A practical and universal framework for generating publicly available medical notes of authentic quality via the power of crowds

Rina Kagawa
Faculty of Medicine
University of Tsukuba
Tsukuba, Japan
kagawa-r@md.tsukuba.ac.jp

Yukino Baba
Faculty of Engineering, Information and Systems
University of Tsukuba
Tsukuba, Japan
baba@cs.tsukuba.ac.jp

Hideo Tsurushima
Center for Innovative Medicine and Engineering
University of Tsukuba Hospital
Tsukuba, Japan
hideotsuru@ybb.ne.jp

***Abstract*— Medical notes written by doctors in hospitals or clinics are information-rich. However, in many countries or cultures, few people have access to them for educational and research purposes, even once anonymized. This is because their contents, including patients' disease information, are sensitive and require confidentiality. Therefore, publicly available pseudo-medical notes are needed. Authentic pseudo-medical notes must meet two requirements: (1) medical consistency, and (2) informal descriptions and specific sub-language; however, these are empirical knowledge, even for medical doctors, and are not clarified specifically. We combat this by harnessing the power of crowds. We propose a human-in-the-loop framework for generating publicly available professional medical notes utilizing human cognitive traits with a small dataset. The practical and universal framework has three steps. In Step 1, crowd workers imitated actual notes. In Step 2, crowds and algorithms collaboratively identified notes' characteristics based on comparisons between actual and dummy notes. In Step 3, the texts generated in Step 1 that exhibited the characteristics from Step 2 were evaluated as authentic medical notes that met all requirements. We demonstrated this framework with a total of 1,662 crowds' power. All data were preprocessed to protect patients' privacy before the experiments. The crowds' generated 9,756 notes were evaluated as the most realistic compared to dummy medical notes written by doctors. These crowd-generated medical notes, which are the largest publicly available dataset of Japanese medical notes, are published. This study was the first challenge for the crowds to solve the medical expert-level task.**

Keywords—human capabilities, medical notes, privacy protection

I. Introduction

A. Background and the Necessity of Pseudo-medical Notes

In hospitals or clinics, after a doctor sees a patient, the doctor describes the patient's emotional complaints, their own thinking process leading to diagnosis, future tests, and treatment policies, and the upcoming treatment schedule in narrative texts (Fig. 1). These are called medical notes and are often stored in electronic health records these days. Doctors and clinical professionals require medical notes for help with patient information collection, problem assessments, communication between clinical professionals, clinical management, and administrative purposes [1]. Therefore, natural language processing (NLP) research using medical notes for clinical decision support has been attracting attention [2, 3]. It is also important for students to learn about interprofessional communication through a wide variety of medical notes [4] in clinical education courses.

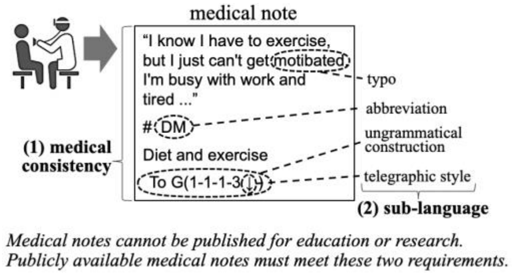
medical note
"I know I have to exercise, but I just can't get motibated, I'm busy with work and tired ..."
#DM
Diet and exercise
To G(1-1-1-3(↓)
typo
abbreviation
ungrammatical construction
telegraphic style
(1) medical consistency
(2) sub-language
Medical notes cannot be published for education or research.
Publicly available medical notes must meet these two requirements.

Fig. 1. Two characteristics of medical notes written by dotctors.

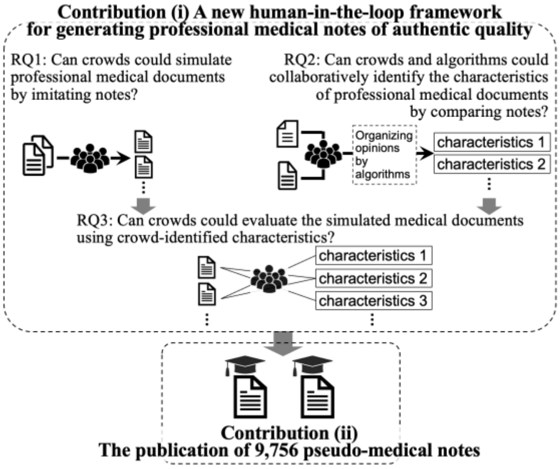
Contribution (i) A new human-in-the-loop framework for generating professional medical notes of authentic quality
RQ1: Can crowds could simulate professional medical documents by imitating notes?
RQ2: Can crowds and algorithms could collaboratively identify the characteristics of professional medical documents by comparing notes?
Organizing opinions by algorithms
characteristics 1
characteristics 2
RQ3: Can crowds could evaluate the simulated medical documents using crowd-identified characteristics?
characteristics 1
characteristics 2
characteristics 3
Contribution (ii)
The publication of 9,756 pseudo-medical notes

Fig. 2. Research question and contributions of this study focusing on human capabilities and cognitive traits.

However, in many countries or cultures [5, 6], only a few limited medical staff members can access medical notes, even for educational or research purposes, because the contents of the notes, including patients' disease information, are sensitive and must be kept confidential, even if identifiable information, such as patients' names, is de-identified. Some anonymized texts in English are freely accessible [7], but these are limited to data from intensive care units (one of the many medical departments or units that exist in hospitals) in the United States. This is because the reproduction of NLP research using the same dataset of medical notes and education using medical notes are virtually impossible. Therefore, pseudo-medical notes that are publicly available independent of the definitions of patient information, even if the definition of "patient information" changes depending on particular countries, cultures, and/or the times, are required for educational and research purposes.

This study was supported in part by the Japan Science and Technology Agency (JST)-Mirai Program (JPMJMI19G8), the Japan Society for the Promotion of Science (JSPS), Grants-in-Aid for Scientific Research (Nos. JP18H06363, JP19K19347), R&D Center for Frontiers of MIRAI in Policy and Technology, and TIA nano KAKEHASHI Tsukuba Innovation Grants.

B. Two Requirements for Pseudo-medical Notes

Generating pseudo-medical notes is difficult. We imposed two requirements relate to empirical knowledge, even for medical doctors, to create a dataset of pseudo-medical notes that are worthy of being used for educational and research purposes (Fig. 1).

(i) The notes should be consistent in terms of medical practice: The pseudo-medical notes must be valid and consistent for each patient in terms of medical practices and relevant healthcare systems to which they are applied; that is, the medical notes for each patient must consistently contain their chief complaints, list of problems, summaries of statuses or laboratory test results, family history, clinical professionals' treatment policies or strategies, and so on.

(ii) The informal descriptions and specific sub-languages characterized in the medical notes should be valid: Some real medical notes are written in sub-languages, such as a telegraphic style of expression (i.e., with displaced or missing words), abbreviations that could have various implications depending on the context [8, 9], ungrammatical constructions, and mixed language use, including Latin terms [10–12], for a wide variety of contents [13], contrary to the ideal description framework [14]. At times, some real medical notes include very narrative descriptions. These characteristics are known to be independent of diseases, skills of the doctors, medical departments, languages, countries, and cultures, but have not been exhaustively researched.

Some pseudo-medical notes exist [4, 15–20], but there are no studies that have fully evaluated the two abovementioned requirements [21]. It is still a great challenge to generate pseudo-medical notes that meet the two requirements.

First, automatic data generation technology is inadequate. Recent years have witnessed the rise of automatic data generation technology, such as the generative adversarial network (GAN) [22–31], GPT-3 [32], automatic generation techniques for images or laboratory tests [33–35], short-length chief complaints (<20 tokens) [36], histories of present illnesses (<40 words) [37], and summaries of each patient's mental health [38]. However, it is difficult to generate long documents that meet the two abovementioned requirements using these techniques. While these techniques can automatically generate documents, they also require large datasets. This is also difficult because privacy protection issues complicate the collection of large-scale medical notes. The challenge entailed in the generation of pseudo-medical notes cannot be solved by applying semantic templates [39].

Second, the two requirements relate to empirical knowledge, even for medical doctors, because it is virtually impossible to view medical notes comprehensively given privacy protection issues. They are difficult to clarify directly (e.g., by using the Delphi method or think-aloud protocol). In addition, the number of texts in the studied datasets was not adequately large for medical research applications using approaches such as natural language processing.

Finally, the specific components of the two requirements are context-dependent, and they vary according to local culture, legal requirements, and medical education. It is virtually impossible to create a dataset of pseudo-medical notes that will fit all contexts globally. We assumed that a flexible framework for generating authentic pseudo-medical notes in a specific culture (e.g., datasets that reflect Japanese culture and legal requirements) would be useful.

C. Contribution of this Study

We propose a framework that harnesses the power of crowds to solve medical expert-level challenges, so that it can be implemented practically and universally with a small dataset, independent of languages, diseases, cultures, and healthcare systems. We focused on two human cognitive traits: **(1) imitation**, which encourages humans without background knowledge to reproduce objects with diverse characteristics [40], and **(2) comparing and contrasting subjects** to highlight similarities and differences and encourage deeper thinking [41]. These human cognitive traits suggested that a large number of medical notes could be simulated by imitation and that wide varieties of informal descriptions and specific sub-languages could be identified by comparison (Fig. 2).

The research questions of this study were as follows.

RQ1: Can crowds simulate professional medical documents by imitating notes?

RQ2: Can crowds and algorithms collaboratively identify the characteristics of professional medical documents by comparing notes?

RQ3: Can crowds evaluate the simulated medical documents using crowd-identified characteristics?

To demonstrate our framework effectively, we focused on crowdsourcing as a human computation platform to solve problems that computers cannot yet clarify [42]. We used general crowdsourcing platforms like Amazon Mechanical Turk to leverage crowds without background knowledge of medicine. This paper makes the following contributions.

(i) A new human-in-the-loop framework is presented for generating authentic professional medical notes.

(ii) The publication of 9,756 pseudo-medical notes, of which about 83% are estimated to be more similar to real medical notes than documents written by medical doctors. These notes are the first freely and the largest publicly available dataset of Japanese medical notes (Table I).

TABLE I. DATASETS CONTAINING JAPANESE CLINICAL TEXTS

Dataset	**Documents**	**Morphologies**	**Available to the public**
Kajiyama+, 2020 [15]	64	154,132	**X**
Aramaki+, 2014 [16]	670	39,589	✔ [43]
Kagawa+, 2021	9,756	2,602,069	✔ [44]

II. RELATED WORKS

Crowds have generated data for wide scientific research [45–49]. In the field of medicine, crowdsourcing has been used for tasks that do not necessarily require in-depth medical knowledge, including surveying methods [50], data processing [51], surveillance or monitoring [52], and problem solving [53, 54]. One challenge remains to be solved in the context of crowdsourcing, namely, understanding how crowds can contribute to solving expert-level tasks, and the generation of medical text data containing detailed medical information. by crowds has not been reported elsewhere.

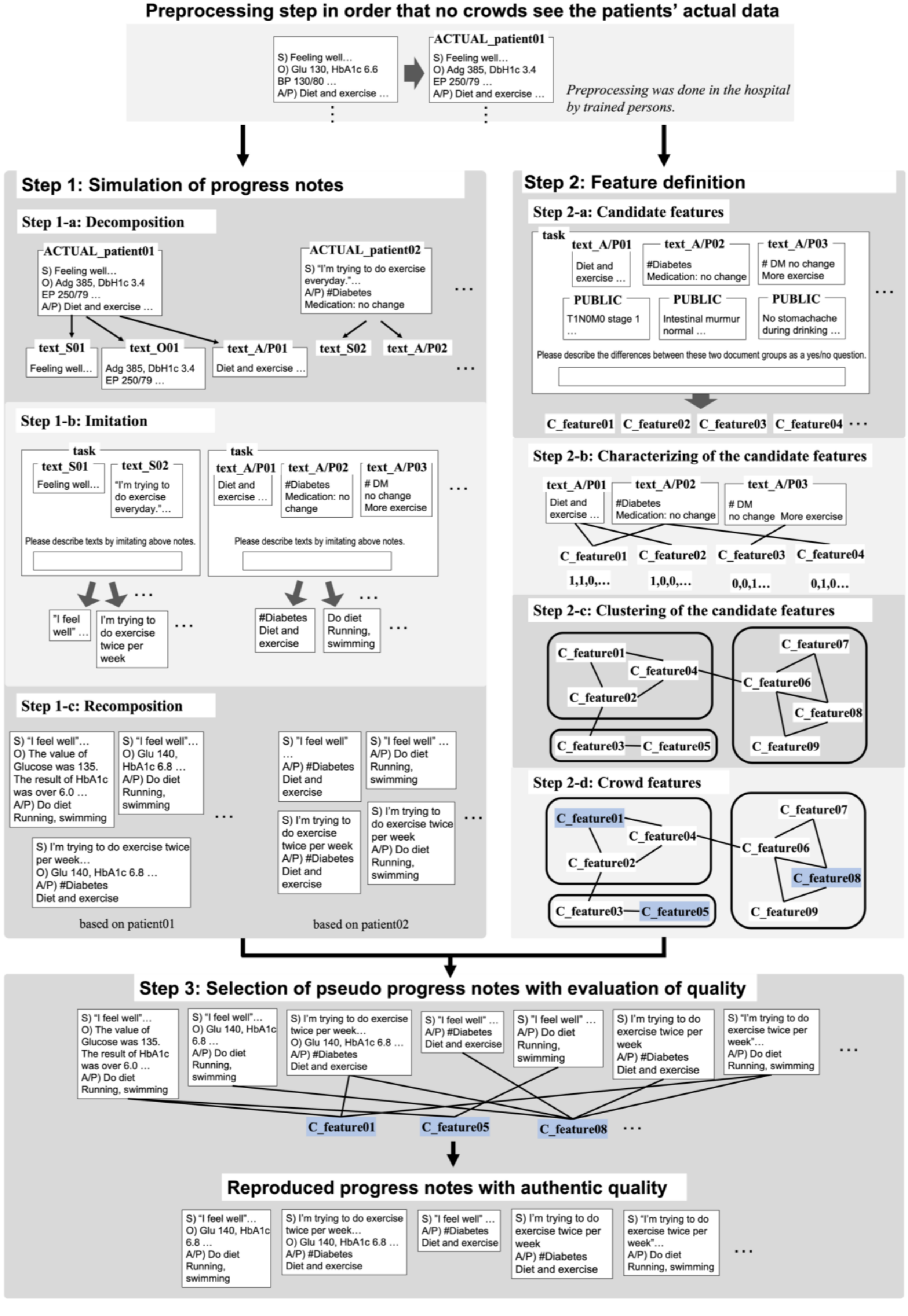

Fig. 3. Overview of our proposed framework. *Note:* ***The fact that the texts were based on medical notes was not revealed to the crowds. In this figure, the number of patinets or texts (e.g., ACTUAL_patient01) are shown for explanation purposes.*** *ACTUAL: Actual medical notes preprocessed for privacy protection, as described in the Appendix 1. Text_S: The decomposed real notes on subjective data. Text_O: The decomposed real notes on objective data. Text_A/P: The decomposed real notes on assessments and plans. PUBLIC: Public dummy medical notes; C_feature: Candidate features.*

III. Datasets

We used medical notes, medication data, laboratory test results, and injection records of 245 Japanese patients (female: 48.2%, mean age: 60.1 years, standard deviation (SD): 16.2 years). These patients were randomly selected from the 53,246 patients who visited any of 11 departments of University of Tsukuba Hospital at least once between January 1, 2013, and September 30, 2018. Detailed settings are shown in Appendix 1[1].

IV. Proposed Framework

We proposed a new framework (Figure 3) consisting of preprocessing to ensure the protection of patient privacy in order that no crowds see the patients' raw data and the three main steps utilizing collaboratively the power of crowds and algorithms. We demonstrated the applicability and effectiveness of this framework with the crowdsourcing platform. Detailed settings of each step of experiments such as how to determine the number of crowd workers, the thresholds of metrics, and postprocessing are shown in Appendix 2.

We used Lancers [55], a commercial Japanese crowdsourcing platform, to process our experiments in Japanese, because most medical notes in Japan are written in Japanese[2]. In this study, for each crowdsourced experiment, a set of multiple microtasks was defined as a "job." Only the crowds who completed all the tasks in each job were paid. The time taken to complete each job was estimated based on the time taken by the first author to complete a pilot run for the job. The wage for each job was designed so that the worker's hourly wage would be approximately US $6.50. The fact that the texts were based on medical notes was not revealed to the crowds in any of the experiments. We recruited a total of 1,662 crowds, 97.3% of whom did not belong to the medical or healthcare field. Among those who were healthcare workers, 0.42% were nurses, 0.30% were pharmacists, 0.10% were healthcare students, and 1.9% worked in other healthcare positions. None of the crowds were medical doctors or medical school students.

A. *Preprocessing of Medical Notes*

Before all the experiments, all data from University of Tsukuba Hospital were preprocessed to ensure the protection of patient privacy (for details, please see Appendix 3). Preprocessing was done in the hospital by the trained persons. In this paper, we refer to these preprocessed medical notes as the "actual medical notes."

B. *Step 1: Generation of Medical Notes Simulated by Crowds via Imitation*

The purpose of this step was to ask crowds to simulate medically consistent medical notes.

1) Step 1-a: Decomposition

The actual medical notes were decomposed into three categories, namely S, O, and A/P according to the SOAP format [14], which is a standard semantic category in the worldwide medical field, in order to turn the challenge of imitating actual medical notes into simpler microtasks that would enable crowds to imitate them. Then, 260 decomposed medical notes on 140 patients were randomly selected. Hereafter, we refer to these 260 notes as the "decomposed real notes."

2) Step 1-b: Imitation

The crowds were asked to write one new text by imitating the decomposed real notes that were displayed. For each task, the decomposed real notes of the same category (S, O, or A/P) and from the same medical department (e.g., cardiology) were shown to the crowd to help them organize the variations in the contents of each decomposed real note. One, two, or three decomposed real notes of each type (such as S and cardiology) were randomly selected to be shown to each worker; it is because if the crowds would have had written one new text by imitating too many of the decomposed real notes shown to them, the texts would have contained contradictory content describing a single patient, and medical validity would not have been guaranteed for that patient. Each worker repeated the above-stated task 20 times. The notes shown to the crowd in each repetition were selected randomly. Given that five workers (on average) checked each task, 35 workers were required and were paid approximately US $5.50 per job.

As a result, all 35 workers responded to all 20 tasks, and 700 imitated texts were generated.

3) Step 1-c: Recomposition

The imitated texts (Step 1-b) based on the same real patient were shuffled and recomposed in the order of S, O, and A/P. Finally, the medical notes were simulated[3]. Not all the actual medical notes included all of the components (S, O, and A/P), and this reality was reproduced by the recomposition of imitated texts. Medical validity and consistency of the recomposed notes were guaranteed by recomposing based on the same patient.

As the details of the numbers in most imitated texts were expected to have changed (e.g., "BP 130" could be changed to "BP 120"; for more information, please see Appendix 3), and most of the recomposed imitated texts were based on the medical notes of multiple patients, the imitated text included descriptions that were not observed in the actual patients. This ensured better protection of patient privacy.

Finally, 9,856 simulated medical notes were generated.

C. *Step 2: Characterization of the Medical Notes*

Here, crowds and algorithms collaboratively identified the characteristics of the actual medical notes, namely informal descriptions and specific sub-languages.

1) Step 2-a: Candidate features by crowds

The crowds compared and differentiated actual medical notes and existing dummy medical notes [16] (details are shown in Appendix 1), enabling crowds to generate the features that characterize actual medical notes or existing dummy medical notes [56].

The 260 decomposed real notes defined in Step 1-a were used. For each task, the decomposed real and dummy medical notes of the same category (S, O, or A/P) and from the same

[1] All appendix files are shown in https://github.com/rinabou_k/HMData2021.

[2] All texts are shown in English in this manuscript for explanation purposes.

[3] For example, an imitated text containing S data based on two S texts of patients 01 and 02, an imitated text containing O data based on one O text of patient 01, and an imitated text containing A/P data based on three A/P texts of patients 01, 02, and 03 were recomposed because these three imitated texts were based on the same patient (patient 01). If any of the S, O, or A/P texts were missing for a patient (e.g., the imitated O texts based on patient 02 were missing), only the existing imitated texts were recomposed.

medical department (such as cardiology) were shown to the crowds, and they differentiated between these two types of notes by writing descriptions as yes/no questions (e.g., “Do the notes mention details about the patient’s concerns?”). In each job, each worker repeated this 50 times, and notes presented to the worker were randomly arranged for each task. In all, 100 workers were sought for the project, but only 72 applied within one week, after which the recruitment was discontinued. The workers were paid approximately US $5.00 for each job. We must mention here that the candidate features did not necessarily characterize only the decomposed real notes. The obtained yes/no questions were post-processed by omitting duplicate descriptions and so on. Then, the remaining yes/no questions were defined as candidate features.

A total of 3,938 yes/no questions were created. After postprocessing, 3,197 candidate features were ultimately used in the subsequent experiment.

2) Step 2-b: Characterizing of the candidate features

For clustering the candidate features in the subsequent analysis, the candidate features were characterized by decomposed real notes related to the candidate features utilizing collaboratively the power of crowds and algorithms.

First, crowds checked the relationships between the candidate features and the decomposed real notes. For each task, a decomposed real note and 150 candidate features were presented to a worker who then selected an arbitrary number of candidate features that they believed characterized the decomposed real note displayed. In each job, the worker repeated this task 25 times. In each repetition, the decomposed real note displayed was randomly selected from 100 decomposed real notes, and the 150 candidate features were also randomly selected from all the candidate features collected in Step 2-a. For an average of 10 workers to check the pairing of one decomposed real note and one candidate feature, 850 workers were paid approximately US $4.25 for each job.

As a result so far, 126,200 pairs of candidate features and decomposed real notes were created, and we obtained 49,430 unique pairs.

Then, selection of the pairs of candidate features and decomposed real notes was accomplished automatically. The pairs with an appearance count of either one or two were excluded. In addition, we tried to extract the pairs of candidate features and decomposed real notes, even if the frequency of their appearance was low. To achieve this extraction, we set a threshold for the lift (Equation (1)) [57] and excluded the pairs below this threshold. f represents each candidate feature, r represents each decomposed real note, and $P(f, r)$ represents the ratio of the number of the pairs of f and r compared to the number of all pairs of candidate features and decomposed real notes. $P(f)$ represents the ratio of the number of pairs of f and any decomposed real note, and $P(r)$ represents the ratio of the number of the pairs of r and any candidate feature.

$$lift(f, r) = {P(f, r)}\Big/{P(f)P(r)} \qquad (1)$$

The threshold of lift was set to five, and 5,723 pairs of candidate features and decomposed real notes were selected.

3) Step 2-c: Clustering of candidate features

Some candidate features were semantically similar and were aggregated in this step via clustering based on their similarities using network analysis. Each cluster is considered to contain the characteristics of the medical notes.

Similarities between candidate features were determined by calculating the similarity of sets of decomposed real notes that characterized each candidate feature. The Jaccard coefficient was used for this task. If the Jaccard coefficient between the candidate features exceeded the Jaccard coefficient threshold, the candidate features shared an edge. The Jaccard coefficient threshold was set to 0.5, and the network of the candidate features was defined. Then, the Louvain algorithm [58] was used to cluster the network.

As a result, 176 communities were extracted.

4) Step 2-d: Selection of crowd features

The content of each candidate feature community created in the experiments thus far was considered to represent the characteristics of decomposed real notes. The crowds identified the most representative feature for each cluster as a label. The labels of each cluster were accounted for to include the characteristics of the informal descriptions and specific sub-languages of the actual medical notes.

For each community, the workers selected the most comprehensive candidate feature from the ones included in the community. The candidate features with the highest number of votes within each community were adopted as the community’s name. These features were referred to as “crowd features.” Approximately 30 workers, on average, worked on each community, 105 workers worked on 50 randomly selected candidate feature communities, resulting in each job costing approximately US $4.25.

Finally, 176 crowd features (Appendix 4(1)) were created.

TABLE II. EXAMPLES OF CROWD FEATURES AND CRC-NOMINATED FEATURES.

Crowd features	**CRC-nominated features**	**Examples**
✓	✓	The patient’s suffering is recorded. The treatment schedule is recorded. Specific details of the patient’s condition are recorded. Detailed interview results for each symptom are written down.
✓	X	Records related to blood pressure are present. Information about awareness of hypoglycemia is recorded. English and Japanese words are mixed together in the descriptions of vital sign items. Half-width numbers and English letters are used for order and readability.
X	✓	The doctor’s opinion on the interview should also be stated. The descriptions do not reveal the doctor’s opinion or guidance. The statements contain only minimal information and do not express the clinician’s views. As a statement of opinion is missing, we do not know what the symptoms indicate.

Note: In Japanese, two types of characters exist for numbers and alphabets: half-width and full-width. Half-width characters present a horizontal to vertical length ratio of 1:2.

5) Evaluation of the validity of crowd features

Evaluation A: Three clinical research coordinators (CRCs) who read medical notes as part of their jobs and who had at least three years of experience were asked whether they thought that some medical notes satisfied each crowd feature using a five-point Likert scale ranging from “medical notes that satisfied the feature do exist” (+2) to “medical notes that satisfied the feature do not exist” (−2). The reason for using such questions was that we assumed that few professionals could answer a direct question such as “Is this feature a valid characteristic of medical notes?” because the characteristics of actual notes were medical doctors’ empirical knowledge.

The average score of each feature ranged from −0.67 to 1.67 (average: 0.61, SD: 0.48). Of the 176 crowd features, 10, 20, and 146 features had negative, zero, and positive average values, respectively. From this result, we judged that the 10 crowd features with negative average values did not realistically reflect the characteristics of actual medical notes.

Evaluation B: Three CRCs with more than three years of experience were shown only the decomposed real notes used in Step 2-a and then asked to describe the features that characterized the notes shown. A total of 165 features were described. Duplicate features were omitted like as Step 2-a, and 105 unique features (Appendix 4(2)) were obtained.

CRC-nominated features and crowd features were compared qualitatively, and they showed no obvious differences. However, some differences were identified. The CRCs noted what they believed should have been described in the medical notes but had not been in reality (e.g., “The doctor’s opinion on the interview should also be stated.”), but the crowds did not. Conversely, the crowds noted notations (abbreviations, etc.) and specific conditions (hypoglycemia, etc.), but the CRCs did not (Table II).

From these two evaluations, we judged that there are no obvious differences between crowd features and the characteristics of medical notes CRC could identified.

D. Step 3: Selection of Pseudo-medical Notes

In Step 3, crowds checked whether the pseudo-medical notes generated by the crowds obtained in Step 1 exhibited the characteristics identified in Step 2, and only those texts with the aforementioned characteristics were judged as “quality-guaranteed medical notes” that met the two requirements. For the selection, we used 100 randomly selected notes from the pseudo-medical notes generated by the crowds in Step 1.

For each task, the notes generated in Step 1 and all crowd features were presented to a worker who then selected an arbitrary number of crowd features the worker believed characterized the displayed note. The worker repeated this task 10 times for each job. During each repetition, one note displayed was randomly selected from 100 notes, and all crowd features were arranged and presented randomly. On average, approximately 50 workers checked each note. The workers were paid approximately US $5.50 per job. Crowd features with negative average values assessed by the CRCs in the *Evaluation A* after Step 2-d, were deemed inappropriate as a feature of the medical notes. The notes judged by more than 25 workers as being characterized by any inappropriate crowd feature were deemed unsuitable for publication. The reason for using such criteria is described in Appendix 2.

Of 100 randomly selected notes from the 9,856 generated in Step 1, 17 were judged to be unsatisfactory. The remaining 83 were published as clinically consistent pseudo-medical notes that exhibit the characteristics of informal descriptions and specific sub-languages found in actual medical notes, which we refer to as “crowd medical notes.” (Table III) [43].

TABLE III. EXAMPLES OF PUBLISHED PSEUDO-MEDICAL NOTES JUDGED TO BE AUTHENTIC AND THOSE JUDGED TO BE INAUTHENTIC (UNPUBLISHED).

Published	**Unpublished**
(S) Feeling good. The stress is better. I started swimming once a week and taking a 10-minute walk around the pool. I think I’m getting lighter. (O) Estimated salt intake: 6.11 g/day Weight: 66.7 kg Skeletal muscle: 23.2 kg Skeletal muscle mass: 25.4 kg No awareness of hypoglycemia. 74-199-177- (A/P) Three units of insulin glargine were started last night, and this morning, hypoglycemia occurred. Insulin glargine will be reduced to one unit. In addition, the patient will be hospitalized in the surgery department until his blood glucose control stabilizes. We will ask the patient to be examined in conjunction with our department to continue blood glucose control. When discharge from the hospital is expected, we will contact the surgery department. To Q(1-1-1) G(0-0-0-1↓) The patient’s life is stressful, and it is difficult to coordinate nutrition and exercise. Vitamin B12 = numbness	(S) Looks good. “I’m fine. No change.” “I smoke two cigarettes a week.” I advised him to quit smoking. (O) He could barely eat, and his blood sugar level was somewhat high. At the time of admission, she weighed 93.5 kg, but she has now reduced to 89.0 kg. An echocardiogram showed decreased wall motion in the apex of the heart, but there was no significant change compared to September 29, 1991. (A/P) Before the PK surgery, we used the Lap-DP manual to explain if it was possible to use Lap-DP. As for laparoscopy, he explained that consultation was necessary, but there was no hope. Dialysis, consultation with nephrology. Call me when the surgery is scheduled, and we’ll coordinate the dialysis. After PCI, Cardiology Consultant Scheduled for Aug. 21. Lap-DP scheduled for the end of this month. 6/14 Lap-DP consultation with Dr. Tanikawa. Blood test/X-rays were fine, and he is ready to be discharged.

V. Evaluation

A. Overview of the Evaluation

To evaluate the crowd medical notes, we quantitatively and qualitatively assessed the similarities between the crowd medical notes and the actual medical notes. We evaluated whether the **crowd** medical notes generated using our framework were more similar to the **actual notes** than the existing **public** dummy medical notes [16] and the notes medical **doctors** wrote based on actual patients. In particular, we focused on whether the crowd medical notes generated using our framework were more similar to the actual notes than the notes medical doctors wrote based on actual patients.

For the evaluation, two doctors with at least five years of clinical experience wrote dummy medical notes for 10 dummy patients on an experimental computer screen because the environment in which the public dummy medical notes [16] were created was not shown. Data of 10 dummy patients were generated from actual patients in University of Tsukuba Hospital that had not been used in previous experiments using the experimental computer screen, which was similar to the one used for Japan's actual electronic health records (Appendix 5(1)). One medical doctor viewed the patient information for one dummy patient, but skipped writing the corresponding dummy medical note; therefore, only 19 dummy medical notes were created. We used these dummy medical notes by medical doctors for comparison purposes.

B. Methods of Quantitative and Qualitative Evaluation

We compared the 257 existing dummy medical notes, 19 dummy medical notes written by the medical doctors, 83 crowd medical notes, and 100 randomly selected actual medical notes. The 100 actual medical notes had not been used in other experiments.

Quantitative Evaluation: The quantitative evaluation was based on three viewpoints: the number of standard **disease** names [59] based on the International Classification of Diseases (ICD)-10 codes described in each note; the number of **morphologies** described in each note; and the **readability** of each note. For the analysis of the morphologies, we used a popular Japanese morphological analyzer, MeCab-0.996 [60], and the mecab-ipadic-2.7.0 dictionary [61]. The readability scores based on the T-13 model was ranged from 1 to 13, with a high score signifying that the text was difficult to read [62].

Qualitative Evaluation: The **absolute reality** of the notes was evaluated. For the four types of medical notes, four medical doctors with at least 10 years of experience and who did not participate in the other experiments in this study were asked whether or not they thought that each note was an actual medical note. They were asked to use a 10-point Likert scale in their evaluation (10: I think that this note is an actual medical note–1: I don't think that this note is an actual medical note). To reduce the psychological and time burden for the four doctors, each was only asked to answer questions on 19 randomly selected notes from the four note types; thus, each doctor answered questions on 76 notes in total.

The **relative reality** of the notes was also evaluated by medical doctors using more complex questions. Ten medical doctors who did not participate in the other experiments in this study were shown two notes; one was the actual note, and the other was one of the four types of medical notes. The doctors were asked "One is a real note and another is a fake note. Which do you think is the real note?" For each note, the percentage of doctors who correctly chose the real note was evaluated. 0.5 means that the real note and the other note couldn't be distinguished at all. A higher value from 0.5 means that the doctor correctly evaluated the actual note as an actual note. To reduce the psychological and time burden for the four doctors, each was only asked to answer questions on 12 randomly selected texts from the four note types; thus, each doctor answered questions on 48 pairs of notes in total.

Analysis: We compared the average of each metric between the actual medical notes and the other three types of medical notes. The details of significant tests are shown in Appendix 5(2). We calculated the Kullback–Leibler (KL) divergence of the distribution of each metric of the actual medical notes from that of each of the three other types of medical notes to analyze the discrepancies between them. $p(x)$ refers to the distribution of each metric of the actual notes, and $q(x)$ denotes the distribution of each metric for each of the three other types of medical notes. When $p(x)$ or $q(x)$ was zero, 1.0×10^{-5} was inserted to $p(x)$ or $q(x)$.

C. Results

The means and SDs of the scores of each metric are shown in Table IV. The mean score of crowd medical notes was the nearest to that of the actual medical notes for all five metrics. As shown in Table V and Appendix 5(3), for the crowd medical notes only, the null hypothesis that the median score would be equal to the median score of the actual medical notes was **not** rejected for all five metrics. The average score of relative reality was 0.417; this was assumed to be due to that the number of notes used in this experiment was small. Additionally, KL divergences from the actual medical notes were the smallest for the crowd medical notes for the two metrics. One medical doctor (the first author) also qualitatively confirmed that the crowd medical notes did not include any numerical values that could never occur clinically (e.g., "Glucose -20 mg/dL").

Therefore, the crowd medical notes were judged to be the most similar to the actual medical notes.

TABLE IV. AVERAGE (SD) OF EACH METRIC SCORE

	Diseases	**Morphemes**	**Readability**	**Absolute Reality**	**Relative Reality**
Crowd	45.18 (37.96)	164.04 (106.94)	10.22 (2.39)	5.59 (2.18)	0.567 (0.210)
Doctors	30.63 (40.21)	88.89 (108.40)	8.05 (3.15)	5.09 (2.51)	0.625 (0.160)
Public	23.56 (28.47)	89.54 (88.75)	9.70 (2.74)	5.01 (2.34)	0.658 (0.168)
Actual notes	60.48 (76.89)	178.34 (205.22)	10.29 (2.61)	6.76 (1.92)	0.417 (0.279)

TABLE V. COMPARISON OF THE THREE TYPES OF MEDICAL NOTES AND THE ACTUAL MEDICAL NOTES. **NOTES WITH P-VALUE ≥ 0.001 WERE JUDGED NOT TO BE SIGNIFICANTLY DIFFERENT FROM ACTUAL MEDICAL NOTES FOR EACH METRIC.**

	Diseases	**Morphemes**	**Readability**	**Absolute Reality**	**Relative Reality**
Crowd	***p* = 0.534** **KL = 0.244**	***p* = 0.36** **KL = 0.115**	***p* = 0.536** KL = 0.633	***p* = 0.001** KL = 1.165	***p* = 0.151** KL = 5.241
Doctors	***p* = 0.022** KL = 0.608	***p* = 0.003** KL = 0.990	*p = 0.0009* KL = 2.648	*p = 1.8e−05* KL = 0.432	***p* = 0.035** **KL = 5.057**
Public	*p = 1.9e−10* KL = 0.385	*p = 1.3e−08* KL = 0.312	***p* = 0.028** **KL = 0.098**	*p = 6.5e−07* **KL = 0.423**	***p* = 0.017** KL = 7.053

Note: Italics p-value means that p-value < 0.001. Bold p-value means that p-value ≥ 0.001. Bold KL means that the smallest KL divergence for each metrics. Crowd: Crowd medical notes. Doctors: Dummy medical notes created by two medical doctors. Public: Public dummy medical notes.

VI. Discussion

We proposed a new practical and universal framework for creating quality-guaranteed medical notes that are clinically consistent and exhibit the characteristics of informal descriptions and specific sub-languages. We did this via the power of crowds and algorithms, using appropriate microtasks designed in accordance with humans' cognitive traits and a small dataset. Crowds simulated the medical notes by imitation, identified characteristics of medical notes based on a comparison between actual and dummy medical notes in human machine collaboration protocols, and evaluated the medical notes. This was achieved in a manner independent of document types, languages, diseases, medical departments, countries, cultures, and healthcare systems. The medical notes generated by the crowds based on our proposed framework were judged to be the most similar to real medical notes, compared to dummy medical notes written by medical doctors and the existing public dummy medical notes. This may seem unsurprising because crowds imitated actual medical notes, but it is novel to use crowdsourcing to realize medical expert-level tasks. Moreover, 9,756 notes that were obtained in Step 1 but not evaluated in Step 3, and 19 dummy medical notes created by doctors for evaluation have also been published[4].

A. *Generalizability and Limitations of the Proposed Framework*

Our proposed framework is generalizable; it is independent of languages, diseases, and other characteristics. Our framework is advantageous, as pseudo-medical notes can be generated based on a small number of real medical notes. We demonstrated our framework using the power of crowds, but other techniques could be used for each step; for example, an automatic data generation technology could be applied for Step 1 in the future. Our framework is widely applicable.

The limitation of our proposed method is that the framework cannot extract features that are considered inappropriate for real medical notes (e.g., a simple English word like drug or heart and the Japanese translation of that word written together in one sentence). This is considered inappropriate for real medical notes, because clinical professionals are familiar with such English words, but such occurrences have been found in existing public dummy medical notes [16]. While our framework cannot extract these features, such descriptions were not found in the crowd medical notes; it might be that the imitation step prevented these inappropriate descriptions.

B. *Generalizability and Limitations of Crowd Medical Notes*

This study showed that crowd medical notes could reflect unique writing styles, which, although not necessarily grammatically perfect, are actually used in practice. "Vitamin B12 = numbness" is one example (see the bottom of the left column in Table III); this entry can be interpreted as follows: "Vitamin B12 was prescribed to treat numbness." Such examples were not found in the existing public dummy medical notes [16], but crowd-generated pseudo-medical notes reflected the reality of the professional medical notes.

The limitation is that unforeseen problems related to the use of Japanese or specific characteristics of the healthcare system in Japan could emerge. Another limitation is that our generated medical notes are not appropriate in research aimed at making new medical discoveries, such as epidemiological studies.

C. *Crowdsourcing to Solve Medical Expert-level Tasks*

Seeing patients and documenting their information is only allowed for medical doctors engaged in daily clinical practice. As a result, writing medical notes is viewed as an "empirical knowledge monopoly" held by doctors, and we assumed that it would be difficult for nonmedical crowds to generate authentic notes. However, our pilot study unveiled the crowd's ability to solve medical-expert level empirical tasks using the appropriate microtasks designed in accordance with human cognitive skills. While questions surrounding the quality of the data generated by crowds have been raised [63, 64], 83 of 100 crowd medical notes met the characteristics of the real medical notes in this work. Compared to the pseudo-medical notes written by medical doctors or existing public dummy medical notes, the notes generated by our framework were judged to be the most similar to real medical notes.

In addition to the above, we should mention that the sub-language characteristics of all possible medical notes in the world are not covered for the extraction of the sub-language characteristics in Step 2. Since it is virtually impossible to view real medical notes (because of privacy protection concerns), we comprehensively acquired the characteristics unique only to the documents available to us; despite this limitation, to the best of our knowledge, this is the first challenge of unveiling the variety of sub-language characteristics in real medical notes.

D. *Future Work*

If our framework was demonstrated by medical doctors instead of nonmedical crowds, pseudo-medical notes more similar to real medical notes could be generated; to demonstrate this is the future work. The sharing of data is necessary to ensure that these data become common social capital [65]; at the same time, each person's privacy must be protected. In the future, we hope that published medical notes are created for medical education or research purposes in many languages, or for various healthcare systems, using our framework.

VII. Conclusion

We proposed a framework that created quality-guaranteed medical notes that were clinically consistent and exhibited the characteristics of informal descriptions and specific sub-languages found in real medical notes by using the power of crowds. This was achieved in a manner independent of languages and diseases. The notes generated by the crowds based on our proposed framework were judged to be the most similar to real medical notes. This study also showed that crowds are able to simulate real medical notes and can identify the characteristics of informal descriptions and specific sub-languages found in these medical notes. Our results open new avenues for solving problems that require in-depth medical knowledge by integrating the contributions of many crowd workers using appropriate microtasks.

Acknowledgment and Ethics

Makoto Ohara, Toshihiko Kawaguchi, Masako Sato, and Mika Sumimoto contributed to the preprocessing step for privacy protection. This research was approved by the Ethics Committee of the University of Tsukuba Hospital (permission

[4] Publicly available medical text data with authentic quality [Internet]. Available from: http://doi.org/10.5281/zenodo.4064153

number: H30-145). Details regarding ethics are shown in Appendix 3.

REFERENCES

[1] Burke HB, Sessums LL, Hoang A, Becher DA, Fontelo, P, Liu F, et al. Electronic health records improve clinical note quality. J Am Med Inform Assoc. 2015;22(1):199-205.

[2] Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resources & Evaluation 2020;54**:** 57–72.

[3] Chen P, Wang S, Liao W, Kuo L, Chen K, Lin Y, et al., Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning. JMIR Med Inform. 2021;9(8):e23230

[4] Conn LG, Lingard L, Reeves S, Miller K, Russell A, Zwarenstein M. Communication channels in general internal medicine: a description of base-line patterns for improved interprofessional communication. Qual Health Res. 2019;19(7):943e953.

[5] Lohr C, Buechel B, Hahn U. Sharing copies of synthetic clinical corpora without physical distribution—a case study to get around IPRs and privacy constraints featuring the German JSYNCC corpus. Proceedings of the 11th International Conference on Language Resources and Evaluation; 2018 May 7-12; Miyazaki, Japan. Paris: European Language Resources Association; 2018. pp. 1259-66.

[6] Grabar N, Dalloux D, Claveau V. CAS: corpus of clinical cases in French. J Biomed Semant. 2020;11(1):1-10.

[7] Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3(1):1-9.

[8] Aramaki A, Imai T, Miyo K, Ohe K. Robust classification for ungrammatical and fragmented texts. Proceedings of the 23rd IEEE International Conference on Data Engineering Workshops; 2007 Apr 15-20: Istanbul. Washington, DC: IEEE; 2007. pp. 195-201.

[9] Kagawa R, Shinohara E, Imai T, Kawazoe Y, Ohe K. Bias of inaccurate disease mentions in electronic health record-based phenotyping. Int Journal of Med Inform. 2019;124(1):90-6.

[10] Grigonyte G, Kvist M, Velupillai S, Wirén M. Improving readability of Swedish electronic health records through lexical simplification: first results. European Chapter of Assoc Comput Linguist. 2014;124:26-30.

[11] Bretschneider C, Zillner S, Hammon M. Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach. Proceedings of the 2013 Workshop on BioNLP-ST; 2013 Aug 8-9; Sofia, Bulgaria. Madison, Wisconsin: Omnipress, Inc.; 2013. pp. 27-35.

[12] Savkov A, Carroll J, Koeling R, Cassell J. Annotating patient clinical records with syntactic chunks and named entities: the Harvey Corpus. Lang Resources & Evaluation. 2016;50(3):523-48.

[13] Moon S, Pakhomov S, Liu N, Ryan JO, Melton GB. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. J Am Med Inform Assoc. 2014;21(2):299-307.

[14] Weed LL. Medical records, medical education and patient care: the problem oriented record as a basic tool. Cleveland (OH): Press of Case Western Reserve University; 1970.

[15] Kajiyama K, Horiguchi H, Okumura T, Morita M, Kano Y. De-identifying free text of Japanese dummy electronic health records. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 31 - Nov 4; Brussels, Belgium. Brussels: Association for Computational Linguistics; 2018. pp. 65-70.

[16] Aramaki E, Morita M, Kano Y, Ohkuma T. Overview of the NTCIR-11 MedNLP-2 task. Proceedings of the 11th NTCIR Conference; 2014 Dec 9-12; Tokyo. Toyko: NII; 2014. pp. 147-54.

[17] Libbi, CA, Trienes, J. Trieschnigg, D, Seifert, C. Generating synthetic training data for supervised de-identification of electronic health records. Future Internet 2021;13(5):1-24.

[18] Li J, Zhou Y, Jiang X, Natarajan K, Pakhomov SV, Liu H, et al. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition, J Am Med Inform Assoc. 2021;28(10):2193–2201.

[19] Amin-Nejad A, Ive J, Velupillai S. Exploring Transformer Text Generation for Medical Dataset Augmentation. Proceedings of the 12th Language Resources and Evaluation Conference: Association for Computational Linguistics; 2020. pp. 4699-4708.

[20] Melamud O. and Shivade C. Towards automatic generation of shareable synthetic clinical notes using neural language models. In Proceedings of the 2nd Clinical Natural Language Processing Workshop; 2019. pp. 35–45.

[21] Kagawa R, Shinohara E, Imai T, Kawazoe Y, Ohe K. Requirements for creating a tagged corpus of clinical notes based on the tacit knowledge of medical doctors. Proceedings of the 24th Annual Meeting of the Association for Natural Language Processing; 2018 Mar 12-16; Okayama, Japan. Tokyo: Association for Natural Language Processing. pp. 757-60.

[22] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S. Generative adversarial nets. Proceedings of the 21st International Conference on Neural Information Processing Systems; 2014 Nov 3-6; Montreal, Canada. Cambridge: MIT Press; 2014. pp. 2672-80.

[23] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R. Show, attend and tell: neural image caption generation with visual attention. Proceedings of the 32nd International Conference on Machine Learning; 2015 Jul 7-9; Lille, France. Lille: Proceedings of Machine Learning Research; 2015. pp. 2048-57.

[24] Liu PJ, Saleh M, Pot E, Goodrich B, Sepassi R, Kaiser L, et al. Generating Wikipedia by summarizing long sequences. Proceedings of the 6th International Conference on Learning Representation; 2018 Apr 30 – May 3; Vancouver BC, Canada. Vancouver: International Conference on Learning Representations; 2018. pp. 1-18.

[25] Radford A, Jozefowicz R, Sutskever I. Learning to generate reviews and discovering sentiment. Proceedings of the 6th International Conference on Learning Representation; 2018 Apr 30 – May 3; Vancouver BC, Canada. Vancouver: International Conference on Learning Representations; 2018. pp. 1-9.

[26] Wang K, Wan X. Sentigan: generating sentimental texts via mixture adversarial networks. Proceedings of the 27th International Joint Conference on Artificial Intelligence; 2018 Jul 13-19; Stockholm, Sweden. Freiburg: International Joint Conferences on Artificial Intelligence, 2018. pp. 4446-52.

[27] Zhang H, Gong Y, Yan Y, Duan N, Xu J, Wang J, et al. Pretraining-based natural language generation for text summarization. Proceedings of the 23rd Conference on Computational Natural Language Learning; 2019 Nov 19-20; Hong Kong, China. Hong Kong: Association for Computational Linguistics; 2019. pp. 789-97.

[28] Budzianowski P, Vuli'c I. Hello, it's gpt-2–how can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3-7; Hong Kong, China. Hong Kong: Association for Computational Linguistics; 2019. pp. 15-22.

[29] Li D, Zhang Y, Gan Z, Cheng Y, Brockett C, Sun MT, et al. Domain adaptive text style transfer. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3-7; Hong Kong, China. Hong Kong: Association for Computational Linguistics; 2019. pp. 3304-13.

[30] Prabhumoye S, Tsvetkov Y, Salakhutdinov R, Black AW. Style transfer through back-translation. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018 Jul 15-20; Melbourne, Australia. Melbourne: Association for Computational Linguistics; 2018. pp. 886-76.

[31] Rao S, Tetreault J. Dear sir or madam, may I introduce the YAFC corpus: corpus, benchmarks and metrics for formality style transfer. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics; 2018 Jun 1-6; New Orleans, USA. New Orleans: Association for Computational Linguistics; 2018. pp. 3168-80.

[32] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv:2005.14165; 2020

[33] Choi E, Siddharth B, Bradley M, Jon D, Walter FS, Jimeng S. Generating multi-label discrete patient records using generative adversarial networks. Proceedings of the 2nd Machine Learning for Healthcare Conference; 2017 Aug 18-19; Boston, USA. Boston: Proceedings of Machine Learning Research; 2017;68(1):286-305.

[34] Lombardo JS, Moniz LJ. A method for generation and distribution of synthetic medical record data for evaluation of disease-monitoring systems. Johns Hopkins APL Technical Digest. 2008;27(4):356-65.

[35] Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018;25(7):230-8.

[36] Lee SH. Natural language generation for electronic health records. NPJ Digit Med. 2018;1(1):1-7.

[37] Guan J, Li R, Yu S, Zhang, X. Generation of synthetic electronic medical record text. Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine; 2018 Dec 3-6; Madrid, Spain. Washington, DC: IEEE; 2018. pp. 374-80.

[38] Ive J, Viani N, Kam J, Yin L, Verma S, Puntis S, et al. Generation and evaluation of artificial mental health records for natural language processing. NPJ Digit Med. 2020;3(1):1-9.

[39] Begoli E, Brown K, Srinivas S, Tamang S. SYNTHNOTES: A generator framework for high-volume, high-fidelity synthetic mental health notes. IEEE Big Data. 2018 Dec 10-13; Seattle, USA. Seattle: IEEE; 2018. pp. 951-8.

[40] Kuniyoshi Y. The science of imitation—towards physically and socially grounded intelligence. RWC Tech Rep 1994;1(1):95-96.

[41] Gentner D, Markman AB. Structure mapping in analogy and similarity. Am Psychol 1997;52(1):45-56.

[42] von Ahn L. Human computation [master’s thesis]. Pittsburgh (PA): Carnegie Mellon University; 2005.

[43] Dummy electronic health record text data [Internet]. Available from: https://www.gsk.or.jp/en/catalog/gsk2012-d.

[44] Publicly available medical text data with authentic quality [Internet]. Available from: http://doi.org/10.5281/zenodo.4064153

[45] Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: challenges and opportunities. Brief. Bioinform. 2016;17(1):23-32.

[46] Raddick MJ, Bracey G, Gay PL, Lintott CJ, Murray P, Schawinski K, et al. Galaxy zoo: exploring the motivations of citizen science volunteers. Astron Educ Rev. 2010;9(1):010103-1-010103-18.

[47] Wiersma Y, Birding 2.0: citizen science and effective monitoring in the web 2.0 world. Avian Conserv Ecol. 2010;5(2):1-9.

[48] Minet J, Curnel Y, Gobin A, Goffart JP, Melard F, Tychon B, et al. Crowdsourcing for agricultural applications: a review of uses and opportunities for a farmsourcing approach. Comput Electron Agric. 2017;142(Part A):126-38.

[49] See L. A review of citizen science and crowdsourcing in applications of pluvial flooding. Front Earth Sci. 2019;7(1):44-50.

[50] Strickland JC, Victor GA. Leveraging crowdsourcing methods to collect qualitative data in addiction science: narratives of non-medical prescription opioid, heroin, and fentanyl use. Int J Drug Policy. 2020;75(1):102587.

[51] Foncubierta RA, Müller H. Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia; 2012 Oct 29; Nara, Japan. New York: ACM; 2012. pp. 9-14.

[52] Créquit P, Mansouri G, Benchoufi M, Vivot A, Ravaud P. Mapping of crowdsourcing in health: systematic review. J Med Int Res. 2018;20(5):187-209.

[53] Lalor JP, Wu H, Chen L, Mazor KM, Yu H. ComprehENotes, an instrument to assess patient reading comprehension of electronic health record notes: development and validation. J Med Int Res. 2018;20(4):139-151.

[54] Lalor JP, Woolf B, Yu H. Improving electronic health record note comprehension with NoteAid: randomized trial of electronic health record note comprehension interventions with crowdsourced participants. J Med Int Res. 2019;21(1):10793.

[55] Lancers [Internet]. Available from: https://www.lancers.jp.

[56] Cheng J, Bernstein MS. Flock: hybrid crowd-machine learning classifiers. Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work; 2015 Mar 14-18; Vancouver BC, Canada. New York: Association for Computing Machinery; 2015. pp. 600-11.

[57] Tan PN, Kumar V, Srivastava J. Selecting the right objective measure for association analysis. Inform Syst. 2004;29(4):293-313.

[58] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008;10008.

[59] Japanese standard disease names based on ICD-10 [Internet]. Available from: https://www2.medis.or.jp/stdcd/byomei/index.html.

[60] Taku Kudoh [Internet]. MeCab. Available from: https://taku910.github.io/mecab/.

[61] IPA dictionary [Internet]. Available from: http://mecab.googlecode.com/files/mecab-ipadic-2.7.0-20070801.tar.gz.

[62] Satoh S. [Internet]. NagoyaObi 3.0.1. Available from: http://kotoba.nuee.nagoya-u.ac.jp/sc/obi3/.

[63] Muller CL, Chapman L, Johnston S, Kidd C, Illingworth S, Foody G, et al. Crowdsourcing for climate and atmospheric sciences: current status and future potential. Int J Climatol. 2015;35(11):3185-203.

[64] Shanley LA, Parker A, Schade S, Bonn A. Policy perspectives on citizen science and crowdsourcing. Citizen Science: Theory and Practice. 2019;4(1):30-34.

[65] Hashimoto H. Medical information systems as social common capital. J Natl Inst Public Health. 2010;59(1):10-6.