Revista Mexicana de Neurociencia

PERMANYER

Check for updates

ORIGINAL ARTICLE

Use of supervised machine learning algorithms to detect pre-existing neurocognitive disorders in hospitalized adults with COVID-19: a classification study

*Juan M. Marquez-Romero[1], Svetlana V. Doubova[2], Flavio Cuellar-Roque[1,3], Carlos A. Prado-Aguilar[4], Alicia Alanis-Ocadiz[5], Jannett Padilla-López[6], Diana C. Navarro-Rodríguez[7], Carolina Quiñones-Villalobos[3], Ángel E. Muñoz-Zavala[8], and Rogelio Salinas-Gutiérrez[8]**

[1]Hospital General de Zona No 2, Instituto Mexicano del Seguro Social (IMSS), OOAD Aguascalientes, Aguascalientes; [2]Unidad de Investigación Epidemiológica y Servicios de Salud Centro Médico Nacional Siglo XXI, IMSS, Mexico City; [3]Programa de Maestría y Doctorado en Ciencias Médicas Odontológicas y de la Salud, Universidad Nacional Autónoma de México (UNAM), Mexico City; [4]Coordinación Medica Auxiliar de Investigación en Salud, IMMS, OOAD Aguascalientes, Aguascalientes; [5]Unidad de Medicina Familiar No 8, IMSS, OOAD Aguascalientes, Aguascalientes; [6]Unidad de Medicina Familiar No 1, IMSS, OOAD Aguascalientes, Aguascalientes; [7]Hospital General de Zona No 1, IMSS, OOAD Aguascalientes, Aguascalientes; [8]Departamento de Estadística, Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes, Aguascalientes. Mexico

Abstract

Objective: *The objective of the study is to determine the performance of three algorithms to classify patients with neurocognitive disorders during the pre-COVID-19 period.* **Methods:** *Data come from the electronic health records (EHRs) of patients hospitalized for COVID-19 aged > 18. EHRS entries before hospitalization were screened by an automated Python script that searched for evidence of neurocognitive testing during the pre-pandemic period. The neurocognitive disorder was determined based on the results of the neurocognitive testing. Machine learning (ML) classifiers (logistic regression [LR], k-nearest neighbors [k-NN], linear discriminant analysis [LDA]) were trained with the frequencies of appearance in the EHRs of a list of nine keywords obtained through a Delphi panel. The ten-fold cross-validated performance of classifiers was evaluated.* **Results***: 3,674 EHRs were screened; 205 featured a pre-pandemic neurocognitive examination, and 127 were positive for neurocognitive disorder. The mean age was 81.8 ± 6.8 years (76% women). The mean number of keywords in the EHRs was 17.4. The most frequent keyword was a referral for suspicion of dementia. The three classifiers performed comparably in 12 out of 13 performance measures. Area under the curve (AUCs) were statistically different, with LR achieving the highest AUC (0.878), followed by k-NN (0.590) and LDA (0.524), p < 0.001.* **Conclusions:** *The three ML algorithms can accurately predict the presence of neurocognitive disorders in a sample of patients hospitalized for COVID-19 with consistent performance. The proposed approach can be used to determine the cognitive status of patients in various medical settings, including the pre-pandemic cognitive status of patients with suspected post-COVID-19 neurocognitive disorders.*

Keywords: *Neurocognitive disorders. COVID-19. Machine learning. Electronic health records.*

***Correspondence:**
Rogelio Salinas-Gutiérrez
E-mail: rogelio.salinas@edu.uaa.mx

Date of reception: 01-10-2025
Date of acceptance: 20-11-2025
DOI: 10.24875/RMN.25000068

Available online: 01-06-2026
Rev Mex Neuroci. 2026;27(4):179-188
www.revmexneurociencia.com

2604-6180 / © 2025 Academia Mexicana de Neurología A.C. Published by Permanyer. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Uso de algoritmos de aprendizaje automático para la detección de trastornos neurocognitivos preexistentes en adultos hospitalizados con COVID-19: un estudio de clasificación

Resumen

Objetivo: *Determinar el desempeño de tres algoritmos para clasificar a pacientes con trastornos neurocognitivos durante el periodo previo a la pandemia de COVID-19.* **Métodos:** *Los datos provienen de los expedientes electrónicos de salud (EES) de pacientes hospitalizados por COVID-19 mayores de 18 años. Las entradas de los EES previas a la hospitalización fueron analizadas mediante un script automatizado en Python que buscó evidencia de evaluaciones neurocognitivas realizadas antes de la pandemia. El diagnóstico de trastorno neurocognitivo se determinó con base en los resultados de dichas evaluaciones. Se entrenaron tres clasificadores de aprendizaje automático, regresión logística (RL), k-vecinos más cercanos (k-NN) y análisis discriminante lineal (ADL), utilizando las frecuencias de aparición en los EES de una lista de nueve palabras clave obtenidas mediante un panel Delphi. Se evaluó el desempeño de los clasificadores mediante validación cruzada de diez iteraciones.* **Resultados:** *Se analizaron 3,674 EES; 205 incluían una evaluación neurocognitiva pre-pandémica y 127 fueron positivas para trastorno neurocognitivo. La edad media fue de 81.8 ± 6.8 años (76 % mujeres). El número medio de palabras clave en los EES fue 17.4. La palabra clave más frecuente fue "referencia por sospecha de demencia." Los tres clasificadores mostraron un desempeño comparable en 12 de 13 métricas de evaluación. Los valores del área bajo la curva (AUC) fueron significativamente diferentes, con la RL alcanzando la AUC más alta (0.878), seguida de k-NN (0.590) y ADL (0.524), $p < 0.001$.* **Conclusiones:** *Los tres algoritmos de aprendizaje automático predijeron con precisión la presencia de trastornos neurocognitivos en una muestra de pacientes hospitalizados por COVID-19, mostrando un desempeño consistente. El enfoque propuesto puede utilizarse para determinar el estado cognitivo de los pacientes en distintos entornos médicos, incluyendo la evaluación del estado cognitivo pre-pandemia en pacientes con sospecha de trastornos neurocognitivos posteriores al COVID-19.*

Palabras clave: *Trastornos neurocognitivos. COVID-19. Aprendizaje automático. Expedientes electrónicos de salud.*

Introduction

In the early aftermath of the COVID-19 pandemic, health service providers have been increasingly concerned that severe acute respiratory syndrome coronavirus 2 infection might be associated with a higher risk for new-onset neurodegenerative diseases in recovered COVID-19 patients[1]. Consequently, the resulting disease burden could disproportionately affect the primary care settings by increasing health services demand[2]. Therefore, it is crucial to identify patients with post-COVID-19 sequelae manifested as non-acute neurocognitive disorders to understand the magnitude of this burden[3]. With this knowledge, it would be possible to escalate the primary care and rehabilitation services to adapt them to the new requirements and achieve adequate resource allocation to improve long-term outcomes, especially in patients requiring hospitalization during the acute phase of COVID-19 infection[4]. Nowadays, the prevalence and the type of post-COVID-19 symptoms vary widely, with the most frequent fatigue and cognitive affectations, yet the exact rate of symptoms remains unknown[5].

One of the critical aspects in determining the true incidence of post-COVID-19 neurocognitive disorders is the patients' pre-pandemic cognitive status. Unfortunately, these data are seldom reported due to the scarcity of pre-pandemic cognitive assessments[6]. Still, there appear to be remarkable similarities between the prevalence and type of symptoms among post-COVID-19 and critical illness survivor cohorts[7].

When applied to electronic health records (EHRs) databases, machine learning (ML) techniques offer a promising avenue for assessing patients' cognitive status. By analyzing EHRs information, including patient demographics and medical histories (e.g., diagnoses, cognitive assessments), ML algorithms help recognize patterns and correlations that may help healthcare professionals identify cognitive status before the onset of the pandemic. This data-driven approach enables early detection of cognitive decline and contributes to a more comprehensive understanding of the factors influencing cognitive health, ultimately facilitating targeted interventions and improving patient healthcare[8].

The present study aimed to determine the performance of ML algorithms to classify pre-pandemic entries of EHRs with and without mild and major neurocognitive disorders in a sample of EHRs of patients who required hospitalization for COVID-19.

Materials and methods

Study design

This was a classification study.

Setting

The study used patients' information from the EHRs system of Family Medicine clinics.

Data source and format

The EHRs system was established in 2003. It currently holds 36 million EHRs from over 1,200 Family Medicine Clinics distributed across the 32 territories. The EHRs system stores clinical records generated with each interaction between a patient and the healthcare providers, including medical diagnoses, referrals to secondary care, laboratory and radiology testing, and drug prescriptions. Data are captured directly by the family physician during the consultation. Data fields contain free-text fields for clinical notes, numeric fields for age, weight, blood pressure, etc., and pre-coded, menu-driven entries fields, such as the International Classification of Diseases, Tenth Revision code for patients' diagnoses, menu-driven electronic prescriptions that use generic product name and linked to the institutional essential list of medicines and clinic's pharmacy inventory, as well as laboratory database connected to the menu-driven, coded catalog of institutional laboratory tests.

Selection of the study population and data acquisition

To identify the study population, we first obtained a complete list of all hospitalized COVID-19 patients in one of the hospitals. We selected patients older than 18 years, with at least 6 months of pre-pandemic records (up to 18 months), including cognitive testing: Cognistat Cognitive Assessment, Addenbrooke's Cognitive Examination III (ACE-III), Montreal cognitive assessment (MoCA), or mini-mental state examination (MMSE). A Neurology, Psychiatry, or Geriatrics specialist performed the cognitive testing and confirmed or ruled out mild or major neurocognitive disorder following the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition of the American Psychiatric Association (DSM-5)[9].

As part of this study, each selected patient record was accessed using the Selenium package[10] for Python (version 3.7.1). The script opened the file and scanned every HTML element in each entry. After gathering, the data were stored as a Python object of the Beautiful Soup 4 library[11]. Within the object, the script searched for the occurrence of any of the cognitive tests used to select the study population, including all possible variations (abbreviations, misspellings, and acronyms). All the files identified by the script were then manually reviewed by one of the investigators to corroborate the inclusion criteria. In addition, a random sample of the files (1%, n = 36) that, according to the script, did not fulfill the inclusion criteria was manually reviewed. The list of files to manually review was generated using the base R sample function.

Between August of 2020 and July of 2022, 3,674 patients required hospitalization for COVID-19. After the automated review of these patients' EHRs, 3469 did not have standardized neuropsychological testing, and 205 patients (5.6%) had at least one standardized neuropsychological test. Within this group of patients, the cases were defined as those confirmed cases that fulfilled the DSM-V criteria for mild or major neurocognitive disorder before requiring hospitalization for COVID-19. The eligible controls were patients who had undergone standardized neuropsychological testing before requiring hospitalization for COVID-19 and were not diagnosed with neurocognitive disorder. The study sample comprised 127 (62%) cases and 78 (38%) controls.

Procedures

Ten experts in neurocognitive disorders reviewed a starting list of 18 keywords and concepts following the modified Delphi panel method[12]. The starting list contained the elements proposed by Ford et al.[13] In the first round, one keyword was eliminated (MMSE), and three keywords were added, (the translation to Spanish of three standardized neurocognitive tests (MMSE, ACE-III, and MoCA). A 70% or higher consensus was reached in the second round for four keywords and five concepts. The final list included the concepts: prescription of neurocognitive disorder medication, referral for suspicion of cognitive impairment, referral for suspicion of dementia, referral for neurocognitive testing, referral for a specialist in neurocognitive disorders consultation, and the following keywords: cognitive impairment, cognition, cognitive testing, and forgetfulness.

A second Python script searched the 205 Beautiful Soup objects containing the entire electronic file of each patient. The script specifically searched for the nine keywords and concepts of the list generated through expert consensus (the search included all possible variations of the keywords and concepts, such as abbreviations, misspellings, and acronyms). Finally, the script returned a database that contained basic demographics (age, sex), an anonymized identification tag, the case/control tag, and the count of the occurrences for each keyword and concept; this process was completely automatic. Figure 1 summarizes this process.

Analysis

Due to its widespread utilization and acceptance in the ML community, this study tested two statistically based (logistic regression [LR] and linear discriminant analysis [LDA]) and one geometrically based non-parametric classifier (k-nearest neighbors' algorithm [k-NN]). Both LR and LDA are rooted in statistical principles and are widely used in practice for classification tasks in medicine. They provide a formal statistical framework for making inferences about the relationships between variables and assessing their significance. In contrast, k-NN does not make explicit assumptions about the underlying data distribution. The classification process is based on the majority class among the k-NN, allowing it to adapt to intricate patterns in the data. The classifiers were implemented in the R programming language (version 4.2.3).

Stepwise LR was implemented with base R "lgm" function using the Akaike Information Criterion of the "MASS" package[14]. For the k-NN, the number of neighbors consisted of a range from zero to the integer of the square root of n (in this case, 14). The balanced accuracy was calculated for each k-value. The k-value with the highest balanced accuracy was used to evaluate the performance of k-NNN in the analysis. Finally, we used the "MASS" package "lda" function to run the LDA. Multiple k-fold cross-validations were performed with the seed set to 202309 for reproducibility. The number of folds was assigned to 10. Using the "createFolds" function of the "caret" package[15], independent folds were created, and the number of instances in each fold was set dynamically based on the sample size and k, and ranged between 20 and 21 instances per fold to maintain the balance between cases/controls within each fold.

Performance metrics

Overall accuracies with 95% confidence interval (CI), unweighted Cohen's kappa, sensitivity, specificity, positive predictive value, negative predictive value, precision (true positives/true positives + false positives), recall (true positives/true positives + false negatives), F1 (2*precision*recall/precision + recall), prevalence (cases/cases + controls), detection rate (true positives/cases + controls), detection prevalence (true positives + false positives/cases + controls), and balanced accuracy (sensitivity + specificity/2) were calculated for each fold. We calculated and reported the mean for each metric from the values obtained in each fold with the "confusionMatrix" function of the "caret" package. The primary outcome of this study was the cross-validated area under the curve (AUC) for each classifier. AUC was calculated with the package "cvAUC"[16] and plotted with the base R plot function.

Results

The mean age of the sample was 80.7 ± 6.7 years. There were 153 women (74.6%) with a mean age of 80.5 ± 6.6 years and 52 men with a mean age of 81.4 ± 7.1 years. The mean number of concepts recorded per patient was 3.9 for controls and 17.4 for cases (range 0-49). Table 1 shows the frequencies of concepts and keywords among cases and controls.

The most frequent concept was a referral for suspicion of dementia, identified in nearly 100% of all participants. All the other concepts were seen at a much higher rate in the cases group compared to the controls.

After manually reviewing the 205 EHRs identified by the Python script, the 127 identified cases contained more than one instance of the list of concepts and keywords; thus, there were no false positives. In the same manner, none of the 78 EHRs from controls contained any instance from the list, and thus, there were no false negatives within the sample. In addition, to evaluate the presence of false negatives outside the sample, we conducted a manual review of a random sample of 36 EHRs not flagged by the Python script. We did not find any instance from the list.

After training with the frequencies of the concepts and keywords, the three classifiers were similarly capable of classifying the patients' records as with or without neurocognitive disorders of any kind. There were no differences in accuracy between the classifiers, k-NN 0.83 (95% CI 0.60-0.95), LR 0.82 (95% CI 0.60-0.94), and LDA 0.74 (95% CI 0.52-0.89), p = 0.266. The

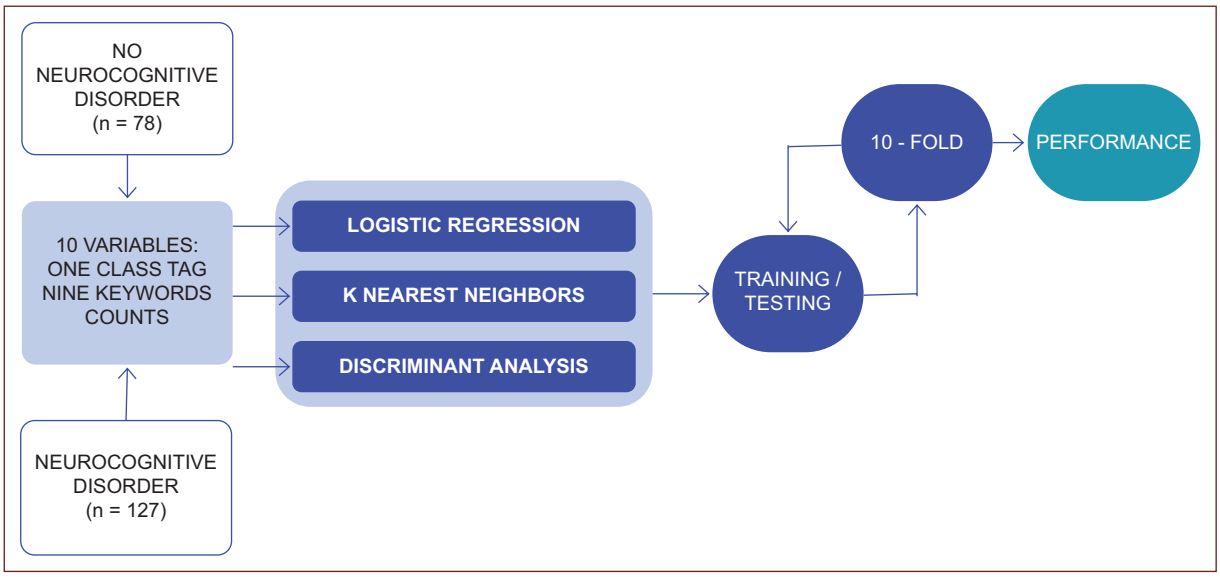


Figure 1. Flowchart of the study.

Table 1. Frequency of keywords among cases and controls

Concepts	Controls (n = 78)	Cases (n = 127)	p
	n (%)	n (%)	
Referral for suspicion of dementia	76 (97.4)	127 (100)	0.853*
Referral for neurocognitive testing	16 (20.5)	108 (85.0)	< 0.001*
Referral for a specialist in neurocognitive disorders consultation	25 (32.1)	79 (62.2)	0.001*
Referral for suspicion of cognitive impairment	2 (2.6)	18 (14.2)	0.004**
Prescription of neurocognitive disorder medication	1 (1.3)	14 (11.0)	0.005**
Keywords			
Cognitive testing	47 (60.3)	112 (88.2)	0.022*
Cognition	13 (16.7)	85 (66.9)	< 0.001*
Forgetfulness	7 (9.0)	50 (39.4)	< 0.001*
Cognitive impairment	6 (7.7)	38 (29.9)	< 0.001*

*Chi-squared test.
**Fisher's exact test.

agreement in the classification between folds was substantial for two of the classifiers (Cohen's kappa for k-NN = 0.64 ± 0.15 and LR = 0.63 ± 0.25) and moderate for LDA (Cohen's kappa = 0.46 ± 0.32) (Table 2).

Overall, LG and k-NN outperformed LDA in all the calculated metrics, as shown in table 2. However, the values obtained by these two classifiers were all equal or within a decimal point of each other, and there were no differences in performance between the three evaluated classifiers.

Regarding the primary outcome, the 10-fold cross-validated AUCs were 0.878, 0.590, and 0.524 for LR, k-NN, and LDA, respectively, p < 0.001. Figure 2 shows the AUC plot for each classifier. Across the folds, the performance of LR was more consistent and homogeneous, accounting for a significantly higher cross-validated AUC. In contrast, LDA's best fold was similar to the fold with the lowest performance of LR. The pattern of k-NN throughout the folds was intermediate between LR and LDA.

Table 2. Comparison of the performance metrics accomplished by each classifier

Performance metric	Logistic regression	k-Nearest neighbors	Linear discriminant analysis	p*
	k = 10	k = 10	k = 10	
Sensitivity	0.82 ± 0.17	0.82 ± 0.15	0.66 ± 0.21	0.084
Specificity	0.82 ± 0.16	0.83 ± 0.08	0.79 ± 0.20	0.845
PPV	0.76 ± 0.16	0.76 ± 0.09	0.70 ± 0.21	0.609
NPV	0.88 ± 0.12	0.89 ± 0.09	0.79 ± 0.13	0.108
Precision	0.76 ± 0.16	0.76 ± 0.09	0.70 ± 0.21	0.609
Recall	0.82 ± 0.17	0.82 ± 0.15	0.66 ± 0.21	0.084
F1	0.78 ± 0.14	0.78 ± 0.10	0.66 ± 0.19	0.137
Prevalence	0.38 ± 0.01	0.38 ± 0.01	0.38 ± 0.01	NA
Detection rate	0.31 ± 0.06	0.31 ± 0.06	0.25 ± 0.08	0.065
Detection prevalence	0.42 ± 0.10	0.41 ± 0.08	0.38 ± 0.12	0.564
Balanced accuracy	0.82 ± 0.13	0.83 ± 0.08	0.73 ± 0.16	0.174

All values mean ± standard deviation.
*One-way ANOVA.
PPV: positive predictive value; NPV: negative predictive value; k: number of folds (data subsets) used in the cross-validation process; N/A: not applicable.

Discussion

In the present study, we were able to classify patients with minor and major neurocognitive disorders from a sample of patients hospitalized for COVID-19 by applying a combination of data scraping of an EHR system and a modified Delphi panel method, and developing and testing a set of concepts and keywords. The statistical models performed with high balanced accuracy, ranging from 73% to 83%, and AUC, ranging from 0.524 to 0.878, after performing 10-fold cross-validation.

The study results underline potential applications for our approach in medical settings, especially in patients with COVID-19, as the pre-pandemic indicators of cognitive function have been previously related to a higher risk of hospitalization in cognitively disabled[17] and non-disabled[18] populations and higher mortality[19]. Nevertheless, the role of COVID-19 in developing new cognitive disorders after the acute phase of the infection is still under research[20,21]. One key factor for establishing a role of COVID-19 in the pathogenesis of non-acute neurocognitive disorders is the patient's pre-pandemic cognitive status, which is unknown in most studies that have associated COVID-19 with neurocognitive disorders[22].

One key factor in studying neurocognitive disorders after COVID-19 is the criteria used to establish the diagnosis; in this regard, it is possible that the results of the present study could be different if alternative diagnostic criteria were used. Nevertheless, the lack of a universal characterization of the cognitive effects of COVID-19 has an impact on the adequate identification of standardized criteria for neurocognitive disorders after COVID-19 and multiple neurological disorders[23]. This, in turn, produces that clinicians working in many clinics (especially in America) are limited to criteria set forth by the American Psychiatric Association in the DSM-5 to establish the diagnosis of neurocognitive disorder, which are known to be broader in their categorization of neurocognitive disorders when compared to other sets of diagnostic criteria such as DSM-IV dementia and MCI or expert diagnosis[24]. This finding has been replicated in other populations with neurological disorders; for example, a study of patients with multiple sclerosis, using the DSM-5 Criteria, identified significantly more patients as impaired compared to current research practices; the authors concluded that this approach could yield diagnoses that are less reliable or stable over time. On the other hand, other investigators have argued that the DSM-5 approach leads to earlier identification of neurocognitive disorders so that interventions can occur sooner[25].

The ML strategy to identify the cognitive status of undiagnosed patients with neurocognitive disorders is a new[26] and valuable approach[27,28]. Recent efforts have also included the automation of the data acquisition; for example, Ford et al.,[13] utilized proprietary software,

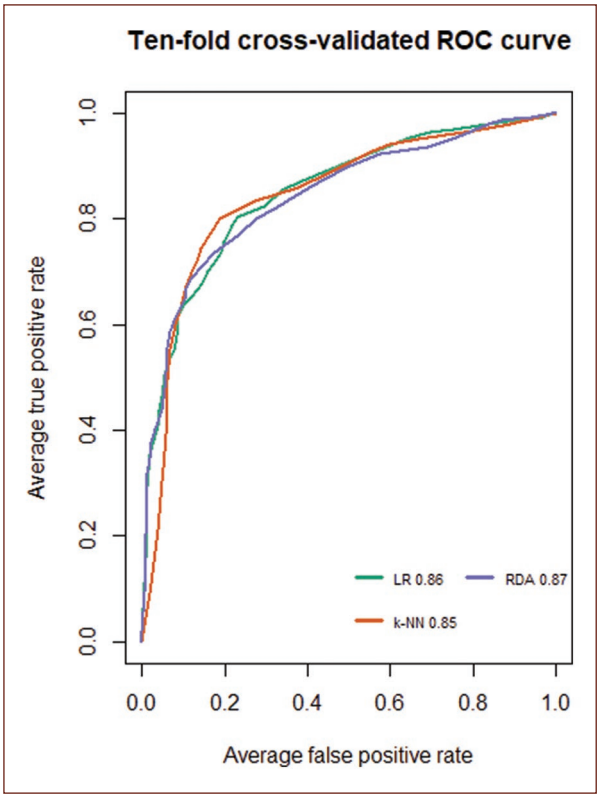


Figure 2. 10-fold cross-validated receiver operating characteristic curves for the classifiers.

automatically searched the UK Clinical Practice Research Datalink for eight coded and nine keyword concepts and used the obtained data to train three ML classifiers (LR, naïve Bayes and, random forest) to discriminate patients with dementia diagnosis code. Their results showed an AUC between 0.90 and 0.94 without differences between classifiers, demonstrating a remarkable performance from the three trained classifiers.

The performance achieved in the present study is lower than reported by Ford et al., with several aspects accounting for the differences. First, the researchers from the United Kingdom used a structured hierarchical vocabulary to capture medical data, compared to our study, where EHRs free-text information was mainly used. The data source could improve the British model's quality and performance. Even though we tried to minimize the impact of the free-text information by including known and unknown variations of the keywords and concepts (including misspellings), some instances might have gone unnoticed by the search script.

We also consider that overreliance on codified data is a significant difficulty for the widespread utilization of automated data collection, especially in EHRs systems that use predominantly free text. Although it is highly desirable for the sake of clarity and systematization of data registry, this approach is yet to be the most prevalent in real-world EHRs systems. A recent literature review of the automated detection of sepsis using data from EMRs showed that all the detection algorithms studied scanned data from discrete fields rather than from free text[29]. In another study, despite the data being extracted from the highly standardized EHRs from a cohort study by the same group of investigators, the trained models achieved similar results as in the present study (accuracy: 68.06%, sensitivity: 88.76%, specificity: 40.74%, F1-score: 75.96%, PPV: 66.36%, and NPV: 73.33%.)[30].

Second, using AUC as the sole measure in the British study shows an incomplete picture of the classifier's performance. In contrast, our results show that out of 13 performance metrics, only AUC showed statistically significant differences between the classifiers. It is worth noting that Ford et al. state that the accuracy of classifiers was assessed using AUC; however, accuracy as a metric is the proportion of correctly classified instances out of the total number of instances in a dataset, and in the context of binary classification, accuracy is calculated as the ratio of correctly predicted instances to the total number of instances[31]. Therefore, although AUC is generally preferred over accuracy as an indicator of model performance, they are not the same metrics, and it is advisable to report both instead of only one[31,32]. Despite our acknowledgment that AUC offers advantages over other metrics, such as accuracy, sensitivity, or predictive values, we consider that calculating and reporting multiple performance metrics should be the rule rather than the exception. In the Despins review[29], only the positive and negative predictive values were reported in 13 studies included, and 10 studies reported sensitivity and specificity in addition to the predictive values.

In addition, including additional performance metrics can aid in the decision-making process, depending on the specific needs of the studies. For example, precision is preferable when the costs of a false positive are high. For instance, labeling a patient with a neurocognitive disorder will lead to the patient receiving a diagnosis with all the psychological and medical implications associated. On the other hand, recall is preferable when a high cost is associated with false negatives. In addition, there is the possibility of finding an optimal balance between false positives and false negatives by calculating and reporting the F1 score, which can also

account for uneven class distribution, as seen in the studied dataset[33]. This is why we believe that reporting multiple classifiers with several performance metrics will eventually be the standard for classification studies. For example, a recent study aimed to develop a ML approach to assist in the detection of patients with systemic lupus erythematosus using genomic data and EHRs trained six classifiers and reported seven performance metrics, finding results similar to ours with no significative differences between the performance of the distinct classifiers (precision ranged from 0.6429 for support vector machines to 0.7746 for random forests)[34].

Finally, in the present study, cross-validation significantly improved the probability of revealing the actual performance of the classifiers, which may, nevertheless, be lower than what would be observed when using a simple train-test split. This can sometimes be perceived as a "lower performance" because it may uncover issues like overfitting that are not apparent with a single train-test split approach[35]. Moreover, overfitting is often a neglected issue associated with selection bias, especially in data from hospital records. For example, in a study of patient records from a single consolidated EHRs to classify patients with diabetes mellitus as type 1 or type 2, an optimized algorithm that incorporated the ratio of type 1 versus type 2 codes, plasma C-peptide, autoantibody levels, and suggestive prescriptions of antidiabetic medications achieved 100% accuracy. Nevertheless, since all data originated from a private multisite practice based in Eastern Massachusetts, it is highly improbable that the same algorithm can be applied to other populations[36]. Another study performed in patients with cardiac amyloidosis had a similar source for EHRs[37].

Besides the previously described strengths, our study also has certain limitations that deserve consideration. First, the accuracy and completeness of EHRs may vary. Furthermore, issues such as missing data, inaccuracies, or biases in the EHRs could have negatively affected the reliability of the results in this study. It is also worth noting that this study's findings are specific to our family medicine clinics and patient population. As mentioned above, important differences exist between the organization and functioning of EHRs across regions and countries. Therefore, we do not advocate extrapolating our results to a different demographic. Another aspect that limits the generalizability of the results is that our study focused on hospitalized patients; therefore, the relevance of the findings to patients with different severities of COVID-19 or those who did not require hospitalization is not addressed. Although cognitive deficits can occur after COVID-19 with mild symptoms, as well as in patients with severe forms of disease and requiring intensive care unit (ICU) admission[38], cognitive symptoms are much more frequent in older patients with severe infection or the presence of multiple comorbidities[39]. Given the apparent differences in the demographics of patients with mild and severe forms of COVID-19, in this study, we decided to limit the population to the most affected, leaving the study of patients with mild forms of COVID-19 to further studies. Still, we believe that the automation of the gathering of data can alleviate the cited limitations in generalizability. Using an automated approach, it is possible to adapt the process to different clinics and populations, changing parameters and accounting for regionalisms, language differences, and other particularities in how data is collected in EHRs. In this regard, although the results of the present study are challenging to generalize, the automated approach and the use of several classifiers are highly generalizable due to their high customizability. This customizability constitutes the novel contribution of our study.

Regarding the data analysis, there are also limitations; for example, there was a slight imbalance (1: 2.6) in the dataset, with the minority class being the class of interest, and as a result, the classifiers struggled to distinguish the minority class from the majority class. However, as shown in table 2, metrics such as precision, recall, and F1, often considered more informative for assessing model performance in class imbalance, showed adequate performance for two of the three implemented classifiers. In addition, the AUC, resilient to class imbalance, showed the same pattern.

A further limitation concerns the temporality of the data collection, which restricts the generalizability of the statistical models tested for future application because, after the COVID-19 pandemic, the awareness of cognitive symptoms as a sequelae of COVID-19 has undoubtedly increased both among healthcare professionals and among the general population; therefore, the propensity of general practitioners to interrogate and register the information related to the possible neurocognitive disorders might have also increased. This could lead to an increase in the frequency with which the general practitioners report the items of our list of concepts and keywords in the EHRs and to a possible diminution of the discriminative power of our models. Therefore, future projects should incorporate additional data generated after the COVID-19 pandemic to reassess the model's performance (retraining).

Conclusion

This study exemplifies a novel approach to automated data collection from EHRs, which, combined with a list of concepts and keywords, can accurately predict the presence of neurocognitive disorders in a sample of patients hospitalized for COVID-19. Additionally, three ML algorithms were trained with the identified data and extensively tested for finding consistent performance between classifiers, with the exception being a higher performance by LR when assessed by AUC. These results support the utilization of the proposed approach for determining the cognitive status of patients in various medical settings, including, in a unique manner, the pre-pandemic cognitive status of patients with neurocognitive disorders attributed to COVID-19.

Funding

The authors declare that this work was carried out with the authors' own resources.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Ethical considerations

Protection of human subjects and animals. The authors declare to have followed the ethical standards of the relevant experimentation committee, according to the World Medical Association and the Declaration of Helsinki. The procedures were approved by the institutional Ethics Committee.

Confidentiality, informed consent, and ethical approval. The authors have obtained approval from the Ethics Committee for the analysis of routinely collected and anonymized clinical data; therefore, individual informed consent was not required. Relevant ethical recommendations have been followed.

Declaration on the use of artificial intelligence. The authors declare that no generative artificial intelligence was used in the writing or creation of the content of this manuscript.

References

1. Rahmati M, Yon DK, Lee SW, Soysal P, Koyanagi A, Jacob L, et al. New-onset neurodegenerative diseases as long-term sequelae of SARS-CoV-2 infection: a systematic review and meta-analysis. J Med Virol. 2023;95:e28909.
2. Chen X, Bakal J, Whitten T, Waldie B, Ho C, Wright P, et al. Assessing the impact of COVID-19 pandemic on the health of residents and the healthcare system in Alberta, Canada: an observational study-the Alberta POST-COVID follow-up study. BMJ Open. 2023;13:e067449.
3. Rahmati M, Udeh R, Yon DK, Lee SW, Dolja-Gore X, Mc EM, et al. A systematic review and meta-analysis of long-term sequelae of COVID-19 2-year after SARS-CoV-2 infection: a call to action for neurological, physical, and psychological sciences. J Med Virol. 2023;95:e28852.
4. Wolff Sagy Y, Feldhamer I, Brammli-Greenberg S, Lavie G. Estimating the economic burden of long-COVID: the additive cost of healthcare utilisation among COVID-19 recoverees in Israel. BMJ Glob Health. 2023;8:e012588.
5. Writing Committee for the COMEBAC Study Group, Morin L, Savale L, Pham T, Colle R, Figueiredo S, et al. Four-month clinical status of a cohort of patients after hospitalization for COVID-19. JAMA. 2021;325:1525-34.
6. Del Brutto OH, Wu S, Mera RM, Costa AF, Recalde BY, Issa NP. Cognitive decline among individuals with history of mild symptomatic SARS-CoV-2 infection: a longitudinal prospective study nested to a population cohort. Eur J Neurol. 2021;28:3245-53.
7. Prescott HC. Outcomes for patients following hospitalization for COVID-19. JAMA. 2021;325:1511-2.
8. Scholte M, Van Dulmen SA, Neeleman-Van Der Steen CW, Van Der Wees PJ, Nijhuis-Van Der Sanden MW, Braspenning J. Data extraction from electronic health records (EHRs) for quality measurement of the physical therapy process: comparison between EHR data and survey data. BMC Med Inform Decis Mak. 2016;16:141.
9. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). 5th ed. Washington, DC: American Psychiatric Association; 2013.
10. Narayani N, Kumar P, Kumar D. Web Scraping and Automation Bot using Python. In: 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). IEEE; 2022. p. 1343-6.
11. Richardson L. Beautiful Soup Documentation [Internet]. 2007 Apr [cited 2026 May 26]. Available from: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
12. Broder MS, Gibbs SN, Yermilov I. An adaptation of the RAND/UCLA modified Delphi panel method in the time of COVID-19. J Healthc Leadersh. 2022;14:63-70.
13. Ford E, Sheppard J, Oliver S, Rooney P, Banerjee S, Cassell JA. Automated detection of patients with dementia whose symptoms have been identified in primary care but have no formal diagnosis: a retrospective case-control study using electronic primary care records. BMJ Open. 2021;11:e039248.
14. Ripley B, Venables B, Hornik K, Gebhardt A, Firth D. MASS: Support Functions and Datasets for Venables and Ripley's MASS. R Package Version 7.3. Berlin: Springer; 2019.
15. Kuhn M. Caret: Classification and Regression Training. R Package Version 6.0-86; 2020. Available from: https://www.rdocumentation.org/packages/caret/versions/6.0-86
16. LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. Electron J Stat. 2015;9(1):1583-607.
17. Clift AK, Coupland CA, Keogh RH, Hemingway H, Hippisley-Cox J. COVID-19 mortality risk in down syndrome: results from a cohort study of 8 million adults. Ann Intern Med. 2021;174:572-6.
18. Batty GD, Deary IJ, Luciano M, Altschul DM, Kivimäki M, Gale CR. Psychosocial factors and hospitalisations for COVID-19: prospective cohort study based on a community sample. Brain Behav Immun. 2020;89:569-78.
19. Batty GD, Deary IJ, Gale CR. Pre-Pandemic Cognitive Function and COVID-19 Mortality: Prospective Cohort Study. medRxiv [Preprint]; 2021.
20. Alnefeesi Y, Siegel A, Lui LM, Teopiz KM, Ho RC, Lee Y, et al. Impact of SARS-CoV-2 infection on cognitive function: a systematic review. Front Psychiatry. 2020;11:621773.
21. Daroische R, Hemminghyth MS, Eilertsen TH, Breitve MH, Chwiszczuk LJ. Cognitive impairment after COVID-19-a review on objective test data. Front Neurol. 2021;12:699582.
22. Xu E, Xie Y, Al-Aly Z. Long-term neurologic outcomes of COVID-19. Nat Med. 2022;28:2406-15.
23. Hancock LM, Hermann B, Schoonheim MM, Hetzel SJ, Brochet B, DeLuca J. Comparing diagnostic criteria for the diagnosis of neurocognitive disorders in multiple sclerosis. Mult Scler Relat Disord. 2022;58:103479.
24. Eramudugolla R, Mortby ME, Sachdev P, Meslin C, Kumar R, Anstey KJ. Evaluation of a research diagnostic algorithm for DSM-5 neurocognitive disorders in a population-based cohort of older adults. Alzheimers Res Ther. 2017;9:15.
25. Chen MH, Chiaravalloti ND, DeLuca J. Neurological update: cognitive rehabilitation in multiple sclerosis. J Neurol. 2021;268:4908-14.
26. Jammeh EA, Carroll CB, Pearson SW, Escudero J, Anastasiou A, Zhao P, et al. Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study. BJGP Open. 2018;2(2):bjgpopen18X101589. doi: 10.3399/bjgpopen18X101589.

27. Ford E, Rooney P, Oliver S, Hoile R, Hurley P, Banerjee S, et al. Identifying undetected dementia in UK primary care patients: a retrospective case-control study comparing machine-learning and standard epidemiological approaches. BMC Med Inform Decis Mak. 2019;19:248.
28. Ford E, Starlinger J, Rooney P, Oliver S, Banerjee S, Van Marwijk H, et al. Could dementia be detected from UK primary care patients' records by simple automated methods earlier than by the treating physician? A retrospective case-control study. Wellcome Open Res. 2020;5:120.
29. Despins LA. Automated detection of sepsis using electronic medical record data: a systematic review. J Healthc Qual. 2017;39:322-33.
30. Ramesh J, Keeran N, Sagahyroon A, Aloul F. Towards validating the effectiveness of obstructive sleep apnea classification from electronic health records using machine learning. Healthcare (Basel). 2021;9:1450.
31. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng. 2005;17:299-310.
32. Ling CX, Huang J, Zhang H. AUC: A Statistically Consistent and More Discriminating Measure than Accuracy. In: International Joint Conferences on Artificial Intelligence; 2003. p. 519-24.
33. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. 2022;12:5979.
34. Chung CW, Chou SC, Hsiao TH, Zhang GJ, Chung YF, Chen YM. Machine learning approaches to identify systemic lupus erythematosus in anti-nuclear antibody-positive patients using genomic data and electronic health records. BioData Min. 2024;17:1.
35. De Rooij M, Weeda W. Cross-validation: a method every psychologist should know. Adv Methods Pract Psychol Sci. 2020;3:248-63.
36. Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. Diabetes Care. 2013;36:914-21.
37. Mazzucato S, Bandini A, Micera S, Vergaro G, Dalmiani S, Emdin M, et al. Classification of patients with cardiac amyloidosis using machine learning models on Italian electronic clinical health records. Annu Int Conf IEEE Eng Med Biol Soc. 2023;2023:1-4.
38. Beaud V, Crottaz-Herbette S, Dunet V, Vaucher J, Bernard-Valnet R, Du Pasquier R, et al. Pattern of cognitive deficits in severe COVID-19. J Neurol Neurosurg Psychiatry. 2021;92:567-8.
39. Asadi-Pooya AA, Simani L. Central nervous system manifestations of COVID-19: a systematic review. J Neurol Sci. 2020;413:116832.