Consultation Checklists: Standardising the Human Evaluation of Medical Note Generation

Aleksandar Savkov[1], Francesco Moramarco[1,2], Alex Papadopoulos Korfiatis[1],
Mark Perera, Anya Belz[2,3], Ehud Reiter[2]

[1]Babylon [2]University of Aberdeen [3]ADAPT Research Centre, Dublin City University

[1]{sasho.savkov, francesco.moramarco, alex.papadopoulos}
@babylonhealth.co.uk
[2]{r01fm20, ehud.reiter, anya.belz}@abdn.ac.uk

Abstract

Evaluating automatically generated text is generally hard due to the inherently subjective nature of many aspects of the output quality. This difficulty is compounded in automatic consultation note generation by differing opinions between medical experts both about which patient statements should be included in generated notes and about their respective importance in arriving at a diagnosis. Previous real-world evaluations of note-generation systems saw substantial disagreement between expert evaluators. In this paper we propose a protocol that aims to increase objectivity by grounding evaluations in Consultation Checklists, which are created in a preliminary step and then used as a common point of reference during quality assessment. We observed good levels of inter-annotator agreement in a first evaluation study using the protocol; further, using Consultation Checklists produced in the study as reference for automatic metrics such as ROUGE or BERTScore improves their correlation with human judgements compared to using the original human note.

1 Introduction

While Electronic Health Record systems are a necessity in modern healthcare, they are burdening primary care clinicians with significant clerical work that distracts them from patient care and increases their dissatisfaction and burnout rates (Arndt et al., 2017). Since a significant part of the required documentation involves note writing, there has been a mounting interest in assisting clinicians by automatically generating consultation notes (Finley et al., 2018; Enarvi et al., 2020; Molenaar et al., 2020; Knoll et al., 2022).

A common approach involves passing the recording of the consultation through a speech-to-text system, then using a sequence-to-sequence model trained on parallel transcript and note datasets to automatically generate the note (Krishna et al., 2020;

Transcript	Note
Clinician: Hello there, it's Dr Smith, and how can I help you this afternoon?	3/7 hx developed headache. Constant, severity 8/10, dull ache with associated sharp pain, gradual onset.
Patient: Hi there. Well, I have this like really crazy headache that's been going on for days.	Progressively worsening. Has tried ibuprofen with limited relief.
Clinician: Ohh dear, OK. When, when did it exactly start, this headache?	Feels nauseous, no vomit. No neck pain/stiffness. No speech disturbances. No arm or leg weakness.
Patient: Eh, around three days ago, maybe.	No head injury. No fevers. No rashes. PMH: Nil.
Clinician: Three days ago, OK. And whereabout in your head, is this pain?	DH: Nil. NKDA FH: mother and sister - migraines
Patient: Um, it kind of feels all over my head, but mainly around my right eye. [...]	SH: lives with housemates, works in IT Socially smoke/EtOH.

Table 1: Abridged version of a mock transcript and human-written note from Papadopoulos Korfiatis et al. (2022).

Joshi et al., 2020; Zhang et al., 2021; Moramarco et al., 2022). An example of a transcript and associated consultation note, taken from the mock consultation dataset released by Papadopoulos Korfiatis et al. (2022), can be seen in Figure 1.

Evaluating the output of such systems is challenging (Gehrmann et al., 2022), as it is often the case in Natural Language Generation (NLG). Widely used automatic metrics, such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), often fail to capture relevant aspects of generated text (Reiter and Belz, 2009), and human evaluation, the best practice in NLG, is not only expensive and hard to reproduce (Belz et al., 2021) but also highly subjective (Howcroft et al., 2020; van der Lee et al., 2021; Gehrmann et al., 2022). Even in the field of Note Generation where evaluators tend to be medical experts rather than crowd-sourced workers, inter-annotator agreement is low, as there is no explicit ground truth and the annotators have dif-

fering opinions on the importance of each patient statement and whether it should be included in a consultation note (Moramarco et al., 2022).

In this work, we propose an evaluation protocol that uses Consultation Checklists (CC), itemised reference of all facts discussed during doctor-patient consultations. We report good agreement between clinicians building CCs from the same consultation, which indicates good consistency of the reference creation process. Since Consultation Checklists act as an approximation of the ground truth, they reduce evaluator subjectivity, which is reflected in the high inter-annotator agreement observed in our first study. We also show that correlation with human judgements increases when using CCs instead of the original clinician note as the reference for automatic evaluation metrics.

2 Related Work

There are a number of different approaches to quantitative human evaluation in NLG.

Rating or Likert scales work well with few criteria, but lack explanatory power and fail to capture text quality (Hastie and Belz, 2014). Adding more criteria partly resolves this, but at the cost of the evaluation task becoming more difficult and subjective (van der Lee et al., 2021). For example, Moen et al. (2016) use a 30 item rating scale and report that subjects found it too difficult to use.

Ranking methods, where evaluators are asked to rank the output of text generation systems along a specified criterion, are an alternative to rating scales. Some studies have shown ranking to be more reliable and consistent (van der Lee et al., 2021); however, ranking methods do not scale well when comparing multiple models.

Extrinsic measures, such as measuring post-edit time of generated text (Moramarco et al., 2022) provide a better estimate of how useful the generated text may be to the final user, but are often expensive and subjective (Lai et al., 2022).

A common shortcoming among all methods described above is that none of them provide granular insights into the text generation systems' errors and how to address them. This is particularly important when evaluating automatically generated medical notes, where the factual accuracy and completeness of the generated note are critical, as well as identifying the situations where the system fails.

As with all summarisation tasks, Note Generation has an element of content selection that is highly subjective. For this reason, evaluating system-generated notes against a single reference summary would penalise those notes that diverge from the reference in their content selection. One way to address this is by using multiple reference summaries[1], as for example in the Pyramid evaluation protocol (Nenkova and Passonneau, 2004). In a similar way to our proposed protocol, it splits the evaluation into two independent steps: extracting Summarization Content Units (SCUs) from multiple references, then using these SCUs to evaluate generated text.

Another way to address the subjectivity of using single references comes from *reference-less* approaches that compare the generated text against the source text directly rather than against reference summaries. For example, Narayan et al. (2019) once more split the evaluation in two steps: highlighting annotations in the source document, then comparing each generated summary to these annotations. In the domain of note generation, however, the format of the source documents (consultation transcripts) and the generated summaries (consultation notes) is different enough (see Table 1) that a highlighting-based approach would not work.

In the medical domain Moramarco et al. (2022) give evaluators the consultation audio recording instead of a reference, and ask them to identify the missing and incorrect items in a generated note, providing the required insight into how the generated text is wrong. However, even though the evaluators are experts (medical practitioners), agreement between them is very low. While this could be improved by better evaluator training, we believe that the evaluation task itself inherently inhibits agreement. There is no standard way of recognising or mapping facts from the audio recording of the consultation, which is the 'ground truth', to the consultation note – generated or otherwise. This makes it very hard to align multiple evaluators and get consistent results.

3 Proposed Protocol

We propose Consultation Checklists — a protocol using an expert-crafted ground truth approximation to evaluate the quality of system-generated medical notes with human raters. The evaluation protocol

---

[1]Multiple reference summaries are also used in some automatic evaluation metrics, such as ROUGE (Lin, 2004).

Checklist	Importance	Present / Absent
PRESENTING COMPLAINT		
Headache	C	Y
↳ 1 day	C	Y
↳ since she woke up yesterday morning	C	N
↳ throbbing	C	Y
↳ left side of head	C	Y
↳ intermittent	C	N
↳ sensitive to light	C	N
↳ headache still present	C	Y
PAST MEDICAL HISTORY		
Nil	C	N
DRUG HISTORY		
States allergy to aspirin	C	Y
↳ Feels sick when she takes aspirin	C	N
FAMILY HISTORY		
Mum - hypertension and migraines	C	Y
SOCIAL HISTORY		
Lives with friends	NC	N
Non-smoker	C	N

Note	Correct / Incorrect	Importance
PC: Pain in the left side of the head.	Y	C
HPC: Onset of a headache for the past day.	Y	C
Safety Netting: Nil abnormal smells or tastes,	Y	C
↳ Nil nausea or vomiting.	Y	C
PC: Headache.	Y	NC
Factors : light,	Y	C
↳ sensitive to light.	Y	C
E&D: Fine.	Y	C
PC: Left sided headache.	Y	NC
Factors : sunlight,	Y	NC
↳ lying flat,	N	NC
↳ bending forward.	N	C
Meds: None.	N	C
Factors : lying in bed,	N	NC
↳ lying flat.	N	NC
Meds : Meds.	N	C
Allergy to aspirin.	Y	C
Safety Netting : Nil changes in the skin,	N	C
↳ Nil weakness or numbness in the arms or legs.	Y	C

Figure 1: Columns A and B are the abridged version of a Consultation Checklist. Column E is the automatically-itemised system generated note. Columns C, F, and G are filled in by the evaluating clinician.

consists of a reference creation step followed by a notes evaluation step (see Figure 2).

3.1 Creation

Given a dataset of consultation audio recordings, one expert clinician is asked to listen to the audio and produce a Consultation Checklist: a structured list of **all the facts** discussed in the consultation. Including both relevant and irrelevant content in the Consultation Checklist is an important feature of the protocol as it eliminates the subjectivity of the content selection characteristic of other human-made references as discussed in Section 2. The list items are organised in sections for clarity and split into sub-lists to allow for more granularity (e.g. 'headache for 1 day' may be item 'Headache' with sub-item '1 day'); see Columns A and B in Figure 1. Following Moramarco et al. (2022), each item is marked for clinical importance as follows:

Critical: Items medico-legally required to document the diagnosis and treatment decisions whose absence or incorrectness may lead to wrong diagnosis and treatment later on, e.g. the symptom 'cough' in a suspected chest infection consultation. This is the key information a note needs to capture correctly in order to not mislead clinicians.

Non-critical: Items that should be documented in a complete note but whose absence will not affect future treatment or diagnosis, e.g. 'who the patient lives with' in a consultation about chest infection.
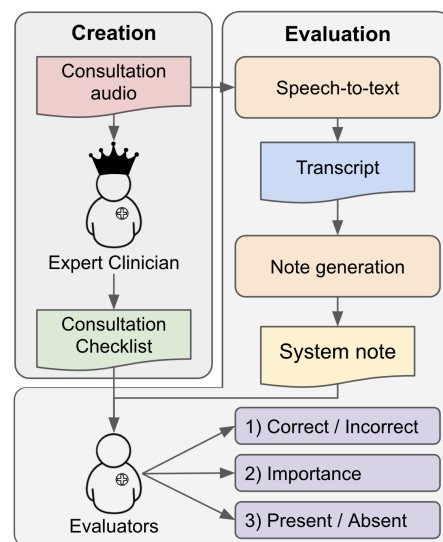

Figure 2: Diagram of the Consultation Checklists evaluation protocol, including the creation stage and the evaluation stage.

Irrelevant: Medically irrelevant information covered in the consultation, e.g. the pet of a patient with a suspected chest infection just died. Including such information in the Consultation Checklist allows for a fair evaluation of the less relevant parts of the generated notes.

Once the Consultation Checklists are created, they can be stored and reused in any future evaluation, thereby making the evaluation cost more scalable.

3.2 Evaluation

The created Consultation Checklists are then used to evaluate one or more system-generated consultation notes by one or more clinicians (or raters). These clinicians are not required to listen to the consultation recording but to only rely on the Consultation Checklist as a common ground when evaluating the notes. Each note is automatically split into sentences, then each sentence is split on punctuation and conjunctions. The first item in each sentence is considered top level and all others are nested in a sub-list (column E in Figure 1). We instruct evaluators to read the full sentence before marking each item as they may be meaningless in isolation. Once familiar with the Consultation Checklist, clinicians are asked to carry out the following sequence of tasks (see Figure 2):

1. Mark each item in the note as *correct* or *incorrect* using the Consultation Checklist as a common reference (column F). Since there is no explicit mapping, the clinicians need to scan the Consultation Checklist to try and find supporting facts for each item in the note. For example, in Figure 1 item 2 in the note ('HPC[2]: Onset of a headache for the past day') is validated by the first two items in the Consultation Checklist ('Headache' and '1 day').

2. Mark each item in the generated note as critical, non-critical, or irrelevant (column G). Even though some of these values could be inferred from the importance of the Consultation Checklist items, asking the evaluating clinicians to fill them in covers two common edge cases: (i) when one item in the note is fact-checked against multiple items in the Consultation Checklist, and (ii) when the item is incorrect. In the case of incorrect items, the importance is established by the effect the presence of the item may have to the clinician using the generated note.

3. Mark each item in the Consultation Checklist as present or absent in the generated note (column C) where 'present' means that the item is fully reported in the generated note.

We define *Precision* and *Recall* in the context of Consultation Checklists as:

$$Precision = \frac{\mid correct\ items \mid}{\mid generated\ items \mid} \quad (1)$$

---

[2]History of Presenting Complaint

$$Recall = \frac{\mid present\ items \mid}{\mid checklist\ items \mid} \quad (2)$$

Both metrics can also be computed for critical items only using the importance level assigned to each item (see Figure 1, column B).

The items in each Consultation Checklist are similar to the 'Summarization Content Units' described by Nenkova and Passonneau (2004) and can be re-used to evaluate any number of generated notes. However, as our method does not use human notes as a reference, the items are extracted from the consultation audio recording rather than from multiple reference summaries. As the evaluation stage involves no writing and only a limited amount of interpretation, the set of clinicians carrying it out could be of lower skill in contrast to the Pyramid approach where the same amount of skill is required for both stages.

4 Checklist Creation Pilot

Based on our initial assumption that a Consultation Checklist should capture the salient points of a consultation in an itemised format, we ran a pilot study with 2 clinicians expert in AI annotation, A & B. The goal of the study was to define best practices for creating Consultation Checklists and to evaluate the consistency of their creation between clinicians.

To investigate the agreement on Checklist creation, we asked the two clinicians to produce Consultation Checklists for the same 10 mock consultations taken from Papadopoulos Korfiatis et al. (2022) (see Figure 3 in the Appendix for an example). In order to quantify the alignment between them, two of the authors checked whether the information of each item in Clinician A's version was present in Clinician B's version, and vice-versa.

Based on this analysis, we define the fact coverage for each Consultation Checklist as the number of matching facts divided by the total number of facts. We found that 93.7% of the items in Clinician A's Consultation Checklists were also present in Clinician B's, and 78.3% the other way around. This may indicate that the two clinicians agree on the basic facts to include but Clinician B tended to add more details. Table 2 shows the values for both annotators, the average, and the agreement computed with Krippendorff's Alpha on the binary values for each statement (present or absent).

As part of the pilot, we refined the following process for creating Consultation Checklists:

Annotator	Checklist A (critical)	Checklist B (critical)
Ann1	94.7% (95.7%)	79.8% (79.1%)
Ann2	92.8% (94.3%)	76.9% (78%)
Avg	93.7% (95%)	78.3% (78.6%)
Agreement	0.624	0.77

Table 2: Results of the alignment between Checklists.

1. Listen to the consultation audio and take notes on every patient statement.

2. Format the notes into an itemised list, splitting longer items to a more atomic level in sub-lists and categorising them using sub-headers (e.g. Presenting History, Past Medical History, Social History, etc.). More examples of sub-headers can be seen in Figure 3 (Appendix A).

3. Read through two system-generated notes to sanity-check the Checklist with regards to item granularity and coverage.

4. Mark each item in the Checklist as critical, non-critical or irrelevant, as defined in Section 3.

5. Re-listen to the consultation audio to ensure no important points have been missed.

The pilot study also highlighted the cognitive effort of producing Consultation Checklists. On average, a Consultation Checklist for a 10 minutes consultation contains 56 items (excluding sub-headings) and it takes the clinician around 1 hour to complete. Also, the clinicians found that they would be able to produce 4 Consultation Checklists in a row before requiring a break, and that the first one would be the quickest to make (in as little as 30 minutes) but the following ones would require progressively more time.

5 Consultation Notes Evaluation

To test the utility of our protocol, we used a single expert clinician to create 20 Consultation Checklists from real-life patient consultations. We then generated History and Examination notes for each of the consultations using a BART (Lewis et al., 2020) encoder-decoder transformer model[3]. The model was pre-trained on the CNN/Dailymail dataset (Hermann et al., 2015), and fine-tuned on a proprietary dataset of 10,000 (speech-to-text generated) consultation transcripts and human-written History and Examination notes.

[3]https://huggingface.co/facebook/bart-large-cnn

Finally, we hired three clinicians (two women and one man from ethnically diverse backgrounds) to evaluate the 20 generated notes by following the evaluation process defined in Section 3. The training for this task involved: (i) task instructions (see Appendix); (ii) two evaluation practice tasks; (iii) an alignment session between the three clinicians, where they investigated all cases of disagreement on the practice tasks and came to a joint decision.

	Present / Absent	Correct / Incorrect	Imp.
# data points	2258	904	904
Eval 1 - Eval 2	0.733	0.690	0.521
Eval 1 - Eval 3	0.729	0.627	0.387
Eval 2 - Eval 3	0.754	0.701	0.697
3-way Agreement	0.739	0.672	0.522

Table 3: Krippendorf's Alpha inter-annotator agreement scores for Present/Absent, Correct/Incorrect and Importance (Imp.).

	CC (pairwise)	CC (count)	No CC (count)
Pre / Abs	0.739	0.969	0.374*
Cor / Inc	0.672	0.726	0.541*

Table 4: Krippendorff's Alpha scores for Present/Absent and Correct/Incorrect values for an evaluation using the Consultation Checklist as common ground and one using the consultation recording. Asterisk(*) denotes scores reported in Moramarco et al. (2022).

In addition, we wanted to quantify how often evaluators misjudge the generated note because of information that is omitted or misrepresented in Consultation Checklists. In order to do this, we asked each evaluator to listen to the audio of the consultation and review whether a generated note item was correct or incorrect.

Metric	Spearman correlation coefficient		Pearson correlation coefficient	
	Ref: Human note	Ref: Checklist	Ref: Human note	Ref: Checklist
Rouge1 Fscore	0.493	0.553	0.509	0.553
Rouge2 Fscore	0.376	0.570	0.445	0.545
Rouge3 Fscore	0.348	0.424	0.408	0.442
RougeL Fscore	0.431	0.636	0.439	0.577
BERTScore	0.375	0.58	0.414	0.563
Levenshtein Distance†	0.003	0.284	0.064	0.224

Table 5: Spearman and Pearson correlation coefficients against the human judgements (an average of precision and recall) from the study. † denotes lack of statistical significance ($p \geq 0.05$).

6 Results & Discussion

6.1 Inter-Annotator Agreement

One of the advantages of the Consultation Checklists protocol is that it allows us to compute inter-annotator agreement at a item level, as opposed to just error counts as done in Moramarco et al. (2022). Table 3 shows Krippendorff's Alpha[4] (Krippendorff, 2018) scores for the raw pairwise values of *Present / Absent*, *Correct / Incorrect*, and *Importance*. We use nominal agreement for the first two, which have binary values and ordinal agreement for Importance by converting *irrelevant*, *non-critical*, and *critical* to integers.

In order to compare agreement to Moramarco et al.'s reported results, we also compute interval agreement on error counts. While the results are not directly comparable since the dataset, sample size and annotator count are different, the agreement using the Consultation Checklists protocol is much higher (Table 4). It is also a considerable increase over the average NLG human evaluation agreement of 0.3 to 0.5 reported by van der Lee et al. (2021).

6.2 Accuracy Trade-off

As mentioned in Section 3, when generating Consultation Checklists, the goal is to capture as much of the consultation as possible. However, it is difficult to capture all points while keeping the Consultation Checklist concise, and some nuances which might be needed to faithfully assess the generated notes could be missed.

This trade-off between evaluation accuracy and standardisation is a limitation of our approach that we quantified by checking how often evaluators changed their Correct / Incorrect answers after listening to the consultation audio (Table 6). On average, this only happened for a small number

of generated note items (3.91%). Most changes were from Incorrect to Correct, which highlights the importance of making sure the Consultation Checklists are a thorough representation of the consultation. For example, consider a checklist that includes the item "was feeling cold" but omits the extra information of "had to wear more clothes than usual". In this case, a generated note item referring to this extra information would be marked as Incorrect based on the Consultation Checklist, but as Correct based on the audio.

Evaluator	# Changes	Correct -> Incorrect	Incorrect -> Correct
Eval 1	44 (4.87%)	9	35
Eval 2	39 (4.31%)	4	35
Eval 3	23 (2.54%)	3	20
Avg	35 (3.91%)	5.33	30

Table 6: Changes in Correct / Incorrect values after listening to the consultation recording.

6.3 Time Efficiency

It took clinicians a self-reported 45 minutes on average to complete each evaluation task: 5 minutes to understand the Consultation Checklist; 15 minutes for each of the two notes to evaluate correct vs. incorrect and present vs. absent items, including item importance; and 10 minutes to listen to the consultation audio and modify any answers. For context, Moramarco et al. (2022) report that their evaluators need 1 hour to listen to a consultation audio recording and evaluate 5 consultation notes.

6.4 Consultation Checklists for automatic evaluation

Since Consultation Checklists aim to be a standardised reference, we expected that their textual

---

[4]https://pypi.org/project/krippendorff/

representation[5] can also be used as a more objective reference than a single clinician's consultation note for automatic metrics. To test this, we computed the scores for a few common NLG metrics: ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019) and Levenshtein distance (Levenshtein et al., 1966) using either clinician-written notes or Consultation Checklists as the reference text. Table 5 reports both Spearman and Pearson correlation coefficients against our human judgements (an average of precision and recall). For all three metrics, using Consultation Checklists as references increases the correlation with human judgements, with BERTScore showing the highest gain at 20.5% Spearman's correlation increase.

7 Conclusion

In this work we proposed a novel reference data structure called Consultation Checklists (CC), and a protocol that uses it for evaluating automatically generated consultation notes. Our experiments show good inter-annotator agreement levels when parallel CCs are created from the same set of clinical consultations. We also report good agreement when the Consultation Checklist protocol is used by different clinicians to evaluate the same consultations. Finally, we showed that expertly-crafted Consultation Checklists are better than human-written notes when used as references for automatic evaluation.

While we have tested our protocol only on note generation for primary care consultations, we postulate that consultation checklists would apply to a number of medical domains, including secondary care and

8 Ethical Considerations

We considered the ethical implications of this work and found no concerns. The study participants are senior clinicians with at least 5-10 years of experience consulting. They are paid £70 an hour, have agreed to work for a maximum of 8 hours per week, and able to withdraw at any time. The consultations they are asked to evaluate are real doctor-patient interactions. These are stored securely following GDPR practices and the patients have consented for their data to be used for research purposes.

Finally, while we hope it can be generalised and applied to other domains, medical and otherwise,

---

[5]We get the textual representation of a Consultation Checklist by concatenating all items in a single string.

the evaluation protocol we propose in this paper has only been tested in the domain of primary care UK consultations.

References

Brian G. Arndt, John W. Beasley, Michelle D. Watkinson, Jonathan L. Temte, Wen-Jan Tuan, Christine A. Sinsky, and Valerie J. Gilchrist. 2017. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *The Annals of Family Medicine*, 15(5):419–426. Publisher: The Annals of Family Medicine Section: Original Research.

Anja Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The reprogen shared task on reproducibility of human evaluations in nlg: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258.

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.

Gregory Finley, Erik Edwards, Amanda Robinson, Michael Brenndoerfer, Najmeh Sadoughi, James Fone, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. An automated medical scribe for documenting clinical encounters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–15, New Orleans, Louisiana. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*.

Helen Hastie and Anja Belz. 2014. A comparative evaluation methodology for nlg in interactive systems. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4004–4011.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

David M Howcroft, Anja 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.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3755–3763.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. *In press: NAACL*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating soap notes from doctor-patient conversations. *arXiv preprint arXiv:2005.01795*.

Vivian Lai, Alison Smith-Renner, Ke Zhang, Ruijia Cheng, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. 2022. An exploration of post-editing effectiveness in text summarization. *In press: NAACL*.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25 – 37.

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, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anja 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.

Shashi Narayan, Andreas Vlachos, et al. 2019. Highres: Highlight-based reference-less evaluation of summarization. *arXiv preprint arXiv:1906.01361*.

Ani Nenkova and Rebecca J Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hltnaacl 2004*, pages 145–152.

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.

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.

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.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech Language*, 67:101151.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Evaluating Clinicians instructions

1. Read through the checklist (on the left side of the spreadsheet) and familiarise yourself with the consultation.

2. Based on the information in the checklist, go through the generated note (on the right side of the spreadsheet) and mark each statement of the note as **correct** (y) / **incorrect** (n). [Column G]

3. Mark each statement in the note as either critical, non critical or irrelevant: [Column I]

    (a) **Critical**: The statement is of critical medical importance. If it's a correct statement, the note would not be medically complete without the statement present; if it's an incorrect statement, its presence in a consultation note would be a medical risk (for example, could lead to a different diagnosis).

    (b) **Non-critical**: The statement is of medical value, but its presence or absence in the note is not critical medically. For example, some doctors might include a non-critical statement in their note but other doctors could skip it.

    (c) **Irrelevant**: The statement is irrelevant; if correct, most doctors would not include it in the note (for example, "Patient reports they prefer to wear green clothes"). If incorrect, its presence in the note is inconsequential.

4. Go through each statement in the checklist and mark it as either present or absent in the generated note. [Column C]

5. Repeat this for each generated note

6. Now, listen to the actual consultation recording. Take notes if you need to, especially if something you hear is different from what you understood through reading the consultation checklist.

7. Finally, fill the "Correct / Incorrect (after audio)" field, essentially amending your earlier answers after listening to the consultation audio. This will allow us to evaluate how much information and context is lost by using the consultation checklist instead of the actual recording. [Columns D,I]

A.1.1 Checklists comparison

Checklist (doctor 1)	Importance	Checklist (doctor 2)	Importance
SETTING		**INTRODUCTION**	
Patient confirmed he was in a private place to speak	C	Doctor's credentials	NC
PRESENTING HISTORY		Confirm name	NC
Elbow symptoms	C	Confirm DOB	NC
↳ Left elbow	C	In a private place	NC
↳ 1 week history	C	Able to speak freely	NC
↳ Pulling sensation	C	**PRESENTING COMPLAINT**	
↳ Initially noted while patient was in the shower	C	Problem with elbow	C
↳ Patient feels like there maybe "some fluid"	C	↳ Left	C
↳ Patient reports he does not know what is going on	NC	↳ Swelling	C
↳ Slightly warm.	C	↳ 1 week	C
↳ Not painful.	C	↳ Not painful	C
BACKGROUND HISTORY		↳ Noticed when in the shower	I
No previous injury	C	↳ Slightly warm	C
Works at a desk	C	↳ Feels fluid like inside	C
↳ not a physical job	NC	↳ No injury	C
No previous episodes	C	↳ Feels strange	NC
PAST MEDICAL HISTORY		**PAST MEDICAL HISTORY**	
possible osteoarthritis	C	Not happened before	C
↳ reportedly confirmed by a GP	C	Thinks has OA - diagnosed by GP previously	C
↳ Confirmed a few years ago.	C	No eczema	I
No known other diagnoses.	C	**DRUG HISTORY**	
No eczema.	NC	No known drug allergies	C
FAMILY HISTORY		Nil meds	C
no known family history of rheumatoid arthritis	C	Allergic to peanuts	I
no known family history of lupus		**FAMILY HISTORY**	
DRUG HISTORY		Unsure if hx of RA or lupus	C
↳ nil regular.	C	**SOCIAL HISTORY**	
↳ Allergic to peanuts.	NC	Sits at a desk	C
↳ No known drug allergies	C	Runs regularly 2/3 times a week	NC
SOCIAL HISTORY		**EXAMINATION**	
↳ Runs 2-3 times/week.	C	Tip of elbow is warm	C
↳ Patient considers himself healthy.	NC	Skin is dry on tip of elbow	C
DIGITAL EXAMINATION:			
Doctor asked patient to demonstrate some movements/actions	C		
↳ Some pain around where the swelling is	C		
↳ Skin was noted to have been dry in this region	C		

Figure 3: Example of two checklists for the same mock consultation.