

Advancing equity in breast cancer care: natural language processing for analysing treatment outcomes in under-represented populations

Jung In Park ¹, Jong Won Park,² Kexin Zhang,³ Doyop Kim⁴

To cite: Park JI, Park JW, Zhang K, *et al.* Advancing equity in breast cancer care: natural language processing for analysing treatment outcomes in under-represented populations. *BMJ Health Care Inform* 2024;**31**:e100966. doi:10.1136/bmjhci-2023-100966

Received 16 November 2023
Accepted 21 June 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹University of California Irvine, Irvine, California, USA

²Yonsei Cancer Center, Yonsei University College of Medicine, Seoul, South Korea

³Donald Bren School of Information & Computer Sciences, University of California Irvine, Irvine, California, USA

⁴Independent Researcher, Irvine, California, USA

Correspondence to

Dr Jung In Park;
junginp@uci.edu

ABSTRACT

Objective The study aimed to develop natural language processing (NLP) algorithms to automate extracting patient-centred breast cancer treatment outcomes from clinical notes in electronic health records (EHRs), particularly for women from under-represented populations.

Methods The study used clinical notes from 2010 to 2021 from a tertiary hospital in the USA. The notes were processed through various NLP techniques, including vectorisation methods (term frequency-inverse document frequency (TF-IDF), Word2Vec, Doc2Vec) and classification models (support vector classification, K-nearest neighbours (KNN), random forest (RF)). Feature selection and optimisation through random search and fivefold cross-validation were also conducted.

Results The study annotated 100 out of 1000 clinical notes, using 970 notes to build the text corpus. TF-IDF and Doc2Vec combined with RF showed the highest performance, while Word2Vec was less effective. RF classifier demonstrated the best performance, although with lower recall rates, suggesting more false negatives. KNN showed lower recall due to its sensitivity to data noise.

Discussion The study highlights the significance of using NLP in analysing clinical notes to understand breast cancer treatment outcomes in under-represented populations. The TF-IDF and Doc2Vec models were more effective in capturing relevant information than Word2Vec. The study observed lower recall rates in RF models, attributed to the dataset's imbalanced nature and the complexity of clinical notes.

Conclusion The study developed high-performing NLP pipeline to capture treatment outcomes for breast cancer in under-represented populations, demonstrating the importance of document-level vectorisation and ensemble methods in clinical notes analysis. The findings provide insights for more equitable healthcare strategies and show the potential for broader NLP applications in clinical settings.

INTRODUCTION

Breast cancer is the second leading cause of cancer deaths in US women, comprising 30% of new female cancer diagnoses.¹ It is the most common cancer across all ethnic

WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ Before this study, it was understood that breast cancer is the most prevalent cancer affecting women of all ethnic groups in the USA, with disparities in outcomes among different racial and ethnic groups.
- ⇒ The widespread use of electronic health records and advances in natural language processing (NLP) offered avenues for improved patient care through detailed data analysis; however, there was a gap in automated, detailed analysis of clinical notes, especially for breast cancer treatment outcomes in women from under-represented populations, necessitating this study.

WHAT THIS STUDY ADDS

- ⇒ This study contributes by developing a robust NLP pipeline to analyse clinical notes for breast cancer treatment outcomes in under-represented populations.
- ⇒ It demonstrates the effectiveness of specific text vectorisation methods (term frequency-inverse document frequency and Doc2Vec) combined with classification models, particularly random forest (RF), in extracting relevant treatment outcome data from clinical notes.
- ⇒ The study also reveals the challenges in achieving high recall rates in predictive models, highlighting the complexity of clinical data and the need for specialised NLP approaches.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- ⇒ This study has significant implications for future research, clinical practice and health policy.
- ⇒ It underscores the potential of NLP in enhancing the understanding of breast cancer treatment outcomes, particularly for under-represented groups, thereby guiding more personalised and equitable healthcare strategies.
- ⇒ The findings could influence policy decisions related to healthcare data management and the integration of NLP techniques in clinical settings.
- ⇒ Moreover, the developed pipeline can be adapted for other clinical NLP applications, potentially broadening its impact beyond breast cancer research.

groups in the USA, but disparities exist in outcomes.² While white women have higher incidence rates, black and Hispanic women face higher mortality rates.^{3,4} Additionally, the incidence is increasing rapidly among Asian/Pacific Islanders and American Indian/Alaska Natives.⁴

The widespread adoption of electronic health records (EHRs) offers promising opportunities for predicting future events using large amounts of data.⁵ Especially, unstructured clinical notes contain important information often not captured in structured, coded formats.⁶ For example, patient-reported outcomes from patients with cancer are often not captured in structured EHRs, but is increasingly found in unstructured or semi-structured text formats within EHRs, facilitating translational research and personalised care.⁷⁻⁹ One common approach in clinical text analysis involves using a rule-based natural language processing (NLP) algorithm that leverages distinct medical keywords from clinical texts.^{10,11} Specifically, with the advancements in neural language modelling, integrating neural networks with features extracted from this rule-based NLP method can be achieved by using word embedding models for feature extraction.¹² This approach allows for building a fully neural network-based pipeline that combines embedding models with supervised learning algorithms.¹³

In cancer research, incorporating clinical notes into analyses is crucial for capturing information on comprehensive symptoms and side effects that patients experience,¹⁴ as it can provide insights into monitoring and individualised symptom management. Several studies have investigated breast cancer treatment outcomes using clinical notes and NLP¹⁴⁻¹⁶; however, research that specifically aims the capture of treatment side effects and patient-reported outcomes in patients with breast cancer from under-represented populations remains sparse. Addressing this research gap is important, because these populations face unique health disparities that impact treatment outcomes and patient care. Understanding these specific challenges and barriers enables the development of targeted interventions to mitigate disparities and enhance health outcomes. There is a clear need for an automated tool to capture symptoms and side effects from clinical notes, enabling accurate symptom management and tailored nursing care planning for those patients from under-represented populations.

The goal of this study was to develop NLP algorithms to automate the knowledge extraction process for patient-centred breast cancer treatment outcomes from clinical notes, aiming to gain valuable insights to improve care for those from under-represented populations. Specifically, we aimed to compare the effectiveness of these algorithms in providing scientific evidence for their use in the care of patients with breast cancer from under-represented populations.

METHODS

To harness the full potential of large health datasets from the EHRs and unique application of NLP techniques, we sourced EHR clinical notes dated 1 January 2010 to 31 August 2021 at a tertiary hospital in the USA, selecting patients who met the following criteria: (1) women from under-represented populations (Hispanic, American Indian or Alaska Native, Asian, black or African-American, Native Hawaiian or Other Pacific Islander or multiple race); (2) aged 18 years or greater; (3) diagnosed with invasive breast cancer; (4) had at least one follow-up visit at the medical centre after breast cancer treatment (ie, surgery, radiation therapy, chemotherapy, endocrine therapy or hormone therapy). We excluded the patients who were not followed up at the medical centre.

Overview of the NLP pipeline

In this study, we developed a classification model to predict a binary outcome: whether a side effect was observed in relation to breast cancer treatment, based on the text within a clinical note. Our approach involved a multistep process, as illustrated in [figure 1](#). The process began with raw clinical notes from which text was extracted to train and test the downstream models. The extracted texts underwent preprocessing to ensure they were clean and normalised. Following preprocessing, the cleaned text corpus was used for text vectorisation. Additionally, we randomly sampled notes and had them annotated by clinical experts. After annotation, the texts were mapped into a feature vector space (vectorisation). We then selected the most impactful features and reduced the feature dimension (feature selection) to train a conventional classifier and predict the outcome using this feature vector. Subsequent sections provide a detailed description of each step involved.

Data preprocessing and annotation

To prepare text data for the NLP process, it must undergo preprocessing. This involves standard NLP cleaning techniques such as removing numbers, special characters and duplicated words; performing word tokenisation; removing stop words and applying stemming.¹⁷ Once cleaned, these text data serve as a corpus to train a vectorisation model that converts input text into numerical form (feature vector). This vectorisation can proceed without explicit document annotation, relying on the text corpus of the clinical notes. In contrast, expert annotations are crucial for the classification phase, making it a supervised learning task. Notes were labelled as positive if they referenced side effects or symptoms of breast cancer treatment, adhering to guidelines from the American Cancer Society and American Society of Clinical Oncology.¹⁸ A clinical expert annotated 100 notes, which were randomly selected from the original texts. Subsequently, the annotated data were divided into training and test sets using a 7:3 ratio.

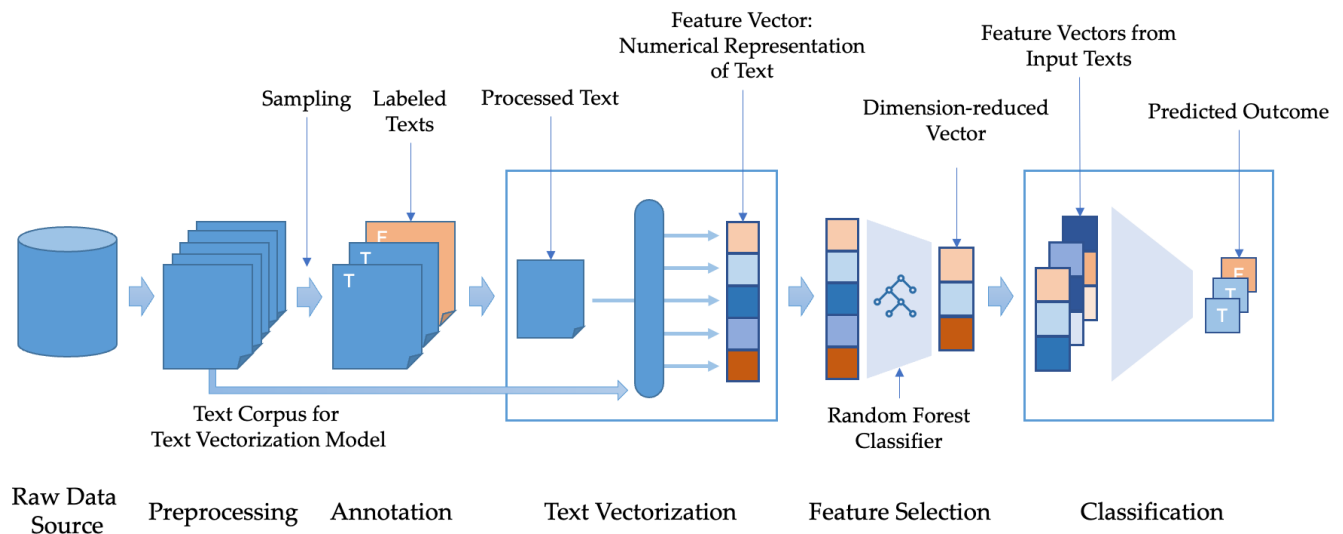


Figure 1 Overview of the natural language processing pipeline. T, true label; F, false label of clinical notes.

Text vectorisation

The texts were converted into a set of numerical values—a vector that represents a given text. We used three different vectorisation approaches—term frequency-inverse document frequency (TF-IDF),¹⁹ Word2Vec²⁰ and Doc2Vec²¹—and compared their performance with different predictive models (text vectorisation step in figure 1).

TF-IDF measures a word's importance in a text by computing its term frequency, indicating the word's relative frequency in a document.¹⁹ This method is effective for assessing word relevance in document queries. Word2Vec vectorises text using a neural network to create word embeddings, mapping words to vectors.²⁰ It employs a sliding window technique, using either the continuous bag-of-words (CBOW) method to predict a word from its context or the skip-gram method to predict context words from a given word. Doc2Vec, a generalised Word2Vec, vectorises entire paragraphs or documents directly into single vectors, bypassing the averaging step required in Word2Vec.²¹ It offers two algorithms: distributed memory (DM) and distributed bag of words (DBOW).²² Figure 2 shows the Word2Vec and Doc2Vec algorithms.

Predictive modelling

After the texts were vectorised, the rows of numerically encoded features for both the training and test sets were prepared. We performed feature selection to filter out features that did not positively contribute to the classification task. This step further reduced the feature dimension, resulting in a more compact space. We trained a random forest (RF) classifier to determine the top relevant features for each text vectoriser (feature selection step in figure 1).

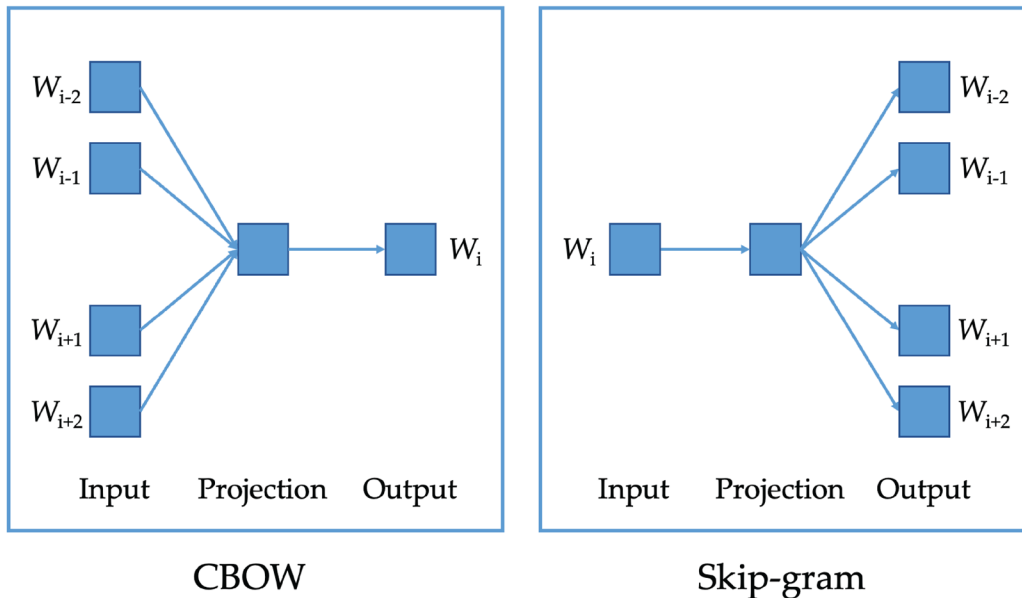
The transformed training set was used to train the predictive models using multiple classification methods

(classification step in figure 1). We used three different classification approaches: support vector classification (SVC), K-nearest neighbours (KNN) and RF. These approaches spanned a wide variety of classifier categories, including support vector machines, non-parametric methods and ensemble methods, enabling us to evaluate a broader spectrum of model performance. All of these methods were supervised learning techniques; therefore, we used the annotated training set, composed of 70 clinical notes, to train each model.

The SVC finds a hyperplane that maximises the margin between the nearest data points of each label, with hyperparameters tuned for optimal separation.²³ KNN classifies by voting among the 'k' nearest training data points to an input query, leading to larger models with more data.^{24 25} RF, an ensemble of decision trees, combines their predictions to reduce overfitting and variance, using moderately tuned hyperparameters for peak performance.²⁶ We chose the hyperparameter set with moderate parameter tuning to maximise model performance and trained an RF model with the same feature-label pairs from the training set to build a classifier.

We performed a random search combined with fivefold cross-validation to determine the optimal parameters for SVC, KNN and RF methods. Random hyperparameter search randomly selects values from predefined ranges or distributions to evaluate model performance. This is typically done using techniques such as k-fold cross-validation, where the training set is further divided into k-folds, and the model is trained and evaluated on different subsets of data, with each fold used as the validation set once. Then the model is trained and tested multiple times with different hyperparameter values to obtain an estimate of its performance.^{27 28}

Word2Vec Algorithms



Doc2Vec Algorithms

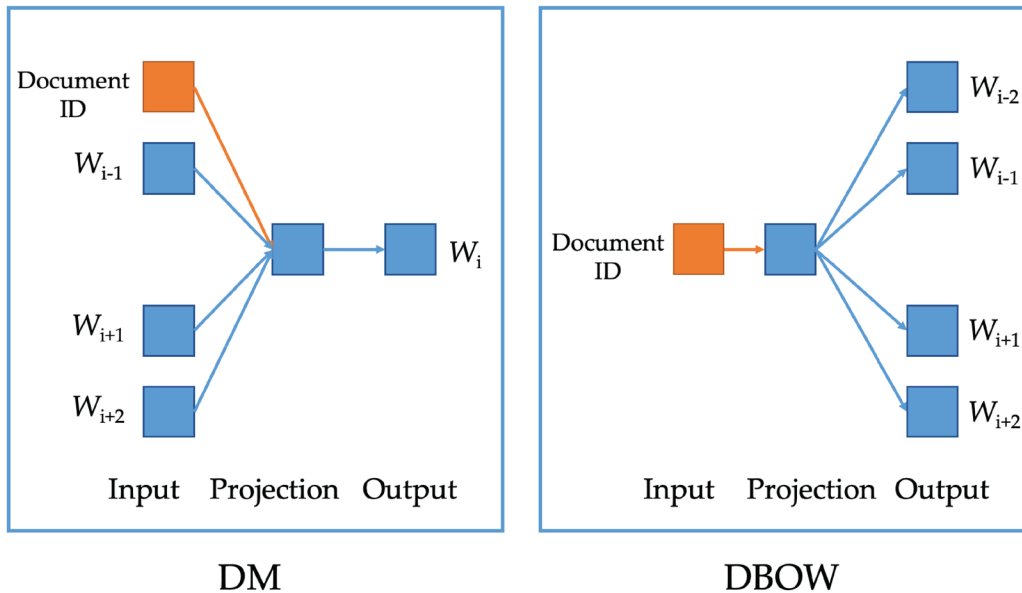


Figure 2 Word2Vec and Doc2Vec algorithms (W_i represents i -th word in a given text). CBOW, continuous bag of words; DBOW, distributed bag of words; DM, distributed memory.

We used the NLTK library²⁹ for text cleaning, Scikit-Learn for data splitting, TF-IDF vectorisation, predictive modelling (SVC, KNN and RF), random search with cross-validation and evaluation and the Gensim library for Word2Vec and Doc2Vec implementation.³⁰

Model evaluation

We used three commonly used performance metrics for model evaluation: precision, recall and area under the receiver operating characteristic curve (AUC). Precision gauges the model's accuracy in predicting positive

classes, aiming to reduce false positives. Recall measures the model's success in identifying actual positives, targeting the reduction of false negatives. AUC reflects the model's ability to differentiate between classes across various thresholds, with higher values denoting greater discrimination.

RESULTS

Among the 1000 clinical notes we collected, 100 were randomly selected and annotated by a clinical expert, while the remaining 900 were used to build the text corpus. We found 41 positive notes and 59 negative notes from these 100 annotated notes. We divided the annotated notes into training and test sets (using random selection of 70 and 30 notes, respectively) for modelling. The training set included 27 positive samples, whereas the test set had 14 due to random selection. The annotated dataset comprised 41% of positive labels. The distribution of positive labels was 39% in the training set and 47% in the test set, closely reflecting the entire dataset. We used the 900 unannotated notes and 70 training notes (970 in total) to build our text corpus in the text vectorisation model for the final analysis. We identified 13 029 unique words after the stemming process¹⁸ among the 970 clinical notes selected for training text vectorisation (embedding) model. The mean value was 657.8, and the SD was 438.0. The minimum value recorded was 8, and the maximum was 2721. The 25th percentile was 372.5, the median (50th percentile) was 619.0 and the 75th percentile was 857.8.

We began by using 970 clinical notes as the corpus input for the TF-IDF model, transforming these notes into vectorised features for training and test sets. The n-gram range was set from 1–3 g, resulting in an output feature dimension of 408 791 for the training set. Similarly, we used the same corpus to train a Word2Vec word embedding model, following the TF-IDF approach. After

training, each word in a note was converted into a vector, and each note was represented by the average of these vectors.

For the Doc2Vec approach, we trained a word-embedding model with the same set of clinical notes, treating each note as a document in the Doc2Vec framework. This enabled us to infer document vectors for each note, which were then used in training predictive models. Both Word2Vec and Doc2Vec models were assigned a feature size of 2000. In the Word2Vec model, the CBOW approach was preferred over Skip-gram due to its superior performance, while for the Doc2Vec model, we chose the DM model over the DBOW method. A window size of three was selected for both models. The hyperparameters for these models are detailed in [table 1](#).

Feature selection is a crucial step in machine learning model development, as it helps identify the most relevant features or variables that contribute to a model's prediction performance. We employed a selection-by-model approach for feature selection after training the vectorisers. In this method, an intermediate model is trained to rank the importance of features based on their impact on the overall accuracy or performance of the model. Specifically, we trained an intermediate RF classifier to rank the importance of features based on their contribution to maximising the accuracy of the classifier. The RF classifier was chosen for its ability to handle non-linearity, interactions and most importantly, its ability to provide feature importance estimation. Then we selected the top 300 features ranked by the RF classifier across all text vectorisation models to balance between capturing relevant information and avoiding overfitting or issues with high-dimensional data.

We performed a random search with fivefold cross-validation to determine the optimal parameters for each model. The hyperparameters used in the random search are listed in [table 1](#). The random search keeps the

Table 1 Text vectoriser classifiers hyperparameters for each text vectorisation model

Text vectoriser hyperparameters		
TF-IDF	n-gram range: 1–3; max document frequency: 1.0; min document frequency count: 1	
Word2Vec	Features size: 2000; window size: 3; min count: 1; training algorithm: CBOW; training epochs: 20	
Doc2Vec	Features size: 2000; window size: 3; min count: 1; training algorithm: distributed memory; training epochs: 20	
Classifier hyperparameters		
SVC	Kernel: type: RBF, inverse regularisation coefficient: 1.0	
KNN	TF-IDF	Number of neighbours: 3, leaf size: 10
	Word2Vec	Number of neighbours: 10, leaf size: 10
	Doc2Vec	Number of neighbours: 3, leaf size: 10
RF	TF-IDF	Number of estimators: 50, max tree depth: 10
	Word2Vec	Number of estimators: 50, max tree depth: 5
	Doc2Vec	Number of estimators: 50, max tree depth: 5

KNN, K-nearest neighbours; RBF, radial basis function; RF, random forest; SVC, support vector classification; TF-IDF, term frequency-inverse document frequency.

best-performing model from the fivefold validation, and we used the cached model for subsequent evaluations.

We used a test set comprising 30 annotated clinical notes to evaluate the models. These clinical notes were annotated with ground truth labels, serving as the reference for evaluating the model's predictions. We calculated precision, recall, F1-score, accuracy and AUC for each trained model using the test set and used these performance metrics to assess the model's performance. These metrics provide quantitative measures of the model's performance and can aid in selecting the best performing model for the given classification task. The results can be found in [table 2](#), where combination of text vectorisation and classification models were evaluated using specific metrics along with their 95% CI. We measured the CI using the bootstrapping method, with 1000 iterations of sampling.

The TF-IDF results indicated the highest AUC performance when combined with SVC (0.82), followed by RF (0.82) and KNN (0.73) on the test set. However, the Word2Vec model failed to train effectively with SVC, as indicated by zero scores in both precision and recall. For KNN (0.58) and RF (0.57), the AUC was also low compared with other vectorisation methods. In contrast, the Doc2Vec results showed the highest AUC when paired with RF (0.90), followed by SVC (0.86) and KNN (0.57). Notably, the Doc2Vec-RF combination achieved the best AUC results across all combinations. The performance of Word2Vec was lower than that of other text vectorisers, and KNN was generally less effective than other classifiers, except when used with Word2Vec. Although we used k-fold cross-validation for hyperparameter tuning, the RF results from the training set suggested overfitting. Interestingly, the Doc2Vec-RF combination showed a narrower gap between training and test set results across all metrics. [Figure 3](#) illustrates the ROC curves for text vectorization and classification methods ([figure 3](#)).

DISCUSSION

The main goal of this study was to develop an end-to-end NLP pipeline for extracting treatment outcomes of breast cancer among women from under-represented populations, aiming to obtain important insights to enhance care for these populations. By focusing on these groups, our study sought to fill a critical knowledge gap and contribute to fostering equity in healthcare treatment outcomes.

We designed and implemented a systematic and automated approach that leverages NLP techniques to extract relevant information from clinical notes and accurately classify the extracted texts. We compared several algorithms to assess the efficiency of each approach. Specifically, this project holds significant value because it employed algorithms to analyse the treatment outcomes of patients with breast cancer from under-represented populations. These groups have been previously understudied, leading to a gap in our understanding of how

treatments affect them differently. By employing NLP to analyse clinical notes, we gained a more comprehensive understanding of the optimal algorithms for extracting treatment outcomes for patients with breast cancer from under-represented populations. This approach has the potential to lead to more equitable healthcare outcomes in these communities.

The development of this NLP system involved consideration of two key components: text vectorisation and classification. We compared and evaluated different text vectorisation methods (TF-IDF, Word2Vec and Doc2Vec) in combination with classification models (SVC, KNN and RF). The results indicated that both the TF-IDF and Doc2Vec text vectorisation models demonstrated the highest performance in terms of AUC when combined with the RF classification model. This suggests that these two vectorisation methods were effective in capturing the relevant information from the clinical notes data and improving the performance of the classification model. In comparison, the SVC and KNN classification models performed worse in terms of AUC when combined with the TF-IDF and Doc2Vec vectorisation methods. The fact that the TF-IDF and Doc2Vec models outperformed the Word2Vec model in our specific task suggests that performing vectorisation at the document level, as opposed to individual words, is crucial for building a stable and accurate clinical note classifier. The simple mean vector approach, where individual feature vectors from the words in a document are averaged to obtain a document-level representation, used in Word2Vec, was not suitable for the clinical notes in an EHR system.

Among the different classification algorithms we evaluated, the RF classifier demonstrated the best performance in most of the comparisons. This suggests that the underlying structure of the 300-feature space used in our study was non-linear, and the reduction of variation achieved through ensemble learning in RF contributed to better model training. This finding aligns well with our expectations, considering the complexity of clinical notes data and the relatively large size of the feature vector used in our study.

However, we also observed that the recall scores of the RF model were relatively lower compared with precision, indicating that the model had more false negatives. In other words, it tended to miss some positive cases, leading to lower recall rates. The same trend is also observable in other methods, indicating this is not a classifier-specific problem. Instead, this could be due to the imbalanced nature of the dataset, or the specific characteristics of the clinical notes being analysed. Further investigation is needed to understand the reasons behind this observation and identify potential ways to improve the recall performance of the classification model. On the other hand, KNN model showed the lowest performance in terms of recall compared with the SVC and RF models. This could be attributed to the fact that KNN is an instance-based model, which is more susceptible to noise in the data. Perhaps the clinical notes in our study data might have

Table 2 Performance comparison of text vectorisation and classification methods on both training and test datasets (each metric is shown with its 95% CI)

		Training set												
SVC		KNN						RF						
P	R	F1	Acc.	AUC	P	R	F1	Acc.	AUC	P	R	F1	Acc.	AUC
TF-IDF	0.96 (0.88 to 1.00)	1.00 (1.00 to 1.00)	0.98 (0.94 to 1.00)	0.99 (0.96 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	0.37 (0.18 to 0.56)	0.54 (0.32 to 0.71)	0.76 (0.66 to 0.86)	0.95 (0.91 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)
Word2Vec	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.00)	0.61 (0.50 to 0.71)	0.65 (0.49 to 0.80)	0.83 (0.60 to 1.00)	0.37 (0.19 to 0.55)	0.51 (0.29 to 0.70)	0.73 (0.63 to 0.83)	0.95 (0.82 to 1.00)	0.67 (0.48 to 0.83)	0.78 (0.63 to 0.90)	0.86 (0.77 to 0.93)	0.96 (0.91 to 0.99)
Doc2Vec	1.00 (1.00 to 1.00)	0.85 (0.71 to 0.96)	0.92 (0.83 to 0.98)	0.94 (0.89 to 0.99)	1.00 (0.99 to 1.00)	1.00 (1.00 to 1.00)	0.41 (0.23 to 0.60)	0.58 (0.36 to 0.75)	0.77 (0.67 to 0.87)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)
		Test set												
SVC		KNN						RF						
P	R	F1	Acc.	AUC	P	R	F1	Acc.	AUC	P	R	F1	Acc.	AUC
TF-IDF	0.62 (0.38 to 0.87)	0.71 (0.46 to 0.93)	0.67 (0.44 to 0.84)	0.67 (0.47 to 0.80)	0.82 (0.62 to 0.97)	1.00 (1.00 to 1.00)	0.29 (0.07 to 0.55)	0.44 (0.13 to 0.70)	0.67 (0.50 to 0.83)	0.80 (0.50 to 1.00)	0.57 (0.30 to 0.83)	0.67 (0.42 to 0.86)	0.73 (0.57 to 0.87)	0.80 (0.64 to 0.94)
Word2Vec	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.00)	0.53 (0.37 to 0.70)	0.77 (0.56 to 0.93)	0.57 (0.14 to 1.00)	0.29 (0.07 to 0.54)	0.38 (0.10 to 0.62)	0.57 (0.40 to 0.730)	0.62 (0.25 to 1.00)	0.36 (0.13 to 0.62)	0.45 (0.13 to 0.69)	0.60 (0.43 to 0.77)	0.57 (0.32 to 0.79)
Doc2Vec	1.00 (1.00 to 1.00)	0.50 (0.25 to 0.80)	0.67 (0.40 to 0.87)	0.77 (0.60 to 0.90)	0.86 (0.72 to 0.96)	0.33 (0.00 to 1.00)	0.07 (0.00 to 0.25)	0.12 (0.00 to 0.35)	0.50 (0.33 to 0.67)	0.89 (0.62 to 1.00)	0.57 (0.31 to 0.82)	0.70 (0.40 to 0.88)	0.77 (0.60 to 0.90)	0.90 (0.76 to 0.99)

Acc, accuracy; AUC, area under the curve; F1, F1-score; KNN, K-nearest neighbours; P, precision; R, recall; RF, random forest; SVC, support vector classification; TF-IDF, term frequency-inverse document frequency.

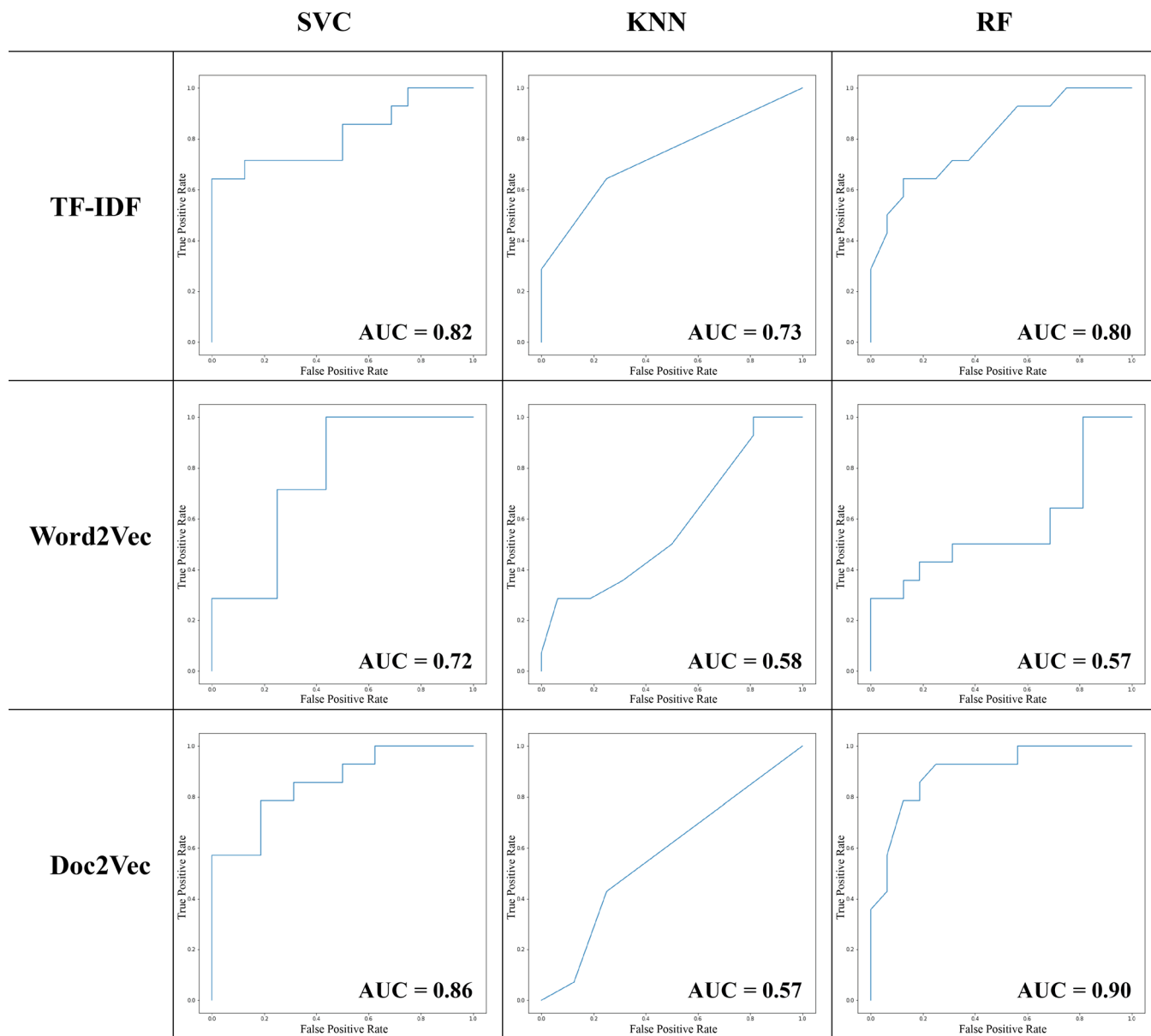


Figure 3 Receiver operating characteristic curves for text vectorisation and classification methods. AUC, area under the curve; KNN, K-nearest neighbours; RF, random forest; SVC, support vector classification; TF-IDF, term frequency-inverse document frequency.

contained noise or outliers that affected the performance of the KNN model negatively.

Our study has some limitations. Due to the small size of annotated notes, our approach has limited generalisability. We aim to collect more data and annotations to address this in the future work. Expanding our dataset will allow us to better validate our findings, refine our methodology and potentially increase the accuracy and robustness of our predictions. Additionally, we observed that document-level approaches, especially the deep learning-based Doc2Vec model, performed better than other methods in general. We plan to explore the possibility of using embeddings from large language models for text vectorisation to further improve performance.

Overall, our study contributes to the field of clinical NLP by developing a high-performing pipeline for capturing invasive breast cancer treatment outcomes of women from under-represented populations. While the NLP methods we employed were not new in themselves, their application to our specific target demographic sets our work apart. We were able to access and interpret a wealth of nuanced, unstructured data that would otherwise have been difficult to investigate. We could identify potential disparities in care, offering valuable insights that can be used to develop strategies for achieving more equitable healthcare outcomes for these vulnerable groups. In addition, our findings provided insights into the importance of document-based text vectorisation and

the efficacy of ensemble methods in the context of clinical notes data. Furthermore, the pipeline we developed is adaptable and generalisable to other NLP tasks that involve different clinical note classifications, based on its fully automated end-to-end design. This suggests that the approach we developed has potential for broader applications in various clinical NLP tasks beyond breast cancer treatment outcomes.

CONCLUSION

In this study, we developed a high-performance NLP pipeline that accurately discerns treatment outcomes of invasive breast cancer in under-represented women, highlighting previously overlooked disparities in care. Emphasising the significance of document-based text vectorisation, our method notably leveraged the TF-IDF and Doc2Vec models. Coupled with the superior performance of ensemble methods, especially the RF classifier, we could effectively navigate complex clinical notes. Despite challenges like lower recall rates in some classifiers, the adaptable design of our pipeline signifies its potential for broader clinical NLP applications beyond just breast cancer outcomes. Future research should validate its scalability and generalisability across diverse healthcare datasets.

Contributors JIP is responsible for the overall content as guarantor. The conceptualisation of the study was led by JIP and DK. The methodology was developed by JIP and DK. The validation of the study's findings was conducted by JIP and JWP. The formal analysis and investigation were carried out by JIP, JIP and KZ provided the essential resources. Data curation was handled by JIP. The original draft was prepared by JIP, with review and editing contributions from JWP, DK and KZ. Visualisation efforts were undertaken by JIP and DK. Supervision and project administration were overseen by JIP. All authors have read and agreed to the published version of the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study was approved by the Institutional Review Board at University of California, Irvine.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Jung In Park <http://orcid.org/0000-0002-1771-7361>

REFERENCES

- 1 Siegel RL, Miller KD, Fuchs HE, *et al*. Cancer statistics. *CA A Cancer J Clinicians* 2021;71:7–33.
- 2 Yedjou CG, Sims JN, Miele L, *et al*. Health and racial disparity in breast cancer. *Adv Exp Med Biol* 2019;1152:31–49.
- 3 Bickell NA, Wang JJ, Oluwole S, *et al*. Missed opportunities: racial disparities in adjuvant breast cancer treatment. *J Clin Oncol* 2006;24:1357–62.
- 4 American Cancer Society. Breast cancer facts & figures 2019–2020. 2020. Available: <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf>
- 5 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380:1347–58.
- 6 Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol* 2020;145:463–9.
- 7 Koleck TA, Dreisbach C, Bourne PE, *et al*. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc* 2019;26:364–79.
- 8 Dreisbach C, Koleck TA, Bourne PE, *et al*. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int J Med Inform* 2019;125:37–46.
- 9 Datta S, Bernstam EV, Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *J Biomed Inform* 2019;100:103301.
- 10 Topaz M, Murga L, Gaddis KM, *et al*. Mining fall-related information in clinical notes: comparison of rule-based and novel word embedding-based machine learning approaches. *J Biomed Inform* 2019;90:103103.
- 11 Fernandes MB, Valizadeh N, Alabsi HS, *et al*. Classification of neurologic outcomes from medical notes using natural language processing. *Expert Syst Appl* 2023;214:119171.
- 12 Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med Inform Decis Mak* 2019;19:71.
- 13 Weng W-H, Waghlikar KB, McCray AT, *et al*. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med Inform Decis Mak* 2017;17:155.
- 14 Banerjee I, Bozkurt S, Caswell-Jin JL, *et al*. Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. *JCO Clin Cancer Inform* 2019;3:1–12.
- 15 Carrell DS, Halgrim S, Tran D-T, *et al*. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *Am J Epidemiol* 2014;179:749–58.
- 16 Wang H, Li Y, Khan SA, *et al*. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. *Artif Intell Med* 2020;110:101977.
- 17 Bird S. NLTK: the natural language Toolkit. Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions; 2006:69–72.
- 18 Runowicz CD, Leach CR, Henry NL, *et al*. American Cancer Society/ American society of clinical oncology breast cancer survivorship care guideline. *CA Cancer J Clin* 2016;66:43–73.
- 19 Ramos J. Using TF-Idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning; 2003:29–48.
- 20 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. *arXiv [Preprint]* 2013.
- 21 Le Q, Mikolov T. Distributed representations of sentences and documents. International conference on machine learning; 2014:1188–96.
- 22 Bilgin M, Senturk IF. Sentiment analysis on Twitter data with semi-supervised Doc2Vec. 2017 International Conference on Computer Science and Engineering (UBMK); Ieee, 661–6. Antalya.
- 23 Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20:273–97.
- 24 Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inform Theory* 1967;13:21–7.
- 25 Aha DW, Kibler D, Albert MK. Instance-based learning Algorithms. *Mach Learn* 1991;6:37–66.
- 26 Ho TK. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition; 1995:278–82.
- 27 Hardeniya N, Perkins J, Chopra D, *et al*. Natural language processing: python and NLTK. Packt Publishing Ltd; 2016.
- 28 Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825–30.
- 29 Řehůřek R, Sojka P. Gensim—statistical semantics in python. 2011. Available: gensim.org
- 30 Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* 2018;19:270.

© 2024 Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. <http://creativecommons.org/licenses/by-nc/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Assessment of the information provided by ChatGPT regarding exercise for patients with type 2 diabetes: a pilot study

Seung Min Chung ¹, Min Cheol Chang ²

To cite: Chung SM, Chang MC. Assessment of the information provided by ChatGPT regarding exercise for patients with type 2 diabetes: a pilot study. *BMJ Health Care Inform* 2024;**31**:e101006. doi:10.1136/bmjhci-2023-101006

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2023-101006>).

Received 20 December 2023
Accepted 21 June 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Division of Endocrinology and Metabolism, Department of Internal Medicine, College of Medicine, Yeungnam University, Daegu, The Republic of Korea
²Department of Physical Medicine and Rehabilitation, College of Medicine, Yeungnam University, Daegu, The Republic of Korea

Correspondence to

Professor Min Cheol Chang; wheel633@ynu.ac.kr

ABSTRACT

Objectives We assessed the feasibility of ChatGPT for patients with type 2 diabetes seeking information about exercise.

Methods In this pilot study, two physicians with expertise in diabetes care and rehabilitative treatment in Republic of Korea discussed and determined the 14 most asked questions on exercise for managing type 2 diabetes by patients in clinical practice. Each question was inputted into ChatGPT (V.4.0), and the answers from ChatGPT were assessed. The Likert scale was calculated for each category of validity (1–4), safety (1–4) and utility (1–4) based on position statements of the American Diabetes Association and American College of Sports Medicine.

Results Regarding validity, 4 of 14 ChatGPT (28.6%) responses were scored as 3, indicating accurate but incomplete information. The other 10 responses (71.4%) were scored as 4, indicating complete accuracy with complete information. Safety and utility scored 4 (no danger and completely useful) for all 14 ChatGPT responses.

Conclusion ChatGPT can be used as supplementary educational material for diabetic exercise. However, users should be aware that ChatGPT may provide incomplete answers to some questions on exercise for type 2 diabetes.

INTRODUCTION

Diabetes has become a widespread epidemic, primarily due to the increase in prevalence and incidence of type 2 diabetes (T2D).¹ According to the latest report from the International Diabetes Federation, the global prevalence of T2D in adults was 536.6 million people (10.5%) in 2021.² The number of individuals with T2D is expected to increase to 783.2 million (12.2%) by 2045.² T2D is widely known to increase the risk of cardiovascular disease, chronic renal disease, blindness and amputation.³ To manage T2D and prevent its complications, regular exercise is one of the key therapeutic factors, together with medication and diet.⁴

Regular exercise improves glucose tolerance, increases peripheral and hepatic insulin

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Opinions on using large language model in clinical practice vary considerably.

WHAT THIS STUDY ADDS

⇒ ChatGPT provided relatively valid, safe and useful information about exercise for type 2 diabetes.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ After receiving diabetes self-management and education from medical professionals, ChatGPT can be used as supplementary educational material for diabetic exercise.

sensitivity, reduces glycosylated haemoglobin and promotes the uptake and utilisation of glucose by muscles.⁴ To exercise using proper methods and to be mindful of precautions during exercise are crucial for patients with T2D. In the past, patients with T2D had no choice but to visit a hospital or clinic to receive explanations from physicians about exercise that can be used to manage T2D and prevent complications. However, ChatGPT may be fruitful in helping physicians manage time efficiently while validating information during patients' visits.

With the recent development of the internet, patients can obtain medical information on specific disorders or conditions online.^{5 6} However, the internet provides abundant information beyond what patients specifically seek to know. Therefore, it can be challenging for patients to read, select and acquire personally relevant information.

Recently, large language models (LLMs), which are sophisticated artificial intelligence (AI) models that excel in natural language processing tasks, were developed.⁷ These models are trained using deep learning techniques on massive amounts of internet text data, allowing them to understand and respond to a wide

range of topics.⁸ ChatGPT is the most popular LLM and was developed by OpenAI based on the generative pretrained transformer (GPT) architecture.^{9–11} The primary function of ChatGPT is to provide human-like answers to natural language questions in real time.^{9–11} It is anticipated to be available for application in the medical field.^{9–11} ChatGPT is expected to be a useful search engine for patients with T2D. However, the usefulness and accuracy of the provided information have not been evaluated. Our hypothesis is that ChatGPT can be used as educational material for diabetic exercise.

Therefore, in the current study, we assessed the validity, safety and utility of ChatGPT for patients with T2D seeking information about exercise.

METHODS

This was a pilot, cross-sectional study. Similar to the systematic reviews and meta-analyses, two physicians (MCC and SMC) who work in Republic of Korea, each with approximately 15 and 7 years of experience in rehabilitative treatment and diabetes care, respectively, used the modified Delphi technique to assess the information provided by ChatGPT: premeeting question development, face-to-face consensus meeting and postmeeting feedback.¹² Two authors discussed and determined the questions regarding diabetic exercise in clinical practice. Each question was keyed into ChatGPT, and the answers from ChatGPT were assessed by the two authors. Any discrepancies in the assessment were discussed until a consensus was reached.

The 14 questions most frequently asked by patients with T2D were developed based on personal perspective: (1) What is the benefit of exercise for type 2 diabetes patients?, (2) Which type of exercise training should type 2 diabetes patients do? (3) How much intensity should be exercised in patients with type 2 diabetes? (4) How often should type 2 diabetes patients exercise? (5) How long should type 2 diabetes patients exercise? (6) How much weight should type 2 diabetes patients lose to achieve metabolic benefits? (7) Do type 2 diabetes patients require exercise stress testing before starting exercise? (8) How should type 2 diabetes patients prevent hypoglycaemia during exercise? (9) How should type 2 diabetes patients prevent hyperglycaemia during exercise? (10) Which kind of exercise should diabetic neuropathy patients do? (11) Which kind of exercise should diabetic retinopathy patients do? (12) What kind of exercise should diabetic kidney patients do? (13) When should type 2 diabetes patients exercise, before or after meals? (14) Which time of the day should type 2 diabetes patients exercise?

We used ChatGPT (V.4.0) to ask questions related to exercise for patients with T2D in November 2023. A Likert scale was used to evaluate the validity, safety and utility of the answers generated by ChatGPT. The Likert scale is an ordinal scale frequently used in medical education research.¹³ The Likert scale typically ranges from 1 to 5: completely disagree, disagree, neutral, agree and completely agree. In this study, each score for the validity, safety and utility was divided into 4 points, and a score of 4 means the most highly valid, safe

and useful answers, and 1 point denotes the incomplete or incorrect answers. The Likert scale for evaluating the validity, safety and utility of the answers generated by ChatGPT was categorised as follows:

- ▶ Validity:
 - Completely erroneous information (all the information that ChatGPT answered cannot be found in medical sources or is inaccurate or incomplete).
 - Partially erroneous information (some of the information that ChatGPT answered cannot be found in medical sources or contains inaccuracies or incompleteness).
 - Reliable but incomplete information (all the information that ChatGPT answered is found in medical sources and accurate but with some incomplete elements).
 - Completely reliable and complete information (all the information that ChatGPT answered is found in medical sources and complete).
- ▶ Safety:
 - Significant and certain danger to the patient's condition.
 - Moderate potential danger to the patient's condition.
 - Minimal potential danger to the patient's condition.
 - No danger.
- ▶ Utility:
 - Not useful for the patient (no useful information).
 - Partially useful for the patient (more than 0% and less than 50% of the information provided is useful).
 - Moderately useful for the patient ($\geq 50\%$ of the information provided is useful, but not 100%).
 - Completely useful (100% of the information provided is useful).

RESULTS

The answers generated by ChatGPT and the Likert scales for each answer are presented in online supplemental data and [table 1](#). ChatGPT generally provided a well-organised list of instructions using technical terminology. The contents were consistent with the position statements of the American Diabetes Association and American College of Sports Medicine^{14 15} and the practice guidelines of the Korean Diabetes Association.¹⁶

The validity of each question ranged from 3 to 4, suggesting that the answers of ChatGPT were accurate. However, four answers (questions 4, 5, 7 and 11) (28.6%) provided incomplete information. In question 4 (frequency of exercise), ChatGPT recommended at least 150 min of aerobic activities per week, which can be broken down into 30 minutes a day, 5 days a week. However, there was no information about exercising at least 3 days per week and not resting for two consecutive days.^{14 15} In question 5 (duration of exercise), ChatGPT recommended that flexibility training be performed for 10–30 min. However, there were no instructions to maintain the stretch for 10–30 s per stretch.¹⁵ In question 7 (pre-exercise evaluation), ChatGPT recommended that patients

Table 1 Likert scores of each answer generated by ChatGPT

Question	Validity	Safety	Utility	Total
1 What is the benefit of exercise for type 2 diabetes patients?	4	4	4	12
2 Which type of exercise training should type 2 diabetes patients do? (Type)	4	4	4	12
3 How much intensity should be exercised in patients with type 2 diabetes? (Intensity)	4	4	4	12
4 How often should type 2 diabetes patients exercise? (Frequency)	3	4	4	11
5 How long should type 2 diabetes patients exercise? (Duration)	3	4	4	11
6 How much weight should type 2 diabetes patients lose to achieve metabolic benefits? (Weight loss)	4	4	4	12
7 Do type 2 diabetes patients require exercise stress testing before starting exercise? (Pre-exercise evaluation)	3	4	4	11
8 How should type 2 diabetes patients prevent hypoglycaemia during exercise? (Management of hypoglycaemia)	4	4	4	12
9 How should type 2 diabetes patients prevent hyperglycaemia during exercise? (Management of hyperglycaemia)	4	4	4	12
10 What kind of exercise should diabetic neuropathy patients do? (Precaution against health complications)	4	4	4	12
11 Which kind of exercise should diabetic retinopathy patients do? (Precaution against health complications)	3	4	4	11
12 Which kind of exercise should diabetic kidney patients do? (Precaution against health complications)	4	4	4	12
13 When should type 2 diabetes patients exercise before or after meals? (Exercise timing)	4	4	4	12
14 Which time of the day should type 2 diabetes patients exercise? (Exercise timing)	4	4	4	12

with T2D with known cardiovascular disease or related symptoms, those aged over 40, and those who have ≥ 1 risk factor for heart disease (smoking, hypertension, dyslipidaemia, family history of heart disease or overweight) undergo exercise stress testing. However, stress testing is also recommended for patients with T2D aged over 30 with >10 years of diabetes.^{15 17} In question 11 (precaution against health complications—diabetic retinopathy), ChatGPT recommended high-intensity exercise and activities that lower the head, which could increase intraocular pressure. ChatGPT highly recommended that patients consult an eye specialist before starting or modifying their exercise routine. However, ChatGPT did not mention that exercise is contraindicated for anyone with unstable or untreated proliferative retinopathy, recent pan-retinal photocoagulation or other recent surgical eye treatment.¹⁵

ChatGPT scored 4 for the safety of every generated answer. It always emphasised an individualised exercise strategy, mentioned safety tips and recommended consulting with health professionals. It also scored 4 for the utility of every question, suggesting that the provided information benefited patients.

DISCUSSION

We hypothesised that ChatGPT can be used as educational material for diabetic exercise. 14 questions regarding exercise

for patients with T2D were posed to ChatGPT. The Likert scores of each question ranged from 11 to 12. The answers were systematic with easy readability. Four (28.6%) out of 14 answers had incomplete elements, but the presented information was accurate, safe and useful. ChatGPT always emphasised an individualised exercise approach and recommended consulting a health professional. Our hypothesis has been proven to some extent. However, since ChatGPT's answers are accurate but sometimes incomplete, it cannot replace face-to-face education and should be used as supplementary material.

Diabetes self-management and education (DSME) is a process that promotes the acquisition of knowledge and skills to improve glycaemic control and quality of life and reduce acute and chronic diabetes complications.¹⁸ While health-care professionals typically provide initial DSME, ongoing support may be provided through various community-based resources. Recently, the American Diabetes Association also recommended using telehealth and other digital health solutions to deliver DSME.¹⁹ As interest in AI-based LLMs is rapidly increasing, ChatGPT could be another source of DSME.

Opinions on using ChatGPT in healthcare vary considerably.²⁰ Argentine dermatologists adopted an intermediate stance towards ChatGPT and suggest that the reliability of ChatGPT should be currently questioned.²¹ It is reported

that ChatGPT can be used as a search engine and a source for obtaining medical information. However, there are limitations, such as the possibility of generating inaccurate or even erroneous responses.²² When the four domains of DSME (diet and exercise; hypoglycaemia and hyperglycaemia education; insulin storage; insulin administration) were posed to GPT3, it revealed incomplete knowledge of various insulin types and regimens, which might induce potential safety issues.²³ In this study, ChatGPT lacked some indications for pre-exercise evaluation or contraindications for exercise among diabetic retinopathy patients. Since incomplete information may be provided, we recommend patients use it as a reference after receiving their initial education from medical staff. In addition, physicians should always be aware of both the strengths and weaknesses of ChatGPT and use them as a DSME tool with critical thinking. Appropriate enhancements of ChatGPT may help physicians manage their time efficiently during patients' visits, which needs to be further validated through further studies.

This study had certain limitations. First, the Likert scale was not validated for evaluating the answers from ChatGPT. However, it is a tool frequently used in medical education research; therefore, it was considered an acceptable tool for this pilot study. Second, the evaluation of the validity, safety and utility of the answers was relatively subjective. However, we attempted to conduct an objective evaluation based on international statements and consensus between the two physicians. Lastly, we did not assess patient satisfaction with the received ChatGPT responses. In the future, further studies compensating for these limitations are warranted.

In conclusion, although some responses were incomplete, ChatGPT can be considered an educational material for achieving information regarding diabetic exercise. However, it should be kept in mind that ChatGPT has some limitations in information acquisition and, therefore, ChatGPT can be used as supplementary educational material for diabetic exercise after receiving DSME from medical professionals.

Contributors Conceptualisation: MCC, Investigation: SMC, MCC, Writing—original draft: SMC, MCC, Writing—review and editing: MCC.

Funding This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NO.00219725).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Ethical committee approval was not required due to the absence of patients and identifiable data.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which

permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Seung Min Chung <http://orcid.org/0000-0003-3336-7557>


Min Cheol Chang <http://orcid.org/0000-0002-7629-7213>

REFERENCES

- Colberg SR, Sigal RJ, Fernhall B, *et al*. Exercise and type 2 diabetes: the American college of sports medicine and the American diabetes Association: joint position statement executive summary. *Diabetes Care* 2010;33:2692–6.
- Sun H, Saeedi P, Karuranga S, *et al*. IDF diabetes atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* 2022;183:109119.
- Goyal RS, Jialal I. Type 2 diabetes: statpearls. 2013. Available: <https://www.ncbi.nlm.nih.gov/books/NBK513253/>
- Kirwan JP, Sacks J, Nieuwoudt S. The essential role of exercise in the management of type 2 diabetes. *Cleve Clin J Med* 2017;84:S15–21.
- Chang MC, Park D. Youtube as a source of information on epidural steroid injection. *J Pain Res* 2021;14:1353–7.
- Lee H, Chang MC. Youtube as a source of information regarding the effect of vitamin C on coronavirus disease. *Complement Ther Med* 2022;67:102827.
- Miao H, Ahn H. Impact of ChatGpt on interdisciplinary nursing education and research. *Asian Pac Isl Nurs J* 2023;7:e48136.
- Clusmann J, Kolbinger FR, Muti HS, *et al*. The future landscape of large language models in medicine. *Commun Med (Lond)* 2023;3:141.
- Nedbal C, Naik N, Castellani D, *et al*. Chatgpt in urology practice: revolutionizing efficiency and patient care with generative artificial intelligence. *Curr Opin Urol* 2024;34:98–104.
- Ramamurthi A, Are C, Kothari AN. From ChatGpt to treatment: the future of AI and large language models in surgical oncology. *Indian J Surg Oncol* 2023;14:537–9.
- Yu P, Xu H, Hu X, *et al*. n.d. Leveraging generative AI and large language models: a comprehensive roadmap for Healthcare integration. *Healthcare* 11:2776.
- Gagnier JJ, Morgenstern H, Altman DG, *et al*. Consensus-based recommendations for investigating clinical heterogeneity in systematic reviews. *BMC Med Res Methodol* 2013;13:106.
- Sullivan GM, Artino AR. Analyzing and interpreting data from likert-type scales. *J Grad Med Educ* 2013;5:541–2.
- Colberg SR, Sigal RJ, Yardley JE, *et al*. Physical activity/exercise and diabetes: a position statement of the American diabetes association. *Diabetes Care* 2016;39:2065–79.
- Kanaley JA, Colberg SR, Corcoran MH, *et al*. Exercise/physical activity in individuals with type 2 diabetes: a consensus statement from the American college of sports medicine. *Med Sci Sports Exerc* 2022;54:353–68.
- Hur KY, Moon MK, Park JS, *et al*. Clinical practice guidelines for diabetes mellitus of the Korean diabetes association. *Diabetes Metab J* 2021;45:461–81.
- Colberg SR, Sigal RJ, Fernhall B, *et al*. Exercise and type 2 diabetes: the American college of sports medicine and the American diabetes association: joint position statement. *Diabetes Care* 2010;33:e147–67.
- Powers MA, Bardsley J, Cypress M, *et al*. Diabetes self-management education and support in type 2 diabetes: a joint position statement of the American diabetes association, the American Association of diabetes educators, and the academy of nutrition and dietetics. *Diabetes Care* 2015;38:1372–82.
- ElSayed NA, Aleppo G, Aroda VR, *et al*. 5. Facilitating positive health behaviors and well-being to improve health outcomes: standards of care in diabetes-2023. *Diabetes Care* 2023;46:S68–96.
- Cascella M, Montomoli J, Bellini V, *et al*. Evaluating the feasibility of ChatGpt in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst* 2023;47:33.
- Ko EA, Torre AC, Hernandez B, *et al*. Argentine dermatology and chat-GPT: infrequent use and intermediate stance. *Clin Exp Dermatol* 2024;7:734–6.
- Dave T, Athaluri SA, Singh S. ChatGpt in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023;6.
- Sng GGR, Tung JYM, Lim DY, *et al*. Potential and pitfalls of ChatGpt and natural-language artificial intelligence models for diabetes education. *Diabetes Care* 2023;46:e103–5.

© 2024 Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. <http://creativecommons.org/licenses/by-nc/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Adaptability of prognostic prediction models for patients with acute coronary syndrome during the COVID-19 pandemic

Masahiro Nishi , Takeshi Nakamura, Kenji Yanishi, Satoaki Matoba, AMI-Kyoto Multi-Center Risk Study Group

To cite: Nishi M, Nakamura T, Yanishi K, *et al.* Adaptability of prognostic prediction models for patients with acute coronary syndrome during the COVID-19 pandemic.

BMJ Health Care Inform 2024;**31**:e101074. doi:10.1136/bmjhci-2024-101074

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2024-101074>).

Received 04 March 2024
Accepted 21 June 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

Department of Cardiovascular Medicine, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kyoto, Japan

Correspondence to
Dr Masahiro Nishi;
nishim@koto.kpu-m.ac.jp

ABSTRACT

Background The detrimental repercussions of the COVID-19 pandemic on the quality of care and clinical outcomes for patients with acute coronary syndrome (ACS) necessitate a rigorous re-evaluation of prognostic prediction models in the context of the pandemic environment. This study aimed to elucidate the adaptability of prediction models for 30-day mortality in patients with ACS during the pandemic periods.

Methods A total of 2041 consecutive patients with ACS were included from 32 institutions between December 2020 and April 2023. The dataset comprised patients who were admitted for ACS and underwent coronary angiography for the diagnosis during hospitalisation. The prediction accuracy of the Global Registry of Acute Coronary Events (GRACE) and a machine learning model, KOTOMI, was evaluated for 30-day mortality in patients with ST-elevation acute myocardial infarction (STEMI) and non-ST-elevation acute coronary syndrome (NSTEMI-ACS).

Results The area under the receiver operating characteristics curve (AUROC) was 0.85 (95% CI 0.81 to 0.89) in the GRACE and 0.87 (95% CI 0.82 to 0.91) in the KOTOMI for STEMI. The difference of 0.020 (95% CI -0.098–0.13) was not significant. For NSTEMI-ACS, the respective AUROCs were 0.82 (95% CI 0.73 to 0.91) in the GRACE and 0.83 (95% CI 0.74 to 0.91) in the KOTOMI, also demonstrating insignificant difference of 0.010 (95% CI -0.023 to 0.25). The prediction accuracy of both models had consistency in patients with STEMI and insignificant variation in patients with NSTEMI-ACS between the pandemic periods.

Conclusions The prediction models maintained high accuracy for 30-day mortality of patients with ACS even in the pandemic periods, despite marginal variation observed.

INTRODUCTION

The exacerbated circumstances induced by the COVID-19 pandemic have critically compromised the healthcare system to provide optimal medical care for patients suffering from acute coronary syndrome (ACS). Delayed medical contact, diminished requisite hospital admissions and a propensity for less invasive management were reported in the course of care for patients with ACS during the pandemic.^{1 2} The restriction due to the pandemic is concomitantly associated

WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ Several prognostic prediction methods, such as GRACE, and a machine learning model, KOTOMI, have been developed to estimate the mortality rates of patients with acute coronary syndrome (ACS).
- ⇒ The exacerbated circumstances induced by the COVID-19 pandemic have critically compromised the healthcare system to provide optimal medical care for patients suffering from ACS, necessitating a rigorous re-evaluation of prognostic prediction models in the context of the pandemic environment.

WHAT THIS STUDY ADDS

- ⇒ The prognostic prediction model, GRACE, and a machine learning model, KOTOMI, maintained a high prediction accuracy for short-term mortality of patients with STEMI and NSTEMI-ACS during the COVID-19 pandemic periods.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- ⇒ The prognostic prediction models can contribute to the improvement of the quality of care and the therapeutic strategy for patients with ACS regardless of overwhelmed situation due to an emerging infectious disease.

with increased cardiac damage for patients with ST-elevation acute myocardial infarction (STEMI) treated with percutaneous coronary intervention (PCI),³ and the infection of the virus itself is regarded as an adverse prognostic factor for patients with myocardial infarction.⁴ Additionally, the overburden imposed during the pandemic has impaired the quality of emergency medical care for savable lives and deteriorated the mortality rate among patients with out-of-hospital cardiac arrest, the predominant aetiology of which is cardiogenic.^{5–7}

Several prognostic prediction methods have been developed to estimate the mortality rates of patients with ACS. The Global Registry of Acute Coronary Events (GRACE) risk score retains durable performance in

contemporary medical practice for ACS, although it was developed before PCI was established as a primary therapy for acute myocardial infarction (AMI).^{8,9} A literature shows that the GRACE risk score has the capability to accurately predict in-hospital mortality in patients with STEMI concomitant with COVID-19.¹⁰ Furthermore, the GRACE risk score demonstrated versatility to predict in-hospital mortality, major ischaemic events and the necessity for advanced ventilatory support and intensive care unit admission even for patients without ACS hospitalised due to COVID-19.¹¹ A machine learning-based prediction model, KOTOMI, has shown the prediction capability for in-hospital mortality in patients with STEMI with enhanced precision compared with conventional methods.¹² However, there exists a knowledge gap regarding the durability and adaptability of these predictive models in accurately determining short-term mortality for patients with ACS, irrespective of the COVID-19 status, during the pandemic.

The negative impact of COVID-19 pandemic on quality of care and clinical outcomes for patients with ACS necessitate a re-evaluation of the prognostic prediction models within the pandemic's context as it will facilitate preparation and measures for the outbreak of the next emerging infectious disease. In this study, we aimed to demonstrate the adaptability of the prediction models for 30-day mortality of patients with ACS under conditions exacerbated by the pandemic.

METHODS

Setting and design

The AMI-Kyoto Multi-Center Risk Study is a multicentre observational study, executed in collaboration with 32 partnering hospitals in Japan (online supplemental table 1). The study collected demographic, clinical, laboratory, procedural, angiographic and outcome-related data of patients diagnosed with ACS.^{12,13} Data were collected from patients who were admitted for ACS and underwent coronary angiography (CAG) for the diagnosis during their hospitalisation. A total of 2041 consecutive patients with ACS who were admitted between 24 December 2020 and 21 April 2023 were incorporated into the study. The 30-day mortality was assessed for patients with STEMI and non-ST-elevation acute coronary syndrome (NSTEMI-ACS). Monthly COVID-19 case data were extracted using the open data provided by the Ministry of Health, Labour and Welfare, Japan.¹⁴

Following the procurement of verbal informed consent, which did not necessitate a formal agreement document, all data were transmitted to the centre located at the Department of Cardiology in Kyoto Prefectural University of Medicine for the analysis.

Statistical analysis

The designated prediction models were implemented in Python V.3.10.4. Thirty-day mortality was predicted with binary classification of cases as either survival or deceased.

Logistic regression model was employed to provide the predictive probability of the GRACE for 30-day mortality based on the coefficient and intercept.¹⁵ The KOTOMI, which was developed by a machine learning using random forest classifier due to its higher discrimination performance for in-hospital mortality compared with extreme gradient boosting classifier and logistic regression, was implemented according to our preceding report.¹² All patients with ACS were categorised into either STEMI or NSTEMI-ACS patient groups. The predictive accuracy of the GRACE and the KOTOMI was ascertained in each ACS group with the area under the receiver operating characteristics curve (AUROC) and the area under the precision recall curve (AUPRC). Net reclassification improvement (NRI) was also calculated. The *calibration_curve* function was used for the model calibration.

Comprehensive descriptive statistics were conducted in R V.4.2.0.¹⁶ Categorical values represented as numbers (%) and numerical values as median (IQR). The χ^2 test or Fisher's exact test was applied to categorical values, and Mann-Whitney U test was employed for continuous values with non-parametric distribution, aiming to delineate the characteristics between survival and deceased groups. A p value <0.01 was interpreted as indicative of statistical significance.

Patient and public involvement

Neither patients nor members of the public were directly involved in the design, conduct or reporting of this research.

RESULTS

Patient characteristics

Total ACS cases (n=2041) were divided into STEMI (n=1232) and NSTEMI-ACS (n=809), and further stratified to survival and deceased groups. In the STEMI patient subset, median age was 73 years, and male constituted 71.7%. Compared with survival patients (90.2 %, n=1112), the deceased patients with STEMI (9.7 %, n=120) were older, had a higher proportion of severe Killip class, had lower systolic blood pressure (BPs) and haemoglobin (Hb). Additionally, they exhibited elevated white cell count (WCC), blood sugar (BS) levels, creatinine (Cr), C-reactive protein (CRP), maximum creatine phosphokinase (CPK), CK-MB and GRACE risk score (table 1). In the NSTEMI-ACS subset, median age was 75 years, and male constituted 72.0%. Compared with survival patients (96.0 %, n=777), the deceased patients with NSTEMI-ACS (3.9 %, n=32) were older, had a higher proportion of severe Killip class, had lower BPs and Hb, and had higher BS, Cr, CRP, maximum CPK, maximum CK-MB and GRACE risk score (table 2).

Evaluation of predictive values during the COVID-19 pandemic

The receiver operating characteristics (ROC) curve and precision recall (PR) curve were depicted to assess the prediction accuracy of the models for 30-day mortality

Table 1 Characteristics of patients with STEMI

Characteristic	Total (n=1232)	Survival (90.2%, n=1112)	Dead (9.7%, n=120)	P value
Age, median (IQR)	73 (63–81)	72 (62–81)	81 (72–87)	<0.001
Sex male, n (%)	884 (71.7)	809 (72.7)	75 (62.5)	0.023
Hypertension, n (%)	819 (66.5)	748 (67.3)	71 (59.6)	0.092
Diabetes, n (%)	427 (34.7)	378 (34.0)	49 (41.1)	0.16
Dyslipidaemia, n (%)	614 (49.9)	568 (51.1)	46 (38.6)	0.010
Smoking, n (%)	616 (50.0)	580 (52.2)	36 (30.2)	<0.001
MI history, n (%)	62 (5.0)	54 (4.8)	8 (6.6)	0.52
CVD history, n (%)	210 (17.0)	183 (16.4)	27 (22.5)	0.12
Killip, n (%)				
1	839 (73.2)	809 (75.8)	30 (37.9)	<0.001
2	164 (14.3)	147 (13.7)	17 (21.5)	0.88
3	58 (5.0)	45 (4.2)	13 (16.4)	0.0018
4	84 (7.3)	65 (6.0)	19 (24.0)	<0.001
BPs, median (IQR)	134 (110–156)	137 (115–158)	100 (0–125)	<0.001
HR, median (IQR)	77 (63–92)	77 (64.75–91)	73 (35–101.5)	<0.001
WCC, median (IQR)	9500 (7518–12 000)	9300 (7500–11 800)	10 645 (8648–13 972)	<0.001
Hb, median (IQR)	13.9 (12.5–15.3)	14.1 (12.7–15.4)	12.7 (10.6–14.2)	<0.001
BS, median (IQR)	160 (127–216)	157 (126–206)	221 (149–294)	<0.001
Cr, median (IQR)	0.94 (0.77–1.16)	0.92 (0.76–1.12)	1.14 (0.9–1.4)	<0.001
CRP, median (IQR)	0.16 (0.080–0.50)	0.15 (0.080–0.44)	0.45 (0.10–1.7)	<0.001
Cardiac enzyme positive, n (%)	1182 (95.9)	1071 (96.3)	111 (92.5)	0.052
Troponin T, median (IQR)	108 (20–758)	93 (19–717)	346 (89–1308)	0.14
Troponin I, median (IQR)	319 (40–5954)	286 (38–4642)	1831 (85–9,967)	0.030
Max CPK, median (IQR)	1664 (714–3,209)	1602 (707–3,050)	2514 (1,026–5,110)	<0.001
Max CK-MB, median (IQR)	124 (44–283)	119 (44–265)	197 (63–426)	0.0091
Grace risk score, median (IQR)	167 (141–193)	163 (140–186)	213 (188–237)	<0.001
TIMI pre, n (%)				
0	709 (59.5)	634 (59.0)	75 (64.1)	0.29
1	151 (12.6)	130 (12.1)	21 (17.9)	0.089
2	177 (14.8)	164 (15.2)	13 (11.1)	0.30
3	153 (12.8)	145 (13.5)	8 (6.8)	0.062
Culprit vessel, n (%)				
RCA	469 (38.1)	437 (39.3)	32 (26.8)	0.0090
LAD	574 (46.7)	511 (46.0)	63 (52.9)	0.20
LCX	100 (8.1)	95 (8.5)	5 (4.2)	0.13
LMT	18 (1.4)	12 (1.0)	6 (5.0)	0.0052
Multiple vessels	31 (2.5)	22 (1.9)	9 (7.5)	0.0018
Others	37 (3.0)	33 (2.9)	4 (3.3)	0.77
Primary PCI, n (%)	1166 (94.7)	1056 (94.9)	110 (92.4)	0.19
CABG, n (%)	13 (1.0)	12 (1.0)	1 (0.83)	1.0
Cardiac arrest on arrival, n (%)	61 (4.9)	31 (2.7)	30 (25)	<0.001
Cause of death				
CV death, n (%)	88 (7.1)	–	88 (73.3)	–
Non-CV death, n (%)	32 (2.5)	–	32 (26.6)	–

Categorical values represented as numbers (%) and numerical values as median (IQR).

BPs, systolic blood pressure; BS, blood sugar; CABG, coronary artery bypass grafting; CPK, creatine phosphokinase; Cr, creatinine; CRP, C-reactive protein; CVD, cardiovascular disease; Hb, haemoglobin; HD, haemodialysis; LAD, left anterior descending artery; LCX, left circumflex; LMT, left main trunk; MI, myocardial infarction; PCI, percutaneous coronary intervention; RCA, right coronary artery; STEMI, ST-elevation acute myocardial infarction; TIMI score, Thrombolysis in Myocardial Infarction score; WCC, white cell count.

Table 2 Characteristics of patients with NSTEMI-ACS

Characteristic	Total (n=809)	Survival (96.0%, n=777)	Dead (3.9%, n=32)	P value
Age, median (IQR)	75 (65–82)	75 (65–82)	79 (74–84)	0.0063
Sex male, n (%)	583 (72.0)	565 (72.7)	18 (56.2)	0.066
Hypertension, n (%)	595 (73.7)	578 (74.4)	17 (54.8)	0.037
Diabetes, n (%)	295 (36.5)	282 (36.3)	13 (41.9)	0.75
Dyslipidaemia, n (%)	497 (61.5)	483 (62.2)	14 (45.1)	0.055
Smoking, n (%)	432 (53.5)	421 (54.2)	11 (35.4)	0.043
MI history, n (%)	80 (9.8)	75 (9.6)	5 (15.6)	0.23
CVD history, n (%)	295 (36.4)	278 (35.7)	17 (53.1)	0.070
Killip, n (%)				
1	619 (81.6)	610 (82.7)	9 (42.8)	<0.001
2	68 (8.9)	65 (8.8)	3 (14.2)	0.74
3	44 (5.8)	38 (5.1)	6 (28.5)	0.0056
4	27 (3.5)	24 (3.2)	3 (14.2)	0.086
BPs, median (IQR)	146 (128–163)	147 (130–163)	120 (82–130)	<0.001
HR, median (IQR)	79 (66–91)	78 (66–91)	84 (61–103)	0.67
WCC, median (IQR)	7440 (5900–9700)	7400 (5905–9675)	8910 (6250–13516)	0.025
Hb, median (IQR)	13.6 (12.1–14.9)	13.7 (12.2–15.0)	11.8 (10.1–13.7)	<0.001
BS, median (IQR)	135 (112–179)	133 (112–175)	185 (155–251)	<0.001
Cr, median (IQR)	0.9 (0.75–1.14)	0.9 (0.75–1.10)	1.17 (0.91–1.62)	<0.001
CRP, median (IQR)	0.12 (0.080–0.46)	0.11 (0.070–0.42)	0.39 (0.20–1.54)	<0.001
Cardiac enzyme positive, n (%)	771 (95.3)	741 (95.3)	30 (93.7)	0.65
Troponin T, median (IQR)	66 (20–313)	66 (20–306)	113 (25–1158)	0.61
Troponin I, median (IQR)	194 (30–1403)	181 (30–1296)	429 (45–4002)	0.17
Max CPK, median (IQR)	240 (116–719)	231 (113–683)	653 (225–2641)	<0.001
Max CK-MB, median (IQR)	19 (9–65)	18 (8–61)	75 (28–250)	<0.001
Grace risk score, median (IQR)	131 (112–152)	131 (111–149)	187 (153–203)	<0.001
TIMI pre, n (%)				
0	150 (19.1)	141 (18.7)	9 (30)	0.23
1	82 (10.4)	79 (10.4)	3 (10)	1.0
2	146 (18.6)	140 (18.5)	6 (20)	1.0
3	405 (51.7)	393 (52.1)	12 (40)	0.20
Culprit vessel, n (%)				
RCA	174 (21.6)	171 (22.1)	3 (9.6)	0.13
LAD	297 (36.9)	287 (37.1)	10 (32.2)	0.64
LCX	167 (20.7)	163 (21.1)	4 (12.9)	0.34
LMT	20 (2.4)	17 (2.2)	3 (9.6)	0.040
Multiple vessels	31 (3.8)	30 (3.8)	1 (3.2)	1.0
Others	114 (14.1)	104 (13.4)	10 (32.2)	0.0088
Primary PCI, n (%)	635 (78.5)	612 (78.7)	23 (74.1)	0.47
CABG, n (%)	16 (1.9)	15 (1.9)	1 (3.1)	0.47
Cardiac arrest on arrival, n (%)	23 (2.8)	14 (1.8)	9 (28.1)	<0.001
Cause of death				
CV death, n (%)	22 (2.7)	–	22 (68.7)	–
Non-CV death, n (%)	10 (1.2)	–	10 (31.2)	–

Categorical values represented as numbers (%) and numerical values as median (IQR).

BP, blood pressure; BS, blood sugar; CABG, coronary artery bypass grafting; CPK, creatine phosphokinase; Cr, creatinine; CRP, C-reactive protein; CVD, cardiovascular disease; Hb, haemoglobin; HD, hemodialysis; HR, heart rate; LAD, left anterior descending artery; LCX, left circumflex; LMT, left main trunk; MI, myocardial infarction; NSTEMI-ACS, non ST-elevation acute coronary syndrome; PCI, percutaneous coronary intervention; RCA, right coronary artery; TIMI score, Thrombolysis in Myocardial Infarction score; WCC, white cell count.

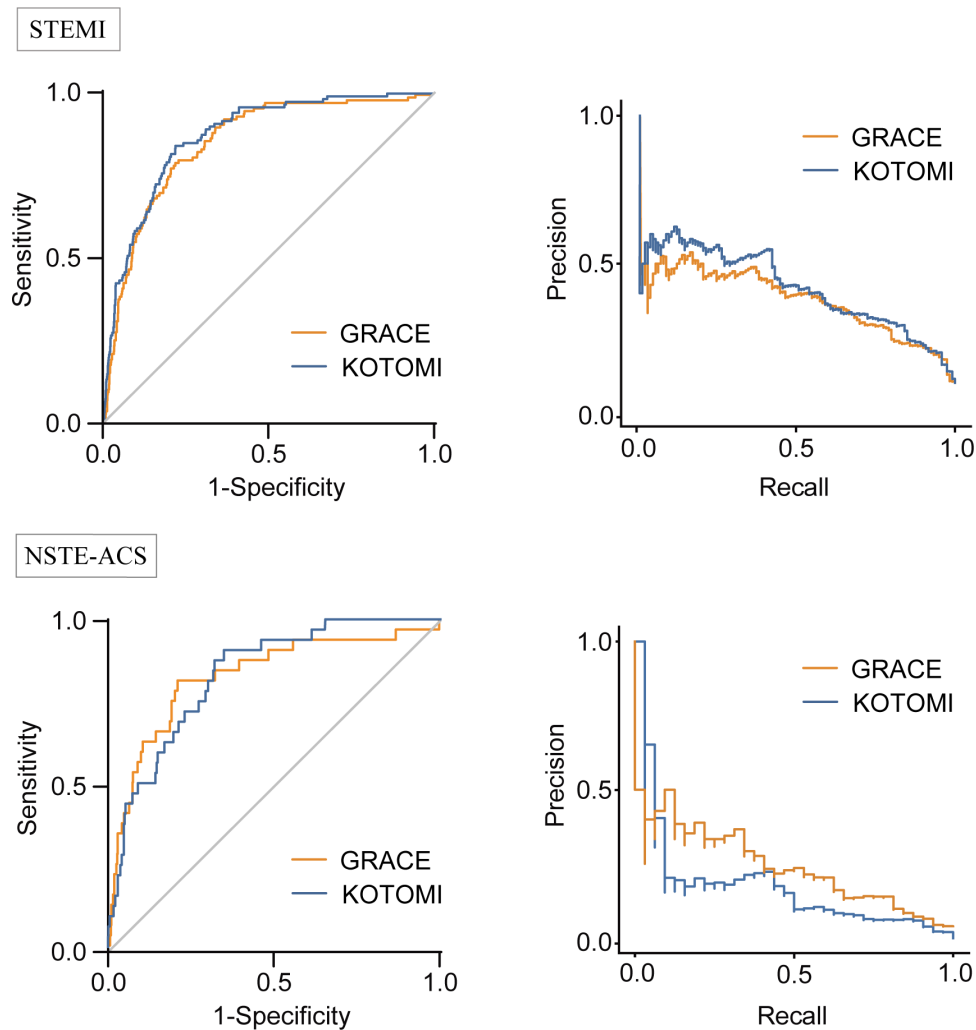


Figure 1 Model accuracy for 30-day mortality of patients with ACS. Left panel shows receiver operating characteristics curve, and right panel shows precision recall curve. ACS, acute coronary syndrome; NSTEMI-ACS, non-ST-elevation acute coronary syndrome; STEMI, ST-elevation acute myocardial infarction.

in patients with STEMI and NSTEMI-ACS (figure 1). The AUROC was 0.85 (95% CI 0.81 to 0.89) for the GRACE and 0.87 (95% CI 0.82 to 0.91) for the KOTOMI in cases of STEMI, reflecting no significant disparity between the models (the difference in AUROC was 0.020 (95% CI -0.098 to 0.13)). The AUROC was 0.82 (0.73–0.91) for the GRACE and 0.83 (0.74–0.91) for the KOTOMI in cases of NSTEMI-ACS, revealing no significant disparity between the models (the difference in AUROC was 0.010 (95% CI -0.023 to 0.25)). The AUPRC was 0.36 for the GRACE and 0.40 for the KOTOMI in cases of STEMI, and 0.22 for the GRACE and 0.20 for the KOTOMI in cases of NSTEMI-ACS. Calibration plot was depicted to assess the consistency between the actual mortality rate and the predictive probability yielded by the prediction models (figure 2). The mortality rate (fraction of positives) was increased along with the average prediction probability in both the GRACE and the KOTOMI for STEMI and NSTEMI-ACS. Nonetheless, for patient groups with high average prediction probability, the probability was inclined to be understated compared with the actual

mortality rate. Overall, both prediction models exhibited high prediction accuracy for patients with ACS including STEMI and NSTEMI-ACS during the COVID-19 pandemic.

Variation of model accuracy across pandemic phases

Subsequently, we analysed the deviations in the prediction accuracy correlating with the COVID-19 cases. Aligning with the monthly COVID-19 case counts in Japan, the pandemic timeline was divided into two distinct periods, with December 2021 serving as the boundary (online supplemental figure 1). The AUROC of both the GRACE and the KOTOMI was compared between the first and second period for patients with STEMI and NSTEMI-ACS, respectively (figure 3). For STEMI, the AUROC was 0.87 (95% CI 0.81 to 0.93) in the first period, 0.83 (95% CI 0.76 to 0.89) in the second, yielding an insignificant disparity of 0.041 (95% CI -0.13 to 0.21) in the GRACE model; concurrently, the KOTOMI model exhibited an AUROC of 0.88 (95% CI 0.82 to 0.93) in the first period and 0.85 (95% CI 0.79 to 0.91) in the second, with a non-significant disparity of 0.028 (95% CI -0.13 to 0.19). For

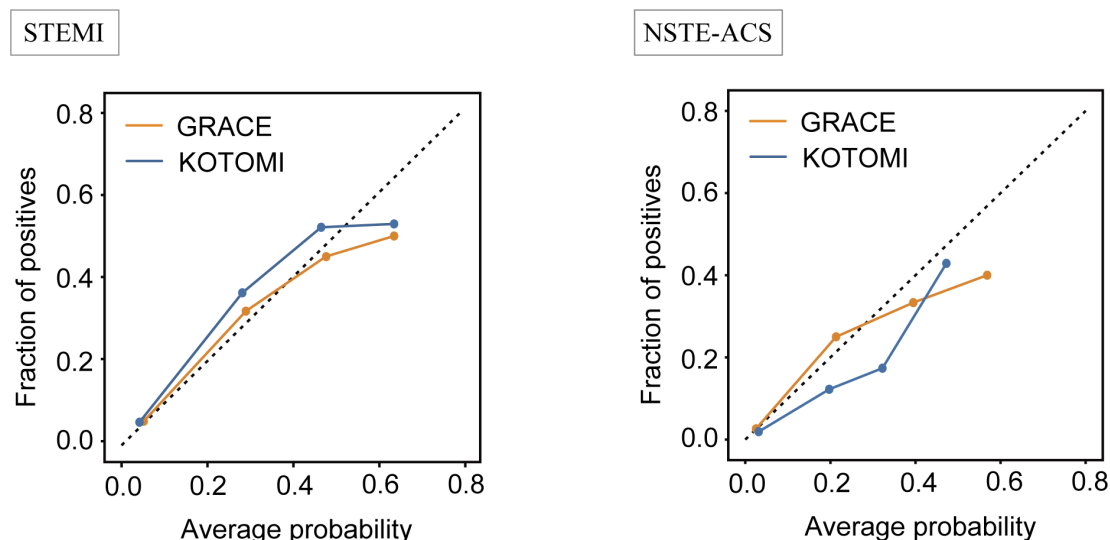


Figure 2 Model calibration for patients with STEMI and NSTEMI-ACS. Samples were divided into four bins according to probability. NSTEMI-ACS, non-ST-elevation-acute coronary syndrome; STEMI, ST-elevation acute myocardial infarction.

NSTEMI-ACS, the AUROC was 0.85 (95% CI 0.74 to 0.95) in the first period, 0.78 (95% CI 0.62 to 0.93) in the second, with an insignificant disparity of 0.072 (95% CI -0.29 to 0.44) in the GRACE model; 0.88 (95% CI 0.78 to 0.97) in the first period, 0.73 (95% CI 0.57 to 0.90) in the second, with an insignificant disparity of 0.14 (95% CI -0.22 to 0.51) in the KOTOMI model. NRIs of the KOTOMI and the GRACE were not significant in any pandemic phases for each ACS group (online supplemental table 2). Collectively, the prediction accuracy of both models had consistency in patients with STEMI, and non-significant variation between the distinct phases of the pandemic periods in patients with NSTEMI-ACS.

DISCUSSION

The present study elucidated that the prediction models, the GRACE and the KOTOMI, have durable prediction

accuracy for 30-day mortality of patients with STEMI and NSTEMI-ACS during the COVID-19 pandemic, and the variation was not significant between the pandemic periods. The model accuracy of the KOTOMI marginally surpassed that of the GRACE for STEMI as well as NSTEMI-ACS, although the disparity of the prediction accuracy was not significant. These findings indicate that the accurate prediction of 30-day mortality of patients with ACS is feasible during the overwhelmed situation under such COVID-19 pandemic, despite its negative impact on medical service and clinical outcomes for patients with ACS.

The GRACE and the KOTOMI model were formulated based on different combination of prediction factors. The GRACE is inclusive of eight predictors, including age, Killip class, BPs, HR, Cr, ST-segment deviation, cardiac arrest during presentation and positive initial

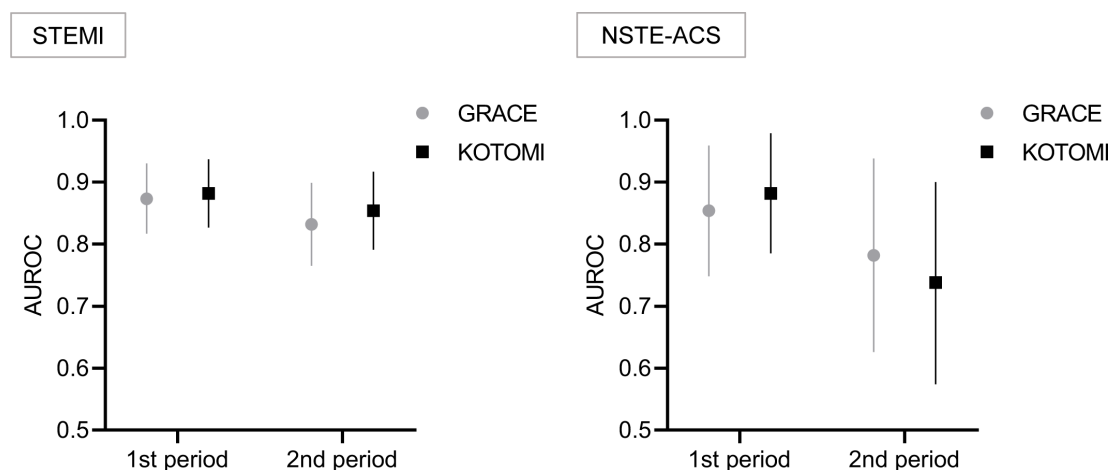


Figure 3 Variation of model accuracy between the pandemic periods. The pandemic period was divided into two periods, with December 2021 as the boundary, and the models' accuracy was evaluated. Error bar indicates 95% CI. AUROC, area under the receiver operating characteristics curve; NSTEMI-ACS, non-ST-elevation-acute coronary syndrome; STEMI, ST-elevation acute myocardial infarction.

cardiac enzyme findings. Conversely, the KOTOMI integrates 10 predictors, including age, Killip class, BPs, HR, WCC, Hb, BS, Cr, CRP and max CPK. Both models were superior to the conventional TIMI (Thrombolysis in Myocardial Infarction) risk index regarding the prediction of in-hospital mortality of patients with STEMI.¹² The GRACE can be used to predict in-hospital mortality as well as 6-month prognosis in patients with STEMI and NSTEMI-ACS.^{15 17} It has enhanced discriminatory power for in-hospital mortality of NSTEMI-ACS compared with the traditional TIMI risk score.^{18 19} Conversely, the KOTOMI was developed to predict in-hospital mortality of patients with STEMI. Machine learning-based prediction model for 1-year adverse events among patients following ACS have been previously developed and validated using data collected before the COVID-19 outbreak.²⁰ Indeed, the evaluation of short-term prognosis is also necessary, given the poor 30-day mortality rate—approximately 10% for STEMI and 4% for NSTEMI-ACS, as shown by our results. In this context, the present study demonstrated the adaptability and robustness of both prediction models for 30-day mortality of patients with STEMI and NSTEMI-ACS even in the COVID-19 pandemic.

The types of biomarkers and their times of measurement are different between the GRACE and the KOTOMI. In contemporary clinical practice for patients with ACS, cardiac troponin has emerged as a pivotal biomarker for early diagnosis due to its high sensitivity.^{21–23} Cardiac enzymes including CPK or CK-MB, as well as troponin, are useful prognostic markers for ACS, reflecting cardiac damage and infarct size.^{24 25} The GRACE uses positive initial cardiac enzyme including troponin, whereas the KOTOMI employs maximum CPK for prediction of mortality of patients with ACS. The GRACE can be used at admission for patients with ACS, while the KOTOMI demonstrates the strength after peak out of CPK levels following the diagnosis by CAG. The present study observed that the KOTOMI marginally surpasses the GRACE without significant difference for 30-day mortality of patients with STEMI and NSTEMI-ACS. However, the types of biomarkers and their times of measurement need to be taken into consideration.

There are several limitations in the study. First, the validation study in other countries is imperative because the pandemic impact on medical system differed between countries. Second, the validation was performed for prediction of 30-day mortality of patients with ACS. Therefore, further study is needed to assess the models' adaptability for longer term outcomes during the pandemic. In addition, it will be warranted to establish a prediction model for antithrombotic risk following ACS, adaptable to situations overwhelmed by emerging infectious disease.

CONCLUSIONS

The GRACE and the KOTOMI model maintained high prediction accuracy for 30-day mortality of patients with STEMI and NSTEMI-ACS during the COVID-19 pandemic.

These prediction models contribute to the improvement of the quality of care and the therapeutic strategy for patients with ACS regardless of overwhelmed situation due to an emerging infectious disease.

Acknowledgements We are grateful to all the participant of AMI-Kyoto Multi-Center Risk Study Group.

Contributors MN was responsible for overall content as guarantor. MN and SM conceived and designed the study. TN and KY curated data. MN analysed data.

Funding This study was supported by Japanese Circulation Society Grant for Future-Pioneering Doctors for Clinical Research.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by the ethics committee of Kyoto Prefectural University of Medicine with the approval number (ERB-C-1865) and adhered to the principles articulated in the Declaration of Helsinki. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Masahiro Nishi <http://orcid.org/0000-0001-8593-3835>

REFERENCES

- Altobelli E, Angeletti PM, Marzi F, *et al*. Impact of SARS-Cov-2 outbreak on emergency department presentation and prognosis of patients with acute myocardial infarction: a systematic review and updated meta-analysis. *J Clin Med* 2022;11:2323.
- Grundeken MJ, Claessen BEPM. Impact of the COVID-19 outbreak on the treatment of myocardial infarction patients. *Curr Treat Options Cardiovasc Med* 2023;2023:1–11.
- Lechner I, Reindl M, Tiller C, *et al*. Impact of COVID-19 pandemic restrictions on ST-elevation myocardial infarction: a cardiac magnetic resonance imaging study. *Eur Heart J* 2022;43:1141–53.
- Cheema HA, Ehsan M, Ayyan M, *et al*. In-hospital mortality of COVID-19 patients hospitalized with ST-segment elevation myocardial infarction: a meta-analysis. *Int J Cardiol Heart Vasc* 2022;43:101151.
- Baldi E, Sechi GM, Mare C, *et al*. Out-of-hospital cardiac arrest during the COVID-19 outbreak in Italy. *N Engl J Med* 2020;383:496–8.
- Chan PS, Girotra S, Tang Y, *et al*. Outcomes for out-of-hospital cardiac arrest in the United States during the Coronavirus disease 2019 pandemic. *JAMA Cardiol* 2021;6:296–303.
- Marijon E, Karam N, Jost D, *et al*. Out-of-hospital cardiac arrest during the COVID-19 pandemic in Paris, France: a population-based, observational study. *Lancet Public Health* 2020;5:e437–43.
- Abu-Assi E, Ferreira-González I, Ribera A, *et al*. Do GRACE (global registry of acute coronary events) risk scores still maintain their performance for predicting mortality in the era of contemporary management of acute coronary syndromes. *Am Heart J* 2010;160:826–34.
- Komiyama K, Nakamura M, Tanabe K, *et al*. In-hospital mortality analysis of Japanese patients with acute coronary syndrome using

- the Tokyo CCU network database: applicability of the GRACE risk score. *J Cardiol* 2018;71:251–8.
- 10 Wójcik M, Karpiak J, Zaręba L, *et al.* The GRACE risk score in patients with ST-segment elevation myocardial infarction and concomitant COVID-19. *Arch Med Sci Atheroscler Dis* 2022;7:e116–23.
 - 11 Dönmez E, Özcan S, Tuğrul S, *et al.* Prognostic value of GRACE risk score in patients hospitalized for Coronavirus disease 2019. *Coron Artery Dis* 2022;33:465–72.
 - 12 Nishi M, Uchino E, Okuno Y, *et al.* Robust prognostic prediction model developed with integrated biological markers for acute myocardial infarction. *PLoS One* 2022;17:e0277260.
 - 13 Shiraishi J, Kohno Y, Sawada T, *et al.* Predictors of in-hospital prognosis after primary percutaneous coronary intervention for acute myocardial infarction requiring mechanical support devices. *Circ J* 2010;74:1152–7.
 - 14 Ministry of Health, Labour and Welfare. Visualizing the data: information on COVID-19 infections. JAPAN, Available: <https://covid19.mhlw.go.jp/> [accessed 18 Sep 2023]
 - 15 Granger CB, Goldberg RJ, Dabbous O, *et al.* Predictors of hospital mortality in the global registry of acute coronary events. *Arch Intern Med* 2003;163:2345–53.
 - 16 R Core Team. R: A language and environment for statistical computing. Austria R Foundation for Statistical Computing V; 2022. Available: <https://www.R-project.org/>
 - 17 Eagle KA, Lim MJ, Dabbous OH, *et al.* A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. *JAMA* 2004;291:2727–33.
 - 18 Yan AT, Yan RT, Tan M, *et al.* Risk scores for risk stratification in acute coronary syndromes: useful but simpler is not necessarily better. *Eur Heart J* 2007;28:1072–8.
 - 19 Antman EM, Cohen M, Bernink PJ, *et al.* The TIMI risk score for unstable angina/non-ST elevation MI: a method for prognostication and therapeutic decision making. *JAMA* 2000;284:835–42.
 - 20 D’Ascenzo F, De Filippo O, Gallone G, *et al.* Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. *Lancet* 2021;397:199–207.
 - 21 Kimura K, Kimura T, Ishihara M, *et al.* JCS 2018 guideline on diagnosis and treatment of acute coronary syndrome. *Circ J* 2019;83:1085–196.
 - 22 Byrne RA, Rossello X, Coughlan JJ, *et al.* 2023 ESC guidelines for the management of acute coronary syndromes. *Eur Heart J* 2023;44:3720–826.
 - 23 Gulati M, Levy PD, Mukherjee D, *et al.* 2021 AHA/ACC/AASE/CHEST/SAEM/SCCT/SCMR guideline for the evaluation and diagnosis of chest pain: executive summary: a report of the American college of cardiology/American heart association joint committee on clinical practice guidelines. *Circulation* 2021;144:e368–454.
 - 24 Halkin A, Stone GW, Grines CL, *et al.* Prognostic implications of creatine kinase elevation after primary percutaneous coronary intervention for acute myocardial infarction. *J Am Coll Cardiol* 2006;47:951–61.
 - 25 Chia S, Senatore F, Raffel OC, *et al.* Utility of cardiac biomarkers in predicting infarct size, left ventricular function, and clinical outcome after primary percutaneous coronary intervention for ST-segment elevation myocardial infarction. *JACC Cardiovasc Interv* 2008;1:415–23.

© 2024 Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. <http://creativecommons.org/licenses/by-nc/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.