Use of Natural Language Processing to Infer Sites of Metastatic Disease From Radiology Reports at Scale

See Boon Tay, MBBS[1,2]; Guat Hwa Low, BS[1,3]; Gillian Jing En Wong, MBBS[2]; Han Jieh Tey, MSc[1,3]; Fun Loon Leong, BS[1,3]; Constance Li, PhD[3]; Melvin Lee Kiang Chua, MBBS, PhD[3,4,5]; Daniel Shao Weng Tan, MBBS, PhD[1,4,6]; Choon Hua Thng, MBBS[4,7]; Iain Bee Huat Tan, MBBS, PhD[1,3,4]; and Ryan Shea Ying Cong Tan, MBBS, MTec[1,3,4,8]

DOI https://doi.org/10.1200/CCI.23.00122

ABSTRACT

PURPOSE To evaluate natural language processing (NLP) methods to infer metastatic sites from radiology reports.

METHODS A set of 4,522 computed tomography (CT) reports of 550 patients with 14 types of cancer was used to fine-tune four clinical large language models (LLMs) for multilabel classification of metastatic sites. We also developed an NLP information extraction (IE) system (on the basis of named entity recognition, assertion status detection, and relation extraction) for comparison. Model performances were measured by F1 scores on test and three external validation sets. The best model was used to facilitate analysis of metastatic frequencies in a cohort study of 6,555 patients with 53,838 CT reports.

RESULTS The RadBERT, BioBERT, GatorTron-base, and GatorTron-medium LLMs achieved F1 scores of 0.84, 0.87, 0.89, and 0.91, respectively, on the test set. The IE system performed best, achieving an F1 score of 0.93. F1 scores of the IE system by individual cancer type ranged from 0.89 to 0.96. The IE system attained F1 scores of 0.89, 0.83, and 0.81, respectively, on external validation sets including additional cancer types, positron emission tomography-CT ,and magnetic resonance imaging scans, respectively. In our cohort study, we found that for colorectal cancer, liver-only metastases were higher in de novo stage IV versus recurrent patients (29.7% *v* 12.2%; *P* < .001). Conversely, lung-only metastases were more frequent in recurrent versus de novo stage IV patients (17.2% *v* 7.3%; *P* < .001).

CONCLUSION We developed an IE system that accurately infers metastatic sites in multiple primary cancers from radiology reports. It has explainable methods and performs better than some clinical LLMs. The inferred metastatic phenotypes could enhance cancer research databases and clinical trial matching, and identify potential patients for oligometastatic interventions.

ACCOMPANYING CONTENT

Data Supplement

Accepted April 1, 2024
Published May 24, 2024

JCO Clin Cancer Inform
8:e2300122

INTRODUCTION

Pattern of metastatic spread is an important diagnostic and prognostic factor in advanced-stage cancer.[1,2] Different primary cancers display distinct metastatic organotropism, which may lend insight into disease biology.[3] For instance, bone-only tropism is seen more frequently in breast and prostate cancers, and often carries a better prognosis than those with visceral metastases.[4-6] Metastatic phenotype and its evolution over time guide clinical decision making and treatment selection. For example, surgical and radiotherapy treatments for oligometastatic disease can be discussed while clinical trials may have specific criteria such as excluding patients with active brain metastases.[7,8] Thus, it is important for oncology electronic health records to be able to accurately capture this information for every patient.

Computed tomography (CT) is a common imaging modality for identifying sites of metastatic disease.[9] Data extraction from CT radiology reports using natural language processing (NLP) has been explored to gain insights from this rich source of unstructured clinical data.[10-12] Earlier work to identify sites of metastases used conventional machine learning methods (eg, XGBoost classifiers) and early transformer models (ie, BERT). Although these achieved promising accuracies ranging between 0.84 and 0.99,[9] such approaches predicted only one site of metastases per model. Newer NLP approaches using large language models (LLMs) have since emerged and achieved state-of-the-art performance on many clinical NLP benchmarks.[13-15] They present a promising new approach to apply to this task of inferring metastatic sites of cancer from radiology reports.

CONTEXT

Key Objective

To evaluate natural language processing (NLP) methods for inferring various metastatic sites from unstructured radiology reports in multiple primary cancers.

Knowledge Generated

Using a cohort of 4,522 computed tomography reports from 550 patients spanning 14 cancer types, we developed an information extraction system that uses a named entity recognition-assertion-relation extraction-term normalization pipeline. It accurately infers metastatic sites in multiple cancers with explainable methods and performs better than some clinical large language models.

Relevance *(J.L. Warner)*

The authors have developed a highly performing model to address the important task of determining metastatic sites from unstructured radiology reports, and have made their software and models available with a Spark NLP Healthcare license. Readers are encouraged to evaluate the model on their own datasets to establish utility and generalizability.*

*Relevance section written by *JCO Clinical Cancer Informatics* Editor-in-Chief Jeremy L. Warner, MD, MS, FAMIA, FASCO.

In this paper, we aim to (1) evaluate multiple fine-tuned LLMs trained on clinical corpora for prediction of multiple sites of metastases for multiple primary cancers; (2) develop an information extraction (IE) system on the basis of traditional NLP methods customized for this task for comparison; and (3) demonstrate use of the best-performing model for large-scale labeling of CT reports to extract metastatic phenotypes over patient treatment journeys and discuss how these could enrich data insights for cancer research and clinical care.

METHODS

Data Source and Study Participants

This study recruited patients with any cancer type and stage seen at the National Cancer Centre Singapore or Singapore General Hospital into a prospectively maintained research database. Prospective and retrospective CT radiology reports across various anatomic regions, including the brain, neck, paranasal space, thorax, abdomen, and pelvis, between May 2000 and February 2022 were retrieved and deidentified. Patient age, sex, self-reported ethnicity, cancer stage, and primary site were also collected. This study was approved by Singapore Health Services' Institutional Review Board (CIRB Ref: 2019/2401, 2018/2795, 2018/3046, 2019/2170). All participants provided written informed consent.

Report Annotation

Radiology reports were extracted and processed by a research data management team supervised by a medical oncologist. Each report included the reporting date, modality, region, and four sections: Report, Finding, Other Finding, and Conclusion. Reports without Conclusion section were excluded. Only the Conclusion section was used for downstream model development. In our recruiting institutions, the Conclusion section consistently reported with the greatest certainty and almost always included all identified sites of metastases.

The Conclusion sections were first annotated by medical research staff using INCEpTION, an open-source semantic annotation platform.[16] All annotations were curated by a team consisting of an oncology research officer, a consultant medical oncologist, and a senior consultant radiologist. These annotations were done according to a consensus classification dictionary developed by the curation team (Data Supplement, Table S1 and Fig S1).

The key named entity for sites of metastases was cancer imaging finding. This entity was given an assertion status to assess level of confidence in an imaging finding being metastases using a four-point Likert scale: definite (probability-high), probable (probability-medium), indeterminate (probability-uncertain), and low (probability-low).[17,18] More granular anatomic detail (eg, segment V, upper lobe) were labeled using the named entity anatomic descriptor. Relationships between named entities were also identified by relation labels.

Annotated reports were divided into patient-level train (90%; including a 10% development set) and hold-out test sets (10%), maintaining proportional representation of major tumor types.

Clinical LLMs

We fine-tuned four transformed-based multilabel classification models of varying parameter sizes trained on clinical or biomedical texts: RadBERT (110 million parameters, 4.42 million radiology reports),[13] BioBERT (110 million parameters, English Wikipedia, BooksCorpus, PubMed),[14] GatorTron-base (345 million parameters, >290 million clinical notes), and GatorTron-medium (3.9 billion parameters, >290 million clinical notes).[15] GatorTron-large was not publicly available at the time of our experiments. These were implemented using Hugging Face Transformers library and PyTorch version 1.11 framework.[19,20] RadBERT was our baseline model. To prevent model overfitting and reduce compute time, we used the early-stopping callback parameter available in PyTorch trainer to monitor the metric on validation loss, which stopped training once there was no more improvement observed. Other hyperparameters tuned for each model included learning rate, batch size, and maximum sequence length. Overall model performance was reported using micro average precision, recall, and F1 score.

IE System

For comparison with more traditional NLP methods, we developed a pipeline consisting of three models developed using Spark NLP library version 3.4.2 and a postprocessing step for term normalization.[21] We termed this our IE system. Figure 1A depicts the IE system, which contains

1. Named entity recognition (NER) model for identifying and labeling named entities such as anatomic descriptor, body part, and cancer imaging finding, among others.
2. Assertion status detection model for determining if the named entity (cancer imaging finding) refers to cancerous lesions and to label its probability on the basis of the four-point Likert scale.
3. Relation extraction (RE) model for identifying and extracting three key relations required to determine sites of metastases. Only high- or medium-probability cancer imaging finding were used for REs for metastatic sites.
4. Postprocessing for term normalization. This was required as radiologists often use a variety of terms to describe different body parts and anatomic structures, and typographical errors or variations in spelling can occur in free-text reports.[22] Thus, postprocessing using rule-based logic was iteratively developed by our curation team and applied to normalize extracted chunks to 63 terms for final model predictions. An example is shown in Figure 1B.

The NER model uses a BiLSTM-CNN-Char architecture to automatically detect word and character-level features.[23] The Assertion status detection model is a modified version of the Bi-LSTM framework proposed by Fancellu and Lopez.[24] The RE model uses a BioBERT architecture with a 128-token sequence length to extract and classify instances of relations between named entities.[25] To prevent model overfitting and reduce compute time, we monitor the metric on validation loss and stopped training once there was no more improvement observed. This was implemented using the early-stopping callback parameter available for NER model and manual review of training logs for the assertion and RE model as this parameter was not supported. Other hyperparameters tuned included learning rate and batch size. For maximum sequence length, we used 128 tokens for NER and RE as this is the maximum value supported in Spark NLP version 3.4.2. Details of hyperparameters can be found in the Data Supplement (Table S2). Overall model performance was reported using micro average precision, recall, and F1 score.

Model Validation on External Data Sets

To evaluate how well our language models could generalize to other primary sites and radiology scan modalities, we validated them in three external data sets consisting of a cohort of patients from a molecular tumor board, positron emission tomography CT (PET-CT), and magnetic resonance imaging (MRI) reports. These were across various anatomic regions, including the brain, breast, liver, rectum, abdomen, and pelvis. Micro average precision, recall, and F1 score were used for evaluation. Code used for model training and evaluation is available on GitHub.[26]

Model Inference on a Cohort Study

To illustrate clinical utility, we used the best-performing model to predict metastatic phenotypes for a large cohort study. After excluding cases with synchronous cancers, sites of metastases were identified for each CT report. Frequency of each metastatic site was calculated for each type of primary cancer (eg, number of colorectal cancer patients with lung metastases divided by total number of patients with colorectal cancer in the cohort).

To illustrate metastatic patterns over time, we chose patients with colorectal cancer, the largest group in the cohort study, and analyzed the first site of metastatic spread in both initial stage I-III patients and stage IV colorectal cancer. We classified de novo (stage IV) or recurrent (stages I, II, and III) metastases as well as identified the first metastatic site using manual chart review. For identification of the first metastatic site, we reviewed physician notes as well as the CT scan reports of patients at diagnosis and follow-up. We then performed an analysis to identify the top four most frequent first sites of non-nodal single-organ metastases, as well as the combinations of these four sites. Fisher's exact test was used to compare the difference in proportions of the four most common metastasis sites in de novo versus recurrent patient groups.

RESULTS

Cohort Descriptions

Our cohort for training and testing consisted of 550 patients and 4,522 reports. The three external validation sets included 85, 70, and 73 patients; 491 CT reports, 100 PET-CT, and 100 MRI reports for the molecular tumor board, PET-CT, and MRI cohorts respectively. The large cohort study consisted of 53,838 CT reports obtained from 6,555

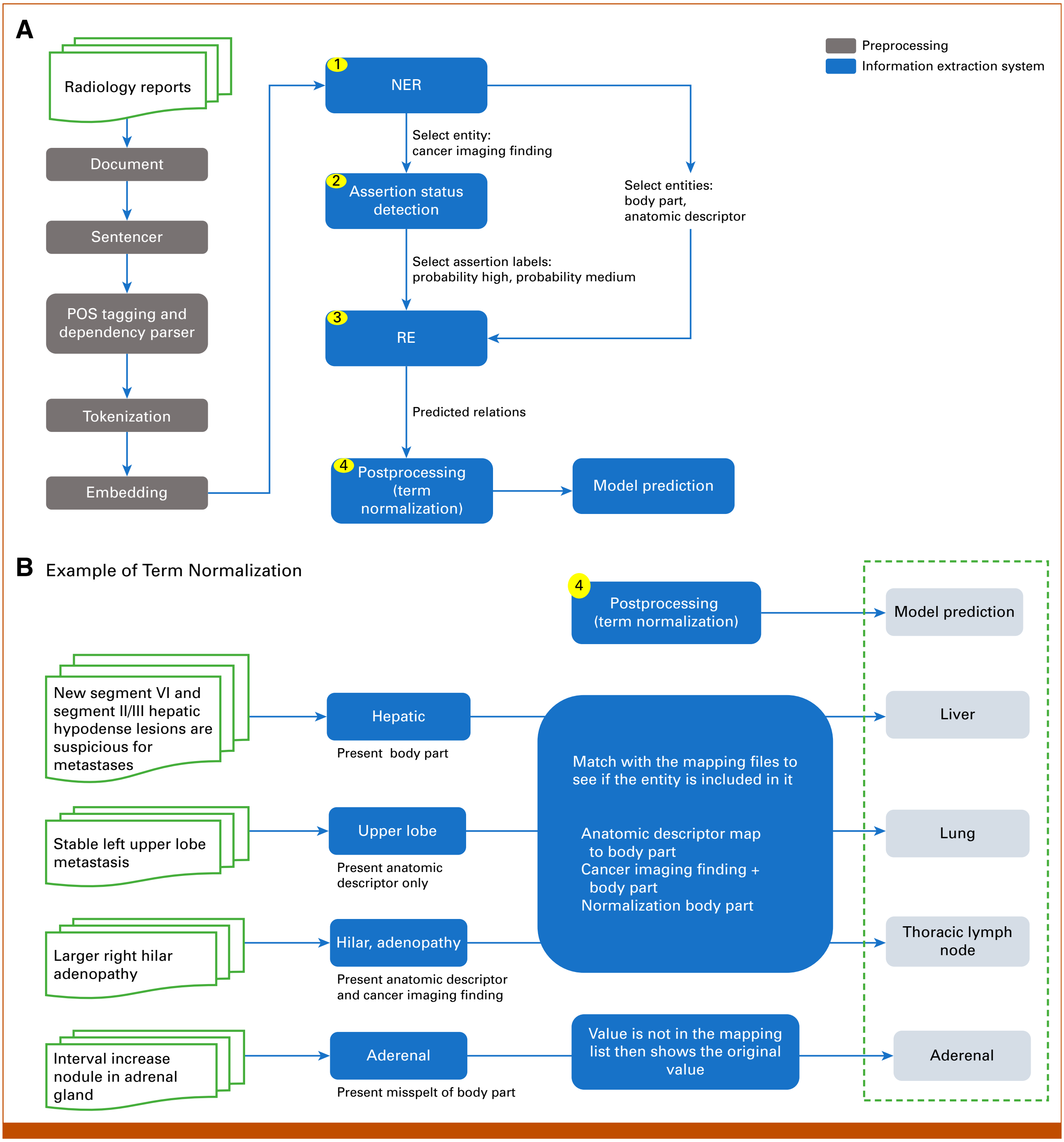
A
Preprocessing
Information extraction system
Radiology reports
Document
Sentencer
POS tagging and dependency parser
Tokenization
Embedding
1 NER
Select entity: cancer imaging finding
2 Assertion status detection
Select assertion labels: probability high, probability medium
Select entities: body part, anatomic descriptor
3 RE
Predicted relations
4 Postprocessing (term normalization)
Model prediction
B Example of Term Normalization
4 Postprocessing (term normalization)
Model prediction
New segment VI and segment II/III hepatic hypodense lesions are suspicious for metastases
Hepatic
Present body part
Liver
Stable left upper lobe metastasis
Upper lobe
Present anatomic descriptor only
Lung
Larger right hilar adenopathy
Hilar, adenopathy
Present anatomic descriptor and cancer imaging finding
Thoracic lymph node
Match with the mapping files to see if the entity is included in it
Anatomic descriptor map to body part
Cancer imaging finding + body part
Normalization body part
Interval increase nodule in adrenal gland
Aderenal
Present misspelt of body part
Value is not in the mapping list then shows the original value
Aderenal

FIG 1. (A) Overview of the information extraction system. (B) Example flowchart of term normalization process. NER, named entity recognition; RE, relation extraction.

patients (Table 1; Data Supplement, Table S3) Baseline characteristics of patients in training, test set, and three external validation sets are listed in Data Supplement (Table S4). This study comprised 21 primary cases, with colorectal, hepatobiliary, and breast cancers being the most commonly identified types in the training cohort. Within the molecular tumor board cohort, there were 16 different primary cases, six of which were not included in the training set (anus, brain, thymus, thyroid, urachus, and uveal melanoma). Anus and appendix cancers were part of the large cohort but were not in the training set (Table 1). The average number of radiology reports per patient was 8.22. The various types of CT, PET-CT, and MRI scans used for this study are presented in Table 2. Average word count for CT reports was 67 in both training and test sets and 100 in the molecular tumor board set. PET-CT and MRI reports had an average word count of 85 and 54 words, respectively.

TABLE 1. Number of Patients and Different Reports in Training Set, Validation Sets, Larger Cohort, and Distribution by Cancer Type

	No. of Patients						No. of CT/PET-CT/MRI Reports					
			External Validation Sets						External Validation Sets			
Cancer Type	Training Set	Test Set	MTB	PET-CT	MRI	Large Cohort	Training Set	Test Set	MTB	PET-CT	MRI	Large Cohort
Anus	0	0	1	0	0	19	0	0	4	0	0	124
Appendix	0	0	0	0	0	40	0	0	0	0	0	393
Brain	0	0	1	0	0	0	0	0	1	0	0	0
Breast	97	8	9	10	24	868	698	82	57	26	35	6,691
Colorectal	146	23	1	31	29	3,565	1,130	160	2	40	36	30,155
Esophagus	3	0	2	0	0	122	18	0	10	0	0	950
Gastric	55	8	0	11	2	510	518	77	0	13	2	4,225
Gynecologic	1	0	3	0	0	234	11	0	18	0	0	1,684
Head and neck	1	0	4	0	0	0	2	0	10	0	0	0
Hepatobiliary	113	11	5	11	16	804	951	80	22	11	25	6,378
Lung	5	2	36	0	0	0	70	35	242	0	0	0
Lymphoma	10	1	0	0	0	0	88	5	0	0	0	0
Pancreatic	46	2	1	7	2	369	353	21	2	10	2	3,020
Prostate	5	0	0	0	0	0	50	0	0	0	0	0
Renal	1	0	2	0	0	0	29	0	8	0	0	0
Sarcoma	11	0	5	0	0	0	138	0	38	0	0	0
Small bowel	1	0	0	0	0	24	6	0	0	0	0	218
Thymus	0	0	1	0	0	0	0	0	10	0	0	0
Thyroid	0	0	9	0	0	0	0	0	33	0	0	0
Urachus	0	0	1	0	0	0	0	0	9	0	0	0
Uveal melanoma	0	0	3	0	0	0	0	0	20	0	0	0
Cancer of unknown primary	0	0	1	0	0	0	0	0	5	0	0	0
Total	495	55	85	70	73	6,555	4,062	460	491	100	100	53,838

Abbreviations: CT, computed tomography; MRI, magnetic resonance imaging; MTB, molecular tumor board; PET-CT, positron emission tomography-CT.

Model Performance

The performance of the IE system and four LLMs on both the training and test sets is presented in Data Supplement (Table S5). The IE system performed the best in the hold-out test set and external molecular tumor board validation set, with F1 scores of 0.93 and 0.89, respectively. For the external PET-CT and MRI validation sets, GatorTron-medium performed best with both F1 scores of 0.86, while the IE system achieved F1 scores of 0.83 and 0.81, respectively (Table 3; Data Supplement, Table S6). The IE system performed well across most sites of metastases, with F1 scores on the test set above 0.8 in 13 of the 18 sites (Table 4). It also performed well across the different cancer types, achieving F1 scores between 0.89 and 0.96 on the test set (Data Supplement, Table S7). F1 scores of the NER, assertion, and RE models in the IE system were 0.91, 0.96, and 0.99, respectively (Data Supplement, Tables S8-S10).

Model Inference on a Cohort Study

We analyzed metastatic patterns for 10 primary cancer types and visualized results using a heatmap. Figure 2 and the Data Supplement (Table S11) demonstrate distribution of various sites of metastases in our cohort. Lung metastases were most frequently seen in esophageal cancers (45.1%), followed by colorectal, hepatobiliary, and breast cancers (all at least 34.4%). Liver metastases were most common in pancreatic, small bowel, and colorectal cancers (56.1%-35.8%). Breast cancers were found to have the highest frequency of bone metastasis (37.0%). Peritoneal metastases were most frequently reported in appendix cancers (62.5%), followed by gynecologic, small bowel, and gastric cancers (all at least 37.1%). Thoracic node metastases were observed most in esophageal (56.6%) and gynecologic (25.6%) cancers. Brain metastases were rarely observed in this cohort (<4.6%).

TABLE 2. Distribution of Imaging Reports in Training, Internal and External Validation, and Large Cohort Sets

	No. of Imaging Reports (%)					
			External Validation Sets			
Imaging Report Type	Training Set	Test Set	MTB	PET-CT	MRI	Large Cohort
CT of abdomen and pelvis	317 (7.8)	29 (6.3)	1 (0.2)	0	0	6,782 (12.6)
CT of brain	201 (4.9)	23 (5.0)	0	0	0	3,315 (6.2)
CT of brain, chest, and abdomen	3 (0.1)	0	82 (16.7)	0	0	59 (0.1)
CT of chest	147 (3.6)	19 (4.1)	2 (0.4)	0	0	3,311 (6.1)
CT of chest, abdomen, and pelvis	3,339 (82.2)	382 (83.0)	405 (82.5)	0	0	39,052 (72.5)
CT of neck	47 (1.2)	7 (1.5)	0	0	0	1,131 (2.1)
CT of neck, chest, and abdomen	4 (0.1)	0	0	0	0	70 (0.1)
CT of paranasal sinuses	4 (0.1)	0	1 (0.2)	0	0	118 (0.2)
PET-CT FDG whole body	0	0	0	98 (98.0)	0	0
PET-CT of gallium 68 dotatate whole body	0	0	0	1 (1.0)	0	0
PET-CT of Ga68 PSMA	0	0	0	1 (1.0)	0	0
MRI of abdomen	0	0	0	0	22 (22.0)	0
MRI of brain	0	0	0	0	21 (21.0)	0
MRI of breast	0	0	0	0	6 (6.0)	0
MRI of pancreas	0	0	0	0	2 (2.0)	0
MRI of liver	0	0	0	0	26 (26)	0
MRI of pelvis	0	0	0	0	18 (18)	0
MRI of rectum	0	0	0	0	5 (5)	0
Total	4,062	460	491	100	100	53,838

Abbreviations: CT, computed tomography; FDG, fluorodeoxyglucose; Ga68 PSMA, gallium 68 prostate-specific membrane antigen; MRI, magnetic resonance imaging; MTB, molecular tumor board; PET-CT, positron emission tomography-CT.

Metastatic patterns in 1,415 patients with stage I-III colorectal cancer with distant recurrences and 889 patients with de novo stage IV colorectal cancer are shown in the Data Supplement (Fig S2). The four most common sites of the first non-nodal, distant metastases in both recurrent and de novo stage IV were liver, lung, peritoneum, and bone metastases. Occurrence of liver metastases was significantly higher in de novo stage IV compared with recurrent patients (29.7% *v* 12.2%; *P* < .001). Conversely, lung metastases were more frequently observed in recurrent compared with de novo stage IV patients (17.2% *v* 7.3%; *P* < .001). Frequency of peritoneal and bone metastases was low in both groups (<5% and <2% respectively). The most frequent combination of these 4 first sites of metastases was lung and liver metastases, with 6.1% in de novo stage IV and 1.2% in recurrent patients.

DISCUSSION

In this study, we successfully developed an IE system with traditional NLP methods, which performed better than four comparator LLMs (GatorTron-medium, GatorTron-base, BioBERT, and RadBERT) in accurately inferring sites of

TABLE 3. Performance of Clinical Large Language Models and IE System on Different Validation Sets

	Validation Sets															
	Test Set				Molecular Tumor Board				PET-CT				MRI			
Model	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
RadBERT	0.73	0.92	0.78	0.84	0.63	0.86	0.70	0.77	0.56	0.80	0.65	0.72	0.62	0.77	0.75	0.76
BioBERT	0.77	0.92	0.82	0.87	0.67	0.87	0.74	0.80	0.64	0.84	0.73	0.78	0.60	0.76	0.74	0.75
GatorTron-base	0.81	0.93	0.86	0.89	0.73	0.89	0.80	0.84	0.70	0.88	0.78	0.82	0.70	0.84	0.82	0.83
GatorTron-medium	0.83	0.93	0.88	0.91	0.74	0.90	0.81	0.85	0.76	0.88	0.84	**0.86**	0.75	0.87	0.85	**0.86**
IE system	0.88	0.95	0.92	**0.93**	0.80	0.89	0.89	**0.89**	0.71	0.87	0.79	0.83	0.68	0.79	0.83	0.81

NOTE. Results in bold indicate the best F1 scores in the test and validation sets.

Abbreviations: Acc, accuracy; F1, F1 score; IE, information extraction; MRI, magnetic resonance imaging; PET-CT, positron emission tomography-computed tomography; Prec, precision; Rec, recall.

TABLE 4. Performance of Information Extraction System by Sites of Metastases in Test Set

Site of Metastases	True Positives	False Positives	False Negatives	Acc	Prec	Rec	F1
Abdominopelvic node	68	4	12	0.81	0.94	0.85	0.89
Adrenal	13	6	1	0.65	0.68	0.93	0.79
Axillary node	54	2	2	0.93	0.96	0.96	0.96
Bone	72	1	2	0.96	0.99	0.97	0.98
Brain	4	0	2	0.67	1.00	0.67	0.80
Cervical node	44	1	6	0.86	0.98	0.88	0.93
Liver	102	1	4	0.95	0.99	0.96	0.98
Omentum	7	1	2	0.70	0.88	0.78	0.82
Ovary	8	0	1	0.89	1.00	0.89	0.94
Pancreatic	4	4	4	0.33	0.50	0.50	0.50
Peritoneum	41	0	2	0.95	1.00	0.95	0.98
Pleural	11	0	1	0.92	1.00	0.92	0.96
Lung	111	4	8	0.90	0.97	0.93	0.95
Kidney	2	0	0	1.00	1.00	1.00	1.00
Spleen	2	2	0	0.50	0.50	1.00	0.67
Thoracic node	56	1	6	0.89	0.98	0.90	0.94
Thyroid	1	1	0	0.50	0.50	1.00	0.67
Ureter	1	0	1	0.50	1.00	0.50	0.67
Total	601	28	54	—	—	—	—

Abbreviations: Acc, accuracy; F1, F1 score; Prec, precision; Rec, recall.

metastatic disease. The superior performance was consistent across test set and the external molecular tumor board validation set. However, GatorTron-medium performed better for the external PET-CT and MRI validation sets.

Accuracies of 0.88 and 0.83 achieved by the IE system and GatorTron-medium, respectively, on the test set for prediction of sites of metastases are comparable with previous work. For example, Karimi et al[11] developed XGBoost classifier models that had accuracies between 0.79 and 0.97. In

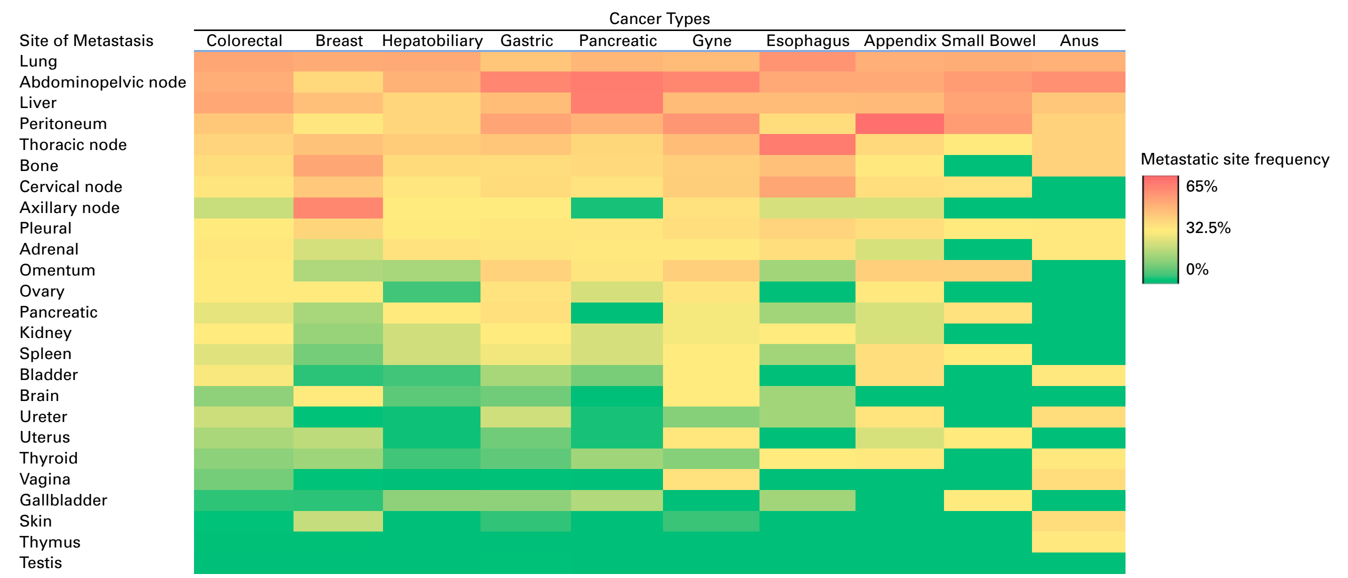

FIG 2. Frequency of various sites of metastatic disease among 6,555 patients stratified according to cancer types. Frequency was calculated by the number of patients of the same primary cancer with the same site of metastases (eg, colorectal cancer patients with lung metastases) divided by the total number of patients of that primary cancer (ie, patients with colorectal cancer).

addition, the BERT transformer model by Do et al[9] attained accuracies ranging from 0.93 to 0.99. Although these binary classifiers may achieve higher accuracies for some sites of metastases, the main advantage of our demonstrated approaches over previous work is that multiple sites of metastases can be extracted simultaneously with a single model, as opposed to multiple organ-specific models, which need to be run repeatedly for every site of metastases.[9,27,28] Compared with LLMs, the IE system carries two additional key advantages. First, run time for our IE system was shorter at 6 minutes per 1,000 reports on a single graphics processing unit (Nvidia RTX 3090), compared with 8-80 minutes per 1,000 reports for the LLMs. Second, the IE system is more explainable through reviewing stepwise predictions of the NER, assertion, and RE models.

GatorTron-medium's superior performance in the PET-CT and MRI sets could be attributed to its ability to use additional information, such as the fluorodeoxyglucose avidity of the lesion, which may not have been described in CT reports. Compared with the IE system, which was customized for a specific task in a specific context (ie, sites of metastases prediction in CT reports), LLMs exhibit more versatility when assessed across a broader spectrum of related tasks (eg, the same task but in PET and MRI reports), particularly as the models expand in size. Larger LLMs likely possess a greater capacity to identify the context of new terms used in different modalities. This finding highlights the importance of assembling representative training sets for NLP models as even subtle differences in terminology can significantly influence the performance and accuracy of models. It is also worth highlighting that all the LLMs we chose had transformer encoder architectures that are best suited for our multilabel classification task that required NER and an understanding of full sentences.[29] This is in contrast to transformer decoder architectures (eg, generative pretrained transformers), which are better for text generation.[30]

Our model demonstrates potential for transfer learning between cancer types as shown in the external validation set of molecular tumor board patients, which consisted of six tumor types were not present in the training set (Data Supplement, Table S6). This could be because the anatomic regions of metastases and language used to describe them are largely similar across cancer types for the same imaging modality. This is valuable, given the long-tail challenge in this task: a few very common primary cancer types but also a fair number of rarer cancer types encountered in clinical practice.

Our study illustrated the application of NLP methods in analyzing extensive radiology data sets to deduce metastatic patterns and enrich clinical research databases. This can offer a broad high-level overview of research data, enabling researchers to gain initial insights, such as identifying the most commonly affected organs at a specific metastatic site in their research database. These preliminary observations (eg, differences in proportion of lung metastases in de novo *v* recurrent colorectal patients) can then undergo further hypothesis testing. Our results concur with previous large-scale studies, such as a previous autopsy series of over 1,000 patients with solid malignancies where the five most common organs for metastatic disease were also the liver, lung, bones, pleura, and peritoneum.[6,31] These NLP models could be deployed prospectively for detection of differences in longitudinal metastatic trajectories as illustrated by our findings for patients with colorectal cancer. This could lead to better understanding of biologic basis of metastases, organotropism, and resistance when combined with temporal multiomic analysis.[3,4] Such models could also be used in clinical care to flag up oligometastatic findings during surveillance, given timely surgery or radiation therapy may be viable treatment options that could increase survival rates for some cancers.[32,33] Finally, these models could be used to provide relevant information for identification of potential patients for clinical trials (eg, presence of brain or visceral metastases). However, clinical trial matching is complex,[34,35] and NLP may just be one component of a successful solution.[36-38]

Limitations of this study include radiology reports from a limited number of centers over a long period, and generalizability of these models to other hospitals remains to be seen. The development of the IE system also required significant domain expertise to assist with data curation, which may limit reproducibility of this approach in resource-constrained settings. Our model performed well in predicting most metastatic sites for primary tumors not encompassed in the training or test sets. However, it faced difficulties in accurately predicting specific types of primary tumors (eg, head and neck) that exhibit complex anatomic features and are closely situated to adjacent organs as well as some rarer sites of metastases (ie, pancreatic, spleen, thyroid, and ureter). To overcome these challenges, future work should prioritize incorporating additional training data specifically tailored to handle the complexities associated with these particular primary types and enrich them with cases of rarer sites of metastases. In addition, our IE system has been primarily trained on CT scans of the chest, abdomen, and pelvis. Future work should incorporate additional anatomic sites as well as other scan modalities to improve generalizability of our model.

In conclusion, we developed an IE system that can accurately infer sites of metastases in multiple primary cancers from free-text radiology reports. It has explainable methods and can perform better than clinical LLMs. The inferred metastatic phenotypes could be used to enrich cancer research databases and clinical trial matching, and identify potential patients for oligometastatic interventions.

AFFILIATIONS

[1]Division of Medical Oncology, National Cancer Centre Singapore, Singapore, Singapore
[2]NUS Yong Loo Lin School of Medicine, Singapore, Singapore
[3]Data and Computational Science Core, National Cancer Centre Singapore, Singapore, Singapore
[4]Singapore Duke-NUS Medical School, Singapore, Singapore
[5]Division of Radiation Oncology, National Cancer Centre Singapore, Singapore, Singapore
[6]Division of Clinical Trials and Epidemiological Sciences, National Cancer Centre Singapore, Singapore, Singapore
[7]Division of Oncologic Imaging, National Cancer Centre Singapore, Singapore, Singapore
[8]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY

CORRESPONDING AUTHOR

Ryan Shea Ying Cong Tan, MBBS, MTec; e-mail: gmstycrs@duke-nus.edu.sg.

SUPPORT

Supported by A*STAR under its IAF-PP program (award H18/01/a0/019) and Singapore Health Services under the Singhealth Duke-NUS Oncology ACP Program (award 08/FY2020/EX(SL)/74-A150).

AUTHOR CONTRIBUTIONS

Conception and design: Guat Hwa Low, Choon Hua Thng, Iain Bee Huat Tan, Ryan Shea Ying Cong Tan
Financial support: Iain Bee Huat Tan, Ryan Shea Ying Cong Tan
Administrative support: See Boon Tay, Gillian Jing En Wong, Fun Loon Leong, Melvin Lee Kiang Chua, Iain Bee Huat Tan, Ryan Shea Ying Cong Tan
Provision of study materials or patients: See Boon Tay, Fun Loon Leong, Melvin Lee Kiang Chua, Daniel Shao Weng Tan, Iain Bee Huat Tan, Ryan Shea Ying Cong Tan
Collection and assembly of data: See Boon Tay, Guat Hwa Low, Gillian Jing En Wong, Han Jieh Tey, Fun Loon Leong, Daniel Shao Weng Tan, Choon Hua Thng, Iain Bee Huat Tan, Ryan Shea Ying Cong Tan
Data analysis and interpretation: See Boon Tay, Guat Hwa Low, Fun Loon Leong, Constance Li, Melvin Lee Kiang Chua, Daniel Shao Weng Tan, Iain Bee Huat Tan, Ryan Shea Ying Cong Tan
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Han Jieh Tey
Employment: National Cancer Centre Singapore

Melvin Lee Kiang Chua
Leadership: Digital Life Line
Stock and Other Ownership Interests: Digital Life Line, BeiGene
Honoraria: Janssen Oncology, Varian Medical Systems
Consulting or Advisory Role: Janssen Oncology, Merck Sharp & Dohme, ImmunoSCAPE, Telix Pharmaceuticals, IQVIA, BeiGene, PVmed
Speakers' Bureau: AstraZeneca, Bayer, Janssen
Research Funding: PVmed, Decipher Biosciences, EVYD Technology, MVision, BeiGene
Patents, Royalties, Other Intellectual Property: High Sensitivity Lateral Flow Immunoassay For Detection of Analyte in Samples (10202107837T), Singapore (Danny Jian Hang Tng, Chua Lee Kiang Melvin, Zhang Yong, Jenny Low, Ooi Eng Eong, Soo Khee Chee)
Uncompensated Relationships: Alice's Arc

Daniel Shao Weng Tan
Honoraria: Takeda (Inst), Novartis (Inst), Roche (Inst), Pfizer (Inst)
Consulting or Advisory Role: Merck (Inst), AstraZeneca (Inst), Roche (Inst), Pfizer (Inst), Amgen (Inst), DKSH (Inst), Bayer (Inst)
Research Funding: Novartis (Inst), GlaxoSmithKline (Inst), AstraZeneca (Inst), ACM Biolabs (Inst), Pfizer (Inst)
Travel, Accommodations, Expenses: Pfizer, Boehringer Ingelheim, Roche

Iain Bee Huat Tan
Honoraria: Amgen, Roche, Merck Serono, MSD, BMS GmbH & Co KG, Guardant Health, Novartis
Consulting or Advisory Role: Amgen, Roche, Merck Serono, MSD, Novartis
Research Funding: MSD
Travel, Accommodations, Expenses: Merck Serono, Amgen, Roche

Ryan Shea Ying Cong Tan
Stock and Other Ownership Interests: Abbott Laboratories, AstraZeneca, Becton Dickinson, Edwards Lifesciences, Intuitive Surgical, Johnson & Johnson/MedTech, Medtronic
Honoraria: Novartis
Travel, Accommodations, Expenses: Merck

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

The authors thank Ira Jiale Teh, Megan Jing Yi Yeo, Natalie Rui Lin Kuan, and Lynn Wee Lin Khoo for their invaluable support in the completion of this project. Additionally, the authors express appreciation to Hong Qi Tan and Wen Long Nei for their technical support and guidance. The authors are also grateful to the contribution by General Electric Healthcare on knowledge in natural language processing.

REFERENCES

1. Jin J, Gao Y, Zhang J, et al: Incidence, pattern and prognosis of brain metastases in patients with metastatic triple negative breast cancer. BMC Cancer 18:446, 2018
2. Dillekås H, Rogers MS, Straume O: Are 90% of deaths from cancer caused by metastases? Cancer Med 8:5574-5576, 2019
3. Fares J, Fares MY, Khachfe HH, et al: Molecular principles of metastasis: A hallmark of cancer revisited. Signal Transduct Target Ther 5:28, 2020
4. Gao Y, Bado I, Wang H, et al: Metastasis organotropism: Redefining the congenial soil. Dev Cell 49:375-391, 2019
5. Ban J, Fock V, Aryee DNT, et al: Mechanisms, diagnosis and treatment of bone metastases. Cells 10:2944, 2021
6. Budczies J, von Winterfeld M, Klauschen F, et al: The landscape of metastatic progression patterns across major human cancers. Oncotarget 6:570-583, 2015
7. Krug D, Vonthein R, Illen A, et al: Metastases-directed radiotherapy in addition to standard systemic therapy in patients with oligometastatic breast cancer: Study protocol for a randomized controlled multi-national and multi-center clinical trial (OLIGOMA). Clin Transl Radiat Oncol 28:90-96, 2021
8. Pfannschmidt J, Dienemann H: Surgical treatment of oligometastatic non-small cell lung cancer. Lung Cancer 69:251-258, 2010

9. Do RKG, Lupton K, Causa Andrieu PI, et al: Patterns of metastatic disease in patients with cancer derived from natural language processing of structured CT radiology reports over a 10-year period. Radiology 301:115-122, 2021
10. Pons E, Braun LM, Hunink MG, et al: Natural language processing in radiology: A systematic review. Radiology 279:329-343, 2016
11. Karimi YH, Blayney DW, Kurian AW, et al: Development and use of natural language processing for identification of distant cancer recurrence and sites of distant recurrence using unstructured electronic health record data. JCO Clin Cancer Inform 10.1200/CCI.20.00165
12. Sun S, Lupton K, Batch K, et al: Natural language processing of large-scale structured radiology reports to identify oncologic patients with or without splenomegaly over a 10-year period. JCO Clin Cancer Inform 10.1200/CCI.21.00104
13. Yan A, McAuley J, Lu X, et al: RadBERT: Adapting transformer-based language models to radiology. Radiol Artif Intell 4:e210258, 2022
14. Lee J, Yoon W, Kim S, et al: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36:1234-1240, 2019
15. Yang X, Chen A, PourNejatian N, et al: A large language model for electronic health records. NPJ Digit Med 5:194, 2022
16. Klie J-C, Bugert M, Boullosa B, et al: The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. Presented at Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, August 20-26, 2018
17. Harada T, Abe T, Kato F, et al: Five-point Likert scaling on MRI predicts clinically significant prostate carcinoma. BMC Urol 15:91, 2015
18. Claessens YE, Debray MP, Tubach F, et al: Early chest computed tomography scan to assist diagnosis and guide treatment decision for suspected community-acquired pneumonia. Am J Respir Crit Care Med 192:974-982, 2015
19. Wolf T, Debut L, Sanh V, et al: Hugging Face's Transformers: State-of-the-art natural language processing. ArXiv191003771 Cs. http://arxiv.org/abs/1910.03771
20. Paszke A, Gross S, Massa F, et al: PyTorch: An imperative style, high-performance deep learning library. ArXiv191201703 Cs Stat. http://arxiv.org/abs/1912.01703
21. Kocaman V, Talby D: Spark NLP: Natural language understanding at scale. Softw Impacts 8:100058, 2021
22. Tahmasebi AM, Zhu H, Mankovich G, et al: Automatic normalization of anatomical phrases in radiology reports using unsupervised learning. J Digit Imaging 32:6-18, 2019
23. Chiu J, Nichols E: Named entity recognition with bidirectional LSTM-CNNs. Trans Assoc Comput Linguist 4:357-370, 2016
24. Fancellu F, Lopez A: Neural networks for negation scope detection. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, August 12, 2016, pp 495-504
25. Haq HU, Kocaman V, Talby D: Deeper clinical document understanding using relation extraction, 2021. ArXiv:211213259 Cs. https://arxiv.org/abs/2112.13259
26. Repository for information extraction (IE) system and instructions for fine tuning of clinical large language models. Github. https://github.com/nccsnlp/sites_of_metastases
27. Swaminathan KK, Mendonca E, Mukherjee P, et al: A deep learning model using NLP to identify metastatic patients from unstructured notes. Presented at the American Society of Clinical Oncology (ASCO) 2020, ConcertAI, May 19, 2020
28. Kehl KL, Elmarakeby H, Nishino M, et al: Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. JAMA Oncol 5:1421-1429, 2019
29. Encoder Models: Hugging Face. https://huggingface.co/learn/nlp-course/chapter1/5?fw=pt
30. Radford A, Narasimhan K, Salimans T et al: Improving language understanding by generative pre-training, 2018. https://paperswithcode.com/paper/improving-language-understanding-by
31. Disibio G, French SW: Metastatic patterns of cancers: Results from a large autopsy study. Arch Pathol Lab Med 132:931-939, 2008
32. Kitano Y, Hayashi H, Matsumoto T, et al: Borderline resectable for colorectal liver metastases: Present status and future perspective. World J Gastrointest Surg 13:756-763, 2021
33. Fode MM, Høyer M: Survival and prognostic factors in 321 patients treated with stereotactic body radiotherapy for oligo-metastases. Radiother Oncol 114:155-160, 2015
34. Kirshner J, Cohn K, Dunder S, et al: Automated electronic health record-based tool for identification of patients with metastatic disease to facilitate clinical trial patient ascertainment. JCO Clin Cancer Inform 10.1200/CCI.20.00180
35. Jain NM, Culley A, Micheel CM, et al: Learnings from precision clinical trial matching for oncology patients who received NGS testing. JCO Clin Cancer Inform 10.1200/CCI.20.00142
36. Savova GK, Danciu I, Alamudun F, et al: Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Res 79:5463-5470, 2019
37. Alexander M, Solomon B, Ball DL, et al: Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients. JAMIA Open 3:209-215, 2020
38. Haddad T, Helgeson JM, Pomerleau KE, et al: Accuracy of an artificial intelligence system for cancer clinical trial eligibility screening: Retrospective pilot study. JMIR Med Inform 9:e27767, 2021