. Querying of five databases resulted in a total of 1237 sources of evidence eligible for screening. This number was reduced to 374 after deduplication and removal based on publication year. Eventually, 34 studies were included in this review after completion of the screening process.

Table 2 | Countries of publication

Country of publication	Frequency	References
USA	15	21,23,26,29,30,32,33,35–37,40,48,57,59,60
China	6	20,28,38,39,51,54
UK	3	18,19,27
Germany	3	34,41,63
Canada	2	22,25
Japan	2	24,46
Austria	1	53
Spain	1	31
The Netherlands	1	107

Model

In the following, details regarding the reported model architectures and implementations are described, including base models, (further) pre-training and fine-tuning methods, hyperparameters, performance measures, external validation and hardware details.

For an overview of applied model architectures, see Table 3. 28 out of 34 papers (82%) describe at least one transformer-based architecture, while the remaining six studies apply various adaptions of the Bidirectional Long Short-Term Memory (Bi-LSTM) architecture. Out of the 28 studies that describe transformer-based architectures, 27 are based on the BERT architecture[8] and one is based on the ERNIE architecture[49]. Eight studies (24%) describe further pre-training of a BERT-based, pre-trained model on in-house data. Eighteen studies (53%) use a BERT-based, pre-trained model without further pre-training. One study applies pre-training to other layers than the LLM. Two studies do not provide any details regarding the architecture of the BERT models. One study combines both BERT- and BiLSTM-based architectures[28]. Out of six studies that describe only BiLSTM-based architectures, two studies apply pre-training of word vectors based on word2vec[50]. 31 studies (91%) provide sufficient details about the fine-tuning process. Three studies do not provide details[24,39,51].

Reported performance measures vary between included studies, including traditional measures like precision, recall, and accuracy as well as different variations of the F1 score (micro, macro, averaged, weighted, pooled). The performance of studies reporting a F1-score variation (including micro-, macro-, pooled- generalized, exact match and weighted F1) is compared in Table 4. If a study describes multiple models, the score of the best model was chosen. If two or more datasets are compared, the higher

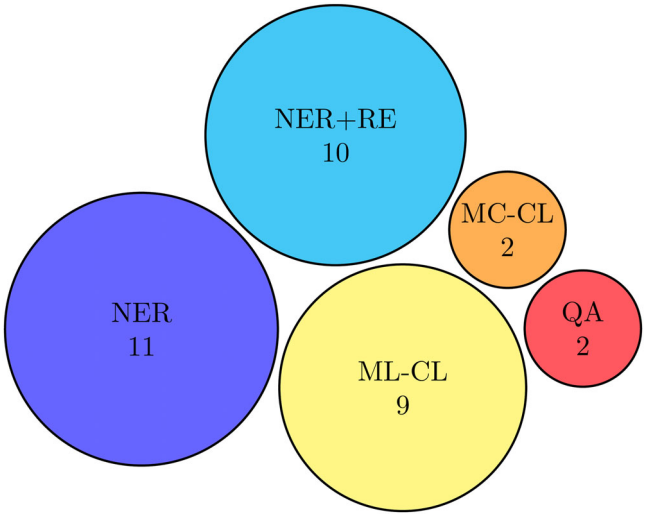

Fig. 2 | Distribution of reported NLP tasks. The circles contain the absolute number of studies per task. NER Named entity recognition, RE Relation extraction, ML-CL Binary multi-label classification, MC-CL Multi-class classification, QA Question answering.

Table 3 | Overview of reported BERT-based model architectures

Architecture	Frequency	References
BERTBase[8]	5	22,23,26,33,59
MIMIC BERTBase[107]	4	37,40,48,60
MIMIC BERTLarge[107]	3	48,57,60
RoBERTa[108]	2	26,41
BioBERT[109]	2	18,19
Clinical BERT[110]	2	22,23
German BERT[111]	2	34,63
BERTLarge[8]	1	35
PubMedBERT[112]	1	26
DistilBERT[113]	1	26
DeBERTa[104]	1	26
BERT WWM[114]	1	38
BlueBERT[115]	1	29
G-BERT[116]	1	41
GM-BERT[117]	1	41
MULTI-BERTs[118]	1	63
R-BERT[119]	1	53
SciBERT[120]	1	27
SpERT[121]	1	21
XLNet Large[122]	1	35

score was chosen. If applicable, the result of external validation is also presented. 22 studies (65%) report having conducted statistical tests, including cross-validation, McNemar test, Mann-Whitney *U* test and Tukey-Kramer test.

Hyperparameters used to train the models (e.g., learning rate, batch size, embedding dimensions) are described in 28 studies (82%), however with varying degree of detail. Six studies (18%) do not report any details on hyperparameters. Seven studies (21%) describe a validation of their algorithm on training data from an external institution. Seven studies (21%) include details about hardware and computational resources spent during the training process.

Table 4 | Performance overview of studies reporting averaged F1-scores

Performance (%, F1 score)	Performance external (%)	Extracted concepts[a](*n*)	Reference
68.76		9	57
70.00		1	34
74.00	39.00	3	62
75.93		9	48
80.10 (macro-F1, exact matching)		18	39
80.40 (weighted average)		13	29
81.10		5	37
82.66		4	40
83.97		27	41
85.60		5	35
85.96 (macro-F1)		14	38
86.00 (micro-average, pooled)		5	27
90.49		5	36
93.53		75	54
93.80		5	24
95.00		5	25
95.36 (micro-average)	94.62	7	46
96.00	92.84	64	30
97.72	92.63	7	31
98.00 (weighted average)	85.00	1	33

[a]Including entities, relations, and attributes.

Data sets

In this section, we describe the study characteristics related to data sets, encompassing number of reports, data splits, modalities, anatomic regions, origin, language, and ethics approval.

Data set size used for fine-tuning ranges from 50 to 10,155 reports. The amount of external validation data ranges from 10% to 31% of the amount of data used for fine-tuning. For further pre-training of transformer-based architectures, 50,000 up to 3.8 million reports are used. Jantscher et al. additionally use the content of a public clinical knowledge platform (*DocCheck Flexicon*[52])[53]. Zhang et al. only report the amount of data (3 GB)[54]. Jaiswal et al. performed further pre-training on the complete MIMIC-CXR corpus[29]. Two studies that described pre-training of word embeddings for Bi-LSTM-based architectures used 3.3 million and 317,130 reports, respectively[24,32].

Data splits vary widely; the majority of studies ($n = 23$, 68%) divide their data into three sets, namely train-, validation- and test-set, with the most common split being 80/10/10, respectively. This split variation is reported in eight studies (24%). Seven studies (21%) use two sets only, four studies (12%) apply cross-validation-based methods.

19 studies (56%) describe the timeframe within which reports had been extracted. Dada et al. report the longest timeframe of 22 years, using reports between 1999 and 2021 for further pre-training[41]. The shortest timeframe reported is less than one year (2020–2021)[26].

Several studies are based on publicly available datasets: MIMIC-CXR[55] was used once[29] while MIMIC[56] was used by two studies[40,57]. MIMIC-III[58] was used by six studies (18%)[37,40,48,57,59,60]. The Indiana chest X-ray collection[61] was used twice[35,36]. For external validation, MIMIC-II was applied by Mithun et al.[62] and MIMIC-CXR by Lau et al.[23]. While some of these studies use the datasets as-is, some perform additional annotation. Other studies use data from hospitals, hospital networks, other tertiary care

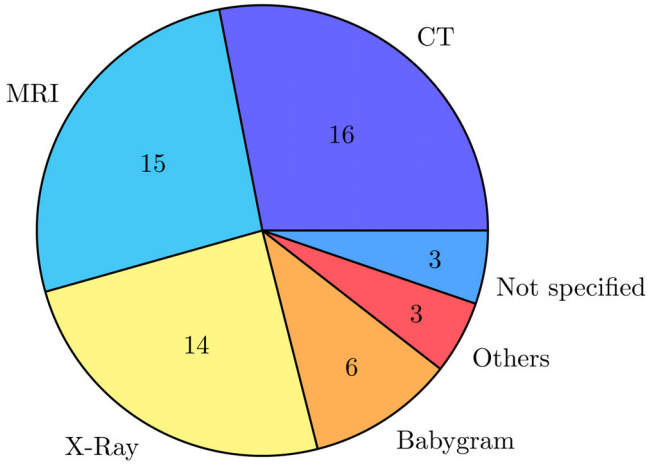

Fig. 3 | Distribution of modalities. The diagram shows absolute numbers of mentioned modalities. Several studies use reports obtained from multiple modalities. Other modalities include positron emission tomography-computed tomography (PET-CT) ($n = 1$) and ultrasound ($n = 2$). Three studies did not explicitly mention associated modalities. Abbreviations: CT Computer tomography, MRI Magnetic resonance imaging.

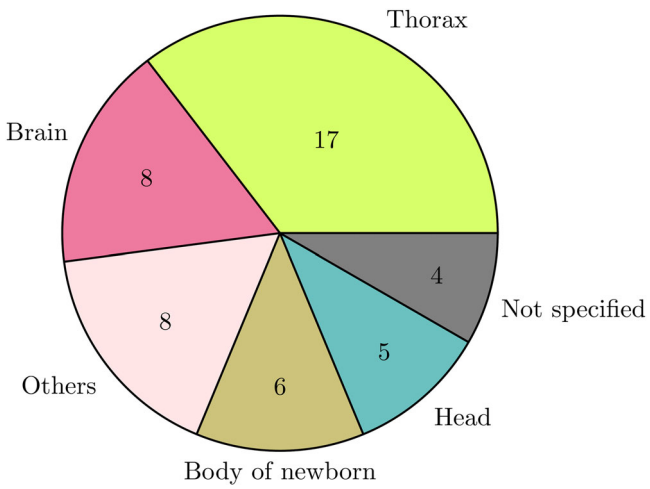

Fig. 4 | Distribution of anatomical regions. The diagram shows absolute numbers of mentioned anatomical regions. Several studies use reports corresponding to multiple anatomical regions. Other anatomical regions include the heart, abdomen, pelvis, "all body regions", nose, thyroid ($n = 1$ each) and breast ($n = 2$). Four studies did not explicitly mention associated anatomical regions.

institutions, medical big data companies, research centers, care centers or university research repositories.

Figures 3 and 4 show the frequencies of modalities and anatomical regions, respectively. Note that frequencies were counted on study-level and not weighted by the number of reports.

Report language was inferred from the location of the institution of the corresponding author: Most studies use English reports ($n = 21$, 62%) followed by Chinese ($n = 6$, 18%), German ($n = 4$, 12%), Japanese ($n = 2$, 6%) and Spanish ($n = 1$). The corresponding author address of one study is located in the Netherlands but using data from an Indian Hospital[62].

19 studies (56%) explicitly state that the endeavor was approved by either a national committee or agency ($n = 3$, 9%) or a local institutional or hospital review board or committee ($n = 15$, 44%). One study reports approval only for in-house data, but not for the external validation set from another institution[33].

Annotation process

28 studies (82%) describe an exclusively manual annotation process. Five studies (15%) explicitly state that each report was annotated by two persons independently. Lau et al. use annotated data to train a classifier that supports the annotation process by proposing only documents that contain potential annotations[32]. Two studies use tools for automated annotation with manual correction and review[29,31]. Lybarger et al. do not provide details on their augmentation of an existing dataset[21], three others do not report details as they either extract information available in the hospital information system[33] or exclusively use existing annotated datasets[36,59].

Annotation tagging schemes mentioned include IOB(2), BISO and BIOES (short for beginning, inside, outside, start, end). The number of involved annotators ranges from one to five, roles include clinical coordinators, radiologists, radiology residents, medical and graduate students, medical informatics engineers, neurologists, neuro-radiologists, surgeons, radiological technologists and internists. Existing annotation guidelines are reported by three studies, four studies mention that instructions exist but do not provide details. 23 studies (68%) do not mention information regarding annotation guidelines.

Inter-annotator-agreement (IAA) is reported by 23 (68%) studies. Measures include F1 score variants ($n = 8$, 24%), Cohen kappa ($n = 7$, 21%), Fleiss kappa ($n = 19$, 56%) and the intraclass correlation coefficient ($n = 1$). IAA results are reported by 16 studies (47%) and range, for Cohen kappa, from 81% to 93.7%. Eleven studies (32%) mention the tool used for annotation, including Brat[23,37,39,48,53,60], Doccano[34], TagEditor[30], Talen[46] and two self-developed tools[19,63].

Data and source code availability

Five studies (15%) state that data is available upon request. One study claims availability, although there is no data present in the referenced online repository[57]. One study published its dataset in a GitHub repository[35]. One study only uses annotations provided within a dataset with credentialed access[59]. The remaining 22 studies (65%) do not mention whether data is available or not. Regarding source code availability, ten studies (29%) claim their code to be available. The remaining 24 studies (71%) do not mention whether the source code is available or not.

Challenges and limitations

Various aspects related to limitations and challenges are described. The most common mentioned limitation is that studies use only data from a single institution[21,22,24,30,36,51,53]. Similarly, multiple studies mention validation on external or multi-institutional data as a future research direction[19,26,59]. Two studies mention the need of semantic enrichment or normalization of extracted information[48,54].

Many studies report intentions to augment their described approaches to other report types[21,28,30,37], other report sections[22], to include other or more data sources[35,39,54] or entities[32,62], body parts[46], clinical contexts[34] or modalities[35,53,59].

Additional limitations include the application to only a single modality or clinical area[21,46,53], small dataset size[27,32,54], technical limitations[27,63], no negation detection[35,62], few extracted entities[24,28] or result degradation upon evaluation on external data[19] or more recent reports[25]. Missing interpretability is mentioned by two studies[28,41].

Discussion

Performance measures reported in Table 4 cannot be compared due to differences in datasets, number of extracted concepts and the heterogeneity of applied performance measures. External validation performed by six studies shows in general lower performance of the algorithm applied to external data, so data from a source different from the one used for training. The largest performance drop of 35% (overall F1 score) was reported in a Bi-LSTM-based study, performing multi-label binary classification of only three entities on the document-level[62]. On the contrary, Torres-Lopez et al. extracted a total of 64 entities with a performance drop of only 3.16% (F1 score), although not providing details on their model architecture. The

smallest performance drop amounts to only 0.74% (Micro F1) for extracting seven entities based on a further pre-trained model[46]. However, it cannot be assumed that further pre-training increases model generalizability and therefore performance.

Upon analysis of performance, several inconsistencies between included studies impairs comparability: First, there is no standardized measure or best-practice to assess model performance for information extraction. Although in general, the F1 score is most often applied and well known, there exist many variations, including micro-, macro-, exact and inexact match scores, weighted F1 score and 1-Margin F1 scores. On the contrary, Zaman et al. argue that macro-averaged F1 score or overall accuracy are not suited as performance measures when class imbalances are present[27]. For the same reason, F1 score is only used to assess binary classification and not for multi-class classification by Wood et al.[19].

While 22 studies apply some variation of cross-validation to assess model performance, 12 studies apply simple split validation methods. Singh et al. show that if data sets are small, simple split validation shows significant differences of performance measures compared to cross-validation[64].

Specific statistical tests to compare performance of different models include DeLong's test to compare Area under the ROC Curves[19,27], the Tukey-Kramer method for multiple comparison analysis[46] and the McNemar test to compare the agreement between two models[22]. However, appropriateness of each test method remains unclear, as shown by Demner et al.[65].

In general, equations on how performance metrics are computed should always be included in the manuscript to improve understandability, e.g., as done by[22] or[30]. To improve comparability of studies, scores for each class as well as a reasonable aggregated score over all classes should be reported.

This review identified only decoder-based architectures or pre-transformer architectures and no generative models, such as GPT-4 (released in March 2023). The majority of the described models is based on the encoder-only BERT architecture, first described by Devlin et al.[8]. We envision multiple reasons: First, while having been available since 2018[66], generative models first needed time to be established as a new technology to be investigated and applied in the healthcare sector. Second, early generative models might have demonstrated poor performance due to their relatively small size and lack of domain-specific data for pre-training[67]. Third, poor performance might also entail model hallucinations: Farquhar et al. define hallucination as "answering unreliably or without necessary information"[68]. Hallucinations include, among others, provision of wrong answers due to erroneous training data, lying in pursuit of a reward or errors related to reasoning and generalization[68]. On the contrary, encoder-only models like the BERT architecture cannot hallucinate as they provide only context-aware embeddings of input data; the actual NLP task (e.g., sequence labeling, classification or regression) is performed by a relatively simple, downstream neural network, rendering this architecture more transparent and verifiable than generative models.

An advantage of LLMs is their capability to be customized to a specific language or general domain (e.g., medicine): First, a base version of the model is trained using a large amount of unlabeled data: This process is called pre-training. The concept of transfer-learning enables researchers to further customize a pre-trained model to a more specific domain (e.g., clinical domain, another language or from a certain hospital). This is also referred to as further pre-training. The process of training the model to perform a particular NLP task (e.g., classification) based on labeled data is called fine-tuning. These definitions (pre-training, further pre-training, transfer learning and fine-tuning) tend to be confused by authors or replaced by other term variants, e.g., "supervised learning". However, it is imperative to use clear and concise language to distinguish between the concepts mentioned above.

Seven included studies apply further pre-training as defined above. The effect of further pre-training depends on various factors, including specifications of the input model used or amount and quality of the data used for further pre-training. Interestingly, further pre-training of a pre-trained model to another language was not reported.

Opposed to the traditional further pre-training as described above, Jaiswal et al. show how BERT-based models achieve higher performance when little data is available based on contrastive pre-training[29]. The authors claim that their model achieves better results than conventional transformers when the number of annotated reports is limited.

Only two studies solve the task of information extraction based on extractive question answering[41,59]. Extractive question answering was already described in the original BERT paper[8]: Instead of generating a pooled embedding of the input text or one embedding per input token, a BERT model fine-tuned for question answering takes an answer as an input and outputs the start and end token of the text span that contains the answer to the posed question - this is also possible if no answer or multiple answers are contained within the text as shown by Zhang et al.[69].

The most common modalities for which reports of findings were used in the included studies are CT ($n = 16$), MRI ($n = 15$) and X-Ray ($n = 14$). CT reports appear to be the most common source when using in-house data. According to data provided by the Organisation for Economic Cooperation and Development (OECD), the availability of CT scanners and MRI machines has increased steadily during the past decades. Furthermore, there has been a general upwards trend in the number of performed CT and MRI interventions worldwide[70]. CT exams are fast and cheap compared to MRI.

The most common anatomical regions studied are thorax ($n = 17$) and brain ($n = 8$). There might be different reasons for this distribution. First, chest X-Ray is one of the most frequently performed imaging examinations. Second, six studies used reports obtained from MIMIC datasets, including thorax X-Ray, brain MRI and babygram examinations. Two studies used thorax X-Ray reports obtained from publicly available datasets. Furthermore, a report on the annual exposure from medical imaging in Switzerland shows that the thorax region is the third most common anatomical region of CT procedures (11.8%), preceded by abdomen and thorax (16.4%) and abdomen only (17.7%)[71].

We identified several aspects that showed different interpretations in the included studies. One of the major ambiguities discovered is the clear definition of the terms test set and validation set: Some studies use these two very distinct terms interchangeably. However, agreement is needed upon which set is used during parameter optimization of a model and which set is used for evaluation of the final model. Furthermore, studies either report number of sentences or number of documents, hindering comparability. It also remains unclear, whether the stated dataset size includes documents without annotation or annotated data only. Report language is never explicitly stated.

Regarding annotation, it becomes apparent that there is no standard for IAA calculation, recommended number of annotator and their backgrounds, number of reports, number of reconciliation rounds and especially, IAA calculation methods. All these aspects differ widely in the included papers.

Good practices observed in the included papers include reporting of descriptive annotation statistics[35] and conducting complexity analysis of the report corpus[29,34]: These complexity metrics include e.g., unique n-gram counts, lexical diversity as measured with the Yule 1 score and the Type-Token-Ratio, as reported in ref. 46. Wood et al. highlight the importance of splitting data on patient-level instead of report level[19].

Last, we want to highlight interesting approaches: Fine et al. first use structured reports for fine-tuning and then apply the resulting model on unstructured reports[34]. Jaiswal et al. introduce three novel data augmentation techniques before fine-tuning their model based on contrastive learning[29]. Pérez-Díez et al. developed a randomization algorithm to substitute detected entities with synthetic alternatives to disguise undetected personal information[31].

The mentioned challenges and limitations are manifold and diverse. Ten papers in total address the topic of generalizing to data from other institutions. Another challenge are the limitations of every study, be it a limited number of entities and usually a single modality and clinical domain.

Every included study is based on a pre-defined information model and fine-tuned on annotated data. This means, that by August 2023, no truly generalized approach for IE has been described in the identified literature.

Upon interpretation of the above-mentioned results, several limitations of this review can be mentioned. First, the definition of *information extraction* proved to be challenging. We defined information extraction as a collective term for the NLP tasks of document-level multi-label classification (including binary or multiple classes for each label), NER (including RE), as well as question answering approaches. We excluded binary classification on the document level. While a narrow definition of IE would possibly only include NER and RE, whereas the widest definition would also include binary document classification. With our approach, we wanted to ensure a balanced level of task complexity.

Furthermore, the definition of an LLM was also unclear. In the protocol for this review, LLMs are defined as "deep learning models with more than one million parameters, trained on unlabeled text data"[72]. Although BiLSTM-based architectures are not trained on text, the applied context-aware word embeddings like fastText and word2-vec stipulate the inclusion of these architectures into this review. An additional argument for including BiLSTM-based architectures is ELMO, a BiLSTM-based architecture with ~ 13M parameters, and referred to as one of the first LLMs. However, we decided not to include BiGRU-based architectures, as information on their parameter count was usually not available. A more narrow definition would only include transformer-based architectures, having billions of parameters. This definition seems to have recently reached consensus among researchers and in industry. As of the time of submission in June 2024, LLMs tend to be defined even more narrow, only including generative models based on autoregressive sampling[73]. This might be due to generative models currently being the most common and frequent model architecture. On the contrary, a wider definition would also potentially include BiGRU-based, CNN-based and other architectures. It also remains subject to discussion whether summarization can be regarded as information extraction—for this study, summarization was not included, potentially missing studies of interest, e.g., ref. 74. Likewise, image-to-text report generation was excluded.

Regarding the search strategy, we decided not to include numerous model names to keep the complexity of the search term low. Instead, we initially only included the terms *transformers* and *Bert*. Eventually, only two search dimensions were used because otherwise, the number of search results would have been too small. To minimize the number of missed studies, the forward search of references of included studies was carried out, eventually leading to nine additionally included studies that were not covered by the search strategy. Nevertheless, our search strategy was not exhaustive: Studies that used terms related to *transformation* or *structuring* of reports, e.g.,refs. 75,76, were missed as these terms are missing in the search strategy.

No generative models and therefore no approaches based on generative models (including few-, single- or zero-shot learning) are included in the search results. This might be due to the fact that generative models have only started to become widely accessible with the publication of chatGPT in November 2022. Only later, open-source alternatives became available. However, due to the sensitive nature of patient data, utilization of publicly serviced models, e.g., GPT-4, is restricted due to data protection rules. Until the cut-off time of this review, state-of-the-art, open-source generative models, e.g., LLama 2 (70B), had still required vast computational resources, restricting the possibilities of on-premise deployment within hospital infrastructures. Furthermore, early studies might so far only be published without peer-review (e.g., on arXiv), excluding them for this review, e.g., ref. 77. As no search updates were performed for this review, arXiv papers that were later peer-reviewed were also not included, e.g.,[78]. Relevant papers published in the ACL Anthology were also not included, potentially missing papers describing generative approaches, e.g., by Agrawal et al.[79] and Kartchner et al.[80]. Sources that did not mention "information extraction", "named entity recognition" or "relation extraction" in the title or abstract and were not referred to by other papers were also not included, e.g., ref. 81.

Given the diverse nature of the included studies alongside discrepancies in both the quality and quantity of reported data, a comprehensive analysis of the extracted information was deemed impossible. Future systematic reviews could enhance this comparison by refining the research question and subquestions to a more specific scope. However, according to the protocol for this scoping review, a purely descriptive presentation of findings was conducted.

Another potential limitation is the fact that data extraction was performed by one author (DR) only. However, prior to data extraction, two studies were extracted by two authors, and the resulting information compared. This led to the addition of six additional aspects to the original data extraction table, including details on hardware specification, hyperparameters, ethical approval, timeframe of dataset and class imbalance measures.

Last, we want to highlight that this scoping review strictly adheres to the PRISMA-ScR and PRISMA-S guidelines. Our search strategy of five databases resulted in over 1200 primary search results, minimizing the risk of missing relevant studies. This risk was further minimized by carefully choosing a balanced definition of both IE and LLMs. As only peer-reviewed studies were taken into account, a certain study quality was furthermore ensured.

Due to the current rapid technical progress, we summarize the latest developments regarding LLMs in general, their application in medicine, as well with regard to this review's topic. We give an overview on studies published outside the scope of our review (published after August 1st 2023) as well as on the application of LLMs in clinical domains and tasks different from IE from radiology reports.

As of June 2024, the majority of recently published LLMs, be it commercial or open-source, are generative models, based on the decoder-block of the original transformer architecture. Two development strategies can be observed to increase model performance: The first strategy is about simply increasing the amount of model parameters (and therefore, model size), leading also to an increased demand for training data. The second strategy, on the other hand, is about optimizing existing models based on different strategies, including model pruning, quantization or distillation, as shown by Rohanian et al.[82]. Recent models include the Gemini family (2024)[83], the T5 family[84], LLama 3 (2024)[85] and Mixtral (2024)[86]. Moreover, research has increasingly been focussing on developing domain-specific models, e.g., Meditron, Med-PaLM 2, or Med-Gemini for the healthcare domain[87–89].

In the broad clinical domain, these recent, generative LLMs show impressive capabilities, partly outperforming clinicians in test settings regarding, e.g., medical summary generation[90], prediction of clinical outcomes[91] and answering of clinical questions[92]. Dagdelen et al. have recently demonstrated that, in the context of structured information extraction from scientific texts, even generative models require a few hundred training examples to effectively extract and organize information using the open-source model Llama-2[93].

For the specific topic of structured IE from radiology reports, several papers and pre-prints have been published since August 2023: In general, it becomes apparent that resource-demanding generative models seem not to show better results compared to encoder-based approaches, as shown by the following studies: When applying the open-source model Vicuna[94] to binary label 13 concepts on document-level of radiology reports, Mukherjee et al. showed only moderate to substantial agreement with existing, less resource-demanding approaches[95]. Document-level binary level was also investigated by Adams et al., who compared GPT-4 to a BERT-based model further pre-trained on German medical documents[75]. In this comparison, the smaller, open-source model[96] outperformed GPT-4 for five out of nine concepts. The authors also tested GPT-4 on English radiology reports, however not providing any detailed performance measures. Similarily, Hu et al. used ChatGPT as a commercial platform to extract eleven concepts from radiology reports without further fine-tuning or provision of examples[97]. The

results show inferiority of ChatGPT upon comparison with a previously described approach (BERT-based multiturn question answering[98]) as well as a rule-based approach (averaged F1 scores: 0.88, 0.91, 0.93, respectively). Mallio et al. qualitatively compared several closed-source generative LLMs for structured reporting, although lacking clear results[99]. Additionally, several key gaps remain with the application of above-mentioned generative models. For example, closed-source models continue getting larger, requiring an increasing extent of scarce hardware resources and training data. Moreover, although large generative models currently show the best performance, they are less explainable than, e.g., encoder-based architectures prevalent in this review's results[100].

Generative models and encoder-based models each offer unique advantages and disadvantages. Yang et al. show that generative models might excel at generalizing to external data by applying in-context learning[101]. Generative models are by design able to aggregate information, and might be therefore more suitable to extract more complex concepts. Recently, open-source models are becoming more efficient and compact, as seen in recent advancements, e.g., the Phi 3 model family[102]. However, generative models are usually computationally intensive and require substantial resources for training and deployment. While still facing issues regarding hallucination, this behavior might be improved by combining LLMs with knowledge graphs, as introduced by Gilbert et al.[103].

On the other hand, encoder-based models, such as BERT, are highly effective at understanding and generating bidirectional contextual embeddings of input data, which makes them particularly strong in tasks requiring precise comprehension or annotation of text, such as extractive question answering or NER. They tend to be more resource-efficient during inference compared to generative models. However, encoder-based models often struggle with generating coherent text, a task where generative models excel. Additionally, while encoder-based models can be fine-tuned for specific tasks, they may not generalize as well as generative models. Moreover, research and industry currently focus on the development of generative models, as the last encoder-based architecture was published in 2021[104]. In summary, while generative models currently offer flexibility and powerful aggregation capabilities, encoder-based models provide efficiency and precision.

In this review, we provide a comprehensive overview of recent studies on LLM-based information extraction from radiology reports, published between January 2018 and August 2023. No generative model architectures for IE from radiology reports were described in literature. After August 2023, generative models have been becoming more common, however tending not to show a performance increase compared to pre-transformer and encoder-based architectures. According to the included studies, pre-transformer and encoder-based models show promising results, although comparison is hindered by different performance score calculation methods and vastly different data sets and tasks. LLMs might improve generalizability of IE methods, although external validation is performed in only seven studies. The majority of studies used pre-trained LLMs without further pre-training on their own data. So far, research has focused on IE from reports related to CT and MRI examinations and most frequently on reports related to the thorax region. We recognize a lack of publicly available datasets. Furthermore, a lack of standardization of the annotation process results in potential differences regarding data quality. The source code is made available by only ten studies, limiting reproducibility of the described methods. Most common challenges reported are missing validation on external data and augmentation of the described method to other clinical domains, report types, concepts, modalities and anatomical regions.

No generative model architectures for IE from radiology reports were described in literature. After August 2023, generative models have been becoming more common, however tending not to show a performance increase compared to pre-transformer and encoder-based architectures. According to the included studies, pre-transformer and encoder-based models show promising results, although comparison is hindered by different performance score calculation methods and vastly different data sets and tasks. LLMs might improve generalizability of IE methods, although external validation is performed in only seven studies.

We conclude by highlighting the need to facilitate comparability of studies and to review generative AI-based approaches. We therefore plan to develop a reporting framework for clinical application of NLP methods. This need is confirmed by Davidson et al. who also state that available guidelines are limited[14]; journal-specific guidelines already exist[105]. Considering the periodical publication of larger, more capable generative models, transparent and verifiable reporting of all aspects described in this review is essential to compare and identify successful approaches. We furthermore suggest future research to focus on the optimization and standardization of annotation processes to develop few-shot prompts. Currently, the correlation between annotation quality, quantity and model performance is unknown. Last, we recommend the development and publication of standardized, multilingual datasets to foster external validation of models.

Methods

This scoping review was conducted according to the JBI Manual for evidence synthesis and adheres to the PRISMA extension for scoping reviews (PRISMA-ScR). Regarding methodological details, we refer to the published protocol for this review[72]. In this section, we give an overview on the applied methodology and explain the adaptations made to the protocol. The completed PRISMA-ScR checklist is provided in Supplementary Table 1.

Search strategy

The search strategy comprised three steps: First, a preliminary search was conducted by searching two databases (Google Scholar and PubMed), using keywords related to this review's research question. Based on the results, a list of relevant search and index terms was retrieved, which in turn served as a basis for the iterative development of the full search query.

During search query development, different combinations of terms and dimensions of the research topic were combined to build query combinations that were run on PubMed. Balancing of search results and relevance showed that the inclusion of only two dimensions, "radiology" and "information extraction", showed the best balance regarding the quantity and quality of results and was therefore chosen as the final search query.

Second, a systematic search was carried out using the final version of the search query. The PubMed-based query was adapted to meet syntactical requirements of the other four databases, comprising IEEE Xplore, ACM Digital Library, Web of Science Core Collection and Embase. The systematic search was conducted on 01/08/2023, and included all sources of evidence (SOE) since database inception. No additional limits, restrictions, or filters were applied. The full query for each database as well as a completed PRISMA-S extension checklist are shown in Supplementary Table 2 and Supplementary Table 3. Third, reference lists of included studies were manually checked for additional sources of evidence and included if fulfilling all inclusion criteria. No search updates were performed.

Inclusion criteria

Inclusion criteria were discussed among and agreed on by all three authors. No separation was made between exclusion and inclusion criteria; reports were included upon fulfillment of all the following six aspects:

- C.01: The full-text SOE is retrievable.
- C.02: The SOE was published after 31/12/2017.
- C.03: The SOE is published in a peer-reviewed journal or conference proceeding.
- C.04: The SOE describes original research, excluding reviews, comments, patents and white papers.
- C.05: The SOE describes the application of NLP methods for the purpose of IE from free-text radiology reports.
- C.06: The described approach is LLM-based (defined as deep learning models with more than one million parameters, trained on unlabeled text data).

Screening and data extraction

Record screening was performed by two authors (KD, DR), using the online-platform Rayyan[106]. To improve alignment regarding inclusion criteria between reviewers, a first batch of 25 records was screened individually. Two conflicting decisions were discussed and clarified, leading to the consensus that BiLSTM-based architectures might also classify as LLMs and should therefore be included. In order to validate this change, a second batch of 25 records was screened and compared. Three conflicting decisions helped to clarify that, when a LLM-based architecture is not explicitly stated in the title or abstract, the record should still be marked as included to maximize overall recall of relevant papers.

Upon clarification of the inclusion criteria, each remaining record (title, abstract) was screened twice. After completion of the screening process, conflicts (comprising differing decisions or records marked as "maybe") were resolved by including all records that are marked at least once as "included".

After screening, records were sought for full-text retrieval. Data extraction was performed by one author (DR). During the extraction phase, reports were ex post excluded when a violation of inclusion criteria became apparent from the full-text. Reference lists of included papers were screened for further reports to include. Changes to the published protocol for this review are documented in Supplementary Table 4, including its description, reason, and date.

Data availability

The complete list of extracted documents for all queried databases as well as the completed data extraction table are available in the OSF repository, see https://doi.org/10.17605/OSF.IO/RWU5M.

Code availability

For data screening, the publicly available online platform rayyain.ai was used (free plan), see https://www.rayyan.ai.

Received: 21 February 2024; Accepted: 9 August 2024;
Published online: 24 August 2024

References

1. Müskens, J. L. J. M., Kool, R. B., Van Dulmen, S. A. & Westert, G. P. Overuse of diagnostic testing in healthcare: a systematic review. *BMJ Qual. Saf.* **31**, 54–63 (2022).
2. Nobel, J. M., Van Geel, K. & Robben, S. G. F. Structured reporting in radiology: a systematic review to explore its potential. *Eur. Radiol.* **32**, 2837–2854 (2022).
3. Khurana, D., Koli, A., Khatter, K. & Singh, S. Natural language processing: state of the art, current trends and challenges. *Multimed. Tools Appl.* **82**, 3713–3744 (2023).
4. Jurafsky, D. & Martin, J. H. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Pearson Education, 2024).
5. Birhane, A., Kasirzadeh, A., Leslie, D. & Wachter, S. Science in the age of large language models. *Nat. Rev. Phys.* **5**, 277–280 (2023).
6. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30 (Curran Associates, Inc., 2017).
7. Peters, M. E. et al. Deep contextualized word representations 1802. 05365 (2018).
8. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C. & Solorio, T. (eds.) In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).
9. Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, vol. 33, 1877–1901 (Curran Associates, Inc., 2020).
10. OpenAI et al. GPT-4 Technical Report 2303.08774. (2023).
11. Shoeybi, M. et al. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism 1909.08053 (2020).
12. Pons, E., Braun, L. M. M., Hunink, M. G. M. & Kors, J. A. Natural language processing in radiology: a systematic review. *Radiology* **279**, 329–343 (2016).
13. Casey, A. et al. A systematic review of natural language processing applied to radiology reports. *BMC Med. Inform. Decis. Mak.* **21**, 179 (2021).
14. Davidson, E. M. et al. The reporting quality of natural language processing studies: systematic review of studies of radiology reports. *BMC Med. Imaging* **21**, 142 (2021).
15. Saha, A., Burns, L. & Kulkarni, A. M. A scoping review of natural language processing of radiology reports in breast cancer. *Front. Oncol.* **13**, 1160167 (2023).
16. Gholipour, M., Khajouei, R., Amiri, P., Hajesmaeel Gohari, S. & Ahmadian, L. Extracting cancer concepts from clinical notes using natural language processing: a systematic review. *BMC Bioinform.* **24**, 405 (2023).
17. Gorenstein, L., Konen, E., Green, M. & Klang, E. Bidirectional encoder representations from transformers in radiology: a systematic review of natural language processing applications. *J. Am. Coll. Radiol.* **21**, 914–941 (2024).
18. Wood, D. A. et al. Automated labelling using an attention model for radiology reports of MRI scans (ALARM). In Arbel, T. et al. (eds.) *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, vol. 121 of *Proceedings of Machine Learning Research*, 811–826 (PMLR, 2020-07-06/2020-07-08).
19. Wood, D. A. et al. Deep learning to automate the labelling of head MRI datasets for computer vision applications. *Eur. Radiol.* **32**, 725–736 (2022).
20. Li, Z. & Ren, J. Fine-tuning ERNIE for chest abnormal imaging signs extraction. *J. Biomed. Inform.* **108**, 103492 (2020).
21. Lybarger, K., Damani, A., Gunn, M., Uzuner, O. Z. & Yetisgen, M. Extracting radiological findings with normalized anatomical information using a span-based BERT relation extraction model. *AMIA Jt. Summits Transl. Sci. Proc.* **2022**, 339–348 (2022).
22. Kuling, G., Curpen, B. & Martel, A. L. BI-RADS BERT and using section segmentation to understand radiology reports. *J. Imaging* **8**, 131 (2022).
23. Lau, W., Lybarger, K., Gunn, M. L. & Yetisgen, M. Event-based clinical finding extraction from radiology reports with pre-trained language model. *J. Digit. Imaging* **36**, 91–104 (2023).
24. Sugimoto, K. et al. End-to-end approach for structuring radiology reports. *Stud. Health Technol. Inform.* **270**, 203–207 (2020).
25. Zhang, Y. et al. Using recurrent neural networks to extract high-quality information from lung cancer screening computerized tomography reports for inter-radiologist audit and feedback quality improvement. *JCO Clin. Cancer Inform.* **7**, e2200153 (2023).
26. Tejani, A. S. et al. Performance of multiple pretrained BERT models to automate and accelerate data annotation for large datasets. *Radiol. Artif. Intell.* **4**, e220007 (2022).
27. Zaman, S. et al. Automatic diagnosis labeling of cardiovascular MRI by using semisupervised natural language processing of text reports. *Radiol. Artif. Intell.* **4**, e210085 (2022).
28. Liu, H. et al. Use of BERT (bidirectional encoder representations from transformers)-based deep learning method for extracting evidences in chinese radiology reports: Development of a computer-aided liver cancer diagnosis framework. *J. Med. Internet Res.* **23**, e19689 (2021).
29. Jaiswal, A. et al. RadBERT-CL: factually-aware contrastive learning for radiology report classification. In *Proc. Machine Learning for Health*, 196–208 (PMLR, 2021).
30. Torres-Lopez, V. M. et al. Development and validation of a model to identify critical brain injuries using natural language processing of

text computed tomography reports. *JAMA Netw. Open* **5**, e2227109 (2022).
31. Pérez-Díez, I., Pérez-Moraga, R., López-Cerdán, A., Salinas-Serrano, J. M. & la Iglesia-Vayá, M. De-identifying Spanish medical texts - named entity recognition applied to radiology reports. *J. Biomed. Semant.* **12**, 6 (2021).
32. Lau, W., Payne, T. H., Uzuner, O. & Yetisgen, M. Extraction and analysis of clinically important follow-up recommendations in a large radiology dataset. *AMIA Summits Transl. Sci. Proc.* **2020**, 335–344 (2020).
33. Santos, T. et al. A fusion NLP model for the inference of standardized thyroid nodule malignancy scores from radiology report text. *Annu. Symp. Proc. AMIA Symp.* **2021**, 1079–1088 (2021).
34. Fink, M. A. et al. Deep learning–based assessment of oncologic outcomes from natural language processing of structured radiology reports. *Radiol. Artif. Intell.* **4**, e220055 (2022).
35. Datta, S. et al. Understanding spatial language in radiology: representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning. *J. Biomed. Inform.* **108**, 103473 (2020).
36. Datta, S. & Roberts, K. Spatial relation extraction from radiology reports using syntax-aware word representations. *AMIA Jt. Summits Transl. Sci. Proc.* **2020**, 116–125 (2020).
37. Datta, S. & Roberts, K. A Hybrid deep learning approach for spatial trigger extraction from radiology reports. In *Proc. Third International Workshop on Spatial Language Understanding*, 50–55 (Association for Computational Linguistics, Online, 2020).
38. Zhang, H. et al. A novel deep learning approach to extract Chinese clinical entities for lung cancer screening and staging. *BMC Med. Inform. Decis. Mak.* **21**, 214 (2021).
39. Hu, D. et al. Automatic extraction of lung cancer staging information from computed tomography reports: Deep learning approach. *JMIR Med. Inform.* **9**, e27955 (2021).
40. Datta, S., Khanpara, S., Riascos, R. F. & Roberts, K. Leveraging spatial information in radiology reports for ischemic stroke phenotyping. *AMIA Jt. Summits Transl. Sci. Proc.* **2021**, 170–179 (2021).
41. Dada, A. et al. Information extraction from weakly structured radiological reports with natural language queries. *Eur. Radiol.* **34**, 330–337 (2023).
42. Eisenhauer, E. et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
43. Rosen, R. D. & Sapra, A. TNM Classification. In *StatPearls* (StatPearls Publishing, 2023).
44. University of California Berkeley. HIPAA PHI: definition of PHI and List of 18 Identifiers. https://cphs.berkeley.edu/hipaa/hipaa18.html# (2023).
45. Stanford NLP Group. Stanfordnlp/stanza. Stanford NLP (2024).
46. Sugimoto, K. et al. Extracting clinical terms from radiology reports with deep learning. *J. Biomed. Inform.* **116**, 103729 (2021).
47. US National Institutes of Health. NationalCancer Institute. NCI Thesaurus. https://ncit.nci.nih.gov/ncitbrowser/.
48. Datta, S., Godfrey-Stovall, J. & Roberts, K. RadLex normalization in radiology reports. *AMIA Annu. Symp. Proc.* **2020**, 338–347 (2021).
49. Zhang, Z. et al. ERNIE: Enhanced Language Representation with Informative Entities In *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics (2019).
50. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space 1301.3781 (2013).
51. Huang, X., Chen, H. & Yan, J. D. Study on structured method of Chinese MRI report of nasopharyngeal carcinoma. *BMC Med. Inform. Decis. Mak.* **21**, 203 (2021).
52. DocCheck. DocCheck Flexicon. https://flexikon.doccheck.com/de/Hauptseite (2024).
53. Jantscher, M. et al. Information extraction from German radiological reports for general clinical text and language understanding. *Sci. Rep.* **13**, 2353 (2023).
54. Zhang, X. et al. Extracting comprehensive clinical information for breast cancer using deep learning methods. *Int. J. Med. Inform.* **132**, 103985 (2019).
55. Johnson, A. E. W. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
56. Moody, G. B. & Mark, R. G. The MIMIC Database (1992).
57. Datta, S. & Roberts, K. Weakly supervised spatial relation extraction from radiology reports. *JAMIA Open* **6**, ooad027 (2023).
58. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
59. Datta, S. & Roberts, K. Fine-grained spatial information extraction in radiology as two-turn question answering. *Int. J. Med. Inform.* **158**, 104628 (2022).
60. Datta, S. et al. Rad-SpatialNet: a frame-based resource for fine-grained spatial relations in radiology reports. In Calzolari, N. *et al*. (eds.) *Proc. Twelfth Language Resources and Evaluation Conference*, 2251–2260 (European Language Resources Association, Marseille, France, 2020).
61. Demner-Fushman, D. et al. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **23**, 304–310 (2016).
62. Mithun, S. et al. Clinical concept-based radiology reports classification pipeline for lung carcinoma. *J. Digit. Imaging* **36**, 812–826 (2023).
63. Bressem, K. K. et al. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics* **36**, 5255–5261 (2021).
64. Singh, V. et al. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Sci. Rep.* **11**, 14490 (2021).
65. Demler, O. V., Pencina, M. J. & D'Agostino, R. B. Misuse of DeLong test to compare AUCs for nested models. *Stat. Med.* **31**, 2577–2587 (2012).
66. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training (2018).
67. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
68. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024).
69. Zhang, Y. & Xu, Z. BERT for question answering on SQuAD 2.0 (2019).
70. OECD. Diagnostic technologies (2023).
71. Viry, A. et al. Annual exposure of the Swiss population from medical imaging in 2018. *Radiat. Prot. Dosim.* **195**, 289–295 (2021).
72. Reichenpfader, D., Müller, H. & Denecke, K. Large language model-based information extraction from free-text radiology reports: a scoping review protocol. *BMJ Open* **13**, e076865 (2023).
73. Shanahan, M., McDonell, K. & Reynolds, L. Role play with large language models. *Nature* **623**, 493–498 (2023).
74. Liang, S. et al. Fine-tuning BERT Models for Summarizing German Radiology Findings. In Naumann, T., Bethard, S., Roberts, K. & Rumshisky, A. (eds.) *Proc. 4th Clinical Natural Language Processing Workshop*, 30–40 (Association for Computational Linguistics, Seattle, WA, 2022).
75. Adams, L. C. et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* **307**, e230725 (2023).

76. Nowak, S. et al. Transformer-based structuring of free-text radiology report databases. *Eur. Radiol.* **33**, 4228–4236 (2023).
77. Košprdić, M., Prodanović, N., Ljajić, A., Bašaragin, B. & Milošević, N. From zero to hero: harnessing transformers for biomedical named entity recognition in zero- and few-shot contexts 2305.04928 (2023).
78. Smit, A. et al. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In Webber, B., Cohn, T., He, Y. & Liu, Y. (eds.) *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1500–1519 (Association for Computational Linguistics, Online, 2020).
79. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large language models are few-shot clinical information extractors. In Goldberg, Y., Kozareva, Z. & Zhang, Y. (eds.) *Proc. Conference on Empirical Methods in Natural Language Processing*, 1998–2022 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).
80. Kartchner, D., Ramalingam, S., Al-Hussaini, I., Kronick, O. & Mitchell, C. Zero-shot information extraction for clinical meta-analysis using large language models. In Demner-fushman, D., Ananiadou, S. & Cohen, K. (eds.) *Proc. 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, 396–405 (Association for Computational Linguistics, Toronto, Canada, 2023).
81. Jupin-Delevaux, É. et al. BERT-based natural language processing analysis of French CT reports: application to the measurement of the positivity rate for pulmonary embolism. *Res. Diagn. Interv. Imaging* **6**, 100027 (2023).
82. Rohanian, O., Nouriborji, M., Kouchaki, S. & Clifton, D. A. On the effectiveness of compact biomedical transformers. *Bioinformatics* **39**, btad103 (2023).
83. Gemini Team, Google. Gemini: a family of highly capable multimodal models. https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf (2024).
84. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer 1910.10683 (2023).
85. Llama-3. Meta (2024).
86. Jiang, A. Q. et al. Mixtral of experts 2401.04088 (2024).
87. Chen, Z. et al. MEDITRON-70B: scaling medical pretraining for large language models 2311.16079 (2023).
88. Singhal, K. et al. Towards expert-level medical question answering with large language models 2305.09617 (2023).
89. Saab, K. et al. Capabilities of Gemini models in medicine 2404.18416 (2024).
90. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
91. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
92. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
93. Dagdelen, J. et al. Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
94. Zheng, L. et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Adv. Neural Inf. Process Syst.* **36**, 46595–46623 (2023).
95. Mukherjee, P., Hou, B., Lanfredi, R. B. & Summers, R. M. Feasibility of using the privacy-preserving large language model Vicuna for labeling radiology reports. *Radiology* **309**, e231147 (2023).
96. Bressem, K. K. et al. MEDBERT.de: a comprehensive German BERT model for the medical domain. *Expert Syst. Appl.* **237**, 121598 (2024).
97. Hu, D., Liu, B., Zhu, X., Lu, X. & Wu, N. Zero-shot information extraction from radiological reports using ChatGPT. *Int. J. Med. Inform.* **183**, 105321 (2024).
98. Hu, D., Li, S., Zhang, H., Wu, N. & Lu, X. Using natural language processing and machine learning to preoperatively predict lymph node metastasis for non–small cell lung cancer with electronic medical records: development and validation study. *JMIR Med. Inform.* **10**, e35475 (2022).
99. Mallio, C. A., Sertorio, A. C., Bernetti, C. & Beomonte Zobel, B. Large language models for structured reporting in radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing. *La Radiol. Med.* **128**, 808–812 (2023).
100. Zhao, H. et al. Explainability for large language models: a survey. *ACM Trans. Intell. Syst. Technol.* **15**, 1–38 (2024).
101. Yang, H. et al. Unveiling the generalization power of fine-tuned large language models. In *Proc. of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers) (eds Duh, K., Gomez, H. & Bethard, S.) 884–899 (Association for Computational Linguistics, Mexico City, Mexico, 2024). https://doi.org/10.18653/v1/2024.naacl-long.51.
102. Abdin, M. et al. Phi-3 technical report: a highly capable language model locally on your phone 2404.14219 (2024).
103. Gilbert, S., Kather, J. N. & Hogan, A. Augmented non-hallucinating large language models as medical information curators. *npj Digital Med.* **7**, 1–5 (2024).
104. He, P., Liu, X., Gao, J. & Chen, W. DeBERTa: decoding-enhanced BERT with disentangled attention 2006.03654 (2021).
105. Kakarmath, S. et al. Best practices for authors of healthcare-related artificial intelligence manuscripts. *NPJ Digit. Med.* **3**, 134 (2020).
106. Rayyan - AI Powered Tool for Systematic Literature Reviews (2021).
107. Si, Y., Wang, J., Xu, H. & Roberts, K. Enhancing clinical concept extraction with contextual embeddings. *J. Am. Med. Inform. Assoc.* **26**, 1297–1304 (2019).
108. Liu, Y. et al. RoBERTa: a robustly optimized BERT pretraining approach 1907.11692 (2019).
109. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
110. Alsentzer, E. et al. Publicly Available Clinical BERT Embeddings. In Rumshisky, A., Roberts, K., Bethard, S. & Naumann, T. (eds.) *Proc. 2nd Clinical Natural Language Processing Workshop*, 72–78 (Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019).
111. Deepset. German BERT. https://huggingface.co/bert-base-german-cased (2019).
112. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 2:1–2:23 (2021).
113. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter 1910.01108 (2020).
114. Cui, Y., Che, W., Liu, T., Qin, B. & Yang, Z. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **29**, 3504–3514 (2021).
115. Peng, Y., Yan, S. & Lu, Z. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proc. of the 18th BioNLP Workshop and Shared Task* (eds Demner-Fushman, D., Cohen, K. B., Ananiadou, S. & Tsujii, J.) 58–65 (Association for Computational Linguistics, Florence, Italy, 2019). https://doi.org/10.18653/v1/W19-5006.
116. Chan, B., Schweter, S. & Möller, T. German's next language model. In *Proc. of the 28th International Conference on Computational Linguistics* (eds Scott, D., Bel, N. & Zong, C.) 6788–6796 (International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020). https://doi.org/10.18653/v1/2020.coling-main.598.
117. Shrestha, M. *Development of a Language Model for the Medical Domain*. Ph.D. thesis (Rhine-Waal University of Applied Sciences, 2021).

118. The MultiBERTs: BERT reproductions for robustness analysis. In Sellam, T. et al. (eds.) *ICLR 2022* (2022).
119. Wu, S. & He, Y. Enriching pre-trained language model with entity information for relation classification. In *Proc. of the 28th ACM International Conference on Information and Knowledge Management*, 2361–2364 (Association for Computing Machinery, New York, NY, USA, 2019). https://doi.org/10.1145/3357384.3358119.
120. Beltagy, I., Lo, K. & Cohan, A. SciBERT: a pretrained language model for scientific text. In *Proc. Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (eds Inui, K., Jiang, J., Ng, V. & Wan, X.), 3615–3620 (Association for Computational Linguistics, Hong Kong, China, 2019).
121. Eberts, M. & Ulges, A. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020*, 2006–2013 (IOS Press, 2020).
122. Yang, Z. et al. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).

Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. We thank Cornelia Zelger for her support during the search query definition process.

Author contributions

D.R. conceptualized the study, defined the methodology (incl. the search strategy), performed the database searches and managed the screening process. D.R. also performed data extraction and authored the original draft. K.D. focused on reviewing and editing the manuscript. K.D. also participated in the screening process. H.M. provided supervision and contributed to writing review.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01219-0.

Correspondence and requests for materials should be addressed to Daniel Reichenpfader.