

Prediction of high-risk emergency department revisits from a machine-learning algorithm: a proof-of-concept study

Chih-Wei Sung,^{1,2} Joshua Ho,^{3,4} Cheng-Yi Fan,¹ Ching-Yu Chen,⁵
Chi-Hsin Chen ,¹ Shao-Yung Lin,⁶ Jia-How Chang ,^{1,2} Jiun-Wei Chen,¹
Edward Pei-Chuan Huang ^{1,2,6}

To cite: Sung C-W, Ho J, Fan C-Y, *et al.* Prediction of high-risk emergency department revisits from a machine-learning algorithm: a proof-of-concept study. *BMJ Health Care Inform* 2024;**31**:e100859. doi:10.1136/bmjhci-2023-100859

Received 18 July 2023
Accepted 09 March 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY. Published by BMJ.

¹Department of Emergency Medicine, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu, Taiwan

²Department of Emergency Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan

³Institute of Information Science, Academia Sinica, Taipei, Taiwan

⁴Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan

⁵Department of Emergency Medicine, National Taiwan University Hospital Yun-Lin Branch, Douliou, Taiwan

⁶Department of Emergency Medicine, National Taiwan University Hospital, Taipei, Taiwan

Correspondence to

Dr Edward Pei-Chuan Huang;
edward56026@gmail.com

ABSTRACT

Background High-risk emergency department (ED) revisit is considered an important quality indicator that may reflect an increase in complications and medical burden. However, because of its multidimensional and highly complex nature, this factor has not been comprehensively investigated. This study aimed to predict high-risk ED revisit with a machine-learning (ML) approach.

Methods This 3-year retrospective cohort study assessed adult patients between January 2019 and December 2021 from National Taiwan University Hospital Hsin-Chu Branch with high-risk ED revisit, defined as hospital or intensive care unit admission after ED return within 72 hours. A total of 150 features were preliminarily screened, and 79 were used in the prediction model. Deep learning, random forest, extreme gradient boosting (XGBoost) and stacked ensemble algorithm were used. The stacked ensemble model combined multiple ML models and performed model stacking as a meta-level algorithm. Confusion matrix, accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUROC) were used to evaluate performance.

Results Analysis was performed for 6282 eligible adult patients: 5025 (80.0%) in the training set and 1257 (20.0%) in the testing set. High-risk ED revisit occurred for 971 (19.3%) of training set patients vs 252 (20.1%) in the testing set. Leading predictors of high-risk ED revisit were age, systolic blood pressure and heart rate. The stacked ensemble model showed more favourable prediction performance (AUROC 0.82) than the other models: deep learning (0.69), random forest (0.78) and XGBoost (0.79). Also, the stacked ensemble model achieved favourable accuracy and specificity.

Conclusion The stacked ensemble algorithm exhibited better prediction performance in which the predictions were generated from different ML algorithms to optimally maximise the final set of results. Patients with older age and abnormal systolic blood pressure and heart rate at the index ED visit were vulnerable to high-risk ED revisit. Further studies should be conducted to externally validate the model.

INTRODUCTION

Emergency department (ED) revisit is a well-known quality index for ED medical care and

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ High-risk emergency department (ED) revisits can potentially be prevented in advance. However, the predictive model for high-risk ED return is yet to be determined in this particular cohort.

WHAT THIS STUDY ADDS

⇒ Four machine-learning models were employed to predict high-risk ED revisits. Among these, the stacked ensemble algorithm demonstrated superior predictive performance compared with the other artificial intelligence (AI) models, achieving an area under the receiver operating characteristic curve value of 0.82. Notably, all AI models outperformed the traditional logistic regression model in terms of predictive accuracy.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ ED physicians should identify patients potentially at high risk for ED revisits to prevent further deterioration of their condition.

patient safety, in which revisit rates >5% may reflect poor quality of care, and those <1% indicate undue risk aversion.¹ Previous studies indicated that ED revisit may increase medical costs, ED crowding and poor prognosis, particularly in patients who require hospital admission, often due to rapid deterioration after ED discharge.²⁻⁴ Over the past decade, this concept has become challenging because the factors that influence ED revisit are multifactorial, such as issues related to diagnosis, management, procedural complications and medical adverse effects.^{3 5 6} Most issues are preventable and do not result in severe outcomes.⁷ Recent studies of intrinsic factors for high-risk ED revisit focused on patients who received hospital admission or intensive care unit (ICU) care.⁸⁻¹⁰

To identify potential risk factors for either high-risk ED revisit or unscheduled ED revisit within 72 hours, logistic linear regression models are widely used. Well-known factors for high-risk ED revisit include age, male sex, ambulance transport for return visit, longer ED length of stay, symptoms of dyspnoea or chest pain on ED presentation, triage level 1 or 2, acute change in levels of consciousness and unstable vital signs (tachycardia and/or fever), among others.^{8 11 12} However, a study of ED revisit has been limited by the use of linear algorithms, such as logistic regression routine, use of administrative data and small sample sizes, in part because the assessment of risk factors is more complicated than that possible with linear association.¹³

With the development of artificial intelligence, the machine-learning (ML)-based prediction model is used now as a clinical classifier. Lee *et al* developed an ML framework combining a particle swarm optimisation feature selection algorithm and an optimisation-based discriminant analysis model, to predict ED revisit. Hong *et al* indicated that gradient-boosting models that leveraged clinical data were superior to traditional logistic regression models built on administrative data to predict ED revisit.¹⁴ Hsu *et al* developed an ML model, the voting classifier model, to predict ED revisit in patients with abdominal pain.¹⁵ These works shed light on the use of a prediction model for ED revisit based on an ML algorithm.

In previous ML-based studies, the work by Lee and Hong focused on building a prediction model for general ED revisit, whereas that of Hsu focused on ED revisit and abdominal pain symptoms. All prediction models showed superior prediction performance than that with a traditional logistic regression model. Expanding on these previous works, in the current study, we specifically predict high-risk ED revisit in 72 hours using a large dataset of adult ED revisits, with more than 150 variables extracted per visit from each medical record. Our study used a powerful classification algorithm—the stacked ensemble model. Also, a comprehensive comparison between models and previous reports is presented.

MATERIAL AND METHODS

Study design, participants and setting

This study recruited patients who demonstrated unscheduled ED revisit within 72 hours between January 2019 and December 2021 from National Taiwan University Hospital Hsin-Chu Branch (NTUH-HCH), a tertiary centre with 829-bed capacity and more than 1700 staff. About 60 000 patients visit the ED each year; on average, 4.5% of these patients demonstrated an ED revisit after the index discharge. Patients were eligible for recruitment and analysis if they were age 20 years or older and demonstrated an ED revisit within 72 hours, whereas those who demonstrated ED revisit simply for diagnostic certificate or legal issue were immediately excluded.

Data source, features and preprocessing

For data acquisition, independent ED attending physicians retrospectively reviewed the medical charts rather than extracting information from the integrated medical database to minimise the biases and errors in the original medical record. For data dimensions, 150 features were initially included, such as age, sex, pre-existing diseases, diagnosis, final disposition and two sets of covariates from the ED index and revisit. Each set contained triage level, vital signs, chief concern, management, medication and laboratory data. Pre-existing diseases were hypertension, diabetes mellitus, coronary artery disease, cerebrovascular disease, chronic kidney disease, malignancy, chronic obstructive pulmonary disease and previous documented surgery.

Triage level was determined by the Taiwan Triage and Acuity Scale computerised triage system, which has been validated with levels 1–5 to indicate resuscitation, emergent, urgent, less urgent and non-urgent.¹⁶ Vital signs included body temperature, respiratory rate, heart rate, blood pressure and oxygen saturation.

Chief concerns, originally written on medical charts, were recorded and classified by ED attending physicians into 30 common concerns, such as headache, vertigo, chest pain, short of breath, cough, rhinorrhoea, abdominal pain, nausea, vomiting, diarrhoea, dysuria, frequent urination, retention of urine, chills, limb oedema and tube malfunction, among others.

Management included electrocardiography, chest radiography, CT, MRI, panendoscopy, colonoscopy and specialist consultation with any formal consultation from surgeons, radiologists or intensivists. Medications included analgesics and antibiotics, either orally or intravenously. Laboratory data included serum concentrations of white cell count, haemoglobin, sodium, potassium and C reactive protein; blood gas analysis; and liver function and renal function tests. Diagnosis was categorised into infection, neurological diseases, circulation diseases, respiratory disease, gastrointestinal diseases, genitourinary diseases and musculoskeletal diseases. To predict high-risk ED revisit, the features in the index visit should be included, and those in the revisit should be reasonably excluded. A total of 79 features were used for data training (online supplemental figure 1).

For data cleaning, nonsense records were first removed. For unreasonable values for the feature, we re-examined the medical record to confirm the correctness. Because the rate of missing data was 4.3%, with most missing variables missing at random, mean imputation was used to replace missing values for a specific feature by the mean of non-missing cases for that feature. For data aggregation, we aggregated the feature according to its characteristic. We set body temperature as a binary feature based on whether it ranged between 36.0°C and 37.4 °C or not. In addition, blood gas features (eg, pH value, partial pressure of carbon dioxide) were also aggregated for analysis.

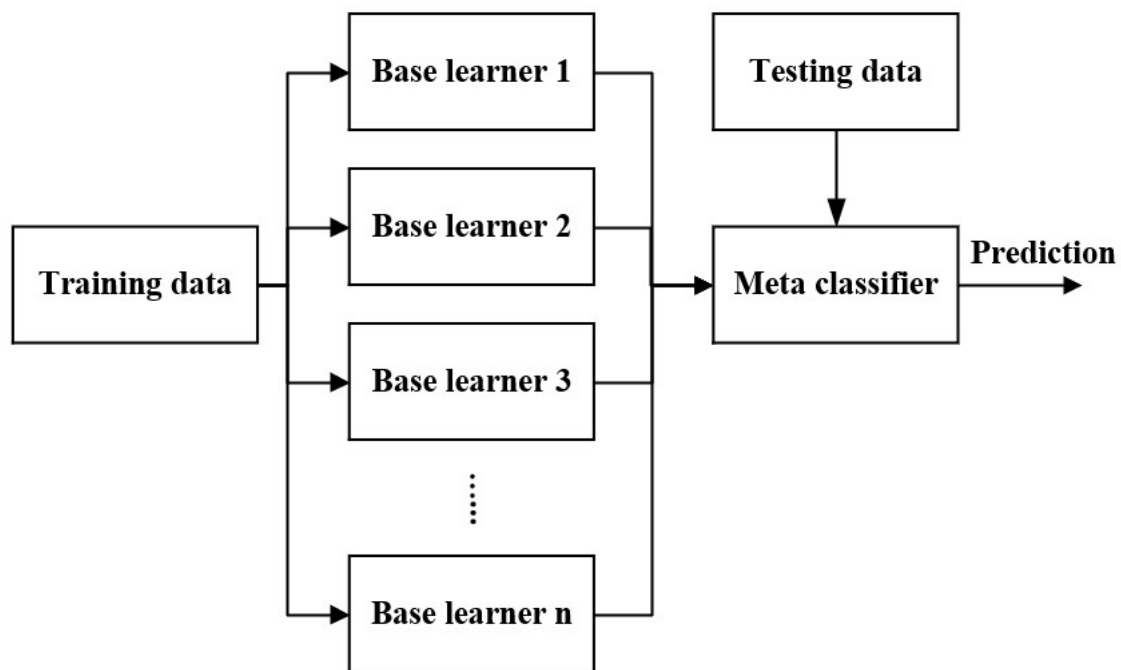


Figure 1 The stacked ensemble algorithm.

Stacked ensemble algorithm

In ML, the ensemble method uses multiple learning algorithms, to achieve better predictive performance than that from any single-constituent learning algorithm.¹⁷ The principle of the ensemble method is to combine the predictions from multiple existing models or algorithms of the same or different types named after base learners, to further fine-tune the model. This approach creates a more robust system that combines the predictions from all base learners. By stacking multiple layers of ML models, each model carries its prediction to the layer above it, and the top layer model takes the final decision (figure 1).

ML model and training

The ML model in this study included deep learning, random forest, extreme gradient boost (XGBoost) and stacked ensemble. The training set included 80% of subjects, while the remaining 20% of subjects were included in the testing set. To select the best model for the final testing dataset, we trained each model by 10-fold cross-validation. To increase the performance and prediction capacity according to our selected best base models, we proposed a stacked ensemble algorithm for the experiments in the base model. We performed hyperparameter tuning for each model. For deep learning, we used Bayesian optimisation based on the Gaussian process. For the random forest model, a random search algorithm was used because the decision tree was complex. For XGBoost, we tuned the hyperparameters based on Bayesian optimisation.

Outcome measurement

For the ‘final disposition’ feature, high-risk ED revisit was defined as when a patient was admitted to hospital,

including ICU admission or died, whereas low-risk ED revisit indicated a direct discharge after the return. Patients who were discharged against medical advice or transferred to other hospitals were excluded from the analysis.

Statistical analysis

The features were computed by using SAS V.9.4 (SAS Institute). The Wilcoxon rank-sum test was applied to examine the significant differences among features when the features were continuous type, and the χ^2 test was used for those that were categorical. A two-sided $p < 0.05$ indicated statistical significance.

To compare performance between models, the area under the receiver operating characteristic curve (AUROC), accuracy, sensitivity and specificity were used. Accuracy indicated the number of high-risk and low-risk ED revisits that were correctly predicted. Sensitivity indicated the number of cases that were correctly predicted as high-risk ED revisit among all true high-risk ED revisits. Specificity indicated the number of cases that were correctly predicted as low-risk ED revisit among all low-risk ED revisits.

RESULTS

Study flow and ML assignment

Figure 2 demonstrates the study population and assignment to the training and testing sets. A total of 7699 preliminary patients who demonstrated an unscheduled ED revisit within 72 hours were recorded. Patients aged younger than 20 years ($n=1365$, 17.7%) and those who demonstrated a discharge against medical advice ($n=29$, 0.4%) or hospital transfer ($n=23$, 0.3%) were excluded.

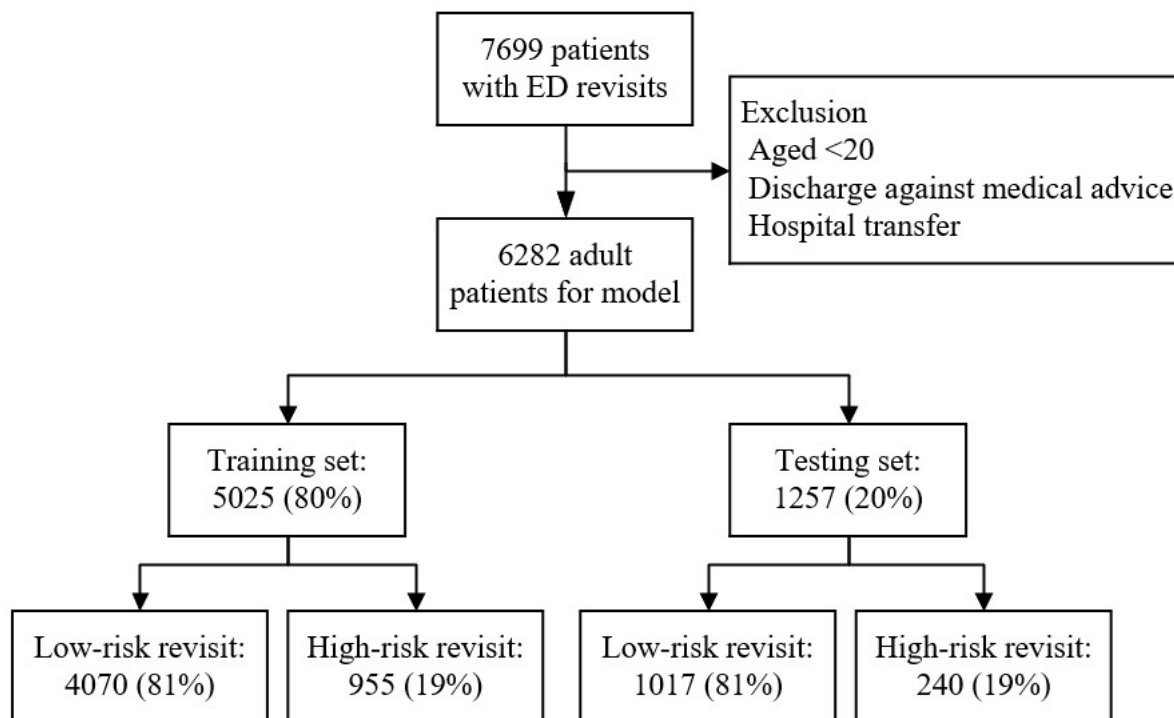


Figure 2 Flow chart for the inclusion of eligible subjects. ED, emergency department.

After exclusion, 6282 adult patients were divided into two subgroups: 5025 for training (80.0%) and 1257 for testing (20.0%). High-risk ED revisit was found for 971 (19.3%) patients in the training set vs 252 (20.1%) in the testing set.

Data from the training and testing cohort

Table 1 presents a comparison of characteristics between the training and testing cohorts. These cohorts were randomly selected from the population at an 80:20 ratio. In the training cohort, 971 patients (19.32%) experienced a high-risk revisit, compared with 252 patients (20.05%) in the testing cohort. There were no significant differences between the two cohorts in various aspects, including age, sex, year and month of enrolment, time from discharge to ED revisit, common pre-existing diseases, triage information, complaints and laboratory data. The process of randomisation achieved a balanced distribution across both cohorts.

Variable extraction and importance

The preliminary 151 features were included for screening, such as demographic data, pre-existing diseases and information on the index visit (month, visit time, triage level, vital signs and chief concerns) and the revisit (revisit time, triage level, vital signs, chief concerns, laboratory data and disposition). To provide a rationale and optimise the prediction for high-risk ED revisit, a total of 79 features, including the data from the index visit, were eventually included in the model for prediction.

Figure 3 shows the scaled importance of each variable. The importance variables were proposed based on the XGBoost model. Age was the most important feature for

predicting high-risk ED revisit. In general, the leading features in the index visit other than age were systolic blood pressure, heart rate, month for visit, diastolic blood pressure and body temperature. For serum tests at the index visit, concentrations of neutrophils, creatinine and white cell count were important biomarkers to predict high-risk ED revisit, whereas alanine transaminase, sodium, potassium and glucose were minor ones. For chief concerns, only skin-related concerns or medical device issues contributed to high-risk ED revisit. For the physician's management at the index visit, the feature of oral analgesic administration after discharge was a factor for high-risk ED revisit, but with low-scale importance. Patient sex was less relevant than other factors for high-risk ED revisit.

Performance comparison of each model

Figure 4 is a diagram that displays the true-positive versus false-positive rates and the AUROC that was then calculated. The stacked ensemble model exhibited the highest AUROC (0.82), which was significantly higher than that in the XGBoost model (0.79), random forest model (0.78) and deep-learning model (0.69). Table 2 further summarises the performance of each model in terms of accuracy, sensitivity and specificity. Other than AUROC, the four models demonstrated a similar level of accuracy, which ranged from 0.85 to 0.87. All models demonstrated almost the same sensitivity of 0.45. For specificity, the stacked ensemble model achieved 0.90, followed by the random forest model (0.88), XGBoost model (0.88) and deep-learning model (0.75).

Table 1 Comparison of demographics and medical information between training and testing cohorts

Variables	Training cohort (n=5025)	Testing cohort (n=1257)	P value
High-risk revisit	971 (19.32)	252 (20.05)	0.875
Age	58.6±19.6	58.1±19.8	0.382
Males	2650 (52.74)	664 (52.82)	0.923
Year			0.492
2019	1975 (39.30)	511 (40.65)	
2020	1547 (30.79)	389 (30.95)	
2021	1433 (28.52)	338 (26.89)	
Month			0.189
January	458 (9.11)	123 (9.79)	
February	409 (8.14)	102 (8.11)	
March	412 (8.20)	93 (7.40)	
April	374 (7.44)	107 (8.51)	
May	429 (8.54)	90 (7.16)	
June	384 (7.64)	105 (8.35)	
July	436 (8.68)	103 (8.19)	
August	433 (8.62)	136 (10.82)	
September	418 (8.32)	107 (8.51)	
October	449 (8.94)	92 (7.32)	
November	394 (7.84)	93 (7.40)	
December	359 (7.14)	87 (6.92)	
Return to ED			0.196
<24 hours	2357 (46.91)	554 (44.07)	
24–48 hour	1559 (31.02)	406 (32.30)	
48–72 hours	1039 (20.68)	278 (22.12)	
Pre-existing diseases			
Hypertension	1774 (35.30)	432 (34.37)	0.737
Diabetes mellitus	1067 (21.23)	268 (21.32)	0.931
Coronary artery disease	528 (10.51)	115 (9.15)	0.159
Cerebrovascular disease	221 (4.40)	45 (3.58)	0.201
Malignancy	802 (15.96)	213 (16.95)	0.386
Chronic kidney disease	374 (7.44)	86 (6.84)	0.471
COPD	153 (3.04)	50 (3.98)	0.093
Triage			
Glasgow Coma Scale (=15)	4783 (95.18)	1197 (95.23)	0.783
Triage level 1 or 2	807 (16.06)	195 (15.51)	0.647
Systolic blood pressure (mm Hg)	149.6±31.9	150.2±31.6	0.584
Diastolic blood pressure (mm Hg)	82.4±16.7	82.6±16.7	0.705
Pulse rate	91.6±19.6	92.1±20.0	0.493
Breath rate	20.2±2.8	20.3±2.8	0.791
Body temperature	36.9±0.8	36.9±0.8	0.362
Symptoms or complaints			
Headache	329 (6.55)	81 (6.44)	0.901
Chest pain	432 (8.60)	95 (7.56)	0.239
Dyspnoea	354 (7.04)	94 (7.48)	0.586
Abdominal pain	1097 (21.83)	253 (20.13)	0.194
Vomiting	503 (10.01)	116 (9.23)	0.412

Continued

Table 1 Continued

Variables	Training cohort (n=5025)	Testing cohort (n=1257)	P value
Diarrhoea	253 (5.03)	66 (5.25)	0.748
Skin disorders	330 (6.57)	87 (6.92)	0.644
Leg oedema	128 (2.55)	28 (2.23)	0.518
Tube malfunction	241 (4.80)	65 (5.17)	0.575
Examination and blood data			
Electrocardiography	1554 (30.93)	415 (33.02)	0.144
Chest radiograph	2672 (53.17)	628 (49.96)	0.436
White cell count (x10 ⁹ /L)	9.3±4.7	9.1±3.9	0.324
Neutrophil (%)	73.5±13.2	74.4±30.9	0.252
Haemoglobin (g/L)	128±26	129±25	0.192
Serum creatinine (mg/dL)	1.4±1.9	1.4±1.8	0.648

COPD, chronic obstructive pulmonary disease; ED, emergency department.

DISCUSSION

In the current study, we developed an ML approach to predict high-risk ED revisit within 72 hours. Age, systolic blood pressure and heart rate were leading features contributing to high-risk ED revisit, followed by diastolic

blood pressure and body temperature. In brief, age and vital signs in the index visit could predict high-risk ED revisit. Our results differ from those of previous reports because the features in the prediction model were cleansed and corrected after the data were reviewed and

▼ VARIABLE IMPORTANCES

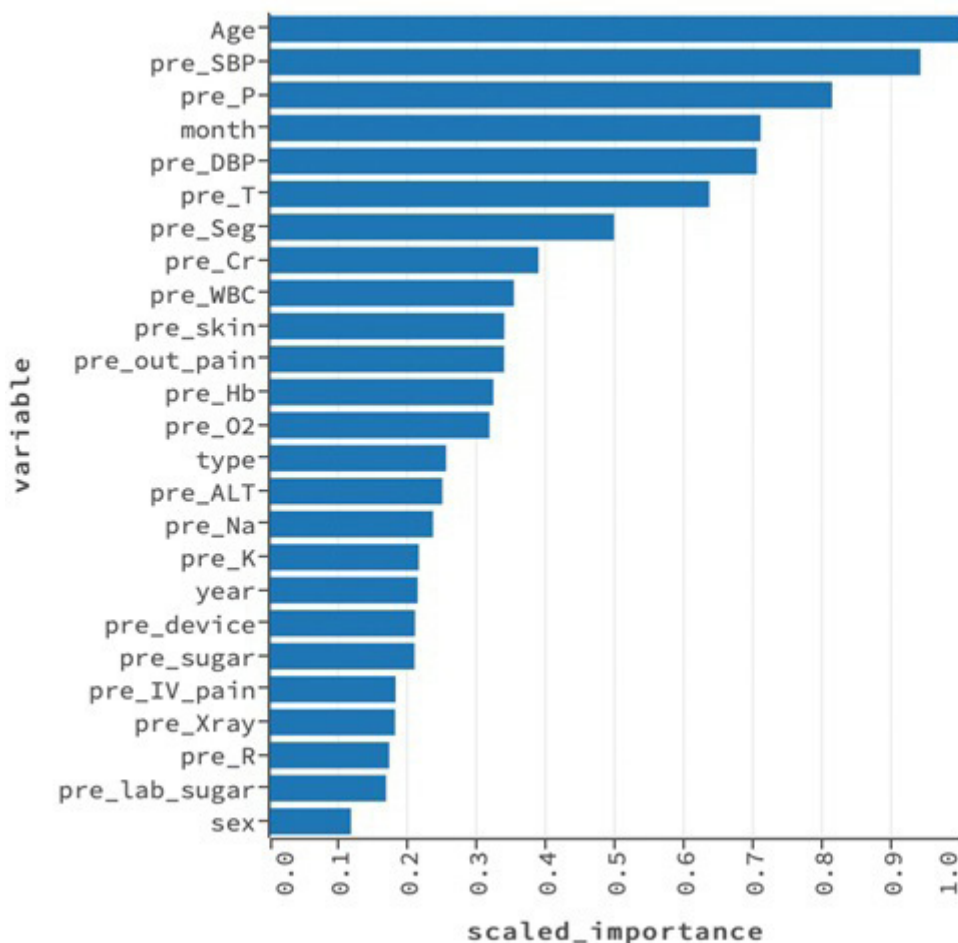
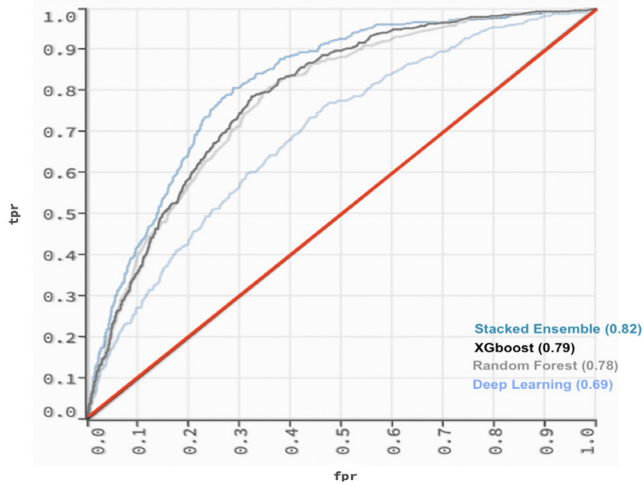


Figure 3 The scaled importance of the features.

A ROC CURVE - CROSS VALIDATION METRICS , AUC = 0.815861



B

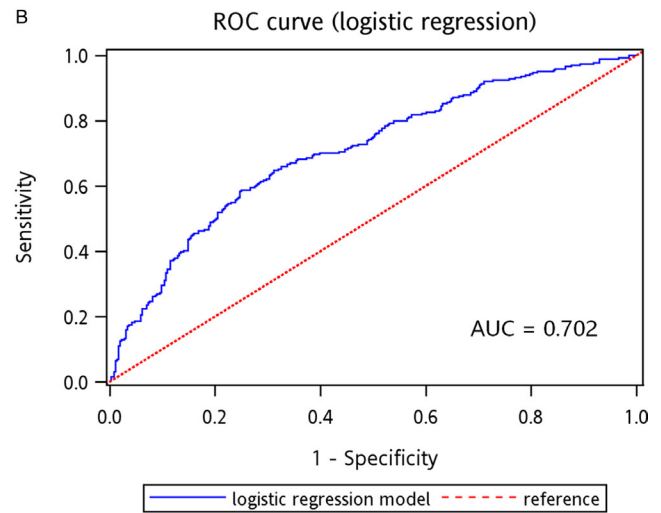


Figure 4 (A) Performance comparison of each AI model. (B) Performance of the logistic regression model. AI, artificial intelligence; AUC, area under the curve; ROC, receiver operating characteristic curve.

examined by physicians for each patient, rather than using diagnostic coding or database extraction. In addition, we used an ML approach to overcome the non-linearity and high dimensional features that was difficult to process in the traditional logistic regression. We used the stacked ensemble method, which combined several base learners. As a super learner, the stacked ensemble algorithm increased the prediction performance compared with deep learning, XBG and random forest. The AUROC in this method achieved 0.82. Significantly, our findings revealed that the specificity of ML models greatly surpassed that of traditional logistic regression, making them more effective as predictive tools for users.

Our study indicated that only age and vital signs in the index ED visit could predict high-risk ED revisit, which resulted in some consistent and some controversial findings, compared with previous reports.^{7 8 11 12 18} First, vital signs are important at each ED visit because they reflect the disease condition. In a previous case-crossover study, arrival by ambulance, dyspnoea or chest pain on ED presentation, high triage levels, acute change in levels of consciousness, tachycardia (>90 beats/min) and high fever (>39°C) were associated with high-risk ED revisit.⁸ Vital signs including heart rate and body temperature were important risk factors and features in the regression model and prediction model, respectively. Patients

with unstable or abnormal vital signs not only reflected immediate urgency but also a potential high-risk revisit if the patient was discharged from the ED. As for other variables, such as high triage level or vital sign-related symptoms, the association was undoubtedly made. Another retrospective study indicated that male sex, ambulance transport at return visit and longer length of stay were associated with higher risks of admission among ED 72-hour return visits.¹¹ Age was adjusted after multivariate regression. However, these factors associated with high-risk ED visit may not reflect the disease condition because the chief concerns, diagnosis and laboratory data were not obtained. Whether the factors of male sex, ambulance transport or longer ED stay were associated with specific diseases remains unknown. In addition, another study included older age, multiple comorbidities and worsening severity index as prognostic factors for poor outcome in high-risk ED revisit. In that study, the overall mortality rate was almost 20%.¹²

Also, our report demonstrated that ML would be a better approach to provide a prediction model than multivariate logistic regression, which mainly focuses on the association between dependent and independent variables. The AUROC in this study was almost the same as that in previous studies that used ML technique, approximately 0.74–0.83 with a different algorithm.^{13–15} However,

Table 2 The comparison of performance in each model

	AUC	Accuracy	Sensitivity	Specificity
Logistic regression	0.64	0.81	0.53	0.70
Deep learning	0.69	0.85	0.45	0.75
Random forest	0.78	0.85	0.46	0.88
XGBoost	0.79	0.85	0.45	0.88
Stacked ensemble	0.82	0.87	0.45	0.90

AUC, area under the curve.

Table 3 Logistic regression model with stepwise selection

Variable	aOR	95% CI	P value
Age	1.019	(1.014 to 1.024)	<0.001
Male	1.305	(1.095 to 1.555)	0.003
Systolic blood pressure	0.992	(0.990 to 0.995)	<0.001
Body temperature (fever)	1.496	(1.360 to 1.645)	<0.001
White cell count (x10 ⁹ /L)	1.086	(1.061 to 1.111)	<0.001
Serum creatinine (mg/dL)	1.073	(1.024 to 1.124)	0.003
Return to ED			
<24 hours	Ref		
24–48 hours	0.957	(0.787 to 1.164)	0.155
48–72 hours	0.683	(0.540 to 0.864)	0.005
Malignancy	1.426	(1.131 to 1.801)	0.003
COPD	1.891	(1.229 to 2.908)	0.004
Triage level 1 or 2	1.551	(1.234 to 1.948)	0.001
Chest pain	0.576	(0.425 to 0.781)	0.001
Abdominal pain	1.388	(1.140 to 1.689)	0.001
Leg oedema	2.004	(1.211 to 3.316)	0.007
Tube malfunction	0.344	(0.146 to 0.809)	0.014

aOR, adjusted OR; COPD, chronic obstructive pulmonary disease; ED, emergency department.

in some studies with multivariate logistic regression, the concordance (C)-statistic, which is often used to assess the ability of a risk factor to predict outcome, ranged from 0.55 to 0.74.⁴ This finding is not surprising because the cause of high-risk ED revisit was multifactorial. The ML approach was suitable for dealing with high dimensional features, non-linear and complicated features, not simply applied to administrative data.^{19 20} To enhance the performance of the current model, subgroup analysis may be considered, especially considering that certain cohorts, such as patients aged over 75 years old. In our comparison with traditional multivariable logistic regression, we observed notable differences in the variables identified as significant compared with those in the ML model. [Table 3](#) presents the results of the multivariable logistic regression model, which was developed using a stepwise selection process. Several factors were consistently associated with high-risk ED revisits across both the logistic regression and ML models. These included age, sex, systolic blood pressure, the presence or absence of fever, serum levels of white cell count or creatinine, time until return to the ED and issues related to tube malfunction (device issue). However, it was important to highlight that some symptoms such as chest pain, abdominal pain and leg oedema, which might be intuitively assumed as critical, did not emerge as key features in the ML models. Furthermore, certain pre-existing diseases such as chronic obstructive pulmonary disease and malignancy, identified as risk factors for high-risk ED revisits in the logistic regression model, were obviously absent in the ML model. Additionally, our analysis revealed that some features were deemed

significant in the ML model but did not demonstrate a similar impact in the logistic regression model. These features included serum levels of electrolytes, the month of the ED visit and liver function tests. Such disparities underscore the distinct analytical perspectives offered by different model approaches. The ML model, with its ability to capture complex interactions and non-linear relationships, may identify subtle patterns not apparent in traditional logistic regression analysis. This difference in model sensitivity and specificity highlighted the need to carefully interpret and understand the implications of each model's findings, particularly in the context of predicting high-risk ED revisits.

Study limitations

This study demonstrated several limitations. First, it was a single-centre study, which may cause selection bias. Although the study duration was 3 years and sample size was sufficient, the patient condition or disease type may be restricted to the local region. In addition, if a patient with a potential high-risk ED revisit chose another hospital after ED discharge, the study may fail to include these cases. Second, because high-risk ED revisit is multifactorial, some features were not collected, which may cause information bias, particularly for qualitative features. One study indicated that the patient being told to 'return if unwell' (22.7%) and being seen faster after returning to the ED (12.5%) were associated with ED revisit²¹; however, this information could not be obtained from medical records. Third, the patients in this study were those who had both the index and return ED visit in our

hospital. Some patients if they had return visit to other hospitals could not be controlled. Fourth, high-risk ED revisit was also associated with uncertain or missed diagnoses, causing inappropriate dispositions,²² but we did not further follow the misdiagnosis rate. Fifth, the study faced a challenge with an unbalanced population, where only 4% were rehospitalised. Consequently, the positive predictive value, a critical metric for users, was merely 0.16, indicating a need for improvement in future studies. Sixth, the first 26 characteristics accounted for only 8% of the model's performance, as illustrated in figure 3. This suggests that the remaining 126 features contribute at most 12% to the model's efficacy. Considering this, alternative analytical approaches, such as SHAPE analysis, may be worth exploring. Lastly, this model was not validated in another new cohort. Further external validation would be warranted.

CONCLUSION

In this study, we used ML to predict high-risk ED revisit. The stacked ensemble algorithm exhibited better prediction performance compared with random forest, XGB and deep-learning models. The leading features in the prediction model were age, systolic blood pressure and heart rate in the index ED visit. To determine whether this ML model can be externally validated in other clinical settings with the same performance, further evaluation is required.

Acknowledgements We thanked the centre of intelligent healthcare, NTUH, for data cleansing, data processing, model training and testing.

Contributors All authors contributed substantially to this work. C-WS and EP-CH originally conceived of the project. C-YC, C-HC, S-YL, J-HC and J-WC initially collected data. JH and C-WS performed data analysis. C-WS wrote the manuscript. EP-CH and J-WC gave critical revisions. EP-CH took full responsibility for the work and/or the conduct of the study, had access to the data, and controlled the decision to publish. All authors have read and approved this manuscript.

Funding This work was funded by National Taiwan University Hospital Hsin-Chu Branch (grant number 110-HCH019 and 111-HCH061).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Written informed consent was waived because of the provision of minimal interventions. The study was approved by the Institutional Review Board of the NTUH-HCH, and it was conducted in accordance with Declaration of Helsinki and International Council on Harmonisation Good Clinical Practice guidelines.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on 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 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any

purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iDs

Chi-Hsin Chen <http://orcid.org/0000-0003-1826-6598>

Jia-How Chang <http://orcid.org/0000-0003-3545-4507>

Edward Pei-Chuan Huang <http://orcid.org/0000-0002-4800-2561>

REFERENCES

- Heyworth J. Emergency medicine-quality indicators: the United Kingdom perspective. *Acad Emerg Med* 2011;18:1239–41.
- Hu K-W, Lu Y-H, Lin H-J, et al. Unscheduled return visits with and without admission post emergency department discharge. *J Emerg Med* 2012;43:1110–8.
- Calder L, Pozgay A, Riff S, et al. Adverse events in patients with return emergency department visits. *BMJ Qual Saf* 2015;24:142–8.
- Pellerin G, Gao K, Kaminsky L. Predicting 72-hour emergency department revisits. *Am J Emerg Med* 2018;36:420–4.
- Chang C-S, Lee K-H, Su H-Y, et al. Physician-related factors associated with unscheduled revisits to the emergency department and admission to the intensive care unit within 72 H. *Sci Rep* 2020;10.
- Cozzi G, Ghirardo S, Fiorese I, et al. Risk of hospitalisation after early-revisit in the emergency department. *J Paediatr Child Health* 2017;53:850–4.
- Lin C-F, Huang Y-S, Tsai M-T, et al. In-hospital outcomes in patients admitted to the intensive care unit after a return visit to the emergency department. *Healthcare (Basel)* 2021;9:431.
- Sung C-W, Lu T-C, Fang C-C, et al. Factors associated with a high-risk return visit to the emergency department: a case-crossover study. *Eur J Emerg Med* 2021;28:394–401.
- Hiti EA, Tamim H, Makki M, et al. Characteristics and determinants of high-risk unscheduled return visits to the emergency department. *Emerg Med J* 2020;37:79–84.
- Miller KEM, Duan-Porter W, Stechuchak KM, et al. Risk stratification for return emergency department visits among high-risk patients. *Am J Manag Care* 2017;23:e275–9.
- Liu SW. Risk factors of admission in 72-H return visits to emergency department. *Tzu Chi Med J* 2021;33:169–74.
- Fan J-S, Kao W-F, Yen DH-T, et al. Risk factors and prognostic predictors of unexpected intensive care unit admission within 3 days after ED discharge. *Am J Emerg Med* 2007;25:1009–14.
- Lee EK, Yuan F, Hirsh DA, et al. A clinical decision tool for predicting patient care characteristics: patients returning within 72 hours in the emergency department. *AMIA Annu Symp Proc* 2012;2012:495–504.
- Hong WS, Haimovich AD, Taylor RA. Predicting 72-hour and 9-day return to the emergency department using machine learning. *JAMIA Open* 2019;2:346–52.
- Hsu C-C, Chu C-C, Lin C-H, et al. A machine learning model for predicting unscheduled 72 H return visits to the emergency department by patients with abdominal pain. *Diagnostics (Basel)* 2021;12.
- Ng C-J, Yen Z-S, Tsai JC-H, et al. Validation of the Taiwan triage and acuity scale: a new computerised five-level triage system. *Emerg Med J* 2011;28:1026–31.
- Polikar R. Ensemble based systems in decision making. *IEEE Circuits Syst Mag* 2006;6:21–45.
- Guo D-Y, Chen K-H, Chen I-C, et al. The association between emergency department revisit and elderly patients. *J Acute Med* 2020;10:20–6.
- McRae AD, Rowe BH, Usman I, et al. A comparative evaluation of the strengths of association between different emergency department crowding metrics and repeat visits within 72 hours. *CJEM* 2022;24:27–34.
- Ayubi E, Safiri S. Predicting 72-hour emergency department revisits: methodological issues. *Am J Emerg Med* 2018;36:320–1.
- Hutchinson CL, Curtis K, McCloughen A, et al. Clinician perspectives on reasons for, implications and management of unplanned patient returns to the emergency department: a descriptive study. *Int Emerg Nurs* 2022;60:101125.
- de Groot B, Stolwijk F, Warmerdam M, et al. The most commonly used disease severity scores are inappropriate for risk stratification of older emergency department sepsis patients: an observational multi-centre study. *Scand J Trauma Resusc Emerg Med* 2017;25:91.

© 2024 Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY. Published by BMJ. <https://creativecommons.org/licenses/by/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Impact of a pandemic shock on unmet medical needs of middle-aged and older adults in 10 countries

Chao Guo ,^{1,2} Dianqi Yuan,¹ Huameng Tang,¹ Xiyuan Hu,³ Yiyang Lei¹

To cite: Guo C, Yuan D, Tang H, *et al*. Impact of a pandemic shock on unmet medical needs of middle-aged and older adults in 10 countries. *BMJ Health Care Inform* 2024;**31**:e100865. doi:10.1136/bmjhci-2023-100865

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2023-100865>).

Received 01 August 2023
Accepted 09 March 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.

¹Institute of Population Research, Peking University, Beijing, China

²APEC Health Science Academy, Peking University, Beijing, China

³Department of Population Health Sciences, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, Wisconsin, USA

Correspondence to
Dr Chao Guo;
chaoguo@pku.edu.cn

ABSTRACT

Objective The objective is to explore the impact of the pandemic shock on the unmet medical needs of middle-aged and older adults worldwide.

Methods The COVID-19 pandemic starting in 2020 was used as a quasiexperiment. Exposure to the pandemic was defined based on an individual's context within the global pandemic. Data were obtained from the Integrated Values Surveys. A total of 11 932 middle-aged and older adults aged 45 years and above from 10 countries where the surveys conducted two times during 2011 and 2022 were analysed. We used logistic regression models with the difference-in-difference method to estimate the impact of pandemic exposure on unmet medical needs by comparing differences before and after the pandemic across areas with varying degrees of severity.

Results Among the 11 932 middle-aged and older adults, 3647 reported unmet medical needs, with a pooled unmet rate of 30.56% (95% CI: 29.74% to 31.40%). The pandemic significantly increased the risk of unmet medical needs among middle-aged and older adults (OR: 2.33, 95% CI: 1.94 to 2.79). The deleterious effect of the pandemic on unmet medical needs was prevalent among middle-aged adults (2.53, 2.00 to 3.20) and older adults (2.00, 1.48 to 2.69), as well as among men (2.24, 1.74 to 2.90) and women (2.34, 1.82 to 3.03). The results remained robust in a series of sensitivity analyses.

Conclusion These findings suggest that efforts should be made by policymakers and healthcare professionals to balance healthcare resources to adequately address the comprehensive healthcare demands of individuals regarding multiple health issues, taking into account the challenges posed by pandemics.

INTRODUCTION

Although the current leading cause of human disease and death has shifted from infectious and parasitic diseases to chronic non-communicable and degenerative diseases according to the theory of epidemiological transition,¹ as some scholars have pointed out, this shift should not obscure the ongoing threat posed by infectious diseases.² In recent decades, outbreaks of new infectious diseases have occurred in some regions of the world. New infectious diseases are daunting due to their unexpected appearance and rapid spread.³ Severe outbreaks of new infectious

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Some previous studies have reported a decline in medical services utilisation among older patients without coronavirus after COVID-19, but most studies only observe changes in outcomes before and after the pandemic, without differentiating whether these changes are specifically attributed to the effects of the pandemic or reflect general temporal trends over the same period due to other factors.

WHAT THIS STUDY ADDS

⇒ This study contributes to the literature pool by providing trustworthy evidence about the impact of COVID-19 on medical services utilisation among middle-aged and older adults at the global level based on reliable data and methods

⇒ This study demonstrates that pandemic shocks have a negative impact on the fulfilment of medical needs among middle-aged and older adults of different age groups and sexes.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ These findings suggest that policymakers and healthcare professionals, while prioritising pandemic-related measures and response, should not overlook the healthcare needs of individuals, particularly middle-aged and older adults, for other medical services during such outbreaks.

⇒ In addition to the investment of resources for prevention and control directly related to pandemic prevention and control, other medical services for people, especially middle-aged and older adults with high needs and vulnerabilities for disease treatment and rehabilitation, should be further strengthened in strategies to address the emerging infectious diseases transmission for a better health promotion and high-quality development in an ageing world.

diseases often become public health emergencies, even international ones, such as the outbreak of severe acute respiratory syndrome (SARS) in 2003,⁴ the influenza (H1N1) pandemic in 2009,⁵ the Ebola virus in 2014–2016,⁶ the Zika virus in 2016⁷ and COVID-19 recognised by the WHO as a public

health emergency of international concern (PHEIC) in March 2020.⁸

This COVID-19 pandemic is a global public health and safety challenge, and the crisis has brought disruptive effects on health, social, economic, political and even cultural macroscopic areas. A UN framework for the immediate socioeconomic response to COVID-19 states that the COVID-19 pandemic is not just only a health crisis, but is also affecting the social and economic core and that while the extent of the pandemic varies from country to country, it is likely to increase poverty and inequality globally and affect the achievement of the Sustainable Development Goals.⁹ Studies have shown that middle-aged and older adults are undoubtedly vulnerable to this pandemic event due to their higher susceptibility to COVID-19 and the risk of death and secondary disease following infection,^{10–13} people aged 50 and above in some countries were more likely to have medical services postponed¹⁴ and were more likely than younger adults to experience impairment in general,¹⁵ as well as their relatively lower resilience to other life and behavioural effects beyond infection in the pandemic.^{16 17} Are middle-aged and older adults experiencing a shortage of health services due to the global COVID-19 pandemic in the context of the large number of health resources that have to be devoted to prevention and treatment in response to the event of an outbreak? This is an important issue for policymakers and medical services professionals in the demographic context of increasing global ageing, which is crucial for targeting medical services to the middle-aged and older population, promoting the rehabilitation of geriatric diseases and preventing middle-aged and older adults from falling into a vicious cycle of increased disease susceptibility due to unmet medical needs.

Some previous studies have reported a decline in medical services utilisation among middle-aged and older patients without coronavirus after COVID-19. In Europe, a study showed substantial increases in the number of avoidable cancer deaths in England as a result of diagnostic delays due to the COVID-19 pandemic in the UK.¹⁸ In Asia, middle-aged and older Singaporeans' healthcare utilisation and the diagnosis of chronic conditions substantially decreased among non-COVID-19 patients during the first peak period of the COVID-19 outbreak.¹⁹ A study in Japan showed that the total number of hospitalisations and outpatient visits decreased by 27% and 22%,²⁰ respectively, after the first wave of COVID-19. Studies assessing the effect of the COVID-19 pandemic on health services utilisation in China showed that health facility visits were observed significant reduction and the impact still existed 2 years later.^{21 22} A study in Hong Kong, China, showed that the number of missed medical appointments among older adults during COVID-19 increased from 16.5% a year ago to 22.0% after the outbreak.²³ Studies in Latin America showed a similar pattern that a majority (83%) of patient advocacy organisations reported their patients experienced delays in receiving their treatment and care services.²⁴ And the same is true with many multicountry

analyses. A study including six low-income and middle-income countries, such as Zimbabwe, showed that people with disabilities experienced additional difficulties accessing healthcare during the pandemic.²⁵ And a review summarising literature from Africa, Australia and New Zealand, China, Europe, Latin America and the USA showed that individuals with rheumatic diseases during the pandemic faced disruptions in healthcare and medication supply shortages.²⁶ However, many of these studies rely on small local samples or people with certain diseases and do not explore whether the decline in service utilisation is a result of reduced or unmet demand. Additionally, most studies only observe changes in outcomes before and after the pandemic, without differentiating whether these changes are specifically attributed to the effects of the pandemic or reflect general temporal trends over the same period due to other factors.

Given this, this study employed the global pandemic of COVID-19 as a quasiexperiment, combined with international large-scale survey data, to estimate the impact of the pandemic on the medical services utilisation of middle-aged and older adults worldwide. By constructing difference-in-difference (DID) models that considered both exposure time and severity, the study aims to provide robust evidence regarding the imbalances in medical services during public health emergencies. The findings would offer valuable insights for policymakers and healthcare practitioners, enabling them to avoid neglecting and proactively address the utilisation of routine medical resources for middle-aged and older individuals in future pandemics. This facilitates the development of comprehensive and targeted contingency plans to effectively tackle population health challenges arising from global public health emergencies, including infectious disease outbreaks.

METHODS

Data source and participants

The study used the global pandemic COVID-19 starting in 2020 as a quasiexperiment. Data on the global pandemic COVID-19 were obtained from the WHO COVID-19 Detailed Surveillance Data.²⁷ Daily COVID-19 case numbers of each country, area or territory were collected for further analysis. Individual information on medical needs and other demographic statuses was obtained from the Integrated Values Surveys (IVS), which were constructed based on repeated questions from the European Value Study (EVS) from 1981 to 2021 and the World Value Survey (WVS) from 1981 to 2022.^{28 29} EVS and WVS are both renowned, international, large-scale, repeated cross-sectional surveys that are dedicated to gathering extensive information on the social, political, economic, religious and cultural values of individuals across the globe.

While the IVS covered a wide range of surveyed countries, our evaluation of the pandemic's impact is based on comparing differences in unmet medical needs before

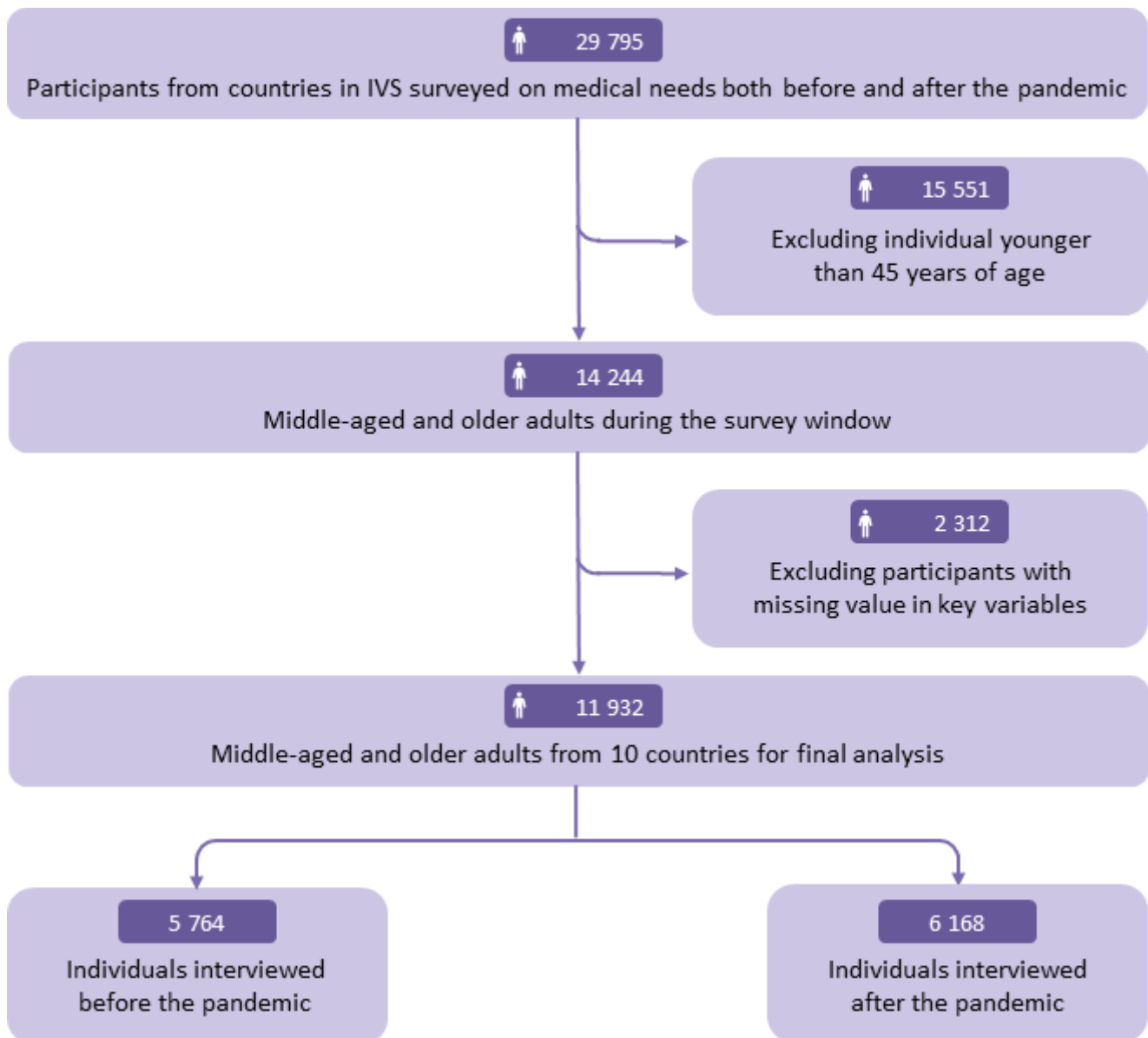


Figure 1 Flowchart of samples.

and after the pandemic across regions with varying COVID-19 severity levels. Therefore, our analysis only included participants residing in countries surveyed both before and after 2020, when the pandemic outbreak occurred. We combined COVID-19 data for countries surveyed during both periods and included participants with available information on medical needs. We focused on participants aged 45 years and above, excluding those with missing data on the outcome measure or any covariate. In the final analysis, a total of 11 932 middle-aged and older adults from 10 countries were included. Each country was surveyed two times between 2011 and 2022. Among the participants, 5764 were interviewed between 2011 and 2014, while 6168 were interviewed between 2020 and 2022 (online supplemental table S1). [Figure 1](#) illustrates the process we followed to derive our analytical sample.

Exposure

Exposure to the pandemic was defined based on an individual's context within the global pandemic, rather than their infection status. It was measured by both exposure time relative to the outbreak and exposure severity. All samples surveyed after 2020 were considered part of the after-pandemic group (exposure group), indicating that they had experienced the pandemic. Samples surveyed before 2020 were classified as the before-pandemic group (reference group). Regarding severity, we tentatively assumed that amidst the global outbreak, residents living in a specific country have a consistent perception of the severity of the outbreak within their country relative to other countries, given the reduced international travel. Consequently, we used country-level average data as an estimation of the pandemic's severity within each country.

The incidence of confirmed cases from 2020 to the survey year in a country or region was used in this study to measure the severity of the pandemic in a country and was standardised to mitigate dimensional influences. Let $C_{cumulative}^i$ represent the cumulative number of COVID-19 cases in the i th country from 2020 to the survey year of this country, and P_{total}^j denotes the total population of the i th country in the survey year j . The incidence of COVID-19 by the survey year can be calculated as a ratio of $C_{cumulative}^i$ and P_{total}^j . Next, let μ and σ denote the mean and SD of the afore-mentioned ratio, respectively. Then, the standardised incidence (SI) of the i th country by the survey year can be obtained using the following formula:

$$SI_i = \left(\frac{C_{cumulative}^i}{P_{total}^j} \times 100\% - \mu \right) / \sigma$$

where a larger value indicates a more severe pandemic. In the analysis, the SI was initially treated as a continuous variable. It was then divided into high and low groups using bisection to create a dichotomous variable, which replaced the continuous SI for sensitivity analysis. Additionally, a sensitivity analysis was conducted using the SI in the survey year as a proxy for the SI by the survey year, substituting the cumulative COVID-19 cases in the survey year for the cumulative cases from 2020 to the survey year.

Outcome

The main outcome of this study was whether participants reported any unmet medical needs during the survey year. The original survey question was, 'What is the frequency you or your family gone without needed medicine or treatment during the last 12 months?'. Respondents could choose one of four options: 'often', 'sometimes', 'rarely' or 'never'.

In this study, to assess the overall situation of unmet medical needs, the outcome event of 'unmet' was defined by combining the three categories of 'often', 'sometimes' and 'rarely'. The category of 'never' indicated the absence of unmet needs, resulting in the creation of a dichotomous variable indicating whether there were any unmet medical needs (yes or no). Additionally, we retained the original four-category approach to measure the severity of unmet medical needs based on the frequency of their occurrence (never, rarely, sometimes or often).

Covariates

According to previous studies, the medical services utilisation of middle-aged and older adults is influenced by various factors, such as age, gender, education, marriage, income and health insurance status.^{30 31} Thus, the following covariates were included in the analysis: age (continuous), sex (male or female), marital status (single or having a partner), religious denomination (do not belong to a denomination, Roman Catholic, Protestant, Orthodox, other Christian, Jew, Muslim, Hindu, Buddhist or other), educational level (lower, medium, upper), employment status (unemployed, full-time employed, part-time employed, self-employed,

retired, housewife, students or other) and income level (comprising a total of ten steps). Self-rated health (very good, good, fair, poor) was also included as a potential confounder since individual health status is a key determinant of the demand for care. Additionally, a control variable indicating international immigrant status (yes or no) was included to account for potential confounding, as inclusive healthcare coverage often relates to national status. Furthermore, the analysis also took into account the nation and year in which participants were surveyed as control variables. More detailed information on these variables is available in online supplemental table S2.

Statistical analysis

The descriptive analysis used frequencies and percentages to describe the demographic and socioeconomic characteristics of the sample, along with the unmet medical need status. The χ^2 test was used to compare the characteristics before and after the pandemic, as well as the prevalence of unmet medical needs among samples with different characteristics.

The inferential analysis used logistic regression models with the DID method. This approach aimed to estimate the impact of pandemic exposure on unmet medical needs by comparing differences before and after the pandemic across areas with varying degrees of severity.

The logistic regression model based on DID estimation was developed as follows:

$$\ln \left(\frac{p}{1-p} \right) = \alpha_0 + \beta_{jk} (Period_j \times Severity_k) + \gamma_j Period_j + \theta_k Severity_k + \delta X_{ijk} + \varepsilon_{ijk}$$

where $p = P(y_{ijk} = 1|x)$ denotes the probability of experiencing unmet medical needs (1=yes, 0=no) for the i th participant interviewed in period j and with severity k . $Period_j$ denotes the survey time (before or after the pandemic) and $Severity_k$ represents the severity of the pandemic measured by SI. X_{ijk} denotes covariates if any. ε_{ijk} represents the random error, and α_0 denotes the constant term. Then, β_{jk} as the interaction coefficient between exposure time and exposure severity is the DID estimate of the pandemic's effect on unmet medical needs for middle-aged and older adults.

Furthermore, we conducted multinomial logistic regressions using the severity of unmet medical needs as the dependent variable. This allowed us to assess the impact of the pandemic across all the range of potential unmet needs.

In addition, subgroup analyses were conducted to examine the heterogeneity across age groups and sexes. The same models were applied for analysis among two age groups: middle-aged adults (aged 45–64 years) and older adults (aged 65 years and above), as well as for both men and women, separately.

To test the robustness of the results, the following sensitivity analyses were conducted in this study. First, models were repeated substituting the SI in the survey year for the SI by the survey year as a measure of the pandemic severity. Next, models were reanalysed by replacing the

Table 1 Characteristics of samples

Characteristics	Pooled, n (%)	By period		P value
		Before pandemic, n (%)	After pandemic, n (%)	
Total sample	11 932 (100.00)	5764 (48.31)	6168 (51.69)	
Sex				0.468
Female	6656 (55.78)	3235 (56.12)	3421 (55.46)	
Male	5276 (44.22)	2529 (43.88)	2747 (44.54)	
Age group				0.001
Middle-aged	8174 (68.50)	4032 (69.95)	4142 (67.15)	
Older	3758 (31.50)	1732 (30.05)	2026 (32.85)	
Marital status				0.003
Single	3725 (31.22)	1725 (29.93)	2000 (32.43)	
Having a partner	8207 (68.78)	4039 (70.07)	4168 (67.57)	
Religious denomination				<0.0001
Do not belong to a denomination	2925 (24.51)	1300 (22.55)	1625 (26.35)	
Roman Catholic	1398 (11.72)	642 (11.14)	756 (12.26)	
Protestant	1072 (8.98)	660 (11.45)	412 (6.68)	
Orthodox	2186 (18.32)	1116 (19.36)	1070 (17.35)	
Other Christian	298 (2.50)	22 (0.38)	276 (4.47)	
Jew	29 (0.24)	14 (0.24)	15 (0.24)	
Muslim	2745 (23.01)	1423 (24.69)	1322 (21.43)	
Hindu	103 (0.86)	47 (0.82)	56 (0.91)	
Buddhist	602 (5.05)	279 (4.84)	323 (5.24)	
Other				
Educational level				<0.0001
Lower	3262 (27.34)	1570 (27.24)	1692 (27.43)	
Medium	4823 (40.42)	2891 (50.16)	1932 (31.32)	
Upper	3847 (32.24)	1303 (22.61)	2544 (41.25)	
Employment status				<0.0001
Full time	3577 (29.98)	1475 (25.59)	2102 (34.08)	
Part-time	1370 (11.48)	634 (11.00)	736 (11.93)	
Self-employed	899 (7.53)	488 (8.47)	411 (6.66)	
Retired	3772 (31.61)	1890 (32.79)	1882 (30.51)	
Housewife	1281 (10.74)	741 (12.86)	540 (8.75)	
Students	30 (0.25)	24 (0.42)	6 (0.10)	
Unemployed	813 (6.81)	410 (7.11)	403 (6.53)	
Other	190 (1.59)	102 (1.77)	88 (1.43)	
Income level				<0.0001
Lower step	1055 (8.84)	458 (7.95)	597 (9.68)	
2nd step	1015 (8.51)	586 (10.17)	429 (6.96)	
3rd step	1472 (12.34)	801 (13.9)	671 (10.88)	
4th step	1608 (13.48)	823 (14.28)	785 (12.73)	
5th step	2734 (22.91)	1151 (19.97)	1583 (25.66)	
6th step	1596 (13.38)	798 (13.84)	798 (12.94)	
7th step	1124 (9.42)	577 (10.01)	547 (8.87)	
8th step	640 (5.36)	313 (5.43)	327 (5.30)	
9th step	271 (2.27)	123 (2.13)	148 (2.40)	
10th step	417 (3.49)	134 (2.32)	283 (4.59)	

Continued

Table 1 Continued

Characteristics	Pooled, n (%)	By period		P value
		Before pandemic, n (%)	After pandemic, n (%)	
International migration				0.787
No	10 745 (90.05)	5195 (90.13)	5550 (89.98)	
Yes	1187 (9.95)	569 (9.87)	618 (10.02)	
Self-rated health				<0.0001
Very good	1866 (15.64)	974 (16.90)	892 (14.46)	
Good	5038 (42.22)	2494 (43.27)	2544 (41.25)	
Fair	3737 (31.32)	1664 (28.87)	2073 (33.61)	
Poor	1291 (10.82)	632 (10.96)	659 (10.68)	
Severity of the pandemic				<0.0001
Lower	4499 (37.71)	1958 (33.97)	2541 (41.20)	
Upper	7433 (62.29)	3806 (66.03)	3627 (58.80)	

continuous variable SI with a binary classification. Then, models were reanalysed through multilevel logistic regression, incorporating additional adjustments at the national level including economic indicators and relevant information on health systems.

Crude ORs and 95% CIs were initially calculated for models without control variables and then the estimates were adjusted by including control variables. In this study, two-sided p values less than 0.05 were considered statistically significant. STATA V.17 (STATA Corp, College Station, Texas, USA) software was used for the statistical analysis of all data.

RESULTS

Sample characteristics

A total of 11 932 middle-aged and older adults aged 45 years and above from 10 countries were included in this study. Among them, 5764 (48.31%) were interviewed before the pandemic, while 6168 (51.69%) were interviewed after the pandemic. In terms of pandemic severity, 4499 (37.71%) participants resided in areas with lower severity, while 7 433 (62.29%) were in areas with upper severity. Regarding demographic characteristics, 55.78%

of the participants were women and 31.50% were older adults. The majority of participants (68.78%) had a partner and 24.51% reported not belonging to a religious denomination. 72.66% had a medium or upper-range education level, 31.61% were retired and 9.95% were international migrants. Only a few of the participants reported poor health (10.82%). [Table 1](#) presents more detailed information on sample characteristics by period.

Prevalence of unmet medical needs among middle-aged and older adults

Among all the participants, a total of 3647 reported any unmet medical needs, with a pooled unmet rate of 30.56% (95% CI: 29.74 to 31.40). Overall, the prevalence of unmet medical needs among middle-aged and older adults in the 10 countries after the pandemic (27.25, 26.15 to 28.38) was significantly lower than that before the pandemic (34.11, 32.88 to 35.35) ($p < 0.0001$). However, a significantly higher prevalence was found in areas with an upper pandemic severity (32.45, 31.39 to 33.53) compared with areas with a lower severity (27.45, 26.15 to 28.78) ($p < 0.0001$). [Table 2](#) and [figure 2](#) present the prevalence of unmet medical needs by period and severity of the pandemic. For more detailed information

Table 2 Prevalence of unmet medical needs, by period and severity of the pandemic

	Pooled			By period						P value
	Sample, n	Unmet, n	Prevalence (% 95% CI)	Before pandemic			After pandemic			
				Sample, n	Unmet, n	Prevalence (% 95% CI)	Sample, n	Unmet, n	Prevalence (% 95% CI)	
Pooled	11 932	3647	30.56 (29.74 to 31.40)	5764	1966	34.11 (32.88 to 35.35)	6168	1681	27.25 (26.15 to 28.38)	<0.0001
By Severity										
Lower	4499	1235	27.45 (26.15 to 28.78)	1958	605	30.90 (28.86 to 33.00)	2541	630	24.79 (23.12 to 26.52)	<0.0001
Upper	7433	2412	32.45 (31.39 to 33.53)	3806	1361	35.76 (34.23 to 37.31)	3627	1051	28.98 (27.5 to 30.48)	<0.0001
P value			<0.0001			<0.0001			<0.0001	

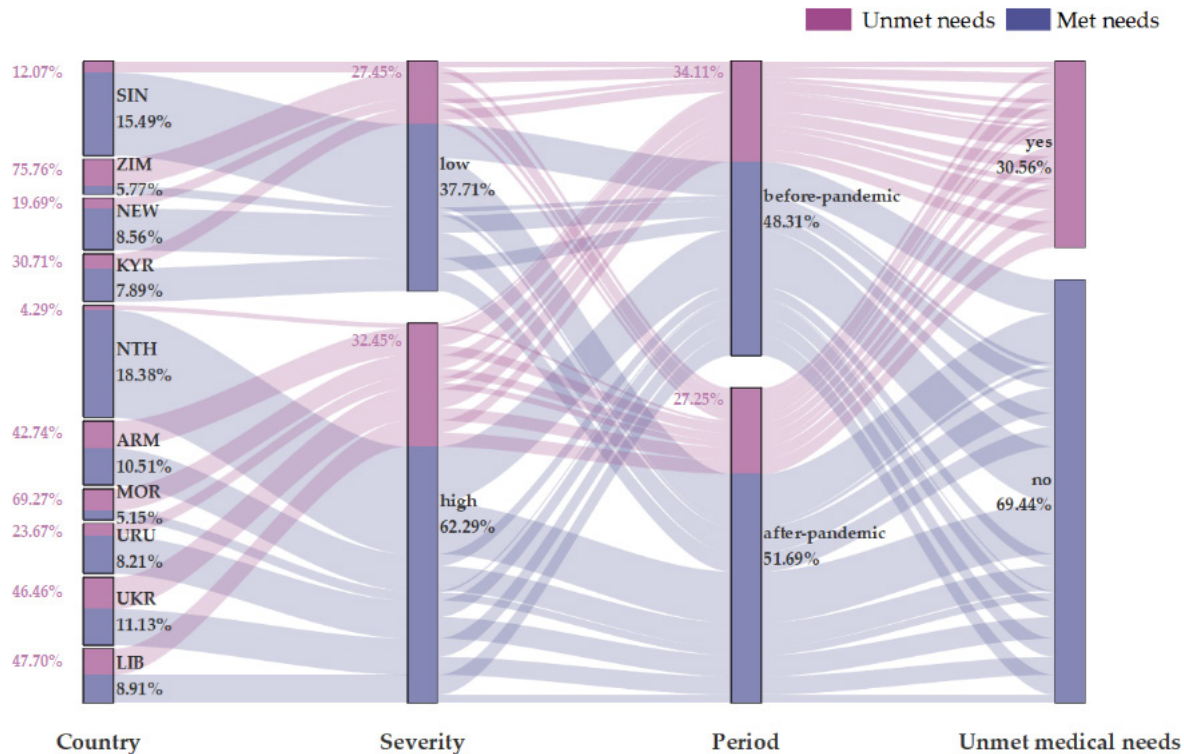


Figure 2 Prevalence of unmet medical needs by country, pandemic severity and period. The black numbers are the proportion of participants with a certain characteristic to the total participants. The pink numbers represent the prevalence of unmet needs among participants with certain characteristics. The abbreviations in the figure adhere to the ISO 3166-1 alpha-3 code standard, which assigns three-letter alphabetic codes to countries, including ARM for Armenia, KGZ for Kyrgyzstan, LBY for Libya, MAR for Morocco, NLD for the Netherlands, NZL for New Zealand, SGP for Singapore, UKR for Ukraine, URY for Uruguay, and ZWE for Zimbabwe.

on the prevalence by other demographic and socioeconomic characteristics, please refer to online supplemental table S3.

The impact of the pandemic on unmet medical needs among middle-aged and older adults

After estimating the change in unmet medical needs related to the pandemic beyond the background trends by doing a DID analysis (figure 3A), we found that the pandemic significantly increased the risk of any unmet medical needs among middle-aged and older adults (OR: 1.17, 95% CI: 1.07 to 1.27). This effect was partially increased and remained significant after controlling for multiple covariates (2.33, 1.94 to 2.79).

In the analysis of heterogeneity based on age (figure 3B) and sex (figure 3C), we found that the deleterious effect of the pandemic on unmet medical needs was prevalent among middle-aged adults (2.53, 2.00 to 3.20) and older adults (2.00, 1.48 to 2.69), as well as among men (2.24, 1.74 to 2.90) and women (2.34, 1.82 to 3.03), without heterogeneity in age groups (P for interaction=0.913) and sexes (P for interaction=0.615).

The results from the multinomial models indicated that the impact of the pandemic on the increased risk of unmet medical needs among middle-aged and older adults intensified with higher frequencies of occurrences of unmet medical needs. Relative to never reporting any

unmet medical needs, the OR and 95% CI for reporting unmet medical needs rarely, sometimes and often were 1.65 (1.34 to 2.02), 3.75 (2.69 to 5.23) and 4.88 (2.67 to 8.91), respectively. This trend was observed across the samples of middle-aged adults, older adults, men and women (table 3).

Sensitivity analysis

The series of sensitivity analyses we conducted indicated a certain level of robustness in the study findings. First, the effects of the pandemic were still observed in the overall participants (2.13, 1.77 to 2.57) as well as the subpopulations by age groups and sexes in models that repeated substituting the SI in the survey year for the SI by the survey year as a measure of the pandemic severity (online supplemental table S4). Second, models by replacing the continuous variable SI with a binary classification also yielded similar results. The pandemic also exhibited a significant increase in the risk of unmet medical needs across all participants (2.85, 2.25 to 3.62), and this effect remained significant when analysing middle-aged adults, older adults, men and women separately (online supplemental table S5). Third, the effect observed among the participants remained statistically significant (2.77, 1.66 to 4.61) in the multilevel models. Similar trends were also identified when examining subpopulations based on age groups and sexes (online supplemental table S6).

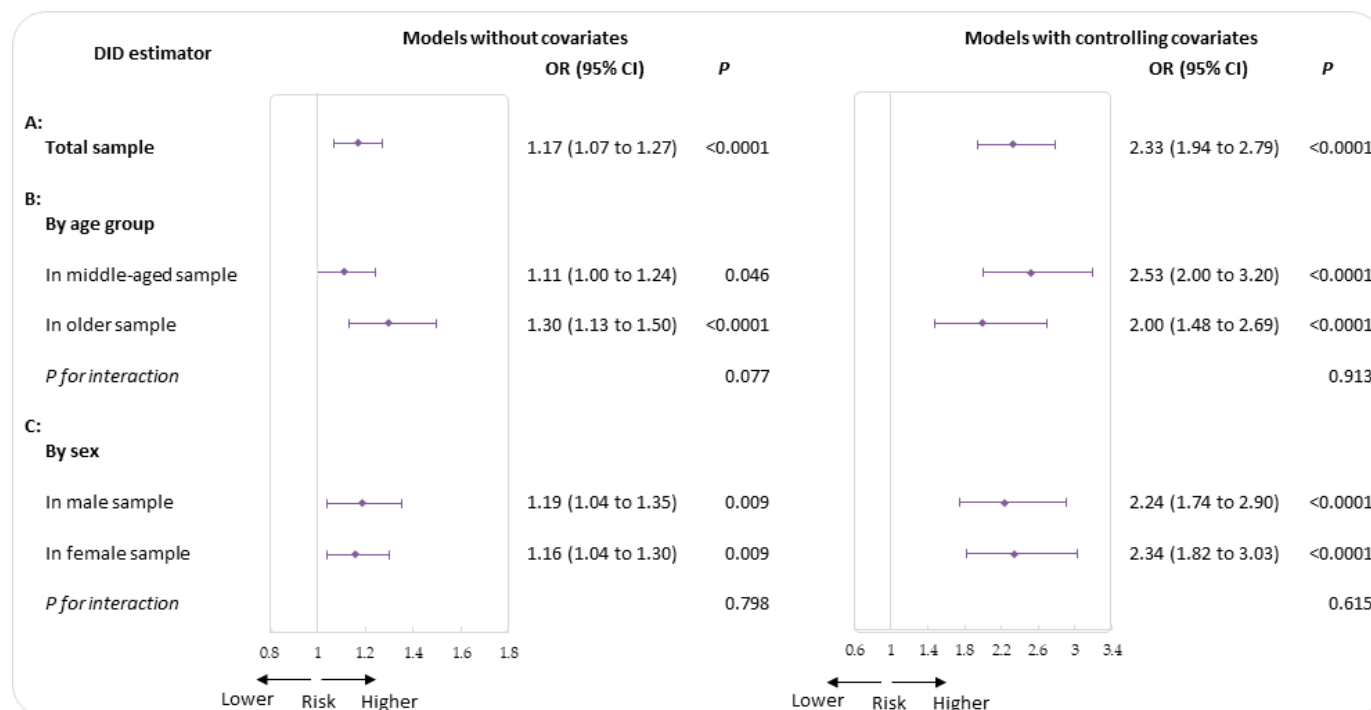


Figure 3 The impact of the pandemic on unmet medical needs among middle-aged and older adults. (A) Total sample, (B) subsamples by age group and (C) subsamples by sex. In models with controlling covariates, exact age, sex, marital status, religious denomination, educational level, employment status, income level, international migration, self-rated health, nation and survey year were controlled in the total sample; control variables in the age-specific models were the same as above; control variables in the sex-specific models were the same as the above model except for sex. DID, difference-in-difference.

DISCUSSION

This study conducted a comprehensive and robust analysis to investigate the influence of the pandemic of COVID-19 on the medical services utilisation of middle-aged and older adults in multiple countries. The results indicated that the pandemic shock has significantly increased the risk of unmet medical needs of middle-aged and older adults, regardless of age or sex. These findings not only support the results of previous studies but also provide further clarification regarding the role of the pandemic in this particular context. As suggested by WHO in implications of the COVID-19 pandemic for patient safety, severe

disruptions in all major health areas have led to delays in the diagnosis and treatment of diseases, especially in countries experiencing fragility, social and economic instability, conflict and violence.³²

The potential mechanisms underlying the negative effect of the pandemic on the medical utilisation of middle-aged and older adults may be wide-ranging. On the one hand, the outbreak and rapid spread of COVID-19 inevitably crowded out the limited resources of medical services, resulting in a diversion of substantial health resources including human and material resources towards COVID-19 prevention, virus detection

Table 3 The impact of the pandemic on various severity of unmet medical needs among middle-aged and older adults

DID estimators	Rarely		Sometimes		Often	
	OR (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value
Total sample	1.65 (1.34 to 2.02)	<0.0001	3.75 (2.69 to 5.23)	<0.0001	4.88 (2.67 to 8.91)	<0.0001
By age group						
Middle-aged sample	1.74 (1.35 to 2.24)	<0.0001	4.08 (2.70 to 6.16)	<0.0001	6.00 (3.33 to 10.80)	<0.0001
Older sample	1.50 (1.04 to 2.16)	0.031	3.00 (1.72 to 5.25)	<0.0001	3.51 (0.79 to 15.61)	0.099
P for interaction		0.903		0.793		0.618
By sex						
Male	1.53 (1.15 to 2.05)	0.004	3.79 (2.30 to 6.26)	<0.0001	5.46 (2.54 to 11.71)	<0.0001
Female	1.66 (1.24 to 2.21)	0.001	3.63 (2.29 to 5.74)	<0.0001	5.40 (2.11 to 13.85)	<0.0001
P for interaction		0.517		0.316		0.074

OR, OR after controlling covariates.

and patient care. As a consequence, there was a significant reduction in resources available for the management and care of other diseases.^{33–35} At the same time, general medical resources have been further reduced by the suspension of hospitals to deal with the potential risk of nosocomial infections, the inability of medical services workers to work due to infections and the emergence of strikes by medical services workers in some countries or regions.^{36–38} These have objectively reduced the supply of geriatric care in some regions where healthcare systems have reached the point of exhaustion,³⁹ especially in the severe early days of COVID-19.

On the other hand, in response to a sudden outbreak of a new infectious disease, countries and regions have been experimenting and changing their coping strategies, such as some emergency measures such as community closure, traffic control and social distancing to prioritise the response to the spread of the pandemic. Some of the countries analysed in our study also adopted such strategies such as the stay-at-home orders in Singapore,⁴⁰ which may not only lead to active or passive changes in daily life behaviour and social interaction but also undoubtedly reduces the accessibility of medical services resources, especially in cross-regional medical treatment.⁴¹ This is particularly evident among middle-aged and older adults, who may put on hold non-acute or urgent medical needs. In contrast, the impact of the pandemic and social distance can have a significant negative impact on the physical and psychological well-being of older adults.⁴² For example, studies have shown that the pandemic may increase anxiety, depression, poor sleep quality, nutritional deficiencies and physical inactivity among older adults,^{43–45} which in turn further amplifies the demand for medical services among the older population, leading to a greater gap between demand and utilisation.

After SARS, the last major pandemic with a significant impact on the population,⁴⁶ COVID-19 is a wake-up call for humanity at the beginning of entering the 20s of the 21st century, when governments, industries and families are once again aware of the challenges of the emerging disease in this new era, in addition to the traditional disease threats. However, just as we should not overlook emerging infectious diseases due to the increasing prevalence of chronic diseases during epidemiological transitions, we should also not neglect the healthcare needs for chronic and other conventional diseases during a pandemic. With the WHO declaring that the COVID-19 pandemic is no longer a PHEIC, governments worldwide are reflecting on lessons learnt and developing preparedness plans for future pandemics. The increased medical utilisation gaps, particularly among middle-aged and older individuals resulting from the COVID-19 pandemic as discovered in this study, should undoubtedly be given full consideration by policymakers and clinical healthcare professionals. Declines in essential health service utilisation could even result in more deaths than the disease

outbreak itself.⁴⁷ Measures should be taken to reduce the neglect of healthcare needs for other diseases during a pandemic and formulate effective strategies to balance the allocation of healthcare resources.

It is clear that, our research is based on countries with varying levels of socioeconomic development and healthcare resources, and overall, consistent with previous studies,^{48–49} higher levels of socioeconomic status and healthcare resources at the country level were found to be associated with a lower risk of unmet medical needs in the sample included in this study (see online supplemental table S6). However, even after controlling for these country-level covariates, the impact of the pandemic shock on unmet medical needs remains significant. While this is an ‘averaged’ outcome, such estimates provide support and basis for advocating international attention to ensuring basic healthcare service provision from a more macroscopic global perspective during public health emergencies. Indeed, the WHO has released a position paper calling on countries and the international community to build resilient health systems by integrating universal health cover and health security efforts during COVID-19 pandemic and beyond in 2021.⁵⁰ In the postpandemic era, the WHO also needs to assume greater international responsibilities in this field and rebuild trust among the people to prepare for the next pandemic.⁵¹ The results of our study once again highlights the need for countries all over the world to take every opportunity to build resilient health systems and all-hazards emergency risk management based on a strong primary healthcare foundation and rebuild the health systems sustainably, more equitably and closer to communities.⁵²

There are also some shortcomings in this study. First, several potential confounders, such as the objective medical conditions of participants that were not controlled because of data accessibility, may have had some impact on the results. Second, although the cumulative confirmed infected cases were obtained from the WHO, they were based on the integration of official reports from various countries or regions and the different criteria in each region may produce some bias. Third, the results should be interpreted with caution given that the exposure period groupings in our analysis are in years and the results reflect the average long-term effect over that period. Furthermore, as our data were aggregated at the country level, all individuals within a country were grouped together. This might introduce bias stemming from regional variations within each country. The limited number of countries also poses a potential threat to the external validity when making global generalisations of our research findings and presents challenges in deriving policy implications and recommendations for specific nations. In addition, self-rated health might have a bidirectional relationship with our outcome variable. However, we opted to retain it as a covariate due to the lack of a more appropriate exogenous health condition variable. Additionally, it is unfortunate that we lack further relevant variables pertaining to healthcare access for migrant

populations in each country. Consequently, we have solely considered migrant status as a regression factor. Moreover, we did not distinguish between the specific types of medical needs of the participants because there was no such information in the database. Nevertheless, to the best of our knowledge, this study contributes to the literature pool by providing trustworthy evidence about the impact of COVID-19 on medical services utilisation among middle-aged and older adults at the global level based on reliable data and methods for the first time.

The findings of this study on the global pandemic on the medical services utilisation of middle-aged and older adults in multiple countries emphasise the importance of balancing medical resources in the response to outbreaks. In addition to the investment of resources for prevention and control directly related to pandemic prevention and control, other medical services for people, especially middle-aged and older adults with high needs and vulnerabilities for disease treatment and rehabilitation, should be further strengthened in strategies to address the emerging infectious diseases transmission for a better health promotion and high-quality population development in an ageing world.

Contributors CG formulated the research questions, designed the study and wrote the first draft of the manuscript. CG, DY and HT analysed the data. CG, DY, HT, XH and YL revised the manuscript together. All the authors had access to the data and were responsible for the decision to submit the manuscript for publication. CG was responsible for the overall content as the guarantor.

Funding This study was funded by the National Natural Science Foundation of China (grant number 82103955) and the Clinical Medicine Plus X—Young Scholars Project, Peking University, the Fundamental Research Funds for the Central Universities (grant number 7100604313). The funders had no role in the study design, data collection, data analysis, the decision to publish or the preparation of the manuscript.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants. The Integrated Values Surveys, including European Value Study and World Value Survey, was reviewed and approved by the ethical review board in each country. The detailed approval numbers can be obtained by contacting the principal investigator of each national team through the e-mail address wvsa.secretariat@gmail.com. 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 in a public, open access repository. This study is based on publicly available datasets, and the data were released to the researchers without access to any personal information from the website: <https://www.worldvaluessurvey.org/WVSEVTrend.jsp>.

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

Chao Guo <http://orcid.org/0000-0003-4343-4848>

REFERENCES

- 1 Adogu P, Ubajaka C, Emelumadu O, *et al*. Epidemiologic transition of diseases and health-related events in developing countries: a review. *Am J Med Med Scien* 2015;5:150–7.
- 2 McKeown RE. The epidemiologic transition: changing patterns of mortality and population dynamics. *Am J Lifestyle Med* 2009;3:19S–26S.
- 3 World Health Organization. Severe acute respiratory syndrome (SARS): status of the outbreak and lessons for the immediate future. Geneva World Health Organization; 2003.
- 4 Zhong NS, Zheng BJ, Li YM, *et al*. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, people's Republic of China, in February, 2003. *Lancet* 2003;362:1353–8.
- 5 Gerberding JL. Influenza in 2009: new solutions, same old problems. *JAMA* 2009;302:1907–8.
- 6 Green A. West Africa struggles to contain Ebola outbreak. *Lancet* 2014;383:1196.
- 7 Lucey DR, Gostin LO. The emerging Zika pandemic: enhancing preparedness. *JAMA* 2016;315:865–6.
- 8 Phelan AL, Katz R, Gostin LO. The novel Coronavirus originating in Wuhan, China: challenges for global health governance. *JAMA* 2020;323:709–10.
- 9 UNDP. COVID-19 socio-economic impact. Available: <https://www.undp.org/coronavirus/socio-economic-impact-covid-19> [Accessed 10 Jan 2024].
- 10 World Health Organization. World health statistics 2022: monitoring health for the SDGs, sustainable development goals. Geneva World Health Organization; 2022.
- 11 Marcon G, Tettamanti M, Capacci G, *et al*. COVID-19 mortality in Lombardy: the vulnerability of the oldest old and the resilience of male centenarians. *Ageing (Albany NY)* 2020;12:15186–95.
- 12 Damian AJ, Gonzalez M, Oo M, *et al*. A national study of community health centers' readiness to address COVID-19. *J Am Board Fam Med* 2021;34:S85–94.
- 13 Curtis AF, Schmiedeler A, Musich M, *et al*. COVID-19-related anxiety and cognition in middle-aged and older adults: examining sex as a moderator. *Psychol Rep* 2023;126:1260–83.
- 14 Smolić Š, Čipin I, Medimurec P. Access to healthcare for people aged 50+ in Europe during the COVID-19 outbreak. *Eur J Ageing* 2022;19:793–809.
- 15 Salthouse TA. When does age-related cognitive decline begin? *Neurobiol Aging* 2009;30:507–14.
- 16 Gietel-Basten S, Matus K, Mori R. COVID-19 as a trigger for innovation in policy action for older persons? Evidence from Asia. *Policy and Society* 2022;41:168–86.
- 17 Tur-Sinai A, Bentur N, Lamura G. Perceived deterioration in health status among older adults in Europe and Israel following the first wave of the COVID-19 pandemic. *Eur J Ageing* 2022;19:1243–50.
- 18 Maringe C, Spicer J, Morris M, *et al*. The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study. *Lancet Oncol* 2020;21:1023–34.
- 19 Ahn S, Kim S, Koh K. Associations of the COVID-19 pandemic with older individuals' healthcare utilization and self-reported health status: a longitudinal analysis from Singapore. *BMC Health Serv Res* 2022;22:66.
- 20 Yamaguchi S, Okada A, Sunaga S, *et al*. Impact of COVID-19 pandemic on healthcare service use for non-COVID-19 patients in Japan: retrospective cohort study. *BMJ Open* 2022;12:e060390.
- 21 Liu J, Zhai X, Yan W, *et al*. Long-term impact of the COVID-19 pandemic on health services utilization in China: a nationwide longitudinal study. *Glob Transit* 2023;5:21–8.
- 22 Xiao H, Dai X, Wagenaar BH, *et al*. The impact of the COVID-19 pandemic on health services utilization in China: time-series analyses for 2016–2020. *Lancet Reg Health West Pac* 2021;9:100122.
- 23 Wong SYS, Zhang D, Sit RWS, *et al*. Impact of COVID-19 on loneliness, mental health, and health service utilisation: a prospective cohort study of older adults with Multimorbidity in primary care. *Br J Gen Pract* 2020;70:e817–24.
- 24 Kruse MH, Durstine A, Evans DP. Effect of COVID-19 on patient access to health services for noncommunicable diseases in Latin America: a perspective from patient advocacy organizations. *Int J Equity Health* 2022;21:45.
- 25 Hunt X, Hameed S, Tetali S, *et al*. Impacts of the COVID-19 pandemic on access to healthcare among people with disabilities:

- evidence from six low- and middle-income countries. *Int J Equity Health* 2023;22:172.
- 26 Hsieh E, Dey D, Grainger R, *et al*. Global perspective on the impact of the COVID-19 pandemic on rheumatology and health equity. *Arthritis Care Res (Hoboken)* 2024;76:22–31.
 - 27 World Health Organization. WHO Coronavirus (COVID-19). Available: <https://covid19.who.int/data> [Accessed 16 May 2023].
 - 28 EVS. EVS Trend File 1981-2017. Data File Version 3.0.0. Cologne ZA 7503: GESIS Data Archive 2022, 2022.
 - 29 Haerpfner C, Inglehart R, Moreno A, eds. World Values Survey Trend File (1981-2022) Cross-National Data-Set. Data File Version 3.0.0. Madrid, Spain & Vienna, Austria: JD Systems Institute & WWSA Secretariat, 2022.
 - 30 Guo C, Du W, Hu C, *et al*. Prevalence and factors associated with healthcare service use among Chinese elderly with disabilities. *J Public Health (Oxf)* 2016;38:e345–51.
 - 31 Liu TY, Qiu DC, Chen T. Effects of social participation by middle-aged and elderly residents on the utilization of medical services: evidence from China. *Front Public Health* 2022;10:824514.
 - 32 World Health Organization. Implications of the COVID-19 pandemic for patient safety. Available: <https://www.un.org/zh/node/96680> [Accessed 02 Jan 2024].
 - 33 McWilliams JM, Russo A, Mehrotra A. Implications of early medical services spending reductions for expected spending as the COVID-19 pandemic evolves. *JAMA Intern Med* 2021;181:1118–20.
 - 34 Nab M, van Vehmendahl R, Somers I, *et al*. Delayed emergency healthcare seeking behaviour by Dutch emergency department visitors during the first COVID-19 wave: a mixed methods retrospective observational study. *BMC Emerg Med* 2021;21:56.
 - 35 Olivera MJ. Opportunity cost and COVID-19: a perspective from health economics. *Med J Islam Repub Iran* 2020;34:177.
 - 36 Du Q, Zhang D, Hu W, *et al*. Nosocomial infection of COVID-19: a new challenge for healthcare professionals (review). *Int J Mol Med* 2021;47:31.
 - 37 Cox-Ganser JM, Henneberger PK, Weissman DN, *et al*. COVID-19 test positivity by occupation using the Delphi US COVID-19 trends and impact survey, September–November 2020. *Am J Ind Med* 2022;65:721–30.
 - 38 Titov N, Staples L, Kayrouz R, *et al*. Rapid report: early demand, profiles and concerns of mental health users during the Coronavirus (COVID-19) pandemic. *Internet Interv* 2020;21:100327.
 - 39 Rashedi V, Farvahari A, Sabermahani M, *et al*. Integrated geriatric health care services at the level of primary health care: a comparison study during COVID-19 pandemic. *J Public Health (Berl)* 2023.
 - 40 Hakim AJ, Victory KR, Chevinsky JR, *et al*. Mitigation policies, community mobility, and COVID-19 case counts in Australia, Japan, Hong Kong, and Singapore. *Public Health* 2021;194:238–44.
 - 41 Belchior CA, Gomes Y. Liquidity constraints, cash transfers and the demand for medical services in the Covid-19 pandemic. *Health Econ* 2022;31:2369–80.
 - 42 Sepúlveda-Loyola W, Rodríguez-Sánchez I, Pérez-Rodríguez P, *et al*. Impact of social isolation due to COVID-19 on health in older people: mental and physical effects and recommendations. *J Nutr Health Aging* 2020;24:938–47.
 - 43 Stephenson E, Tu K, Ji C, *et al*. Effects of COVID-19 pandemic on anxiety and depression in primary care: a cohort study in Ontario, Canada. *Ann Fam Med* 2022;20:2911.
 - 44 Lee ATC, Mo FYM, Lam LCW. Higher psychogeriatric admissions in COVID-19 than in severe acute respiratory syndrome. *Int J Geriatr Psychiatry* 2020;35:1449–57.
 - 45 Nguyen PH, Kachwaha S, Pant A, *et al*. COVID-19 disrupted provision and utilization of health and nutrition services in Uttar Pradesh, India: insights from service providers, household phone surveys, and administrative data. *J Nutr* 2021;151:2305–16.
 - 46 Guo C, Zheng X. Prenatal exposure to the SARS epidemic emergency and risk of cognitive impairment in toddlers. *Science Bulletin* 2021;66:2153–6.
 - 47 Ballard M, Olsen HE, Millier A, *et al*. Continuity of community-based healthcare provision during COVID-19: a multicountry interrupted time series analysis. *BMJ Open* 2022;12:e052407.
 - 48 McKee M, Suhrcke M, Nolte E, *et al*. Health systems, health, and wealth: a European perspective. *Lancet* 2009;373:349–51.
 - 49 Peters DH, Garg A, Bloom G, *et al*. Poverty and access to health care in developing countries. *Ann N Y Acad Sci* 2008;1136:161–71.
 - 50 World Health Organization. WHO position paper: building health systems resilience for universal health coverage and health security during the COVID-19 pandemic and beyond. Geneva World Health Organization; 2021. Available: <http://www.jstor.org/stable/resrep44393> [accessed 10 Jan 2024]
 - 51 Guo C, Hu X, Yuan D, *et al*. The effect of COVID-19 on public confidence in the World Health Organization: a natural experiment among 40 countries. *Global Health* 2022;18:77.
 - 52 Gomes DJ, Hazim C, Safstrom J, *et al*. Infection prevention and control initiatives to prevent healthcare-associated transmission of SARS-CoV-2, East Africa. *Emerg Infect Dis* 2022;28:S255–61.

© 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.

Personalised prediction of maintenance dialysis initiation in patients with chronic kidney disease stages 3–5: a multicentre study using the machine learning approach

Anh Trung Hoang ¹, Phung-Anh Nguyen ^{2,3,4}, Thanh Phuc Phan,^{5,6}
Gia Tuyen Do,^{1,7} Huu Dung Nguyen,¹ I-Jen Chiu,^{8,9,10} Chu-Lin Chou,^{9,10,11,12}
Yu-Chen Ko,¹³ Tzu-Hao Chang,¹⁴ Chih-Wei Huang,^{14,15} Usman Iqbal ^{16,17},
Yung-Ho Hsu,^{8,9,10,11} Mai-Szu Wu,^{8,9,10} Chia-Te Liao^{8,9,10}

To cite: Hoang AT, Nguyen P-A, Phan TP, *et al.* Personalised prediction of maintenance dialysis initiation in patients with chronic kidney disease stages 3–5: a multicentre study using the machine learning approach. *BMJ Health Care Inform* 2024;**31**:e100893. doi:10.1136/bmjhci-2023-100893

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2023-100893>).

ATH and P-AN contributed equally., M-SW and C-TL contributed equally.

Received 06 September 2023
Accepted 16 April 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Chia-Te Liao;
ctliao19386@tmu.edu.tw

Professor Mai-Szu Wu;
maiszuwu@tmu.edu.tw

ABSTRACT

Background Optimal timing for initiating maintenance dialysis in patients with chronic kidney disease (CKD) stages 3–5 is challenging. This study aimed to develop and validate a machine learning (ML) model for early personalised prediction of maintenance dialysis initiation within 1-year and 3-year timeframes among patients with CKD stages 3–5.

Methods Retrospective electronic health record data from the Taipei Medical University clinical research database were used. Newly diagnosed patients with CKD stages 3–5 between 2008 and 2017 were identified. The observation period spanned from the diagnosis of CKD stages 3–5 until the maintenance dialysis initiation or a maximum follow-up of 3 years. Predictive models were developed using patient demographics, comorbidities, laboratory data and medications. The dataset was divided into training and testing sets to ensure robust model performance. Model evaluation metrics, including area under the curve (AUC), sensitivity, specificity, positive predictive value, negative predictive value and F1 score, were employed.

Results A total of 6123 and 5279 patients were included for 1 year and 3 years of the model development. The artificial neural network demonstrated better performance in predicting maintenance dialysis initiation within 1 year and 3 years, with AUC values of 0.96 and 0.92, respectively. Important features such as baseline estimated glomerular filtration rate and albuminuria significantly contributed to the predictive model.

Conclusion This study demonstrates the efficacy of an ML approach in developing a highly predictive model for estimating the timing of maintenance dialysis initiation in patients with CKD stages 3–5. These findings have important implications for personalised treatment strategies, enabling improved clinical decision-making and potentially enhancing patient outcomes.

INTRODUCTION

Chronic kidney disease (CKD) and end-stage renal disease (ESRD) are significant global

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Early prediction of dialysis initiation in patients with chronic kidney disease (CKD) is invaluable for tailoring personalised treatment plans. Despite several prediction models that have been developed, they almost lack sufficient accuracy and fail to include several crucial factors related to CKD progression.

WHAT THIS STUDY ADDS

⇒ Developing a machine learning predictive model that incorporates comprehensive clinical data and standardises the selection of the index date may lead to even more accurate predictions. So far, this study has involved the largest patient enrolment and the broadest range of clinical parameters.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ These findings enable highly accurate estimation of the timing for initiating maintenance dialysis in patients with CKD stages 3–5. As a result, they have significant implications for personalised treatment strategies, facilitating improved clinical decision-making and potentially enhancing patient outcomes.

health problems that burden healthcare systems worldwide. In Taiwan, the overall prevalence of CKD was 8.2%, and the incidence of treated ESRD was 529 per million population, according to the US Renal Data System annual report published in 2021.¹ International guidelines recommend referring patients with CKD to nephrology for pre-ESRD care upon reaching an advanced stage to improve the quality of care and reduce costs.^{2,3} One critical component of pre-ESRD care is counselling patients on choosing kidney replacement therapy (KRT) following shared decision-making. It may involve

preparing vascular access for haemodialysis (HD) at least 6 months before HD initiation, placing a peritoneal dialysis (PD) catheter at least 2 weeks before PD initiation, or identifying suitable donors for pre-emptive kidney transplantation before dialysis is required to replace failing kidney function. Therefore, personalising the timing of referral for maintenance dialysis initiation is essential for each patient. However, early personalised estimation of the optimal timing for maintenance dialysis initiation presents a significant challenge for patients with CKD. This decision relies not only on the glomerular filtration rate (GFR) level but also on symptoms of uraemia syndrome and the ability to manage complications such as electrolyte imbalance, acid–base disturbances and fluid overload through medical treatment.^{3–5}

Numerous traditional and artificial intelligence (AI) models have been developed and evaluated to estimate the duration until the initiation of KRT among patients with CKD.^{6–10} Among these models, machine learning (ML) approaches that use complex computer algorithms are effective in identifying the most critical factors and developing predictive models with superior performance.^{11 12} However, existing models rely primarily on laboratory data and comorbidities to inform their analyses and predictions, neglecting important indicators of CKD progression such as an annual decline in GFR, proteinuria and medication use. Furthermore, the timing (index date) chosen for these models is heterogeneous due to the diverse stages of CKD represented in the patient cohorts, resulting in relatively low predictive power. Therefore, developing an accurate and reliable predictive model that incorporates comprehensive clinical data and standardises index date selection is crucial to improving personalised decision-making for patients with CKD.

In this study, we aim to develop and validate an ML model for early personalised prediction of maintenance dialysis initiation within 1-year and 3-year timeframes

among patients with CKD stages 3–5 using a multicentre longitudinal cohort.

MATERIALS AND METHODS

Study data source

We conducted a retrospective analysis of the Taipei Medical University (TMU) clinical research database (TMUCRD), which comprises comprehensive medical claims data for patients across three affiliated hospitals: TMU Hospital (TMUH), Wan Fang Hospital (WFH) and Shuang Ho Hospital (SHH). The TMUCRD encompasses structured and unstructured data for over 4 million patients, spanning the period from 1998 to 2021. All data were fully anonymised prior to analysis, with patients' identity codes and medical facility information scrambled to ensure patient privacy. This study was authorised by the joint institutional review board (IRB) committee of TMU (IRB#: N202105032).

Study population

We identified patients diagnosed with CKD between 1 January 2008 to 31 December 2017, based on the International Classification of Disease, ninth revision (ICD-9) code 585 and 10th revision (ICD-10) code N18. First-time patients diagnosed with CKD stages 3–5 and who had not undergone KRT, defined as the index date, were included in the study. We confirmed the diagnosis based on the G-stages, in which the estimated glomerular filtration rate (eGFR) was calculated using the CKD-EPI creatinine equation.¹³ To define correctly the stage of CKD, only patients who had been diagnosed with CKD prior to the index dates and had their eGFRs observed within 3–6 months after the index date were included. Patients who were younger than 20 years at the time of diagnosis of CKD stages 3–5 or underwent pre-emptive kidney transplantation were excluded from the study (figure 1).

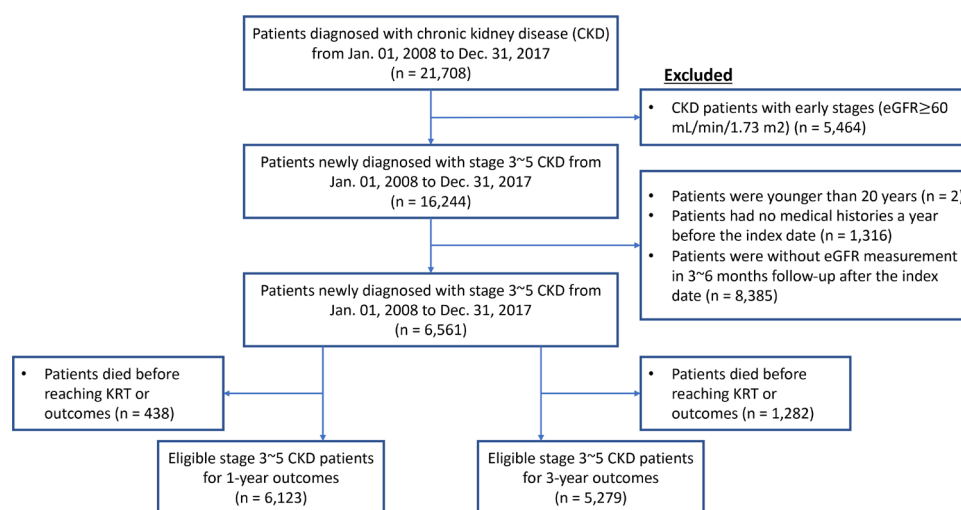


Figure 1 Enrollment process of the first study. The index date is the point of time at which patients with CKD were first diagnosed with stages 3–5. CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate; KRT, kidney replacement therapy.

Observation period

Patients were followed up from the time of stages 3–5 CKD diagnosis (ie, the index date), and their data were censored at the start of maintenance dialysis, loss to follow-up, termination of insurance or the end of the study period (ie, 3 years from the index date). Additionally, patients who died before initiating KRT during the follow-up were excluded from the study.

Outcome measurement

The primary outcome of this study was the initiation of maintenance dialysis among patients with CKD stages 3–5. The initial dialysis point was defined as the first day of maintenance dialysis treatment (eg, the first day of long-term HD, the day of catheter insertion for PD or the first day of PD) based on the procedure-related codes under Taiwan's National Health Insurance (online supplemental table S1). The patients who underwent maintenance dialysis were defined as 'receiving dialysis' and others as 'living without dialysis'.

Variables and data processing

Patient demographics, comorbidities, medications and laboratory data were collected from the database within 1 year before the diagnosis date. The major comorbidities were identified using diagnostic codes (ICD-9 and ICD-10) from outpatient and inpatient databases. The analysis included all diseases listed in the Charlson Comorbidity Index (CCI),¹⁴ along with additional conditions such as essential hypertension, glomerular diseases, lipid metabolism disorders and septicaemia. These diseases were considered confirmed if at least one outpatient or inpatient visit was documented within 1 year prior to the diagnosis date.

The TMUCRD contains comprehensive information on prescribed medications from three affiliated hospitals. Patients' medication prescription claims were tracked using the Anatomical Therapeutic Chemical codes for 1 year preceding the diagnosis date of CKD stages 3–5 (online supplemental table S2).

We also retrieved some routine blood tests from laboratory datasets, including haemoglobin (Hgb), white blood cells (WBC), neutrophils, platelets (PLT), blood urea nitrogen (BUN), creatinine (CREA), cholesterol (CHOL), triglyceride (TG), albumin, calcium, phosphorus, sodium and potassium. The average value of each blood test was calculated based on the results collected within 1 year before the diagnosis date. Blood tests that had over 50% missing values were excluded from the analysis. To manage the missing continuous features, we used the Multiple Imputation by Chained Equations method to fill in these gaps in the data.¹⁵ The eGFR was measured at the diagnosis date (the baseline eGFR) and compared with the previous eGFR, from which the decline of eGFR was calculated using the formula:

The decline of eGFR = (Previous eGFR–Baseline eGFR)/the day interval

In addition, we also collected urine tests and classified albuminuria based on albuminuria categories according to the Kidney Disease Improving Global Outcomes (KDIGO) classification¹³ (online supplemental table S3). Patients with missing values on albuminuria were defined as the group 'unknown'.

Modelling

The classification models were used to predict the initial dialysis, including logistic regression, linear discriminant analysis, gradient boosting machine (GBM), light GBM, AdaBoost, random forest (RF), extreme gradient boosting machine (Xgboost) and artificial neural networks (ANN). A detailed description of different models and their parameters is shown in online supplemental appendix S1.

Model training and testing

To ensure robust model development and account for sample selection bias, we divided the dataset into two parts: the training set and the test set. The training set consisted of patient data from two hospitals, TMUH and WFH, and was used for model development. To evaluate the performance of different ML models and estimate generalisation errors, we applied the stratified fivefold cross-validation method within the training set. This involved dividing the patients into five groups while ensuring that each group represented a proportional distribution of patient characteristics. Each group was then used as the internal validation set for one of the five replications. On the other hand, the test set comprised patient data obtained from SHH and served as an independent dataset for external model validation.

Statistical analysis and evaluation of model performance

Continuous variables were provided as the mean±SD, and categorical variables were provided as absolute (n) and relative (%) frequency; it is described in table 1.

The area under the receiver operating characteristic (ROC) curve (AUC), accuracy, sensitivity (recall), specificity, positive predictive value (precision), negative predictive value and F1 score were computed to evaluate and compare the performance of all prediction models. The model with the highest AUC was selected as the best model through comparison using the external testing set. Furthermore, the impact of features in the best model was analysed using Shapley Additive exPlanations (SHAP) values.¹⁶

Data processing was conducted using SQL Server Management Studio V.18.6 (Redmond, Washington, USA), while model training and testing were performed using Python V.3.8.8 software (Wilmington, Detroit, USA) with scikit-learn V.1.1 (Paris, France).

RESULTS

Data extraction

We identified 16244 eligible patients with CKD stages 3–5 who were diagnosed for the first time in three

Table 1 Demographic characteristics of patients with CKD stages 3–5

	1-year prediction model*		3-year prediction model*	
	Training (n=4496)	Testing (n=1627)	Training (n=3879)	Testing (n=1400)
Age, years, N (%)				
Age <65	1235 (27.5)	478 (29.4)	1185 (30.5)	451 (32.2)
Age ≥65	3261 (72.5)	1149 (70.6)	2694 (69.5)	949 (67.8)
Mean (SD)	72.1 (13.4)	71.3 (13.0)	70.8 (13.3)	70.1 (12.9)
Median (IQR)	74(63 - 82)	73(62 - 81)	73(62 - 81)	71(61 - 80)
Gender, female, N (%)	1872 (41.6)	656 (40.3)	1610 (41.5)	564 (40.3)
Baseline G-stages, N (%)				
G3a	1196 (26.6)	314 (19.3)	1067 (27.5)	270 (19.3)
G3b	1328 (29.5)	390 (24.0)	1142 (29.4)	319 (22.8)
G4	1219 (27.1)	462 (28.4)	995 (25.7)	387 (27.6)
G5	753 (16.7)	461 (28.3)	675 (17.4)	424 (30.3)
Baseline eGFR, mL/min/1.73 m ² , mean (SD)	32.2 (15.5)	27.2 (15.8)	32.4 (15.7)	26.7 (16.0)
Decline of eGFR, mL/min/1.73 m ² , mean (SD)	-0.173 (1.43)	-0.127 (0.75)	-0.166 (1.42)	-0.104 (0.75)
Patients with maintenance dialysis, N (%)††	341 (7.6)	216 (13.3)	752 (19.4)	403 (28.8)
Comorbidities, N (%)				
Diabetes mellitus	2308 (51.3)	850 (52.2)	1994 (51.4)	733 (52.4)
Essential hypertension	2747 (61.1)	782 (48.1)	2385 (61.5)	657 (46.9)
Glomerular diseases	1055 (23.5)	395 (24.3)	951 (24.5)	345 (24.6)
Septicaemia	203 (4.5)	143 (8.8)	157 (4.0)	101 (7.2)
Malignant neoplasm	460 (10.2)	102 (6.3)	356 (9.2)	75 (5.4)
Disorders of lipid metabolism	2124 (47.2)	554 (34.1)	1897 (48.9)	488 (34.9)
Ischaemic heart disease	1541 (34.3)	534 (32.8)	1302 (33.6)	440 (31.4)
Cardiac dysrhythmias	739 (16.4)	187 (11.5)	591 (15.2)	145 (10.4)
Congestive heart failure	918 (20.4)	446 (27.4)	728 (18.8)	366 (26.1)
Cerebrovascular disease	912 (20.3)	326 (20.0)	717 (18.5)	258 (18.4)
Peripheral vascular disease	202 (4.5)	52 (3.2)	165 (4.3)	41 (2.9)
Chronic pulmonary disease	625 (13.9)	218 (13.4)	496 (12.8)	162 (11.6)
Chronic liver disease	391 (8.7)	111 (6.8)	333 (8.6)	91 (6.5)
CCI, N (%)				
CCI <3	1046 (23.3)	367 (22.6)	975 (25.1)	350 (25.0)
CCI ≥3	3450 (76.7)	1260 (77.4)	2904 (74.9)	1050 (75.0)
Mean (SD)	3.84 (1.75)	3.80 (1.57)	3.72 (1.65)	3.67 (1.50)
Median (IQR)	4 (3–5)	4 (3–5)	3 (2–5)	3 (2.25–5)
Medications, N (%)				
Antacids	1256 (27.9)	481 (29.6)	1055 (27.2)	407 (29.1)
H2-receptor antagonists	745 (16.6)	261 (16.0)	622 (16.0)	212 (15.1)
Proton pump inhibitors	569 (12.7)	335 (20.6)	451 (11.6)	273 (19.5)
Laxatives	1484 (33.0)	552 (33.9)	1177 (30.3)	428 (30.6)
Insulins and analogues	906 (20.2)	364 (22.4)	763 (19.7)	314 (22.4)
Sulfonylureas	838 (18.6)	285 (17.5)	754 (19.4)	254 (18.1)
Dipeptidyl peptidase-4 inhibitors	1028 (22.9)	349 (21.5)	904 (23.3)	300 (21.4)
Antiplatelets	2264 (50.4)	1024 (62.9)	1947 (50.2)	872 (62.3)
Vitamin B12 and folic acid	1370 (30.5)	181 (11.1)	1195 (30.8)	165 (11.8)
Organic nitrates	995 (22.1)	470 (28.9)	831 (21.4)	387 (27.6)
Diuretics	2031 (45.2)	789 (48.5)	1682 (43.4)	646 (46.1)

Continued

Table 1 Continued

	1-year prediction model*		3-year prediction model*	
	Training (n=4496)	Testing (n=1627)	Training (n=3879)	Testing (n=1400)
Purine derivatives	1567 (34.9)	629 (38.7)	1386 (35.7)	558 (39.9)
Beta-blocking agents	1894 (42.1)	739 (45.4)	1652 (42.6)	642 (45.9)
Calcium channel blockers	2381 (53.0)	889 (54.6)	2059 (53.1)	767 (54.8)
Agents acting on the renin-angiotensin system	2250 (50.0)	864 (53.1)	1974 (50.9)	748 (53.4)
Statins	1738 (38.7)	576 (35.4)	1547 (39.9)	515 (36.8)
Corticosteroids	802 (17.8)	323 (19.9)	665 (17.1)	250 (17.9)
Beta-lactam antibacterial	1535 (34.1)	657 (40.4)	1243 (32.0)	524 (37.4)
Non-steroids	1287 (28.6)	516 (31.7)	1113 (28.7)	431 (30.8)
Antigout preparations	1736 (38.6)	490 (30.1)	1519 (39.2)	424 (30.3)
Cough and cold preparations	1335 (29.7)	510 (31.3)	1074 (27.7)	410 (29.3)
Antihistamines	718 (16.0)	329 (20.2)	592 (15.3)	261 (18.6)
Laboratory tests, mean (SD)				
Haemoglobin, g/dL	113 (19.3)	110 (21.0)	114 (19.2)	110 (21.2)
White blood cells k/uL	7.55 (2.61)	7.94 (3.32)	7.52 (2.56)	7.90 (3.22)
Neutrophils, %	70.3 (8.74)	71.7 (9.55)	70.3 (8.50)	71.7 (9.22)
Platelets, mL	206 (65.4)	215 (75.4)	208 (64.3)	214 (71.4)
Blood urea nitrogen, mg/dL	37.3 (21.1)	42.2 (24.9)	37.2 (21.1)	42.9 (25.4)
Creatinine, mg/dL	2.45 (1.73)	3.06 (2.17)	2.49 (1.80)	3.18 (2.25)
Aspartate transferase, U/L	18.7 (17.9)	26.9 (24.7)	18.3 (17.2)	25.9 (20.4)
Cholesterol, mg/dL	179 (39.3)	184 (42.8)	180 (39.7)	185 (43.7)
Triglycerides, mg/dL	147 (103)	157 (273)	148 (106)	160 (289)
Albumin, g/dL	3.96 (0.43)	3.98 (0.20)	3.98 (0.42)	3.98 (0.20)
Calcium, mg/dL	8.98 (0.52)	8.96 (0.53)	8.99 (0.52)	8.96 (0.53)
Phosphorous, mg/dL	3.83 (0.62)	4.04 (0.75)	3.84 (0.64)	4.08 (0.77)
Sodium, mmol/L	139 (3.72)	138 (3.8)	139 (3.56)	138 (3.68)
Potassium, mmol/L	4.40 (0.60)	4.41 (0.68)	4.41 (0.59)	4.43 (0.68)
Albuminuria, N (%)				
A1	845 (18.8)	262 (16.1)	751 (19.4)	225 (16.1)
A2	651 (14.5)	183 (11.2)	565 (14.6)	158 (11.3)
A3	1494 (33.2)	679 (41.7)	1320 (34.0)	609 (43.5)
Unknown	1506 (33.5)	503 (30.9)	1243 (32.0)	408 (29.1)

*Study aimed to develop two prediction models that observed patients for 1 year and 3 years.

†Patients with KRT or maintenance dialysis is the outcome of the study.

CCI, Charlson Comorbidity Index; eGFR, estimated Glomerular filtration rate; G-stages, glomerular filtration rate stages; KRT, kidney replacement therapy.;

TMU-affiliated hospitals from 1 January 2008 to 31 December 2017. Among those, we excluded 9703 patients due to being younger than 20, having no medical histories a year before the index date or lacking evidence for the diagnosis of CKD. Furthermore, we excluded 438 and 1282 patients for the 1-year and 3-year prediction models, respectively, who had died prior to undergoing maintenance dialysis. Finally, 6123 and 5279 patients who met all inclusion criteria were included in model development for 1-year and 3-year prediction performances, respectively (figure 1).

Study population characteristics

The patients' demographics, comorbidities, medications and laboratory data were summarised in table 1. The study population was over 65 years, with a mean (SD) ages of 71.7 (13.2) and 70.5 (13.1) years for the 1-year and 3-year prediction models, respectively. Most patients were male (59.1%) and had baseline G-stages of G3 (eg, 56% for the training set and 43% for the testing set). Patients had a high prevalence of comorbidities such as diabetes (51%), hypertension (61%), disorders of lipid metabolism (47%) and glomerular disease (24%). The outcome of initiation

Table 2 Summary of different classification models

Classifiers	Training AUC	Testing AUC	Accuracy	Sensitivity	Specificity	Precision	F1 score
1-year prediction model performances							
Logistic regression	0.92	0.90	0.80	0.89	0.79	0.39	0.71
Linear discriminant analysis	0.92	0.90	0.80	0.87	0.79	0.39	0.67
Gradient boosting classifier	0.98	0.89	0.81	0.83	0.81	0.40	0.63
LGBM classifier	1.00	0.86	0.76	0.83	0.75	0.34	0.60
Ada Boost classifier	0.97	0.81	0.83	0.69	0.85	0.41	0.51
Random forest classifier	1.00	0.89	0.82	0.85	0.82	0.42	0.65
XGB classifier	1.00	0.87	0.82	0.80	0.82	0.41	0.63
ANN*	0.99	0.96	0.89	0.88	0.75	0.39	0.60
3-year prediction model performances							
Logistic regression	0.91	0.90	0.80	0.88	0.77	0.61	0.82
Linear discriminant analysis	0.92	0.91	0.83	0.86	0.82	0.66	0.81
Gradient boosting classifier	0.96	0.91	0.82	0.84	0.81	0.65	0.81
LGBM classifier	1.00	0.90	0.80	0.89	0.76	0.60	0.82
Ada Boost classifier	0.95	0.89	0.79	0.84	0.78	0.60	0.80
Random forest Classifier	1.00	0.90	0.80	0.88	0.77	0.61	0.82
XGB classifier	1.00	0.90	0.82	0.83	0.82	0.65	0.80
ANN*	0.95	0.92	0.82	0.87	0.79	0.63	0.73

*Best model based on AUC values.

.ANN, artificial neural network; AUC, area under the receiver operating characteristic curve; LGBM, light gradient boosting machine; XGB, extreme gradient boosting.

of maintenance dialysis was observed at 10.5% and 24.1% for the 1-year and 3-year prediction models, respectively.

The performances of different prediction models

The performances of different prediction models are shown in table 2. For a 1-year prediction of successful dialysis treatment, the highest AUC value of 0.96 was observed for the ANN model (ie, sensitivity, 0.88; specificity, 0.75; precision, 0.39 and F1 score, 0.6), followed by the GBM and RF models with an AUC value of 0.89. Likewise, the ANN model was performed with a better AUC value of 0.92 (ie, sensitivity, 0.87; specificity, 0.79; precision, 0.63 and F1 score, 0.73) than other ML models in the 3-year prediction of receiving maintenance dialysis. The ROC curves of varying prediction models for 1-year and 3-year successful dialysis treatment are shown in figure 2.

Features importance

The lists of the top 20 important features that might impact the prediction model's performance for 1-year and 3-year successful dialysis are shown in figure 3. The essential features of the 1-year and 3-year follow-up models were baseline eGFR, BUN, creatinine, triglyceride, age, gender, Hgb, CHOL, PLTs, albuminuria, diabetes disease, hypertension and related medications (eg, diuretics, insulin, dipeptidyl peptidase-4 and calcium channel blockers).

DISCUSSION

Our study findings demonstrated that ML classification models are well suited for a meaningful prediction of the initiation of maintenance dialysis in patients with CKD stages 3–5. The ANN method showed a better performance level for 1-year and 3-year prediction of dialysis commencement with a higher AUC (0.96 and 0.92), good sensitivity (0.88 and 0.87) and specificity (0.75 and 0.79).

In previous studies, AI models have been applied to predict CKD progression and start KRT. In 2015, Jamshid Norouzi *et al* used an adaptive neuro-fuzzy inference system to predict renal failure progression. Their model could accurately (>95%) predict the GFR for 6-month to 18-month intervals. However, only 465 patients with CKD were included in their study, and it was noted that proteinuria was not an important feature in their model.⁸ In 2019, Jing Xiao *et al* developed ML models to predict CKD progression. Their model used only the patient's demographics and biochemical blood features, not features derived from a urinalysis. Besides, the predictive power of the model was not high (AUC: 0.873, sensitivity: 0.83 and specificity: 0.82).¹⁷ Another model was performed using only comorbidity data from 8492 patients to predict the onset of KRT, and their results were even lower (AUC, sensitivity and specificity were only 0.773, 0.623 and 0.781, respectively).⁷ Recently, Qiong Bai *et al* also conducted an ML model to predict

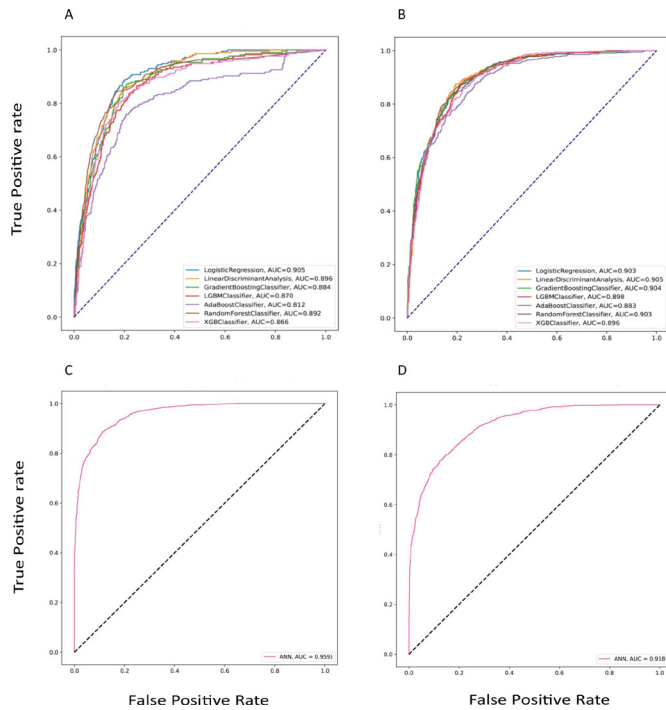


Figure 2 The performance of the prediction models in the testing dataset. (A and C) 1-year prediction with machine learning and ANN models; (B and D) 3-year prediction with machine learning and ANN models. ANN, artificial neural network; AUC, area under the receiver operating characteristic curve; LGBM, light gradient boosting machine; XGB, extreme gradient boosting.

the risk of ESRD. This model selected many important factors associated with the progression of CKD, including demographics, blood tests and comorbidities, but not proteinuria. However, the predictability has not improved compared with previous models.¹⁸ It could be explained by the fact that many patients in that study were in the early stages of CKD, resulting in a low percentage of those who progressed to ESRD when followed for a short period of time. The imbalance in the outcome can significantly affect the model's predictive power.

In this study, we only focused on patients with CKD stages 3–5, and their risk of progression to ESRD is high. Hence, predicting the time of their dialysis commencement is very practical in our daily clinical care. Moreover,

we carefully identified the model features associated with CKD progression and KRT based on the clinical setting and traditional logistic regression analysis. Forty-five significant prognostic factors were selected, including patient demographics, comorbidities, routine blood and urine tests and commonly used medications. Therefore, the predictive ability of our model has higher accuracy.

In further analysis of the ANN model, we also ranked all predictors according to their influence on the 1-year and 3-year models using SHAP values.¹⁹ Notably, several distinct features have been identified, respectively. For example, age, comorbidity (CCI score), PLT counts and WBC counts were important contributing factors in the 1-year prediction model, whereas gender and other medications such as proton pump inhibitors, beta-lactam antibacterial agents, organic nitrates and H2-receptor antagonists were relevant factors in the 3-year model. Common important factors identified in both models included eGFR at baseline, blood urea, serum creatinine and albuminuria (see figure 3). These are also key determinants for the risk classification of CKD according to the 2012 KDIGO guidelines.^{13 20} Other contributing factors in both models included serum Hgb level, TG or CHOL levels, hypertension, diabetes mellitus, diuretic use, anti-hypertensive agents and medications for controlling blood glucose levels (see figure 3). Anaemia typically develops during the course of CKD; a decrease in serum Hgb is significantly associated with the progression of CKD.^{21 22} Diabetic nephropathy is the leading cause of ESRD in adults.^{23 24} In patients with diabetic CKD, blood glucose levels are associated with poor outcomes such as serum creatinine doubling, ESRD and mortality, and intensive glycaemic control could reduce these risks.^{25–29} Additionally, several studies have demonstrated that certain levels of dyslipidaemia is independently associated with rapid renal progression, KRT, all-cause mortality and cardiovascular death in predialysis patients.^{30–33} Hypertension may occur early during the course of CKD and is related to a more rapid decline of kidney function, the development of cardiovascular disease and death in patients with CKD.^{34 35} Early intervention and tight control of blood pressure could lessen the risk of CVD and all-cause death in patients with and without CKD.^{36 37} Diuretics are an important part of guideline-directed medical therapy for patients with CKD with hypertension, oedema and hyperkalaemia.³⁸ In terms of adverse effects, whether diuretics are an independent risk factor for CKD progression remains controversial. However, these medicines played important roles in both our models.^{39–41} Therefore, diuretics should be used with caution in patients with CKD stages 3–5. Finally, the GFR decline rate is also influenced by some immutable patient factors. The Kidney Disease Outcomes Quality Initiative guideline has provided ample evidence that African-American race (not justified in this study), male gender and older age are related to a more rapid GFR reduction.²⁰ In summary, our models take advantage of the important factors involved in the progression of CKD, are consistent with

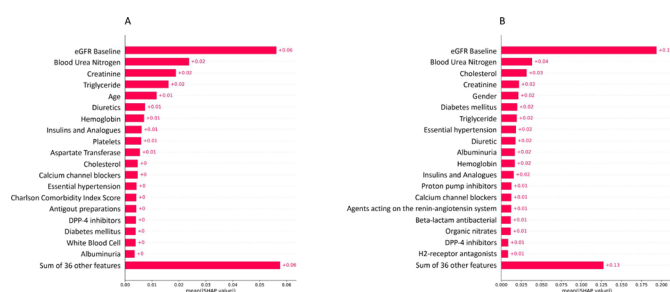


Figure 3 Feature importance of the ANN prediction model. (A) 1-year prediction model and (B) 3-year prediction model. ANN, artificial neural network; eGFR, estimated glomerular filtration rate; DPP-4, dipeptidyl peptidase

current clinical practice guidelines and are highly applicable. They could be a good screening tool to determine the likelihood of initiating long-term dialysis by using the available clinical data on the patient. Several limitations need to be addressed. First, due to the patient's lack of weight and height, the body surface area was not adjusted for eGFR. As a result, the determination of G-stage using unadjusted eGFR may be inaccurate for oversized patients. Second, using the decline in eGFR between baseline eGFR and previous eGFR may not accurately capture the progression of CKD when compared with the annual decline in eGFR during the follow-up period. Consequently, this factor did not significantly contribute to our model. Third, we only used retrospective data from three hospitals in Taipei to create our models, and it is widely recognised that racial and regional variables also influence CKD progression. Further work should involve training and validating the models through multinational and multiracial data before the clinical application is generalised. Fourth, we incorporated all important features into the prediction model, acknowledging that this approach might not be practical for clinical implementation. Nevertheless, these features underwent meticulous screening and hold varying degrees of significance in relation to CKD progression. Additionally, we assessed the model using only the top 10 important features and obtained comparable results (online supplemental tables S4 and S5, online supplemental figures S1 and S2).

CONCLUSION

We have shown that using the machine learning approach can develop a highly predictive model for estimating the timing of maintenance dialysis initiation in patients with CKD stages 3–5, which provides a further step towards personalised treatment in this population.

Author affiliations

¹Nephro-Urology and Dialysis Center, Bach Mai Hospital, Hanoi, Vietnam

²Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan

³Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei, Taiwan

⁴Research Center of Health Care Industry Data Science, College of Management, Taipei Medical University, Taipei, Taiwan

⁵International PhD program of Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan

⁶University Medical Center, Ho Chi Minh City, Vietnam

⁷Department of Internal Medicine, Hanoi Medical University, Hanoi, Vietnam

⁸Division of Nephrology, Department of Internal Medicine, Shuang Ho Hospital, Taipei Medical University, New Taipei City, Taiwan

⁹Division of Nephrology, Department of Internal Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

¹⁰TMU-Research Center of Urology and Kidney (TMU-RCUK), Taipei Medical University, Taipei, Taiwan

¹¹Division of Nephrology, Department of Internal Medicine, Hsin Kuo Min Hospital, Taipei Medical University, Taoyuan City, Taiwan

¹²Division of Nephrology, Department of Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

¹³Division of Cardiovascular Surgery, Department of Surgery, Shuang Ho Hospital, Taipei Medical University, New Taipei City, Taiwan

¹⁴Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan

¹⁵International Center for Health Information Technology, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan

¹⁶School of Population Health, Faculty of Medicine and Health, University of New South Wales (UNSW), Sydney, New South Wales, Australia

¹⁷Global Health & Health Security Department, College of Public Health, Taipei Medical University, Taipei, Taiwan

Acknowledgements The authors thank the staffs at Office of Data Science, Taipei Medical University for their assistance in data collection and processing.

Contributors THA, CTL, Y-HH and M-SW conceptualised the study and wrote the first draft. THA, PAN and C-TL completed and edited the final manuscript. THA, PAN, PPT, I-JC, C-LC and T-HC collected and assimilated necessary data. CTL is guarantor.

Funding This research was funded by Taipei Medical University (TMU109-AE1-B31) and the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly 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 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iDs

Anh Trung Hoang <http://orcid.org/0000-0001-5748-5035>

Phung-Anh Nguyen <http://orcid.org/0000-0002-7436-9041>

Usman Iqbal <http://orcid.org/0000-0002-0614-123X>

REFERENCES

- Johansen KL, Chertow GM, Gilbertson DT, *et al.* US renal data system 2021 annual data report: epidemiology of kidney disease in the United States. *Am J Kidney Dis* 2022;79:A8–12.
- Farrington K, Warwick G. Renal Association clinical practice guideline on planning, initiating and withdrawal of renal replacement therapy. *Nephron Clin Pract* 2011;118 Suppl 1:c189–208.
- Daugirdas JT, Depner TA, Inrig J, *et al.* KDOQI clinical practice guideline for hemodialysis adequacy: 2015 update. *Am J Kidney Dis* 2015;66:884–930.
- Chan CT, Blankestijn PJ, Dember LM, *et al.* Dialysis initiation, modality choice, access, and prescription: conclusions from a kidney disease: improving global outcomes (KDIGO) controversies conference. *Kidney Int* 2019;96:37–47.
- Nesrallah GE, Mustafa RA, Clark WF, *et al.* Canadian Society of Nephrology 2014 clinical practice guideline for timing the initiation of chronic dialysis. *CMAJ* 2014;186:112–7.
- Lee M-J, Park J-H, Moon YR, *et al.* Can we predict when to start renal replacement therapy in patients with chronic kidney disease using 6 months of clinical data *PLoS ONE* 2018;13:e0204586.
- Dovgan E, Gradišek A, Luštrek M, *et al.* Using machine learning models to predict the initiation of renal replacement therapy among chronic kidney disease patients. *PLoS One* 2020;15:e0233976.
- Norouzi J, Yadollahpour A, Mirbagheri SA, *et al.* Predicting renal failure progression in chronic kidney disease using integrated intelligent fuzzy expert system. *Comput Math Methods Med* 2016;2016:6080814.

- 9 Chang H-L, Wu C-C, Lee S-P, *et al.* A predictive model for progression of CKD. *Medicine (Baltimore)* 2019;98:e16186.
- 10 Zacharias HU, Altenbuchinger M, Schultheiss UT, *et al.* A predictive model for progression of CKD to kidney failure based on routine laboratory tests. *Am J Kidney Dis* 2022;79:217–30.
- 11 Chen T, Li X, Li Y, *et al.* Prediction and risk stratification of kidney outcomes in IgA nephropathy. *Am J Kidney Dis* 2019;74:300–9.
- 12 Akbilgic O, Obi Y, Potukuchi PK, *et al.* Machine learning to identify dialysis patients at high death risk. *Kidney Int Rep* 2019;4:1219–29.
- 13 Stevens PE. Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. *Ann Intern Med* 2013;158:825.
- 14 Quan H, Sundararajan V, Halfon P, *et al.* Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;43:1130–9.
- 15 Azur MJ, Stuart EA, Frangakis C, *et al.* Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011;20:40–9.
- 16 Sphinx. Welcome to the SHAP documentation. 2018. Available: <https://shap.readthedocs.io/en/latest/index.html>
- 17 Xiao J, Ding R, Xu X, *et al.* Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *J Transl Med* 2019;17:119.
- 18 Bai Q, Su C, Tang W, *et al.* Machine learning to predict end stage kidney disease in chronic kidney disease. *Sci Rep* 2022;12:8377.
- 19 Shapley LS. 17. A value for N-person games. In: *Contributions to the theory of games (AM-28)*. Princeton University Press, 2016: Volume II. 307–18.
- 20 National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Am J Kidney Dis* 2002;39:S1–266.
- 21 Astor BC, Coresh J, Heiss G, *et al.* Kidney function and anemia as risk factors for coronary heart disease and mortality: the Atherosclerosis risk in communities (ARIC) study. *Am Heart J* 2006;151:492–500.
- 22 Kovesdy CP, Trivedi BK, Kalantar-Zadeh K, *et al.* Association of anemia with outcomes in men with moderate and severe chronic kidney disease. *Kidney Int* 2006;69:560–4.
- 23 Centers for Disease Control and Prevention. *Chronic kidney disease in the United States, 2021*. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, 2021.
- 24 Van Dijk PCW, Jager KJ, Stengel B, *et al.* Renal replacement therapy for diabetic end-stage renal disease: data from 10 registries in Europe (1991–2000). *Kidney Int* 2005;67:1489–99.
- 25 Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. *N Engl J Med* 2008;358:2560–72.
- 26 Ismail-Beigi F, Craven T, Banerji MA, *et al.* Effect of intensive treatment of Hyperglycaemia on Microvascular outcomes in type 2 diabetes: an analysis of the ACCORD randomised trial. *Lancet* 2010;376:419–30.
- 27 Duckworth W, Abraira C, Moritz T, *et al.* Glucose control and vascular complications in veterans with type 2 diabetes. *N Engl J Med* 2009;360:129–39.
- 28 Intensive blood-glucose control with Sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK prospective diabetes study (UKPDS) group. *Lancet* 1998;352:837–53.
- 29 Jung HH. Evaluation of serum glucose and kidney disease progression among patients with diabetes. *JAMA Netw Open* 2021;4:e2127387.
- 30 Chen S-C, Hung C-C, Kuo M-C, *et al.* Association of dyslipidemia with renal outcomes in chronic kidney disease. *PLoS One* 2013;8:e55643.
- 31 Kovesdy CP, Anderson JE, Kalantar-Zadeh K. Inverse association between lipid levels and mortality in men with chronic kidney disease who are not yet on dialysis: effects of case mix and the malnutrition-inflammation-Cachexia syndrome. *J Am Soc Nephrol* 2007;18:304–11.
- 32 Chawla V, Greene T, Beck GJ, *et al.* Hyperlipidemia and long-term outcomes in nondiabetic chronic kidney disease. *Clin J Am Soc Nephrol* 2010;5:1582–7.
- 33 Shlipak MG, Fried LF, Cushman M, *et al.* Cardiovascular mortality risk in chronic kidney disease: comparison of traditional and novel risk factors. *JAMA* 2005;293:1737–45.
- 34 USRDS System. *USRDS annual data report: epidemiology of kidney disease in the United States*. Bethesda, MD: National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, 2020.
- 35 Thompson S, James M, Wiebe N, *et al.* Cause of death in patients with reduced kidney function. *J Am Soc Nephrol* 2015;26:2504–11.
- 36 Cheung AK, Rahman M, Reboussin DM, *et al.* Effects of intensive BP control in CKD. *J Am Soc Nephrol* 2017;28:2812–23.
- 37 The SPRINT Research Group. Final report of a trial of intensive versus standard blood-pressure control. *N Engl J Med* 2021;384:1921–30.
- 38 Ku E, Lee BJ, Wei J, *et al.* Hypertension in CKD: core curriculum 2019. *Am J Kidney Dis* 2019;74:120–31.
- 39 Levi TM, Rocha MS, Almeida DN, *et al.* Furosemide is associated with acute kidney injury in critically ill patients. *Braz J Med Biol Res* 2012;45:827–33.
- 40 Khan YH, Sarriff A, Mallhi TH, *et al.* Is diuretic use beneficial or harmful for patients with chronic kidney disease *Eur J Hosp Pharm* 2017;24:253–4.
- 41 Fitzpatrick JK, Yang J, Ambrosy AP, *et al.* Loop and thiazide diuretic use and risk of chronic kidney disease progression: a Multicentre observational cohort study. *BMJ Open* 2022;12:e048755.

© 2024 Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY. Published by BMJ. <https://creativecommons.org/licenses/by/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Definitions of digital biomarkers: a systematic mapping of the biomedical literature

Ana Karen Macias Alonso,^{1,2} Julian Hirt ,^{2,3,4} Tim Woelfle,^{2,5} Perrine Janiaud,^{2,3} Lars G Hemkens^{2,3,6,7}

To cite: Macias Alonso AK, Hirt J, Woelfle T, *et al*. Definitions of digital biomarkers: a systematic mapping of the biomedical literature. *BMJ Health Care Inform* 2024;**31**:e100914. doi:10.1136/bmjhci-2023-100914

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2023-100914>).

Received 27 September 2023
Accepted 06 March 2024

ABSTRACT

Background Technological devices such as smartphones, wearables and virtual assistants enable health data collection, serving as digital alternatives to conventional biomarkers. We aimed to provide a systematic overview of emerging literature on ‘digital biomarkers,’ covering definitions, features and citations in biomedical research.

Methods We analysed all articles in PubMed that used ‘digital biomarker(s)’ in title or abstract, considering any study involving humans and any review, editorial, perspective or opinion-based articles up to 8 March 2023. We systematically extracted characteristics of publications and research studies, and any definitions and features of ‘digital biomarkers’ mentioned. We described the most influential literature on digital biomarkers and their definitions using thematic categorisations of definitions considering the Food and Drug Administration Biomarkers, EndpointS and other Tools framework (ie, data type, data collection method, purpose of biomarker), analysing structural similarity of definitions by performing text and citation analyses.

Results We identified 415 articles using ‘digital biomarker’ between 2014 and 2023 (median 2021). The majority (283 articles; 68%) were primary research. Notably, 287 articles (69%) did not provide a definition of digital biomarkers. Among the 128 articles with definitions, there were 127 different ones. Of these, 78 considered data collection, 56 data type, 50 purpose and 23 included all three components. Those 128 articles with a definition had a median of 6 citations, with the top 10 each presenting distinct definitions.

Conclusions The definitions of digital biomarkers vary significantly, indicating a lack of consensus in this emerging field. Our overview highlights key defining characteristics, which could guide the development of a more harmonised accepted definition.

INTRODUCTION

Biomarkers are defined as a set of characteristics that are objectively measured and used as indicators of normal biological processes, pathogenic processes or biological responses that appear due to exposure or therapeutic interventions.¹ This comprises physiological, molecular, histologic and radiographic measurements.² The US Food and Drug Administration (FDA)

subclassifies susceptible/risk, diagnostic, monitoring, prognostic, predictive, response and safety biomarkers.¹ They highlight that a full biomarker description must include the source or matrix, the measurable characteristic(s) and the methods used to measure the biomarker.¹ The digitalisation of our world impacting daily living and healthcare broadens the spectrum of the possible source and methods used to measure biomarkers and introduces a novel dimension of measurable characteristics. This allows digital devices used daily, such as smartphones, wearable devices, sensors and smart home devices, to provide a new category of biomarkers, often called ‘digital biomarkers’. In recent years, digital biomarkers became increasingly present in routine care and in research in many areas of medicine, such as cardiology, oncology or COVID-19. For example, smartphone recorded cough sounds have been used as a digital biomarker to detect asthma and respiratory infections in clinical trials,^{3 4} or deep learning was applied to data from a three-axis accelerometer to predict sleep/wake patterns.^{4 5} Moreover, such digital biomarkers have spread in the field of neurology, which has a large unmet need for non-invasive and objective biomarkers reflecting cognitive and motor functions that are traditionally assessed with specific tests performed by neurologists.⁶ Beyond monitoring health and disease status, predicting the occurrence and development of diseases would be promising applications of such novel approaches.⁷

Thus, digital biomarkers have the potential to offer valuable insights on the health of patients. They usually have high temporal resolution (up to (quasi-)continuous), are usually objective (and not subject to interobserver variability) and can have high external validity as they may be applied in the patient’s routine environment (as opposed to, eg, the clinic or a research environment).⁸



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Lars G Hemkens;
lars.hemkens@usb.ch

Many everyday digital tools used mainly for entertainment/leisure purposes (eg, fitness trackers) are increasingly considered as a source of helpful information that may be transformed into digital biomarkers. Yet, with all this diversity in application and complex interaction with rapidly evolving technology, it becomes necessary to provide a clear and precise definition of the fundamental underlying concepts to facilitate research and decision-making with and on these novel approaches.

One of the first definitions of this novel type of biomarker was provided by Dorsey *et al*, who defined digital biomarkers as ‘the use of a biosensor to collect objective data on a biological (eg, blood glucose, serum sodium), anatomical (eg, mole size) or physiological (eg, heart rate, blood pressure) parameter obtained using sensors followed by algorithms to transform these data into interpretable outcome measures, helping to address many of the shortcomings in current measures.’ Furthermore, they stated that these new measures ‘include portable (eg, smartphones), wearable, and implantable devices, and are by their nature largely independent of raters.’⁹ A later definition given in 2020 by the European Medicines Agency (EMA) was based on ‘digital measures’ (‘measured through digital tools’) and did not include the requirement of algorithms as a defining feature: ‘a digital biomarker is an objective, quantifiable measure of physiology and/or behaviour used as an indicator of biological, pathological process or response to an exposure or an intervention that is derived from a digital measure. [...]’¹⁰

Others gave broader definitions including further defining features, for example, defining digital biomarkers as ‘objective, quantifiable, quantitative, physiological and behavioural data that are collected and measured by means of digital devices such as portables, wearables, implantables or digestibles. The data collected are used to explain, influence and/or predict health-related outcomes’.^{2 6 11}

Overall, such a disagreement between definitions used by regulators and in articles published in high-impact biomedical journals raised concerns that no clear consensus exists among researchers and users of this novel approach and terminology, increasing the risk for miscommunication. There are numerous examples where differences in definitions have been recognised as critical cause of inefficiencies and delay in health research and avoidable controversy, uncertainty and potential harm in clinical care and public health.^{12–15} The Biomarkers, EndpointS and other Tools (BEST) framework developed by the FDA and US National Institutes of Health with ‘the goals of improving communication, aligning expectations, and improving scientific understanding’ highlights that ‘unclear definitions and inconsistent use of key terms can hinder the evaluation and interpretation of scientific evidence and may pose significant obstacles to medical product development programmes’.¹ We aimed to provide a systematic overview of the emerging literature on digital biomarkers and characterisation of

the definitions of digital biomarkers that are provided in biomedical journal articles by performing a systematic mapping and citation analysis of all articles that prominently used the term ‘digital biomarker’. We sought to determine differences in characteristics of common definitions to provide a foundation for subsequent activities to develop clearer and consistent definitions that ensure improved application of digital biomarkers in research and healthcare decision-making.

METHODS

Design

We analysed all articles published at any time in PubMed that prominently used the term ‘digital biomarker’, that is, either in title or abstract.

We systematically explored definitions of digital biomarkers that are provided and/or referred to in the biomedical literature, that is, journal articles that are indexed in PubMed, in a mapping review without a formal assessment of included studies.¹⁶ We structured our review report to the ‘Preferred Reporting Items for Systematic Reviews and Meta-Analyses’ guidance, where applicable.¹⁷ We did not use a prespecified protocol.

Eligibility criteria, information source and search strategy

We searched PubMed and included all articles mentioning ‘digital biomarker’ or ‘digital biomarkers’ in their title or abstract (by searching PubMed for ‘digital biomarker*(-tiab)’; date of last search: 8 March 2023). We excluded animal research.

Study selection

One reviewer (AKMA) screened titles, abstracts and full texts for eligibility. Confirmation by a second reviewer (JH or LGH) was planned for situations where the reviewer was unsure, but this case never occurred given the clear and objective selection criteria.

Data extraction

We developed a spreadsheet to structure the data extraction process. One reviewer (AKMA) extracted data with confirmation by a second reviewer (JH or LGH) in case of any uncertainty.

We extracted from every article: author(s), publication year, title, journal, corresponding author, and country of correspondence, article type (ie, primary research, review or other type (eg, editorial, comment, opinion-based letter)). Of primary research articles, we additionally extracted definitions of digital biomarkers that are provided and/or referred to (based on a semantic search for indicators of definition such as ‘digital biomarkers are’, ‘... are defined as’, ‘... can be defined’, ‘the definition of ... is’), medical context, and whether the article is about the development and/or validation of a digital biomarker. The number of global citations was obtained by using metadata from OpenAlex¹⁸; accessed via the Local Citation Network¹⁹ (as of 26 June 2023).

Data analysis and categorisation of definition components

We considered the BEST framework to derive components of definitions for digital biomarkers.¹ We analysed the identified digital biomarker definitions by assessing if they contained descriptions that fall within three key components, that is, the (1) type of data that is measured (eg, whether data were measured objectively, continuously or quantitatively), (2) data collection method (eg, whether sensors, computers, portables, wearables, implantables or ingestibles were used to collect data) and (3) purpose of the digital biomarker (eg, whether a biomarker was used as measure of disease progression or to predict health-related outcomes). We defined definitions as duplicates when they used the same sequence of words. We illustrate the frequency of various terminologies used in all provided definitions with a word cloud.²⁰ We analysed the structural similarity of definitions that were provided without a reference by performing hierarchical clustering on the distance-matrix containing pairwise 'Indel'-distances, that is, 'the minimum number of insertions and deletions required to change one (definition) into the other'.²¹ Since we aimed at exploring how digital biomarkers are defined in the biomedical literature, we did not critically assess the included articles and studies. For the analysis of citations, we calculated the quotient of number of global citations (retrieved by the Local Citation Network¹⁹) and years since publication per article. To create a citation network of citing and cited relationships between the articles, we used the Local Citation Network with the OpenAlex scholarly index.^{19 22}

We used descriptive statistics by reporting numbers and percentages. For all analyses, we used R (V.4.2.2) or Python (V.3.11.4).

RESULTS

We identified 415 articles that had 'digital biomarker' in their title or abstract (online supplemental S1). The first article was published in 2014 (median publication year 2021; [figure 1](#); online supplemental S2). Most articles described primary studies (n=283; 68%) and were published in digital medicine specialty journals, including *Digital Biomarkers* (n=35; 8%), *Journal of Medical Internet Research* (n=21; 5%) or *npj Digital Medicine* (n=19; 4%; [table 1](#)). Of the 415 articles, 128 (31%) provided at least 1 definition of a digital biomarker.

Characteristics of articles providing a definition of digital biomarker

The 128 articles with a definition of digital biomarker were published between 2015 and 2023 (median: 2021). Of them, 59 articles were primary studies, 50 were reviews and 19 were other types of articles ([table 1](#)).

Almost all primary studies described the development of one or more digital biomarkers (53 of 59 articles), and many described a validation process of biomarkers (35 of 59 articles). The most frequent medical field of the primary research articles that described the development of one or more digital biomarkers was neurology (25 of 53), while the spectrum of medical fields was overall very wide ([table 1](#)). The most frequent diseases

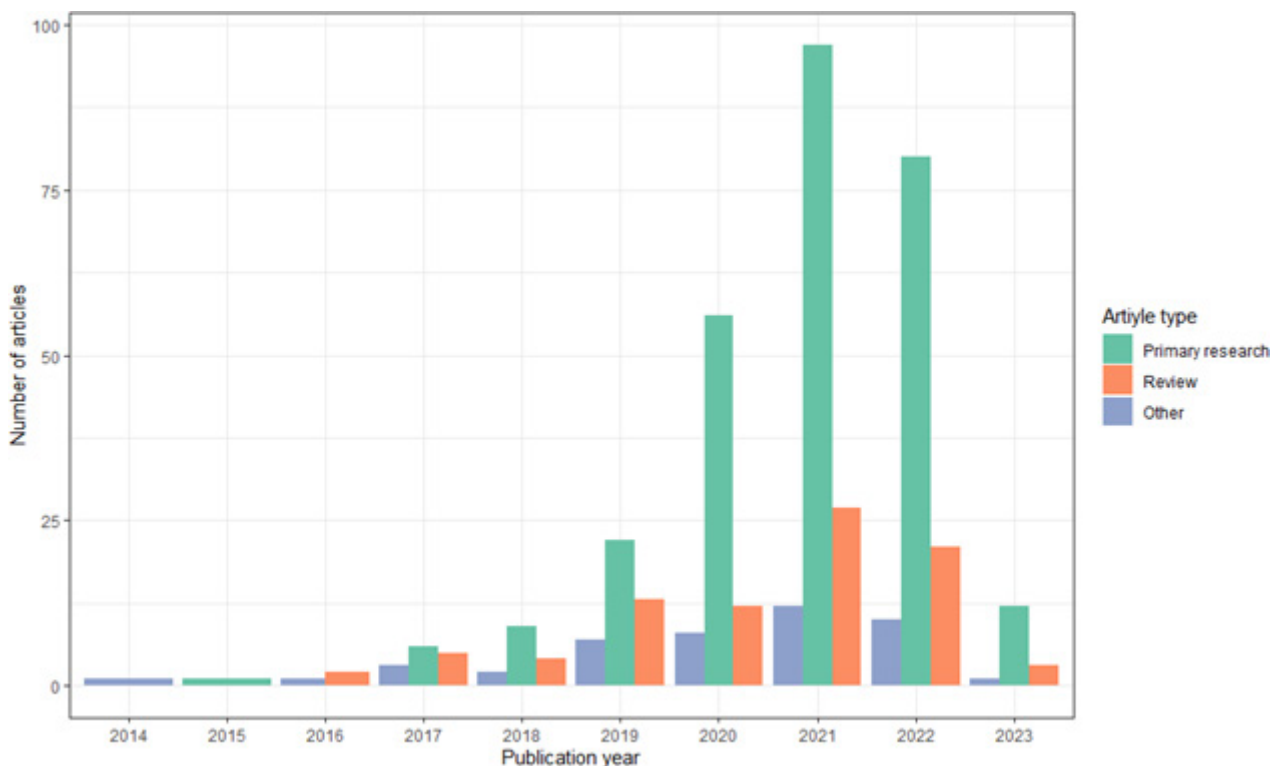


Figure 1 The annual number of published article types referring to digital biomarkers as of 8 March 2023 (n=415).

Table 1 Characteristics of all 415 articles in PubMed using 'digital biomarker' in title or abstract

	All articles (n=415)	Articles with a definition of digital biomarker (n=128)
	n (%)	n (%)
Publication year: median, range	2020, 2014–2023	2021, 2015–2023
Type of articles		
Primary research	283 (68.2)	59 (46.1)
Development of a digital biomarker*	–	53 (41.4)
Medical context		
Neurology	–	25 (19.5)
Cardiology	–	3 (2.3)
Endocrinology	–	3 (2.3)
Geriatrics	–	3 (2.3)
Psychiatry	–	3 (2.3)
Sleep medicine	–	3 (2.3)
Infectiology	–	2 (1.6)
Oncology	–	2 (1.6)
Psychology	–	2 (1.6)
Rheumatology	–	2 (1.6)
Addiction medicine	–	1 (0.8)
Not specified	–	7 (5.5)
Disease specific		
Dementia/MCI/CI	–	16 (12.5)
Parkinson's disease	–	5 (3.9)
Diabetes	–	3 (2.3)
Alcohol use disorder	–	2 (1.6)
Arthritis	–	2 (1.6)
COVID-19	–	2 (1.6)
Multiple sclerosis	–	2 (1.6)
Not specified	–	14 (10.9)
Others*	–	8 (6.2)
Validation of a digital biomarker†	–	35 (27.3)
Reviews	87 (21.0)	50 (39.1)
Editorials, opinions, perspectives, etc	45 (10.8)	19 (14.8)
Journals		
<i>Digital Biomarkers</i>	35 (8.4)	15 (11.7)
<i>Journal of Medical Internet Research</i>	21 (5.1)	5 (3.9)

Continued

Table 1 Continued

	All articles (n=415)	Articles with a definition of digital biomarker (n=128)
	n (%)	n (%)
<i>npj Digital Medicine</i>	19 (4.6)	8 (6.3)
<i>Sensors (Basel, Switzerland)</i>	18 (4.3)	2 (1.6)
<i>Frontiers in Digital Health</i>	16 (3.8)	9 (7.0)
<i>JMIR mHealth and uHealth</i>	14 (3.4)	7 (5.5)
<i>Scientific Reports</i>	12 (2.9)	–
<i>Frontiers in Psychiatry</i>	10 (2.4)	6 (4.7)
Other	270 (65.0)†	76 (59.4)‡
Affiliated country of corresponding authors§		
USA	–	69 (53.9)
Switzerland	–	22 (17.2)
Germany	–	16 (12.5)
UK	–	16 (12.6)
Canada	–	11 (8.6)
France	–	10 (7.8)
Other	–	90 (70.3)
All extracted data are provided in online supplemental S2.		
*Fewer than 2 articles.		
†Fewer than 10 articles.		
‡Fewer than 5 articles.		
§More than 1 category possible.		
.MCI, mild cognitive impairment.		

were dementia and related disorders (16 of 53 articles, ie, (mild) cognitive impairment or Alzheimer's disease), Parkinson's disease (5 of 53 articles) and diabetes (3 of 53 articles), with numerous other conditions addressed in one or two studies (eg, atrial fibrillation, cervical cancer, depression, heart failure and muscular dystrophy; online supplemental S2).

The corresponding authors were mostly from the USA (69 of 128 articles), Switzerland (22 of 128 articles), Germany (16 of 128 articles) and the UK (16 of 128 articles; table 1).

The articles were cited a median of 6 times (range 0–517, IQR 2–20, overall 2,705); on average two times per year (range 0–86, IQR 1–5; online supplemental S2). We show the citation network (ie, citing and cited relationships within the sample of these 128 articles) online (<https://LocalCitationNetwork.github.io/?fromJSON=Digital-Biomarker-Definitions.json>).

Definitions of digital biomarkers

Overall, 128 articles reported between 1 and 7 definitions (median 1, IQR 1–2). In 91 articles, at least 1 reference

Table 3 Definitions of digital biomarkers that include three key components: type of data, data collection method and purpose of a digital biomarker (n=23)

Authors (year), reference	Definition (original quote)	Three key components classification: (1) type of data, (2) data collection method and (3) intended use/purpose
Andrade <i>et al</i> ³²	'Digital biomarkers may have a place as an objective, accurate, and low-cost patient metric to support risk stratification and clinical planning. Digital biomarkers use digital information to objectively measure biological and pathological processes and have the potential to overcome some of the above-mentioned limitations of conventional prognostic tools. Digital data, in particular data from accelerometers and other wearable sensors, are a non-invasive, passively collected low-cost source of individual information. Further exploration of clinical uses for these data may improve clinical decision-making with minimal risk and cost.'	<ol style="list-style-type: none"> 1. '... an objective, accurate, and low-cost patient metric ...' 2. '... use digital information; Digital data, in particular data from accelerometers and other wearable sensors, are a non-invasive, passively collected low-cost source of individual information.' 3. '... to support risk stratification and clinical planning; objectively measure biological and pathological processes; Further exploration of clinical uses for these data may improve clinical decision-making with minimal risk and cost.'
Babrak <i>et al</i> ⁶	'Digital biomarkers are objective, quantifiable, physiological, and behavioral measures that are collected by means of digital devices that are portable, wearable, implantable, or ingestible. These data are often used to explain, influence, and/or predict health-related outcomes. Digital biomarkers fall within the scope of traditional biomarkers in relation to addressing health related questions, with use of a digital and portable technology that adds new dimensions, unique features, and challenges. digital biomarkers are usually less or non-invasive, modular, and often cheaper to measure. They can produce qualitative and quantitative measurements, but most importantly, they provide easier and cheaper access to continuous and longitudinal measurements.'	<ol style="list-style-type: none"> 1. '... objective, quantifiable, physiological, and behavioral measures; Digital biomarkers fall within the scope of traditional biomarkers ...' 2. '... collected by means of digital devices that are portable, wearable, implantable, or ingestible; with use of a digital and portable technology ...' 3. 'These data are often used to explain, influence, and/or predict health-related outcomes; in relation to addressing health related questions ...'
Bartolome and Prioleau ³³	'Digital biomarkers refer to objective, quantifiable physiological, and behavioral measures that are collected by means of digital devices, such as wearable devices, for the purpose of outcomes explaining, influencing, or predicting health. However, unlike traditional biomarkers that provide a "snapshot view" based on limited measurements collected over time, digital biomarkers are often derived from longitudinal and continuous measurements, and thus can capture dynamic changes in health and related outcomes.'	<ol style="list-style-type: none"> 1. '... objective, quantifiable physiological, and behavioral measures ...' 2. '... that are collected by means of digital devices, such as wearable devices ...' 3. '... for the purpose of outcomes explaining, influencing, or predicting health; thus can capture dynamic changes in health and related outcomes.'
Bijlani <i>et al</i> , Nam <i>et al</i> , Parziale and Mascalzoni, Phillips <i>et al</i> , and Wright and Jones ³⁴⁻³⁸	'Digital biomarkers are consumer-generated physiological and behavioral measures collected through connected digital tools that can be used to explain, influence and/or predict health-related outcomes. Health-related outcomes can vary from explaining disease to predicting drug response to influencing fitness behaviors. In our definition of digital biomarkers, we exclude patient-reported measures (eg, survey data), genetic information, and data collected through traditional medical devices and equipment. These data types, though still a key component of research and clinical care that may be stored digitally, are not digitally measured or truly dependent on software.'	<ol style="list-style-type: none"> 1. '... consumer-generated physiological and behavioral measures ...' 2. '... collected through connected digital tools ...' 3. '... can be used to explain, influence and/or predict health-related outcomes. Health-related outcomes can vary from explaining disease to predicting drug response to influencing fitness behaviors.'
Dillenseger <i>et al</i> ³⁹	'... digital biomarkers—digital health technologies— to explain, influence and/or predict health-related outcomes. Digital biomarkers stem is quite broad, and range from wearables that collect patients' activity during digitalized functional tests to digitalized diagnostic procedures and software-supported magnetic resonance imaging evaluation. With the increasing digitalization of healthcare, medicine now gains access to a new type of biomarker. So-called digital biomarkers enable the translation of up-to-date new data sources into informative, actionable knowledge. Digital biomarkers are basically collected by digital tools. Digital biomarkers mean objective, quantifiable physiological and behavioral data that are measured and collected by digital devices. The data collected by, for example, portables, wearables, implantables or ingestibles are typically used to generate, influence and/or predict health-related outcomes, and thus represent deep digital phenotyping, collecting clinically meaningful and objective digital data.'	<ol style="list-style-type: none"> 1. '... objective, quantifiable physiological and behavioral data; represent deep digital phenotyping, collecting clinically meaningful and objective digital data.' 2. '... from wearables that collect patients' activity during digitalized functional tests to digitalized diagnostic procedures and software-supported magnetic resonance imaging evaluation; are basically collected by digital tools; measured and collected by digital devices; data collected by, for example, portables, wearables, implantables or ingestibles ...' 3. '... to explain, influence and/or predict health-related outcomes; typically used to generate, influence and/or predict health-related outcomes ...'
Dorsey <i>et al</i> ⁹	'Digital biomarkers—the use of a biosensor to collect objective data on a biological (eg, blood glucose, serum sodium), anatomical (eg, mole size), or physiological (eg, heart rate, blood pressure) parameter followed by the use of algorithms to transform these data into interpretable outcome measures can help address many of the shortcomings in current measures. These new measures, which include portable (eg, smartphones), wearable, and implantable devices, are by their nature largely independent of raters. They are, therefore, not prone to rater bias. The goal of digital biomarkers is to maximize the ecological validity and temporal and spatial resolution of capturing motor and nonmotor phenomena that are expected to change over time.'	<ol style="list-style-type: none"> 1. '... objective data on a biological (eg, blood glucose, serum sodium), anatomical (eg, mole size), or physiological (eg, heart rate, blood pressure) parameter ...' 2. '... use of a biosensor to collect; portable (eg, smartphones), wearable, and implantable devices ...' 3. '... a biological, anatomical, or physiological parameter; interpretable outcome measures ...'

Continued

Table 3 Continued

Authors (year), reference	Definition (original quote)	Three key components classification: (1) type of data, (2) data collection method and (3) intended use/purpose
Gielis <i>et al</i> ⁴⁰	'Complementary to their biological counterparts, digital biomarkers are "user-generated physiological and behavioral measures collected through connected digital devices to explain, influence and/or predict health-related outcomes.'	<ol style="list-style-type: none"> 1. '... user-generated physiological and behavioral measures ...' 2. '... collected through connected digital devices ...' 3. '... to explain, influence and/or predict health-related outcomes.'
Harms <i>et al</i> ⁴¹	'Digital biomarkers are defined as objective, quantifiable physiological and behavioral data that are collected and measured by means of digital devices. Their use has revolutionized clinical research by enabling high-frequency, longitudinal, and sensitive measurements. Digital biomarkers are that the latter are collected via digital devices and can be collected outside of traditional clinical settings. The digital devices collecting these biomarkers can include wearables, implantables, ingestible devices, and smartphones and tablets. Examples of digital biomarkers are objective consumer-grade data such as voice, temperature, activity, gait, blood oxygen, heart rate, touch, and augmented reality, all collected via mobile and wearable technologies. As opposed to standard clinical measures, digital biomarkers enable high-frequency, longitudinal, and objective measurements, largely independent of the clinical rater. Digital biomarkers can continuously monitor patients to assess therapy response and disease progression without the need for clinical assessment. Moreover, they often exhibit higher sensitivity than traditional clinically used methods, enabling early predictive diagnostics by identifying patients at risk of overt clinical disease.'	<ol style="list-style-type: none"> 1. '... objective, quantifiable physiological and behavioral data ...' 2. '... collected and measured by means of digital devices; collected via digital devices and can be collected outside of traditional clinical settings. The digital devices collecting these biomarkers can include wearables, implantables, ingestible devices, and smartphones and tablets.' 3. '... can continuously monitor patients to assess therapy response and disease progression without the need for clinical assessment; Moreover, they often exhibit higher sensitivity than traditional clinically used methods, enabling early predictive diagnostics by identifying patients at risk of overt clinical disease.'
Hartl <i>et al</i> ⁴²	'Digital biomarkers are defined as physiological and behavioral measures collected via digital devices (such as portables, wearables, implantables and digestibles) that characterize, influence, or predict health-related outcomes. Digital biomarkers offer several potential advantages compared to traditional clinical assessments. Digital biomarker products are usually the result of the combination of multiple individual hardware (sensors) and software (operating systems and algorithms) components. Digital biomarkers as clinical endpoints provide objective and quantitative measures yet still require broader clinical use and health authority acceptance.'	<ol style="list-style-type: none"> 1. '... physiological and behavioral measures; clinical endpoints provide objective and quantitative measures ...' 2. '... collected via digital devices (such as portables, wearables, implantables and digestibles) ...' 3. '... that characterize, influence or predict health-related outcomes.'
Hartl <i>et al</i> ⁴²	'Digital biomarkers: Physiological and behavioral measures collected by means of digital devices such as portables, wearables, implantables, or digestibles that characterize, influence, or predict health-related outcomes.'	<ol style="list-style-type: none"> 1. 'Physiological and behavioral measures ...' 2. '... collected by means of digital devices such as portables, wearables, implantables, or digestibles ...' 3. '... that characterize, influence, or predict health-related outcomes.'
Katsaros <i>et al</i> ⁴³	'Digital biomarkers are objective measurements of physiological, pathologic, or anatomic characteristics continuously collected outside the clinical environment via home-based connected devices. Passively collecting data from patients' mobile or wearable devices potentially offers a convenient and unobtrusive method to prospectively identify psychosocial burden and deliver tailored social support to the right patients at the right time.'	<ol style="list-style-type: none"> 1. '... objective measurements of physiological, pathologic, or anatomic characteristics ...' 2. '... continuously collected outside the clinical environment via home-based connected devices; Passively collecting data from patients' mobile or wearable devices.' 3. '... offers a convenient and unobtrusive method to prospectively identify psychosocial burden and deliver tailored social support to the right patients at the right time.'
Motahari-Nezhad <i>et al</i> ⁴⁴	'Sensors and digital devices have revolutionized the measurement, collection, and storage of behavioral and physiological data, leading to the new term digital biomarkers. Digital biomarkers are measured across multiple layers of the hardware (eg, sensors) and software of medical devices that capture signals (behavioral and physiological data) from patients. Digital biomarkers can increase diagnostic and therapeutic precision in the modern health care system by remotely and continuously measuring reliable clinical data and allowing continuous monitoring and evaluation. Captured by wearable, implantable, and ingestible devices and sensors, digital biomarkers can be used at home to provide clinical data, collecting data that is not possible in the clinical setting. This information can improve physicians' and patients' decisions, personalize the treatment, and predict diseases' current and future status.'	<ol style="list-style-type: none"> 1. '...behavioral and physiological data; signals (behavioral and physiological data) from patients; remotely and continuously measuring reliable clinical data ...' 2. 'Sensors and digital devices have revolutionized the measurement, collection, and storage; measured across multiple layers of the hardware (eg, sensors) and software of medical devices; Captured by wearable, implantable, and ingestible devices and sensors ...' 3. '... increase diagnostic and therapeutic precision in the modern health care system; allowing continuous monitoring and evaluation; used at home to provide clinical data, collecting data that is not possible in the clinical setting; This information can improve physicians' and patients' decisions, personalize the treatment, and predict diseases' current and future status.'
Nam <i>et al</i> ³⁵	'In terms of IoT, the digital biomarker represents digitized data acquired from patients via IoT devices. Therefore, the digital biomarker can be defined as a biomarker that is objectively and quantitatively measured using digital devices and be used to explain or predict health-related outcomes. Digital biomarker is measured using the digital tools that include portable, wearable, implantable or ingestible devices, and exclude data obtained via patient-reported measurements or traditional devices and equipment. In a broad sense, digital biomarker include all human data that can be measured using digital tool.'	<ol style="list-style-type: none"> 1. '... digitized data; a biomarker that is objectively and quantitatively measured; digital biomarker include all human data ...' 2. '... acquired from patients via IoT devices; using digital devices; measured using the digital tools that include portable, wearable, implantable or ingestible devices, and exclude data obtained via patient-reported measurements or traditional devices and equipment; measured using digital tool.' 3. '... used to explain or predict health-related outcomes.'

Continued

Table 3 Continued

Authors (year), reference	Definition (original quote)	Three key components classification: (1) type of data, (2) data collection method and (3) intended use/purpose
Palanica <i>et al</i> ⁴⁵	'Digital biomarkers are digitally collected data, such as heart rate from a wearable device, that are transformed through mathematical models into indicators of health outcomes like prediabetes. Some digital biomarkers have been found to outperform traditional clinical methods, for example, for arrhythmia detection, because of their ability to continuously monitor patients outside of the clinic. The most successful digital biomarkers have been developed based on supervised, unsupervised, and semi-supervised machine learning models.'	<ol style="list-style-type: none"> 1. '... digitally collected data ...' 2. '... from a wearable device; developed based on supervised, unsupervised and semi-supervised machine learning models.' 3. '... indicators of health outcomes like prediabetes.'
Petersen <i>et al</i> ⁴⁶	'The use of remotely collected data that monitors health and behavior is an emerging area of research. Such data could be considered digital biomarkers objective information that can be used to predict changes in health status and the use of digital biomarkers offers a more efficient method of identifying such markers as the use of devices continuously collecting data increases. One critical requirement in the development of digital biomarkers is connecting these novel measurements to health outcomes.'	<ol style="list-style-type: none"> 1. '... remotely collected data; objective information; novel measurements ...' 2. '... devices continuously collecting data ...' 3. '... monitors health and behavior; can be used to predict changes in health status; health outcomes.'
Phillips <i>et al</i> ⁴⁷	'Digital biomarker technologies, which fall into the category of 'wearables and biosensing devices', use consumer-generated physiological and behavioral measures collected through connected digital tools that can be used to explain, influence, and/or predict health-related outcomes. These technologies may focus on measurements for consumer use only, or clinical measurements that are transmitted to clinicians for health care decision-making. They may passively monitor ongoing activities (such as steps taken) or be used to actively collect specific measurements (such as blood glucose).'	<ol style="list-style-type: none"> 1. '... consumer-generated physiological and behavioral measures ...' 2. '... technologies, which fall into the category of "wearables and biosensing devices"; collected through connected digital tools ...' 3. '... can be used to explain, influence, and/or predict health-related outcomes; These technologies may focus on measurements for consumer use only, or clinical measurements that are transmitted to clinicians for health care decisionmaking; They may passively monitor ongoing activities or be used to actively collect specific measurements ...'
Piau <i>et al</i> ⁴⁷	'Digital biomarker definition. Objective, quantifiable, physiological, and/or behavioral data that are collected and measured by means of digital devices such as embedded environmental sensors, portables, wearables, implantables, or ingestibles, and which opens up opportunities for the remote collection and processing of ecologically valid, real-life, continuous, long-term, health-related data.'	<ol style="list-style-type: none"> 1. 'Objective, quantifiable, physiological, and/or behavioral data ...' 2. '... collected and measured by means of digital devices such as embedded environmental sensors, portables, wearables, implantables, or ingestibles ...' 3. '... which opens up opportunities for the remote collection and processing of ecologically valid, real-life, continuous, long-term, health-related data.'
Sahandi Far <i>et al</i> ⁴⁸	'Digital biomarkers (DB), as captured using sensors embedded in modern smart devices, are a promising technology for home-based sign and symptom monitoring in Parkinson disease (PD). The emergence of new technologies has led to a variety of sensors (ie, acceleration, gyroscope, GPS, etc) embedded in smart devices for daily use (ie, smartphone, smartwatch). Such sensor data, alongside other digital information recorded passively or when executing prespecified tasks, may provide valuable insight into health-related information. Such applications are now commonly referred to as digital biomarkers (DB). DB being collected frequently over a long period of time can provide an objective, ecologically valid, and more detailed understanding of the inter- and intra-individual variability in disease manifestation in daily life.'	<ol style="list-style-type: none"> 1. '... sensor data; objective, ecologically valid, and more detailed understanding of the inter- and intra-individual variability ...' 2. '... captured using sensors embedded in modern smart devices; alongside other digital information recorded passively or when executing prespecified tasks; The emergence of new technologies has led to a variety of sensors (ie, acceleration, gyroscope, GPS, etc) embedded in smart devices for daily use (ie, smartphone, smartwatch).' 3. '... promising technology for home-based sign and symptom monitoring in Parkinson disease (PD); may provide valuable insight into health-related information; disease manifestation in daily life.'
Seyhan and Carini ⁴⁹	'Digital biomarkers (BMs) can have several applications beyond clinical trials in diagnostics—to identify patients affected by a disease or to guide treatment. Digital BMs present a big opportunity to measure clinical endpoints in a remote, objective, and unbiased manner. Digital BMs are defined as an objective, quantifiable physiological and behavioral data that are collected and measured by means of digital devices. The data collected is typically used to explain, influence and/or predict health-related outcomes.'	<ol style="list-style-type: none"> 1. '... measure clinical endpoints; objective, quantifiable physiological and behavioral data; remote, objective and unbiased manner.' 2. '... collected and measured by means of digital devices.' 3. '... can have several applications beyond clinical trials in diagnostics—to identify patients affected by a disease or to guide treatment; The data collected is typically used to explain, influence and/or predict health-related outcomes.'
Shandhi <i>et al</i> ⁵⁰	'Multiple studies suggest the utility of digital biomarkers, objective and quantifiable digitally collected physiological and behavioral data (eg, resting heart rate (RHR), step count, sleep duration, and respiratory rate), collected by consumer devices along with patient-reported symptoms to monitor the progression of respiratory and influenza-like illnesses.'	<ol style="list-style-type: none"> 1. '...objective and quantifiable digitally collected physiological and behavioral data ...' 2. '... collected by consumer devices along with patient-reported symptoms ...' 3. '... to monitor the progression of respiratory and influenza-like illnesses.'
Tavabi <i>et al</i> ⁵¹	'Digital biomarkers are physiological and behavioral measures collected from participants through digital tools that can be used to explain, influence, or predict health-related outcomes.'	<ol style="list-style-type: none"> 1. '... physiological and behavioral measures ...' 2. '... collected from participants through digital tools ...' 3. '... can be used to explain, influence, or predict health-related outcomes.'

Continued

Table 3 Continued

Authors (year), reference	Definition (original quote)	Three key components classification: (1) type of data, (2) data collection method and (3) intended use/purpose
van den Brink <i>et al</i> ⁶²	'Wearable technologies, including smartphones and smartwatches, are increasingly utilized in the healthcare domain for the development of so-called digital biomarkers. This novel type of biomarker is characterized by being measured non-invasively, continuously, and under real-world conditions using digital technology, allowing for a more holistic and personal insight into someone's health. Therefore, digital biomarkers enable accessible health and behavioral feedback to the user and are particularly suited for driving the healthcare transition towards prevention, empowering people in the self-management of health and disease. Furthermore, digital biomarkers can provide users with more frequent and detailed contextual information and continuously update personal lifestyle recommendations.'	<ol style="list-style-type: none"> 1. '... type of biomarker is characterized by being measured non-invasively, continuously, and under real-world conditions ...' 2. 'Wearable technologies, including smartphones and smartwatches, are increasingly utilized in the healthcare domain for the development; using digital technology ...' 3. '...allowing for a more holistic and personal insight into someone's health; accessible health and behavioral feedback to the user and are particularly suited for driving the healthcare transition towards prevention, empowering people in the self-management of health and disease; can provide users with more frequent and detailed contextual information and continuously update personal lifestyle recommendations.'
Zetterström <i>et al</i> ⁶³	'We define a DB as patient-generated physiological and behavioural measures collected through sensors and other connected digital tools that can be used to monitor, predict and/or influence health-related outcomes.'	<ol style="list-style-type: none"> 1. '...patient-generated physiological and behavioural measures ...' 2. '... collected through sensors and other connected digital tools ...' 3. '... monitor, predict and/or influence health-related outcomes.'

reference (for those with a reference, similarities such as paraphrasing are expected; online supplemental S4).

DISCUSSION

We systematically searched and characterised the biomedical literature that used the term digital biomarker and analysed the provided definitions of the concept. We identified 415 articles using 'digital biomarker' in title and/or abstract that were published between 2014 and 2023. Of them, 128 articles provided 127 different definitions. By comparing the defining features, we aimed to better understand what those who use this term in the context of biomedical research or healthcare mean by 'digital biomarker' and which components are deemed the essence of it.²⁶

The first definition of a digital biomarker is from 2015.²⁷ Within 8 years, more than 127 definitions have been used, with none of them clearly being the most widely used; indicating a high heterogeneity of the concept of digital biomarkers. The definitions often cover different aspects of definitional components that are traditionally used to describe more conventional biomarkers. Authors have created their own concepts and gave an identity to this type of biomarker. The variation in these definitions and the fact that only 23 of them provide a full description containing all components of FDA's BEST framework, shows how broad the current understanding of this fundamental concept is.

Digital biomarkers emerged as a concept in medical and technological domains, although with a diverse terminology across different academic journals. In the medical field, digital biomarkers are often referred to as biomarkers of health or disease obtained through digital health technologies. In the technical field, these biomarkers are viewed as data-driven indicators collected from sensors, wearables and other portable

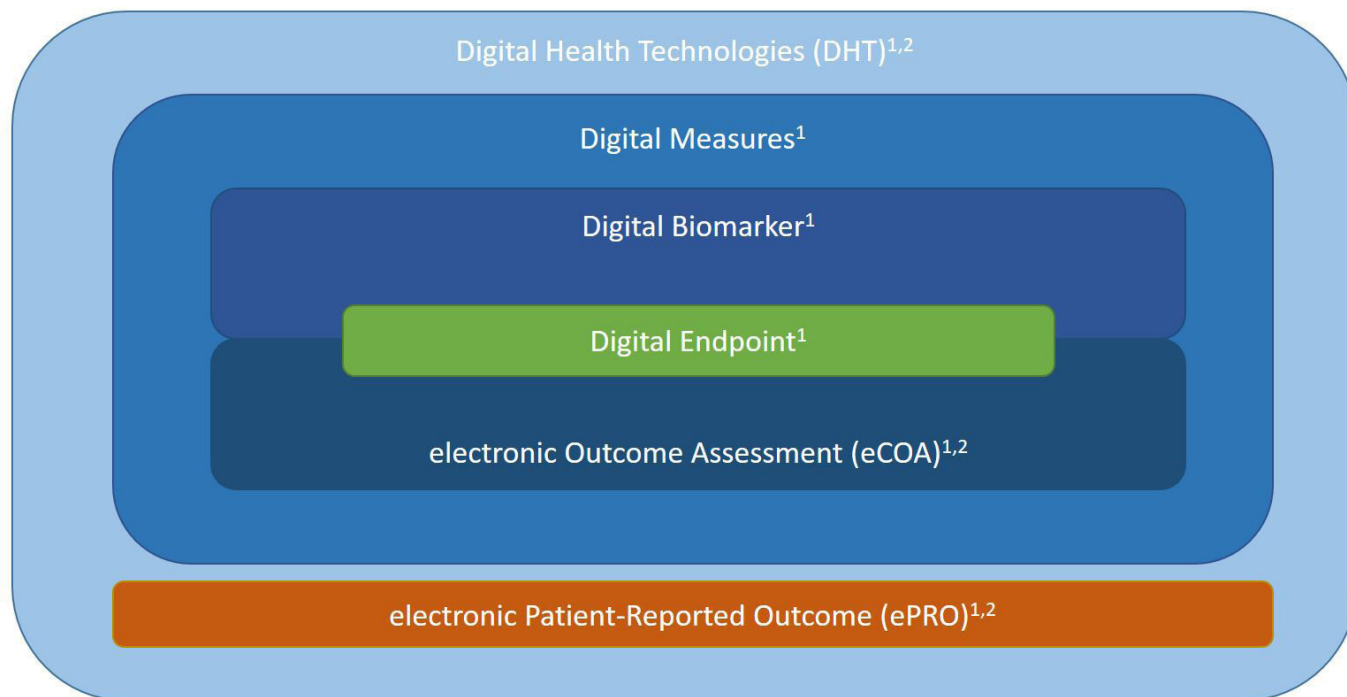
digital technologies that provide an assessment of the health status. These diverse terminologies and definitions reflect the interdisciplinary nature of digital biomarkers with their application in a broad spectrum of biomedicine which underlines the importance of unified concepts to enhance the communications and cross-disciplinary collaborations on this evolving field.

Regulatory perspectives

The EMA has defined digital biomarkers in 2020 in their draft guidance 'Questions and answers: Qualification of digital technology-based methodologies to support approval of medicinal products', stating their 'clinical meaning is established by a reliable relationship to an existing, validated endpoint'.¹⁰ EMA draws a clear line to electronic clinical outcome assessments (eCOA), whose 'clinical meaning is established de novo'. According to EMA's terminology, both digital biomarkers and eCOA are derived from 'digital measures' and can be used as 'digital endpoints'.¹⁰

On the other hand, the term 'digital biomarker' cannot be found in the FDA draft guidance 'Digital Health Technologies for Remote Data Acquisition in Clinical Investigations', which instead features eCOA as examples of digital health technologies.²⁸ Figure 3 contains our semantic interpretation of the terminology used by EMA and FDA.

This distinction can rarely be observed in the medical literature—we found this term in 8 of the 415 articles analysed and a PubMed search for 'electronic clinical outcome assessment*' returned also only 8 articles mentioning it in title or abstract (as of 31 August 2023), compared with the 415 for our search term 'digital biomarker*'. As Vasudevan *et al* stated in 2022: 'There are currently multiple definitions of the term digital biomarker reported in the scientific literature, and some seem to conflate established definitions of a biomarker and a clinical outcomes assessment (COA)'.¹¹



- (1) EMA: https://www.ema.europa.eu/en/documents/other/questions-answers-qualification-digital-technology-based-methodologies-support-approval-medical_en.pdf
 (2) FDA: <https://www.fda.gov/media/155022/download>

Figure 3 Semantic overview of terminology used by EMA and FDA. Digital health technologies obtain digital measures, which include digital biomarkers and electronic clinical outcome assessment (eCOA). Digital biomarkers and eCOAs both can provide digital endpoints. EMA, European Medicines Agency; FDA, Food and Drug Administration.

This divergency in the terminology of digital biomarkers between the academic literature and the regulators' language raises challenges and ambiguity. Consequently, a more cohesive and comprehensive framework within the digital biomarker field is needed to strengthen the clarity and continue growing the potential that this data could bring for health.

The development of a substantive and unified definition of digital biomarkers would be an important step in shaping a conceptual framework for the development, assessment and reporting of digital biomarkers. Our results may inform this process by using the existing understanding of digital biomarkers systematically analysed in this study as a basis. To achieve a common and more unified understanding of what digital biomarkers are—and are not—a Delphi study could be useful.^{29 30} Such a study would aim to combine multiple views and expectations on the existing definitions of digital biomarkers and their components until a consensus is reached. Ideally, that would be achieved by an international panel with expert's representative of all relevant stakeholders covering a range of medical fields (eg, cardiology, neurology), professional backgrounds (eg, clinical care/rehabilitation/nursing, software developers, device manufacturer, editors, guideline developers), and professional perspectives (eg, academia, regulatory, industry/technology, publishing) and involving patients.

Limitations

There are some limitations to our study.

First, we used a limited search only in a single database using the single term of 'digital biomarker*', which may have overlooked some other relevant studies. PubMed was chosen as literature database given its outstanding role, reflecting the most impactful journals in biomedicine.³¹ We focused on this single term because we assume it to be the most central and widely used term describing the concept of 'digital biomarker'. It is very unlikely that the definitions would be much more uniform in potentially overlooked studies or would we have included other potential concepts, and it is quite possible that many more different definitions would emerge, especially from digital biomarker developments contained in technical literature databases (such as IEEE Explore or ACM Digital Library). Therefore, we may have even underestimated the large number of different definitions.

Second, the screening and data extraction were performed by a single reviewer only. This may have resulted in some studies that were overlooked and some misclassifications, but it is unlikely that our main interpretation would change. Third, we developed a simple framework with three key elements of definitions based on a well-established framework (BEST), but the categorisation of elements is subjective to some degree. However,

we employed a structured analysis that confirmed the observed heterogeneity across definitions.

CONCLUSIONS

Clear and unambiguous communication and research reporting is essential for the effective implementation of scientific innovations and developments. This requires clear definitions and consistent use and understanding of key terms and concepts. A lack of clarity and consistency can lead to research waste, delay or even misdirection of promising developments and potential. Digital biomarkers offer the opportunity to collect objective, meaningful, patient-relevant data cost-effectively with unprecedented granularity. An exact understanding of what they are and how they are described in biomedical literature is essential to let them shape the future of clinical research and enable them to provide most useful evidence for research and care. Our study can inform the development of a harmonised and more widely accepted definition, for example, with a Delphi study.

Author affiliations

¹Department of Applied Natural Sciences, Technische Hochschule Lübeck, Lübeck, Germany

²Pragmatic Evidence Lab, Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland

³Department of Clinical Research, University Hospital Basel and University of Basel, Basel, Switzerland

⁴Department of Health, Eastern Switzerland University of Applied Sciences, St. Gallen, Switzerland

⁵Department of Neurology and MS Center, University Hospital Basel and University of Basel, Basel, Switzerland

⁶Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, USA

⁷Meta-Research Innovation Center Berlin (METRIC-B), Berlin Institute of Health, Berlin, Germany

Acknowledgements We thank Saido Haji Abukar (Medical Bachelor student at ETH Zurich) for her support with the study selection and data extraction.

Contributors All authors made substantial contributions to the conception and design of the work; all authors have drafted the work or substantively revised it; all authors have approved the submitted version; all authors have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved and the resolution documented in the literature.

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 RC2NB (Research Center for Clinical Neuroimmunology and Neuroscience Basel) is supported by Foundation Clinical Neuroimmunology and Neuroscience Basel. One of the main projects of RC2NB is the development and evaluation of a digital biomarkers which is supported by grants from Novartis, Roche and Innosuisse (Swiss Innovation Agency). All authors declare no competing interests.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as online supplemental information.

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 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iD

Julian Hirt <http://orcid.org/0000-0001-6589-3936>


REFERENCES

- 1 FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and Other Tools) Resource. Silver Spring (MD): Food and Drug Administration (US), 2016.
- 2 Motahari-Nezhad H, Péntek M, Gulácsi L, *et al*. Outcomes of digital biomarker-based interventions: protocol for a systematic review of systematic reviews. *JMIR Res Protoc* 2021;10:e28204.
- 3 Moschovis PP, Sampayo EM, Cook A, *et al*. The diagnosis of respiratory disease in children using a phone-based cough and symptom analysis algorithm: the Smartphone recordings of cough sounds 2 (SMARTCOUGH-C 2) trial design. *Contemp Clin Trials* 2021;101:106278.
- 4 Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ Digit Med* 2019;2:14.
- 5 Wolz R, Munro J, Guerrero R, *et al*. [P3–200]: predicting sleep/wake patterns from 3-axis accelerometry using deep learning. *Alzheimer's & Dementia* 2017;13.
- 6 Babrak LM, Menetski J, Rebhan M, *et al*. Traditional and digital biomarkers: two worlds apart *Digit Biomark* 2019;3:92–102.
- 7 Buegler M, Harms R, Balasa M, *et al*. Digital biomarker-based individualized prognosis for people at risk of dementia. *Alzheimers Dement (Amst)* 2020;12:e12073.
- 8 Woelfle T, Bourguignon L, Lorscheider J, *et al*. Wearable sensor technologies to assess motor functions in people with multiple sclerosis: systematic scoping review and perspective. *J Med Internet Res* 2023;25:e44428.
- 9 Dorsey ER, Papapetropoulos S, Xiong M, *et al*. The first frontier: digital biomarkers for neurodegenerative disorders. *Digit Biomark* 2017;1:6–13.
- 10 European Medicines Agency. Questions and answers: qualification of Digital technology-based Methodologies to support approval of medicinal products 2020. 2022. Available: https://www.ema.europa.eu/en/documents/other/questions-answers-qualification-digital-technology-based-methodologies-support-approval-medicinal_en.pdf
- 11 Vasudevan S, Saha A, Tarver ME, *et al*. Digital biomarkers: convergence of digital health technologies and biomarkers. *NPJ Digit Med* 2022;5:36.
- 12 Cook DJ, Reeve BK, Guyatt GH, *et al*. Stress ulcer prophylaxis in critically ill patients. resolving discordant meta-analyses. *JAMA* 1996;275:308–14.
- 13 Oliveira ML, Lucchetta RC, Bonetti A de F, *et al*. Efficacy outcomes reported in trials of multiple sclerosis: a systematic Scoping review. *Mult Scler Relat Disord* 2020;45:102435.
- 14 Soriano JB, Murthy S, Marshall JC, *et al*. A clinical case definition of post-COVID-19 condition by a Delphi consensus. *Lancet Infect Dis* 2022;22:e102–7.
- 15 Verghis R, Blackwood B, McDowell C, *et al*. Heterogeneity of surrogate outcome measures used in critical care studies: a systematic review. *Clin Trials* 2023;20:307–18.
- 16 Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Info Libr J* 2009;26:91–108.
- 17 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *J Clin Epidemiol* 2021;134:178–89.
- 18 OpenAlex. Openalex 2023. Available: <https://openalex.org/> [Accessed 16 Jun 2023].

- 19 Woelfle T. Local citation network 2019. Available: <https://LocalCitationNetwork.github.io> [Accessed 16 Jun 2023].
- 20 Zygomatic. Wordclouds.Com 2023. Available: <https://www.wordclouds.com/> [Accessed 25 Apr 2023].
- 21 Bachmann M. RapidFuzz 3.2.0: Indel 2021, Available: <https://maxbachmann.github.io/RapidFuzz/Usage/distance/Indel.html> [Accessed 3 Aug 2023].
- 22 Priem J, Piwowar H, Orr R. Openalex: a fully-open index of scholarly works, authors, venues, institutions, and concepts 2022. n.d. Available: <https://arxiv.org/abs/2205.01833>
- 23 Califf RM. Biomarker definitions and their applications. *Exp Biol Med (Maywood)* 2018;243:213–21.
- 24 Piau A, Wild K, Mattek N, et al. Current state of digital biomarker technologies for real-life, home-based monitoring of cognitive function for mild cognitive impairment to mild Alzheimer disease and implications for clinical care. *J Med Internet Res* 2019;21:e12785.
- 25 Coravos A, Goldsack JC, Karlin DR, et al. Digital medicine: a primer on measurement. *Digit Biomark* 2019;3:31–71.
- 26 Gupta A, Mackereth S. Definitions: the stanford encyclopedia of philosophy 2023, Available: <https://plato.stanford.edu/entries/definitions/> [Accessed 31 Jan 2024].
- 27 Gerbelot R, Koenig A, Goyer C, et al. A Wireless patch for sleep respiratory disorders applications. *Annu Int Conf IEEE Eng Med Biol Soc* 2015;2015:2279–82.
- 28 Food and Drug Administration (US). Digital health Technologies for remote data acquisition in clinical investigations. FDA; 2021. Available: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations> [Accessed 20 Jul 2021].
- 29 Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. *Information & Management* 2004;42:15–29.
- 30 Khodyakov D, Grant S, Kroger J, et al. Disciplinary trends in the use of the Delphi method: a bibliometric analysis. *PLoS One* 2023;18:e0289009.
- 31 National Cancer Institute at the National Institutes of Health. NCI dictionary of cancer terms: biomedicine 2024. Available: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomedicine> [Accessed 5 Feb 2025].
- 32 Andrade AQ, Lim R, Kelly T-L, et al. Wrist accelerometer temporal analysis as a prognostic tool for aged care residents: a sub-study of the remindar trial. *J Am Geriatr Soc* 2023;71:1124–33.
- 33 Bartolome A, Prioleau T. A computational framework for discovering digital biomarkers of glycemic control. *NPJ Digit Med* 2022;5:111.
- 34 Bijlani N, Nilforooshan R, Kouchaki S. An unsupervised data-driven anomaly detection approach for adverse health conditions in people living with dementia: cohort study. *JMIR Aging* 2022;5:e38211.
- 35 Nam KH, Kim DH, Choi BK, et al. Internet of things, digital biomarker, and artificial intelligence in spine: current and future perspectives. *Neurospine* 2019;16:705–11.
- 36 Parziale A, Mascalonzi D. Digital biomarkers in psychiatric research: data protection qualifications in a complex ecosystem. *Front Psychiatry* 2022;13:873392.
- 37 Phillips KA, Douglas MP, Trosman JR, et al. What goes around comes around": lessons learned from economic evaluations of personalized medicine applied to digital medicine. *Value Health* 2017;20:47–53.
- 38 Wright JM, Jones GB. Harnessing the digital exhaust: incorporating wellness into the pharma model. *Digit Biomark* 2018;2:31–46.
- 39 Dillenseger A, Weidemann ML, Trentzsch K, et al. Digital biomarkers in multiple sclerosis. *Brain Sci* 2021;11:1519:11:.
- 40 Gielis K, Vanden Abeele M-E, De Croon R, et al. Dissecting digital card games to yield digital biomarkers for the assessment of mild cognitive impairment: methodological approach and exploratory study. *JMIR Serious Games* 2021;9:e18359.
- 41 Harms RL, Ferrari A, Meier IB, et al. Digital biomarkers and sex impacts in Alzheimer's disease management - potential utility for innovative 3p medicine approach. *EPMA J* 2022;13:299–313.
- 42 Hartl D, de Luca V, Kostikova A, et al. Translational precision medicine: an industry perspective. *J Transl Med* 2021;19:245.
- 43 Katsaros D, Hawthorne J, Patel J, et al. Optimizing social support in oncology with digital platforms. *JMIR Cancer* 2022;8:e36258.
- 44 Motahari-Nezhad H, Fgaier M, Mahdi Abid M, et al. Digital biomarker-based studies: scoping review of systematic reviews. *JMIR Mhealth Uhealth* 2022;10:e35722.
- 45 Palanica A, Docktor MJ, Lieberman M, et al. The need for artificial intelligence in digital therapeutics. *Digit Biomark* 2020;4:21–5.
- 46 Petersen CL, Weeks WB, Norin O, et al. Development and implementation of a person-centered, technology-enhanced care model for managing chronic conditions: cohort study. *JMIR Mhealth Uhealth* 2019;7:e11082.
- 47 Piau A, Rumeau P, Nourhashemi F, et al. Information and communication technologies, a promising way to support pharmacotherapy for the behavioral and psychological symptoms of dementia. *Front Pharmacol* 2019;10:1122.
- 48 Sahandi Far M, Eickhoff SB, Goni M, et al. Exploring test-retest reliability and longitudinal stability of digital biomarkers for Parkinson disease in the M-power data set: cohort study. *J Med Internet Res* 2021;23:e26608.
- 49 Seyhan AA, Carini C. Are innovation and new technologies in precision medicine paving a new era in patients centric care *J Transl Med* 2019;17:114.
- 50 Shandhi MMH, Cho PJ, Roghanizad AR, et al. A method for intelligent allocation of diagnostic testing by leveraging data from commercial wearable devices: a case study on COVID-19. *NPJ Digit Med* 2022;5:130.
- 51 Tavabi N, Stück D, Signorini A, et al. Cognitive digital biomarkers from automated transcription of spoken language. *J Prev Alzheimers Dis* 2022;9:791–800.
- 52 van den Brink WJ, van den Broek TJ, Palmisano S, et al. Digital biomarkers for personalized nutrition: predicting meal moments and interstitial glucose with non-invasive, wearable technologies. *Nutrients* 2022;14:4465.
- 53 Zetterström A, Hämäläinen MD, Karlberg E, et al. Maximum time between tests: a digital biomarker to detect therapy compliance and assess schedule quality in measurement-based eHealth systems for alcohol use disorder. *Alcohol Alcohol* 2019;54:70–2.

© 2024 Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY. Published by BMJ. <https://creativecommons.org/licenses/by/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Generative artificial intelligence and non-pharmacological bias: an experimental study on cancer patient sexual health communications

Akiko Hanai ^{1,2}, Tetsuo Ishikawa,^{1,2,3,4} Shoichiro Kawauchi,¹ Yuta Iida,¹ Eiryō Kawakami^{1,2}

To cite: Hanai A, Ishikawa T, Kawauchi S, *et al*. Generative artificial intelligence and non-pharmacological bias: an experimental study on cancer patient sexual health communications. *BMJ Health Care Inform* 2024;**31**:e100924. doi:10.1136/bmjhci-2023-100924

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2023-100924>).

Received 03 October 2023
Accepted 09 March 2024

ABSTRACT

Objectives The objective of this study was to explore the feature of generative artificial intelligence (AI) in asking sexual health among cancer survivors, which are often challenging for patients to discuss.

Methods We employed the Generative Pre-trained Transformer-3.5 (GPT) as the generative AI platform and used DocsBot for citation retrieval (June 2023). A structured prompt was devised to generate 100 questions from the AI, based on epidemiological survey data regarding sexual difficulties among cancer survivors. These questions were submitted to Bot1 (standard GPT) and Bot2 (sourced from two clinical guidelines).

Results No censorship of sexual expressions or medical terms occurred. Despite the lack of reflection on guideline recommendations, 'consultation' was significantly more prevalent in both bots' responses compared with pharmacological interventions, with ORs of 47.3 ($p<0.001$) in Bot1 and 97.2 ($p<0.001$) in Bot2.

Discussion Generative AI can serve to provide health information on sensitive topics such as sexual health, despite the potential for policy-restricted content. Responses were biased towards non-pharmacological interventions, which is probably due to a GPT model designed with the 's prohibition policy on replying to medical topics. This shift warrants attention as it could potentially trigger patients' expectations for non-pharmacological interventions.

INTRODUCTION

With the recent development of generative artificial intelligence (AI), particularly large language models which uses billions of parameters, a growing discussion exists about its usefulness and risks as a healthcare tool.¹ Generative AI is expected to facilitate cross-cultural communication between patients with real-life experiences and medical professionals with rich medical knowledge. However, disadvantages such as bias in training data, a proliferation of false, harmful responses and ambiguous reasoning behind responses have been pointed out to using AI-generated information in healthcare.¹

Although cancer survivors have sexual problems, they are particularly hard to communicate

between patients and healthcare providers.² Clinical guidelines provide practical ways to deal with sexual problems, and the first step is to connect the patient to a medical consultation.^{3,4} However, it is difficult for patients to confess their sexual problems to the doctor before them, and we hypothesised that patients would initially consult AI about this difficult-to-convey issue. For the Generative Pre-trained Transformer (GPT), the policy of its provider (Open AI) states that the model is not fine-tuned to provide medical information or adult content.⁵ We performed a generative experiment with a hypothetical cancer survivor to examine the characteristics of medical and sexual consultations, two areas not covered in this fine-tuning.

METHODS

We conducted a dialogue generation experiment and performed an exploratory analysis. We used GPT-3.5 (Open AI) as the generative AI and DocsBot (docsbot.ai) to refer to specific documents (the latest version as of June 2023 in Japanese). The prompt 'I am a cancer survivor. Please create a question about a problem that is hard to consult' generated 100 questions by DocsBot that had learnt a survey on sexual problems among cancer survivors.⁶ The generated questions were categorised into seven topics based on the symptom categories specified in the clinical guidelines: sexual response, body image, intimacy, sexual functioning, vasomotor symptoms, genital symptoms and others. These questions were presented to Bot1 (standard GPT) and Bot2 (sourced from two clinical guidelines^{3,4}).

The collected conversational data from Bot1 and Bot2 were tokenised into individual words, and linguistic features were extracted from the text data, including lemmatised and stop-word-removed text, noun phrases as keywords and



© 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.

For numbered affiliations see end of article.

Correspondence to

Dr Akiko Hanai;
hanaaki0803@gmail.com

verb lemmas. We then calculated a similarity score between the responses from Bot1 and Bot2 using word vectors to measure semantic similarity. Frequency and sentiment were also analysed. Fisher's exact test was used to compare the rate of non-pharmacological and pharmacological interventions which were defined by GPT. We used Python V.3.11 package (sklearn and spaCy) for the analyses. All data and codes can be obtained from this link https://github.com/AkikoHanai/LLM_CancerConsul_Trial

RESULTS

The topics of the generated questions were, in order of frequency, sexual functioning (N=24), sexual response (N=13), body image (N=17), intimacy (N=8) and others (N=38), including general lifestyle or health check-up in cancer survivorship (online supplemental file 1). The mean similarity score between Bot1 and Bot2 responses was 0.93 (ranging from 0.77 to 0.98); the less the guideline mentioned the topics, the lower its concordance rate. For sexual response and sexual function, the guidelines recommended both pharmacological and non-pharmacological intervention for them, non-pharmacological intervention (counselling) was significantly more frequently than pharmacological intervention (OR=47.3 in Bot1 (95% CI 9.55 to 233.81, $p<0.001$), 97.2 in Bot2 (95% CI 11.72 to 806.04, $p<0.001$)). Sentiment analysis showed a slightly positive polarity (Bot1 mean=0.18 (SD=0.12), Bot2 mean=0.19 (SD=0.15)).

DISCUSSION

When disseminating information about cancer treatment and sexual health issues faced by cancer survivors, the generated chatbots functioned without refusing to answer, with or without training sources of medical guidelines. GPT responses have been noted to be as reliable as web searches and are closer to clinical guidelines, making it a promising tool to support medical communication.^{7,8} In this study, the GPT returned useful results comparable to the guidelines, not calling for excessive pessimism or optimism. However, GPT-based questions and answers tended to return counselling over pharmacological treatment options, with many responses encouraging consultation with medical staff. The advertising policies of consumer search engines or usage policy of Open AI limit the accessibility of information about medical contents or specific drugs, depending on the legal restrictions in each country. As a result, the AI created may have been biased toward medical consultation, which lies within the realm of 'specific medical information' subject to such legal restrictions.

Given the potential use of generative AI to address issues that patients may be hesitant to discuss with medical staff, such as sexual issues, generative AI may help patients clarify their concerns and facilitate shared decision-making. The limitations of this study include adjustments of the prompts and no actual trial with patients or providers to maintain that reliability or validity—however, the situation in which bias due to medical information regulations likely to be universal.

Healthcare providers would need to consider the possibility that patients who use consumer web tools, including generative AI, may have expectations for non-pharmacological interventions such as counsellings.

Author affiliations

¹Medical Data Mathematical Reasoning Team, Advanced Data Science Project, RIKEN Information R&D and Strategy Headquarters, RIKEN, Yokohama, Japan

²Department of Artificial Intelligence Medicine, Graduate School of Medicine, Chiba University, Chiba, Japan

³Department of Extended Intelligence for Medicine, The Ishii-Ishibashi Laboratory, Keio University School of Medicine, Tokyo, Japan

⁴Collective Intelligence Research Laboratory, Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan

Acknowledgements We thank Manami Kato and Ayaka Mori for their assistance with data management. This work was supported by a RIKEN grant.

Contributors All authors proposed the study concept and reviewed the overall direction and manuscript; AH designed the model and computational framework, analysed the data and wrote the manuscript.

Funding This study was supported in part by RIKEN, which was not involved in the study design, data collection and analysis, decision to publish or manuscript preparation.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

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

Akiko Hanai <http://orcid.org/0000-0003-4468-1488>

REFERENCES

- Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023;6:120.
- Reese JB, Soric K, Beach MC, *et al*. Patient-provider communication about sexual concerns in cancer: a systematic review. *J Cancer Surviv* 2017;11:175–88.
- Melisko ME, Narus JB. Sexual function in cancer survivors: updates to the NCCN guidelines for survivorship. *J Natl Compr Canc Netw* 2016;14:685–9.
- Carter J, Lacchetti C, Andersen BL, *et al*. Interventions to address sexual problems in people with cancer: American society of clinical oncology clinical practice guideline adaptation of cancer care Ontario guideline. *J Clin Oncol* 2018;36:492–511.
- Open AI. Usage policies. Available: <https://openai.com/policies/usage-policies> [Accessed 15 Jul 2023].
- Raggio GA, Butryn ML, Arigo D, *et al*. Prevalence and correlates of sexual morbidity in long-term breast cancer survivors. *Psychol Health* 2014;29:632–50.
- Ayoub NF, Lee Y-J, Grimm D, *et al*. Head-to-head comparison of ChatGPT versus Google search for medical knowledge acquisition. *Otolaryngol Head Neck Surg* 2023.





8 Walker HL, Ghani S, Kuemmerli C, *et al.* Reliability of medical information provided by ChatGPT: assessment against clinical

guidelines and patient information quality instrument. *J Med Internet Res* 2023;25:e47479.

© 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.

Development of a scoring system to quantify errors from semantic characteristics in incident reports

Haruhiro Uematsu ¹, Masakazu Uemura,² Masaru Kurihara ², Hiroo Yamamoto,² Tomomi Umemura,² Fumimasa Kitano,² Mariko Hiramatsu,² Yoshimasa Nagao^{1,2}

To cite: Uematsu H, Uemura M, Kurihara M, *et al.* Development of a scoring system to quantify errors from semantic characteristics in incident reports. *BMJ Health Care Inform* 2024;**31**:e100935. doi:10.1136/bmjhci-2023-100935

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2023-100935>).

Received 11 October 2023
Accepted 02 April 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.

¹Department of Quality and Patient Safety, Nagoya University Graduate School of Medicine, Nagoya, Aichi, Japan
²Department of Patient Safety, Nagoya University Hospital, Nagoya, Aichi, Japan

Correspondence to

Dr Haruhiro Uematsu;
hiro_uematsu@hotmail.com

ABSTRACT

Objectives Incident reporting systems are widely used to identify risks and enable organisational learning. Free-text descriptions contain important information about factors associated with incidents. This study aimed to develop error scores by extracting information about the presence of error factors in incidents using an original decision-making model that partly relies on natural language processing techniques.

Methods We retrospectively analysed free-text data from reports of incidents between January 2012 and December 2022 from Nagoya University Hospital, Japan. The sample data were randomly allocated to equal-sized training and validation datasets. We conducted morphological analysis on free text to segment terms from sentences in the training dataset. We calculated error scores for terms, individual reports and reports from staff groups according to report volume size and compared these with conventional classifications by patient safety experts. We also calculated accuracy, recall, precision and F-score values from the proposed 'report error score'.

Results Overall, 114 013 reports were included. We calculated 36 131 'term error scores' from the 57 006 reports in the training dataset. There was a significant difference in error scores between reports of incidents categorised by experts as arising from errors ($p < 0.001$, $d = 0.73$ (large)) and other incidents. The accuracy, recall, precision and F-score values were 0.8, 0.82, 0.85 and 0.84, respectively. Group error scores were positively associated with expert ratings (correlation coefficient, 0.66; 95% CI 0.54 to 0.75, $p < 0.001$) for all departments.

Conclusion Our error scoring system could provide insights to improve patient safety using aggregated incident report data.

INTRODUCTION

Many healthcare organisations have endorsed patient safety measures over the years.¹ However, the rates of medical errors and adverse events continue to be of serious concern.² Measures of quality are relatively well established, but the measurement and monitoring of safety continue to be problematic.³

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Incident reporting systems have an important role in patient safety. Incidents caused by errors have a particularly significant influence on patient safety. There are various methods to analyse errors in healthcare settings, but to our knowledge, no studies have explored methods to quantify errors or analyse organisational trends by using the scores developed from free text in incident reports.

WHAT THIS STUDY ADDS

⇒ We developed error scores that partly rely on natural language processing techniques to obtain quantitative information about the presence of error factors in incident reports. Group error scores, representing averaged error scores in a certain group, were positively associated with manual ratings of patient safety experts. Our error scoring system could provide insights to improve patient safety using aggregated incident report data.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The proposed model will be used to monitor chronological trends in errors in groups and increase the awareness of workers and general risk managers. This system will also potentially be helpful for preventing future incidents by providing a warning (score changes), as well as for educational purposes. In future, a useful tool to improve patient safety may be developed by combining and balancing multiple factors to produce scores using the same methodology applied herein.

Incident reporting systems allow healthcare workers to voluntarily disclose adverse events and 'near misses'.^{4,5} These systems function as barometers of risk in the healthcare setting and provide a foundation for organisational learning and improvement.⁶ In addition, voluntary confidential submission is thought to deepen our understanding of events and promote a safe environment.^{5,7} The use of incident reports is strongly recommended by the Institute of Medicine.⁸ However, their

varying quality is considered suboptimal for organisational learning.⁹ Reports submitted by frontline workers can provide valuable insights,⁴ especially the free-text sections used to describe incidents in greater detail,¹⁰ but interpretation of incidents is challenging for various reasons, including inadequate use of evolving health information technology.¹¹

The integration of artificial intelligence into patient safety measures has gained greater attention in recent years.^{12,13} Studies have explored how to obtain better value from incident reports using health information technologies. One recent study proposed an original decision-making model that partly relies on natural language processing (NLP) techniques to quantify the severity of incidents from aggregated big data and measure organisational trends using the central limit theorem.¹⁴ This model was novel in two ways. First, it used an original vectorisation approach to weigh terms from a bag of words. This enabled conversion of narrative free-text data into quantitative measures. Second, the model aimed to investigate organisational patterns and trends using a computer-assisted decision-making model. Generally, techniques using NLP help to answer binary questions or classify incident types for individual reporting.¹⁵ However, the WHO recommends collecting systemic insights from aggregated incident data,⁶ and this model was helpful in investigating particular factors in incident reports at the organisational level. However, it is not clear whether it could also be used to measure other factors in incident reports.

Incidents caused by errors potentially have a significant influence on patient safety. The occurrence of errors could lead to malpractice suits, which have an impact on healthcare costs.¹⁶ Errors also create a serious public health problem¹⁷ and are associated with stress for healthcare professionals.¹⁸ We therefore attempted to extract information about errors from incident reports using the model from a previous study. To date, various methods have been applied in healthcare settings to analyse errors.^{19,20} However, to our knowledge, no studies have used models to quantify errors or analyse organisational trends in incident reports.

This study aimed to develop error scores to quantify errors in incidents using semantic characteristics in incident reports, and to confirm the criterion-related validity of these error scores by comparing them with manual ratings of patient safety experts.

METHODS

Data sources

Incident reporting systems

All incident reports were collected at Nagoya University Hospital (NUH), Japan. NUH is a 1080-bed hospital that contributes to advanced medical care, education and research. NUH is the only national university hospital in Japan accredited by the Joint Commission International, an accreditation body for healthcare quality and

safety. NUH has used an incident reporting system since 2000 and a reporting culture is well established.²¹ Every employee can report incidents anonymously through the electronic health record system. The system collects background data about incidents using a structured format and free-text descriptions. Collected reports are reviewed by trained general risk managers (GRMs), a multidisciplinary group including physicians, nurses, a pharmacist and lawyers. Our hospital has been making considerable efforts to eliminate severe error-containing events, and GRMs sort incident reports according to information such as the severity and nature of errors. Severity is classified into five categories using the grading system developed at NUH: 'Near Miss', 'No Harm', 'Low Harm', 'Severe Harm' and 'Catastrophic/Fatal Event'. It has similarities with other grading systems such as those of the WHO and National University Hospital Council of Japan.¹⁴

Manual data labelling (definition of error)

The term 'error' has various meanings depending on the context; its precise meaning is actively debated,²² and no universal grading scale is used in healthcare. The WHO defines an error as a failure to carry out a planned action as intended or the application of an incorrect plan.²³ Errors are divided into active and latent. Active errors are caused by unsafe acts committed by personnel resulting from slips, lapses, mistakes or violations.²⁴ Latent errors may provoke further errors or create inherent weaknesses in the system. In NUH, GRMs review all incident reports, and incidents are labelled as containing errors if they are associated with any types of system, process or human errors, regardless of whether a patient was harmed. Reports are classified as error free when the incident is considered very unlikely to have occurred as the result of an error.

Generating training and validation datasets

We retrospectively extracted free-text data from incident reports dating from January 2012 to December 2022 at NUH. We included all free-text data from the submitted incident reports, regardless of the type of incident or the length of the text. Data were randomly allocated to equal-sized training and validation datasets. The training dataset was used to generate error scores according to the GRM classification, and the validation dataset was used to test the scores.

Semantic feature representation

We segmented the original free-text data to the smallest unit using morphological analysis to define the semantic characteristics in incident reports. Of the possible parts of speech, we used only nouns in the analysis because they appear most frequently in the text and represent the smallest unit of meaning. We did not preprocess the data because we prioritised applying the method to real-world data. During segmentation, sentences were processed into an unordered collection of words, known as a bag of words. This analysis was performed by an

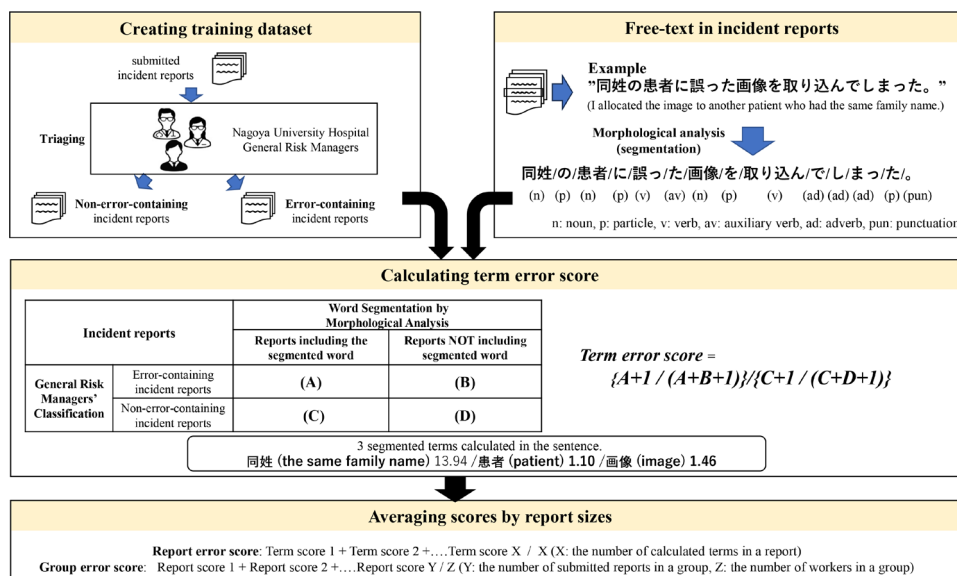


Figure 1 Process for calculating error scores from original text in incident reports.

open-source engine, MeCab, which was equipped with two commercially available medical dictionaries, MANBYO (MANBYO_201907_Dic-sjis) and Comejisyo (ComeJisyo Sjis-2), for application to medical writing.

Calculating error scores

Figure 1 provides an overview of how the error scores were developed from incident reports. After segmentation, the bag of words was transformed into a numerical representation using the original vectorisation ($\frac{A+1}{A+B+1} \Big/ \frac{C+1}{C+D+1}$), inspired by the epidemiological concept of relative risk. All segmented terms were examined using the χ^2 test in terms of the relative frequency of their use in error-containing reports and other reports, as classified by GRMs. We modified the formula by adding one to both the numerator and denominator to pick up more terms from free-text data and avoid zero probability. When a term appeared more frequently in reports of error-free incidents ($\frac{C+1}{C+D+1} > \frac{A+1}{A+B+1}$), implying that it is less important in incidents arising from errors, we reversed the numerator and denominator and replaced the plus sign with a minus sign ($-\frac{C+1}{C+D+1} \Big/ \frac{A+1}{A+B+1}$). A term score having more impact on the likelihood of errors being associated with the incident therefore becomes greater than 1, and one with the opposite effect becomes less than -1.

Once the error scores for segmented terms had been calculated, we averaged the scores for each report unit. Then, the scores for a certain group (clinical or non-clinical departments/wards) were calculated by averaging the scores for individual reports in that group, adjusted by its number of workers. 'Term error score', 'report error score' and 'group error score' were defined in this manner. We also analysed the score distributions according to report volume size.

Statistical analysis

We used the Wilcoxon signed-rank test to compare the 'report error score' with the manual GRM ratings. In addition, we calculated accuracy, recall, precision and F-score values from the report error score to evaluate the performance. For this analysis, we set a cut-off value on the basis of the receiver operating characteristic (ROC) curve using the training dataset. To determine the association of group error score level with manual ratings, we used Pearson's product-moment correlation test. A validation dataset not used for generating the training dataset was analysed using R software (V.4.3.0; R Project for Statistical Computing, Vienna, Austria).

RESULTS

Sample characteristics

Overall, 116786 incident reports were collected during the study period. The incident reports by year and reporter occupation are shown in table 1. Reports in all years were made most often by nurses or midwives, accounting for 73.4% of all reports. After 2018, when 'other healthcare professionals' were subdivided by profession, physicians, pharmacists and rehabilitation therapists were the most likely (after nurses) to submit reports.

We included 114013 reports in the study, with 2773 being excluded because the GRMs made no assessment about whether or not they were associated with errors. Of these reports, 71038 (62.3%) were determined to contain errors; 57006 reports were included in the training dataset and 57007 in the validation dataset.

Development of error scores

Using morphological analysis, error scores were calculated for 36131 terms in the incident reports in the training dataset. The median term error score was -1.36 (IQR: -3.27 to 1.22). The highest term error score (45.25)

Table 1 Incident reports by year and occupation of reporter

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	Total
Total submitted incident reports	9183	9752	10082	11443	11109	10333	10086	10676	10939	11883	11300	116786
By reporters												
Physician	635	674	768	958	822	700	746	987	1218	1033	971	9512
Nurse/midwife	7603	7899	7734	8457	7975	7379	7074	7357	7491	8517	8268	85755
Other healthcare professionals	914	1142	1542	1985	2263	2212						10058
Pharmacist							458	803	853	915	636	3665
Radiological technologist							221	172	170	222	266	1051
Medical technologist							310	259	156	145	188	1058
Rehabilitation therapist							314	387	395	413	381	1890
Orthoptist							51	45	55	56	33	240
Biomedical equipment technician							222	201	194	197	145	959
Nursing assistant							1	2	1	1	2	7
Dental hygienist							0	0	0	3	5	8
Nutritionist							219	226	178	143	144	910
Administrative assistant							447	186	191	200	238	1262
Security guard							0	0	0	2	2	4
Others	31	37	38	43	49	42	23	51	36	36	21	407

The category of 'other healthcare professionals' was subdivided after 2018.

was for ‘temoto joho’ (patient identifiers that healthcare workers can access to prevent misidentification), followed by ‘jikanme’ (hour(s) passed since an action was taken; 37.91), ‘kansasha’ (a person who inspects compounded medications; 30.58), ‘shokusatsu’ (a diet card with a patient identifier; 27.82) and ‘kanjagonin’ (patient misidentification; 23.85). The terms with high error scores are shown in online supplemental appendix 1.

The median report error score was 0.50 (IQR: -1.26 to 1.46). The median group error score, which is the total report error score of a group divided by the number of workers therein, was 0.06 (IQR: -0.46 to 0.61). Group error scores were high for the clinical nutrition (2.86), administration (2.73) and hospital pharmacy (2.20) departments, and low for the geriatrics ward (-2.38) and rehabilitation department (-2.09).

The SDs of these scores steadily decreased by level (3.45 for the term error score, 2.31 for the report error score and 0.85 for the group error score) (online supplemental appendix 2).

Validation and performance of the report error score

The median report error score was -1.66 (IQR: -3.66 to -0.05) for error-free reports and 1.11 (IQR: 0.35-1.80) for reports of incidents that GRMs labelled as error containing; the difference was significant ($p < 0.001$, $d = 0.73$ (large)) (figure 2). Regarding the performance metrics, accuracy, which indicates the model’s ability to correctly predict the outcome (error containing or non-error containing) on the basis of all reports, was 0.8. Recall, that is, the probability of identifying error-containing reports among the GRM-classified error-containing reports, was 0.82. Precision, that is, the concordance between GRM-categorised error-containing reports and error-containing reports as determined by the model, was 0.85. Finally, the F-score, which reflects the balance between precision and recall, was 0.84. These results were obtained using an optimal cut-off score of ≥ 0.037 for error-containing incident reports derived from the ROC analysis (figure 3).

Correlation of group error scores with manual classifications

A total of 119 organisational units were eligible to submit incident reports, including all departments and wards in NUH. Incident reports that GRMs rated as error containing were plotted against the group error scores for these 119 organisational units, and the correlation coefficient of 0.66 was highly significant (95% CI 0.54 to 0.75, $p < 0.001$) (figure 4A). In the subgroup analysis focusing on the clinical departments in which incident reports were only submitted by physicians ($n = 58$), the correlation coefficient was 0.71 (95% CI 0.55 to 0.82, $p < 0.001$) (figure 4B).

DISCUSSION

Main findings and implications

The writing quality of individual reports widely varied because the reporters sometimes used the same term

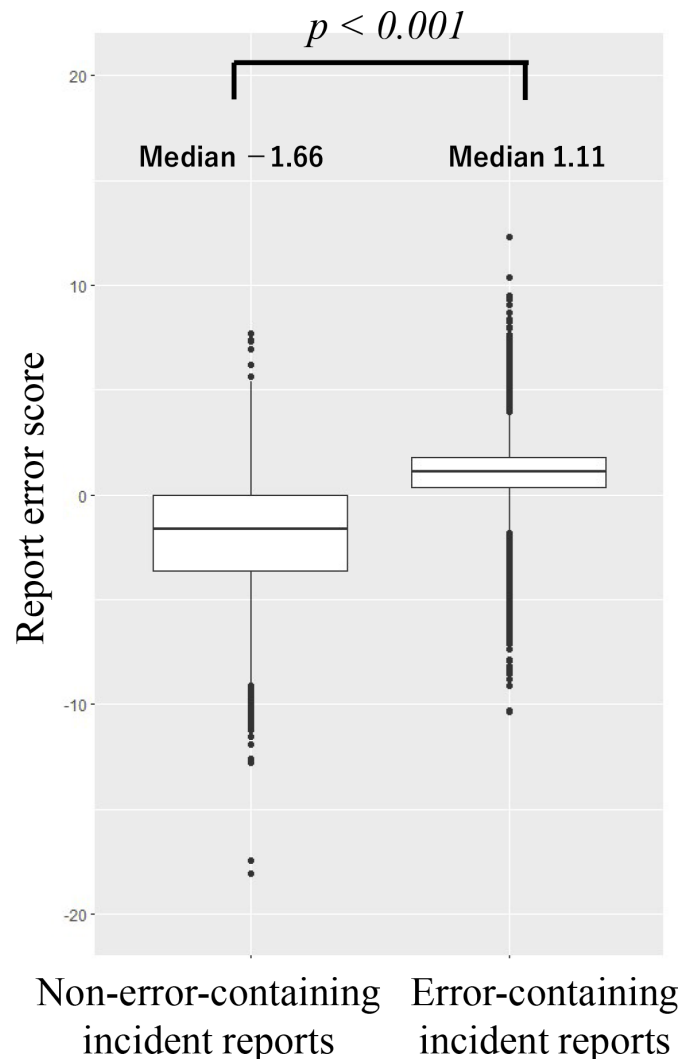


Figure 2 Box plot of report error scores among incident reports manually categorised by general risk managers (GRMs).

in different contexts or used different expressions to describe the same event. In addition, in medical dictionaries, ‘error’ appeared in terms such as ‘Human error’ and ‘Error message’, which may have affected the scores. However, the results became more reliable as the volume of reports increased, in line with the central limit theorem.

Notably, the report error score demonstrated that the model could more effectively identify reports of incidents arising from errors compared with manual categorisation by GRMs; the model’s performance metrics were good. These findings suggest that our model could be useful to analyse errors documented in individual reports, but we emphasise that it was designed to evaluate organisational trends in aggregated reports.

More importantly, higher error scores for departments were associated with a higher submission rate of error-containing incident reports. This phenomenon was also observed for group severity scores which indicate the severity of incidents using this model.¹⁴ Severe events rarely occur, but events associated with errors are

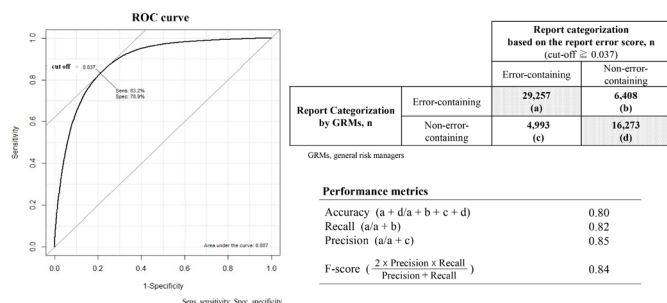


Figure 3 Performance metrics of report error score. ROC, receiver operating characteristic.

relatively common. The results suggest that our model is able to analyse factors involved in incidents regardless of their frequency of occurrence.

Departments with higher error scores, such as the clinical nutrition, administration and hospital pharmacy departments, tended to submit more reports. However, their reports included many near-miss and less severe events. The error score simply indicates the existence of error in association with an incident, not the severity of the error or its consequences. Each department provides their own services, and scores therefore cannot be compared directly among departments. The scores are also influenced by whether departments are correctly submitting reports of all incidents, including those arising from errors and other reasons. Although we are aware that the outcomes would have been more accurate had outliers been removed, the results are nevertheless considered robust given the sufficient data volume.

Comparison with previous related work

When artificial intelligence-enabled decision support systems are implemented correctly, they can improve patient safety.¹³ Researchers have explored the potential of applying NLP techniques to incident reports, often in conjunction with machine learning.¹⁵ Most studies used a binary classification, but research aiming to identify multiclass classifications is emerging gradually.²⁵ These studies were designed to answer questions

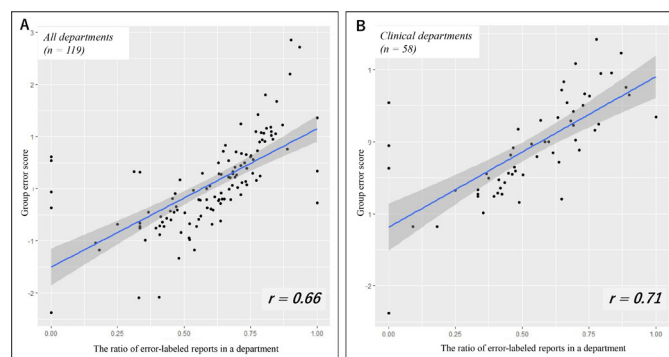


Figure 4 Scatter plot of the ratio of general risk manager (GRM)-labelled error-containing incident reports against averaged group error scores in each organisational unit at Nagoya University Hospital (NUH). In A, all departments were plotted. Only the clinical departments were included in B.

about individual incident reports. However, the writing quality (ie, complexity and length) of incident reports varies greatly.²⁶ Our model is unique in that we aimed to analyse groups of reports to understand organisational patterns and trends. We performed statistical analysis to compare the results between groups, but we could not find adequate classifiers to evaluate groups in the context of machine learning and NLP. We therefore adopted rank-based tests, which are sometimes used in NLP.^{27 28}

The drawback of rank-based tests is their relatively weak statistical power, but our sample size was large enough to overcome this limitation.

Various vectorisation methods, such as binary, term frequency, thresholding and term frequency-inverse document frequency methods, are generally used to transform segmented terms into numerical representations.²⁹ We adopted the same vectorisation method to weigh semantic characteristics as was applied to the severity score, which is used to quantify event severity on the basis of training data and GRM classifications.¹⁴ The severity score can also be used to predict organisational trends. A study on severity scores highlighted that many terms used in reports of severe incidents did not appear in reports of non-severe incidents. However, that study had a huge number of non-severe incident reports and far fewer severe reports.¹⁴ To alleviate this imbalance, the formula was updated in this study by adding one. This method reduced the number of words with a zero probability and has been used in other vectorisations, such as term frequency-inverse document frequency³⁰ and Bayesian vectorisation.³¹ However, direct comparison with other vectorisation models was outside the scope of this research.

Limitations

This study had several limitations. First, it used data from a single facility in Japan. All incident reports were written in Japanese, and the results may vary by language. Moreover, we applied a consensus method to triage reports using our institutional definition of 'error'. Unfortunately, the inter-rater reliability of the GRMs in terms of error scores was not confirmed, although we consider the quality of our safety department to be high. In addition, the judgements of multiple trained GRMs were considered, including legal experts. The number of incident reports may vary among hospitals depending on the reporting culture. As incident reports share similarities, we believe that this model is widely applicable, although additional research is required to confirm its applicability to other languages or institutions.

Second, we did not perform any qualitative analysis of the segmented terms generated by the morphological analysis, and the narrative descriptions in the reports were not included in the analysis. Although these factors would have influenced the quality of the scores, we nevertheless consider the study useful because it included a large sample of real-world data, including incomplete reports and ones with inaccurate event descriptions. However,

some measures, such as maintenance of dictionaries for morphological analysis and preprocessing of raw free-text data to correct typing errors, could improve the results.

Challenges for future work

In future, our scoring model could be used to monitor chronological trends in errors at the group level, as well as to increase the awareness of workers and GRMs. It might therefore provide data that could help prevent future incidents. We also expect this system to be useful for educating new GRMs.

We will continue to try to improve the performance of the model. We modified the vectorisation formula to increase calculable terms in free-text data; other possible measures include data preprocessing, updating dictionaries and identifying the optimal number of incident reports to assess group error scores.

In addition to severity and error, other factors are involved in incidents; we will aim to quantify these factors using the same methodology applied herein. In future, a useful tool could be developed to enhance organisational patient safety by combining multiple scores, including severity and error scores, in a balanced manner. This study represents a useful step towards that goal.

CONCLUSIONS

We developed a decision-making model to quantify errors by analysing the semantic characteristics of free-text data in incident reports. Analysing scores by organisational unit revealed strong correlations with expert ratings. By expanding the scope of this model, an incident reporting system promoting patient safety could be obtained.

Acknowledgements We thank the staff of the Department of Patient Safety at Nagoya University Hospital and members of ASUISHI and CQSO. We also thank Atsushi Okawa, Nobuyuki Toyama, Yasuyuki Nasuhara, Toshihiro Kaneko, Masashi Uramatsu, Yumi Arai, Kouichi Tanabe and Tatsuya Fukami for their help in this research project. We thank Melissa Leffler, MBA, and Michael Irvine, PhD, from Edanz (<https://jp.edanz.com/ac>) for editing a draft of this manuscript.

Contributors YN and MU conceived the idea for the study. HU led the writing of this paper. MU, MK, HY, TU, MH, FK and YN analysed and interpreted the data. MU, MK, HY and YN contributed to the writing of the paper as well as participated in revising this manuscript. All authors contributed substantially to the writing of the paper, and all reviewed and approved the final draft. HU and YN are the guarantors of this study.

Funding This research was supported by the Ministry of Health, Labour and Welfare Policy Research Grants, Research on Region Medical (201A2001).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the Nagoya University Hospital Research Ethics Committee (2020-0181).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The data that support the findings of this study are available from the corresponding author 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 iDs

Haruhiro Uematsu <http://orcid.org/0000-0003-2800-6802>

Masaru Kurihara <http://orcid.org/0000-0001-9195-4202>

REFERENCES


- 1 Dzau VJ, Shine KI. Two decades since to err is human: progress, but still a "chasm". *JAMA* 2020;324:2489–90.
- 2 Makary MA, Daniel M. Medical error—the third leading cause of death in the US. *BMJ* 2016;353:i2139.
- 3 Vincent C, Burnett S, Carthey J. Safety measurement and monitoring in Healthcare: a framework to guide clinical teams and healthcare organisations in maintaining safety. *BMJ Qual Saf* 2014;23:670–7.
- 4 Pham JC, Girard T, Pronovost PJ. What to do with healthcare incident reporting systems. *J Public Health Res* 2013;2:e27.
- 5 Evans SM, Smith BJ, Esterman A, et al. Evaluation of an intervention aimed at improving voluntary incident reporting in hospitals. *Qual Saf Health Care* 2007;16:169–75.
- 6 World Health Organization. *Patient safety incident reporting and learning systems: technical report and guidance*. Geneva: World Health Organization, 2020.
- 7 Stavropoulou C, Doherty C, Tosey P. How effective are incident-reporting systems for improving patient safety? A systematic literature review. *Milbank Q* 2015;93:826–66.
- 8 Kohn KT, Corrigan JM, Donaldson MS. *To err is human: building a safer health system*. Washington, DC, 1999.
- 9 Scott J, Dawson P, Heavey E, et al. Content analysis of patient safety incident reports for older adult patient transfers, handovers, and discharges: do they serve organizations, staff, or patients? *J Patient Saf* 2021;17:e1744–58.
- 10 Howell A-M, Burns EM, Bouras G, et al. Can patient safety incident reports be used to compare hospital safety? Results from a quantitative analysis of the English national reporting and learning system data. *PLoS One* 2015;10:e0144107.
- 11 Mitchell I, Schuster A, Smith K, et al. Patient safety incident reporting: a qualitative study of thoughts and perceptions of experts 15 years after 'to err is human'. *BMJ Qual Saf* 2016;25:92–9.
- 12 Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digit Med* 2021;4:54.
- 13 Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. *JMIR Med Inform* 2020;8:e18599.
- 14 Uematsu H, Uemura M, Kurihara M, et al. Development of a novel scoring system to quantify the severity of incident reports: an exploratory research study. *J Med Syst* 2022;46:106.
- 15 Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *Int J Med Inform* 2019;132:103971.
- 16 Hoshi T, Nagao Y, Sawai N, et al. Assessment of medical malpractice cost at a Japanese national University hospital. *Nagoya J Med Sci* 2021;83:397–405.
- 17 Rodziewicz TL, Houseman B, Hipskind JE. Medical error reduction and prevention. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing Copyright © 2023, StatPearls Publishing LLC, 2023.
- 18 Higham H, Vincent C. Human error and patient safety. In: Donaldson L, Ricciardi W, Sheridan S, et al, eds. *Textbook of patient safety and clinical risk management*. Cham (CH): Springer Copyright, 2021: 29–44.
- 19 Thomas EJ, Petersen LA. Measuring errors and adverse events in health care. *J Gen Intern Med* 2003;18:61–7.
- 20 Benn J, Koutantji M, Wallace L, et al. Feedback from incident reporting: information and action to improve patient safety. *Qual Saf Health Care* 2009;18:11–21.
- 21 Fukami T, Uemura M, Nagao Y. Significance of incident reports by medical doctors for organizational transparency and driving forces for patient safety. *Patient Saf Surg* 2020;14:13.



- 22 Fondahn E, Lane M, Vannucci A. *The Washington manual of patient safety and quality improvement*. Philadelphia, Pennsylvania: Wolters Kluwer, 2016.
- 23 World Health Organization. Patient safety curriculum guide: multi-professional edition. 2011. Available: https://apps.who.int/iris/bitstream/handle/10665/44641/9789241501958_eng.pdf;jsessionid=E3E8BA7049BED778EACF49D665F9FCD4?sequence=1
- 24 Reason J. *Human error*. Cambridge: Cambridge University Press, 1990.
- 25 Wang Y, Coiera E, Magrabi F. Using convolutional neural networks to identify patient safety incident reports by type and severity. *J Am Med Inform Assoc* 2019;26:1600–8.
- 26 Fong A, Hettinger AZ, Ratwani RM. Exploring methods for identifying related patient safety events using structured and unstructured data. *J Biomed Inform* 2015;58:89–95.
- 27 Rousseau JF, Ip IK, Raja AS, et al. Can automated retrieval of data from emergency department physician notes enhance the imaging order entry process? *Appl Clin Inform* 2019;10:189–98.
- 28 Donnelly LF, Grzeszczuk R, Guimaraes CV, et al. Using a natural language processing and machine learning algorithm program to analyze inter-Radiologist report style variation and compare variation between radiologists when using highly structured versus more free text reporting. *Curr Probl Diagn Radiol* 2019;48:524–30.
- 29 Ong MS, Magrabi F, Coiera E. Automated identification of extreme-risk events in clinical incident reports. *J Am Med Inform Assoc* 2012;19:e110–8.
- 30 van Zaanen M, Kanters P. *Automatic mood classification using TF*IDF based on lyrics*. ISMIR, 2010.
- 31 Sueno HT, Gerardo BD, Medina RP. Converting text to numerical representation using modified Bayesian vectorization technique for multi-class classification. *International Journal* 2020;9:10.30534.

© 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.

'If you build it, they will come...to the wrong door: evaluating patient and caregiver-initiated ethics consultations via a patient portal'

Liz Blackler ¹, Amy E Scharf,¹ Konstantina Matsoukas,^{1,2} Michelle Colletti,^{1,3} Louis P Voigt^{1,4}

To cite: Blackler L, Scharf AE, Matsoukas K, *et al*. 'If you build it, they will come...to the wrong door: evaluating patient and caregiver-initiated ethics consultations via a patient portal'. *BMJ Health Care Inform* 2024;**31**:e100988. doi:10.1136/bmjhci-2023-100988

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2023-100988>).

Received 05 December 2023
Accepted 17 April 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.

¹Ethics Committee, Memorial Sloan Kettering Cancer Center, New York, New York, USA

²Technology Division, Library Services, Memorial Sloan Kettering Cancer Center, New York, New York, USA

³Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, New York, USA

⁴Department of Anesthesiology, Pain, and Critical Care Medicine, Memorial Sloan Kettering Cancer Center, New York, New York, USA

Correspondence to

Ms Liz Blackler;
blacklel@mskcc.org

ABSTRACT

Objectives Memorial Sloan Kettering Cancer Center (MSK) sought to empower patients and caregivers to be more proactive in requesting ethics consultations.

Methods Functionality was developed on MSK's electronic patient portal that allowed patients and/or caregivers to request ethics consultations. The Ethics Consultation Service (ECS) responded to all requests, which were documented and analysed.

Results Of the 74 requests made through the portal, only one fell under the purview of the ECS. The others were primarily requests for assistance with coordinating clinical care, hospital resources or frustrations with the hospital or clinical team.

Discussion To better empower patients and caregivers to engage Ethics, healthcare organisations and ECSs must first provide them with accessible, understandable and iterative educational resources.

Conclusion After 19.5 months, the 'Request Ethics Consultation' functionality on the patient portal was suspended. Developing resources on the role of Ethics for our patients and caregivers remains a priority.

INTRODUCTION

Clinical ethics consultation services exist to support patients, families, clinicians and hospital administrators who are facing ethical or moral challenges related to patient care. Patients are the common denominators in most ethics consultations.¹ However, in the USA in general and our institution specifically, ethics consultations are overwhelmingly requested by physicians and other clinicians.²⁻⁵ In our 2021 *BMJ Journal of Medical Ethics* article 'Call to action: empowering patients and families to initiate clinical ethics consultations,' we hypothesised on the reasons for this trend and the many potential benefits that reversing this phenomenon could impart to patients, families, clinicians and entire institutions.⁶

In January 2022, Memorial Sloan Kettering Cancer Center (MSK) launched a programme

to enable patients/caregivers to request ethics consultations through our electronic patient portal. Previously, they could only do so by telephone, while MSK staff could request consultations either through an electronic health record order or by telephone.⁷ In this Implementer Report, we describe how the patient portal-initiated ethics consult programme was designed, implemented and received by patients, caregivers, and MSK staff.

METHODS

MSK's Ethics Committee (EC) and Digital Informatics & Technology Solutions developed functionality on MSK's secure electronic patient portal that allowed patients or their designated caregivers to learn about the Ethics Consultation Service (ECS) and then submit a request for an ethics consultation by briefly explaining their 'reason for consult.' Portal secure messages alerted the EC leadership about each request, providing the requestor's name, contact information, relationship to patient and reason for request. The EC leadership contacted requestors by the end of the next business day, assessed their needs, addressed any concerns and facilitated appropriate next steps and referrals. The ECS documented all requests, relevant data and remediation processes. Anticipating that there would likely be requests outside the scope of clinical ethics, the EC leadership met with Directors of Case Management, Patient Financial Services, Patient Representative/Advocate and Social Work in advance of the launch to secure their support and identify pathways for assistance.

A 'Request Consultation' functionality was added to the patient portal directly below the two already-existing Ethics Committee

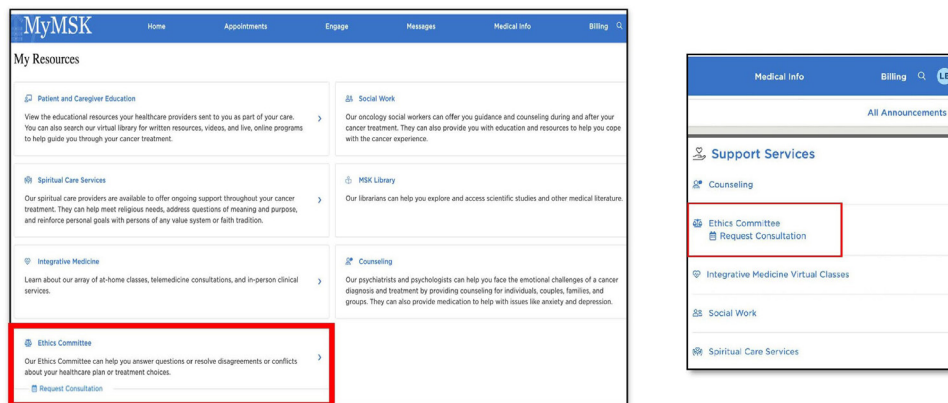


Figure 1 Ethics consultation requests on the patient portal.

references on the ‘My Resources’ and ‘Support Services’ pages (figure 1). When users clicked on the ‘Request Consultation’ link, they were directed to a new page where they were asked to provide their name, contact information and reason for consultation request (online supplemental file 1).

The ‘My Resources’ page included a one-sentence description of the EC: ‘Our Ethics Committee can help you answer questions or resolve disagreements or conflicts about your healthcare plan or treatment choices.’ The portal’s space constraints precluded our providing additional information about the EC or ECS, and therefore, we created a hyperlink from ‘Ethics Committee’ that took users to the EC page on MSK’s public, external-facing website—<https://www.mskcc.org/experience/patient-support/ethics-msk>. These pages provide comprehensive information on our EC and ECS, including both written and video content explaining what we do and the circumstances/events for which it is appropriate—and even recommended—for patients or caregivers to request an ethics consultation. However, our external website does not allow users to request an ethics consultation. For this, patients and caregivers must return to the patient portal.

RESULTS

Between January 2022 and September 2023 (19.5 months), 74 requests were made through the patient portal, with 62% (n=46) originating from patients and 38% (n=28) from caregivers. 93 (93%) (n=69) of requests were for patients being treated in the out-patient setting.

Of the 74 ethics consultation requests, only one was for assistance with an issue that fell under the customary scope of our ECS—the need to facilitate a goals of care discussion between the spouse of an incapacitated patient admitted to the intensive care unit and the clinical team (online supplemental file 2). The remaining 73 requests (98%) were for assistance with issues that Ethics Consultants are not trained to address or remediate. These requests fell into five major categories:

1. Assistance in coordinating clinical care: 33% (n=24), such as appointment scheduling, symptom management and treatment decision-making.

2. Complaints about hospital processes or systems: 27% (n=20) including frustrations with scheduling delays, securing information and guidance, and receiving return phone calls.
3. Frustrations about the clinical team: 19% (n=14), where patients/caregivers reported dissatisfaction with one or more of their care providers, and/or requested transfers of care.
4. Requests for hospital resources: 12% (n=9), including assistance with billing, travel/local accommodations, and referrals for emotional support and home care and/or hospice services.
5. The remaining 9% of requests involved non-specific concerns, for which the consultant offered active listening and emotional support (n=4) and/or general guidance (n=3).

In responding to these 73 requests, Ethics Consultants referred 56% (n=41) to MSK’s Patient Representative Department, which is tasked with addressing patient and caregiver concerns. The 24 requests (33%) regarding treatment and symptom management were referred to the patient’s primary service. The remaining eight requests (11%) prompted Ethics Consultants to offer patients and caregivers direct educational and psychosocial support but required no additional referrals.

DISCUSSION

We continue to maintain that patients and caregivers should be empowered to take a more proactive role in requesting ethics consultations. But as our 19.5-month experience demonstrated, solely providing them with a technological platform is not sufficient. Of the 74 consultation requests, 73 did not address ‘ethical’ issues, illustrating patients’ and caregivers’ limited understanding of the ‘jurisdictions’ of ECs and ECSs. Healthcare organisations and ECs must provide patients and caregivers with accessible, understandable, and iterative resources and education on ECs and ECSs so they can appropriately use a ‘Request Ethics Consultation’ portal function to address concerns and challenges that are truly ethical in nature and within the purview of an ECS. Space constraints on our organisation’s patient platform severely limited the

amount of information we are able to include about our services.

Patients and caregivers should not be held responsible for incorrectly requesting assistance on issues that fall outside the expertise of the ECS. They were confronting physically and emotionally stressful periods of their lives and a complex healthcare system. We should not expect them to intuitively understand the roles and responsibilities of the ECS. We surmise that patients and caregivers contacted us through the patient portal for two primary, interrelated reasons. First, many may have perceived that their questions, experiences and complaints were 'ethical' in nature, given their perceptions that they felt 'wronged,' 'not heard' or 'not supported' by members of the clinical teams. These perceptions may have been reinforced by what we now recognise was an overly vague explanation of the Ethics Committee on the patient portal, which may have been misinterpreted. Second, the ethics patient portal presented an 'actionable' and technologically expedient platform to document their concerns and receive a timely response.

This need for enhanced understanding of the role of the ECS within an institution is not without precedent. Over the past 7 years, the EC at MSK has undertaken multiple, overlapping staff educational programmes to raise awareness and increased comfort with the role, function and potential contributions of the EC and ECS to patient care. The multipronged endeavours have resulted in a steady increase in the number and variety of ethics consultations and may serve as models for programmes geared toward patients and caregivers. We acknowledge the significant resources that such education would demand, particularly for an ever-changing cohort of patients and families.

The difficult decision to suspend this functionality on the patient portal

Recognising the limitations of this initiative, EC leadership deliberated over continuing the patient portal consultation functionality (while making adjustments where possible) or suspending it and focusing on ethics-related resources and programming for patients and caregivers. We recognised that the 73 'non-ethics' requests were relevant, in that they reflected existing gaps in communication between patients/caregivers and their providers, and that the ECS did provide patients and caregivers with an avenue for having their concerns acknowledged and potentially addressed. Our ultimate decision to suspend the patient portal functionality was based on two primary considerations:

- ▶ First, we concluded that comprehensive educational resources for patients and caregivers about the EC and ECS was a prerequisite to a successful presence on the patient portal. Our efforts would require designing multifaceted and iterative programmes in conjunction with multiple institutional stakeholders,

including our Patient and Family Advisory Committee for Quality.

- ▶ Second, we were confident that we were not leaving our patients and caregivers without access to the supports that they needed. The majority of issues that had been raised with Ethics were best addressed by other institutional services, primarily Patient Representative and the patients' own clinical care teams—all of which are accessible through the patient portal. Our patients and caregivers are comfortable using the hospital's patient portal (approximately 80% of MSK patients are subscribed, and the portal receives 6000–8000 messages daily from 200 000 active users). Patient Representative receives approximately 10–15 messages per day from patients and caregivers. Moreover, our institution's EC and Patient Representative Department have a longstanding and collaborative relationship, and we continue to work together when issues arise that are relevant to both our services.

It is important to note that decision to suspend Ethics Consultation requests on the patient portal did NOT leave patients and caregivers without the means to request ethics consultations. Our public-facing site prominently displays our ECS phone number, which is available 24/7 for all constituents.

Limitations

Implementing and maintaining this programme required a meaningful dedication of time and resources by both the ECS and other interdisciplinary teams, particularly Patient Representative. The ECS was committed to responding to all requests by the end of the next business day and to appropriately document referrals of consult requests to relevant institutional services. Not all institutions have the resources to staff such an endeavour, especially in the setting of a large and active ethics consultation workflow.

Finally, we are fortunate that our colleagues within other institutional services were receptive to our calls and emails about, and quickly and professionally assumed responsibility for the relevant issues that patients and caregivers raised. We recognise that at other institutions, these professional relationships may not be well-established or work as seamlessly.

CONCLUSION

After 19.5 months, the Ethics leadership made the difficult decision to suspend the 'Request Ethics Consultation' functionality on the patient portal. We nevertheless remain committed to empowering patients and caregivers to access our services. To achieve such a lofty goal, MSK and the EC leadership must provide patients and caregivers with sufficient and ongoing education and support that helps them understand the mission, benefits and limitations of the clinical ethics consultation process. We



are currently working with other services within our institution to (a) address the institutional deficiencies related to care coordination, delays in returning phone calls and complaints about providers; (b) provide patients and caregivers with a sustained and robust programme aimed at enhancing their understanding of the ethics consultation process.

Acknowledgements The authors thank Memorial Sloan Kettering's Digital Informatics & Technology Solutions (DigITs) team and Rozina Merchant for her support and oversight of the Patient Portal build. We also thank Claire Murray for her tireless efforts collecting and maintaining the Ethics Consultation database.

Contributors All authors contributed equally.

Funding This work was supported by the Ethics Committee at Memorial Sloan Kettering Cancer Center and by the National Institutes of Health Core Grant P30 CA008748 to Memorial Sloan Kettering Cancer Centre, New York, NY, USA.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

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

Liz Blackler <http://orcid.org/0000-0003-3008-5941>

REFERENCES

- 1 Ulrich CM. The moral distress of patients and families. *Am J Bioeth* 2020;20:68–70.
- 2 Cho HL, Grady C, Tarzian A, *et al*. Patient and family descriptions of ethical concerns. *Am J Bioeth* 2020;20:52–64.
- 3 DuVal G, Clarridge B, Gensler G, *et al*. A national survey of U.S. internists' experiences with ethical dilemmas and ethics consultation. *J Gen Intern Med* 2004;19:251–8.
- 4 Hurst SA, Hull SC, DuVal G, *et al*. How physicians face ethical difficulties: a qualitative analysis. *J Med Ethics* 2005;31:7–14.
- 5 Wocial LD, Molnar E, Ott MA. Values, quality, and evaluation in ethics consultation. *AJOB Empir Bioeth* 2016;7:227–34.
- 6 Blackler L, Scharf AE, Matsoukas K, *et al*. Call to action: empowering patients and families to initiate clinical ethics consultations. *J Med Ethics* 2023;49:240–3.
- 7 Marathe PH, Zhang H, Blackler L, *et al*. Ethics consultation requests after implementation of an electronic health record order. *JCO Oncol Pract* 2022;18:e1505–12.

© 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.

Association between daily step counts and healthy life years: a national cross-sectional study in Japan

Masahiro Nishi ,^{1,2} Reo Nagamitsu,^{3,2} Satoaki Matoba¹

To cite: Nishi M, Nagamitsu R, Matoba S. Association between daily step counts and healthy life years: a national cross-sectional study in Japan. *BMJ Health Care Inform* 2024;**31**:e101051. doi:10.1136/bmjhci-2024-101051

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2024-101051>).

Received 12 February 2024
Accepted 16 April 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.

¹Department of Cardiovascular Medicine, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kyoto, Japan
²Department of Health and Welfare, Kyoto Prefectural Government, Kyoto, Japan
³Department of Epidemiology for Community Health and 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 Despite accumulating evidence concerning the association between daily step counts and mortality or disease risks, it is unclear whether daily step counts are associated with healthy life years.

Methods We used the combined dataset of the Comprehensive Survey of Living Conditions and the National Health and Nutrition Survey conducted for a randomly sampled general population in Japan, 2019. Daily step counts were measured for 4957 adult participants. The associations of daily step counts with activity limitations in daily living and self-assessed health were evaluated using a multivariable logistic regression model. The bootstrap method was employed to mitigate uncertainties in estimating the threshold of daily step counts.

Results The median age was 60 (44–71) years, and 2592 (52.3%) were female. The median daily step counts were 5650 (3332–8452). The adjusted OR of activity limitations in daily living for the adjacent daily step counts was 0.27 (95% CI 0.26 to 0.27) for all ages and 0.25 (95% CI 0.25 to 0.26) for older adults at the lowest, with the thresholds of significant association at 9000 step counts. The OR of self-assessed unhealthy status was 0.45 (95% CI 0.44 to 0.46) for all ages and 0.42 (95% CI 0.41 to 0.43) for older adults at the lowest, with the thresholds at 11 000 step counts.

Conclusion Daily step counts were significantly associated with activity limitations in daily living and self-assessed health as determinants of healthy life years, up to 9000 and 11 000 step counts, respectively. These results suggest a target of daily step counts to prolong healthy life years within health initiatives.

INTRODUCTION

Healthy life years, also known as healthy life expectancy, are holistic health indicator encompassing life, health, disease, disability, activity limitation and overall well-being. Estimations of healthy life years are derived from national health surveys that incorporate questionnaires to assess the presence of activity limitations in daily living and self-assessed health in the USA and the UK as well as Japan.^{1–3} In addition to conventional risk factors of lifestyle-related diseases,^{4 5} several non-fatal conditions, such as mental health diseases, orthopaedic problems and

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Healthy life years, also known as healthy life expectancy, have come to be focused on as a holistic health indicator in an ageing society. Estimations of healthy life years are derived from national health surveys to assess the presence of activity limitations in daily living and self-assessed health in the USA and the UK as well as Japan. Despite accumulating evidence concerning the association between daily step counts and mortality or disease risks, there is little knowledge regarding the association between daily step counts and healthy life years.

WHAT THIS STUDY ADDS

⇒ Daily step counts were significantly associated with activity limitations in daily living and self-assessed health as key determinants of healthy life years, up to 9000 and 11 000 step counts, respectively.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The findings suggest the optimal target of daily step counts to prolong healthy life years within health initiatives aimed at increasing physical activity and raising health awareness of people. An effective health promotion, such as an increase in daily step counts in the general population, will prolong healthy life years and narrow the existing health disparities.

neurological disorders, substantially influence healthy life years.^{6–8} Effective health policies for the disease prevention and health promotion are necessary to prolong healthy life years and narrow the differences between life expectancy and healthy life years.

Physical activities and exercises are expected to be beneficial for healthy life years.^{9–11} Walking is a cost-effective aerobic physical activity in daily life, and its metric, step count, is readily measured by simple devices. While the WHO has issued guidelines recommending regular physical activity to promote a healthy life, it has refrained from specifying the optimal target of daily step counts.¹² Notably, an increase in daily step counts ameliorates cardiovascular disease

(CVD) and mortality risk.^{13–15} In this context, target step counts of 7200 and 8800 per day have been suggested for mitigating the risk of CVD and all-cause mortality, respectively.¹⁶

Despite accumulating evidence concerning the association between daily step counts and mortality or disease risks, there is limited knowledge regarding the association between daily step counts and healthy life years. The purpose of this study is to elucidate the associations of daily step counts with activity limitations in daily living and self-assessed health, as key determinants of healthy life years, and provide the optimal target of daily step counts to prolong healthy life years.

METHODS

Study design and setting

The Comprehensive Survey of Living Conditions (CSLC), a cross-sectional national survey, has been conducted every 3 years by Japanese Ministry of Health, Labour and Welfare (MHLW) to investigate the fundamental dimensions of the nation's livelihood, such as health, medical care, welfare, pension and income.³ In health questionnaire of the CSLC, subjective symptoms, health problems in daily life, disease or injury under treatment, subjective health assessment, worries and stress, mental state, and receiving rate of health check-ups are surveyed. The response rate for CSLC in 2019 was recorded at 72.5%. The National Health and Nutrition Survey (NHNS), another cross-sectional national survey, has been conducted with a random sample of participants drawn from the CSLC by the MHLW to comprehensively investigate the nation's physical status, nutrition intakes and lifestyle.¹⁷ The response rate for NHNS in 2019 was recorded at 63.5%. In the NHNS, daily step counts in a single day were measured with a pedometer which was distributed to the participants, accompanied by detailed instruction on the measurement procedure. These surveys were conducted for a randomly sampled general population through face-to-face interviews. An online survey format was used together for questionnaires regarding physical status and lifestyle in NHNS.

The data of CSLC and NHNS were integrated using common identifier for prefecture, region, unit, household and household member. Among the combined dataset of NHNS and CSLC in 2019, data from 22 respondents with missing values of activity limitations in daily living or self-assessed health were excluded. Consequently, data from 4957 adult responders aged ≥ 18 years were prepared for the analysis (online supplemental figure 1). The data were analysed from 30 September 2023 to 20 November 2023.

Outcomes

In the CSLC, activity limitations in daily living of responders were investigated using responses to the questions, 'Do you have any health problem which limits your daily activity?' Respondents who answered 'yes' were

categorised into the 'activity limitations' group, and those who answered 'no' were categorised into the 'no activity limitation' group. In the NHNS, self-assessed health was investigated using responses to the questions, 'How would you rate your current health status?'. Respondents who answered 'excellent', 'good', or 'fair' were categorised into the 'self-assessed healthy status' group, and those who answered 'poor' or 'bad' were categorised into the 'self-assessed unhealthy status' group. The associations of daily step counts with activity limitations in daily living and self-assessed health were evaluated for all age group ≥ 18 years and older adults group ≥ 65 years. In addition, the health condition without activity limitation (HCAL), a machine-learning-based integrated health index reflecting on healthy life years,⁶ was examined.

Statistical analysis

General statistics were performed in R V.4.2.0.¹⁸ Categorical values are represented as numbers (along with percentages), and numerical values are represented as medians with IQRs. A $p < 0.05$ was considered statistically significant. The HCAL was computed based on predictive probabilities for activity limitations in daily living employing a machine-learning model deployed in Python V.3.10.6. according to our preceding report.⁶ Spline curve and its slope curve were plotted with a 95% CI.

The bootstrap method followed by bootstrap aggregating, referred to as bagging, was employed to mitigate uncertainties in estimating the threshold of daily step counts by reducing the prediction error.¹⁹ A large sample group was generated by repeatedly drawing samples with replacement from the original sample. Age, sex and several kinds of diseases or injuries under treatment are known to be important predictors for healthy life years.⁶ After 1000 rounds of bootstrapping, multivariable logistic regression model was fitted for each sample, incorporating variables such as age, sex, increments of 1000 daily step counts and the presence of diseases or injuries under treatment as predictive factors. Subsequently, mean adjusted ORs were calculated from the set of predicted values of each model, and regression curve fitting was performed. The OR for adjacent daily step counts was calculated to define the threshold at which the upper limit of the 95% CI reached 1.0.

Patient and public involvement

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

RESULTS

Participants characteristics

Among the combined dataset of the CSLC ($n=481\,255$) and the NHNS ($n=6820$) in 2019, the data of adult responders were extracted (online supplemental figure 1). The baseline characteristics of participants were described for all-age group ≥ 18 years ($n=4957$) and older adults group

Table 1 Baseline characteristics of study participants

Characteristics	All aged ≥18 years (n=4957)	Older adults aged ≥65 years (n=2024)
Age, years	60 (44–71)	72 (69–78)
Sex (female)	2592 (52.3)	1046 (51.7)
BMI	22.7 (20.5–25.2)	23.0 (20.9–25.3)
Daily step counts	5650 (3332–8452)	4351 (2314–7023)
<2000	478 (12.3)	342 (21.1)
2000–3999	757 (19.4)	400 (24.7)
4000–5999	842 (21.6)	339 (20.9)
6000–7999	714 (18.3)	241 (14.9)
8000–9999	468 (12.0)	147 (9.0)
10 000–11 999	284 (7.2)	73 (4.5)
12 000–13 999	156 (4.0)	34 (2.1)
14 000–15 999	93 (2.3)	24 (1.4)
16 000+	107 (2.7)	22 (1.3)
Activity limitations in daily living	686 (13.8)	452 (22.3)
Self-assessed health status (unhealthy)	681 (13.7)	405 (20.1)
HCAL	93.2 (79.6–96.1)	83.4 (62.6–91.3)
Obesity	34 (0.68)	24 (1.1)
Hypertension	913 (18.4)	675 (33.3)
Diabetes	364 (7.3)	276 (13.6)
Dyslipidaemia	399 (8.0)	278 (13.7)
Gout	94 (1.9)	52 (2.5)
Depression or other mental diseases	103 (2.0)	24 (1.1)
Dementia	34 (0.68)	31 (1.5)
Parkinson disease	15 (0.30)	13 (0.64)
Other neurological disorders, pain or paralysis	46 (0.92)	27 (1.3)
Stroke, cerebral haemorrhage or infarction	72 (1.4)	58 (2.8)
Angina, myocardial infarction	114 (2.3)	94 (4.6)
Other cardiovascular disease	116 (2.3)	89 (4.4)
Malignant neoplasm or cancer	68 (1.3)	47 (2.3)
Anaemia or blood disease	31 (0.62)	14 (0.69)
Thyroid disease	89 (1.8)	47 (2.3)
Allergic rhinitis	123 (2.4)	60 (2.9)
Acute nasopharyngitis, common cold	19 (0.38)	11 (0.54)
Chronic obstructive pulmonary disease	16 (0.32)	14 (0.69)
Asthma	68 (1.3)	34 (1.6)
Other respiratory disease	78 (1.5)	54 (2.6)
Stomach or duodenum disease	101 (2.0)	77 (3.8)
Liver or gallbladder disease	68 (1.3)	40 (1.9)
Other digestive disease	65 (1.3)	40 (1.9)
Rheumatoid arthritis	39 (0.78)	28 (1.3)
Arthritis	160 (3.2)	105 (5.1)
Stiff shoulder	149 (3.0)	93 (4.5)
Back pain	316 (6.3)	220 (10.9)
Bone fracture	39 (0.78)	32 (1.5)

Continued

Table 1 Continued

Characteristics	All aged ≥ 18 years (n=4957)	Older adults aged ≥ 65 years (n=2024)
Osteoporosis	116 (2.3)	101 (4.9)
Other injury or burns	37 (0.74)	21 (1.0)
Kidney disease	65 (1.3)	46 (2.2)
Eye disease	389 (7.8)	319 (15.8)
Ear disease	67 (1.3)	55 (2.7)
Prostatic hypertrophy	108 (2.1)	99 (4.8)
Dental disease	319 (6.4)	194 (9.5)
Atopic dermatitis	45 (0.90)	6 (0.29)
Other skin disease	112 (2.2)	49 (2.4)
Menopausal or postmenopausal disorder	11 (0.22)	1 (0.049)

Categorical values represented as numbers (%) and numerical values as median (IQR).
BMI, body mass index; HCAL, health condition without activity limitations.

≥ 65 years (n=2024) (table 1). The median age was 60 years vs 72 years, and female sex comprised 2592 (52.3%) vs 1046 (51.7%) in the all-age and older adults group, respectively. The median daily step counts were 5652 step counts in the all-age group, compared with 4358 step counts in the older adults group. The prevalence rate of activity limitations in daily living was 686 (13.8%) vs 452 (22.3%), and the rate of self-assessed unhealthy status was 681 (13.7%) vs 405 (20.1%) in the all-age and older adult groups, respectively. The median value of HCAL was 93.2 vs 83.4. The older adults group exhibited higher prevalence rate of various kinds of diseases or injuries under treatments compared with the all-age group, except for depression or other mental diseases, atopic dermatitis and menopausal or postmenopausal disorder.

Association between daily step counts and activity limitations in daily living

To ascertain the association between daily step counts and healthy life years, spline curve and its slope curve were depicted for daily step counts and HCAL (figure 1). In the all-age group, as daily step counts increased, HCAL increased with a marked rise at fewer daily step counts although the increase was gradually diminished in higher daily step counts. The lower limit of the 95% CI of slope descended below 0 at approximately 12 000 step counts per day. Similar to the all-age group, in the older adults group, as daily step counts increased, HCAL showed a similar trend, and the lower limit of the 95% CI of slope descended below 0 within the range of 10 000–11 999 step counts per day. The sex difference was not detected in the association between daily step counts and HCAL in both the all-age and older adult groups (online supplemental figure 2).

The associations between daily step counts and activity limitations in daily living were assessed (figure 2 and online supplemental figure 3). In the all-age group, as daily step counts increased, the unadjusted rate of activity

limitations in daily living declined, and the adjusted odds of activity limitations in daily living steadily decreased. The adjusted OR for the adjacent daily step counts was 0.27 (95% CI 0.26 to 0.27) at the lowest, and the upper limit of the 95% CI reached 1.0 at 9000 step counts (OR 0.95 (95% CI 0.91 to 1.00)). Similarly, the older adults group exhibited a parallel trend in the association between daily step counts and activity limitations in daily living. The OR was 0.25 (95% CI 0.25 to 0.26) at the lowest, and the upper limit of the 95% CI reached 1.0 at 9000 step counts (OR 0.92 (95% CI 0.85 to 1.00)). The impact of sex was not significant in both the all-age and older adult groups (online supplemental figure 4 and 5). Collectively, daily step counts were significantly associated with activity limitations in daily living, with the threshold at 9000 step counts in both the all-age and older adult groups.

Association between daily step counts and self-assessed health

The associations between daily step counts and self-assessed health were also assessed (figure 3 and online supplemental figure 6). In all-age group, as daily step counts increased, the unadjusted rate of self-assessed unhealthy status declined, and the adjusted odds of self-assessed unhealthy status steadily decreased. The OR for the adjacent daily step counts was 0.45 (95% CI 0.44 to 0.46) at the lowest, and the upper limit of the 95% CI reached 1.0 at 11 000 step counts (OR 0.95 (95% CI 0.89 to 1.00)). The older adults group exhibited similar trends of the association between daily step counts and self-assessed unhealthy status. The OR was 0.42 (95% CI 0.41 to 0.43) at the lowest, and the upper limit of the 95% CI reached 1.0 at 11 000 step counts (OR 0.91 (95% CI 0.80 to 1.03)). The impact of sex was not significant in both the all-age and older adult groups (online supplemental figures 7 and 8). Thus, daily step counts were significantly associated with self-assessed health, with the threshold

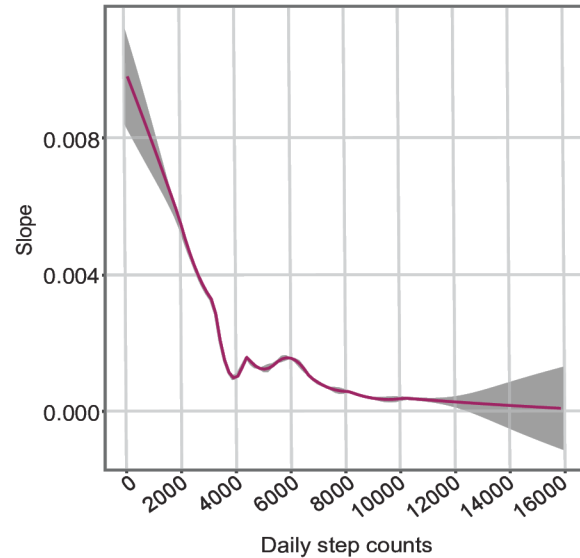
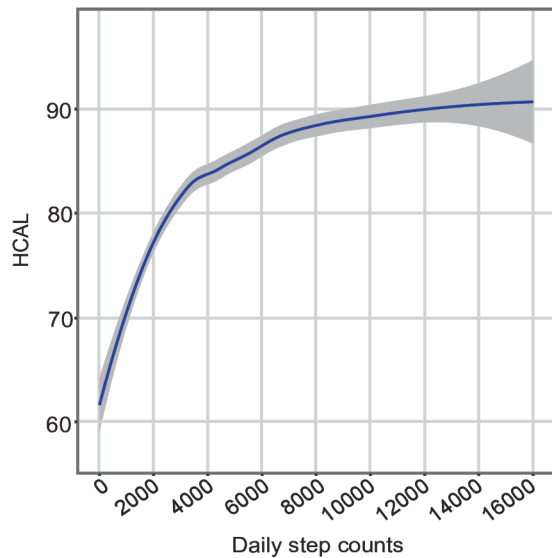
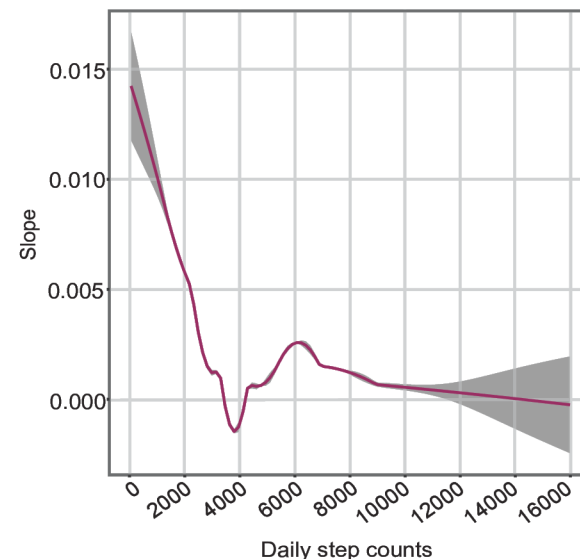
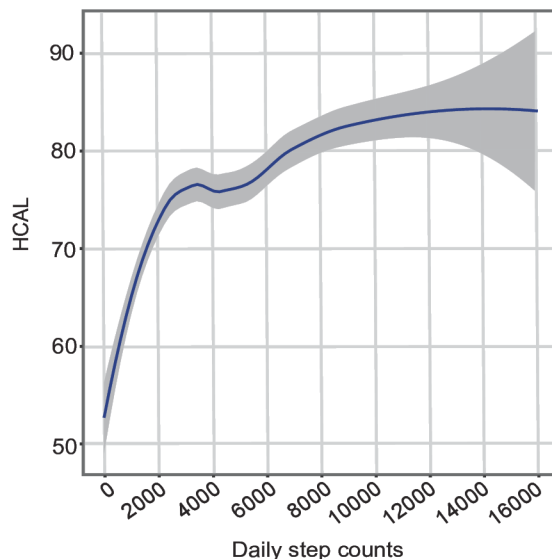
Age \geq 18 years

 Age \geq 65 years


Figure 1 Association between daily step counts and integrated health index reflecting on healthy life years. Spline curve and slope curve were depicted by plotting daily step counts and health condition without activity limitations (HCAL) for all-age and older adult groups. Shaded areas represent 95% CI.

at 11 000 step counts in both the all-age and older adult groups.

DISCUSSION

This study using the combined national health survey data has provided the insights into the associations of daily step counts with activity limitations in daily living and self-assessed health, as key determinants of healthy life years. An increase in daily step counts was significantly associated with the improvement of activity limitations in daily living and self-assessed health, with thresholds of significant association at 9 000 and 11 000 step counts, respectively, regardless of age.

This study suggests the optimal target of daily step counts to prolong healthy life: 9 000 step counts for activity limitations in daily living and 11 000 step counts for self-assessed health. Increase in daily step counts reduces CVD and mortality risks, whereas the effect of step intensity is controversial.^{13–15 20 21} A meta-analysis has introduced 7 200 and 8 800 step counts per day as potential targets for reducing CVD and all-cause mortality risks, respectively.¹⁶ Given that healthy life years are shorter than life expectancy, it is conceivable that the optimal targets of daily step counts, for prolonging healthy life years, may necessitate more step counts than what are required for the reduction of mortality and CVD risk.

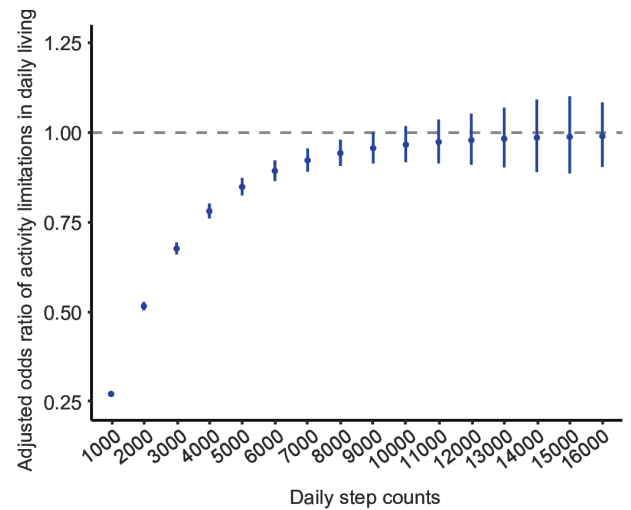
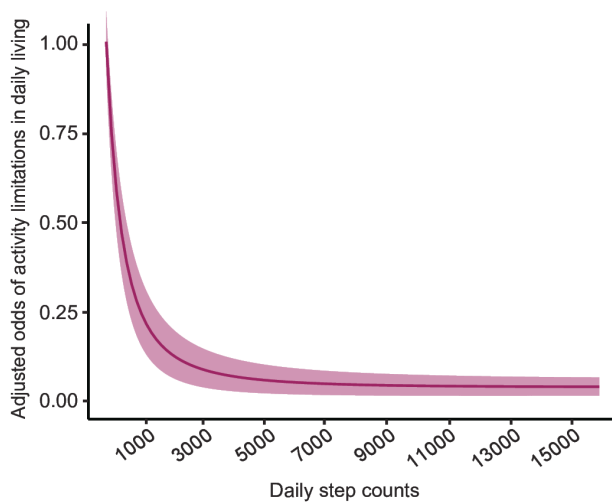
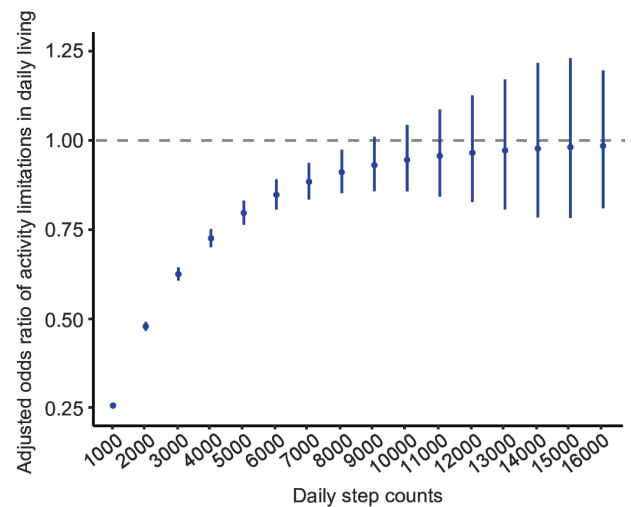
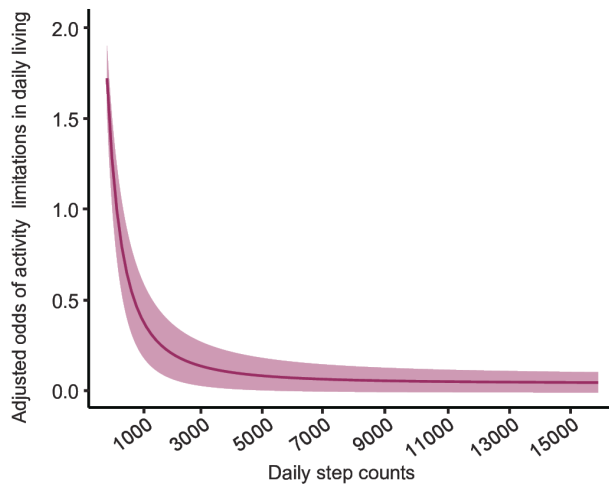
Age ≥ 18 yearsAge ≥ 65 years

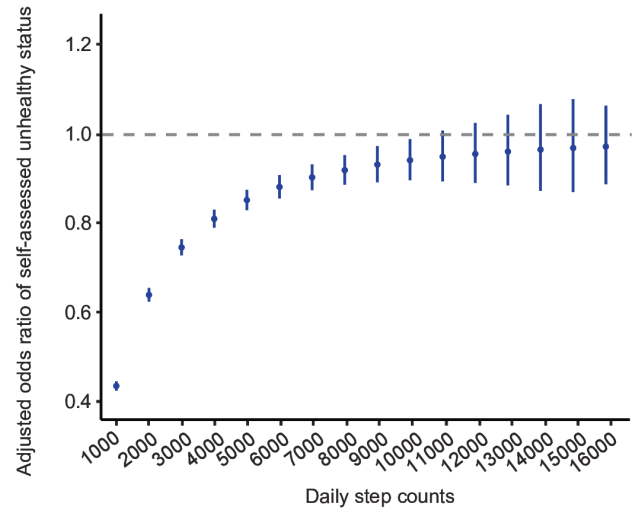
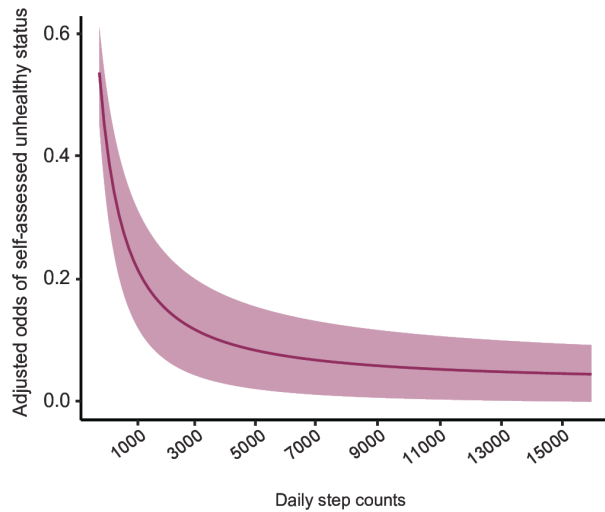
Figure 2 Association between daily step counts and activity limitations in daily living. Adjusted ORs and ORs of activity limitation in daily living were evaluated for daily step counts. Once the daily step counts reached 9000, an increase in 1000 step counts per day was no longer significantly associated with reduced odds of activity limitations in daily living for all-age and older adult groups. Error bars and shaded areas represent 95% CI.

We evaluated activity limitations in daily living and self-assessed health status—metrics that do not include mortality, as alternatives to healthy life years. Indeed, healthy life years are estimated based on a health survey regarding activity limitations in daily living and self-assessed health in the USA and the UK as well as Japan.^{1–3} The prevalence rates of activity limitations in daily living or self-assessed unhealthy status in each age groups are incorporated into a life table to estimate healthy life years.²² Self-assessed health has been shown to be a predictor for mortality and morbidity.^{23 24} In addition, racial and ethnic disparities of self-assessed health status have been reported.²⁵ While life expectancy has

been increasing, healthy life years have not kept pace in the world.^{26 27} An improvement of activity limitations in daily living and self-assessed health by an effective health promotion, such as an increase in daily step counts in the general population, will prolong healthy life years, and narrow the existing health disparities.

The data used in this study are derived from national cross-sectional surveys, and no causal inferences can be deduced. Population-based cohort will be needed to further investigate the prospective effect of daily step counts on healthy life years. Step intensity was not evaluated because the investigation did not include it. Potential confounding factors, such as income, that could affect

Age ≥ 18 years



Age ≥ 65 years

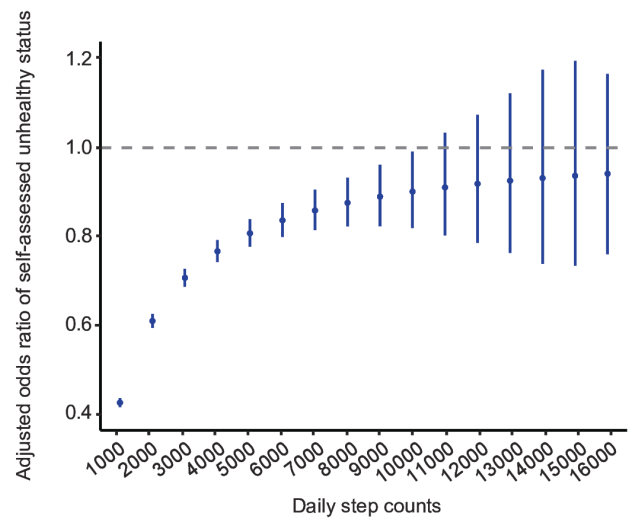
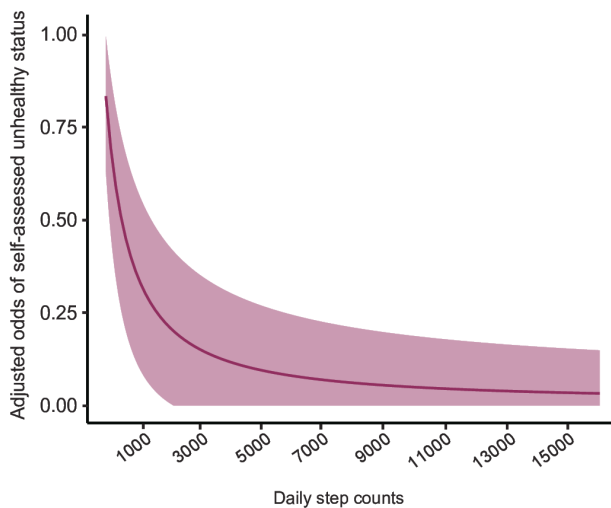


Figure 3 Association between daily step counts and self-assessed health. Adjusted ORs and ORs of self-assessed unhealthy status were evaluated for daily step counts. Once the daily step counts reached 11 000, an increase in 1000 step counts per day was no longer significantly associated with reduced odds of self-assessed unhealthy status for all-age and older adult groups. Error bars and shaded areas represent 95% CI.

the association between daily step counts and healthy life years were not considered.

CONCLUSION

Daily step counts were significantly associated with activity limitations in daily living and self-assessed health as determinants of healthy life years, up to 9000 and 11 000 step counts, respectively. These findings suggest the optimal target of daily step counts to prolong healthy life years within health initiatives aimed at increasing physical activity and raising health awareness of people.

Acknowledgements We are grateful to Tomoyuki Yamamoto, Mika Yamashita, Satoko Kumagai and Kumiko Katsuyama from the Department of Health and Welfare, Kyoto Prefectural Government.

Contributors MN was responsible for conception of the study and overall content as guarantor. MN and RN had full access to all of the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. All authors participated in manuscript writing and approved the final manuscript. SM provided overall supervision. All the authors were responsible for the decision to submit the manuscript for publication.

Funding This study was supported by Foundation for Total Health Promotion.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study received ethical approval from the ethics committee of Kyoto Prefectural University of Medicine with the approval number ERB-C-2878. This study conformed to the principles outlined in the Declaration of Helsinki. Given that this study uses pre-existing national survey data, informed consent from study participants was exempted.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. We are prohibited from publicly opening the data. Data can be accessed through the Household Statistics Office of the Japanese Ministry of Health, Labour and Welfare (<https://www.mhlw.go.jp/toukei/itiran/eiyaku.html>).

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

- National Center for Health Statistics. Foundation health measures - technical notes. n.d. Available: https://www.cdc.gov/nchs/healthy_people/hp2020/fhm-technical-notes.htm
- Office for National Statistics. health state life Expectancies, UK. n.d. Available: <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/healthstatelifeexpectanciesuk/2018to2020#measuring-the-data>
- Ministry of Health, Labour and Welfare, Japan. Comprehensive survey of living conditions. 2019. Available: <https://www.mhlw.go.jp/toukei/list/20-21kekka.html>
- Stenholm S, Head J, Kivimäki M, et al. Smoking, physical inactivity and obesity as predictors of healthy and disease-free life expectancy between ages 50 and 75: a Multicohort study. *Int J Epidemiol* 2016;45:1260–70.
- Willcox BJ, He Q, Chen R, et al. Midlife risk factors and healthy survival in men. *JAMA* 2006;296:2343–50.
- Nishi M, Nagamitsu R, Matoba S. Development of a prediction model for healthy life years without activity limitation: national cross-sectional study. *JMIR Public Health Surveill* 2023;9:e46634.
- Salomon JA, Vos T, Hogan DR, et al. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the global burden of disease study 2010. *Lancet* 2012;380:2129–43.
- Myojin T, Ojima T, Kikuchi K, et al. Orthopedic, ophthalmic, and psychiatric diseases primarily affect activity limitation for Japanese males and females: based on the comprehensive survey of living conditions. *J Epidemiol* 2017;27:75–9.
- Monma T, Takeda F, Noguchi H, et al. The impact of leisure and social activities on activities of daily living of middle-aged adults: evidence from a national longitudinal survey in Japan. *PLoS One* 2016;11:e0165106.
- Yamada M, Arai H. Self-management group exercise extends healthy life expectancy in frail community-dwelling older adults. *Int J Environ Res Public Health* 2017;14:531.
- Monma T, Takeda F, Noguchi H, et al. Exercise or sports in Midlife and healthy life expectancy: an ecological study in all prefectures in Japan. *BMC Public Health* 2019;19:1238.
- World Health Organization. *WHO Guidelines on Physical Activity and Sedentary Behaviour*. Geneva, Switzerland: World Health Organization, 2020.
- Del Pozo Cruz B, Ahmadi MN, Lee IM, et al. Prospective associations of daily step counts and intensity with cancer and cardiovascular disease incidence and mortality and all-cause mortality. *JAMA Intern Med* 2022;182:1139–48.
- Mañas A, Del Pozo Cruz B, Ekelund U, et al. Association of accelerometer-derived step volume and intensity with hospitalizations and mortality in older adults: a prospective cohort study. *J Sport Health Sci* 2022;11:578–85.
- Saint-Maurice PF, Troiano RP, Bassett DR Jr, et al. Association of daily step count and step intensity with mortality among US adults. *JAMA* 2020;323:1151–60.
- Stens NA, Bakker EA, Mañas A, et al. Relationship of daily step counts to all-cause mortality and cardiovascular events. *J Am Coll Cardiol* 2023;82:1483–94.
- Ministry of Health, Labour and Welfare, Japan. National health and nutrition survey. 2019. Available: <https://www.mhlw.go.jp/toukei/itiran/gaiyo/k-eisei.html>
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing V, Austria; 2022. Available: <https://www.R-project.org/>
- Breiman L. Bagging predictors. *Mach Learn* 1996;24:123–40.
- Lee I-M, Shiroma EJ, Kamada M, et al. Association of step volume and intensity with all-cause mortality in older women. *JAMA Intern Med* 2019;179:1105–12.
- Tudor-Locke C, Schuna JM Jr, Han HO, et al. Step-based physical activity Metrics and Cardiometabolic risk: NHANES 2005–2006. *Med Sci Sports Exerc* 2017;49:283–91.
- Sullivan DF. A single index of mortality and morbidity. *HSMHA Health Rep* 1971;86:347–54.
- Idler EL, Benyamini Y. Self-rated health and mortality: a review of twenty-seven community studies. *J Health Soc Behav* 1997;38:21–37.
- DeSalvo KB, Fan VS, McDonnell MB, et al. Predicting mortality and Healthcare utilization with a single question. *Health Serv Res* 2005;40:1234–46.
- Mahajan S, Caraballo C, Lu Y, et al. Trends in differences in health status and health care access and Affordability by race and Ethnicity in the United States, 1999–2018. *JAMA* 2021;326:637–48.
- Kyu HH, Abate D, Abate KH, et al. Dalys and HALE collaborators. global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet* 2018;392:1859–922.
- Salomon JA, Wang H, Freeman MK, et al. Healthy life expectancy for 187 countries, 1990–2010: a systematic analysis for the global burden disease study 2010. *Lancet* 2012;380:2144–62.

© 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.

Building a house without foundations? A 24-country qualitative interview study on artificial intelligence in intensive care medicine

Stuart McLennan ^{1,2}, Amelia Fiske ¹, Leo Anthony Celi ^{3,4,5}

To cite: McLennan S, Fiske A, Celi LA. Building a house without foundations? A 24-country qualitative interview study on artificial intelligence in intensive care medicine.

BMJ Health Care Inform 2024;**31**:e101052. doi:10.1136/bmjhci-2024-101052

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2024-101052>).

Received 15 February 2024
Accepted 08 April 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY. Published by BMJ.

¹Institute of History and Ethics in Medicine, Department of Preclinical Medicine, TUM School of Medicine and Health, Technical University of Munich, Munich, Bavaria, Germany

²Institute for Biomedical Ethics, University of Basel, Basel, Switzerland

³Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

⁴Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA

⁵Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

Correspondence to

Dr Stuart McLennan;
stuart.mclennan@tum.de

ABSTRACT

Objectives To explore the views of intensive care professionals in high-income countries (HICs) and lower-to-middle-income countries (LMICs) regarding the use and implementation of artificial intelligence (AI) technologies in intensive care units (ICUs).

Methods Individual semi-structured qualitative interviews were conducted between December 2021 and August 2022 with 59 intensive care professionals from 24 countries. Transcripts were analysed using conventional content analysis.

Results Participants had generally positive views about the potential use of AI in ICUs but also reported some well-known concerns about the use of AI in clinical practice and important technical and non-technical barriers to the implementation of AI. Important differences existed between ICUs regarding their current readiness to implement AI. However, these differences were not primarily between HICs and LMICs, but between a small number of ICUs in large tertiary hospitals in HICs, which were reported to have the necessary digital infrastructure for AI, and nearly all other ICUs in both HICs and LMICs, which were reported to neither have the technical capability to capture the necessary data or use AI, nor the staff with the right knowledge and skills to use the technology.

Conclusion Pouring massive amounts of resources into developing AI without first building the necessary digital infrastructure foundation needed for AI is unethical. Real-world implementation and routine use of AI in the vast majority of ICUs in both HICs and LMICs included in our study is unlikely to occur any time soon. ICUs should not be using AI until certain preconditions are met.

INTRODUCTION

Intensive care medicine has long been at the forefront of efforts to use routinely collected digital health data to improve patient care,^{1–3} and it is seen to be particularly well positioned to use the advances in artificial intelligence (AI) given the amount of data typically generated in intensive care units (ICUs).⁴ It is expected that applications of AI in ICUs will primarily be focused on machine learning to assist in disease identification, prediction

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Existing research on intensive care professionals' views about artificial intelligence remains limited and only includes participants from four high-income countries.

WHAT THIS STUDY ADDS

⇒ This is one of the largest qualitative studies to date to examine the views of intensive care professionals regarding the use and implementation of AI technologies in intensive care units, involving 59 participants from 24 countries. It shows that the vast majority of ICUs neither have the technical capability to capture the necessary data or run AI algorithms, nor the staff with the right knowledge and skills to use the technology as designed.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Pouring massive amounts of resources into developing AI without first building the necessary digital and knowledge infrastructure foundation needed for AI is unethical and needs to change.

of disease progression, disease phenotyping, recognising unique patterns within complex data and guiding clinical decision-making.^{5–7} Other potential applications include algorithms taking a physically embodied presence, such as in smart autonomous ventilators or infusion pumps.^{8,9} Despite the anticipated benefits AI technology, a large 'implementation gap' between what has been developed and what is used in clinical practice continues to grow, with most developed ICU AI models remaining in testing and prototyping.^{10,11} Challenges for the successful development and implementation of AI tools in ICUs have been increasingly researched and discussed in recent years,^{4–17} including: (1) various technological challenges around obtaining high-quality data; ICU data is often heterogeneous and noise-prone, and de-identifying, standardising, cleaning and structuring the

data can be difficult^{4-6 11 14}; (2) a number of general ethical, legal and regulatory issues, particularly around data protection and sharing^{5 6 18}; (3) the vast majority of ICU AI models are not robust or ready for clinical use; they have been developed using retrospective data, without external validation or prospective evaluation^{6 15}; and (4) obtaining the trust and acceptance of clinicians and other stakeholders.^{5 6} Indeed, it is important to better understand intensive care professionals' views and acceptance of AI to help identify key barriers and facilitators to AI technology being implemented and adding value to intensive care medicine. At the time of designing and initiating this study, there was a lack of empirical studies on intensive care professionals' views about AI. However, in the past 2 years, a few quantitative and qualitative studies have been published.^{13 14 19 20} These studies have found general positive attitudes and expectations of ICU professionals towards the use of AI, but also primarily identified technical barriers to the implementation of AI in ICUs. In addition, they identified some non-technical factors (a lack of AI knowledge among ICU professionals, high clinical workload, no clear AI policy, a lack of funding for digitalisation and a culture of doctor-knows-best). However, these studies have consisted of three small survey studies involving one centre^{13 20} or two centres¹⁹ from the Netherlands or the USA, and an interview study including participants from the USA and three European countries (the Netherlands, Belgium and the UK).¹⁴ Existing research on intensive care professionals' views about AI therefore remains limited and only includes participants from four high-income countries (HICs). Furthermore, HICs have so far dominated the discussion over AI and related ethical issues.²¹ In an era of increasing global collaborative health research efforts, this imbalance is problematic. Lower-to-middle-income countries (LMICs) are also increasingly using healthcare data science and AI.²²⁻²⁵ This study therefore aims to explore the views of intensive care professionals in both HICs and LMICs regarding the use and implementation of AI technologies in ICUs.

METHODS

This study is presented in accordance with the Consolidated Criteria for Reporting Qualitative Research reporting guideline.²⁶ See online supplemental information 1 for additional details on methods used in the study. Intensive care professionals were primarily selected through purposive sampling to ensure that participants were from different backgrounds and regions.²⁷ The classification of a country as an HIC or an LMIC was taken from the Statistical Annex of the World Economic Situation and Prospects 2022.²⁸ Additional participants were identified using snowball sampling.²⁹ 59 intensive care professionals (physicians, nurses, pharmacists, physical therapists) from 24 countries agreed to participate. Interviews were held via telephone or video call between December 2021 and August 2022. All interviews were conducted in English, except for seven interviews which

were held in Spanish. A researcher-developed semi-structured interview guide was developed to guide the discussion (see online supplemental information 2). It should be noted that the interviews were conducted prior to the release of ChatGPT and other chatbots powered by large language models (LLMs).³⁰ Interviews were audio recorded and transcribed and were analysed in their original language using conventional content analysis with the assistance of the qualitative software MAXQDA (VERBI Software).³¹

RESULTS

Among the 59 intensive care professionals who participated in the study, 69.5% were physicians (41/59), 18.6% were nurses (11/59), 6.8% were pharmacists (4/59) and 3.4% were physical therapists (2/59). Overall, 23.7% of participants were from Europe (14/59), 16.9% were from Asia (10/59), 15.3% were from North America (9/59), 13.6% were from South America (8/59), 11.9% were from the Middle East (7/59), 10.2% were from Australasia (6/59) and 6.8% were from sub-Saharan Africa (4/59). Furthermore, 66.1% (39/59) of participants were male presenting (table 1).

Status quo—patient data collection, documentation and utilisation

Most participants described a pervasive lack of digital data collection and documentation, and a chronic underutilisation of patient data in ICUs in both HICs and LMICs. In relation to patient data collection and documentation, most ICUs were reported to be paper-based or partially digitalised. Although patient data may be being collected with electronic monitors in these ICUs, it is typically documented manually either in paper-based records or in electronic health records. Consequently, the amount of available digital data was reported to be limited in most ICUs. With regard to the use of patient data for purposes other than patient care, although most ICUs are using data for national quality benchmarking data sets, the secondary use of patient data was reported to be extremely limited or non-existent by most participants.

Only a few participants working in a small number of large tertiary hospitals in HICs reported that patient data in their ICUs were primarily being automatically collected and documented digitally and being extensively used for secondary purposes. However, these were outliers and participants reporting that most other ICUs within the same country or even city as these fully digitalised ICUs were only paper-based or partially digitalised. Furthermore, even in most fully digitalised ICUs, it was reported that data is still required to be manually verified at regular intervals due to regulatory requirements to ensure data validity. Nevertheless, participants noted that in practice large amounts of data would often be confirmed without detailed verification. A minority of participants reported that verification is not required in their ICU; they want the raw data and did not think that nurses at the bedside

Table 1 Participants demographics

Characteristic	Total
Gender	
Male	39/59 (66.1)
Female	20/59 (33.9)
Position	
ICU physicians	41/59 (69.5)
ICU nurses	11/59 (18.6)
ICU pharmacist	4/59 (6.8)
ICU therapist	2/59 (3.4)
Other	1/59 (1.7)
Region	
Europe	14/59 (23.7)
Germany	3
France	2
Switzerland	1
Spain	1
The Netherlands	4
UK	4
Asia	10/59 (16.9)
China	2
Hong Kong	2
India	2
Japan	2
Philippines	2
North America	9/59 (15.3)
Canada	3
USA	5
Mexico	1
South America	8/59 (13.6)
Argentina	3
Columbia	5
Middle East	7/59 (11.9)
Qatar	4
Israel	2
Jordan	1
Australasia	6/59 (10.2)
Australia	4
New Zealand	2
Sub-Saharan Africa	4/59 (6.8)
Botswana	2
Malawi	1
Rwanda	1

ICU, intensive care unit.

were best placed to check data validity and that their time would be best spent on other tasks (table 2).

Views about using AI in ICUs

Perceived opportunities

Although there were large variations in knowledge of AI among participants, and the vast majority are currently not using AI technology in practice, all participants in

both HICs and LMICs had a generally positive view of AI. Participants saw huge potential for the technology to be very helpful and improve patient outcomes in the ICU, although not all participants had a clear idea of what or how benefits would happen. Many participants, however, highlighted the potential benefits of AI in relation to their workload given the number of patients they needed to simultaneously look after and the impossibility of keeping track of all the information being constantly generated in the ICU. AI was seen as a tool to support intensive care professionals deal with this data overload and to do their jobs more effectively and efficiently; by providing an early warning system for patients deteriorating, predicting which patients are at greatest risk and reducing errors. Many participants also noted the potential for AI to improve workflows, such as helping to manage ICU bed capacity or improving the accuracy of documentation (table 3).

Concerns about use

Most intensive care professionals, however, also held some well-known concerns about the use of AI in clinical practice. There were no important differences regarding the concerns expressed by participants from HICs and LMICs. Five key concerns emerged from the interviews:

Validity

A major concern raised by participants was regarding the risk of AI technology being biased and not generalisable. Participants were very concerned about AI applications not being applicable in real life to the majority of patients, particularly in ICU where there is such a heterogeneous group of patients. Participants were also concerned that AI technology would not work as well with minorities who are already disadvantaged (eg, Indigenous communities or those with limited healthcare access) if those groups are not sufficiently present in the training data set.

Explainability

Some participants thought explainable AI was necessary as they always needed to understand exactly why they were doing something when working with critically ill patients, and that a lack of understanding could generate fear and undermine the trust of clinicians and patients. However, most participants were not concerned about ‘blackbox’ AI applications and thought that evidence that an application was helpful and safe was far more important than explainability. These participants noted that they did not understand how many other technologies used in the ICU worked and that clinical judgement should not be based purely on an algorithm but should combine a range of patient information and professional expertise.

Responsibility

Most participants saw the issue of responsibility being dependent on how AI was used. If AI was used in place of a clinician, making changes to patient care independently, then the question of who would be responsible if things went wrong was seen as very problematic by

Table 2 Status quo—patient data collection, documentation and utilisation

Theme		
Code	Subcode	Example quote
Data collection and documentation		
Implementation stage	Paper-based	<p>'We collect it manually. So, we have an admission book. So, when a patient come, we collect the personal information of the patient, there is like the name of the patient, where the patient is coming from...But then, on our daily monitoring...we have now the observation where we record all the vitals signs...we do that in the admission book, as well as the patient files. We do it manually. We don't have like, electronic documentation.' P2 ICU Nurse LMIC</p> <p>'Most of the systems in [Country] are still paper based. Certainly, in the ICU, we are probably well, 10 years behind our [Country] cousins and probably 15, 20 years behind the US in terms of the way that we manage data.' P47 ICU Physician HIC</p>
	Partial digitalisation	<p>'But the way that is transferred, there is that all the information is stored in the monitor, for example, or in a ventilator. So, it's in there. You won't lose. It's in there. But then you have to go write down the numbers and then move to the computer and transfer those numbers in there. So, for us, that's the big limitation. Because first, you cannot do it minute by minute. And second, it's very time consuming for a person to transfer that. And third, you are not sure that the number that she's transferring is a real number.' P5 ICU Physician LMIC</p> <p>'Okay. So, we have an electronic medical record...It's introduced manually. So, we don't really have like an automatic process where the data is stored. So, basically doctors and nurses put the data in the Electronic Medical Record. So, that's the way we have to restore information.' P18 ICU Physician LMIC</p>
	Full digitalisation	<p>'Yeah, 95% of it is now electronic. So, starting from the vital signs, these are imported through-, collected through a central monitor, which is monitoring every single patient bed. And from that central monitor, it goes into a centralized database. And we're using the [Company name] system. And it's recording minute by minute data. But for verification, and the nurse would chart the data every hour. And if there's an event, which requires more frequent charting, for example, patients deteriorating or some sudden event, the nurse can then chart more data between hours. In terms of the lab data, it is collected via the hospital electronic database. So, our database goes and fetches data from the hospital database...So, we have to use-, we have to juggle a few systems at one time.' P3 ICU Physician HIC</p> <p>'So, I can say confidently at this point, it's 100% electronic documentation as far as vital signs goes. We have a, like a background software that transports patient's vital signs, for ICU patients almost minute-by-minute to like a secondary software that we have that's called [Name]...So, all the vital signs get automatically transcribed. Labs usually gets also documented from electronic medical records, also to that software. So, they're all on the same place. The-, like all the drip rates are manually entered by the nurses when they're started, ended, up titrated or down titrated. All the nursing assessment...they all go in there by manual entry.' P6 ICU Pharmacist HIC</p>
Variations within countries and regions		<p>'So, I've experienced a really wide variation...I've worked in four different hospitals throughout [Country]. And on one end, the hospital has been almost completely paper based, with a separate computer system for pathology values, a separate one for discharge summaries, all the vital signs are recorded manually. All the blood gases are recorded manually. And those are all sort of electronic data storage apart from blocks of text. The other extreme has been, the hospital I'm working in that moment, which is just fully integrated. So, everything is all on one system and includes all the observations which are recorded manually by the nurses. So, there's huge sort of data and variables available for where I am working at the moment. And it's all integrated across the sort of lifespan of the person in the hospital.' P8 ICU Physician HIC</p> <p>'Yeah. So, the online EMR that they use at [Hospital] it's called [Software name], and that has everything in it, it's like your bloods, medications, vital signs, everything is integrated into the one system. And even you'll take a blood sugar, and it will automatically go across to the online system. Whereas [City] was pretty much all paper-based, all of the lab systems and everything were all segregated systems, and often things were then transcribed, the blood results will be transcribed onto paper. So, if you're going to collect data about obs and things, you have to go around and individually, look at each patient.' P28 ICU Nurse HIC</p>
Secondary use of data		
Types	None	<p>'At the moment, no. So, 100% is for clinical care.' P3 ICU Physician HIC</p> <p>'Not that I've seen. So mainly, it's patient care, follow up of patients.' P25 ICU Physician LMIC</p>
	Quality benchmarks	<p>'In places where there's manual data collection, it's primarily been for benchmarking reports...But most ICUs, probably about 95% of ICUs in [Country] collect data that's submitted to a central body that provides a benchmarking report. And so, that would be the only sort of routinely collected clinical data that's sent off for benchmarking and reporting back. I'd say in the hospitals I've worked in, which had EMRs, or, you know, yeah, like databases, definitely it all got used for secondary purposes and quite frequently and quite a lot. And in places where it was manually extracted, then hardly ever.' P8 ICU Physician HIC</p>
	Research	<p>'The answer is yes. So, especially being a-, like an academic medical centre, we have like ongoing research all times within our critical care.' P6 ICU Pharmacist HIC</p>
HIC, high-income country ; ICU, intensive care unit; LMIC, lower-to-middle-income country.		

many. However, if AI was used as just another tool to help clinical decision-making, then participants thought that there was no significant problem and responsibility would remain with the clinician.

Dependency

Many participants raised concerns about clinicians becoming too dependent and trusting of AI technology in the ICU and not using their own clinical judgement

or skills. Participants saw this as part of a wider problem related to increasing digitalisation. Although this technology potentially has benefits, participants reported many junior staff becoming too reliant on technology, which was leading to (1) deskilling of staff, who can no longer do certain tasks themselves (eg, calculate dosages) because the system is down, and (2) a dehumanisation of care, with staff spending too much time looking at the

Table 3 Views about using AI in ICUs

Theme		
Code	Subcode	Example quote
Perceived opportunities		
Potential to improve outcomes		<p>'Yeah. So, I do think that the artificial intelligence is needed in the ICU. And I'll tell you why. Because as a critical care physician, normally I have between ten and 12 patients that I have to take care. And then for me, it's almost impossible to keep track all of all the information that is being generated every single minute in the ICU...Every single minute, what is happening with this patient. There are many, many, many little variations in the vital signs, many little variations in the dose of the medications. But is practically impossible for a human to keep track of all of that. And then when that's in a 24 hours period then even more impossible...And sometimes in the ICU, we have this data overload. So, we really cannot handle. So, I do think that is important.' P5 ICU Physician LMIC</p> <p>'I mean, I would say there's definitely, definitely a huge, huge potential of improving patient, the outcomes by incorporating AI and patient data. And I mean, I can see it through some of the research that I'm doing. I can see it through some of the research that others are doing in the field of critical care in AI. There's so much data. It's probably one of the few fields that has so much data for individual patients and also for various patients.' P7 ICU Pharmacist LMIC</p> <p>'I think it's hugely useful. I think it's going to potentially improve efficiency, improve outcomes, standardize management a lot more.' P45 ICU Physician HIC</p>
Concerns about use		
Validity	Bias	<p>'I think some things that people might worry about whether it's representative, I guess, particularly in some of the communities that I have worked in, that whether AI is actually applicable to you know, indigenous people or people from different backgrounds. And I think that would provide some resistance to its uptake as well. And it would be a concern that I'd have as a clinician in terms of its validity and the people that I'm using it for.' P8 ICU Physician HIC</p> <p>'If you have developed your predictive model on a subset of patients that is some and-, it's somehow biased, it doesn't reflect all patients. You know, there can be racial or biases, or all sorts of potential ways that your predictive model doesn't apply to everybody. And so, but that's just about getting the science right.' P55 ICU Physician HIC</p>
	Generalisability	
Explainability	Essential	<p>'What's necessary in healthcare is explainable AI, we really need to know why the conclusion came up. Both for our own understanding for trust for the patient and the rest of the healthcare ecosystem, and also medical legal purposes, we need to know why the machine thought this was the right answer.' P1 ICU Physician LMIC</p> <p>'Whenever people don't understand how something works, it generates fear, it generate, you know, this feeling that they are not in charge. And that's something for a critical care physician, you need to allow them to be in charge, you know. And in general in medicine, you don't want to take away the decision capacity of the doctor. You don't want to do that. You want to provide a tool, you know, that you are not replacing the doctor. Just making sure that they're aware of the different alternatives. And at the end, they're the ones making the shots.' P5 ICU Physician LMIC</p> <p>'I think you should always understand exactly why you're doing something especially when you're working with people who are really, really unwell. If I always have to be able to explain why I've made a decision then probably the computer also should be able to explain why it's made a decision.' P28 ICU Nurse HIC</p>
	Not a key concern if it works	<p>'Well, let's put it this way. If the back box works...everybody's going to be happy. People aren't going to be happy once the black box stops working or once people use the back box for unintended cases.' P15 ICU Physician HIC</p> <p>'I mean, implementation is a difficult subject because AI, the better AI gets and the more powerful AI gets, the less we'll be able to understand why it comes to certain conclusions. And honestly, to me, this is actually one of the big benefits of AI, that AI can do things that we can't understand.' P20 ICU Physician HIC</p> <p>'This is a hard one, because it depends. I think it depends more about the trust that people have in the in the in the technology. I think if you have evidence that the system works, even if it's not explainable, I mean, I would personally be comfortable in using it. I think for most users, they would...I think it's because I'm biased. I'm biased towards AI. I believe there was some studies looking at how you influence people using AI based on how much explainability they have, and people tended to request more explainability to trust their recommendations. I think as a general rule, it is probably important to try to provide explanations.' P32 ICU Physician HIC</p> <p>'Coming from anesthesia, we have these black boxes. If you look at them for anesthesia monitoring, the precise algorithms are patented and we don't have a clue. We just get a number. If you're good, you can look at the raw EEG and sort of get an opinion whether the number is way off or the ballpark figure is correct. There are now monitors coming out looking at pain levels interoperatively. Same thing essentially, that's a black box. Being an anesthetist, I don't mind black boxes. I need to be aware when the black box could give wrong information. If you have a pain monitor that tells you everything is fine, and you've got a patient who's hypertensive and tachycardic, is this a situation where the monitor might have a problem or is the patient actually in pain or is it just a hemodynamic problem and it's not a pain issue? Being able to trust the machine to take these things apart would be extremely valuable. So I think the black box, if you're used to using it, if you have a feeling for the limits, I'm not too reluctant to use something like that. As I said, we are used to that.' P38 ICU Physician HIC</p> <p>'No, I don't think so. A lot of clinicians aren't data scientists and just don't have the fundamental knowledge to be able to interpret an AI machine learning model. We could say a lot about the current methods we use, for example, of patient monitoring. I couldn't accurately explain all the technology that goes into, for example, an arterial line. Yet I use the output for that and can understand fundamentally how it works. But I don't necessarily understand the full physics and everything that goes behind it. I would say the same is with AI.' P44 ICU Physician HIC</p> <p>'No, I don't. I think if there was enough data to prove that it was safe, then I would say I don't need to understand how it's doing it for it to work. There's so many things that we do at work that I don't understand. I don't really understand how a pulmonary artery catheter gives me data or how a dialysis machine does almost anything that it does. But I'm reassured by its safety profile and the rigorous processes of following up, its continuous safety monitoring. It wouldn't concern me that I didn't understand how it was doing what it's doing.' P45 ICU Physician HIC</p>

Continued

Table 3 Continued

Theme		
Responsibility	Dependent on use	<p>'I think if you look at the AI tool as a tool that would function in place of a clinician then definitely liability would be an issue. But if you look at it as a tool, in addition to all of the other tools that helps in decision-making, and helps in patient management, then there should not be that liability issue. It's kind of like saying, well, you know, you did the labs for the patient, and there's a lab error. You as a clinician, you should look at the full picture and make that decision that this doesn't match with the rest of the pieces that I have. So, I think again, if you look at it as that tool coming in and making that decision of saying, well, this patient has lung cancer or doesn't have lung cancer, and then you start chemotherapy right away based on that machine then that's where there is a liability issue that ideally, I think they should go hand in hand with the clinical, with the clinician's decision or with the full assessment of the patient.' P7 ICU Pharmacist LMIC</p> <p>'I think ultimately yeah, clinicians will have that accountability. So, it's about how we would be applying AI just as it would be the same as how we would be applying any other technology that we use in intensive care...But I think yeah, I think ultimately, it's still a technology. It's not a sort of sentient being. So, ultimately, I think there yeah, in the intensivists will still be responsible for whatever happens.' P8 ICU Physicians HIC</p>
Dependency	Deskilling	<p>'First thing that popped in my mind is probably the recent deskilling. A lot of things, even from paper charts to electronic system, is a big step up...We made a lot of things very automated almost. So say, for example, on paper chart, the doctors would want us to have to write down exactly what amount of drug and in what diluent to put in and what the dose range would be. So by practice, they would then be familiar with it. Whereas right now, all they have to do is select the drop click, and everything is prebuilt on the system for them. Probably one of the worries I would get is if we take that tool away from them, then would they then struggle to then perform what they should have been doing in the first place?' P33 ICU Pharmacist HIC</p> <p>'Yes. Drug calculation before everyone's doing everything on paper, and I think if the system was down, I personally I think I forgotten how to do some of the like noradrenaline, how to calculate it, like quickly on the spot, like I don't think we're-, before it was something we would do it every single hour. So, it was basically drilled in your head. And now on the computer, suddenly, if the system is done, it's like, what do I do now?' P35 and P36 ICU Nurses HICs</p>
	Dehumanisation	<p>'But I guess the other concern would be the dehumanization of it. So, I worry sometimes that the junior doctors are nowadays quite reliant on technology and looking at the screen rather than at the patient. And I think there is a risk of sort of going the other way. Forgetting that it's the patient there.' P3 ICU Physician HIC</p> <p>'Let's see, I think the dehumanization? I don't know if the term is clear, but that fear of being attended by machines, let's say, people always expect the decision or the face to be given to them by someone else. And we believe that, I think that, if used properly, artificial intelligence is going to help us so that people can take care of those things that have to be done because we don't have time, but people in general think that if we use artificial intelligence, it dehumanizes care. They distance themselves from the patient, and the patient feels that the caregivers are distancing themselves from the patient.' P10 ICU Nurse LMIC</p> <p>'I think efficiency's improved. It's so easy to search for a keyword in someone's medical record and find a specific entry from three months ago made by one person. You can access things from home or from your office without waiting physically for a file. As soon as patients come in through ED, you've got their history available. Lots of things like that. I think environmentally it's, I presume it's better, although I'm not actually sure what the environmental impact is of all the computers that are required to manage it. But certainly not wasting paper's useful. I think there's an efficiency in ordering a test and knowing that the receiver is going to get it immediately, and not waiting for a piece of paper to find its way down to radiology or something. Legibility of notes, and particularly medication charts I think has improved a lot. The things that I think have deteriorated, so I noticed junior medical staff and nursing staff spend so much time at the computer, often to the detriment of what's actually happening with the patient. Particularly patients who are awake or not as ill, that really just need more of a personal touch. I think that the computer becomes partly a distraction, but also just a job that takes a lot of time. Which speaks to the inefficiency of the system, I guess, and I'm sure there are better systems than ours. But ours, the nursing staff spend a lot of time looking at the computer screen rather than the patient and the surroundings.' P45 ICU Physician HIC</p>
Disparity	Will widen gap between rich and poor	<p>'Of course, the richer or affluent places, they're going to have more technology. They're going to have the ability to implement these things. These poor county hospitals and predominantly rural black places in the US, they're not going to be spending the money on that. They're not going to have it. For sure there's going to be disparities in the application and benefit from it. Until there's, like at some point in 50 years, every hospital will have a basic amount of EHR and technology. Once you get to that everyone has a certain basic amount, then these tools will be everywhere. But that's probably a long ways away.' P39 ICU Physician HIC</p>

AI, artificial intelligence; HIC, high-income country; ICU, intensive care unit; LMIC, lower-to-middle-income country.

computer screen to the detriment of personal care of the patient.

Disparity

Some participants were also concerned that there will be large disparities in the application and use of AI technology in ICUs, which is going to widen the gap between richer and poorer settings.

Barriers and challenges to implementing AI in ICUs

Three overarching barriers to implementing AI in ICUs emerged (table 4):

1. Digital infrastructure: Participants from both HICs and LMICs identified the current digital infrastructure of institutions as a major barrier. Most

participants reported that their institution has neither the technical capability (hardware and software) to capture the necessary data or run the algorithms, nor the staff with the right knowledge and skills to use the technology. Some participants in LMICs reported not even having a stable electricity supply. This pointed to ongoing structural problems in the organisation and delivery of healthcare, and many participants in both HICs and LMICs described how they worked in broken healthcare systems where funds were limited to varying degrees, and investing in digitalisation and AI is not a priority. They suggested that many decision-makers either did not understand the value of digital technologies for improving patient care or were

Table 4 Barriers and challenges for implementing AI in ICUs

Theme		
Code	Subcode	Example quote
Digital infrastructure	Technical capability lacking	<p>'We don't have the equipment. We need internet, we need trainings. We need to train people on that. I mean we need time to get used to it. At the same time, we need the computers in the department for that...But yeah, some of the things that would maybe delay implementing it will be like the equipment, maybe the orientation of staff on the equipment.' P2 ICU Nurse LMIC</p> <p>'If you don't have the infrastructure and the ability to gather the data and then run these algorithms on it, it's going to be difficult. I think though that most of Western Europe and the US, and rich countries, are going to have, in the near future, fairly completely electronic systems. [Country] trying to get there, we're just bit far behind...Obviously that's a barrier.' P39 ICU Physician HIC</p> <p>'The main obstacle is actually the availability of a hardware and software together that can make things possible. The second obstacle is the people who will know how to use it, and to activate it.' P43 ICU Physician HIC</p> <p>'In most low- and middle-income settings, I think this is, you know, when I'm working with my public health hat on access to high quality, reliable data is like the number one problem for public health research related, but also in this case, thinking about how to develop AI systems. And I think that is, without a question, the biggest challenge. Because it's not just a problem of like the fact that we're mainly working in paper-based records. But it's also the fact that, you know, our electricity comes in and out. Our monitors when they stop working people don't, you know, there are many, many layers to this to the point where we would be having reliable collection of data that can be used in this way. So, I think it's-, that's a challenge. And it's not simply the fact that we don't have an EHR to document data. There are many other components that feed into that.' P50 ICU Physician LMIC</p> <p>'The thing is if we want to try to integrate artificial intelligence into my hospital in particular, we don't have the technology so far. And the other big limitation, you know, is the electronic healthcare system. Because this piece of software is crazy expensive, and trying to integrate whatever you are generated from data perspective, that could be challenging....Yeah. So, what is missing is a way that the data is transferred directly from the ventilator to the electronic healthcare system. And for doing that, you need a piece of hardware and software that allows you to extract the data from the ventilator and put it in the chart. And the thing is, in Colombia, we have many different makers from for the ventilators, for the fusion bombs, all of that sometimes is tough...' P5 ICU Physician LMIC</p> <p>'But the biggest problem is the missing digitalization of the data, a lot of also University Hospitals in [Country] are working on paper, so they have no data available on a server. That's the biggest problem. (Question: Could you estimate how many hospitals would have currently the possibility of using AI technology?)Probably 20–25% of the hospitals actually...But I think this will change in the next five years, probably. We'll arrive probably to 50 or 60% of the hospitals who are used then have all the data available in digital form. But it depend also from the politicians.' P14 ICU Physician HIC</p> <p>'We probably haven't had the bandwidth to think about the other things. Because as you say, if you don't have digital data the rest is irrelevant. I would love to have digital data...But you know, let's start with some basic stuff first, which is a monitor that doesn't have to be turned into a paper record by a nurse every hour. I guess the advantage of paper is, you know that it's secure. It's physically in one place. It can't be accessed by anyone else. But you know, the disadvantage to that is it's secure and can't be accessed by anyone else. So, you certainly can't do analysis on it. Yeah, I think we're a long way behind.' P47 ICU Physician HIC</p>
	Insufficient funding available	<p>'The [Country] system is, the geography's huge. The patients they have to service, the land mass is huge. They haven't updated and put enough money into our system. We have two year waits for hip replacements, two year waits for cataract surgery, chronically underfunded. We've had aging population with mass migration on top of that...Our system is so broken that the electrification or EHR computerization of our healthcare system was not a high priority in the funding list. Some provinces have been more aggressive, the wealthier provinces...I think [Country], and certainly the less affluent parts of [Country] that don't have this basic infrastructure, AI is not a priority. There's no money going into that at all.' P39 ICU Physician HIC</p> <p>'So, the first thing is that we don't have enough money to implement any additional systems other than increasing capacity to provide critical care. And it's not to say that all these things are considered a nice to have. But if you gave me a million dollars, and said, you can either spend this on an AI learning system, or you can open another two ICU beds, I know what I'm going to choose, and it's not the AI learning system. So, I think the biggest barrier we currently have is simply funding.' P47 ICU Physician HIC</p> <p>'I would say that the biggest barrier in South American in general. Because for you to be able to collect everything into the dataset, then you need to have a piece of software that can extract data, transfer that to the system. And that for doing that, you need to invest money. You can imagine ICUs in South America, they don't have money to buy ventilators. Of course, they prefer not to invest money in a piece of software that they don't see how it can affect the patient care.' P5 ICU Physician LMIC</p>

Continued

Table 4 Continued

Theme		
Knowledge and understanding	Disconnect between clinicians and technical partners	'There are times where I wish as a clinician, I had the abilities to do everything myself, including like, data extrapolation, writing my own data. Because sometimes I feel like there is sometimes-, not a disconnect, but lack of connection between your data analyst or data scientist that helps you develop your project or helps you, like that's their project, and you're trying to help them that the-, my lack of understanding of artificial intelligence vs their lack of knowledge from clinician perspective, which is understandable for both parties, leads to sometimes like two sides not understanding each other as well as they should be.' P6 ICU Pharmacist HIC
	Insufficient focus on what is needed	'I think it would be good to see what people really want from AI. Instead of saying, okay, here's the new technology and here's what it can do. Maybe we need to think about, what is it that we're lacking in our current practice that we feel we need? And then see if AI can quickly fill that void. That may be a better way to push AI forward and gain acceptance. Rather than saying: "Hey, look, here's AI, it can do all these things."' P3 ICU Physician HIC 'So the idea that I have, we know the problems. IT doesn't know-, they don't know what problems we are suffering. So, the IT, they providing solutions for a problem, which are most of the time not existing in our eyes. So, this disconnection that makes things with doesn't match. Why you're advertising? Why you are selling me this machine? I don't need it.' P43 ICU Physician HIC
	Views or data ownership and competition	'Most of them don't share it because of commercial aspects and secrets. I have a lot of anger, it's not ethical, it's not their data, and...I think that it's not a commercial aspect, it's the ego aspect. They want to publish papers in the New England Journal or whatever. They want all the data.' P52 ICU Physician HIC 'I would say both. I would say data protection laws, and it's political. Competition between academic centers. We have the bigger data set. So if we open it, people are going to publish using our data. Very classical.' P32 ICU Physician HIC
Regulatory	Large variations in data protection	'It's not too much of a major issue in [Country]. If it's data that I'm pulling from my own institution, it's de-identified and I'm using it for research purpose. I think the issue becomes when you start doing multicenter studies and how do you kind of pull data from different institutions, and then it goes into one pool. There are some institutions that are more strict than others, and some countries that are more strict than others. So, I think this is where the difficulties come into. And I think countries and institutions are realizing that they need to be less strict about those criteria. Because, yes, we are protecting the privacy of our patients by having all these measures. But then at the same time, there's potential harm if you're doing all this to these restrictions, that there's no-, or there's minimal research that people are doing. So, you don't understand your patients fully. You don't give them the full care. You don't have research. So, it's, I mean, it's you have to kind of weigh things both ways. I'm not saying that you just kind of you know, open things and have all the data freely available. But at the same time, too many unnecessary restrictions, I think makes it difficult to conduct research, and then that has its own issues there. So, it's kind of a balance between both.' P7 ICU Pharmacist LMIC 'Yeah, I think it's a huge issue. Again, there's huge regional variation. So, in some states all the hospitals have a shared system, and you add a baseline to set a clinical level, you've got access to all the other hospital information. So, it makes things like linking data very easy. Because it is already linked within the state. The state that I'm in does not do that. And so, it's all separate, which already just makes it like practically very difficult. Let alone, you know, having to deal with de novo sort of ethics, submissions, and trying to actually get that all together.' P8 ICU Physician HIC

AI, artificial intelligence; HIC, high-income country; ICU, intensive care unit; LMICs, lower-to-middle-income countries.

too burdened by the existing financial strain on their health system.

2. Knowledge and understanding: Participants also identified a lack of knowledge and understanding about AI and the clinical context these tools will be implemented in as a significant barrier. Participants felt that this affected professionals' and patients' acceptance and willingness to use AI, and that the disconnect between clinicians and technical partners too often leads to non-optimal AI tools. Indeed, one participant described most AI applications as 'solutions looking for problems' that do not exist in the view of clinicians. Participants also reported that some colleagues' views about data ownership and competition led them to be unwilling to share data, which was also reported to be a substantial challenge that undermines AI implementation.
3. Regulatory: Large variations in regulations regarding data protection within and across countries were also highlighted by participants as an important barrier. Some institutions and countries were reported to be significantly stricter than others with regard to data

sharing and the secondary use of data. Although participants all agreed that protecting patient privacy was essential, they also felt that the current situation could potentially harm patients because it is undermining research and their ability to improve care.

Facilitators for implementing AI in ICUs

Three key suggestions for facilitating and improving the implementation of AI in ICUs emerged (table 5):

Demonstrating the value/limits of AI

Participants thought that clear and consistent evidence from robust research studies confirming the utility and reliability of AI applications would be the most important facilitator for increasing the acceptance of and willingness to use AI applications in ICUs. Participants also saw a need for a clear explanation of the strengths/weaknesses and advantages/disadvantages of each application.

Closing the gap of understanding

Participants made two main suggestions for improving the current gap of understanding between clinicians and

Table 5 Facilitators for implementing AI in ICUs

Theme		
Code	Subcode	Example quote
Demonstrating the value/limits of AI	Evidence of AI utility and reliability	<p>'I think some strong studies will help of course, if you can show that AI helps tremendously with prediction of mortality or prediction of hypotension or with mechanical ventilation with better outcomes. And those results are constantly produced and coming from different countries, then I think it will be difficult for a doctor to ignore it.' P27 ICU Physician HIC</p> <p>'I think more randomized control trials should be done to prove that these technology can really improve the patient's outcome, not just to choosing a model, but to see whether the model, the usefulness of the model, to improve reduce mortality or reduce costs. If we can prove least in the large RCTs, this will be a very important indication.' P29 ICU Physician LMIC</p> <p>'I think one thing that will help is very, very solid research. And not only by people wearing pink glasses and saying, "Oh it's so good." But really solid research. I think, given the current way we practice medicine, I think that's the best way to establish new techniques. The whole technical side. Well, people are all enthusiastic about it, that will happen. But really validating stuff and also making clear that once validated...What do we need? For instance, how many times do you need to update a model? All these kinds of things are below the surface, whereas I think they're extremely important.' P31 ICU Physician HIC</p> <p>'I think you'd have to prove its value. So whether that's making life easier for clinicians or making outcomes better for patients. I think you'd have to show that there's a value in it.' P45 ICU Physician HIC</p>
	Clear explanation of strengths/weaknesses and advantages/disadvantages of	<p>'As far as implementation goes, I think there has to be a clean explanation of the strengths and weaknesses and disadvantages of the software, that this is not like a God to predict everything for you. Like you still have to use your clinical knowledge and your brain basically, at the end of the day before making a decision. And don't just say purely, well because what the algorithm told me I have to do this. It's really how to utilize the algorithms rather than abusing them for the purposes of clinical decision-making processes.' P6 ICU Pharmacist HIC</p>
Closing the gap of understanding	Training and education	<p>'I think it's very important to start early, and they need to be primers on data science, machine learning and AI, right? From the med school level now, start educating people on what it's all about.' P1 ICU Physician LMIC</p> <p>'The way I see AI it's like a new discipline within medicine, for example. I as an ICU doctor, I will not go to the lab and challenge your potassium result. Right? So, I rely on the lab, making sure that that potassium result is correct. Measuring potassium is very difficult, like it's not an easy thing, you have to think about hemolysis, you have to think about the quality of the blood, you have to think about whether the blood has been standing long enough. You have to think about whether your reagents were correct, whether the controls were correct, whether your machine is competing correctly, whether it was the same, the right patient. So, there are so many things that we take for granted anyway, in reading a simple potassium result. So, I don't think that would be one of the things that I really concern me. So, I would not say: "Okay, I'm not going to use AI unless I understand all the algorithms." I don't think that would be the case, but certainly, I think a new discipline. So, for example, like a doctor trained in AI, who's working in the hospital with good confidence of other teams like about critical care physician to use it. So I would see, the way forward would be a new specialty within medicine, where they have specialist trained in artificial intelligence, who are equipped with the skills required to make valid predictions, make valid algorithms, and then I'll trust that person, rather than me trying to learn through it again, which would be impossible.' P3 ICU Physician HIC</p> <p>'I would say education is a large part of it, you know, getting people, not just clinicians, both data scientists, IT and clinician, everyone, to kind of understand the concept more, see the value of it, see what it means. And how do I actually apply it? How do I develop models? How can I incorporate that? And then that would get people a bit outside their comfort zone so that they can kind of take on the next steps sort of. I would say education would be a big part of it.' P7 ICU Pharmacist LMIC</p> <p>'I think it needs to become part of the curriculum that we teach critical care clinicians. I think clinical informatics in some way needs to be taught. And how, what we don't get taught is we use a lot of these IT systems and electronic medical records as clinicians, but we don't ever get taught around the backend. How do we use it? How do we actually use all this stuff that we put in and this data that we aggregate, but we barely touch. So, what, so for me, it's about gradually teaching people how to do it. And then as a result, if you do have these large data sets, what's the way of working with some expert in a local area or in a collaborative group to work on a project that develops something. And so, they're that sort of, they're the things that would in my mind, help people become, understand the issues, but also build a skill set that helps the next generation do it better.' P19 ICU Physician HIC</p>
	More inclusion of clinicians	<p>'And also, have discussions with people who develop these things so that they can interface and interact with doctors and nurses, and they don't build something which won't get used. So, a lot of the apps that are built are crazy good in terms of the technology. But no doctors were asked what they wanted.' P1 ICU Physician LMIC</p> <p>'It should be designed by the people who actually will use it and will need it. I think it cannot be designed by anybody who's not familiar with the processes...you need the collaboration between the end user and the programmer at the start.' P40 ICU Physician LMIC</p> <p>'This connection between the physicians and the IT guys. This is the most important thing. We have disconnected.' P43 ICU Physician HIC</p> <p>'Additionally, it ultimately depends on who builds the model that is used. But I feel like clinicians definitely should be involved.' P9 ICU Pharmacist HIC</p>

AI, artificial intelligence; HIC, high-income country; ICU, intensive care unit; LMICs, lower-to-middle-income countries .

technical partners, to increase the acceptance and the clinical utility of AI in ICUs:

- ▶ *Training and education:* Many participants noted the need to improve training and education of both clinicians and data scientists, so clinicians have a better understanding of AI concepts and data scientists understand the clinical context better. Although participants saw the need to improve all intensive care professionals' knowledge and skills in this area, some participants also advocated for a new (sub)specialty where clinicians are trained in AI as it was unrealistic to think that all clinicians could be trained to the required level.
- ▶ *More inclusion of clinicians:* Participants strongly felt that there needed to be more consultation and involvement of clinicians from the beginning in the design and development of AI applications for ICUs, to improve the connection between clinicians and developers and the resulting product.

Improving ecosystems

Participants also saw the need to improve the wider ecosystem, including: ensuring that there is a proper system of data collection and documentation, that funding bodies are aware of bottlenecks so funding is directed to efforts to translate research into practice rather than just generating more accurate prediction models, that grant panels have the right expertise to evaluate multidisciplinary research and enhance the potential for academic/commercial partnerships.

DISCUSSION

This is one of the largest qualitative studies to date to examine the views of intensive care professionals regarding the use and implementation of AI technologies in ICUs, involving 59 participants from 24 countries, including countries from Europe, Asia, North America, South America, Middle East, Australasia and sub-Saharan Africa. This study found general agreement among participants' views regarding the use and implementation of AI in ICUs, which were largely in line with existing empirical research with ICU professionals.^{13 14 19 20} Participants had generally positive views about the potential use of AI in ICUs but identified important technical and non-technical barriers to the implementation of AI. A key finding of this study, however, was important differences between ICUs regarding their current readiness to implement AI. It was striking that these differences were not primarily between HICs and LMICs as might be expected. Rather, the key difference was between a small number of ICUs in large tertiary hospitals in HICs, which were reported to have the necessary digital infrastructure for AI, and nearly all other ICUs in both HICs and LMICs, which were reported to neither have the technical capability to capture the necessary data or run AI algorithms, nor the staff with the right knowledge and skills to use the technology. Although technical barriers to implementing

AI in ICUs have been widely discussed,^{4-6 11 14} intensive care medicine needs to be careful not to gloss over the importance of the current readiness of ICUs to implement and use AI, otherwise it will risk building a house of cards. Pouring massive amounts of resources into developing AI without first (or in parallel) building the necessary digital and knowledge infrastructure foundation needed for AI is unethical.³² We do not see the possibility of real-world implementation and routine use of AI in the vast majority of ICUs in both HICs and LMICs included in our study any time soon, and we do not think this 'last mile' of implementation³³ will be reached unless the necessary digital and knowledge infrastructures are built first. We are of the view that ICUs should not be using AI until certain preconditions are met. Intensive care societies from around the world need to come together and reach a consensus on what these preconditions should be.

Limitations

This is a qualitative study that did not collect statistically representative data. However, we included a range of intensive care professionals from 24 HICs and LMICs, which makes it likely that this study has captured key aspects of a multisided issue. A bias might exist toward the reporting of socially desirable attitudes,³⁴ however, given our results that are rather critical of current practice, we believe that such a bias is limited. The study was carried out across 24 countries, and there may be some regional and country-specific differences that might limit the generalisability. Nevertheless, many of the key issues are associated with aspects that are common in all countries (eg, limited digital data collection and documentation, and an underutilisation of patient data in ICUs), these findings are likely to be of wider international interest. There is currently no established definition of what constitutes AI, and a definition of AI in medicine was not provided to participants. As noted in the results section there were large variations in knowledge of AI among participants, and concrete examples were provided where needed. However, this may have affected the ability of some participants with limited knowledge of AI to answer some questions. The study was also undertaken before the explosion of interest in the use of LLMs and the chatbots that they power. The AI discussed in this manuscript therefore does not include LLMs.

X Stuart McLennan @McLennanStuart

Collaborators The authors thank Dr. Beatrice Tiangco with her assistance with two interviews in the Philippines.

Contributors SM and LC developed the idea and design of the study. SM and AF conducted the interviews and analysed the data. The work was initially drafted by SM and revised for important intellectual content by AF and LC. All authors read and approved the final manuscript. SM is the guarantor of the study and accepts full responsibility for the work and/or the conduct of the study, had access to the data, and controlled the decision to publish.

Funding Research reported in this publication was supported by the National Institute Of Biomedical Imaging And Bioengineering of the National Institutes of Health under Award Number R01EB017205. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. LAC is funded by the National Institute of Health through R01

EB017205, DS-I Africa U54 TW012043-01 and Bridge2AI OT2D032701, and the National Science Foundation through ITEST #2148451.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study received approval (621/21 S) from the Technical University of Munich's Research Ethics Committee on 23 November 2021. 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. Our data include pseudonymised transcripts of interviews, which cannot be made publicly available in their entirety because of (1) the terms of our ethics approval; and (2) because participants could be identifiable if placed in the context of the entire transcript. This is in line with current ethical expectations for qualitative interview research. We provide anonymised quotes within the paper to illustrate our findings (corresponding to transcript excerpts), and the complete interview guide used in the study has been included as a Supplementary Information.

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 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iDs

Stuart McLennan <http://orcid.org/0000-0002-2019-6253>

Amelia Fiske <http://orcid.org/0000-0001-7207-6897>

Leo Anthony Celi <http://orcid.org/0000-0001-6712-6626>

REFERENCES

- Celi LA, Mark RG, Stone DJ, *et al*. Big data" in the intensive care unit. closing the data loop. *Am J Respir Crit Care Med* 2013;187:1157–60.
- Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.
- Pollard TJ, Johnson AEW, Raffa JD, *et al*. The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci Data* 2018;5:180178.
- Mamdani M, Slutsky AS. Artificial intelligence in intensive care medicine. *Intensive Care Med* 2021;47:147–9.
- Komorowski M. Artificial intelligence in intensive care: are we there yet? *Intensive Care Med* 2019;45:1298–300.
- Yoon JH, Pinsky MR, Clermont G. Artificial intelligence in critical care medicine. *Crit Care* 2022;26:75.
- Saqib M, Iftikhar M, Neha F, *et al*. n.d. Artificial intelligence in critical illness and its impact on patient care: a comprehensive review. *Front Med* 10:1176192.
- Gutierrez G. Artificial intelligence in the intensive care unit. *Crit Care* 2020;24:101.
- Hwang YJ, Kim GH, Kim MJ, *et al*. Deep learning-based monitoring technique for real-time intravenous medication bag status. *Biomed Eng Lett* 2023;13:1–10.
- Wardi G, Owens R, Josef C, *et al*. Bringing the promise of artificial intelligence to critical care: what the experience with sepsis analytics can teach us. *Crit Care Med* 2023;51:985–91.
- van de Sande D, van Genderen ME, Huiskens J, *et al*. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Med* 2021;47:750–60.
- Smit JM, Krijthe JH, van Bommel J, *et al*. The future of artificial intelligence in intensive care: moving from predictive to actionable AI. *Intensive Care Med* 2023;49:1114–6.
- Mlodzinski E, Wardi G, Viglione C, *et al*. Assessing barriers to implementation of machine learning and artificial intelligence-based tools in critical care: web-based survey study. *JMIR Perioper Med* 2023;6:e41056.
- D'Hondt E, Ashby TJ, Chakroun I, *et al*. Identifying and evaluating barriers for the implementation of machine learning in the intensive care unit. *Commun Med (Lond)* 2022;2:162:162:.
- Fleuren LM, Thoral P, Shillan D, *et al*. Machine learning in intensive care medicine: ready for take-off? *Intensive Care Med* 2020;46:1486–8.
- Komorowski M. Clinical management of sepsis can be improved by artificial intelligence: Yes. *Intensive Care Med* 2020;46:375–7.
- Tabah A, Bassetti M, Kollef MH, *et al*. Antimicrobial de-escalation in critically ill patients: a position statement from a task force of the European society of intensive care medicine (ESICM) and European society of clinical Microbiology and infectious diseases (ESCMID) critically ill patients study group (ESGCIP). *Intensive Care Med* 2020;46:245–65.
- McLennan S, Shaw D, Celi LA. The challenge of local consent requirements for global critical care databases. *Intensive Care Med* 2019;45:246–8.
- van der Meijden SL, de Hond AAH, Thoral PJ, *et al*. Perspectives on artificial intelligence-based clinical decision support tools: Preimplementation survey study Jmir hum factors. *JMIR Hum Factors* 2023;10:e39114.
- van de Sande D, van Genderen ME, Braaf H, *et al*. Moving towards clinical use of artificial intelligence in intensive care medicine: business as usual? *Intensive Care Med* 2022;48:1815–7.
- Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 2019;1:389–99.
- Ciecierski-Holmes T, Singh R, Axt M, *et al*. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. *NPJ Digit Med* 2022;5:162.
- Knight SR, Ots R, Maimbo M, *et al*. Systematic review of the use of big data to improve surgery in low- and middle-income countries. *Br J Surg* 2019;106:e62–72.
- Cinaroglu S. Big data to improve public health in Low- and middle-income countries: big public health data in Lmics. In: *Analytics, Operations, and Strategic Decision Making in the Public Sector*. 1st. IGI global.
- Wahl B, Cossy-Gantner A, Germann S, *et al*. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health* 2018;3:e000798.
- Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care* 2007;19:349–57.
- Palinkas LA, Horwitz SM, Green CA, *et al*. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Adm Policy Ment Health* 2015;42:533–44.
- United Nations. World population prospects 2022. In: *World Economic Situation and Prospects 2022*. New York: United Nations, Available: <https://www.un-ilibrary.org/content/books/9789210014380>
- Marshall MN. Sampling for qualitative research. *Fam Pract* 1996;13:522–5.
- Webster P. Six ways large language models are changing healthcare. *Nat Med* 2023;29:2969–71.
- Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res* 2005;15:1277–88.
- Bak M, Madai VI, Fritzsche M-C, *et al*. You can't have AI both ways: balancing health data privacy and access fairly. *Front Genet* 2022;13:929453.
- Coiera E. The last mile: where artificial intelligence meets reality. *J Med Internet Res* 2019;21:e16323.
- Bergen N, Labonté R. "Everything is perfect, and we have no problems": detecting and limiting social desirability bias in qualitative research. *Qual Health Res* 2020;30:783–92.

© 2024 Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY. Published by BMJ. <https://creativecommons.org/licenses/by/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.