Appendix 4

Appendix 4(1)

The number of notes presented to a crowd worker for each task was randomly determined as 1, 3, 5, 7, or 10, and the same number of decomposed dummy and real progress notes were presented to each worker. For example, in Fig 2, three dummy and three real notes were shown to the worker. The notes shown were randomly selected. In each job, each worker answered 50 yes/no questions, which were randomly arranged. Therefore, the total number of tasks amounted to 50 tasks, and 260 real progress notes were presented once to each worker

The details of the definitions of the number of tasks or crowd worker are as follows:

The total number of tasks amounted to 5 tasks $\times$ 10 = 50 tasks, and $(1 + 3 + 5 + 7 + 10) \times 10 = 260$ real progress notes were presented once to each worker.

We note that 50 tasks consist of five types of 10 tasks in each of the following cases: 1 real progress note was presented, 3 real progress notes were presented, 5 real progress notes were presented, 7 real progress notes were presented, and 10 real progress notes were presented.

Appendix 4(2)

We must mention here that the candidate features did not necessarily characterize only the real progress notes. We instructed the crowd workers to describe only yes/no questions, to convert the challenge of generating candidate features into simpler microtasks that they could accomplish. We did not instruct them to indicate whether one type of note differed from the other at a certain point. To be more specific, the workers would not write “Does document group A mention the details about the patient’s concerns, and does document group B exclude them?” Rather, they would write “Do the notes mention the details about the patient’s concerns?” Therefore, it cannot be denied that some of the candidates’ features characterized only the public dummy progress notes.

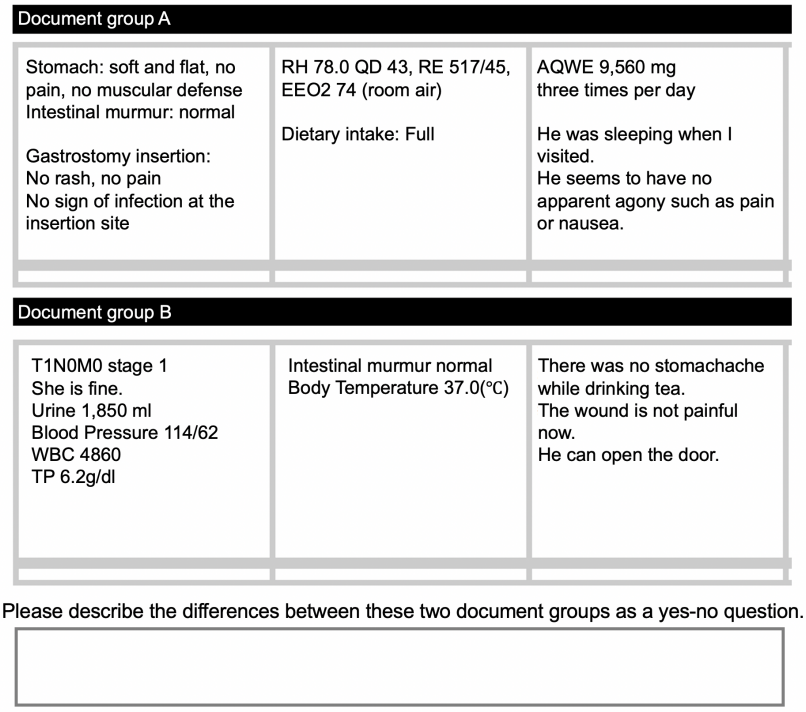

Fig. Example of the web screen for Step 2-a. *Note: In this Figure, three existing dummy and real progress notes were shown to the crowds. These were originally written in Japanese, but are provided in English in this manuscript for explanation purposes. The nonexistent drug name (e.g., "AQWE"), drug dose (e.g., "9560 mg"), and vital signs (e.g., "RH") were mentioned because the character strings were preprocessed for privacy protection*.

Appendix 4(3)

The postprocessing rules for candidate features generated by crowd workers in Step 2-a are as follows:

Features that did not satisfy the yes/no questions were excluded.
(e.g., "no difference," "___ is different.")

The symbols for bullets and blank characters at the beginning of the yes/no questions were deleted.

The duplicate descriptions were omitted. We also excluded features that were the same as other features except for punctuation.
(e.g., "The patient's suffering is recorded." and "The patient's suffering is recorded" were considered as the same, and the latter was excluded.)

"Whether or not ___." " Whether ____." " ____?" → " ____"
(e.g., "Whether or not the documents use the question form." "Whether the documents use the question form." "Do the documents use the question form?"

=> "The documents use question form.")

"There are documents that ___" → "The documents ____"
(e.g., "There are documents that consist only of numbers and alphabets."
=> "The documents consist of numbers and alphabets.")

"Document group A is ___" "Document group A is more ____" → "Document is ____"
(e.g., "Document group A is detailed." "Document group A is more detailed."
=> "Documents are detailed.")

“Are ____described?” => “___ are described.”
(e.g., "Are the numbers described?" → "The numbers are described.")

Appendix 4(4)

(25 progress notes × 150 features × 860 workers) / (100 progress notes × 3,197 features [see Section 3]) = 10.08 workers. Therefore, approximately 10 workers checked each progress note and feature pair.

Appendix 4(5)

The candidate features that characterized few progress notes were considered inappropriate for this study, because such candidate features were considered to express the specific contents of a particular patient (e.g., “Are the details of colon cancer, the history of pneumonia, and an egg allergy described?”). Based on these factors, when the lift threshold was 1, 3, 5, 7, or 9 for the pairs of progress notes and candidate features, it represented the maximum value of the lift for which more than half of the candidate features characterized five or more progress notes. The pair of the lift score under the lift threshold would be excluded.

As results, the pairs of real progress notes and candidate features with an appearance count of one or two were excluded, and 28,487 pairs of real progress notes and candidate features were selected. When the lift threshold was set at 1, 3, 5, 7, and 9, the proportions of candidate features with five or more correlated progress notes among all of those selected were 65.62%, 64.85%, 51.47%, 38.91%, and 28.11%, respectively.

Appendix 4(6)

To define a better network, the Jaccard coefficient was increased from 0.1 to 0.9 in increments of 0.1, and community detection was performed using the Louvain algorithm for each network defined with each Jaccard coefficient. As a result, the determination of the Jaccard coefficient threshold was based on the average number of candidate features in the obtained community. If this number is too large, the obtained communities might include candidate features that are minimally related to each other. This situation would be inappropriate for our purposes, because we wanted to create communities including similar candidate features.

When the Jaccard coefficient was increased from 0.1 to 0.9 (in increments of 0.1) for each community’s detection, the number of extracted communities and the average

number of candidate features included in each community also showed progressive increments. For Jaccard coefficient values of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9, the corresponding numbers of the extracted communities were 8, 12, 16, 31, 176, 176, 184, 184, and 174, respectively, and the corresponding average numbers of candidate features in each community were 269.1, 171.3, 125.3, 59.8, 8.07, 7.43, 5.77, 4.98, and 4.67. Using this result, we assumed that, when the Jaccard coefficient was 0.5 or more, the possibility of the obtained communities including candidate features that had a minimal relation to them would be reduced compared to when the Jaccard coefficient was 0.4 or less.

Appendix 4(7)

The details of the definitions of the number of tasks or crowd worker are as follows:

(105 workers) × 50 communities/176 communities (see Section 3) = 29.8 workers; therefore, approximately 30 workers, on average, worked on each community.

Among 176 communities, seven communities presented with the same number of votes, and one feature was selected randomly.

Appendix 4(8)

The details of the definitions of the number of tasks or crowd worker are as follows:

(500 workers) × 10 progress notes/100 progress notes = 50 workers, therefore, approximately 50 workers, on average, worked on each crowd progress note.

Appendix 4(9)

Long notes inevitably contain a considerable amount of content; that is, they may meet many crowd features known to be appropriate for medical notes. Therefore, we did not use the crowd features that were appropriate as a feature of medical notes as an evaluation metric; instead, we used those that were inappropriate as a feature of medical notes as an evaluation metric.