# A preliminary study on evaluating Consultation Notes with Post-Editing

**Francesco Moramarco***
Babylon Health / London, UK
University of Aberdeen / Aberdeen, UK
francesco.moramarco†

**Alex Papadopoulos Korfiatis***
Babylon Health / London, UK
alex.papadopoulos†

**Aleksandar Savkov**
Babylon Health / London, UK
sasho.savkov†

**Ehud Reiter**
University of Aberdeen / Aberdeen, UK
e.reiter@abdn.ac.uk

†@babylonhealth.com

## Abstract

Automatic summarisation has the potential to aid physicians in streamlining clerical tasks such as note taking. But it is notoriously difficult to evaluate these systems and demonstrate that they are safe to be used in a clinical setting. To circumvent this issue, we propose a semi-automatic approach whereby physicians post-edit generated notes before submitting them. We conduct a preliminary study on the time saving of automatically generated consultation notes with post-editing. Our evaluators are asked to listen to mock consultations and to post-edit three generated notes. We time this and find that it is faster than writing the note from scratch. We present insights and lessons learnt from this experiment.

## 1 Introduction

In modern EHR (Electronic Health Records) systems, at the end of a medical consultation the physician is required to file a consultation note detailing symptoms, examination, and treatment discussed. This is a pain point for physicians, who, according to a US study in 2017-2018 (Arndt et al., 2017) spend up to $44.2\%$ of their time on clerical tasks, and this is a major contributor to physician burnout (Medscape, 2018).

A number of recent studies (Kazi and Kahanda, 2019; Molenaar et al., 2020; Krishna et al., 2020) propose to use summarisation systems to automatically generate the consultation note from the transcript of the consultation. Yet there is limited work on how to properly evaluate such a system so that it may be used in the clinical setting. Intrinsic evaluation metrics through Likert scales or ranking methods may help select the best model, but they don't ensure the model will never hallucinate information or miss key items when generating the consultation note.

In this study we propose to evaluate generated consultation notes with an extrinsic measure based on post-editing time. We ask our evaluators (primary healthcare physicians) to listen to a consultation, write a consultation note, and post-edit a number of generated notes. We then compare the timings of each task to determine whether post-editing a note is faster than writing one from scratch.

We focus on post-editing time because (i) it's simple to measure, and (ii) it provides a gate for adoption of the technology (i.e. post-editing a note should be faster than writing it from scratch). There are other extrinsic metrics which we intend to investigate in the future, such as patient satisfaction, doctor cognitive load, doctor-patient engagement, and usefulness for the next doctor accessing the note.

## 2 Related Work

Post-editing has a long history in Machine Translation (MT) (Chander, 1994; Carl et al., 2015; Graham et al., 2017; Koponen, 2016; De Sousa et al., 2011), with a number of production systems and tools using a semi-automatic approach to fix errors and check the output of the system before it is shown to the users (Dowling et al., 2016; Aziz and Specia, 2012).

Outside of MT, Sripada et al. (2005) carry out a study on post-editing an NLG system for generating weather forecast from data.

As an evaluation metric, Allman et al. (2012) define *Productivity* as "the quantity of text an experienced translator could translate in a given period of time [compared] with the quantity of text gen-

*Equal contribution



*Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 62–68
Online, April 19, 2021. 

erated by [the system] that the same person could edit in the given time."

To the best of our knowledge, post-editing is not widely used in document summarisation. We speculate this is partly because a post-editor of document summaries would need to read the entire document in order to accurately post-edit the summary, and this may minimise the benefit of having a generated summary compared to writing it from scratch. This is not the case, however, with consultation note generation, whereby the physician in charge of writing the note is the same physician who has conducted the consultation. Here post-editing may be very valuable in saving physician time.

## 3 Data

We partner with a UK healthcare provider, Babylon Health, which gives us access to a dataset of 800 proprietary consultation transcripts (automatically transcribed) and notes. The consultations span various topics within primary healthcare and are 10 minutes long on average. The notes are written by the physician who carried out the consultation and are in patient-friendly format, meaning they are in the same language the doctor used while talking to the patient and don't contain any abbreviations or acronyms. Each note is made up of three sections: *History & Examination*, *Diagnosis*, and *Management*.

For our evaluation, we design a dataset of 57 mock consultations produced in a similar manner. We ask five Babylon Health physicians working in primary healthcare to act as doctors and a number of lay people (employees at Babylon) to act as patients. Participation is entirely voluntary and all participants sign a consent form explaining what the study would involve and the intended use of the data produced. They are given the choice to withdraw consent at any point.

We give each patient a case card, prepared by a physician, that contains the condition they need help with and a list of medical details and symptoms. We record the audio of each mock consultation and ask the doctor to write a patient-friendly note as described above. Figure 1 shows a mock patient-friendly note.

We then employ a transcription agency to transcribe the recordings on an utterance level. Figure 2 shows a transcript snippet from the same consultation.

For the evaluation reported here, we only use 3 out of the 57 consultations; we are planning to publish the whole dataset at a later stage.

**HISTORY & EXAMINATION**
You developed lower abdominal pain 2 days ago. The pain came on gradually, is burning in nature, constant and is worsening. You have no bowel symptoms or pain on urination, but have noticed a pink colour to your urine. You have not noticed and blood in your urine. You feel some nausea, but have not vomited. You feel hot and sweaty. You are sexually active with a long term partner. Your last sexual health check-up was 6 months ago. You last had unprotected sex 2 days ago. Your last period was 2 weeks ago. You have no other symptoms. You have no past medical history, but use implanon for contraception.

**DIAGNOSIS:** Urinary Tract Infection. Must rule out pregnancy

**MANAGEMENT**
Take a pregnancy test. Give urine sample for a urine dip and to check for bacteria. Treat with antibiotics. Regular paracetamol for pain. Review in 1 - 2 days if no improvement, or earlier if symptoms are worsening.

Figure 1: Mock consultation note written by a locum doctor, from our evaluation dataset.

## 4 Experimental Setup

We use the proprietary dataset of 800 consultations to finetune two automatic summarisation models based on BART (Lewis et al., 2020). We feed the transcripts as inputs and the consultation notes as outputs.

We then apply the models on the mock consultation dataset, using them to generate the *History & Examination* section of the consultation note. For our experiment, we consider the generated notes from these two models (Model A, Model B) together with the original reference note (Ref). We shuffle these for each task and tell the evaluators that all three notes are generated.

The task is presented to the evaluators using Heartex (Tkachenko et al., 2020), a configurable annotation platform that allows us to customise the design of the evaluation task.

Our evaluators are three primary healthcare physicians. They are employed at Babylon Health

[...]
**Doctor**: Hello? Good morning, Tim. Um, how can I help you this morning?
**Patient**: Um, so I'm having some, some pain, uh, in my tummy, like the lower part of my tummy. Um and I've just been feeling, quite, hot and sweaty.
**Doctor**: OK. Right, I'm sorry to hear that. When, when did your symptoms all start?
**Patient**: About two days ago.
**Doctor**: OK. And whereabouts in your tummy is the pain, exactly?
**Patient**: Uh, like below my belly button, it's like quite, sore when I press on it.
**Doctor**: OK. Did the pain come on quite suddenly, or was it more gradual?
**Patient**: it hasn't been, it's more gradual and it's just, it is getting a bit worse now.
**Doctor**: OK, OK. And can you describe the pain to me? [...]

Figure 2: Sample transcript from the mock dataset.

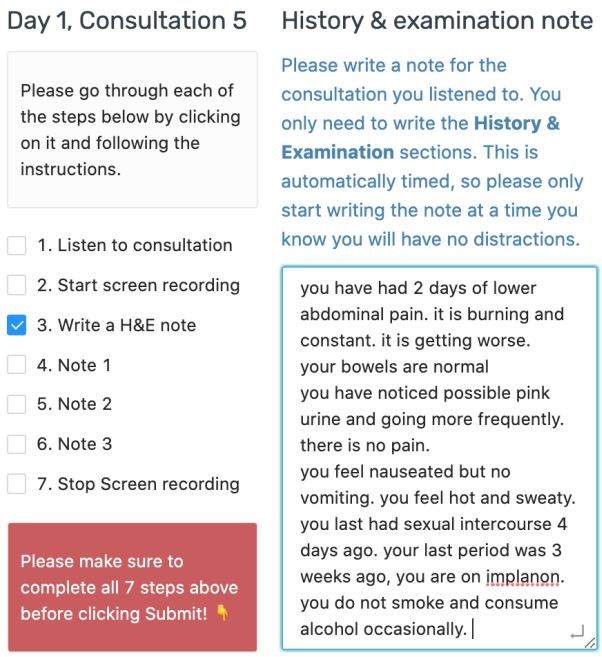


Figure 3: Heartex Annotation interface for writing the History&Examination section of the consultation note.

and have experience in AI research annotation. The task we submit to them consists of the following steps:

1. **Listen to the audio of a mock consultation.** We let evaluators note down any key symptoms on a piece of paper as they would normally do during a consultation.
2. **Write the *History & Examination* sections of the consultation note (this is timed).** Just as they would in the clinical setting, after having listened to a consultation recording we ask them to write the first section of the consultation note. Figure 3 shows an example.
3. **Post-edit three generated notes (this is timed).** The evaluators are presented with the three generated notes (Model A, Model B, Ref, in random order) for the given consultation and are asked to edit incorrect statements and to add missing statements.

   We then present a number of questions to evaluate the quality of the given note. Our criteria are *Correctness*, *Completeness* (Goldstein et al., 2017), and *Coherence*. We agreed these criteria with the lead physician, who drafted definitions and a scoring guidance for the evaluators (Figure 4). Figure 5 shows the questions we ask for scoring these criteria and a sample annotation.

We also ask evaluators to record their screen for the duration of the task. We use these recordings to calculate how long they took to write the note (step 2) and to edit each generated note (step 3). We use the difference of these two timings as our extrinsic measure to check whether editing a generated note is faster than writing one from scratch.

## 5 Results and Discussion

For this experiment, we run our evaluation on 3 of the 57 mock consultations. Table 2 gives a breakdown of the time it took to edit each note and write one from scratch. Here are some observations:

- In almost all cases, post-editing an existing note is faster than writing a note from scratch;
- As expected, post-editing the reference note (written by the consulting physician) is in general faster than post-editing the notes generated by either model. However, there are a number of instances (across all evaluators) where this isn't the case;
- Note-taking style and length is very different amongst physicians (Cohen et al., 2019), and this can be seen in our results as well. Doctor A tends to write shorter, terser notes and only

**Scoring Guidance**

We are scoring the quality of the note based on:

**Correctness**: you will be asked to identify the number of incorrect statements in the note.

**Completeness**: you will be asked to identify the number of major and minor omissions from the note. If an omission is negligible, please do not include it in the omission count. Here's a description of each omission type:

- **Major** = any edit that would be needed before the consultation notes are completed (if not corrected, it would render the note unsatisfactory from a medico-legal and quality perspective) e.g. features of chest pain
- **Minor** = any edit that would be preferable before the notes are completed (satisfied from a medico-legal point of view but deficient from a quality point of view) e.g. alcohol, smoking hx
- **Negligible** = any edit if missed would not pose any issues but if included would improve the quality of the notes (this is information that you may tend not to record but if you had more time, you might record if you remember) e.g. medication hx which is already recorded elsewhere in the record

**Coherence**: you will be asked if the note makes sense, regardless of the content.

Figure 4: Scoring guidance drafted by the lead physician.

| | Incorrect | | | Major Omissions | | | Minor omissions | | | Coherence | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Source** | **Dr A** | **Dr B** | **Dr C** | **Dr A** | **Dr B** | **Dr C** | **Dr A** | **Dr B** | **Dr C** | **Dr A** | **Dr B** | **Dr C** |
| **Ref** | 0.67 | 2 | 1.67 | 0.33 | 0.67 | 1 | 0.33 | 3 | 0.67 | 2 | 2 | 1.67 |
| **Model A** | 1.67 | 2.33 | 1.33 | 0.67 | 3.67 | 3.33 | 1 | 4.67 | 0 | 2 | 1 | 1 |
| **Model B** | 1.67 | 2.33 | 0.67 | 1.67 | 3.33 | 3.33 | 0.33 | 5 | 1 | 1.67 | 1.67 | 1.33 |

Table 1: Aggregated scores for each evaluator, each criterion, and each note. For a full breakdown along tasks, please refer to Table A1 in the Appendices.

| Task | Eval | Write | Mod A | Mod B | Ref |
|---|---|---|---|---|---|
| **1** | **Dr A** | 2:14 | 1:26 | 0:55 | 1:03 |
| | **Dr B** | 4:02 | 4:30 | 4:16 | 2:44 |
| | **Dr C** | 3:51 | 1:43 | 2:35 | 1:45 |
| **2** | **Dr A** | 4:02 | 0:38 | 1:04 | 0:50 |
| | **Dr B** | 3:19 | 2:31 | 2:51 | 1:43 |
| | **Dr C** | 2:26 | 1:10 | 1:16 | 0:42 |
| **3** | **Dr A** | 4:17 | 1:59 | 2:15 | 0:45 |
| | **Dr B** | 4:21 | 4:04 | 3:32 | 4:17 |
| | **Dr C** | 3:53 | - | - | - |

Table 2: A breakdown of the time taken by the evaluators to write the note from scratch and post-edit each of the generated notes (Mod A, Mod B, Ref). The timings are in M:ss for minutes and seconds taken.

edits the generated notes when there are substantial issues. Doctor B on the other hand is more meticulous and edits the generated notes extensively. This is reflected in both their edit times and their note scoring (see Table 1). We report a detailed view of this disagreement in Figure A1 in the Appendices;

- While it's not feasible to compute correlation between post-editing times and note scores given our sample size, there does seem to be a connection between the two: notes that are scored as containing more omissions and/or incorrect statements take longer to edit. For example, both Dr. B's aggregated scores (Table 1) and edit timings (Table 2) are higher than the other two doctors.
- In one instance, one physician was so frustrated by the quality of a specific generated note that they decided to copy the note they wrote from scratch and paste it instead of trying to edit the generated one. This is why we have missing values in Table 2;
- The first task each physician completed took 36 minutes on average, while subsequent tasks were quicker (23 minutes on average).

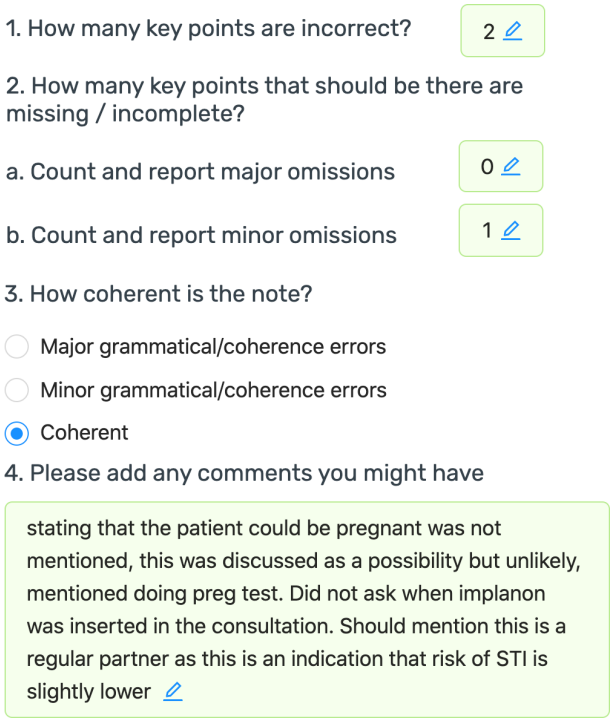


Figure 5: Heartex interface for scoring a generated note.

After watching the recordings and collecting the results, we asked the three evaluators for qualitative feedback regarding the task, the annotation platform, and the generated notes. Here are the key insights we gathered:

- Unlike post-editing, scoring is hard and time-consuming. This is partly due to the interface, which currently doesn't highlight the evaluators' edits on the generated note;
- Familiarity with the interface is key. We shadowed 2 of the 3 physicians through their first few tasks, and that reduced confusion and sped up their work. The physician we did not shadow expressed more difficulty in the evaluation task;
- Our evaluation setup — with physicians asked to listen to a consultation before writing the note — doesn't exactly reproduce the reality of the clinical setting, where they are actually conducting the consultation;
- One physician expressed the worry that even though post-editing a generated note might take less time than writing a note from scratch, it however requires a higher cognitive load. This is because the physician needs to critically read, understand and evaluate the generated note in order to correct it.
- In our experiment, we always ask the evaluators to first write a note from scratch, and then post-edit the generated notes. This specific order may bias our timings. The evaluators may be faster in post-editing after having written the note, or they may be slower if the generated note doesn't follow their style of writing. We plan to address this by shuffling the order of these two tasks.

## 6 Future work

In this paper, we presented our preliminary evaluation study of consultation note generation with post-editing. Based on the insights from this study, we plan to:

- Extend the evaluation to the entire mock consultation dataset and calculate agreement between the evaluators. It would also be interesting to compute agreement between the scores (Correctness, Completeness) and the time taken to post-edit;
- Evaluate the usefulness of auto-generated notes in a live clinical setting;
- Investigate and compare the cognitive load of post-editing notes with that of writing them.

If the issues described in this paper are addressed, we believe post-editing time can be a metric that is both valuable for evaluating model performance and relevant for use in production systems.

Finally, it is important to mention that while automation of medical note taking might help reduce physician burnout and allow the doctors to spend more time with the patients, there are ethical considerations associated to the use of such a technology. For example, time pressures or unwarranted trust in an automated system could potentially result in doctors not properly reviewing and editing the automated notes. Also, post-editing is a very different cognitive task from writing a note from scratch, and that might put extra strain on doctors' already cognitively demanding workflows. In order to mitigate the above concerns in a production system, user experience design, system evaluation, and clinician on-boarding and training are crucially important.

## References

Tod Allman, Stephen Beale, and Richard Denton. 2012. Linguist's assistant: A multi-lingual natural language generator based on linguistic universals, typologies, and primitives. In *INLG 2012 Proceedings*

*of the Seventh International Natural Language Generation Conference*, pages 59–66, Utica, IL. Association for Computational Linguistics.

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.

Wilker Aziz and Lucia Specia. 2012. Pet: a tool for post-editing and assessing machine translation. *LREC*.

Michael Carl, Silke Gutermuth, and Silvia Hansen-Schirra. 2015. Post-editing machine translation. *Psycholinguistic and cognitive inquiries into translation and interpreting*, 115:145.

Ishwar Chander. 1994. Automated Postediting of Documents. In *AAAI-94 Proceedings*, pages 779–784.

Genna R Cohen, Charles P Friedman, Andrew M Ryan, Caroline R Richardson, and Julia Adler-Milstein. 2019. Variation in physicians' electronic health record documentation and potential patient harm from that variation. *Journal of general internal medicine*, 34(11):2355–2367.

Sheila CM De Sousa, Wilker Aziz, and Lucia Specia. 2011. Assessing the post-editing effort for automatic and semi-automatic translations of dvd subtitles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 97–103.

Meghan Dowling, Teresa Lynn, Yvette Graham, and John Judge. 2016. English to irish machine translation with automatic post-editing. In *Proceedings of the second Celtic Language Technology Workshop*, Paris, France.

Ayelet Goldstein, Yuval Shahar, Efrat Orenbuch, and Matan J Cohen. 2017. Evaluation of an automated knowledge-based textual summarization system for longitudinal clinical data, in the intensive care domain. *Artificial intelligence in medicine*, 82:20–33.

Yvette Graham, Qingsong Ma, Timothy Baldwin, Qun Liu, Carla Parra Escartín, and Carolina Scarton. 2017. Improving evaluation of document-level machine translation quality estimation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 356–361.

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.

Maarit Koponen. 2016. Is machine translation postediting worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25:131–148.

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

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.

Medscape. 2018. Medscape national physician burnout depression report 2018.

Sabine Molenaar, Lientje Maas, Verónica Burriel, Fabiano Dalpiaz, and Sjaak Brinkkemper. 2020. Medical dialogue summarization for automated reporting in healthcare. In *International Conference on Advanced Information Systems Engineering*, pages 76–88. Springer.

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)*.

Maxim Tkachenko, Mikhail Malyuk, and Nikolai Liubimov. 2020. Heartex: Data labeling platform for machine learning.

# A Appendices

| | | incorrect | | | major omissions | | | minor omissions | | | coherence | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Task & Source** | **Dr A** | **Dr B** | **Dr C** | **Dr A** | **Dr B** | **Dr C** | **Dr A** | **Dr B** | **Dr C** | **Dr A** | **Dr B** | **Dr C** |
| | **ref** | 2 | 2 | 1 | 1 | 0 | 1 | 0 | 4 | 1 | 2 | 2 | 2 |
| **1** | **model A** | 1 | 2 | 1 | 0 | 3 | 4 | 2 | 6 | 0 | 2 | 1 | 2 |
| | **model B** | 0 | 1 | 1 | 2 | 4 | 4 | 0 | 6 | 1 | 2 | 1 | 1 |
| | **ref** | 0 | 2 | 3 | 0 | 1 | 0 | 0 | 3 | 0 | 2 | 2 | 1 |
| **2** | **model A** | 1 | 2 | 1 | 0 | 3 | 2 | 0 | 5 | 0 | 2 | 1 | 1 |
| | **model B** | 2 | 1 | 1 | 0 | 3 | 2 | 1 | 5 | 0 | 2 | 2 | 1 |
| | **ref** | 0 | 2 | 1 | 0 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 2 |
| **3** | **model A** | 3 | 3 | 2 | 2 | 5 | 4 | 1 | 3 | 0 | 2 | 1 | 0 |
| | **model B** | 3 | 5 | 0 | 3 | 3 | 4 | 0 | 4 | 2 | 1 | 2 | 2 |

Table A1: Scores table.

| **Evaluator 2** | **Evaluator 3** |
|---|---|
| You have been having some problems with your left ear for 3 weeks<br>**Your hearing is muffled on the left side**<br>You noticed that your face has been feeling a bit numb **on the left side of your face, around the ear and the jawline**<br>You have no weakness or numbness in the rest of the body.<br>You have not had any difficulty speaking or swallowing.<br>You have noticed no problems with your vision.<br>The numbness is present on the left side of your face.<br>~~**You have used Cochran spray to clean your ears.**~~<br>~~**You do not have any other illnesses.**~~<br>**No recent fever**<br>You have been told you had polyps in your nose in the past and **occasionally use a prescribe nasal spray to relieve symptoms**.<br>**You have occasional heart burn for which you take over the counter medication.**<br>**You have had labyrinthitis in the past**<br>You are a jockey<br>You get a ringing in your left ear and feel a bit dizzy **which has been affecting your work**<br>You live ~~**alone**~~ with your partner and have no pets at home.<br>**Your brother has neurofibromatosis**<br>**You are allergic to latex** | You have been having some problems with your left ear for the last few days.<br>You noticed that your face has been feeling a bit numb.<br><br>You have no weakness or numbness in the rest of the body.<br>You have not had any difficulty speaking or swallowing.<br>~~**You have noticed no problems with your vision.**~~<br>The numbness is present on the left side of your face.<br>You have used Cochran spray to clean your ears.<br>You do not have any other illnesses.<br><br>You have been told you had polyps in your nose **and labyrinthitis** in the past and occasionally get some kind of funny feeling.<br><br><br>You are a jockey but sometimes you get a ringing and feel a bit dizzy. **Currently this has limited your time in your job.**<br><br>You live alone and have no pets at home. |

Figure A1: Disagreement in editing and scoring a generated note. **Red** marks incorrect statements, **orange** major omissions, and **blue** minor omissions.