Performance of AI-powered chatbots in diagnosing acute pulmonary thromboembolism from given clinical vignettes

B Arslan, MN Sutasir & E Altinbilek

Abstract

Background: Chatbots hold great potential to serve as support tool in diagnosis and clinical decision process. In this study, we aimed to evaluate the accuracy of chatbots in diagnosing pulmonary embolism (PE). Furthermore, we assessed their performance in determining the PE severity.

Method: 65 case reports meeting our inclusion criteria were selected for this study. Two emergency medicine (EM) physicians crafted clinical vignettes and introduced them to the Bard, Bing, and ChatGPT-3.5 with asking the top 10 diagnoses. After obtaining all differential diagnoses lists, vignettes enriched with supplemental data redirected to the chatbots with asking the severity of PE.

Results: ChatGPT-3.5, Bing, and Bard listed PE within the top 10 diagnoses list with accuracy rates of 92.3%, 92.3%, and 87.6%, respectively. For the top 3 diagnoses, Bard achieved 75.4% accuracy, while ChatGPT and Bing both had 67.7%. As the top diagnosis, Bard, ChatGPT-3.5, and Bing were accurate in 56.9%, 47.7% and 30.8% cases, respectively. Significant differences between Bard and both Bing (p=0.000) and ChatGPT (p=0.007) were noted in this group. Massive PEs were correctly identified with over 85% success rate. Overclassification rates for Bard, ChatGPT-3.5 and Bing at 38.5%, 23.3% and 20%, respectively. Misclassification rates were highest in submassive group.

Conclusion: Although chatbots aren't intended for diagnosis, their high level of diagnostic accuracy and success rate in identifying massive PE underscore the promising potential of chatbots as clinical decision support tool. However, further research with larger patient datasets is required to validate and refine their performance in real-world clinical settings.

Keywords

Chatbot, ChatGPT, Bard, Bing, pulmonary embolism, artificial intelligence, emergency medicine

Banu Arslan
M.D., MSc

Mehmet Necmeddin Sutasir,
M.D.

Ertugrul Altinbilek
M.D. Assoc Prof

Department of Emergency Medicine, Republic of Turkey, Ministry of Health, Sisli Hamidiye Etfal Training and Research Hospital, Istanbul, Turkey

Correspondence
Banu Arslan
M.D., MSc
Department of Emergency Medicine, Republic of Turkey, Ministry of Health Sisli Hamidiye Etfal Training and Research Hospital, Istanbul, Turkey
Email: dr.banuarslan@gmail.com
Orcid: 0000-0003-0982-5351

Introduction

The incidence of pulmonary embolism (PE) has been on a steady rise in developed countries.[1-3] This trend may be attributed to the improvement in life expectancy, wider acceptance of validated diagnostic algorithms, increased awareness, and the routine utilization of computed tomographic pulmonary angiography (CTPA).[4,5] Fortunately, deaths from PE have been decreasing over the past years due to the greater proportion of low-risk patients being diagnosed and improvement in PE management.[6] Nevertheless, PE remains a leading cause of cardiovascular mortality after heart attack and stroke.[7]

Pulmonary embolism (PE) is a life-threatening medical emergency characterized by the sudden blockage of one or more arteries in the lungs. Patients can manifest with a range of symptoms, from asymptomatic incidental findings to circulatory collapse and even sudden cardiac arrest. The lack of specific symptoms makes PE challenging to diagnose.

The diagnostic approach depends on the hemodynamic status of patients. Hemodynamically stable patients with suspected pulmonary embolism should be assessed for the combination of validated clinical (pre-test) probability scales with D-dimer levels.[8] Patients with unstable hemodynamic parameters or suspected high-risk PE, bed-side echocardiography or CTPA should be performed whenever possible.[8] Following the diagnose, risk stratification guides the management of acute PE, as it tailors the treatment by classifying PE severity.[8]

Multiple risk stratification tools have been developed to assist health professionals in the management of PE patients. The most common schemes used to classify PE severity are proposed by the American Heart Association (AHA) and the European Society of Cardiology (ESC).[9,10] The AHA/ESC classification guidelines are quite similar and divide PE severity into three categories: low risk, submassive (AHA) or intermediate risk (ESC) and massive (AHA) or high-risk (ESC). In addition, there are some risk scores aim to identify low risk patients who are eligible for outpatient treatment such as

© 2024 Rila Publications Ltd.

Pulmonary Embolism Severity Index (PESI) and the simplified PESI (sPESI).[11,12]

Around 4% of acute PEs are classify as low risk, while 40% is categorized as massive PE.[13] Submassive PE exhibits high heterogeneity in terms of prognosis and the risk of complications. It is estimated that more than half of the acute PEs are classified as submassive PE with an overall 30-day mortality is between 5%-15%.[14,15] Additionally, majority of these normotensive patients can have asymptomatic right ventricle dysfunction (RVD) which may increase 30-day mortality between 3% to 10%.[16] Given these complexities, From this perspective, we believe that more advanced prediction models integrating clinical variables, biomarkers and imaging findings should be developed to identify this subgroup of patients that are at risk for early adverse events.

The recent surge in the field of artificial intelligence (AI), particularly in natural language processing (NLP), has accelerated the healthcare revolution. NLP focuses on the interaction between human language and computers. It Natural language processing (NLP) models enables computers to understand, interpret and generate human language, often providing meaningful response or actions based on the language input. One of the most prominent instances of such NLP models is the Generative Pretrained Transformer (GPT).[17] ChatGPT (OpenAI, San Francisco, California, United States), an AI-powered chatbot, is structured based on GPT and trained on a large data set of conversational text.[18] ChatGPT has been gaining tremendous attention since it was launched in November 2022. Following the success of ChatGPT, Microsoft launched its AI-powered search engine Bing (Microsoft Corporation, Redmond, Washington, United States) in February 2023, and Google AI released its artificial intelligent chatbot Bard (Google LLC, Mountain View, California, United States) in March 2023. Chatbots hold significant potential for improving healthcare practice, research, and education.[19] They can streamline the clinical workflow,[20] improve decision-making process,[21] conduct comprehensive literature reviews, and enhance scientific writing.[22] In medical education, they can support personalized learning,[23] create realistic clinical scenarios[24] and provide accurate and up to date medical information.[25] Recent studies demonstrated that ChatGPT exhibited high diagnostic accuracy by creating differential diagnosis lists based on the given clinical vignettes.[26,27] However, no study was conducted to ascertain the capability of these chatbots to diagnose a specific disease.

In this study, we aimed to assess the accuracy of chatbots in diagnosing a highly significant medical emergency, PE, based on given clinical vignettes. Furthermore, we evaluated the performance of these chatbots in determining the PE severity.

Materials and Methods

Literature Search and Case Selection

On October 2023, two authors systematically searched PubMed database for case reports on PE using the search terms "pulmonary embolism" AND "emergency". For the initial evaluation, 527 case reports involving adult patients and published in the English literature within the last 10 were selected.

First, both authors independently screened the titles and abstracts of the selected publications. Case reports meeting the inclusion criteria were then retrieved for in-depth full-text assessments. The inclusion criteria had contained admission to the emergency department, cases primarily focusing on pulmonary thromboembolism, the presence of sufficient details for crafting clinical vignettes, availability of the full text at no cost, confirmation of the PE through CTPA, and documentation of anatomic the location. Case reports involving postmortem evaluations or cases with a prolonged diagnosis period were excluded from the study. Furthermore, embolization of other materials into the pulmonary circulation such as air, fat, or tumor cells, were not considered in this study. After careful evaluation, 65 case reports were selected for the study. Case selection process is visually presented in Figure 1.

Ethical Considerations

Approval by the ethics committee were not obtained, since this study only used case vignettes from published case reports.

Definition of Outcomes

The primary outcome of this study was the incidence of acute PE the rank of PE in the top 10 differential diagnosis list. All PE patients in this study were diagnosed with CTPA, the gold standard test. The secondary outcome was the classification of PE such as massive, submassive and low risk. Risk stratification is one of the most significant steps in the management of acute PE, since it determines the risk for mortality and choice of treatment strategy. In this study, we used the American Heart Association (AHA) 2011 risk stratification algorithm.[28] The risk stratification in this guideline relies on the three important criteria: systolic blood pressure (SBP), right ventricle functions and biomarker levels including cardiac troponin and brain natriuretic peptides (BNP). Hypotension is defined by SBP inferior to 90 mmHg for at least 15 minutes or requiring vasopressors in the absence of hypovolemia, sepsis, and left ventricle dysfunction. Bradycardia means heart rate is slower than 40 beats in a minute. Risk stratification method that was used in this study is visually presented in Table 1.

© 2024 Rila Publications Ltd.

Performance of AI-powered chatbots in diagnosing acute pulmonary thromboembolism from given clinical vignettes

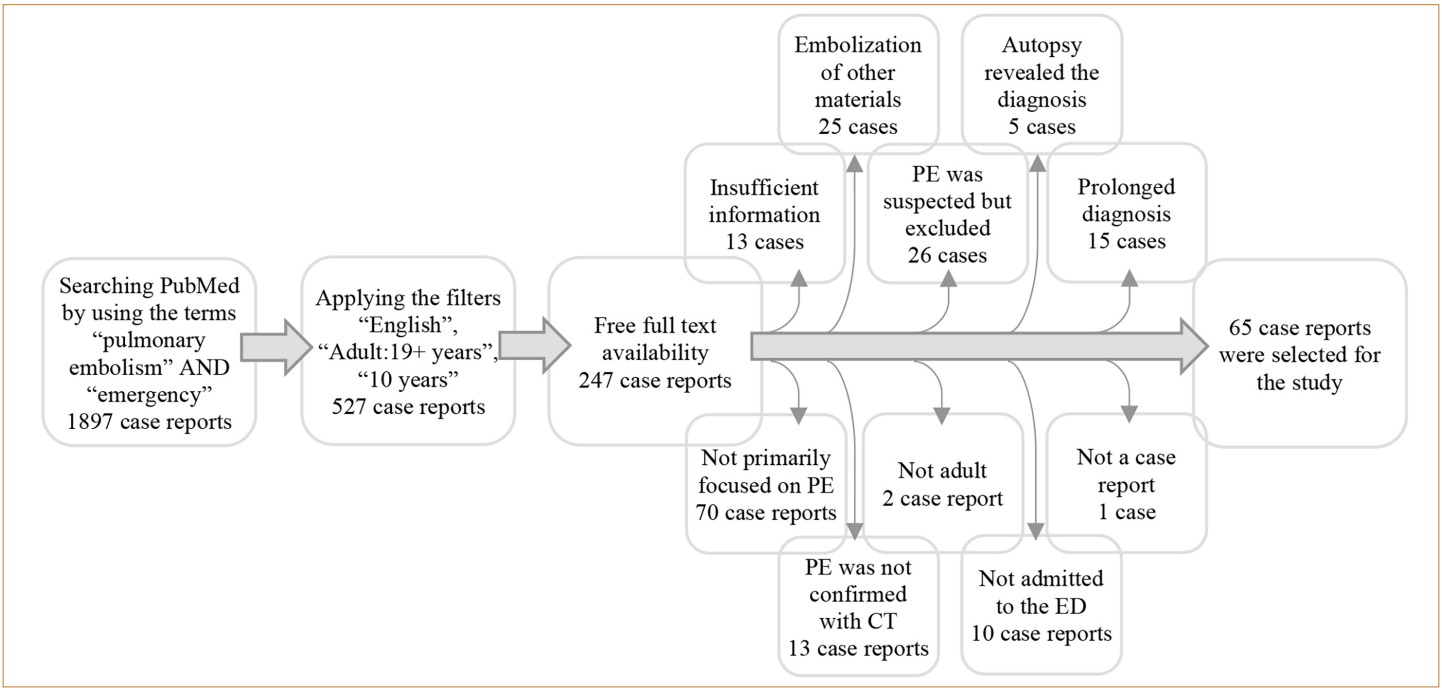

Figure 1. Case report selection flowchart detailing the inclusion and exclusion criteria.

Table 1. Classification of pulmonary embolism.

	Massive	Submassive	Low risk
Presentation	Sustained hypotension (SBP <90 mmHg for at least 15 minutes) Pulselessness Persistent bradycardia (Heart rate <40 bpm with signs or symptoms of shock).	Normotensive (SBP ≥90 mm Hg) Right ventricle dysfunction Myocardial necrosis	Normotensive No right ventricle dysfunction Normal biomarkers

Table 2. Definition of RVD and myocardial necrosis.

Right Ventricle Dysfunction is defined as the presence of at least 1 of the following:
ECG New complete or incomplete right bundle-branch block Anteroseptal ST elevation or depression Anteroseptal T-wave inversion
Biomarkers BNP (>90 pg/mL) N-terminal pro-BNP (>500 pg/mL)
Echocardiography RV dilation (apical 4-chamber RV diameter divided by LV diameter >0.9) RV systolic dysfunction
CT RV dilation (4-chamber RV diameter divided by LV diameter >0.9)
Myocardial dysfunction is defined as either of the following:
Troponin I (>0.4 ng/mL) Troponin T (>0.1 ng/mL)

RVD can be determined based on echocardiography, CT, and ECG findings that was presented in the Table 2. We further defined myocardial injury and right heart failure based on increased serum biomarker levels of cardiac troponins and natriuretic peptides, respectively.

Clinical Vignettes

Two authors extracted the data from the selected case reports and displayed them as detailed clinical vignettes. Extracted data included demographic characteristics, complaints, social habits, medical history, medications, vital parameters, physical examination findings, and ECG interpretations. Information about routine laboratory tests, imaging studies, differential diagnosis, final diagnosis, titles, figures, legends, and tables were removed from the vignettes. Information on the anatomic location of the PE, biomarker levels (BNP, pro-BNP, and troponin I/T), right ventricle functions, treatment of choice, and the outcome extracted from each case and added to the study data separately for further evaluation.

Two independent EM physicians evaluated the 65 case reports separately and determine the severity of the PE according to the risk stratification guideline proposed by

© 2024 Rila Publications Ltd.

Table 3. Example of the differential diagnosis list for a 64-year-old male presented to the ED with a shortness of breath. Patient had a history of HT, PE, and thoracic aortic aneurysm repair with bioprosthetic aortic valve replacement and coronary artery bypass graft. On presentation, the patient was hemodynamically unstable and reported chest and upper back pain.

ChatGPT-3.5	Bard	Bing
Aortic Dissection	**Pulmonary embolism**	Acute Heart Failure
Pulmonary Embolism	Pneumonia	Pneumonia
Cardiac Tamponade	Heart failure	**Pulmonary Embolism**
ACS	Aortic dissection	Aortic Dissection
Heart Failure Exacerbation	Bronchospasm	Myocardial Infarction
Infectious Pneumonia or Sepsis	Pneumothorax	Pericarditis
Hemothorax	ARDS	Pleural Effusion
Pneumothorax	Myocardial infarction	COPD Exacerbation
Arrhythmia or Heart Block	Anaphylaxis	Costochondritis
Anaphylaxis or Allergic Reaction	Anxiety attack	GERD

ACS: Acute Coronary Syndrome. ARDS: Acute respiratory distress syndrome. COPD: chronic obstructive pulmonary disease. GERD: Gastroesophageal reflux disease. HT: Hypertension. PE: Pulmonary embolism

the AHA.[28] Discrepancies between the two physicians were resolved through discussions.

The study employed three distinct chatbots: ChatGPT-3.5 (OpenAI, San Francisco, California, United States), Bard (Google LLC, Mountain View, California, United States), and Bing (Microsoft Corp., Redmond, WA, USA). In November 2023, same authors (A.B. and S.M.N) directed clinical vignettes to these chatbots with adding the following text in the prompt: "Can you list ten possible diagnoses, ordered from most likely to less likely based on the presented information above". Then, the list was checked for the presence of PE and its rank. To prevent any potential interference from prior responses, physicians cleared the preceding conversation every time before introducing new clinical vignette. Table 3 exemplifies a sample differential diagnosis list generated by these three models.

After acquiring all differential diagnosis lists, we enriched clinical vignettes with supplementary details. These details included the anatomic location of the PE, biomarker levels (BNP, pro-BNP, and troponin I/T), and right ventricle functions. The enriched vignettes were then reintroduced to these models. This time the inquiry posed was, "Can you stratify the risk of the PE here?" The level of severity determined by the models were recorded in the study data as massive, submassive and low risk. Cases with differential diagnosis lists did not include PE were not subjected to this further evaluation.

Statistical Analysis

All statistical analysis were performed by using the SPSS 28.0 for Windows (IBM, USA). Descriptive statistics were used as mean ± standard deviation (SD) for continuous variables, and frequency (percentages) for categorical variables. The distribution of variables was measured using the Kolmogorov-Smirnov test. The Wilcoxon test was employed for the analysis of dependent quantitative data. The McNemar test was used for the analysis of dependent qualitative data. $P < .05$ was considered statistically significant.

Results

A total of 65 clinical vignettes were analyzed in this study. These vignettes involved 29 female (44.6%) and 36 male (55.4%) patients. Ages of the patients ranged from 21 years to 86 years with the mean of 53.38 ±17.33. Of these 65 patients, 45 (69%) reported having at least one comorbidity. The most prevalent comorbidity was hypertension (HT) (51%), followed by diabetes mellitus (DM) (15.5%), history of PE (11%), asthma (11%) and history of deep vein thrombosis (DVT) (8%). The most common complaints were dyspnea (69%,) chest pain (49%) and syncope/loss of consciousness (20%). Study population is described in Table 4.

ChatGPT-3.5, Bard and Bing listed PE within the top 10 diagnoses list with an accuracy rate of 92.3% (60/65), 87.6% (57/65) and 92.3% (60/65), respectively. The performance of the chatbots was also evaluated based on the listing PE within the top three diagnoses lists, as well as the first diagnosis. For the top 3 diagnoses lists, Bard achieved 75.4% (49/65) accuracy, while ChatGPT and Bing both listed 67.7% (44/65) of PE cases correctly. In this section, there was no statistically significant differences between Bard and ChatGPT, Bard and Bing or ChatGPT and Bing. Examining the first diagnosis, the rate of PE being ranked first were as follows: Bard 56.9% (37/65), ChatGPT 47.7% (31/65), and Bing 30.8% (20/65). The statistical analysis indicated no significant difference between ChatGPT and Bard (p=0.238). However, a significant difference between Bing and both Bard (p=0.000) and ChatGPT (p=0.007) was noted (Table 5).

For further evaluation, we excluded clinical vignettes without PE in their top 10 diagnoses lists.

© 2024 Rila Publications Ltd.

Table 4. Demographic characteristics.

		Min-Max			Mean±sd/n-%		
Age		21	-	86	53.38	±	17.33
Sex	**Female**				29		44.6%
	Male				36		55.4%
Hypotension or cardiac arrest	**Absent**				44		67.7%
	Present				21		32.3%
RV Dysfunction	**Absent**				18		27.7%
	Present				44		67.7%
	Unknown				3		4.6%
Biomarker elevation	**Absent**				4		6.2%
	Present				43		66.2%
	Unknown				18		27.7%
Anatomic location	**Subsegmental**				2		3.1%
	Segmental				8		12.3%
	Lobar				43		66.2%
	Saddle				12		18.5%
Severity	**Low risk**				3		4.6%
	Submassive				39		60.0%
	Massive				23		35.4%
Treatment	**Anticoagulation**				31		47.7%
	Thrombolytic				16		24.6%
	CDT				10		15.4%
	Surgical embolectomy				8		12.3%
Outcome	**Discharged**				52		80.0%
	ICU				10		15.4%
	Deceased				3		4.6%

Table 5. Relationship between models based on diagnosis list order.

		Mean±sd/n-%			p*		p‡	
1st Diagnosis								
Bard	(-)	28		43.1%				
	(+)	37		56.9%				
ChatGPT	(-)	34		52.3%	0.238	N		
	(+)	31		47.7%				
Bing	(-)	45		69.2%	0.000	N	0.007	N
	(+)	20		30.8%				
Top 3 diagnosis								
Bard	(-)	16		24.6%				
	(+)	49		75.4%				
ChatGPT	(-)	21		32.3%	0.302	N		
	(+)	44		67.7%				
Bing	(-)	21		32.3%	0.359	N	1.000	N
	(+)	44		67.7%				

N: MC Nemar test. p* Difference with Bard / p‡ Difference with ChatGPT

© 2024 Rila Publications Ltd.

Table 6. Success in identifying the severity of PE with underestimation and overestimation rates.

	PE severity based on AHA risk stratification guideline			
	Low risk n (%)	Submassive n (%)	Massive n (%)	All patients n (%)
Bard (n=57)				
- identified	0/3 (0)	7/32 (21.8)	19/22 (86.3)	26/57 (45.6)
- unidentified	0/3 (0)	6/32 (18.7)	3/22 (13.6)	9/57 (15.7)
- overclassified	3/3 (100)	19/32 (59.3)	0/22 (0)	22/57 (38.5)
- underclassified	0/3 (0)	0/32 (0)	0/22 (0)	0/57 (0)
ChatGPT (n=60)				
- identified	3/3 (100)	16/36 (44.4)	19/21 (90.4)	38/60 (63.3)
- unidentified	0 (0)	2/36 (5.5)	0 (0)	2/60 (3.3)
- overclassified	0 (0)	14/36 (38.8)	0 (0)	14/60 (23.3)
- underclassified	0 (0)	4/36 (11.1)	2/21 (9.5)	6/60 (10)
Bing (n=60)				
- identified	3/3 (100)	18/35 (51.4)	19/22 (86.3)	40/60 (66.6)
- unidentified	0 (0)	1/35 (2.8)	1/22 (4.5)	2/60 (3.3)
- overclassified	0 (0)	12/35 (34.2)	0/22(0)	12/60 (20)
- underclassified	0 (0)	4/35 (11.4)	2/22 (9.1)	6/60 (10)

Then, we enriched the rest of the vignettes with supplementary data and reintroduced them to the models with asking the severity of PEs. Bard refused to answer to 9 clinical vignettes with stating this *"I'm unable to help, as I am only a language model and don't have the ability to process and understand that"*. ChatGPT and Bing were also unable to provide a single classification for two cases each. ChatGPT and Bing correctly identified all 3 low-risk PEs, whereas Bard identified none of them. In submassive group, the success rate of Bard, ChatGPT and Bing was 21.8%, 44.4% and 51.4%, respectively. In massive group, the success rate of Bard, ChatGPT and Bing was 86.3%, 90.4% and 86.3%, respectively. Data on overestimation and underestimation rates for each model represented in detail at Table 6.

Discussion

Principal Results

This study included 65 clinical vignettes derived from case reports published in the last 10 years. Case reports in medical literature tend to be highly complex, often involving rare presentations of a disease or unique clinical settings. Even in this challenging environment, all three models in this study generated top 10 differential diagnoses lists with an accuracy rate of over 85%.

In recent years, healthcare has seen significant advancements in diagnosing PE through advanced technologies. One promising method is using ML and AI technics to analyze clinical data. It is shown that AI models can detect patterns that suggest the presence of PE and even assess the risk of complications or mortality.[29-32] For example, studies by Rucco et al. and Ryan et al. have shown impressive accuracy in identifying PE cases and predicting risk using AI algorithms.[33,34] While ML tools specifically designed for PE diagnosis may exist, NLP models offer a unique advantage by providing a comprehensive understanding of patients' conditions. This includes consideration of a broad range of potential differentials, thus more closely mirroring the complexity of clinical practice.

Previous studies have reported that ChatGPT-3 generated the top 10 differentail differential diagnoses lists with 93.3% accuracy, while ChatGPT-3.5 achieved an accuracy of 73% [29,30]. In our study, the rate of PE diagnosis noted as 92.3% for the top 10 diagnoses lists generated by ChatGPT-3.5. It is important to note that those studies relied on a smaller sample size from a single department, encompassing a wide variety of diseases and diagnoses. This may indicate that our study, with its narrower focus and more cases could offer a more precise depiction of the diagnostic capabilities of AI-assisted chatbots. When listing PE 1st, our models exhibited comparatively lower diagnostic accuracy. Specifically, Bard achieved an accuracy rate of 56.9%, ChatGPT achieved 47.7%, and Bing reached 30.8%. Bard and ChatGPT can be considered were better than Bing (p=0.000 vs. p=0.007, respectively) in this group.

Hirosawa et al. has indicated have reported similar results with our study regarding the that ChatGPT's accuracy in providing correct diagnosis as the 1st choice ranged from 42% to 60%.[29,30] It is noteworthy that when they compared the diagnostic capabilities of ChatGPT and physicians, there was no statistically significant difference was

© 2024 Rila Publications Ltd.

found between ChatGPT and physicians in this regard.[29] From this perspective, we believe that AI-powered chatbots have great potential as a clinical decision support tool, providing valuable assistance to healthcare professionals, not only for junior doctors and medical students but also for senior physicians. For example, AI-powered chatbots represent a potent educational resource for junior doctors and medical students. They can be utilized to validate diagnoses and expand their repertoire of potential diagnoses. Additionally, these chatbots serve as effective training aids, enabling learners to refine their diagnostic abilities within a controlled environment. While senior physicians may possess extensive clinical experience, AI-powered chatbots still offer considerable benefits. They can serve as a supplementary resource, providing second opinions or reminders regarding uncommon conditions that may not frequently present in their practice. Moreover, in hectic clinical settings, AI-powered chatbots facilitate the diagnostic process by swiftly generating a list of potential diagnoses based on presented symptoms.

Risk stratification in PE is crucial for several reasons such as choosing the appropriate treatment, assessing the risk for mortality, and providing prognostic information. In this study, all models demonstrated a success rate of over 85% in classifying massive PE correctly. In low-risk PEs, the success rate was 100% for ChatGPT and Bing. However, Bard has showed no success in identifying low risk PEs. Classifying patients with submassive PE and predicting the ones at high risk for developing hemodynamic collapse has always been challenging. Risk scores are increasingly recommended by guidelines for the initial assessment of patients with acute PE.[31,32] The ESC 2019 risk stratification guideline have been suggested for providing more precise classification.[33] This guideline basically relies on hemodynamic status, clinical scores (PESI, sPESI), troponin levels and the presence of RVD. AHA risk stratification guideline was employed in the present study since troponin levels and PESI scores were not available for most of the cases. Following AHA risk stratification guideline, we observed that Bing accurately classified 51.4% of the submassive PEs, followed by ChatGPT-3.5 with 44.4%, and Bard with 21.8%.

Our findings underscored the notable rates of misclassification. Misclassification of PE into the high-risk category was previously reported in the literature.[34] According to the study, using the PESI as the comparative standard, the sPESI misclassified 543 patients (18.1%) in total with overclassifying 443 patients (14.7%). In our research, we observed overclassification rates for Bing, ChatGPT-3.5 and Bard at 20%, 23.3% and 38.5%, respectively. Notably, misclassification rates were highest in our submassive PEs. Bard misclassified 59.3% of submassive PEs, categorizing all as massive PE. ChatGPT-3.5 overclassified 38.8% of submassive PEs, while Bing overestimated 34.2%. The misclassifying PE can result in severe complications. Particularly overestimating the risk may lead to unnecessary thrombolytic treatment, prolonged hospital stays, increased likelihood of complications, and higher financial costs. Tailored trainings are required for improvement in this area.

Strengths and Limitations.

To the best of our knowledge, this is the first study evaluating the performance of AI-powered chatbots in classifying the severity of a medical emergency. We believe that we must study every single clinical tool that is available to identify PE and classify its severity with the ultimate goal of providing the best possible emergency care. In this study, we used three different publicly available chatbots. Even though ChatGPT is considered as the pioneer in the market, there are several chatbots providing promising results. Besides ChatGPT3.5, we included Bard and Bing to assess the performance of chatbots in predicting PE to gain better insight about this novel technology. Our findings highlighted the strengths and weaknesses of these models in classifying PE severity. Another strength of this study was that the case reports we used to create clinical vignettes were from various countries and involved patients from diverse backgrounds, ethnicities, and demographics. We believe that collecting case reports from various countries provides a global perspective on PE patients.

The primary limitation of this study is its reliance on clinical vignettes, even they were derived from real cases. It is essential to note that case reports with positive outcomes are more likely to be published, introducing a potential source of publication bias. However, our study included 3 low risk, 39 submassive and 23 massive PEs, representing 4.6%, 60%, and 35.4%, respectively. These findings indicates that our study population covers a spectrum of acute PEs, with a significant focus on submassive cases, and a distribution of low risk and massive cases as well.

Secondly, we acknowledge that comparing the lists of potential diagnoses generated by the chatbots with those formulated by a group of physicians would have provided deeper insights into the diagnostic process. However, all the clinical vignettes in our study were ultimately diagnosed with PE. This raised concerns that physicians, while creating differential diagnosis lists, might tend to recognize PE as the prevalent diagnosis over time, potentially biasing their decisions. Additionally, Hirosawa T. et al. compared the rates of correct diagnoses in the differential

© 2024 Rila Publications Ltd.

diagnosis lists generated by ChatGPT-3.5 with those by physicians and found no statistically significant difference.[29] However, it is important to note that the diagnostic accuracy of chatbots and physicians can vary significantly depending on the complexity and rarity of the cases presented. More research is needed to evaluate the performance of chatbots across a wider range of medical conditions.

Conclusion

Our study evaluated the performance of AI-powered chatbots in identifying PE within the differential diagnosis lists that was generated for 65 clinical vignettes derived from highly complex case reports. While providing valuable insights into the prevalence of RVD and biomarker elevation in the patient population studied, this study shows that all three models, Bard, ChatGPT-3.5 and Bing, generated top 10 differential diagnosis lists with an accuracy rate of 92.3%, 87.6%, and 92.3%, respectively. The current study also explored the models' success in risk stratification, a critical aspect in managing PE patients. While all models displayed success rates exceeding 85% in correctly classifying massive PE, challenges arose in submassive PE cases, leading to notable misclassification rates. Despite these challenges, our findings highlight the promising potential of AI-powered chatbots as valuable clinical decision support tools. Further research with larger patients is essential to validate and refine the performance of these models in real-world clinical settings.

Acknowledgements

During the preparation of this work the author used ChatGPT-3.5 in order to improve language and readability. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflict of Interest

The authors declare that there is no conflict of interest.

Funding

None Declared

Author contributions

Arslan B. is the research principal. She also wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Acknowledgments

Microsoft Bing, Google AI and Open AI was not involved in the design, implementation, data analysis, or manuscript preparation of the study.

Availability of Data and Materials

The datasets generated during the current study are available from corresponding author.

Animal and Human Rights Statement

No human and animal studies were carried out by the authors for this article.

Abbreviations

AHA: American Heart Association
AI: Artificial intelligence
BNP: Brain natriuretic peptides
CTPA: Computed tomographic pulmonary angiography.
DM: Diabetes mellitus
DVT: Deep vein thrombosis
ESC: European Society of Cardiology
EM: Emergency medicine
GPT: Generative Pretrained Transformer
HT: Hypertension
NLP: Natural language processing
PE: Pulmonary embolism
RVD: Right ventricle dysfunction
SBP: Systolic blood pressure

References

1. Kempny A, McCabe C, Dimopoulos K *et al.* Incidence, mortality and bleeding rates associated with pulmonary embolism in England between 1997 and 2015. *Int J Cardiol.* 1997;**2018**: 10.1016/j.ijcard.2018.10.001
2. Ghanima W, Brodin E, Schultze A *et al.* Incidence and prevalence of venous thromboembolism in Norway 2010–2017. *Thromb Res.* 2020;**195**:165-168. 10.1016/j.thromres.2020.07.011
3. Lee LH, Gallus A, Jindal R, Wang C, Wu CC. Incidence of venous thromboembolism in Asian populations: a systematic review. *Thromb Haemost.* 2017;**117**:2243-2260. 10.1160/TH17-02-0134
4. Schissler AJ, Rozenshtein A, Kulon ME, Pearson GD, Green RA, Stetson PD, Brenner DJ, D'Souza B, Tsai WY, Schluger NW, Einstein AJ. CT pulmonary angiography: increasingly diagnosing less severe pulmonary emboli. *PLoS One.* 2013 Jun 12;**8(6)**:e65669. doi: 10.1371/journal.pone.0065669.
5. Barco S, Valerio L, Gallo A, Turatti G, Mahmoudpour SH, Ageno W, Castellucci LA, Cesarman-Maus G, Ddungu H, De Paula EV, Dumantepe M, Goldhaber SZ, Guillermo Esposito MC, Klok FA, Kucher N, McLintock C, Ní Áinle F, Simioni P, Spirk D, Spyropoulos AC, Urano T, Zhai ZG, Hunt BJ, Konstantinides SV. Global reporting of pulmonary embolism-related deaths in the World Health Organization mortality database: Vital registration data from 123 countries. *Res Pract Thromb Haemost.* 2021 Jun 15;**5(5)**:e12520. doi: 10.1002/rth2.12520.
6. Barco S, Mahmoudpour SH, Valerio L, Klok FA, Münzel T, Middeldorp S, Ageno W, Cohen AT, Hunt BJ, Konstantinides SV. Trends in mortality related to pulmonary embolism in the European Region, 2000-15: analysis of vital registration data from the WHO Mortality Database. *Lancet Respir Med.* 2020 Mar;**8(3)**:277-287. doi: 10.1016/S2213-2600(19)30354-6.

© 2024 Rila Publications Ltd.

7. Goldhaber SZ, Bounameaux H. Pulmonary embolism and deep vein thrombosis. *Lancet.* 2012 May 12;**379**(**9828**):1835-46. doi: 10.1016/S0140-6736(11)61904-1.
8. Konstantinides SV, Meyer G, Becattini C, Bueno H, Geersing GJ, Harjola VP, Huisman MV, Humbert M, Jennings CS, Jiménez D, Kucher N, Lang IM, Lankeit M, Lorusso R, Mazzolai L, Meneveau N, Ní Áinle F, Prandoni P, Pruszczyk P, Righini M, Torbicki A, Van Belle E, Zamorano JL; ESC Scientific Document Group. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS). *Eur Heart J.* 2020 Jan 21;**41**(**4**):543-603. doi: 10.1093/eurheartj/ehz405.
9. 9. Giri J, Sista AK, Weinberg I, Kearon C, Kumbhani DJ, Desai ND, Piazza G, Gladwin MT, Chatterjee S, Kobayashi T, Kabrhel C, Barnes GD. Interventional Therapies for Acute Pulmonary Embolism: Current Status and Principles for the Development of Novel Evidence: A Scientific Statement From the American Heart Association. *Circulation*. 2019 Nov 12;**140**(**20**):e774-e801. doi:10.1161/CIR.0000000000000707.
10. 10. ESC Scientific Document Group . Konstantinides S V, Meyer G, Becattini C. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS) *Eur Heart J.* 2020;**41**(**04**):543–603.
11. 11. Aujesky D, Obrosky DS, Stone RA, *et al.*. Derivation and validation of a prognostic model for pulmonary embolism. *Am J Respir Crit Care Med.* 2005;**172**(**8**):1041-1046. doi: 10.1164/rccm.200506-862OC
12. 12. Bova C, Sanchez O, Prandoni P, *et al.*. Identification of intermediate-risk patients with acute symptomatic pulmonary embolism. *Eur Respir J.* 2014;**44**(**3**):694-703. doi: 10.1183/09031936.00006114
13. Elias A., Mallett S., Daoud-Elias M., Poggi J.-N., Clarke M. Prognostic models in acute pulmonary embolism: A systematic review and meta-analysis. *BMJ Open.* 2016;**6**:e010324. doi: 10.1136/bmjopen-2015-010324.
14. Aujesky D., Obrosky D.S., Stone R.A., Auble T.E., Perrier A., Cornuz J., Roy P.-M., Fine M.J. Derivation and Validation of a Prognostic Model for Pulmonary Embolism. *Am. J. Respir. Crit. Care Med.* 2005;**172**:1041–1046. doi: 10.1164/rccm.200506-862OC.
15. Jiménez D., Aujesky D., Moores L., Gómez V., Lobo J.L., Uresandi F., Otero R., Monreal M., Muriel A., Yusen R.D. Simplification of the Pulmonary Embolism Severity Index for Prognostication in Patients With Acute Symptomatic Pulmonary Embolism. *Arch. Intern. Med.* 2010;**170**:1383–1389.
16. Abrahams-van Doorn P J, Hartmann I J. Cardiothoracic CT: one-stop-shop procedure? Impact on the management of acute pulmonary embolism. *Insights Imaging.* 2011;**2**(**06**):705–715.
17. Zhou B., Yang G., Shi Z., Ma S. Natural language processing for smart healthcare. arXiv. 2021 doi: 10.1109/RBME.2022.3210270.2110.15803
18. Gordijn B, Have HT. ChatGPT: Evolution or revolution?*Med Health Care Philos.* 2023;**26**:1–2.epub ahead of print. doi:10.1007/s11019-023-10136-0.
19. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* (Basel). 2023 Mar 19;**11**(**6**):887. doi: 10.3390/healthcare11060887.
20. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, Landman A, Dreyer K, Succi MD. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *J Med Internet Res.* 2023 Aug 22;**25**:e48659. doi: 10.2196/48659
21. Rao A., Kim J., Kamineni M., Pang M., Lie W., Succi M.D. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. medRxiv. 2023 doi: 10.1101/2023.02.02.23285399
22. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. *Am J Cancer Res.* 2023 Apr 15;**13**(4):1148-1154
23. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. *Pak J Med Sci.* 2023 Mar-Apr;**39**(**2**):605-607. doi: 10.12669/pjms.39.2.7653
24. Benoit JR. ChatGPT for clinical vignette generation, revision, and evaluation. medRxiv 2023.02.04.23285478. [Last accessed on 2023 June 5]. Available from: https://doi.org/10.1101/2023.02.04.23285478
25. Johnson S, King A, Warner E, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectr.* 2023 Mar 01;7(2):1–9. doi: 10.1093/jncics/pkad015
26. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. *Int J Environ Res Public Health.* 2023 Feb 15;**20**(**4**):3378. doi: 10.3390/ijerph20043378.
27. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, Suzuki T, Shimizu T. ChatGPT-Generated Differential Diagnosis Lists for Complex Case-Derived Clinical Vignettes: Diagnostic Accuracy Evaluation. *JMIR Med Inform.* 2023 Oct 9;**11**:e48808. doi: 10.2196/48808.
28. Jaff MR, McMurtry MS, Archer SL, Cushman M, Goldenberg N, Goldhaber SZ, Jenkins JS, Kline JA, Michaels AD, Thistlethwaite P, Vedantham S, White RJ, Zierler BK; American Heart Association Council on Cardiopulmonary, Critical Care, Perioperative and Resuscitation; American Heart Association Council on Peripheral Vascular Disease; American Heart Association Council on Arteriosclerosis, Thrombosis and Vascular Biology. Management of massive and submassive pulmonary embolism, iliofemoral deep vein thrombosis, and chronic thromboembolic pulmonary hypertension: a scientific statement from the American Heart Association. *Circulation.* 2011 Apr 26;**123**(**16**):1788-830. doi: 10.1161/CIR.0b013e318214914f.
29. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, Suzuki T, Shimizu T. ChatGPT-Generated Differential Diagnosis Lists for Complex Case-Derived Clinical Vignettes: Diagnostic Accuracy Evaluation. *JMIR Med Inform.* 2023 Oct 9;**11**:e48808. doi: 10.2196/48808.
30. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. *Int J Environ Res Public Health.* 2023 Feb 15;**20**(**4**):3378. doi: 10.3390/ijerph20043378.
31. Hassine M, Kallala MY, Mahjoub M, Boussaada M, Bouchahda N, Gamra H. Embolie pulmonaire: indice de sévérité de l'embolie pulmonaire (ISEP) score et facteurs prédictifs de mortalité [Pulmonary embolism: the Pulmonary Embolism Severity Index (PESI) score and mortality predictors]. *Pan Afr Med J.* 2023 May 19;**45**:48. French. doi: 10.11604/pamj.2023.45.48.39031.
32. Konstantinides SV, Torbicki A, Agnelli G, *et al.*. 2014 ESC guidelines on the diagnosis and management of acute pulmonary embolism. *Eur Heart J.* 2014;35(43):3033-3069. doi: 10.1093/eurheartj/ehu283
33. Leidi A, Bex S, Righini M, Berner A, Grosgurin O, Marti C. Risk Stratification in Patients with Acute Pulmonary Embolism: Current Evidence and Perspectives. *J Clin Med.* 2022 Apr 30;**11**(**9**):2533. doi: 10.3390/jcm11092533
34. Vinson DR, Ballard DW, Mark DG, Huang J, Reed ME, Rauchwerger AS, Wang DH, Lin JS, Kene MV, Pleshakov TS, Sax DK, Sax JM, McLachlan DI, Yamin CK, Swap CJ, Iskin HR, Vemula R, Fleming BS, Elms AR, Aujesky D; MAPLE investigators of the KP CREST Network. Risk stratifying emergency department patients with acute pulmonary embolism: Does the simplified Pulmonary Embolism Severity Index perform as well as the original? *Thromb Res.* 2016 Dec;**148**:1-8. doi: 10.1016/j.thromres.2016.09.023.

© 2024 Rila Publications Ltd.