Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation

Francesco Moramarco[†‡], Alex Papadopoulos Korfiatis[†], Mark Perera[†], Damir Juric[†], Jack Flann[†], Ehud Reiter[‡], Anya Belz[‡], Aleksandar Savkov[†]
[†]Babylon [‡]University of Aberdeen
[†] {francesco.moramarco, alex.papadopoulos, mark.perera, damir.juric, jack.flann, sasho.savkov}@babylonhealth.co.uk
[‡] {r01fm20, ehud.reiter, anya.belz}@abdn.ac.uk

Abstract

In recent years, machine learning models have rapidly become better at generating clinical consultation notes; yet, there is little work on how to properly evaluate the generated consultation notes to understand the impact they may have on both the clinician using them and the patient's clinical safety.

To address this we present an extensive human evaluation study of consultation notes where 5 clinicians (i) listen to 57 mock consultations, (ii) write their own notes, (iii) post-edit a number of automatically generated notes, and (iv) extract all the errors, both quantitative and qualitative. We then carry out a correlation study with 18 automatic quality metrics and the human judgements. We find that a simple, character-based Levenshtein distance metric performs on par if not better than common model-based metrics like BertScore. All our findings and annotations are open-sourced.

1 Introduction

Modern Electronic Health Records (EHR) systems require clinicians to keep a thorough record of every patient interaction and management decision. While this creates valuable data that may lead to better health decisions, it also significantly increases the burden on the clinicians, with studies showing this is a major contributor to burnout (Arndt et al., 2017).

In most primary healthcare practices, the universal record of a clinician-patient interaction is the SOAP (Subjective, Objective, Assessment, Plan) note, which captures the patient's history, and the clinician's observations, diagnosis, and management plan (Pearce et al., 2016). At the end of a consultation, the clinician is required to write up a SOAP note of the encounter. With the exception of the clinician's internal observations on how the patient looks and feels, most of the SOAP note is verbalised and could be automatically constructed from the transcript of the consultation.

A number of recent studies (Enarvi et al., 2020; Joshi et al., 2020; Zhang et al., 2021a) propose using summarisation systems to automatically generate consultation notes from the verbatim transcript of the consultation—a task henceforth referred to as Note Generation. Yet, there is very limited work on how to evaluate a Note Generation system so that it may be safely used in the clinical setting. Where evaluations are present, they are most often carried out with automatic metrics; while quick and cheap, these metrics were devised for general purpose summarisation or machine translation, and it is unclear whether they work just as well on this new task. In the field of automatic summarisation and Natural Language Generation (NLG) in general, human evaluation is the gold standard protocol. Even in cases where the cost of using human evaluation is prohibitive, it is essential to establish the ground truth scores which automatic metrics should aim for.

Our contributions are: (i) a large-scale human evaluation performed by 5 clinicians on a set of 285 consultation notes, (ii) a thorough analysis of the clinician annotations, and (iii) a correlation study with 18 automatic metrics, discussing limitations and identifying the most suitable metrics to this task. We release all annotations, human judgements, and metric scores.[1]

2 Related Work

Note Generation has been in the focus of the academic community with both extractive methods (Moen et al., 2016b; Alsentzer and Kim, 2018), and with abstractive neural methods (Zhang et al., 2018; Liu et al., 2019; MacAvaney et al., 2019; Zhang et al., 2020; Enarvi et al., 2020; Joshi et al., 2020; Krishna et al., 2021; Chintagunta et al., 2021; Yim and Yetisgen-Yildiz, 2021; Moramarco et al., 2021; Zhang et al., 2021a). Whether these studies

[1]https://github.com/babylonhealth/primock57

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 5739 - 5754

<table>
<tr><th colspan="2">Transcript</th><th>Note</th></tr>
<tr><td>Clinician</td><td>Hello.</td><td rowspan="8">3/7 hx of diarrhea, mainly watery.
No blood in stool. Opening bowels x6/day.
Associated LLQ pain - crampy, intermittent, nil radiation.
Also vomiting - mainly bilous.
No blood in vomit.
Fever on first day, nil since.
Has been feeling lethargic and weak since.
Takeaway 4/7 ago - Chinese restaurant.
Wife and children also unwell with vomiting, but no diarrhea. No other unwell contacts.
PMH: Asthma
DH: Inhalers
SH: works as an accountant.
Lives with wife and children.
Affecting his ADLs as has to be near toilet.
Nil smoking/etOH hx</td></tr>
<tr><td>Patient</td><td>Hello, how are you?</td></tr>
<tr><td>Clinician</td><td>Hello. How can I help you this morning?</td></tr>
<tr><td>Patient</td><td>All right. I just had some diarrhea for the last three days and it's been affecting me. I need to stay close to the toilet. And yeah, it's been affecting my day-to-day activities.</td></tr>
<tr><td>Clinician</td><td>I'm sorry to hear that and when you say diarrhea, what do you mean by diarrhea? Do you mean you're going to the toilet more often or are your stools more loose?</td></tr>
<tr><td>Patient</td><td>Yeah, so it's like loose and watery <u>stole</u> going to the toilet quite often.</td></tr>
<tr><td>Clinician</td><td><u>freak</u></td></tr>
<tr><td></td><td>...</td></tr>
</table>

Table 1: Snippet of a mock consultation transcript and the Subjective part of the corresponding SOAP note. The transcript is produced by Google Speech-to-text[3]; the bold-underlined text shows transcription errors. The note is written by the consulting clinician.

discuss the generation of radiology reports, patient-nurse summaries, discharge summaries, or SOAP notes, they all deal with long passages of text in the medical domain. This is a critical distinction from other application contexts (e.g. news summarisation): here, commonly used and well-studied evaluation criteria such as 'fluency', 'relevance', and 'adequacy' are superseded by other criteria, such as 'omissions of important negatives', 'misleading information', 'contradictions', etc. In addition, common summarisation metrics such as ROUGE (Lin, 2004) or BertScore (Zhang et al., 2019) measure the standalone quality of outputs and are not typically evaluated against more extrinsic criteria, such as post-editing times. Of the 18 studies on the subject that we could identify, 13 present an automatic evaluation (typically based on ROUGE and sometimes on medical entity linking) and 12 carry out a small-scale intrinsic human evaluation. In particular, Moen et al. (2016a) employ three domain experts to review 40 generated notes with Likert scales along 30 criteria (including 'Long-term diagnosis', 'Reason for admission', 'assessment'), but report that the subjects found the 30 item scale too difficult and detailed to assess. MacAvaney et al. (2019) use one domain expert to review 100 notes and report Likert scale values for 'Readability', 'Accuracy', and 'Completeness'. Moramarco et al. (2021) employ three clinicians and compare the times to post-edit generated notes with those of writing them from scratch, reporting that, while faster, post-editing may be more cognitively intensive than writing.

Outside of the medical domain, our work is comparable to Fabbri et al. (2021), who run an automatic metrics correlation study for news article summaries for the CNN/DailyMail dataset (Nallapati et al., 2016). They also release code[2] for evaluating text with a suite of common metrics, some of which we include in our own list of metrics to evaluate.

3 Dataset and Models

Our evaluation study is based on a dataset of 57 pairs of mock consultation transcripts and summary notes (Papadopoulos Korfiatis et al., 2022).[3] The data was produced by enacting consultations using clinical case cards. The clinicians that conducted the mock consultations also wrote the corresponding SOAP note. The consultations span common topics within primary healthcare and are about 10 minutes long.

To mimic a live clinical environment, the audio

[2] https://github.com/Yale-LILY/SummEval

[3] The dataset is available at: https://github.com/babylonhealth/primock57

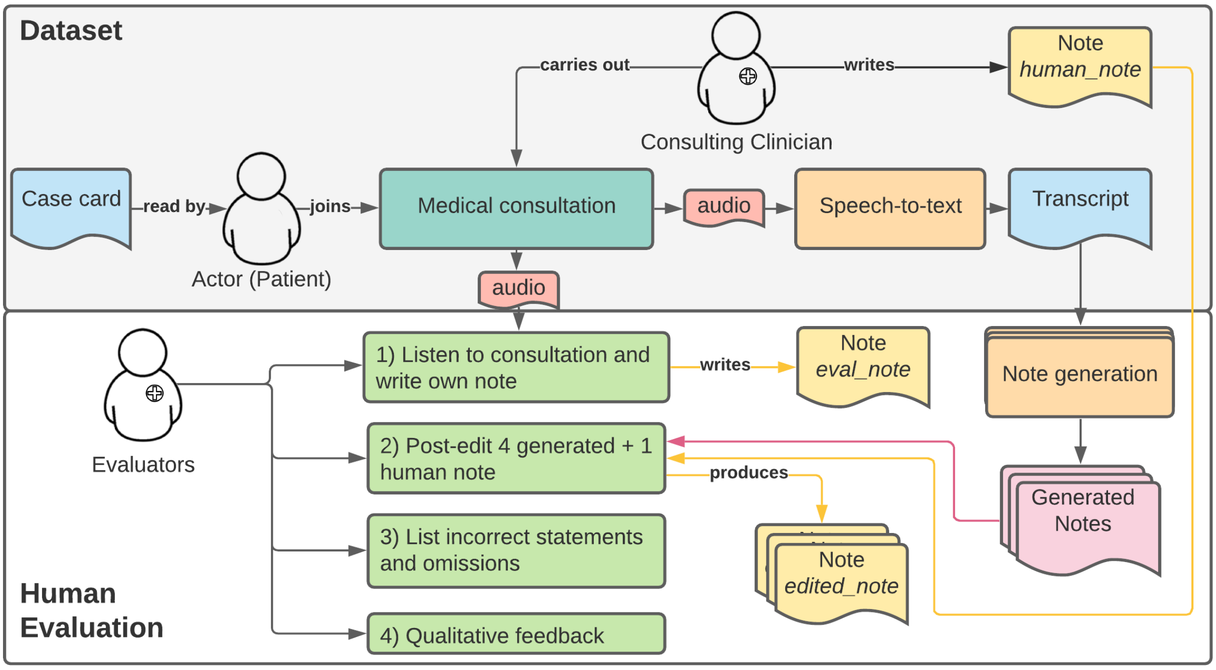

Figure 1: Diagram of the dataset creation and the four tasks involved in the human evaluation.

of the consultations was transcribed with Google Speech-to-text engine[4]. These transcripts form the input to the Note Generation models. The aim is to generate the Subjective part of a SOAP note. Table 1 shows an example transcript and respective note. Figure 1 describes the creation of the dataset and how the data feeds into the human evaluation tasks described below.

In a fashion similar to Chintagunta et al. (2021); Moramarco et al. (2021); Zhang et al. (2021a), we fine-tune 10 neural summarisation models based on BART (Lewis et al., 2020) on a proprietary dataset of 130,000 real consultation notes and transcripts. In accordance with our evaluation dataset, the training set consists of automatic Google Speech-to-text transcripts as inputs and the Subjective part of the corresponding notes as outputs.

The base models are large BART architectures pretrained on the CNN/Dailymail dataset[5]. Since our focus is on evaluation, the aim was to obtain models which would produce different outputs to cover a wider range of errors. The differences between the models included: fine-tuning on different sized datasets; using pre-processing techniques such as filtering the transcripts for relevant sentences; and using post-processing techniques such as filtering the generated notes for irrelevant sentences.

[4]https://cloud.google.com/speech-to-text

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

4 Human Evaluation Setup

Under the supervision of one of the authors (a clinician expert in AI development henceforth referred to as the Lead Clinician) we design the following evaluation tasks:

1. **Listen to the mock consultation audio and take notes (*eval_notes*).** These *eval_notes* appear on the evaluator screen throughout to help reduce the cognitive load of remembering what was discussed in the consultation.

2. **Relying on the *eval_notes* and the consultation audio, read 5 different notes and post-edit each one of them.** Post-editing consists of correcting an imperfect note to produce a factually accurate and relevant note (Sripada et al., 2005). It mimics how a synthetic note could be used in clinical practice while also bootstrapping the error identification (Moramarco et al., 2021). For this purpose, the evaluation platform includes a track-changes interface, which highlights insertions and deletions (Figure 2), and records the time taken to post-edit.

3. **For each note, classify the errors into two categories: 'incorrect statements' and 'omissions', by copying the spans of text from the post-editing interface and pasting them in the appropriate table (as in Figure**

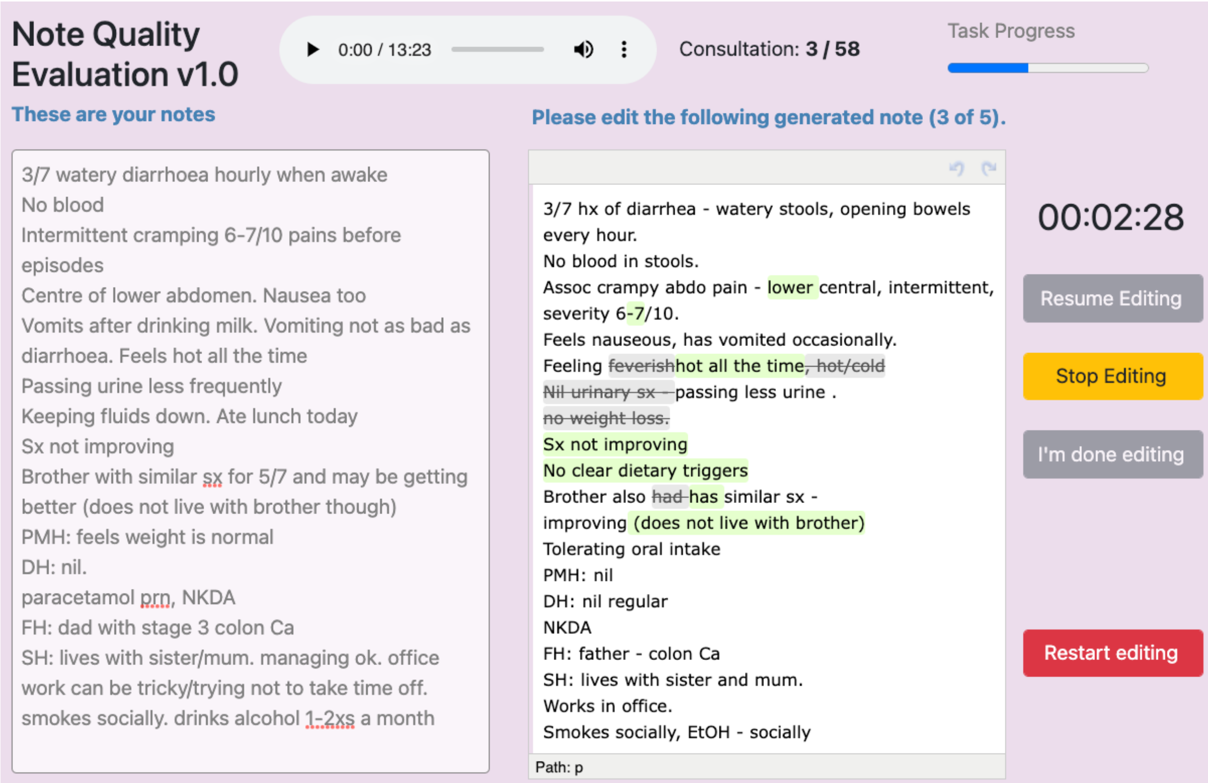

Figure 2: Screenshot of the post-editing task where the evaluator is correcting a note with the track-changes interface. Text in green shows what they have added, and text in grey (strikethrough) what they have deleted.

3). We define 'incorrect statements' as sentences in the generated notes which contain one or more factual errors (compared to the consultation audio). Conversely, 'omissions' are medical facts which should be recorded in a consultation note and were omitted by the model. Examples and edge cases (which were given to the evaluators for training) can be found in the Appendix, Figure A.4. Each error is also tagged as 'critical' if the information contained has essential clinical importance. Specifically, if the error would lead to medico-legal liability.

4. **Report any qualitative feedback (e.g. regarding order of statements, repetition) in the 'Other issues' box.** Figure 1 (bottom half) shows a diagram of the human evaluation workflow.

The subjects of the study were 5 regularly practising clinicians (GPs) with a minimum of 3 years experience. As part of our ethical consideration, all clinicians were paid the UK standard GP working rate and were free to cease participation at any time if they wished. For diversity and inclusion, 2 male clinicians and 3 female clinicians were enlisted from a range of ethnic backgrounds.

Following the tasks described above, each clinician evaluated the entire dataset of 57 mock consultations. Each consultation included 5 notes to evaluate, 4 of which were sampled from our 10 models and 1 was written by the consulting doctor (*human_note*). We shuffled these for every consultation and—to avoid biases—did not specify that one of the notes was not synthetic.

The evaluation study took circa 30 working hours per evaluator to complete over a period of 8 weeks. Before commencing, each evaluator went through a training and practice process conducted by the Lead Clinician, who explained the evaluation interface and guided them through the annotation of a practice note. A copy of the evaluator instructions can be found in Appendix A. Throughout the study, the authors and the Lead Clinician held weekly sessions with each evaluator where we shadowed the evaluation tasks through screen sharing. This helped us understand the difficulties in performing the tasks while ensuring the evaluators followed the guidelines set out for them.

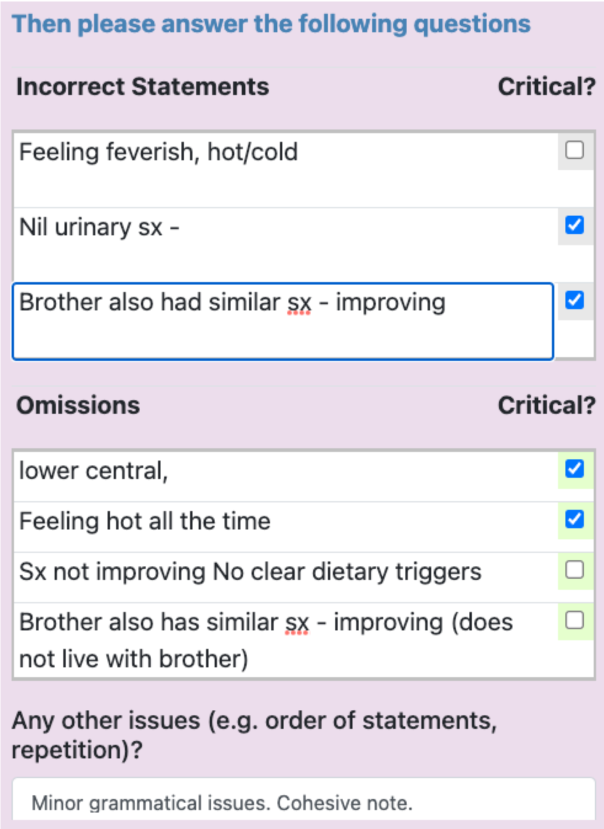

Figure 3: Screenshot of the scoring task, where the clinician is asked to quantify the incorrect statements and omissions in the generated note.

Criterion	Agree.	Word Overlap
Post-edit times	0.542^*	–
Incorrect statements	0.541^*	$0.431^\dagger$
Omissions	0.374^*	$0.268^\dagger$

Table 2: Inter Annotator Agreement. *Krippendorff's Alpha. †Word-level F1 score.

5 Results analysis

5.1 Agreement

The result of the human evaluation consists of 285 evaluator notes (57 consultations x 5 evaluators), 1,425 post-edited notes (285 x 5 notes per consultation), post-editing times, count and spans of incorrect statements, count and spans of omissions, whether they are critical, and qualitative comments. When compared with more common evaluation approaches such as Likert scales and ranking methods, we believe our set-up provides a more granular and more interpretable set of judgements, albeit at the price of lowering the inter-annotator agreement. To compensate for this, the 5 evaluators annotate the same 57 tasks (Sheng et al., 2008) and the scores

Criterion	Human	Generated
Post-edit times	96.5s	136.4s
Number of Incorrect	1.3	3.9
Number of Omissions	3.9	6.6
Note length	16.9	21.5

Table 3: Aggregated judgements for *human* notes and generated notes. Note length is in number of sentences.

are averaged in the correlation study (see Section 6).

As shown in Table 2, we compute inter-annotator agreement on the post-editing times, incorrect statements, and omissions. The absolute post-editing times are converted to a list of rankings for each evaluator, and agreement is computed with Krippendorff's alpha (Krippendorff, 2018) with 'ordinal' level of measurement. This ensures only the ranking of each note is captured in the agreement and not the editing speed of each evaluator. For example, where evaluator 1 takes 60 seconds and 120 seconds to post-edit two given notes and evaluator 2 takes 180 seconds and 240 seconds respectively, their agreement would be perfect because they both agreed that note 1 is quicker to edit than note 2. Conversely, for incorrect statements and omissions we calculate 'interval' Krippendorff's Alpha on the counts of errors identified by the evaluators. As the counts don't ensure that two evaluators have selected the same statements, we also compute word overlap F1 score as suggested by Popović and Belz (2021). As shown in Table 2, the agreements for times and incorrect statements are not very strong (Krippendorff (2018) indicate that $\alpha \geq 0.667$ is the lowest conceivable limit). We investigate the source of disagreement and attribute it to two main factors: (i) human error due to the difficulty inherent in the task, and (ii) stylistic differences in note writing. Examples of human error can be found in subsection 5.2. As for stylistic differences, these are especially evident in the Omissions category, where some clinicians are thorough in their note taking and others only document the most important facts. See Appendix B for more details on pairwise agreement.

5.2 Human error

To compare the accuracy of the models against human-written notes, we average all the judgements for our criteria (post-edit times, incorrect statements, and omissions), aggregate by the generated notes and the *human_notes* respectively, and report the results in Table 3. As expected, the *human_notes* performed better for all criteria; in particular, they contain fewer omissions while being on average 4.6 sentences shorter. However, the evaluators found imperfections in human notes too: it takes over 1.5 minutes on average to read and post-edit a *human_note*, and it contains over 1 incorrect statement and almost 4 omissions on average. While the omissions can be reconciled as stylistic differences among evaluators, the incorrect statements are potentially more impactful. To investigate, we select two human notes and ask the Lead Clinician to post-edit them, comparing the results with those of the evaluators. In the first case, the Lead Clinician agrees with the evaluators in that the human note contains the following two incorrect statements:

Inc. statement	**Correction**
Also vomiting – mainly bilous	Also vomiting
Wife and children also unwell with vomiting, but no diarrhea.	One child had some vomiting, but no other symptoms in wife and other child.

Upon inspecting the consultation recording, the Lead Clinician found that the word 'bilious' was not stated by the patient. However, the consulting clinician may have used this term due to a personal habitual documentation style (as clinically, vomit with no red flags can conventionally be referred to as bilious). The words 'Wife and children also unwell with vomiting, but no diarrhea' were not stated by the patient. Instead, the patient made a tangential statement summarised here: 'One child had some vomiting but no other symptoms in wife and other child.' Therefore, it is inferred that this clinician likely made a normal human error due to excessive patient detail (non-critical).

In the second case, the Lead Clinician found no issues with the human note. Upon inspecting the corrections from the evaluators, he concluded that what they selected as incorrect statements were medical conditions inferred by the consulting clinician yet not specifically stated by the patient. We highlight this to show that it is unclear whether the task has a single ground truth, as even human experts don't completely agree; well thought-out evaluation tasks can mitigate this and produce one or more good ground truth approximations. Detailed examples can be found in Appendix C.

Criterion 1	**Criterion 2**	**Pears.**	**Spear.**
Post-edit times	Incorrect	0.543	0.599
Post-edit times	Omissions	0.769	0.781
Post-edit times	Inc + Omi	0.8	0.829
Post-edit times	Note length	0.38	0.413
Incorrect	Omissions	0.384	0.467
Incorrect	Note length	0.537	0.52
Omissions	Note length	0.122	0.183

Table 4: Correlation coefficients between the criteria; all numbers statistically significant (p value < 0.001). A darker shade of green means higher correlation.

5.3 Analysis of criteria

To understand the interdependence between our criteria, we compute Pearson's correlation (Freedman et al., 2007) and Spearman's rank correlation (Zar, 2005) coefficients between each pair.

Table 4 shows a moderately strong correlation between the time it takes to post-edit a note and the number of incorrect statements it contains. The correlation between post-edit times and omissions is stronger, which could be explained by the fact that it takes longer to type an omitted statement than to delete or edit an incorrect one. Finally, the correlation between post-edit times and *incorrect+omissions* is strong, which suggests that post-edit times is a function of the number of edits and that one of these criteria could be a proxy for the other.

We also compute the correlation between each criterion and the length of the generated note. These numbers can be used as a benchmark for automatic metric correlation; for example, the 0.413 Spearman's correlation between post-edit times and note length indicates that any automatic metric needs to surpass this value in order to be more useful than simply counting the number of sentences in the note.

5.4 Qualitative results

As introduced in Section 4, the evaluators provide qualitative feedback about the generated notes in the 'Other Issues' field. When analysing these comments, a number of repeated patterns emerged, highlighting common pitfalls in the generated notes. Based on these we defined a small taxonomy (Table 5; issues in the *human_notes* are excluded), providing examples and occurrences of each issue type. Aside from incorrect statements and omissions, the most significant issues revolve around repetition,

Issue	Examples	Occ.
Discourse level		
Contradiction	*no family history of bowel issues.* *father has history of colon cancer*	25
Contradiction not reported	patient corrected herself but note did not pick it up.	4
Symptom mentioned is reported as fact	*tingling of hands* stated by clinician (not refuted/confirmed by pt)	18
Misleading statement	statement: *not working* when patient has been off ill for a few days due to current sickness reads like patient is unemployed	9
Factual errors		
Hallucination	*at home and in a private place* was not mentioned in the consultation	17
Incorrect statement	statement: *brother has diabetes*. Correction: *mother has diabetes*	34
Nonsensical	*No recent unwell with diarrhoea*	18
Stylistic errors		
Repetition	*loose and watery stools* *stool is mainly watery*	93
Incorrect order of statements	*heart attack* should be in PMH; structure of history a bit jumbled; recorded social smoker in alcohol section.	38
Use of not universally recognised acronyms	NRS/EMS/DOA are not standard acronyms	7
Omissions		
Generic	no mention of *unable to open bowels*	57
Omissions of important negatives	*No fever* *No shortness of breath*	5
Other		
Good behaviour	Contains all the history that was covered in the audio and follows a logical structure.	21

Table 5: Taxonomy of errors gathered through the qualitative feedback from the evaluators.

disjointed notes, and contradiction. Upon investigating, we believe that all three are related to the tendencies of the models to generate the consultation note following the chronological order of the transcript. While that is an intuitive behaviour, consultations are seldom carried out in the order of SOAP note sections (Subjective, Objective, Assessment, Plan), with the patient providing relevant information whenever they can, sometimes after the clinician has discussed assessment and plan.

6 Correlation with Automatic Metrics

Borrowing from the field of Automatic Summarisation, most studies on Note Generation rely on ROUGE and fact-extraction based metrics to evaluate the generated notes (Section 2 for more details). While some studies carry out a small human evaluation, there is little effort to investigate whether the scores from ROUGE or the other metrics employed correlate well with the human judgements, especially extrinsic criteria such as post-edit times. However, scores from these metrics are featured on leaderboards[6] for summarisation tasks, driving future research. To address this, we carry out a correlation study of automatic metrics for the task of Note Generation. A total of 18 automatic metrics are tested against statistics produced by the human judgements of our criteria: post-edit times, number of incorrect statements, and number of omissions. Following the taxonomies reported by Celikyilmaz et al. (2020) and Sai et al. (2020), the metrics considered can be loosely grouped in:

- **Text overlap metrics.** These are based on string matching, whether character based, word based, or n-gram based. Some use stemming, synonyms, or paraphrases. They include: ROUGE (Lin, 2004), CHRF (Popović, 2015), METEOR (Lavie and Agarwal, 2007),

[6] https://nlpprogress.com/english/summarization.html

Criterion:	Post-edit times					Inc+Omi		Incorrect		Omissions	
Reference:	human	edited	eval	avg	max	avg	max	avg	max	avg	max
ROUGE-1-F1*	0.334	0.627	0.160	0.443	0.550	0.580	0.704	0.378	0.505	0.561	0.651
ROUGE-2-F1*	0.384	0.653	0.166	0.551	0.570	0.694	**0.731**	0.501	0.557	0.641	0.654
ROUGE-3-F1*	0.366	0.645	0.117	**0.576**	0.565	**0.734**	**0.731**	0.555	**0.568**	0.663	0.646
ROUGE-4-F1*	0.342	0.632	0.076	0.575	0.557	**0.745**	0.726	**0.581**	**0.573**	0.661	0.636
ROUGE-L-Pr*	0.348	0.471	0.169	0.366	0.427	0.500	0.613	**0.607**	**0.745**	0.306	0.375
ROUGE-L-Re*	0.409	0.614	**0.300**	0.520	0.551	0.640	0.680	0.374	0.416	0.660	**0.688**
ROUGE-L-F1*	0.384	0.646	0.285	0.538	0.564	0.661	0.719	0.479	0.534	0.610	0.655
CHRF*	0.341	0.460	-0.075	0.463	0.438	0.581	0.560	0.504	0.484	0.483	0.462
METEOR*	**0.415**	**0.667**	0.203	0.529	**0.581**	0.674	0.713	0.429	0.463	**0.668**	**0.699**
BLEU*	0.382	0.642	0.098	0.557	0.565	0.698	0.702	0.447	0.453	**0.685**	0.686
Levenshtein dist.	**0.547**	**0.780**	**0.453**	**0.600**	**0.654**	0.650	**0.760**	0.566	0.555	0.531	**0.697**
WER	0.239	0.629	0.059	0.326	0.550	0.412	0.704	0.499	0.535	0.252	0.631
MER	0.392	0.635	0.156	0.565	0.557	0.703	0.706	0.500	0.513	0.659	0.651
WIL	0.394	0.649	0.117	**0.590**	0.566	**0.747**	0.723	**0.578**	0.566	**0.668**	0.638
ROUGE-WE*	0.402	0.624	0.165	0.496	0.549	0.621	0.712	0.415	0.524	0.595	0.650
SkipThoughts*	0.298	0.403	-0.067	0.229	0.375	0.366	0.504	0.338	0.407	0.288	0.439
Embedding Avg*	0.266	0.375	-0.209	0.064	0.412	0.223	0.572	0.147	0.392	0.211	0.543
VectorExtrema*	0.409	0.553	0.127	0.424	0.500	0.550	0.648	0.367	0.468	0.531	0.600
GreedyMatching*	0.308	0.577	-0.041	0.295	0.520	0.436	0.670	0.281	0.479	0.428	0.624
USE*	0.339	0.522	0.201	0.366	0.476	0.474	0.637	0.327	0.452	0.454	0.598
WMD	0.354	0.594	0.154	0.421	0.529	0.561	0.670	0.319	0.414	0.577	0.670
BertScore*	**0.497**	**0.688**	**0.340**	0.571	**0.590**	0.710	**0.744**	0.530	0.552	0.645	0.676
MoverScore*	0.360	0.640	0.246	0.570	0.559	0.687	0.688	0.448	0.467	**0.669**	0.657
Stanza+Snomed*	0.334	0.508	0.118	0.354	0.460	0.528	0.643	0.447	0.533	0.449	0.553

Table 6: Spearman's correlation coefficients for each metric and each criterion. In bold are the top three scores per column. All the metrics marked with an asterisk (*) are inversely correlated with the given criterion (e.g. higher post-edit time means worse note, but higher ROUGE score means better note); the sign of the coefficient is inverted for ease of visualisation. Coefficients in red are not statistically significant (with $p > 0.05$).

and BLEU (Papineni et al., 2002).

- **Edit distance metrics.** These count the number of character or word level transformations required to convert the system output into the reference text. They include: Levenshtein edit distance (Levenshtein et al., 1966), WER (Su et al., 1992), MER and WIL (Morris et al., 2004).

- **Embedding metrics,** including word-level, byte-level, and sentence-level embeddings. These metrics encode units of text with pre-trained models and compute cosine similarity between them. They include: ROUGE-WE (Morris et al., 2004), SkipThoughts, EmbeddingAverage, VectorExtrema (Forgues et al., 2014), GreedyMatching (Sharma et al., 2017), USE[7] (Cer et al., 2018), WMD (Kusner et al., 2015), BertScore (Zhang et al., 2019), and MoverScore (Zhao et al., 2019).

- **Fact-extraction.** The Stanza+Snomed metric extracts medical concept spans with Stanza (Zhang et al., 2021b), then uses similarity measures to map them to entities in the SNOMED CT clinical ontology (Spackman et al., 1997). The metric computes F1 score between reference and hypothesis over the set of extracted entities.

For more details on each metric please refer to their respective papers. All these metrics attempt to measure the accuracy of the generated text by comparing it against a reference text. Our human evaluation study produces three distinct human-curated notes which can be used as reference: the *human_note* is the original note, written by the consulting clinician (and also one of the hypotheses), the *eval_note* is the note written by the evaluators after listening to the consultation audio, and

[7]Cosine similarity between the reference and the hypothesis embeddings. Embeddings computed with Universal Sentence Encoder.

the *edited_note* is the generated note after being post-edited by the evaluators. Table 6 reports the correlation coefficients. When correlating against post-edit times, we consider each reference text (*human_note*, *edited_note*, *eval_note*) separately, then take the average and the maximum of the metric scores for each reference. For count of incorrect statements, omissions, and incorrect+omissions we only report the average and the maximum scores, taking all three references into account as commonly done by the metrics that support multiple references (e.g. BLEU, ROUGE, METEOR, BertScore). We compute Pearson's and Spearman's coefficients and, upon finding similar patterns, only report Spearman's coefficients in Table 6. The Pearson's coefficients can be found in Table A.8 in the Appendix.

As shown in Table 6, all metrics display a strong bias towards the choice of reference. In particular, the correlation scores with the *edited_note* as reference are much higher than those of either *human_note* or *eval_note*. As the *edited_note* is a transformation of the generated note (refer to Figure 2), these high correlations show how reliant all the metrics are on the surface form of the text. The significant difference between taking *human_note* and *eval_note* as reference can be traced to two main factors: (i) the *human_note* is unique per consultation so the human judgements are averaged across evaluators (reducing noise and collapsing disagreement), and (ii) the *eval_note* was not written to replace a SOAP note but is rather a list of the most salient points in the consultation, and sometimes contains more information than would typically be detailed in a SOAP note.

The top three metrics in most scenarios are Levenshtein distance, BertScore, and METEOR. While METEOR and BertScore are established metrics in NLG evaluation, Levenshtein distance is not typically used as a metric in long-form text evaluation. From a semantic point of view, edit distance has the least amount of knowledge and should be very brittle when comparing text that is meaningfully similar but lexically very different. Yet Levenshtein distance has the highest correlation even when the reference is the *eval_note*, which is syntactically very different from the generated note; whereas even contextual metrics like BertScore perform more poorly. A possible explanation for this behaviour may be that our post-editing times and count of incorrect statements/omissions—unlike Likert scales scores—measure the amount of work required to convert a synthetic note into a factually-correct and relevant note, just as Levenshtein distance measures the character-level distance between the synthetic note and the reference.

We notice that all the metrics correlate better with counts of *incorrect+omissions* than with post-edit times, despite the two criteria being strongly correlated with each other (0.829 Spearman's correlation, see Table 4). We believe this is due to post-editing times containing more noise and capturing more of the stylistic differences between evaluators than the number of errors does.

Correlation matrices between the metrics scores can be found in Appendix D.

7 Conclusions

We conducted a human evaluation study for the task of consultation Note Generation, computed agreement between evaluators, and quantified the extent to which human error impacts the judgements. We then carried out a correlation study with 18 automatic metrics, discussing their limitations and identifying the most successful ones.

We found that the choice of human reference has a significant effect on all automatic metrics and that simple character-based metrics like Levenshtein distance can be more effective than complex model-based metrics for the task of Note Generation. Based on our findings, character-based Levenshtein distance, BertScore, and METEOR are the most suitable metrics to evaluate this task. We release all the data and annotations and welcome researchers to assess further metrics.

8 Acknowledgements

The authors would like to thank Rachel Young and Tom Knoll for supporting the team and hiring the evaluators, Vitalii Zhelezniak for his advice on revising the paper, and Kristian Boda for helping to set up the Stanza+Snomed fact-extraction system.

References

Emily Alsentzer and Anne Kim. 2018. Extractive summarization of ehr discharge notes. *arXiv preprint arXiv:1810.12085*.

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.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

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.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, et al. 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.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. volume Vol. 2.

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

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.

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

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.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231.

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, Online. Association for Computational Linguistics.

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

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. IEEE.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W Filice. 2019. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1016.

Hans Moen, Juho Heimonen, Laura-Maria Murtola, Antti Airola, Tapio Pahikkala, Virpi Terävä, Riitta Danielsson-Ojala, Tapio Salakoski, and Sanna Salanterä. 2016a. On evaluation of automatically generated clinical discharge summaries. *Distributional Semantic Models for Clinical Text Applied to Health Record Summarization Doctoral thesis*, page 99.

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

Francesco Moramarco, Alex Papadopoulos Korfiatis, Aleksandar Savkov, and Ehud Reiter. 2021. A preliminary study on evaluating consultation notes with post-editing. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 62–68.

Andrew C. Morris, Viktoria Maier, and Phil D. Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *INTERSPEECH*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. (in press): Primock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the 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.

Patricia F Pearce, Laurie Anne Ferguson, Gwen S George, and Cynthia A Langford. 2016. The essential soap note in an ehr age. *The Nurse Practitioner*, 41(2):29–36.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Maja Popović and Anya Belz. 2021. A reproduction study of an annotation-based human evaluation of mt outputs. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 293–300.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2020. A survey of evaluation metrics used for nlg systems. *arXiv preprint arXiv:2008.12009*.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

Kent A Spackman, Keith E Campbell, and Roger A Côté. 1997. Snomed rt: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association.

Somayajulu Sripada, Ehud Reiter, and Lezan Hawizy. 2005. Evaluation of an nlg system using post-edit data: Lessons learnt. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. A new quantitative quality measure for machine translation systems. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Wen-wai Yim and Meliha Yetisgen-Yildiz. 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.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R Gormley. 2021a. Leveraging pretrained models for automatic summarization of doctor-patient conversations. *arXiv preprint arXiv:2109.12174*.

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*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213.

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.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021b. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Appendix

A Instructions for evaluators

Figure A.4 shows a table of examples provided to the evaluators to help them understand how to list incorrect statements and omissions, especially around edge cases.

B Pairwise Agreement

Pairwise agreement is reported in Table A.7. The agreement on post-edit times is a rank agreement computed by ranking the times and using ordinal Krippendorff Alpha; the number of incorrect statements and number of omissions agreement is computed with interval Krippendorff Alpha; and the word overlap is computed with word-level F1 score.

C Human error examples

Example human note 1.

```
3/7 hx of diarrhea, mainly watery.
No blood in stool.
Opening bowels x6/day.
Associated LLQ pain - crampy, inter-
mittent, nil radiation.
Also vomiting - mainly bilous.
No blood in vomit.
Fever on first day, nil since.
Has been feeling lethargic and weak since.
Takeaway 4/7 ago - Chinese restaurant.
Wife and children also unwell
with vomiting, but no diarrhea.
No other unwell contacts.
PMH: Asthma
DH: Inhalers
SH: works as an accountant.
Lives with wife and children.
Affecting his ADLs as has to be near
toilet often.
Nil smoking/etOH hx
```

Incorrect statements

- Also vomiting - ~~mainly bilous~~, resolved after 1st day (complex incorrect statement, critical)
- ~~Wife and children also unwell with vomiting~~. -> 1 child had some vomiting but no other symptoms in wife and other child. (complex incorrect statement, non-critical)

Example human note 2.

```
PC: Cough and cold.
HPC: 4-5 day hx runny nose, dry cough.
No sputum/haemoptysis.
No epistaxis/sinus pain.
Feels hot, hasn't measured temperature.
No SOB/inspiratory pain/wheeze.
Aches and pains. No vomiting.
E&D ok. PUing ok.
No hx chest problems/recurrent chest
infections/wt loss.
Thinks last BP was fine, can't remember
when last BP/DM check up was.
Doesn't check blood sugars/urine at home.
No increased thirst/urinary freq.
Taking 2-3 lemsips a day which eases sx.
PMH: Hypertension. T2 DM.
DH: Lisinopril. Metformin
Height- 5ft 5in
Weight - 65kg
SH: Non smoker, odd sherry.
Lives with partner and dog.
Office manager.
```

Here the Lead Clinician found no incorrect statements. However, some of the evaluators did.

E1	– dm
E2	
E3	– Thinks last BP was fine, can't remember when last BP/DM check up was. – Doesn't check blood sugars/urine at home.
E4	– can't remember when last DM check up was. – Doesn't check blood sugars/urine at home. – 4 day hx
E5	

Upon inspecting the recording of the consultation, the Lead Clinician found that the words 'Thinks last BP was fine, can't remember when last BP/DM check up was. Doesn't check blood sugars/urine at home' were not stated by the patient. Instead, the patient made a tangential statement summarised here: 'diabetes (generally well controlled, last blood test was 3 weeks ago), hypertension (doesn't remember last blood pressure check)'. Please note that in diabetic and hypertension patients, clinical convention is to indicate severity of diagnosis by whether the patient requires home monitoring. Therefore, the clinician inferred part of the statement. Furthermore, clinical convention is to differentiate diabetes mellitus (DM) from diabetes insipidus (DI). Therefore again, the clinician inferred part of the statement. Any other errors are attributed to normal

Text	Incorrect Statements	Omissions	Explanation	Critical
Omissions				
Lives with wife and children.		Lives with wife and children.	Omission	No
Simple incorrect statements				
3 ~~13~~ day history of diarrhoea	13 day history of headache		Simple incorrect statements can be corrected without labelling as an omission	Yes
Opening bowels 6 ~~x16~~/day	x16 /day		Simple incorrect statements can be corrected without labelling as an omission	Yes
Takeaway 4/7 ago - Mexican ~~Chinese~~ restaurant.	Chinese restaurant.		Simple incorrect statements can be corrected without labelling as an omission	No
Complex incorrect statements				
~~No blood in stool~~ some streaks of fresh blood on wiping only	No blood in stool	some streaks of fresh blood on wiping only	One incorrect statement, one omission (complex incorrect statements require additional detail)	Yes
Vomiting for 6 ~~3~~ days, intermittent	3 days	intermittent	One incorrect statement, one omission (complex incorrect statements require additional detail)	Yes
Wife and 2 children had vomiting but no diarrhoea. ~~also have diarrhoea.~~	also have diarrhoea.	had vomiting but no diarrhoea.	One incorrect statement, one omission (complex incorrect statements require additional detail)	Yes
Edits that do not require logging as incorrect or omission				
Headache for 3 days. ~~Headache~~			Repeated/redundant text is neither incorrect nor omitted	-
~~sometimes when passes urine feels like its hard~~ feels like it's hard to pass urine			Grammatical issues do not enter in either “incorrect statements” or “omissions”	-
PMH: asthma~~.~~			Stylistic additions (PMH, deleting the full stop)	-

Figure A.4: Table of examples given to the evaluators for reference.

human error due excessive patient detail (non-critical).

D Correlation Matrices

We compute correlation matrices with Spearman's and Pearson's coefficients for all automatic metrics, considering all three references aggregated by taking the average and maximum scores respectively (Figures A.5, A.6, A.7, A.8).

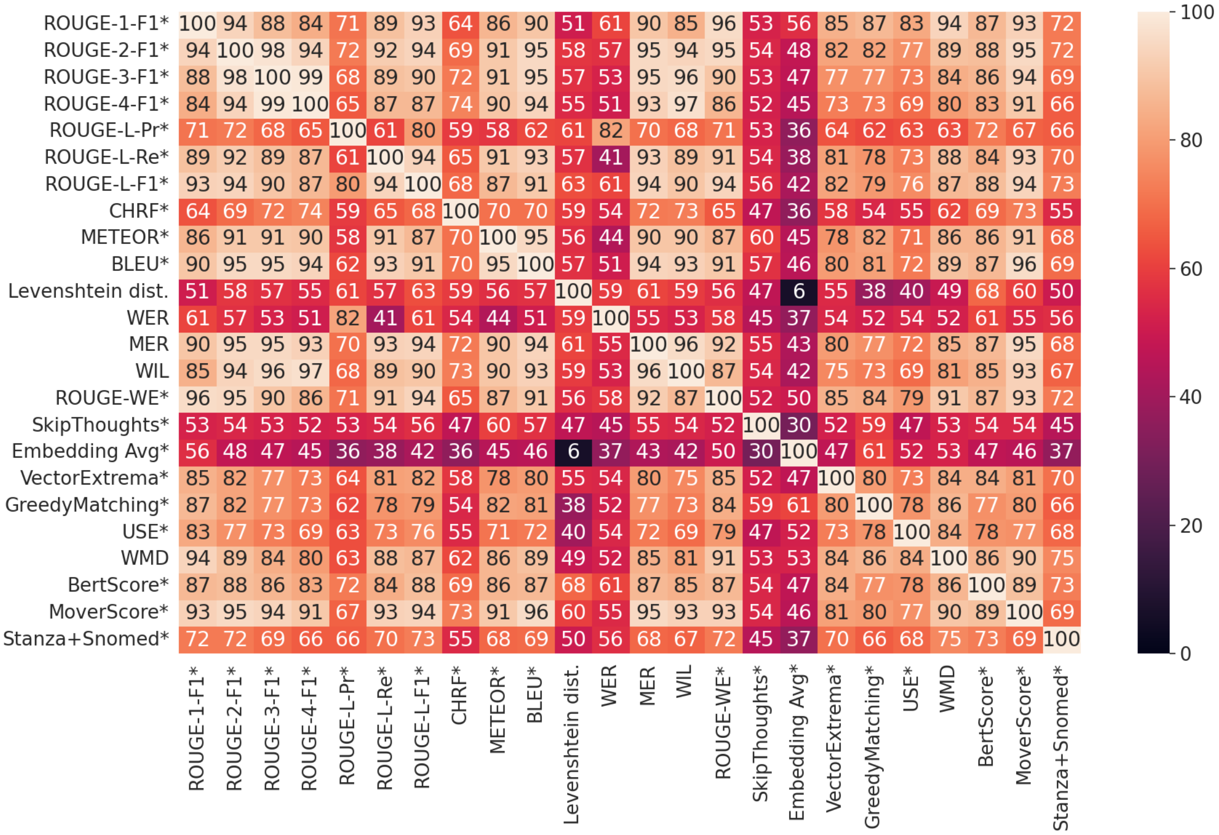

Figure A.5: Spearman's correlation matrix between automatic metrics by using all three references and taking the average score. Values represented as percentages for ease of visualisation.

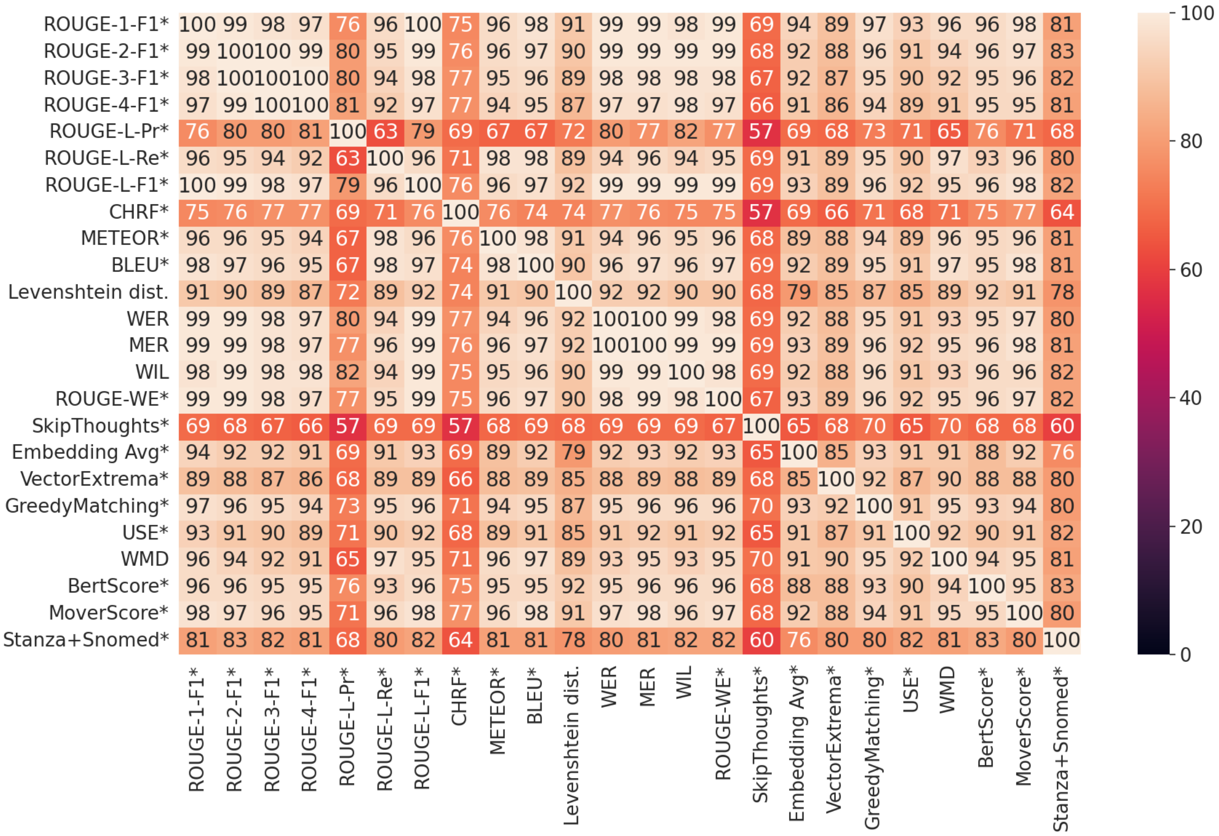

Figure A.6: Spearman's correlation matrix between automatic metrics by using all three references and taking the maximum score. Values represented as percentages for ease of visualisation.

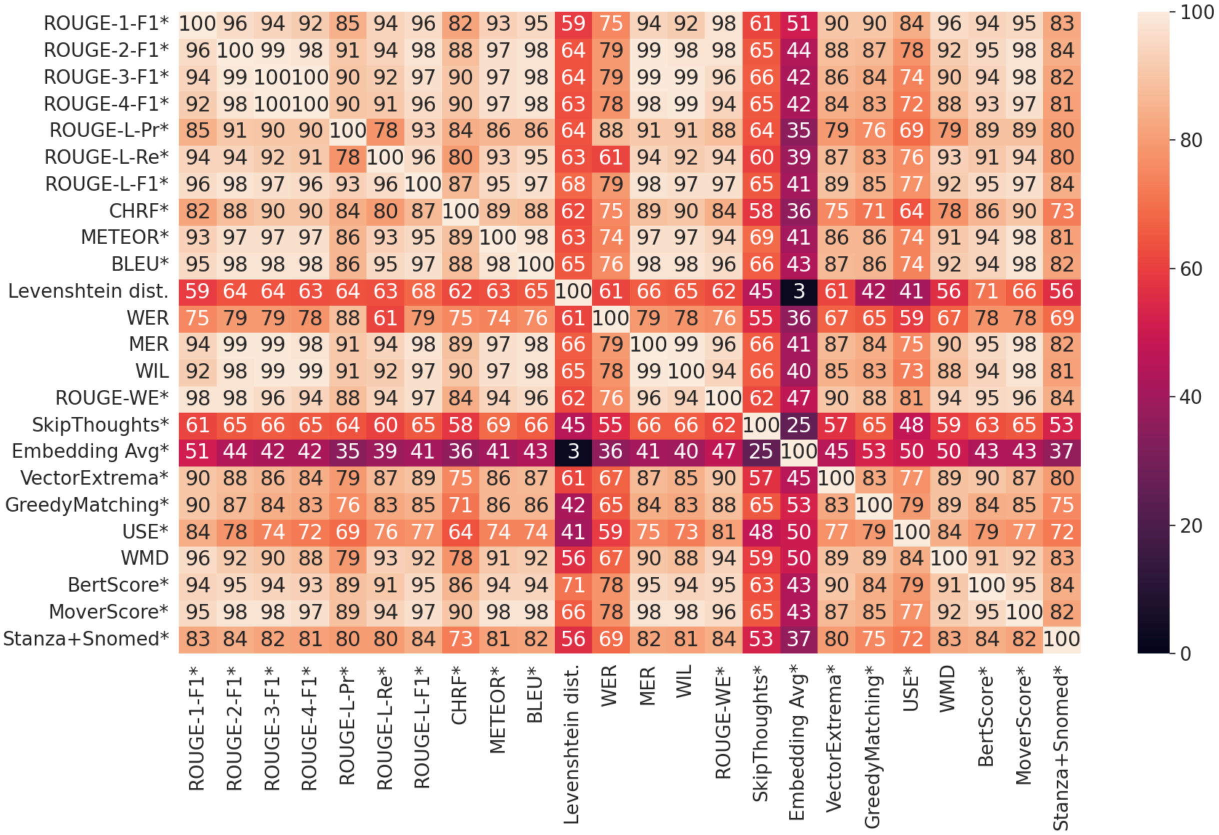

Figure A.7: Pearson's correlation matrix between automatic metrics by using all three references and taking the average score. Values represented as percentages for ease of visualisation.

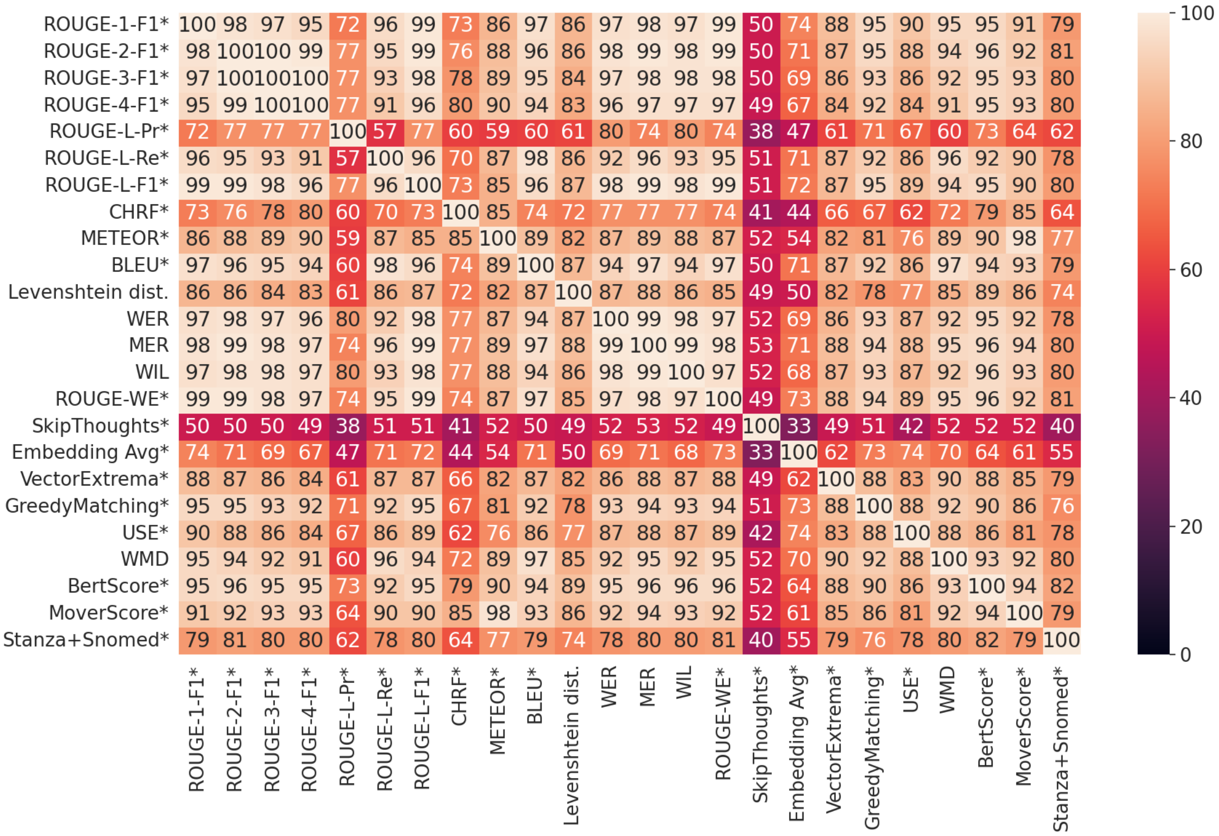

Figure A.8: Pearson's correlation matrix between automatic metrics by using all three references and taking the maximum score. Values represented as percentages for ease of visualisation.

Eval 1	Eval 2	Post-edit times	Incorrect statements		Omissions	
			Count	Word Overlap	Count	Word Overlap
eval1	eval2	0.444	0.573	0.369	0.124	0.240
eval1	eval3	0.534	0.599	0.484	0.452	0.266
eval1	eval4	0.660	0.624	0.471	0.538	0.282
eval1	eval5	0.591	0.639	0.415	0.408	0.260
eval2	eval3	0.408	0.413	0.386	0.232	0.240
eval2	eval4	0.420	0.303	0.389	-0.024	0.243
eval2	eval5	0.634	0.717	0.401	0.543	0.270
eval3	eval4	0.501	0.626	0.531	0.449	0.307
eval3	eval5	0.520	0.512	0.433	0.495	0.269
eval4	eval5	0.664	0.371	0.436	0.294	0.301

Table A.7: Pairwise agreement between evaluators.

Human scores:	Timings					Inc+Omi		Inc		Omi	
Reference:	**human**	**edited**	**eval**	**avg**	**max**	**avg**	**max**	**avg**	**max**	**avg**	**max**
ROUGE-1-F1*	0.374	0.583	0.156	0.417	0.539	0.574	0.693	0.413	0.474	0.537	0.667
ROUGE-2-F1*	0.378	0.601	0.149	0.426	0.545	0.597	**0.717**	0.484	0.538	0.519	0.655
ROUGE-3-F1*	0.368	0.594	0.105	0.417	0.533	0.600	0.711	0.501	0.550	0.511	0.638
ROUGE-4-F1*	0.363	0.583	0.082	0.413	0.518	0.602	0.702	0.510	**0.553**	0.507	0.621
ROUGE-L-Pr*	0.368	0.370	0.125	0.315	0.368	0.492	0.589	**0.547**	**0.685**	0.318	0.36
ROUGE-L-Re*	0.406	0.594	**0.308**	**0.491**	**0.554**	**0.627**	0.674	0.389	0.381	**0.632**	**0.708**
ROUGE-L-F1*	0.387	0.595	0.283	0.445	0.553	0.603	0.714	0.483	0.513	0.529	0.668
CHRF*	0.368	0.445	-0.013	0.397	0.404	0.547	0.542	0.476	0.469	0.451	0.45
METEOR*	0.379	**0.617**	0.210	0.414	0.433	0.580	0.591	0.431	0.447	0.533	0.537
BLEU*	0.370	0.609	0.115	0.443	0.551	0.603	0.678	0.440	0.409	0.560	**0.693**
Levenshtein dist.	**0.503**	**0.772**	**0.472**	**0.612**	**0.667**	**0.663**	**0.765**	**0.550**	0.547	**0.567**	**0.718**
WER	0.280	0.565	0.035	0.291	0.515	0.416	0.687	0.477	0.527	0.258	0.62
MER	0.377	0.590	0.144	0.427	0.531	0.600	0.693	0.485	0.497	0.522	0.649
WIL	0.363	0.599	0.080	0.429	0.534	0.615	0.707	**0.522**	**0.557**	0.517	0.626
ROUGE-WE*	0.392	0.573	0.139	0.432	0.530	0.586	0.698	0.440	0.498	0.535	0.656
SkipThoughts*	0.323	0.301	-0.099	0.213	0.301	0.338	0.403	0.305	0.298	0.270	0.371
Embedding Avg*	0.230	0.260	-0.188	0.039	0.304	0.180	0.406	0.107	0.207	0.186	0.442
VectorExtrema*	**0.408**	0.515	0.095	0.397	0.481	0.543	0.632	0.379	0.434	0.517	0.607
GreedyMatching*	0.371	0.493	-0.046	0.29	0.484	0.43	0.629	0.295	0.433	0.413	0.603
USE*	0.35	0.449	0.178	0.343	0.456	0.459	0.609	0.326	0.417	0.433	0.586
WMD	0.389	0.565	0.146	0.413	0.515	0.561	0.653	0.351	0.389	**0.564**	0.671
BertScore*	**0.413**	**0.646**	**0.325**	**0.471**	**0.56**	**0.643**	**0.73**	**0.522**	0.549	0.558	0.666
MoverScore*	0.368	0.605	0.24	0.434	0.468	0.587	0.62	0.449	0.46	0.53	0.571
Stanza+Snomed*	0.37	0.468	0.106	0.35	0.443	0.529	0.633	0.441	0.504	0.45	0.557

Table A.8: Pearson's correlation coefficients for each metric and each human judgement. In bold are the top three scores per column. All the metrics marked with an asterisk (*) are inversely correlated with the given criterion (e.g. higher post-edit time means worse note, but higher ROUGE score means better note); the sign of the coefficient is inverted for ease of visualisation. Coefficients in red are not statistically significant (with p > 0.05).