A Survey of Evaluation Methods of Generated Medical Textual Reports

Yongxin Zhou **Fabien Ringeval** **François Portet**
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000, Grenoble, France
yongxin.zhou@univ-grenoble-alpes.fr,
{fabien.ringeval, Francois.Portet}@imag.fr

Abstract

Medical Report Generation (MRG) is a subtask of Natural Language Generation (NLG) and aims to present information from various sources in textual form and synthesize salient information, with the goal of reducing the time spent by domain experts in writing medical reports and providing support information for decision-making. Given the specificity of the medical domain, the evaluation of automatically generated medical reports is of paramount importance to the validity of these systems. Therefore, in this paper, we focus on the evaluation of automatically generated medical reports from the perspective of automatic and human evaluation. We present evaluation methods for general NLG evaluation and how they have been applied to domain-specific medical tasks. The study shows that MRG evaluation methods are very diverse, and that further work is needed to build shared evaluation methods. The state of the art also emphasizes that such an evaluation must be task specific and include human assessments, requesting the participation of experts in the field.

1 Introduction

Medical Report Generation (MRG)[1] (Chen et al., 2022) is a subfield of Natural Language Generation (NLG) and aims to present information from various sources in textual form to synthesize salient information, with the goal of reducing the time spent by domain experts in writing medical reports and providing supporting information for decision-making.

MRG includes all systems used to generate medical documentation, such as the generation of radiology reports (Chen et al., 2021), discharge summaries or SOAP notes (Krishna et al., 2021), etc.

The evaluation of automatically generated texts is important for the validity of the systems, especially in the era of ChatGPT and its increased usage in medical domain (Ma et al., 2023). However, it has often been reported that the two main approaches of NLG evaluation – human-based and automatic-based – need to be improved (Reiter and Belz, 2009; van der Lee et al., 2019). On the one hand, the use of automatic metrics for system quality assessment has been criticized for two main points: they are uninterpretable and do not correlate well with human evaluations (Qader et al., 2018; van der Lee et al., 2019). On the other hand, the deployment of human evaluation is sometimes too complex. Indeed, crowdsourcing solutions are not always reliable or adequate while using experts, when available, can lead to high costs. Furthermore, there is a lack of unified framework/criteria (van der Lee et al., 2019).

Although there have been many surveys on NLG evaluation (Gkatzia and Mahamood, 2015; Amidei et al., 2018; van der Lee et al., 2019; Sai et al., 2022), there is no systematic study on report generation evaluation in the medical domain. It is worth mentioning Messina et al. (2021), which has done survey work in the area of automatic report generation from medical images, including an analysis of evaluation methods.

In this paper, we focus on MRG tasks and their evaluations. We include more than 20 papers in this study, classified into two broad categories: text-to-text and data-to-text. We summarize the evaluation methods currently in use and make recommendations for future evaluation of MRG systems.

2 Medical Report Generation

2.1 Paper Search and Selection

To perform a literature review of MRG evaluation, we first followed the paper list introduced in a survey paper on dialogue summarization (Feng et al., 2022). We then extended our search to search engine such as Google Scholar. Papers reviewed in this study were primarily from the major NLP con-

[1] It is also called ‘Clinical Note Generation’.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 447–459
July 14, 2023 ©2023 Association for Computational Linguistics

ferences, including ACL, EMNLP, NAACL, etc. In addition, we also included some articles from journal and domain-specific workshops, including NLPMC, ClinicalNLP, etc. [2]

2.2 MRG Systems and Applications

Since the seminal work of Kukich (1983), there have been several kinds of medical report systems and applications, such as the generation of psychiatric case notes (Kazi and Kahanda, 2019), the generation of consultation notes from transcripts (Papadopoulos Korfiatis et al., 2022), the generation of radiology reports (Chen et al., 2021), nurse-patient summaries (Liu et al., 2019), counseling (conversation) summarization (Srivastava et al., 2022), discharge summaries or clinical notes (Krishna et al., 2021), and even data augmentation for other medical tasks (Kocabiyikoglu et al., 2021).

Joshi et al. (2020) provided a general definition of a medical report in the case of medical dialogue summarization: "*the medical report captures and summarizes the important parts of the medical conversation necessary for clinical decision-making and subsequent follow-ups.*"

Despite the diversity of their tasks, structures and audiences, the main characteristics of MRG remain similar, namely the use of the documentation and the subsequent use of the diagnosis, which can also be used for administration and by institutions, subsequently referenced by clinicians and retained by patients. The main objectives of such systems in clinical practice are to reduce the time spent by clinicians on manual writing and facilitate medical decision-making.

2.3 Main NLG approaches to MRG

According to the different types of input sources, MRG can be divided into two categories: 1) Text-to-text, e.g. summarizing medical dialogues; 2) Data-to-text, e.g. automatically generating reports from medical images or other data. After classifying the different approaches according to the input, we then further categorized different works in the literature according to their tasks, with examples in Table 1.

2.3.1 Text-to-text

There are three main tasks in the Text-to-Text category: 1) summarizing medical dialogues/conversations, including spoken conversations and online medical conversations; 2) summarizing hospital stays/hospitalizations; and 3) summarizing medical reports, where the original reports may come from different domains, such as radiology reports or general clinical reports.

The most common work in text-based MRG is that of **summarizing medical conversations**, where the input source can be either transcripts of clinician-patient spoken conversations (Kazi and Kahanda, 2019; Enarvi et al., 2020; Liu et al., 2019; Krishna et al., 2021; Zhang et al., 2021; Molenaar et al., 2020; Srivastava et al., 2022; Lacson et al., 2006; Moramarco et al., 2022; Yim and Yetisgen, 2021); or online medical conversations (Chintagunta et al., 2021; Nair et al., 2021; Joshi et al., 2020; Song et al., 2020; Chen et al., 2022).

Regarding the **summarization of hospital stays**, some work (Di Eugenio et al., 2014; Acharya et al., 2016) used both physician discharge notes (free text) and the structured nursing documentation (such as nursing plans of care) to generate a unique summary. Others generated summaries from long-form hospital admissions (Adams et al., 2023).

Moreover, work has also been carried out on **summarizing medical reports**. For example, Moramarco et al. (2021) used the *MTSamples* dataset to fill automatically the '*description field*' of a medical report based on the information present in the overall report. In addition, radiology report summarization (Zhang et al., 2020; Karn et al., 2022) is intended to produce a concise and easily comprehensible '*IMPRESSIONS*' section from the rest of the radiology report. The '*IMPRESSIONS*' section of a radiology report is considered a summary of the radiologist's reasoning and conclusions, which helps the referring physician confirm or exclude certain diagnoses (Karn et al., 2022).

2.3.2 Data-to-text

As presented in Table 1, there are different tasks in the Data-to-Text category: 1) generation of reports from medical images, such as radiology images, brain image data.; 2) generation of text summaries from intensive care data; and 3) generation of medical reports from multimodal inputs; 4) other applications such as the generation of tailored smoking cessation letters based on responses to a smoking questionnaire (Reiter et al., 2003).

In order to meet the growing demand of image-based diagnosis from patients using artificial in-

[2] The list of papers about MRG is available at https://github.com/yongxin2020/Medical-Report-Generation-Papers

Category	Task	Description	Examples
Text-to-text	Medical dialogue/conversation summarization	Transcribed conversations	Srivastava et al. (2022); Molenaar et al. (2020); Zhang et al. (2021); Krishna et al. (2021); Liu et al. (2019); Enarvi et al. (2020); Kazi and Kahanda (2019); Lacson et al. (2006); Moramarco et al. (2022); Yim and Yetisgen (2021)
		Online medical conversations	Nair et al. (2021); Chintagunta et al. (2021); Song et al. (2020); Joshi et al. (2020); Chen et al. (2022)
	Summarization of hospital stays/hospitalizations	Sources from physician discharge notes and nursing plans of care	Acharya et al. (2016); Di Eugenio et al. (2014)
		Long-form hospital admissions	Adams et al. (2023)
	Medical report summarization	Radiology report	Zhang et al. (2020); Karn et al. (2022)
		Clinical reports	Moramarco et al. (2021)
Data-to-text	Report Generation from medical images	Radiology	Miura et al. (2021); Chen et al. (2021); Yan et al. (2021); Chen et al. (2020); Lovelace and Mortazavi (2020); Nooralahzadeh et al. (2021); Qin and Song (2022)
		Brain imaging data	Jordan et al. (2014)
	Clinical Data Summarization	Intensive care data	Portet et al. (2009); Reiter et al. (2008)
	Automated Medical Reporting	Sources from multimodal inputs: audio, video and sensor data from medical consultations	Maas et al. (2021)
	Other applications	Generation of tailored smoking cessation letters	Reiter et al. (2003)

Table 1: Categorization of MRG according to system inputs and tasks.

telligence and applying image captioning to the medical field, radiology report generation is the subject of continuous work and growing interest from researchers, which aims to describe radiology images with professional quality reports (Chen et al., 2020, 2021; Lovelace and Mortazavi, 2020; Nooralahzadeh et al., 2021; Yan et al., 2021; Qin and Song, 2022; Miura et al., 2021; Liu et al., 2021). Such research has also been applied to generation of clinician reports from brain imaging data (Jordan et al., 2014).

Besides images, the summarization of physiological data as also been the subject of research. In the *BabyTalk* project, Portet et al. (2009) presented a prototype that generates textual summaries of about 45 minutes of continuous physiological signals and discrete events. Their evaluation with physicians showed that text summaries could be an effective decision-support aids for clinicians.

To cope with the high workload due to the time required for proper documentation, Maas et al. (2021) presented a real-time automated report of the interaction between care provider and patient, taking multimodal inputs that include audio, video, and sensor data from medical consultations, and introducing knowledge graphs – the Patient Medical Graph. They used speech and action recognition technology to first transform multimodal inputs into text before formally representing them and generating reports.

3 Evaluation in NLG

In this section, we will briefly introduce automatic evaluation in NLG and then will look at human evaluation with some current practices.

3.1 Automatic Evaluation

Automatic evaluation is popular because it is cheap and fast and it is widely used in benchmarking activity and for system development.

There is a wide range of automatic evaluation metrics used in NLG (Sai et al., 2022) and we will restrict to the two most popular: 1) the corpus-based metrics and 2) the trainable metrics. The corpus-based metrics rely on a set of reference texts (i.e. gold standard outputs) to which system outputs are compared. For instance, it can be based on n-grams, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004); or on edit distance: WER (Woodard and Nelson, 1982), TER (Snover et al., 2006), etc.

Most automatic metrics require gold references, but these are not always available. Reference-less metrics, where neural models are trained to predict human ratings from texts (e.g. regression models trained on ratings data), are getting more and more attention recently. For example, BLEURT (Sellam et al., 2020) is a learned evaluation metric for English to predict human judgments. It relies on the BERT model using unsupervised techniques with millions of synthetic examples. There are also metrics based on question-answer pairs on a given source document (Scialom et al., 2021; Rebuffel et al., 2021), for example QuestEval (Scialom et al., 2021) uses pre-trained models to assess if two different inputs contain the same information. Note that QuestEval can also be used with references.

To summarize, ROUGE scores assess the similarity between candidates and references based on the overlap of unigrams, bigrams, and the longest common sequence, likewise for BLEU; while BLEU focuses on precision, ROUGE focuses on recall. BERTScore evaluates the similarity between candidates and references at token level, using contextual embeddings from BERT, while QuestEval assesses whether a summary contains all the relevant information from its source document and BLEURT attempts to model human judgments.

However, automatic evaluation metrics have their limitations and do not sufficiently reflect human judgments of system performance (Novikova et al., 2017).

3.2 Human Evaluation

Human evaluation is considered the most informative form of evaluation of NLG systems, but it can be expensive and time-consuming since qualified human evaluators have to be recruited. Hence, human evaluation is difficult to scale up unless using crowd sourcing approaches but these are difficult to apply in medicine for expertise and privacy reasons.

There are several commonly used methods for human evaluation, including the Likert scale scoring and pairwise comparison for general text generation, as well as Pyramid and binary factuality evaluation specifically designed for summarization (Gao et al., 2023). Some other methods consist in evaluating how much information can be extracted back from the text in a formal form (A. Baez Miranda et al., 2015).

It has been argued that human evaluation approaches are difficult to compare (van der Lee et al., 2019; Belz et al., 2020) since different tasks and criteria are used (with different names). Furthermore, only a small number of papers provide full details of human evaluation experiments (Belz et al., 2020). Howcroft et al. (2020) concluded that due to a pervasive lack of clarity in reports and extreme diversity in approaches, human evaluation in NLG presents as extremely confused in 2020, and that the field is in urgent need of standard methods and terminology.

In addition, van der Lee et al. (2019) provided an overview of best practices in human evaluation of automatically generated text based on papers published at INLG (N=51) and ACL (N=38) in 2018, and released a list of best practices on 7 different topics: general, criteria, sampling, annotation, measurement, design and statistics.

4 Evaluation for Text-based Medical Report Generation Systems

In the following subsections, we summarize the automatic measures and human evaluation used in the literature in Table 2.

4.1 Automatic Metrics

We divide automatic metrics into two categories: text quality and medical concept correctness, for medical correctness there are two subcategories: those based on reports and those for auxiliary or intermediate tasks.

4.1.1 Text Quality

For automatic text quality assessment, there are word-overlap-based metrics like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), embedding-based metrics such as BERTScore (Zhang* et al., 2020); learned evaluation metrics like BLEURT (Sellam et al., 2020); and evaluation metrics which rely on question answering like QuestEval (Scialom et al., 2021).

ROUGE (Lin, 2004) has been widely used in MRG tasks: some papers reported only ROUGE-L F1 score (Joshi et al., 2020; Enarvi et al., 2020; Nair et al., 2021), while some others reported ROUGE-1, ROUGE-2, and ROUGE-L scores (Song et al., 2020; Yim and Yetisgen, 2021). Yim and Yetisgen (2021) reported BLEU (Papineni et al., 2002) in addition to ROUGE performances across different note sections.

Category	Subcategory	Metric or evaluation		Used by papers
Automated Metrics	Text quality	ROUGE (-1, -2, -L)		Srivastava et al. (2022); Zhang et al. (2021); Chintagunta et al. (2021); Krishna et al. (2021); Song et al. (2020); Joshi et al. (2020); Enarvi et al. (2020); Liu et al. (2019); Nair et al. (2021); Yim and Yetisgen (2021); Chen et al. (2022); Ben Abacha et al. (2023)
		BLEU		Yim and Yetisgen (2021)
		BERTScore		Ben Abacha et al. (2023)
		QuestEval (QAE) score		Srivastava et al. (2022)
		Bleurt Score (BS)		Srivastava et al. (2022); Ben Abacha et al. (2023)
	Medical correctness (report based)	Medical Concept Coverage	-	Joshi et al. (2020); Chintagunta et al. (2021); Nair et al. (2021); Chen et al. (2022)
			Factual correctness (F1)	Enarvi et al. (2020)
			Concept-based (F1/R/P)	Zhang et al. (2021)
			Items included (P, R, false positives)	Molenaar et al. (2020)
			Fact-based (Fact-Core + Fact-Full)	Ben Abacha et al. (2023)
		Negation Correctness	Negex (Harkema et al., 2009) is used to determine negated concepts	Joshi et al. (2020); Chintagunta et al. (2021); Nair et al. (2021)
		Disease diagnosis	Regex-based Diagnostic Accuracy (RD-Acc)	Chen et al. (2022)
	Medical correctness (auxiliary tasks)	Classification of electronic health record (EHR) categories	AUROC (Area Under the ROC Curve) scale	Kazi and Kahanda (2019)
		Utterances classification	Multilabel classification of noteworthy utterances (Accuracy, Ma-AUC, Ma-F1, Mi-AUC, Mi-F1)	Krishna et al. (2021)
			Utterance tags classification (PD/DT/OT labels) (precision, recall, and F scores)	Song et al. (2020)
			Dialogue turns classification (precision, recall, and F measure)	Lacson et al. (2006)
Human Evaluation	Intrinsic (Text quality)	Relevance		Srivastava et al. (2022); Zhang et al. (2021)
		Consistency		Srivastava et al. (2022)
		Fluency		Srivastava et al. (2022); Zhang et al. (2021); Ben Abacha et al. (2023)
		Coherence		Srivastava et al. (2022)
		Missing		Zhang et al. (2021)
		Hallucination		Zhang et al. (2021)
		Repetition / Non-redundancy		Zhang et al. (2021); Ben Abacha et al. (2023)
		Contradiction		Zhang et al. (2021)
		Extent of verbatim copying from conversation		Krishna et al. (2021)
		Comprehensiveness		Krishna et al. (2021)
		Sentence-level (factually correct, incoherent, irrelevant, redundant, or placed under an inappropriate section)		Krishna et al. (2021)
		Categories relevancy, factual accuracy, writing-style, completeness, and overall		Yim and Yetisgen (2021)
	Intrinsic (Medical Correctness)	Factually correct and medically relevant information		Joshi et al. (2020); Chintagunta et al. (2021)
		Critical Omissions, Hallucinations, Correct Facts, Incorrect Facts based on fact extraction		Ben Abacha et al. (2023)
	Extrinsic	Clinical acceptability framework (Sekhon et al., 2017)		Srivastava et al. (2022)
		List of key questions based on topics that commonly arise between hemodialysis patients and caregivers		Lacson et al. (2006)
		Post-editing (Post-edit times, errors into "incorrect statements" and "omissions")		Moramarco et al. (2022)

Table 2: Summary of evaluation methods used in the articles reviewed.

Commonly used automated metrics, such as ROUGE and BLEU, have their limitations and are known to correlate poorly with human evaluations (van der Lee et al., 2019). Therefore, other measures such as QuestEval (Scialom et al., 2021) and BLEURT (Sellam et al., 2020) which can correlate better with human judgements are used, Srivastava et al. (2022) used these two scores in addition to ROUGE. In addition to ROUGE and BLEURT, Ben Abacha et al. (2023) also reported BERTScore.

4.1.2 Medical Correctness: Report-based

In the study by Joshi et al. (2020) two measures are defined: *Medical Concept Coverage* and *Negation Correctness*. The former captures the coverage of medical terms in the predicted summaries to the gold standard reference, while the latter identifies the negated status of medical concepts. In the healthcare domain, it is crucial to ensure high-quality results in terms of accurate usage of medical terms and capturing negation.

The evaluation of *Concept* involves using specific and in-house extractors and Named Entity Recognition (NER) models. They refer to domain-specific knowledge and compare the match of extracted concepts to standardized health and biomedical vocabularies, such as the Unified Medical Language Systems (UMLS). Several studies have utilized concept correctness measures, such as F1-score, precision, recall, and false positives, at various levels of granularity, including the report level and section level.

For instance, Joshi et al. (2020) used an in-house medical entity extractor to match concepts in the summary to UMLS, and they used Negex (Harkema et al., 2009) to determine negated concepts. Medical concepts in the predicted summary that were not present in the original conversation would be false positives, and vice versa for false negatives. Among the concepts present in the predicted summary, they assessed precision and recall to see whether the predicted negation was accurate for the decoded concepts and computed a Negation F1. The set of automatic metrics proposed was then used in several works (Chintagunta et al., 2021; Nair et al., 2021).

If in-house entity extractor to match concepts in the summary to UMLS have been frequent (Soldaini and Goharian, 2016; Joshi et al., 2020; Zhang et al., 2021), entity extraction using machine learning has appeared recently, which is even more specific to the task. For instance, Enarvi et al. (2020) employed a machine learning-based clinical fact extractor to measure factual correctness by extracting medical facts from both the predicted reports and the ground-truth reports, such as conditions and medications, as well as their attributes such as body part, severity, or dosage, then calculating the F1 score from these two sets. To compute Concept-F1, Chen et al. (2022) used the medical entity extractor – BERT-CRF (Devlin et al., 2019) trained on their NER task to match entities in the predicted summary to the reference summary.

Similarly, Ben Abacha et al. (2023) used "fact-based metrics (Fact Scores)", which is a machine learning-based medical fact extraction system. The Fact Score metric measures the F1-score of medically relevant facts extraction, is used to assess the factual consistency of the generated summaries. The Fact-based metrics consist of two variants: Fact-Core, which relies on the extraction of seven core fact attributes, and Fact-Full, which combines these core facts and five additional attributes.

In addition, there is also work combining the two approaches to extract concepts: Zhang et al. (2021) extracted medical relevant concepts via one of two systems: their in-house rule-based system and quickUMLS (Soldaini and Goharian, 2016) – a Python implementation of UMLS. Their rule-based system was found to be effective in capturing symptom-related findings in clinical reports, and quickUMLS is capable of extracting a wide scope of medical findings such as symptoms, diseases, medication and procedures.

Moreover, Molenaar et al. (2020) measured the quality of the dialogue summarization pipeline for healthcare reporting by establishing the number of items included in the generated and gold standards, using precision, recall and false positives (FPs) as metrics. They followed the SOEP/SOAP format – Subjective (S), Objective (O), Evaluation (E) / Assessment (A) and Plan (P) – commonly used by general practitioners in the Netherlands. It appears that they manually calculated the number of items included in each section of the SOEP format for the eight reports generated.

However, concept-based evaluation can have its own limitations, particularly with regard to false positives errors, Zhang et al. (2021) employed filtering rules to attempt to mitigate this issue.

Additionally, Chen et al. (2022) reported Regex-based Diagnostic Accuracy (RD-Acc), which measures the model's ability to diagnose the disease. Their reference reports written by annotators contain six parts, RD-Acc is calculated using the regex-based approach based on the *diagnosis* part. They calculated for what percentage of the generated reports, the content of their *diagnosis* part contains the actual disease text or key concepts.

4.1.3 Medical Correctness: Auxiliary or Intermediate Tasks

Another subcategory of automatic measures of concept correctness is those that evaluate auxiliary or intermediate tasks, there are two types: classification of electronic health record (EHR) categories and utterances classification.

To generate case notes from digital transcripts of doctor-patient conversations, Kazi and Kahanda (2019) divided the task into two subtasks: (1) predict semantic topics for segments of the transcripts (EHR categories) and then (2) generate a more formal version of the text that goes into the corresponding section of the EHR form. They used the AUROC (Area Under the ROC Curve) scale (Bewick et al., 2005) to assess their first task of predicting EHR categories, which could be any of the following: Client Details, Chief Complaint, Family History, Social History, Medical History and Others. Correct prediction of EHR categories could be useful for subsequent formal text generation.

For utterances classification, there are different types of classification such as classifying noteworthy utterance (Krishna et al., 2021), label prediction for medical conversation utterances (Song et al., 2020), and dialogue turn classification (Lacson et al., 2006).

In detail, Krishna et al. (2021) evaluated the multi-label classification of noteworthy utterances that are relevant to each summary section before clustering related utterances and generating one summary sentence per cluster. Their modular summarization technique outperforms its purely abstractive counterpart, producing much more factual and coherent sentences. Besides, Song et al. (2020) first identified two types of utterances (problem statements and treatment recommendations) and then generated summaries, they showed that for the particular dataset used, high-quality summaries can be generated by extracting these two types of utterances. Thus, in addition to reporting ROUGE scores, they also reported the precision, recall, and F scores of the predicted labels for utterances of medical conversations, compared to the standard labels. In addition, Lacson et al. (2006) also measured precision, recall, and F measure of dialogue turns classification.

4.2 Human Evaluation

The differences between intrinsic and extrinsic evaluation lie in the fact that the former aims to evaluate the properties of the system's output (Graham et al., 2018; Ji et al., 2022), while the latter aims to examine the extent to which the system accomplishes the overarching task for which it was developed. Of the 19 articles reviewed on text-based MRG, 9 included human evaluation (47%), and only 3 of them (16%) included extrinsic evaluation.

4.2.1 Intrinsic Approaches

The intrinsic human evaluation of generated reports comprises two categories as for automated metrics: text quality and medical correctness.

Text quality is important in MRG as in general NLG output. For text quality, a wide variety of properties can be considered and various linguistic parameters can be used, e.g. relevance, consistency, fluency, coherence, missing, hallucination, repetition and contraction. As an example, Srivastava et al. (2022) used four standard linguistic parameters: relevance (selection of relevant content), consistency (factual alignment between the summary and the source), fluency (linguistic quality of each sentence), and coherence (structure and organization of summary). In addition to these commonly used and well-studied criteria, the evaluation of MRG also concludes other medical correctness criteria, such as factually correct and medically relevant information (Joshi et al., 2020; Chintagunta et al., 2021), which are specific to MRG tasks. As another example, Ben Abacha et al. (2023) performed expert-based manual evaluation using NLG criteria such as Fluency and Non-redundancy, and medical criteria such as Critical Omissions, Hallucinations, Correct Facts, Incorrect Facts based on fact extraction.

Furthermore, depending on whether evaluators assess the output directly or by comparing different texts, intrinsic human evaluation can be classified into direct and relative evaluation. As for the articles involving human evaluation, they all used at least direct evaluation, i.e. the evaluators judged the generated texts directly on a defined scale. Some authors also performed relative evaluation in addition to direct evaluation: Joshi et al. (2020) and Chintagunta et al. (2021) performed a comparison task in which, given two summaries generated by different models and the associated dialogue, annotators had to choose which summary was better, they could also choose "both" and "none". Yim and Yetisgen (2021) ranked the four systems against each other, with 1 being the best, in addition to evaluating each system independently with a score from 1 to 5 for the categories relevancy, factual accuracy, writing-style, completeness, and overall.

In general, MRG outputs are evaluated at the report level, however, depending on the design of the model, some are additionally evaluated at the sentence/section/part level. For example, Krishna et al. (2021) divided SOAP notes into several subsections: *Family Medical History*, *Past Surgical History*, *Chief Complaint*, etc. Therefore, they evaluated the generated SOAP notes in two ways: 1) SOAP note sentence level and 2) SOAP note level.

4.2.2 Extrinsic Approaches

As for extrinsic human evaluation: to evaluate generated summaries, a team of mental health experts used clinical acceptability framework (Sekhon et al., 2017), which includes six parameters: affective attitude, burden, ethicality, coherence, opportunity costs, and perceived effectiveness (Srivastava et al., 2022). In addition, to perform a task-based evaluation and measure the usefulness of summaries for preserving important information in the medical setting, Lacson et al. (2006) asked physicians and nurses to create a list of key questions based on topics that commonly arise between hemodialysis patients and caregivers, and then asked five physicians to answer each of the six "yes/no" questions using each of 40 dialogues. Furthermore, in a study evaluating the correlation between human evaluation and automatic metrics in consultation note generation, Moramarco et al. (2022) asked 5 clinicians to post-edit generated notes and extract all errors.

4.2.3 Presence of Domain Experts

Most of the articles we reviewed that included a human evaluation involved domain experts, such as doctors serving patients on their telehealth platform (Chintagunta et al., 2021; Joshi et al., 2020), five licensed physicians (Lacson et al., 2006), three general practice physicians (Moramarco et al., 2021), an annotator with a medical degree (Yim and Yetisgen, 2021), etc. Sometimes, the expertise of the annotators is not specified, e.g. "trained human annotators" (Krishna et al., 2021).

We also note that of the 9 articles including human evaluation, 5 of them reported Inter-Evaluator Agreement: three of the medical dialogue summarization articles (Zhang et al., 2021; Moramarco et al., 2022; Ben Abacha et al., 2023), and two medical (report) summarization articles (Moramarco et al., 2021; Karn et al., 2022). It would be preferable to indicate Inter-Evaluator Agreement in the presence of several annotators.

5 Conclusion

Automating medical report generation can save time for experts and provide crucial information for decision-making. However, the evaluation process is necessary for validation and adoption of MRG systems in the real world. Due to the specificity of domain-specific NLG tasks like MRG, their evaluation requires more investigation and subtlety.

MRG evaluation shares similarities with general NLG evaluation, but it differs in its focus on domain knowledge and task-specific concerns, especially in the assessment of conceptual accuracy of medical concepts. However, the question of which medical facts to pay attention to (correlation, consensus, etc) is an open question, requiring close collaboration with experts in the field.

In addition, the evaluation of MRG systems requires both intrinsic and extrinsic evaluation. Intrinsic evaluation focuses on properties of the system's output, while extrinsic evaluation involves professional experts in the design acceptability process, developing a list of key questions, and post-editing. Future research should prioritize extrinsic evaluation, particularly in scenarios where references are unavailable, and developing efficient, medical task-specific automated measures.

Limitations

In this work, we studied only the evaluation of textual medical report generation from both automatic and human evaluation perspectives, but we did not study the evaluation of data-to-text medical report generation, which has its own specificities.

Acknowledgements

This research was carried out as part of the THERADIA project (funded by BPI) and was partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003) and the COST action Multi3Generation (CA18231). We would also like to thank the anonymous reviewers for their insightful comments.

References

Belén A. Baez Miranda, Sybille Caffiau, Catherine Garbay, and François Portet. 2015. Towards a computational generation of récit: evaluating the perception of the récit plan. In *1st international Workshop on Data to Text Generation (D2T)*.

Sabita Acharya, Barbara Di Eugenio, Andrew D. Boyd, Karen Dunn Lopez, Richard Cameron, and Gail M Keenan. 2016. Generating summaries of hospitalizations: A new metric to assess the complexity of medical terms and their definitions. In *Proceedings of the 9th International Natural Language Generation conference*, pages 26–30, Edinburgh, UK. Association for Computational Linguistics.

Griffin Adams, Jason Zucker, and Noémie Elhadad. 2023. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317, Tilburg University, The Netherlands. Association for Computational Linguistics.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Viv Bewick, Liz Cheek, and Jonathan Ball. 2005. Statistics review 13: Receiver operating characteristic curves. *Critical care (London, England)*, 8:508–12.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. A benchmark for automatic medical consultation system: Frameworks, tasks and datasets.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Barbara Di Eugenio, Andrew Boyd, Camillo Lugaresi, Abhinaya Balasubramanian, Gail Keenan, Mike Burton, Tamara Goncalves Rezende Macieira, Jianrong Li, Yves Lussier, and Yves Lussier. 2014. Patient-Narr: Towards generating patient-centric summaries of hospital stays. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 6–10, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers. In *IJCAI*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60, Brighton, UK. Association for Computational Linguistics.

Yvette Graham, George Awad, and Alan Smeaton. 2018. Evaluation of automatic video captioning using direct assessment. *PLOS ONE*, 13(9):1–20.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J. of Biomedical Informatics*, 42(5):839–851.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad

Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Pamela Jordan, Nancy Green, Chistopher Thomas, and Susan Holm. 2014. TBI-doc: Generating patient & clinician reports from brain imaging data. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 143–146, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Sanjeev Kumar Karn, Ning Liu, Hinrich Schuetze, and Oladimeji Farri. 2022. Differentiable multi-agent actor-critic for multi-step radiology report summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1542–1553, Dublin, Ireland. Association for Computational Linguistics.

Nazmul Kazi and Indika Kahanda. 2019. Automatically generating psychiatric case notes from digital transcripts of doctor-patient conversations. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 140–148, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Nazmul Kazi, Matt Kuntz, Upulee Kanewala, and Indika Kahanda. 2020. Dataset for automated medical transcription.

Ali Can Kocabiyikoglu, Jean-Marc Babouchkine, François Portet, and Raheel Qader. 2021. Neural medication extraction: A comparison of recent models in supervised and semi-supervised learning settings. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 148–152.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Karen Kukich. 1983. Design of a knowledge-based report generator. In *21st Annual Meeting of the Association for Computational Linguistics*, pages 145–150, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Ronilda C. Lacson, Regina Barzilay, and William J. Long. 2006. Automatic analysis of medical dialogue in the home hemodialysis domain: Structure induction and summarization. *Journal of Biomedical Informatics*, 39(5):541–555. Dialog Systems for Health Communications.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021. Contrastive attention for automatic chest X-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280, Online. Association for Computational Linguistics.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F. Chen. 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821.

Justin Lovelace and Bobak Mortazavi. 2020. Learning to generate clinically coherent chest X-ray reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1235–1243, Online. Association for Computational Linguistics.

Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Xi Jiang, Lei Guo, Xiaoyan Cai, Shu Zhang, Tuo Zhang, Dajiang Zhu, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Impressiongpt: An iterative optimizing framework for radiology report summarization with chatgpt.

Lientje Maas, Adriaan Kisjes, Iman Hashemi, Floris Heijmans, Fabiano Dalpiaz, Sandra Van Dulmen, and Sjaak Brinkkemper. 2021. Automated medical reporting: From multimodal inputs to medical reports through knowledge graphs. In *HEALTHINF*.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 735–745, New York, NY, USA. Association for Computing Machinery.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2021. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Comput. Surv.* Just Accepted.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics.

Sabine Molenaar, Lientje Maas, Verónica Burriel, Fabiano Dalpiaz, and Sjaak Brinkkemper. 2020. Medical dialogue summarization for automated reporting in healthcare. *Advanced Information Systems Engineering Workshops*, 382:76 – 88.

Francesco Moramarco, Damir Juric, Aleksandar Savkov, and Ehud Reiter. 2021. Towards objectively evaluating the quality of generated medical summaries. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 56–61, Online. Association for Computational Linguistics.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Varun Nair, Namit Katariya, Xavier Amatriain, Ilya Valmianski, and Anitha Kannan. 2021. Adding more data does not always help: A study in medical conversation summarization with PEGASUS. *CoRR*, abs/2111.07564.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive transformer-based generation of radiology reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7):789–816.

Raheel Qader, Khoder Jneid, François Portet, and Cyril Labbé. 2018. Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation. In *11th International Conference on Natural Language Generation*, Tilburg, Netherlands.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, Dublin, Ireland. Association for Computational Linguistics.

Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8029–8036, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ehud Reiter, Albert Gatt, François Portet, and Marian van der Meulen. 2008. The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 147–156, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Ehud Reiter, Roma Robertson, and Liesl M. Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1):41–58.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandeep Sekhon, Martin Cartwright, and Jill J. Francis. 2017. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. *BMC Health Serv Res*, 17(88):114–133.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *Medical Information Retrieval (MedIR) Workshop, SIGIR*.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aseem Srivastava, Tharun Suresh, Sarah P. Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3920–3930, New York, NY, USA. Association for Computing Machinery.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

J.P. Woodard and J.T. Nelson. 1982. An information theoretic measure of speech recognition performance. In *Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA*.

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2021. Weakly supervised contrastive learning for chest X-ray report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4009–4015, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wen-wai Yim and Meliha Yetisgen. 2021. Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20, Online. Association for Computational Linguistics.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

A Appendix

A.1 Review Structure

We reviewed the MRG articles following the criteria described in van der Lee et al. (2019), the attributes examined are presented in Table 3.

A.2 Resources for Text-based MRG

We identified six datasets used in text-based MRG and summarized them in Table 4. Song et al. (2020) collected medical conversations from online platforms. MTSamples (Moramarco et al., 2021) were also collected from a community platform website. Srivastava et al. (2022) extended data from the publicly available counseling conversation dataset - HOPE (Malhotra et al., 2022), which takes place between therapist and patient. We observed that mock conversations were mentioned several times: Kazi et al. (2020) used transcripts from two different sources to generate audio recordings of simulated doctor-patient conversations, Papadopoulos Korfiatis et al. (2022) simulated primary care consultations. Recently, Ben Abacha et al. (2023) introduced a new collection of simulated doctor-patient conversations from publicly available clinical notes and corresponding clinical summaries.

	Criteria
Task	Sub-task(s) of MRG
Uses Automated Metrics	YES/NO
What kind of Automated Metrics	NLP metrics and/or other specific metrics
Uses Intrinsic (Human) Evaluation	YES/NO
What kind of Intrinsic (Human) Evaluation	Fluency, naturalness, quality, meaning preservation, etc.
Scale	Likert (5-point), preference, rank-based magnitude estimation, etc.
Number of participants	Number of annotators for the Human Evaluation task (including details on annotators)
Uses Extrinsic (Human) Evaluation	YES/NO
What kind of Extrinsic (Human) Evaluation	Task success, etc.
Number of examples	Number of samples evaluated for each system
Examples per participant	Number of examples that each participant is asked to evaluate
Details about design (order, groups)	Methods for selecting human evaluation samples from the original test set and how they are distributed to each annotator
Inter-Annotator Agreement (IAA)	Presence of inter-annotator agreement statistics

Table 3: Attributes studied and their descriptions in our structured review, adapted from van der Lee et al. (2019). *MRG* means *Medical Report Generation*.

Dataset	**Language**	**Description**	**Domain**	**Size**
Medical Conversation (ChiCCo) (Song et al., 2020)	Chinese	The summary has two parts: SUM1 describes the patient's medical problem; SUM2 summarizes the doctor's diagnosis or treatment recommendations.	Medical (online platforms conversation)	44,983 cases, 855,403 utterances
Automated Medical Transcription (Kazi et al., 2020)	English	Used transcripts from two different sources to generate audio recordings of enacted doctor-patient conversations	Medical, psychiatric consultations	71 recordings with transcripts and case notes
MTSamples (Moramarco et al., 2021)	English	From a community platform website, 40 medical specialties. Reports are free text with headings –> to generate the description field of a report	Medical summaries	5,000 sample medical transcription reports
MEMO (Srivastava et al., 2022)	English	Extend data collected from the publicly available counseling conversation (between therapist and patient) dataset - HOPE (Malhotra et al., 2022) to annotate psychotherapy elements and counseling summary	Mental health, Counseling	12.9K utterances, 212 conversations
PriMock57 (Papadopoulos Korfiatis et al., 2022)	English	Mocked primary care consultations, including audio recordings, their manual utterance level transcriptions, and the associated consultation notes	Medical, Primary Care Mock Consultations	57
MTS-Dialog (Ben Abacha et al., 2023)	English	A collection of 1.7k doctor-patient conversations and corresponding clinical notes/summaries.	Doctor-Patient Encounters	1.7k

Table 4: Text-based Medical Report Generation related datasets.