Using Natural Language Processing Techniques to Detect Adverse Events From Progress Notes Due to Chemotherapy

Yukinori Mashima[1,2] (iD), Takashi Tamura[3], Jun Kunikata[1] (iD),
Shinobu Tada[4], Akiko Yamada[2], Masatoshi Tanigawa[2] (iD),
Akiko Hayakawa[3], Hirokazu Tanabe[3] and Hideto Yokoi[1,2,4]

[1]Clinical Research Support Center, Kagawa University Hospital, Kagawa, Japan. [2]Department of Medical Informatics, Kagawa University Hospital, Kagawa, Japan. [3]Pharmacoepidemiology & PMS Department, Daiichi Sankyo Co., Ltd., Tokyo, Japan. [4]Information Network Administration Office, Faculty of Medicine, Kagawa University, Kagawa, Japan.

ABSTRACT

OBJECTIVE: In recent years, natural language processing (NLP) techniques have progressed, and their application in the medical field has been tested. However, the use of NLP to detect symptoms from medical progress notes written in Japanese, remains limited. We aimed to detect 2 gastrointestinal symptoms that interfere with the continuation of chemotherapy—nausea/vomiting and diarrhea—from progress notes using NLP, and then to analyze factors affecting NLP.

MATERIALS AND METHODS: In this study, 200 patients were randomly selected from 5277 patients who received intravenous injections of cytotoxic anticancer drugs at Kagawa University Hospital, Japan, between January 2011 and December 2018. We aimed to detect the first occurrence of nausea/vomiting (Group A) and diarrhea (Group B) using NLP. The NLP performance was evaluated by the concordance with a review of the physicians' progress notes used as the gold standard.

RESULTS: Both groups showed high concordance: 83.5% (95% confidence interval [CI] 74.1-90.1) in Group A and 97.7% (95% CI 91.3-99.9) in Group B. However, the concordance was significantly better in Group B ($P = .0027$). There were significantly more misdetection cases in Group A than in Group B (15.3% in Group A; 1.2% in Group B, $P = .0012$) due to negative findings or past history.

CONCLUSION: We detected occurrences of nausea/vomiting and diarrhea accurately using NLP. However, there were more misdetection cases in Group A due to negative findings or past history, which may have been influenced by the physicians' more frequent documentation of nausea/vomiting.

KEYWORDS: Natural language processing, data mining, electronic health records, progress notes, drug therapy, pharmacovigilance, drug-related side effects and adverse reactions.

Introduction

In cancer treatment, adverse events (AEs) decrease patients' quality of life (QOL) and reduce the treatment completion rate.[1] For example, chemotherapy-induced nausea and vomiting (CINV) is the most distressing symptom for patients.[2] The pathophysiology, risk factors and development of novel antiemetic agents have been researched to relieve patients' suffering.[3–5] Controlling AEs like CINV is the most important factor to maximize the therapeutic effect. In this respect, pharmacovigilance[6] is critically useful in controlling AEs, and it is essential to collect information on AEs accurately. Coded data generated in daily practice at multi-medical institutions have been integrated and used as a data source for pharmacovigilance. Typical examples are the Sentinel Initiative[7,8] by the U.S. Food and Drug Administration and MID-NET,[9] a national database that standardizes and integrates medical information from over 20 hospitals throughout Japan. However, Chan et al[10] reported that the analysis of narrative medical documents was more accurate in detecting symptoms like nausea/vomiting for hemodialysis patients than the analysis of International Classification of Diseases (ICD) codes. This suggests that it may be difficult to collect information accurately on some types of AEs using only coded data such as ICD codes.

Recently, there has been a move to use clinical data pooled in electronic medical records (EMRs) as a data source in pharmacovigilance.[11] In daily practice, structured coded data and unstructured narrative text are recorded in the EMRs of medical institutions.[12] Medical documents written by physicians include progress notes that record the daily changes in a

patient's condition, discharge summaries limited to the information needed to assess and manage a patient's future problems, and patient referral forms that give other physicians a short summary of relevant patient information.[13,14] Comparing these documents in terms of their suitability as data sources for pharmacovigilance, we believe that progress notes are the best document for this purpose. They contain more rich information about a patient's condition at each visit, while discharge summaries and patient referral forms are likely to contain only selective information. Natural language processing (NLP) techniques are used for quantitative analysis of narrative text in progress notes.[15] Symptom detection from medical documents using NLP for pharmacovigilance has been reported,[16] including in the Japanese context. Aramaki et al[17] tried to extract AEs by detecting "drug-symptom pairs" in Japanese discharge summaries. Ujiie et al[18] developed a system to determine the presence of AEs in Japanese case reports, and Shimai et al[19] combined the analysis of Japanese radiology reports and blood test results to detect drug-induced interstitial pneumonia. However, no studies have examined Japanese progress notes as a data source for pharmacovigilance.

To address this gap in the literature, we aimed to detect symptom occurrence from physicians' progress notes using NLP. We focused on patients who had received chemotherapy, and the detection of 2 associated major gastrointestinal toxicity symptoms—nausea/vomiting[20] and diarrhea.[21] Then, we examined the factors affecting NLP performance by analyzing how each symptom was written in the progress note.

Materials and Methods

Study population

We randomly selected 200 out of 5277 patients who had received intravenous injections of cytotoxic anticancer drugs at Kagawa University Hospital (KUH), Japan, for gastrointestinal cancer, pancreas and biliary cancer, breast cancer, or ovarian cancer between January 2011 and December 2018. The observation period for each patient was from the date of the first injection of the anticancer drug (index date) to the end of the regimen, including the injection period and subsequent drug withdrawal period. KUH is the only academic medical center as a national university corporation in Kagawa Prefecture, with 230 000 outpatients and 5000 ambulatory cancer chemotherapies per year.

This study was conducted after approval by the institutional review board of Kagawa University through an ethical review (receipt number: 2019-093) and after confirmation of patient consent through the opt-out approach.

Dataset preparation

The physician progress notes during the observation period were collected for each patient from the EMR (HOPE EGMAIN-GX by Fujitsu Ltd.) at KUH and randomly divided into a trial dataset (n = 30) and an evaluation dataset (n = 170), as shown in Figure 1. The evaluation dataset was randomly divided into Group A (n = 85) for the detection of nausea/vomiting and Group B (n = 85) for the detection of diarrhea.

The notes were written in the 'SOAP' format (subjective data, objective data, assessment, and plan),[13,14] and each one included the date and time when it was written. We deconstructed the progress notes into separate records for each line break. Each record was given the date and time information of the original progress note.

The patients' clinical background (age, sex, inpatient/outpatient, cancer type, injected anticancer drugs, and observation period), the number of progress notes per patient, the number of characters per progress note, and the total number of records in each group were compared. Furthermore, we compared the percentage of 3 types of records containing the following dictionary words (see Natural language processing section): written symptom occurrence (positive finding), written symptom absence (negative finding), or written past history about the symptom.

Progress notes review

Two physicians (Y.M., J.K.) independently reviewed the progress notes of all patients. For both groups (Group A: nausea/vomiting; Group B: diarrhea), we defined patients with symptoms or the therapeutic intervention during the observation period as 'gold standard' (GS) positive; otherwise, they were considered GS negative. In GS positives, the date and time of the progress note written about the first occurrence of each symptom were noted. The shared annotation rules used are shown in Supplemental Table S1. We calculated a simple percentage agreement[22] between the 2 physicians. In cases where the 2 physicians' opinions differed, the judgment was determined through their discussion. Disagreements were resolved by a third physician (H.Y.). All physicians had more than 10 years clinical experience and used EMR in their daily practice.

Natural language processing

Figure 2 shows an overview of the processing for each progress note. First, every record in the progress note was tokenized into morphemes by MeCab 0.996,[23] and keyword matching was performed with dictionary words (described below). If there was no matching, the system determined that the symptom did not occur in the progress note ('No occurrence'). If a matching morpheme was found, the next step, dependency structure analysis was performed using CaboCha 0.69.[24] If all matched morphemes were following a negation, the system determined the progress note as 'No occurrence.' In contrast, if there was a matched morpheme without negation, the system determined the progress note as 'AE-occurred.' The system outputted the date and time information of the 'AE-occurred' progress note closest to the index date per patient.
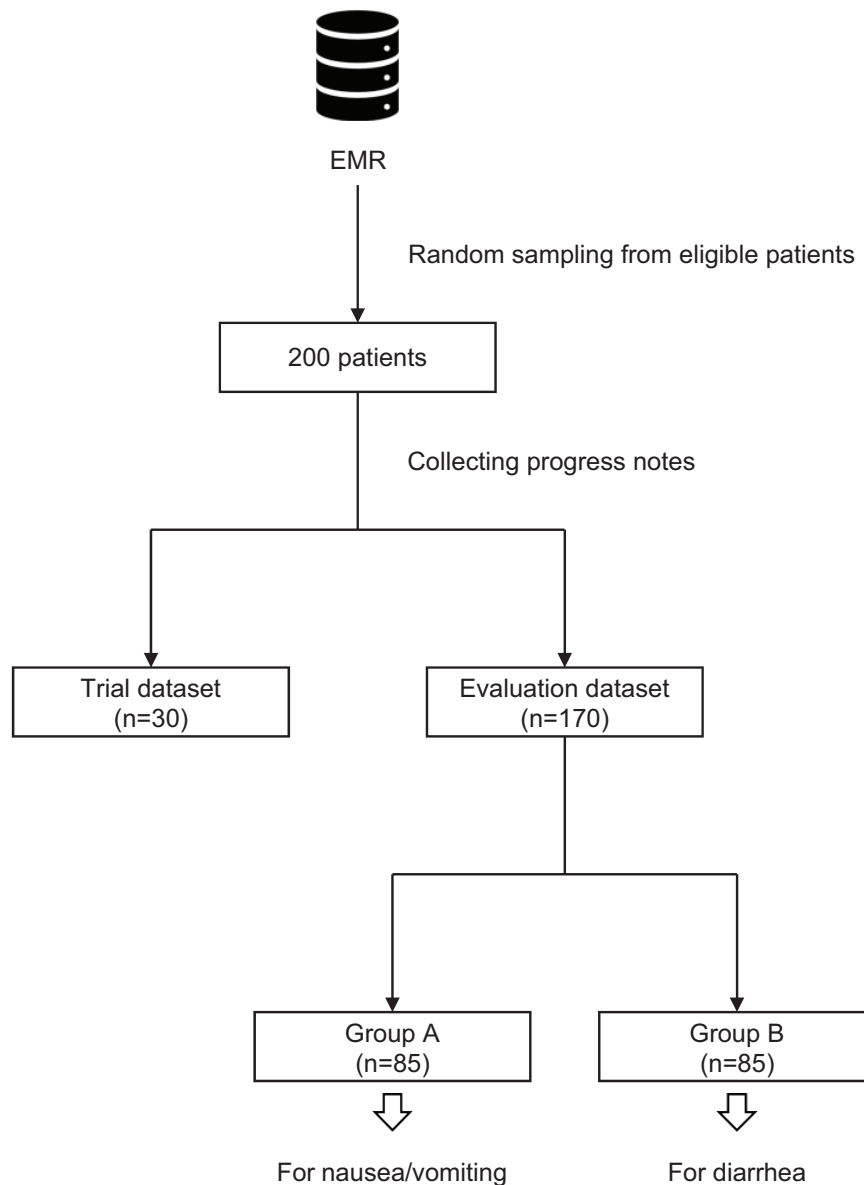
Figure 1. Dataset preparation. EMR, electronic medical record.

After creating the initial dictionary words and the negations based on a heuristic decision, the dictionary words were finalized by repeating the trials using the trial dataset. The finalized dictionary words and the negations were used to analyze the evaluation dataset and are shown in Supplemental Table S2. This NLP system was built on CentOS 7.6.1810 and Python 3.6.8.

NLP performance evaluation

All outputs of the NLP system determined for each patient were compared with the gold standard and classified into the following 5 categories.

a) Correct detection: The system outputted the same date and time as GS positive.

b) True negative: The system outputted the same as GS negative.

c) Early detection: The system outputted date and time that were earlier than GS positive.

d) False positive: The system detected an AE incorrectly when the GS was negative.

e) Delayed detection or False negative: The system outputted date and time later than GS positive, or the system did not detect the AE despite the GS being positive.

Additionally, correct detections and true negatives were treated as 'Matched' cases; early detections and false positives were treated as 'Misdetection' cases, and delayed detections and false negatives were treated as 'Overlooked' cases. We compared the percentage of matched cases, misdetection cases, and

Progress note

Tokenize by line (MeCab)

Dictionary matching by word → No match → No occurrence

Match

Dependency analysis (CaboCha) → With negation → No occurrence
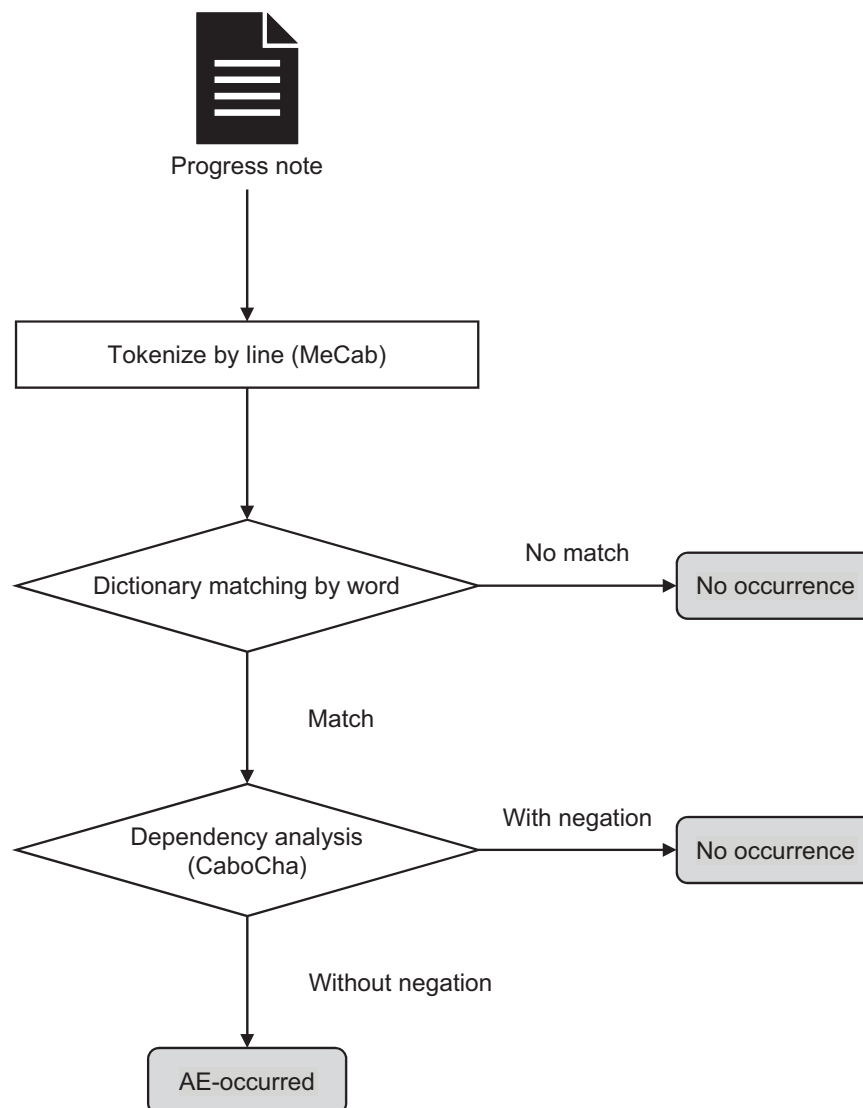
Without negation

AE-occurred

Figure 2. Processing for each progress note. AE, adverse event.

overlooked cases between Group A and Group B. Following that, we analyzed the factors that contributed to errors in the NLP system.

Analysis of GS negatives

For GS negatives, we compared the percentages of the records containing a dictionary word in Group A and Group B that were written as negative findings or as past histories.

Word frequencies of AE-related words

As a supplementary analysis, word frequencies of AE-related words were counted by a physician (Y.M.) to investigate the expressions of nausea/vomiting and diarrhea other than dictionary words. We tokenized all records in each group into morphemes using MeCab and counted word frequencies of AE-related words. AE-related words were defined as words that consist of one morpheme and can be understood to relate to each symptom.

Statistical analysis

Fisher's exact test was used to test for differences between Group A and Group B in the background characteristics of the dataset, the percentage of matched cases, misdetection cases, and overlooked cases, and the percentage of negative findings and past histories in GS negatives. A 2-sided P-value $<.05$ was considered statistically significant. Analysis was performed using R 4.0.3 (https://www.r-project.org/).

Results

Background characteristics

The characteristics of patients and progress notes are shown in Table 1. The median observation period in both groups was 21 days, and the median number of progress notes per patient was 5. There were 20 106 lines of records for Group A and 22 057 lines for Group B. The percentage of records containing the dictionary words as positive findings was 0.8% in Group A and 0.3% in Group B ($P<.001$). Negative findings constituted 1.1% of records in Group A and 0.2% in Group B ($P<.001$).

Table 1. Background characteristics.

	GROUP A FOR NAUSEA/VOMITING (N=85)	GROUP B FOR DIARRHEA (N=85)	*P*-VALUE
Age, years	63 [50-70]	65 [59-70]	n/a
Female	65.9% (55.3-75.1)	58.8% (48.2-68.7)	.43
Outpatient	49.4% (39.0-59.8)	48.2% (37.9-58.7)	1
Cancer type			
Gastrointestinal cancer	23.5% (15.7-33.6)	22.4% (14.7-32.4)	1
Pancreas and biliary cancer	43.5% (33.5-54.1)	47.1% (36.8-57.6)	.76
Breast cancer	24.7% (16.7-34.9)	18.8% (11.8-28.5)	.46
Ovarian cancer	8.2% (3.8-16.3)	11.8% (6.3-20.5)	.61
Total number of anticancer drugs	178	158	n/a
Frequency of the top 5 drugs			
Gemcitabine	14.6% (10.1-20.6)	19.0% (13.6-25.9)	.31
Fluorouracil	14.0% (9.6-20.0)	10.1% (6.2-15.9)	.32
Oxaliplatin	12.9% (8.7-18.7)	15.2% (10.4-21.7)	.64
Cyclophosphamide	11.8% (7.8-17.4)	8.2% (4.8-13.7)	.37
Paclitaxel	9.0% (5.5-14.2)	10.8% (6.7-16.6)	.59
Observation period, days	21 [14–22]	21 [14–28]	n/a
Progress notes per patient	5 [3–10]	5 [3–10]	n/a
Characters per progress note	422.5 [187.75–766]	390 [181.25–675.5]	n/a
Total records in all progress notes, lines	20 106	22 057	n/a
Frequency of records containing dictionary word			
As positive findings	0.8% (0.6-0.9)	0.3% (0.2-0.3)	<.001**
As negative findings	1.1% (1.0-1.3)	0.2% (0.2-0.3)	<.001**
As past history	0.11% (0.07-0.17)	0.03% (0.01-0.06)	.0011*

Values are shown as a median [interquartile range] or percentage (95% confidence interval).
*$P<.005$; **$P<.001$; n/a, not applicable.

Past history was noted in 0.11% of records in Group A and 0.03% in Group B (*P*=.0011).

Gold standard

Thirty patients in Group A (35.3%; 95% confidence interval [CI] 26.0-45.9) and 14 patients in Group B (16.5%; 95% CI 9.95-25.9) were judged as GS positive. The simple percentage agreement between the 2 physicians was 82.4% in Group A and 94.1% in Group B. Only 2 patients (2.4%) in Group A required judgments by the third physician.

NLP performance

The outputs of the NLP system in each group were classified, as shown in Table 2. There were 23 matched cases out of 30 GS positives in Group A, and 13 matched cases out of 14 GS positives in Group B. In the same way, there were 48 matched cases out of 55 GS negatives in Group A, and 70 matched cases out of 71 GS negatives in Group B. As a result, the concordance with the gold standard across Group A was 83.5%, and across Group B, 97.7%, which showed a statistically significant difference (*P*=.0027). The concordance with the GS in both groups denoted the same tendency as the simple percentage agreement in the progress notes review. Additionally, the percentage of misdetection cases across Group A was 15.3%, and that across Group B was 1.2%, another statistically significant difference (*P*=.0012).

Error analysis

Table 3 summarizes the errors of the NLP system. There were 8 misdetection cases for negative findings in Group A, and 1 in

Table 2. NLP system performance.

	GS POSITIVE			GS NEGATIVE			TOTAL		
	PATTERN DIAGRAM	GROUP A (N=30)	GROUP B (N=14)	PATTERN DIAGRAM	GROUP A (N=55)	GROUP B (N=71)	GROUP A (N=85)	GROUP B (N=85)	P-VALUE
Matched	a)	23	13	b)	48	70	71 83.5% (74.1 to 90.1)	83 97.7% (91.3 to 99.9)	.0027*
Misdetection	c)	6	0	d)	7	1	13 15.3% (9.0 to 24.6)	1 1.2% (−0.4 to 7.0)	.0012*
Overlooked	e)	1	1	n/a			1 1.2% (−0.4 to 7.0)	1 1.2% (−0.4 to 7.0)	1

Abbreviations: AE, adverse event; GS, gold standard; n/a, not applicable.
Arrows indicate timeline, and circles mean timing defined as 'GS positive' in progress notes review, inverted triangles mean timing outputted as 'AE-occurred' by the NLP system. Values are shown as a number of cases and percentage (95% confidence interval). a)Correct detection, The system outputted the same date and time as GS positive; b)True negative, The system outputted the same as GS negative; c)Early detection, The system outputted date and time that were earlier than GS positive; d)False positive, The system detected an AE incorrectly when the GS was negative; e)Delayed detection or False negative, The system outputted date and time later than GS positive, or the system did not detect the AE despite the GS being positive.
*P < .005.

Table 3. Error analysis.

ERROR TYPES	SYMPTOMS TO BE DETECTED	ERROR CASES		EXAMPLES
		MISDETECTION (N = 14)	OVERLOOKED (N = 2)	ORIGINAL JAPANESE TEXTS (TRANSLATED TO ENGLISH)
Negative findings	Nausea/ vomiting	8	n/a	吐き気は全くありません† (There is completely zero nausea.)
	Diarrhea	1	n/a	口腔粘膜炎・下痢・末梢神経障害≦Grade2‡ (Oral mucositis, diarrhea, peripheral neuropathy ≤ Grade 2)
Past history	Nausea/ vomiting	4	n/a	10/14 に吐いた。 (Vomited on October 14.)
	Diarrhea	0	n/a	n/a
Others§	Nausea/ vomiting	1	1	便が出そうで出ないのが気持ち悪い‖ (I feel bad because I can't get my stool out.)
	Diarrhea	0	1	便が軟らかい (My stools are loose.)

†The spelling variant of negation could not be recognized; ‡The spelling variant of negation could not be recognized, and the dependency structure analysis for the parallel structure was incorrect; §Other errors were noted due to word-sense ambiguity; ‖The Japanese word '気持ち悪い' can mean either nausea or discomfort; n/a, not applicable.

Table 4. Analysis of GS negatives.

	GS NEGATIVE					
	TRUE NEGATIVE†			FALSE POSITIVE‡		
	GROUP A (N = 19)	GROUP B (N = 11)	*P*-VALUE	GROUP A (N = 7)	GROUP B (N = 1)	*P*-VALUE
Total records containing dictionary word, lines	3024	2198	n/a	2641	612	n/a
As negative findings	1.9% (1.4 to 2.4)	1.3% (0.9 to 1.9)	0.15	1.3% (0.9 to 1.8)	1.0% (0.4 to 2.2)	0.68
As past history	0.07% (0.00 to 0.26)	0	n/a	0.04% (-0.02 to 0.24)	0	n/a

Abbreviations: AE, adverse event; GS, gold standard; n/a, not applicable.
†True negative, The system outputted the same as GS negative; ‡ False positive, The system detected an AE incorrectly when the GS was negative.

Group B. Similarly, there were 4 misdetection cases of past history in Group A, and none in Group B. There was also 1 misdetection case due to a polysemy, and some overlooked cases due to spelling variants.

Analysis of GS negatives

Table 4 shows the analysis of GS negatives. Of true negatives for GS negatives (determined correctly to be the absence of an AE), 3024 lines in Group A and 2198 lines in Group B contained dictionary words. There were 1.9% records written as negative findings in Group A and 1.3% in Group B (*P* = .15). Those with past histories comprised 0.07% in Group A and none in Group B. Similarly, among false positives of GS negatives (incorrectly determined in the absence of an AE), 2641 lines in Group A and 612 lines in Group B contained dictionary words. Records written as negative findings comprised 1.3% of Group A and 1.0% of Group B (*P* = .68). There were 0.04% with past histories in Group A and none in Group B.

Word frequencies of AE-related words

The frequencies of AE-related words are shown in Figure 3. Among nausea/vomiting-related words in Group A, the most frequent was '嘔吐' ('vomiting' in English), counted 134 times. This was followed by '悪心' ('nausea' in English), counted 117 times. There were several words related to nausea/vomiting. In contrast, the only diarrhea-related word in Group B was '下痢' ('diarrhea' in English), which was counted 105 times. No other diarrhea-related words were found.

Discussion

This study showed that the occurrence of 2 gastrointestinal symptoms—nausea/vomiting and diarrhea—could be detected using our NLP system based on physicians' progress notes written in Japanese for patients who had received anticancer drugs. The NLP system performed adequately, as envisaged. Pharmacovigilance in Japan mainly uses a spontaneous reporting system called the Japanese Adverse Drug Event Report database and an administrative database, MID-NET. They are
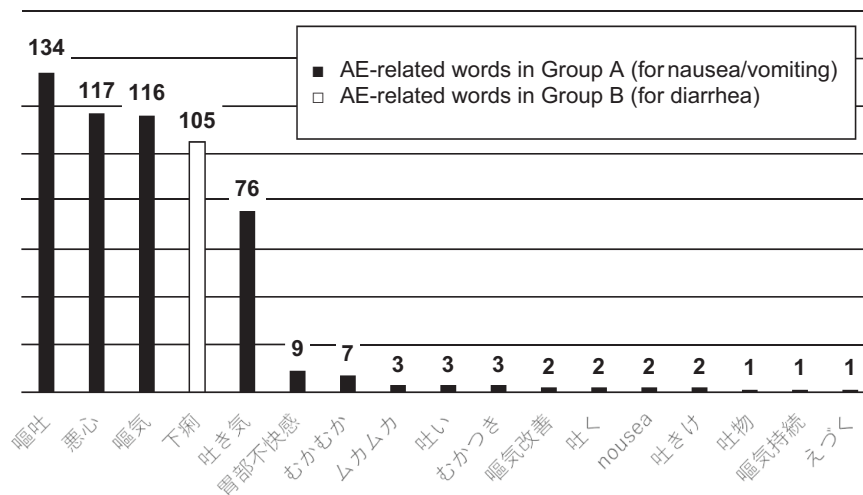
Figure 3. Word frequencies of AE-related words. AE-related words consist of one morpheme and were counted if they were understood to relate to each symptom; Black bars indicate frequencies of nausea/vomiting-related words in Group A, and the white bar indicates the frequency of diarrhea-related words in Group B; Values shown above the bars indicate the frequency of each word; For example, '嘔吐' means vomiting in English, '悪心' means nausea, and '下痢' means diarrhea; Onomatopoeia, Japanese-specific expressions, misspellings and so on are lined up.
Abbreviation: AE, adverse events.

all clusters of structured data. However, one of the weaknesses of structured data analysis is that it can only collect predefined information. In the Japan Chronic Kidney Disease Database[25]—a nationwide database for chronic kidney disease in Japan—basic information such as body mass index and blood pressure is not available because it was not included in the database design. It is difficult to collect additional information that has not been included in the database construction phase. In such situations, it is possible to obtain information from narrative text in progress notes by modifying the NLP system for new detection tasks. This study suggests that narrative text in progress notes could become a leading source for pharmacovigilance.

There was a statistically significant difference in the performance of the NLP system between the nausea/vomiting detection group and the diarrhea detection group. There was also a significant difference in the percentage of misdetection cases between the 2 symptoms, although there was no difference in the percentage of overlooked cases. To accurately interpret the NLP results, we considered it necessary to perform text mining for progress notes as a data source to analyze the factors that contributed to these differences. Error analysis revealed that misdetection cases caused by negative findings in the nausea/vomiting detection group were the most frequent, followed by misdetection cases caused by past histories in the same group. These results suggest that negative findings and past histories for nausea/vomiting had the most influence on the performance of the NLP system. Additionally, the analysis of GS negatives showed no significant difference in true negatives and false positives between the 2 symptoms. In contrast, the background characteristics of the dataset showed that the number of records containing the dictionary words as negative findings or past history was significantly higher in the nausea/vomiting detection group. In this study, we intended to improve the performance of the NLP system by repeating trials to enrich the dictionary words and negations. This is because it is well-known that the processing of negation is crucial when detecting information from medical documents using NLP.[26,27] Cohen et al[28] reported that the distribution of negation varies depending on the type of medical document, which shows that there are more explicit negations in progress notes and more affixal negations in biomedical journal articles. Our NLP system excluded affixal negations by using dictionary words and excluded explicit negations by using dictionary words with the negation. Like this study, when Usui et al[29] tried to detect symptoms from pharmacy medication history data written in Japanese, misdetections mainly occurred because of negative findings. These negative findings are intentionally written in medical documents, including progress notes[13,14] because they are useful for ruling-out diseases and are clinically important.[30] Therefore, the results of this study indicate that more information on negative findings or past histories of nausea/vomiting contributed to differences in the detection performance of the NLP system for each symptom.

Comparing these 2 symptoms in terms of QOL, Morita et al[31] reported that nausea/vomiting is a clinical parameter that affects all domains of QOL-ACD[32] when assessing QOL of patients treated with anticancer drugs. In contrast, diarrhea affects the physical, mental, and psychological domains but not the functional domain. Functional disorders are easily recognized objectively, but physical, mental and psychological disorders are often unrecognizable to anyone other than the patient. Because of this, there may be a discrepancy in the patient's self-assessment and the healthcare provider's assessment based on the patient's general condition. Kobayashi et al[33] reported that cognitive functioning, fatigue and nausea/vomiting in the QLQ-C30[34] items influenced the Karnofsky Performance Status[35] as assessed by medical professionals.

However, diarrhea did not. Although nausea/vomiting and diarrhea both affect patient QOL, nausea/vomiting is more likely to affect the functional domain, be evaluated objectively, and be a critical focus for medical professionals and patients. Boland et al[36] described information heterogeneity in medical documents because physicians' documentation behavior is influenced by patient status (eg, critical or stable), disease status (eg, early or advanced), and also by the experience of the physicians (eg, trainee or expert). Nausea/vomiting, which has a great impact on the functional domain of patients' QOL, tends to lead to physician documentation behavior, and negative findings and past histories are also often included in progress notes. For this reason, the NLP system may have been affected by the information heterogeneity.

The frequency of AE-related words showed that the dataset for the nausea/vomiting detection group contained several words, while the dataset for the diarrhea detection group contained only 1 word. Word-sense ambiguity[37] is known to be problematic, along with negation, in NLP. The more related words, the more difficult it is to achieve consistent mapping. The simple percentage agreement between 2 physicians in the progress notes review was also inferior for nausea/vomiting compared with diarrhea. Because nausea/vomiting is more diverse in expression than diarrhea, it is also difficult to detect the symptom by dictionary matching, resulting in the detection performance difference for each symptom.

This study has several limitations. First, it was conducted at a single institution. KUH is an academic medical center, and patients' backgrounds may have been biased given the institution's characteristics. Second, the cancer types were specified as inclusion criteria. Thus, the physicians who wrote the progress notes and their departments were limited. It is possible that progress notes were influenced by the writing patterns of physicians and their departments. Third, the types of anticancer drugs were also limited. In recent years, conventional cytotoxic anticancer drugs have been used along with drugs with novel mechanisms. The use of drugs with different profiles changes the frequency of AEs, which may change the performance indicators of the NLP system.

Conclusion

Our NLP system could detect the occurrence of nausea/vomiting and diarrhea from Japanese physicians' progress notes for patients who received anticancer drugs. Progress notes may constitute a useful data source for pharmacovigilance, but the detection performance for each symptom may be affected by physicians' documentation behaviors. The performance of this NLP system is expected to be improved by strengthening the processing considering negative findings and past history. Particularly in the nausea/vomiting detection group, reducing misdetection cases using these strategies was required. In future, when progress notes are used as a data source for pharmacovigilance, it would be necessary to interpret the NLP results with more detailed consideration of the associated characteristics.

Author Contributions

Y.M., T.T., J.K., S.T., M.T., A.H., H.T., and H.Y. designed the study protocol. Y.M., T.T., J.K., and S.T. developed the system. Y.M., A.Y., and M.T. interpreted the results and prepared the manuscript. All authors reviewed and approved the final manuscript.

ORCID iDs

Yukinori Mashima [iD] https://orcid.org/0000-0002-7584-0157
Jun Kunikata [iD] https://orcid.org/0000-0002-0089-8311
Masatoshi Tanigawa [iD] https://orcid.org/0000-0003-2676-356X

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Lindley C, Vasa S, Sawyer WT, Winer EP. Quality of life and preferences for treatment following systemic adjuvant therapy for early-stage breast cancer. *J Clin Oncol*. 1998;16:1380-1387.
2. Coates A, Abraham S, Kaye SB, et al. On the receiving end—patient perception of the side-effects of cancer chemotherapy. *Eur J Cancer Clin Oncol*. 1983; 19:203-208.
3. Hesketh PJ. Chemotherapy-induced nausea and vomiting. *N Engl J Med*. 2008;358:2482-2494.
4. Bossi P, Airoldi M, Aloe Spiriti MA, et al. A multidisciplinary expert opinion on CINV and RINV, unmet needs and practical real-life approaches. *Expert Opin Drug Saf*. 2020;19:187-204.
5. Gupta K, Walton R, Kataria SP. Chemotherapy-induced nausea and vomiting: pathogenesis, recommendations, and new trends. *Cancer Treat Res Commun*. 2021;26:100278.
6. World Health Organization. *The Importance of Pharmacovigilance*. World Health Organization; 2002.
7. Platt R, Carnahan RM, Brown JS, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 1):1-8.
8. Robb MA, Racoosin JA, Sherman RE, et al. The US Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 1):9-11.
9. Yamaguchi M, Inomata S, Harada S, et al. Establishment of the MID-NET® medical information database network as a reliable and valuable database for drug safety assessments in Japan. *Pharmacoepidemiol Drug Saf*. 2019;28: 1395-1404.
10. Chan L, Beers K, Yau AA, et al. Natural language processing of electronic health records is superior to billing codes to identify symptom burden in hemodialysis patients. *Kidney Int*. 2020;97:383-392.
11. Osokogu OU, Dukanovic J, Ferrajolo C, et al. Pharmacoepidemiological safety studies in children: a systematic review. *Pharmacoepidemiol Drug Saf*. 2016;25: 861-870.
12. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform*. 2008;77:291-304.
13. Weed LL. *Medical Records, Medical Education, and Patient Care: The Problem Oriented Record as a Basic Tool*. Case Western Reserve University; 1969.
14. Shigeaki H. *POS: The Problem-Oriented System (in Japanese)*. Igaku-Shoin Ltd.; 1973.

15. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc*. 2019;26(4):364-379.

16. Luo Y, Thompson WK, Herr TM, et al. Natural language processing for EHR-based pharmacovigilance: a structured review. *Drug Saf*. 2017;40:1075-1089.

17. Aramaki E, Miura Y, Tonoike M, et al. Extraction of adverse drug effects from clinical records. *Stud Health Technol Inform*. 2010;160:739-743.

18. Ujiie S, Yada S, Wakamiya S, Aramaki E. Identification of adverse drug event-related Japanese articles: natural language processing analysis. *JMIR Med informatics*. 2020;8:e22661.

19. Shimai Y, Takeda T, Okada K, et al. Screening of anticancer drugs to detect drug-induced interstitial pneumonia using the accumulated data in the electronic medical record. *Pharmacol Res Perspect*. 2018;6:e00421.

20. Hesketh PJ, Kris MG, Basch E, et al. Antiemetics: American Society of Clinical Oncology Clinical Practice Guideline Update. *J Clin Oncol*. 2017;35: 3240-3261.

21. Benson AB, Ajani JA, Catalano RB, et al. Recommended guidelines for the treatment of cancer treatment-induced diarrhea. *J Clin Oncol*. 2004;22: 2918-2926.

22. McHugh ML. Interrater reliability: the kappa statistic. *Biochem medica*. 2012;22:276-282.

23. Kudo T, Yamamoto K, Matsumoto Y. Applying conditional random fields to Japanese morphological analysis. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain; July 2004:230-237.

24. Kudo T, Matsumoto Y. Japanese dependency analysis using cascaded chunking. *Proceeding of COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)* Taipei, Taiwan; August 2002:1-7. doi:10.3115/1118853 .1118869

25. Nakagawa N, Sofue T, Kanda E, et al. J-CKD-DB: a nationwide multicentre electronic health record-based chronic kidney disease database in Japan. *Sci Rep*. 2020;10:7351.

26. Wu S, Miller T, Masanz J, et al. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PLoS One*. 2014;9:e112774.

27. Morante R, Blanco E. Recent advances in processing negation. *Nat Lang Eng*. 2021;27:121-130.

28. Cohen KB, Goss FR, Zweigenbaum P, Hunter LE. Translational morphosyntax: distribution of negation in clinical records and biomedical journal articles. *Stud Health Technol Inform*. 2017;245:346-350.

29. Usui M, Aramaki E, Iwao T, Wakamiya S, Sakamoto T, Mochizuki M. Extraction and standardization of patient complaints from electronic medication histories for pharmacovigilance: natural language processing analysis in Japanese. *JMIR Med informatics*. 2018;6:e11021.

30. Bickley Lynn S. *BATES' Guide to Physical Examination and History Taking*. Wolters Kluwer; 2013.

31. Morita S, Kobayashi K, Eguchi K, et al. Influence of clinical parameters on quality of life during chemotherapy in patients with advanced non-small cell lung cancer: application of a general linear model. *Jpn J Clin Oncol*. 2003;33: 470-476.

32. Kurihara M, Shimizu H, Tsuboi K, et al. Development of quality of life questionnaire in Japan: quality of life assessment of cancer patients receiving chemotherapy. *Psychooncology*. 1999;:355-363.

33. Kobayashi K, Takeda F, Teramukai S, et al. A cross-validation of the European Organization for Research and Treatment of Cancer QLQ-C30 (EORTC QLQ-C30) for Japanese with lung cancer. *Eur J Cancer*. 1998;34:810-815.

34. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst*. 1993; 85:365-376.

35. Karnofsky DA, Burchenal JH. The clinical evaluation of chemotherapeutic agents in cancer. In: MacLeod CM, ed. *Evaluation of Chemotherapeutic Agents*. Columbia University Press; 1949:191-205.

36. Boland MR, Hripcsak G, Shen Y, Chung WK, Weng C. Defining a comprehensive verotype using electronic health records for personalized medicine. *J Am Med Inform Assoc*. 2013;20:e232-e238.

37. Navigli R. Word sense disambiguation: a survey. *ACM Comput Surv*. 2009; 41:1-69.