Heliyon 10 (2024) e31750

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

ChatGPT achieves comparable accuracy to specialist physicians in predicting the efficacy of high-flow oxygen therapy

Taotao Liu[a,*,1], Yaocong Duan[b,1], Yanchun Li[c], Yingying Hu[c], Lingling Su[d], Aiping Zhang[d]

[a] *Department of Surgical Intensive Care Unit, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, 100730, China*
[b] *School of Psychology and Neuroscience, University of Glasgow, Glasgow, G12 8QQ, UK*
[c] *The First Affiliated Hospital, and College of Clinical Medicine of Henan University of Science and Technology, Luoyang, 471000, China*
[d] *Department of Respiratory and Critical Care Medicine, Jiangyan Hospital Affiliated to Nanjing University of Chinese Medicine, Taizhou, 225500, China*

ARTICLE INFO

Keywords:
High-flow nasal cannula oxygen
ROX index
ChatGPT
Artificial intelligence

ABSTRACT

Background: The failure of high-flow nasal cannula (HFNC) oxygen therapy can necessitate endotracheal intubation in patients, making timely prediction of the intubation risk following HFNC therapy crucial for reducing mortality due to delays in intubation.

Objectives: To investigate the accuracy of ChatGPT in predicting the endotracheal intubation risk within 48 h following HFNC therapy and compare it with the predictive accuracy of specialist and non-specialist physicians.

Methods: We conducted a prospective multicenter cohort study based on the data of 71 adult patients who received HFNC therapy. For each patient, their baseline data and physiological parameters after 6-h HFNC therapy were recorded to create a 6-alternative-forced-choice questionnaire that asked participants to predict the 48-h endotracheal intubation risk using scale options ranging from 1 to 6, with higher scores indicating a greater risk. GPT-3.5, GPT-4.0, respiratory and critical care specialist physicians and non-specialist physicians completed the same questionnaires ($N = 71$) respectively. We then determined the optimal diagnostic cutoff point, using the Youden index, for each predictor and 6-h ROX index, and compared their predictive performance using receiver operating characteristic (ROC) analysis.

Results: The optimal diagnostic cutoff points were determined to be ≥ 4 for both GPT-4.0 and specialist physicians. GPT-4.0 demonstrated a precision of 76.1 %, with a specificity of 78.6 % (95%CI = 52.4–92.4 %) and sensitivity of 75.4 % (95%CI = 62.9–84.8 %). In comparison, the precision of specialist physicians was 80.3 %, with a specificity of 71.4 % (95%CI = 45.4–88.3 %) and sensitivity of 82.5 % (95%CI = 70.6–90.2 %). For GPT-3.5 and non-specialist physicians, the optimal diagnostic cutoff points were ≥ 5, with precisions of 73.2 % and 64.8 %, respectively. The area under the curve (AUC) in ROC analysis for GPT-4.0 was 0.821 (95%CI = 0.698–0.943), which was the highest among the predictors and significantly higher than that of non-specialist physicians [0.662 (95%CI = 0.518–0.805), $P = 0.011$].

[*] Corresponding author.
E-mail address: taotao20022000@163.com (T. Liu).

[1] T.L. and Y.D. contributed equally as co-first authors.

https://doi.org/10.1016/j.heliyon.2024.e31750
Received 20 February 2024; Received in revised form 17 May 2024; Accepted 21 May 2024
Available online 22 May 2024

Conclusion: GPT-4.0 achieves an accuracy level comparable to specialist physicians in predicting the 48-h endotracheal intubation risk following HFNC therapy, based on patient baseline data and physiological parameters after 6-h HFNC therapy.

1. Introduction

High-flow nasal cannula (HFNC) oxygen therapy, which delivers a heated, humidified, and high-flow air-oxygen mixture to patients, has become increasingly popular in clinical settings for treating hypoxemia due to its ease of use and effectiveness [1,2]. Recent studies have further explored the indications of HFNC therapy [3–5]. Despite its benefits, HFNC therapy sometimes fails, necessitating subsequent endotracheal intubation to reduce mortality risks [6]. Therefore, accurately and timely predicting the need for intubation following HFNC therapy is crucial. While methods like ROX index—a ratio of SpO_2/FiO_2 to respiratory rate—can be used to predict the efficacy of HFNC therapy [7], they offer only moderate predictive accuracy and lack standardized diagnostic cutoff points [8,9].

Artificial intelligence (AI) has shown potential in supporting clinical decision-making. However, the complexity of AI algorithms and their steep learning curve pose significant barriers to physicians without programming experience, limiting their use in AI-assisted clinical decision-making. Recent advancements in large language model (LLM) tools, such as ChatGPT, present an opportunity for physicians to interact with AI through natural language, thereby bypassing the need to deal with complex algorithms. Yet, the performance of using ChatGPT to predict the need for endotracheal intubation after HFNC therapy remains unexplored.

This study aims to assess the accuracy of GPT-3.5 and GPT-4.0 models in predicting the risk of endotracheal intubation within 48 h after the initiation of HFNC therapy. We developed a 6-alternative-forced-choice questionnaire based on patient baseline data and physiological parameters collected 6 h post-HFNC therapy for each of 71 patients prospectively included from multiple centers. GPT-3.5, GPT-4.0, specialist physicians in respiratory and critical care, and non-specialist physicians completed these 71 questionnaires respectively. We then compared the predictive performance of both models against that of specialist physicians and non-specialist physicians. Our findings show that GPT-4.0's predictive accuracy for the 48-h endotracheal intubation risk after HFNC therapy is comparable to that of specialist physicians.

2. Methods

2.1. Patients

This prospective cohort study initially included 73 patients from two Grade-A tertiary care and teaching hospitals. After applying exclusion criteria, the study ultimately included 71 patients who underwent HFNC oxygen therapy (Respircare HUMID BH).

2.2. Inclusion criteria

The study included patients aged 18 years and older who received HFNC oxygen therapy for various clinical needs, including those with or without type 2 respiratory failure.

2.3. Exclusion criteria

Patients were excluded if they received tracheostomy; refused intubation during 48-h HFNC therapy; requested withdrawal from the study; had incomplete data collection; or received intermittent non-invasive ventilation or prone position ventilation during 48-h HFNC therapy.

2.4. Study design

The observation endpoint for HFNC oxygen therapy was defined as any of the following: the initiation of endotracheal intubation or tracheotomy, patient death, or the completion of 48 h of HFNC therapy.

We followed up on clinical outcomes. The primary clinical outcome was the incidence of endotracheal intubation within 48 h. Secondary outcomes included the time until the initiation of endotracheal intubation, time until death, the rate of endotracheal intubation within 28 days, mortality rate within 28 days, and the length of the hospital stay. The endpoint of our follow-up was either the patient's death, their discharge from the hospital, or the completion of 28-day hospitalization.

We recorded the patient baseline data and physiological parameters for 71 patients at the initiation of HFNC therapy and again after 6 h. The patient baseline data included age, gender, body mass index (BMI), mechanical ventilation history, comorbidities, main diagnosis. Physiological parameters included blood gas analysis results, respiratory rate, heart rate, pulse oximetry (SpO_2), blood pressure, fraction of inspired oxygen (FiO_2), and oxygen flow and Glasgow Coma Scale (GCS) score.

These data were integrated into 71 natural language questionnaires, which asked participants to predict the 48-h endotracheal intubation risk after HFNC therapy based on scale options ranging from 1 to 6: 1) Extremely unlikely to undergo endotracheal intubation, 2) Unlikely to undergo endotracheal intubation, 3) Possible not to undergo endotracheal intubation, 4) Possible to undergo endotracheal intubation, 5 Likely to undergo endotracheal intubation, and 6) Extremely likely to undergo endotracheal intubation.

Box 1
The template of natural language questionnaire

A 67-year-old female patient was admitted to the hospital due to respiratory failure. The patient had not received mechanical ventilation treatment within the previous 24 h. The patient had a history of cerebrovascular disease, and had no history of smoking.

The patient received high-flow oxygen therapy. At the beginning of high-flow oxygen therapy, the Glasgow Coma Scale was 8 points, the systolic blood pressure was 94 mmHg, *the diastolic blood pressure was* 47 mmHg, *the respiratory rate was 32 breaths per minute, the heart rate was 130 beats per minute, the pulse oxygen saturation was 88 %, the oxygen flow rate was* 40 L/min, *the oxygen concentration was 55 %, and blood gas analysis showed* pH *7.24,* pO_2 61 mmHg, pCO_2 32 mmHg. *The patient had received vasopressor medication.*

After 6 h of high-flow oxygen therapy, the patient's systolic blood pressure was 83 mmHg, *the diastolic blood pressure was* 56 mmHg, *the respiratory rate was 25 breaths per minute, the heart rate was 127 beats per minute, the pulse oxygen saturation was 91 %, the oxygen flow rate was* 40 L/min, *the oxygen concentration was 55 %, and blood gas analysis showed* pH *7.46,* pO_2 72 mmHg, pCO_2 22 mmHg. *The patient had received vasopressor medication.*

Please predict the risk of endotracheal intubation within 48 h due to the failure of high-flow oxygen therapy according to the following options: 1. extremely unlikely to undergo endotracheal intubation; 2. unlikely to undergo endotracheal intubation; 3. possible not to undergo endotracheal intubation; 4. possible to undergo endotracheal intubation; 5. likely to undergo endotracheal intubation; 6. extremely likely to undergo endotracheal intubation.

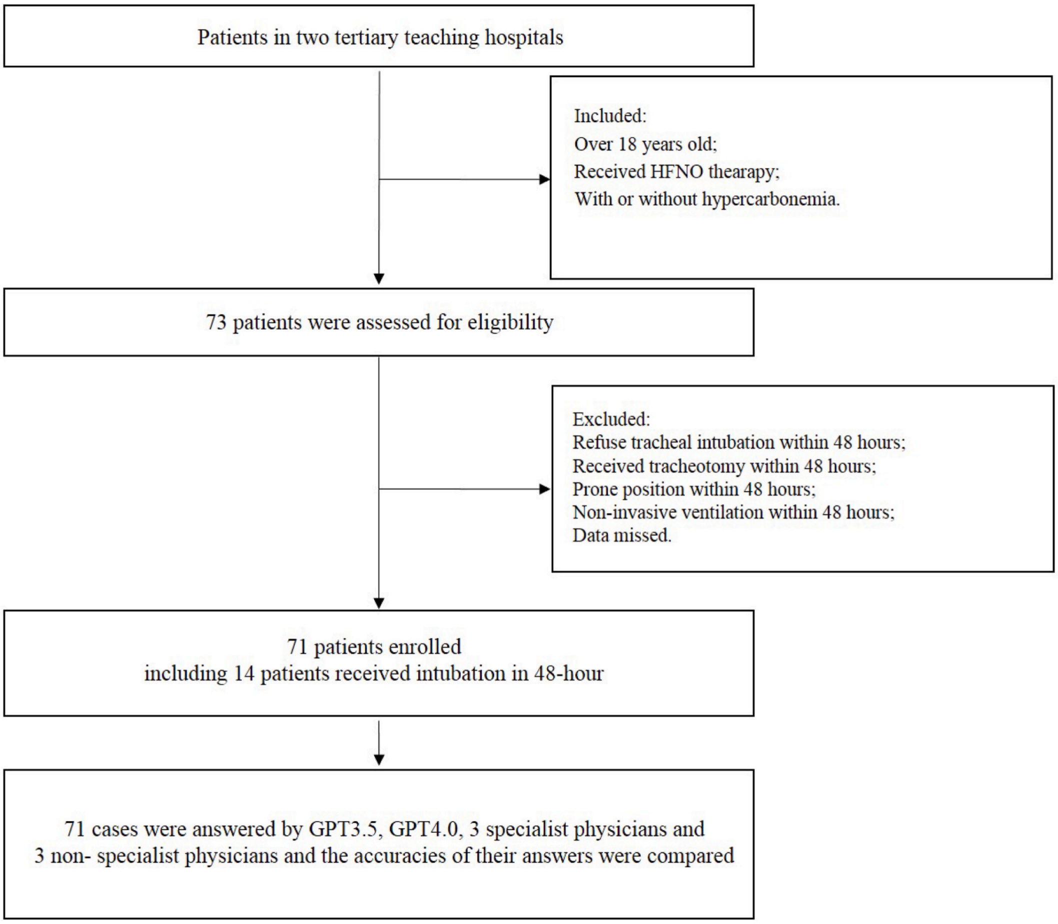
Patients in two tertiary teaching hospitals
Included:
Over 18 years old;
Received HFNO thearapy;
With or without hypercarbonemia.
73 patients were assessed for eligibility
Excluded:
Refuse tracheal intubation within 48 hours;
Received tracheotomy within 48 hours;
Prone position within 48 hours;
Non-invasive ventilation within 48 hours;
Data missed.
71 patients enrolled
including 14 patients received intubation in 48-hour
71 cases were answered by GPT3.5, GPT4.0, 3 specialist physicians and
3 non- specialist physicians and the accuracies of their answers were compared

Fig. 1. Flow chart of study.

One forced choice was required. A template of questionnaire was shown in Box 1.

Both GPT-3.5 and GPT-4.0 were used to predict the 48-h endotracheal intubation risk by prompting the contents of questionnaires. Besides, three respiratory and critical care specialist physicians, aged between 30 and 40, independently completed 23 to 24

Table 1
Clinical characteristics of patients with endotracheal intubation and without intubation within 48 h of treatment.

	all n = 71	Not intubated within 48 h n = 57	Intubation within 48 h n = 14	P
Age, mean (SD)	68.61 ± 15.32	69.42 ± 14.47	65.29 ± 18.66	0.369
Male, n (%)	45 (63.38 %)	36 (63.16 %)	9 (64.29 %)	0.664
BMI, mean (SD)	21.87 ± 3.80	21.73 ± 3.86	22.40 ± 3.59	0.558
Severe pneumonia, n (%)	29 (40.85 %)	21 (36.84 %)	8 (57.14 %)	0.166
Type 1 respiratory failure, n (%)	24 (33.80 %)	19 (33.33 %)	5 (35.71 %)	0.866
Sepsis, n (%)	10 (14.08 %)	9 (15.79 %)	1 (7.14 %)	0.504
Comorbidities				
COPD, n (%)	11 (15.49 %)	11 (19.30 %)	0 (0.00 %)	0.074
Other chronic lung diseases, n (%)	11 (15.49 %)	11 (19.30 %)	0 (0.00 %)	0.074
Coronary heart disease, n (%)	3 (4.23 %)	2 (3.51 %)	1 (7.14 %)	0.545
Heart failure, n (%)	6 (8.45 %)	4 (7.02 %)	2 (14.29 %)	0.381
Chronic kidney disease, n (%)	2 (2.82 %)	1 (1.75 %)	1 (7.14 %)	0.275
Cerebrovascular disease, n (%)	13 (18.31 %)	10 (17.54 %)	3 (21.43 %)	0.736
Active tumor, n (%)	7 (9.86 %)	5 (8.77 %)	2 (14.29 %)	0.535
Smoking history, n (%)	26 (36.62 %)	23 (40.35 %)	3 (21.43 %)	0.188
Mechanical ventilation within the previous 48 h, n (%)	13 (18.31 %)	11 (19.30)	2 (14.29 %)	0.664
When starting high flow oxygen therapy				
GCS, median [Q1, Q3]	13.5 [10.0, 15.0]	14.0 [10.0, 15.0]	13.0 [12.0, 15.0]	0.878
Heart rate, min^{-1}, median [Q1, Q3]	104.0 [91.0, 127.0]	104.0 [89.5, 126.0]	109.0 [96.8, 138.8]	0.227
Respiratory rate, min^{-1}, mean (SD)	27.07 ± 8.36	26.75 ± 8.11	28.36 ± 9.56	0.524
SBP, mmHg, mean (SD)	122.7 ± 23.33	121.77 ± 24.43	126.57 ± 18.45	0.494
DBP, mmHg, mean (SD)	72.27 ± 15.25	72.40 ± 15.99	71.71 ± 12.26	0.881
Receiving vasopressor, n (%)	8 (11.27 %)	6 (10.53 %)	2 (14.29 %)	0.690
pH, mean (SD)	7.41 ± 0.12	7.43 ± 0.10	7.33 ± 0.14	0.013
PaO$_2$, mmHg, median [Q1, Q3]	58.0 [48.0, 68.0]	59.0 [48.0, 68.0]	53.5 [49.3, 66.3]	0.511
PaCO$_2$, mmHg, median [Q1, Q3]	37.0 [29.0, 44.0]	38.0 [31.5, 43.0]	29.0 [23.5, 47.0]	0.211
SpO$_2$, mmHg, median [Q1, Q3]	88.0 [84.0, 93.0]	88.0 [84.5, 95.0]	87.0 [81.0, 90.0]	0.109
FiO$_2$, mmHg, median [Q1, Q3]	50.0 [40.0, 60.0]	50.0 [40.0, 57.5]	52.5 [43.8, 65.0]	0.231
Flow, L/min, mean (SD)	39.96 ± 9.35	39.33 ± 9.80	42.50 ± 7.00	0.259
High flow oxygen therapy at 6 h				
Heart rate, min^{-1}, median [Q1, Q3]	96.0 [88.0, 110.0]	95.0 [87.5, 108.0]	109.0 [96.8, 138.8]	0.099
Respiratory rate, min^{-1}, mean (SD)	22.85 ± 6.90	22.23 ± 6.22	25.36 ± 9.01	0.129
SBP, mmHg, mean (SD)	123.00 ± 18.76	123.12 ± 18.06	122.29 ± 22.14	0.882
DBP, mmHg, mean (SD)	72.41 ± 11.24	72.70 ± 10.80	71.21 ± 13.30	0.661
Receiving vasopressor, n (%)	6 (8.45 %)	3 (5.26 %)	3 (21.43 %)	0.051
pH, mean (SD)	7.41 ± 0.11	7.44 ± 0.07	7.28 ± 0.15	0.002
PaO$_2$, mmHg, median [Q1, Q3]	75.0 [64.0, 101.0]	81.5 [65.0, 115.8]	64.0 [53.0, 67.5]	0.006
PaCO$_2$, mmHg, median [Q1, Q3]	37.0 [31.0, 45.0]	37.0 [32.3, 43.8]	34.0 [29.5, 58.5]	0.602
SpO$_2$, mmHg, median [Q1, Q3]	96.0 [93.0, 98.0]	97.0 [94.0, 99.0]	91.5 [82.5, 95.25]	<0.001
FiO$_2$, mmHg, median [Q1, Q3]	50.0 [45.0, 60.0]	50.0 [42.5, 55.0]	55.0 [50.0, 80.0]	0.010
Flow, L/min, mean (SD)	40.28 ± 7.97	39.82 ± 8.56	42.14 ± 4.69	0.333
ROX at 6 h, mean (SD)	9.35 ± 3.97	9.86 ± 3.52	7.26 ± 5.05	0.027
Outcomes				
LOS in hospital, day, mean (SD)	19.30 ± 18.81	21.16 ± 19.24	12.00 ± 15.49	0.104
Intubation rate in 28 days, n (%)	21 (29.58 %)	7 (12.28 %)	14 (100.00 %)	<0.001
Mortality in 28 days, n (%)	16 (22.54 %)	9 (15.79 %)	7 (50.00 %)	0.006

questionnaires each, totaling 71 questionnaires. Similarly, three non-specialist physicians, also aged between 30 and 40, independently completed 23 to 24 questionnaires each, totaling 71 questionnaires. The 6-h ROX index, which is defined as (SpO$_2$/FiO$_2$)/respiratory rate [7], was calculated as an extra predictor for HFNC failure.

We assessed the performance of each predictor using the Receiver Operating Characteristic (ROC) analysis. Subsequently, patients were stratified into high-risk and low-risk groups based on the prediction values of GPT-4.0 to compare the 28-day cumulative endotracheal intubation rate and mortality between the two groups. The study flow is shown in Fig. 1.

2.5. Statistical analysis

Sample size calculation: This study employs a non-inferiority comparison using rates. According to previous studies, the overall accuracy of 6-h ROX index in predicting the 48-h endotracheal intubation risk in patients is ~0.8. The prediction accuracy by non-specialist physicians is estimated to be ~0.75; The prediction accuracy by GPT-4.0 is estimated to be ~0.85. Therefore, Pt = 0.85, *Pc* = 0.75, δ = 0.1, and the ratio of sample sizes between the two groups is 1:1. The calculation yields Nc = Nt = 62. Considering ~10 % of patients being lost to follow-up, 71 patients were planned to be included to create the questionnaires.

Quantitative data following a normal distribution were presented as the arithmetic mean ± standard deviation (SD). Non-normally distributed data were presented as the median (interquartile range, IQR). Two-independent-sample t-tests and Mann-Whitney U tests were used for intergroup comparisons of continuous variables. And the chi-square test was used for rate comparisons.

Table 2
Prediction results.

	GPT-3.5's prediction	48-h intubation in practice	GPT-4.0's prediction	48-h intubation in practice	Specialist physicians' prediction	48-h intubation in practice	Non-specialist physicians' prediction	48-h intubation in practice
1 extremely unlikely to undergo endotracheal intubation	0	0	0	0	13	2	14	0
2 unlikely to undergo endotracheal intubation	9	0	29	1	22	0	11	3
3 possible not to undergo endotracheal intubation	3	0	17	2	16	2	10	1
4 possible to undergo endotracheal intubation	32	3	17	6	6	2	7	1
5 likely to undergo endotracheal intubation	27	11	7	4	6	3	14	5
6 extremely likely to undergo endotracheal intubation	0	0	1	1	8	5	15	4

After constructing the Receiver Operating Characteristic (ROC) curve, the optimal diagnostic cutoff point was determined using the maximum Youden index. The Log-rank test was used to compare differences in 28-day mortality rates and to construct Kaplan-Meier survival curves.

All data analyses were carried out using SPSS 26.0. GraphPad 8.0 was used for data visualization. A p-value of less than 0.05 was considered statistically significant.

3. Results

Among 71 patients included in the study, the causes of their conditions were as follows: severe pneumonia (29 cases, 40.85 %), type 1 respiratory failure (24 cases, 33.80 %), and sepsis (10 cases, 14.08 %). As a result, 14 (19.72 %) required endotracheal intubation within 48 h following HFNC therapy, 21 (29.58 %) required endotracheal intubation within 28 days, and 16 (22.53 %) died. There were no statistically significant differences in baseline data between the intubation group and the non-intubation group, including age, gender, BMI, comorbidities, and other factors (all $P > 0.05$). However, after 6 h of HFNC oxygen therapy, significantly decreased levels of pH, PaO_2, and SpO_2 (all $P < 0.05$) were observed in the intubation group compared to the non-intubation group (see Table 1).

To determine the accuracy of predictions by GPT-3.5, GPT-4.0, specialist physicians, non-specialist physicians and the ROX index, we compared their prediction results with the actual clinical outcomes in patients (see Table 2 and Fig. 2). Subsequently, we constructed the Receiver Operating Characteristic (ROC) curves for each predictor and compared the area under the curve (AUC) to evaluate their performance (see Fig. 3). The optimal diagnostic cutoff point was determined using the maximum Youden index. And the overall accuracy, specificity, sensitivity, positive predictive value, and negative predictive value were calculated accordingly.

The optimal diagnostic cutoff points were determined to be ≥ 4 for both GPT-4.0 and specialist physicians. GPT-4.0 demonstrated a precision of 76.1 %, with a specificity of 78.6 % (95%CI = 52.4–92.4 %) and sensitivity of 75.4 % (95%CI = 62.9–84.8 %). The positive predictive value was 40.7 %, and the negative predictive value was 93.5 %. In comparison, the precision of specialist physicians was 80.3 %, with a specificity of 71.4 % (95%CI = 45.4–88.3 %) and sensitivity of 82.5 % (95%CI = 70.6–90.2 %). The positive predictive value was 50.0 %, and the negative predictive value was 92.2 %. For GPT-3.5 and non-specialist physicians, the optimal diagnostic cutoff points were ≥5, with precisions of 73.2 % and 64.8 %, respectively.

For GPT-4.0, the area under the curve (AUC) in ROC analysis was 0.821 (95 % CI = 0.698–0.943), marking it as the highest among the predictors. This value, however, was not significantly higher ($p > 0.05$) than the AUCs of GPT-3.5 [0.775 (95%CI = 0.652–0.898)] and specialist physicians [0.782 (95%CI = 0.619–0.945)]. Nonetheless, it was significantly greater than the AUC of non-specialist physicians [0.662 (95%CI = 0.518–0.805), P = 0.011] as illustrated in Table 3 and Fig. 3.

Patients were further divided into high-risk (N = 25) and low-risk (N = 46) groups based on the GPT-4.0's prediction values (≥4 as high-risk). The high-risk group exhibited significantly higher 28-day cumulative intubation rate (56.00 % vs. 15.22 %, $P < 0.001$) and 28-day mortality (44.00 % vs. 10.87 %, $P < 0.001$) compared to the low-risk group, as illustrated in Fig. 4. Furthermore, there were statistically significant differences in the parameters of heart rate, respiratory rate, pH, PaO2, SpO2, FiO2, and oxygen flow rate after 6 h of HFNC therapy (all $P < 0.05$) between the two groups of patients (see Table 4).

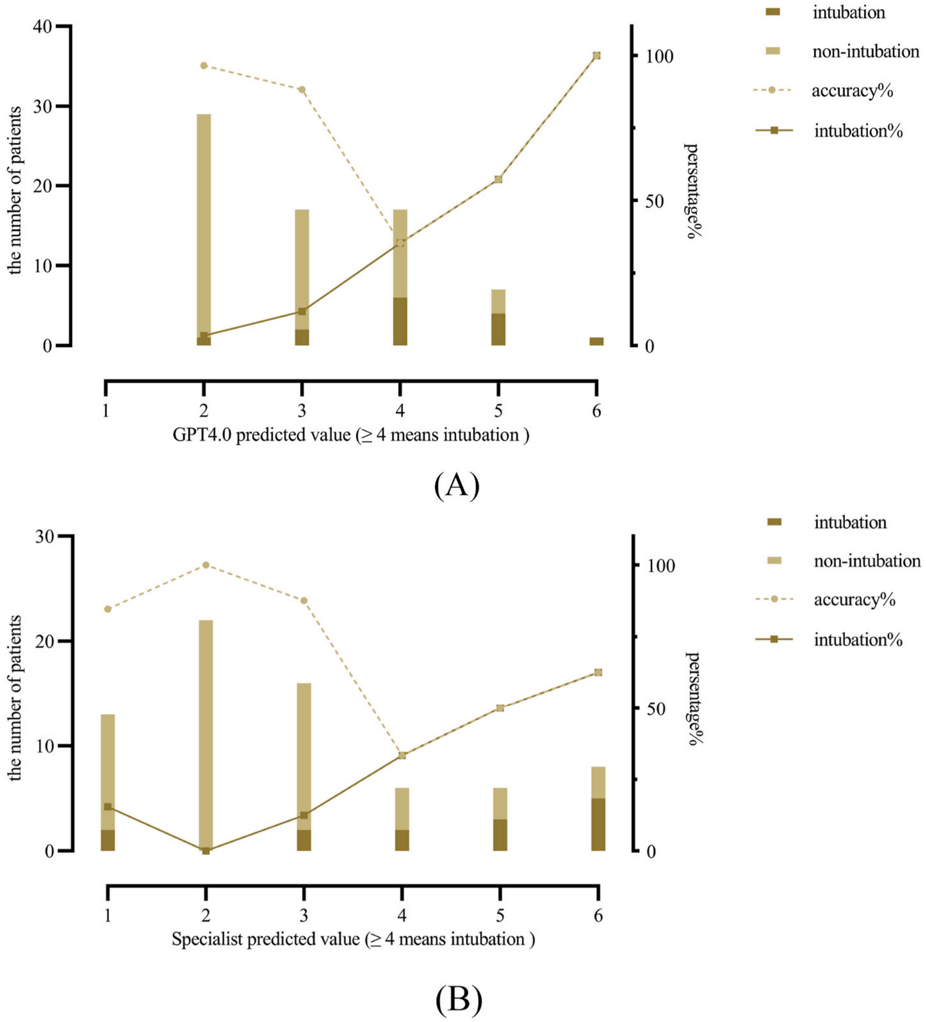

Fig. 2. Distribution of accuracy in predicting endotracheal intubation within 48 h between GPT-4.0 and specialist physicians. (A) The accuracy in predicting endotracheal intubation within 48 h of GPT-4.0. (B) The accuracy in predicting endotracheal intubation within 48 h of specialist physicians.

4. Discussion

The present study assessed the performance of advanced large language models (LLMs), namely GPT-3.5 and GPT-4.0, in comparison with that of both specialist in respiratory and critical care and non-specialist physicians to predict the risk of endotracheal intubation within 48 h following the initiation of high-flow nasal cannula (HFNC) oxygen therapy. We developed a novel questionnaire-based method with scale options ranging from 1 to 6 to guide participants in predicting the likelihood of intubation, and evaluated the predictive accuracy of each predictor using Receiver Operating Characteristic (ROC) analysis. Our findings reveal that the GPT-4.0 model demonstrates predictive accuracy comparable to that of specialist physicians, which marks a significant advancement in the application of AI-assisted clinical decision-making in a natural language manner.

According to the natural language description of the outcomes, the standard cutoff point should be ≥ 4 (3. possible not to undergo endotracheal intubation; 4. possible to undergo endotracheal intubation). We determined the optimal diagnostic cutoff point for each predictor using the maximum Youden index. For both GPT-4.0 and specialist physicians, their optimal diagnostic cutoff points were determined be ≥ 4, aligning with the standard cut-off point based on the outcomes' description. In contrast, the optimal cutoff points for both GPT-3.5 and non-specialist physicians were set at ≥5, suggesting a propensity to overestimate the endotracheal intubation risk due to the lack of clinical experience.

The overall predictive accuracy of GPT-4.0 exhibited a good negative predictive value along with an AUC of 0.821, which was significantly higher than that of non-specialist physicians (AUC = 0.662, P = 0.011). In comparison, the AUCs of GPT-3.5, specialist physicians, and ROX index were all lower than 0.8. Our results suggest that GPT-4.0 demonstrates clinical judgment experience on 48-h endotracheal intubation risk following HFNC therapy, which was at least better than non-specialist physicians and comparable to

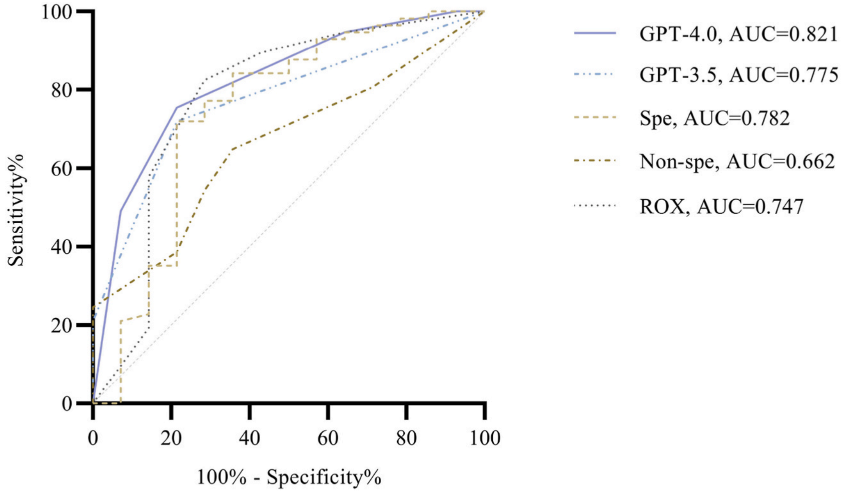

Fig. 3. ROC curves of predicting endotracheal intubation within 48 h for GPT and clinical physicians.

specialist physicians (AUC = 0.782). After grouping the patients according to GPT-4.0's optimal diagnostic cutoff point (i.e., ≥4), there were significant differences between the high-risk and low-risk groups in parameters such as heart rate, respiratory rate, pH, PaO_2, SpO_2, FiO_2, and oxygen flow rate after 6 h of high-flow oxygen therapy, indicating that GPT-4.0 can effectively leverage these clinical features to make accurate judgements.

However, GPT-4.0 tended to avoid extreme judgements (i.e., the options of 1 and 6 in the questionnaire), whereas physicians tended to give a diverse range of answers based on their individual clinical judgment. Moreover, both GPT-4.0 and specialist physicians exhibited low accuracy in identifying patients categorized as '4. possible to undergo endotracheal intubation', which represent a group that is particularly critical for early screening to prevent delayed intubation [9]. It is also worth noting that the physicians involved in the actual treatment could also make errors in clinical judgements in a short treatment period of 48 h given that clinical practice is influenced by various factors. Therefore, it does not necessarily mean that the specialist physicians in our study were wrong when their predictions do not match the actual clinical outcomes. It is also essential to acknowledge that the inclusion of patients with different etiologies introduces heterogeneity that could potentially affect the predictive accuracy of both the AI models and the physician assessments. The small sample size of our study restricts our ability to fully explore the implications of patient heterogeneity on predictive accuracy. Future research with larger sample size would provide a more definitive comparison of the predictive accuracies between clinical physicians and GPT models.

The ROX index can be used to predict the failure of HFNC therapy. However, it offers only a moderate level of predictive accuracy and lacks a unified standard diagnostic cutoff point as the ROX index only includes three parameters, i.e., SpO_2/FiO_2 to respiratory rate [8,10]. Incorporating a broader range of physiological parameters could enhance predictive accuracy [11,12]. Therefore, we aimed to improve the predictive accuracy about the endotracheal intubation risk by collecting a more comprehensive set of baseline data and physiological parameters of patients and leveraging the algorithm models of ChatGPT. We contend that using GPT-4.0 to predict the endotracheal intubation risk following HFNC oxygen therapy holds substantial clinical promise [13]. GPT-4.0 demonstrates a great potential to outperform the specialist physicians in judging endotracheal intubation risk following HFNC therapy [14]. With its rapid development, fast processing speed and ease of use, GPT could serve as a tool for dynamically monitoring patient data, potentially reducing labor costs [15].

Nevertheless, we raise ethical concerns regarding the reliance on GPT for clinical decision-making. GPT's decision was based on accumulated data from actual clinical practice. An overreliance on GPT by physicians for clinical decisions could inadvertently reinforce GPT's decision-making patterns without a corresponding increase in clinical accuracy. This scenario risks transforming GPT into a self-validating system, potentially misaligned with the actual needs of clinical practice. We cannot expect artificial intelligence to " lift itself by its own bootstraps." Therefore, it is crucial to develop corresponding ethical guideline for the clinical application of GPT [16,17].

Limitations: 1. The answers of GPT are not entirely stable and can give different but similar answers for the same questionnaire. 2. This study is a multicenter prospective cohort study including only 71 patients, and a small number of specialist and non-specialist physicians. Subgroup analysis was not performed for these patients due to small sample size. Further analysis can be conducted in subsequent large-scale cohort studies.

Funding statement

T.L. was supported by National High Level Hospital Clinical Research Funding [BJ-2023-173]. T.L. and L.S. were supported by Shenyang RMS Medical Tech Company [20210901].

Table 3

Comparison of ROC area and accuracy for predicting endotracheal intubation.

	AUC (95 % CI)	P	Cut-off	Sensitivity (95 % CI)	Specificity% (95 % CI)	positive predictive value	negative predictive value	Accuracy
GPT-4.0	0.821 (0.698–0.943)	–	≥4	75.4 % (62.9–84.8 %)	78.6 % (52.4–92.4 %)	40.7 %	93.5 %	76.1 %
GPT-3.5	0.775 (0.652–0.898)	0.484	≥5	71.9 % (59.2–81.9 %)	78.6 % (52.4–92.4 %)	40.7 %	93.2 %	73.2 %
Specialist	0.782 (0.619–0.945)	0.475	≥4	82.5 % (70.6–90.2 %)	71.4 % (45.4–88.3 %)	50.0 %	92.2 %	80.3 %
Non-Specialist	0.662 (0.518–0.805)	0.011	≥5	64.9 % (51.9–76.0 %)	71.4 % (45.4–88.3 %)	31.0 %	88.1 %	64.8 %
ROX index	0.746 (0.576–0.916)	0.296	≤7.90	71.9 % (59.2–81.9 %)	78.6 % (52.4–92.4 %)	40.7 %	93.2 %	73.2 %

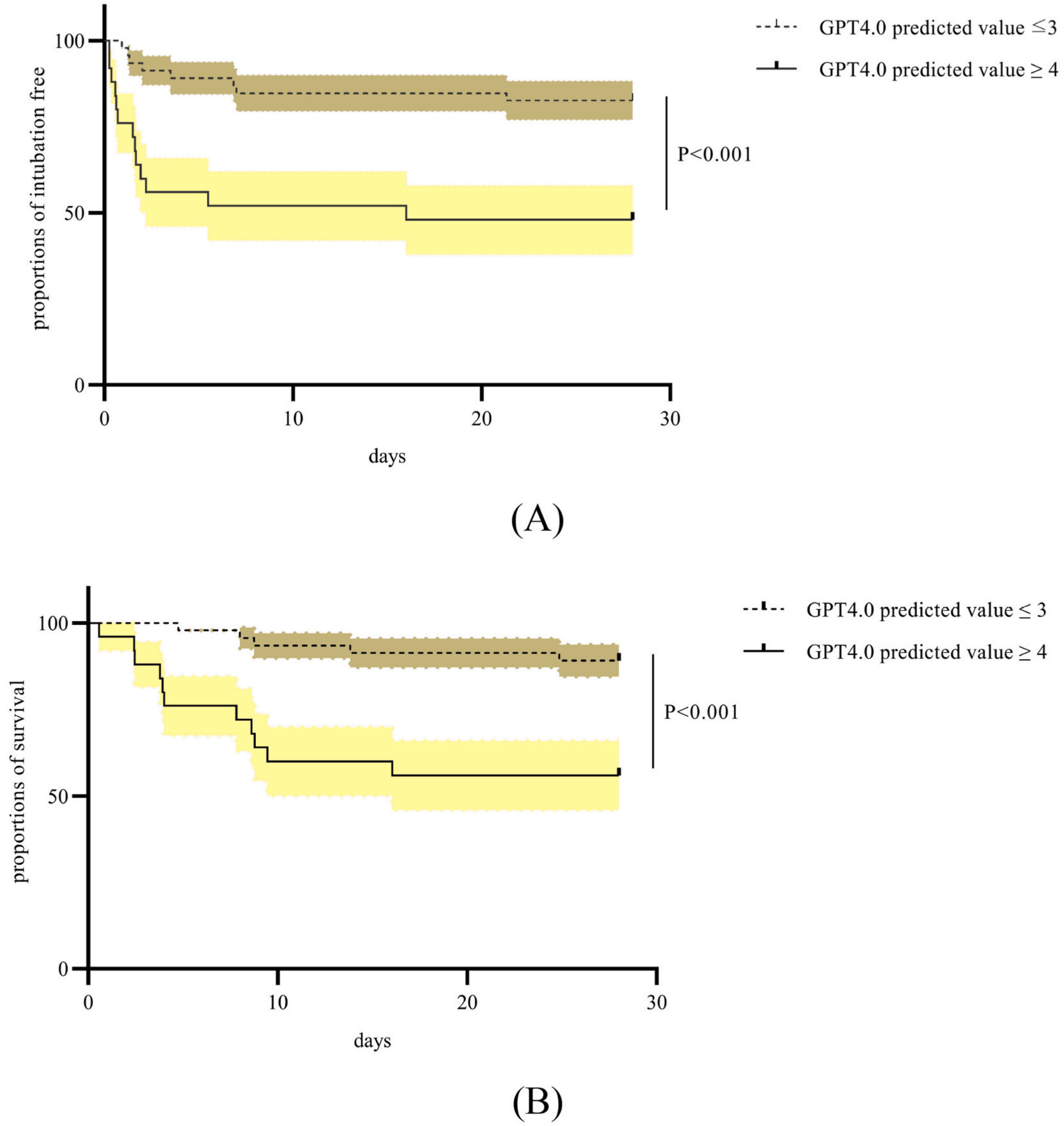

Fig. 4. Cumulative endotracheal intubation free curve and cumulative survival curve of patients grouped by prediction values of GPT-4.0 over 28 days of treatment. (A) Cumulative endotracheal intubation free curve grouped by prediction values of GPT-4.0 over 28 days of treatment. (B) Cumulative survival curve of patients grouped by prediction values of GPT-4.0 over 28 days of treatment.

Ethics statement

This study was reviewed and approved by the First Affiliated Hospital of Henan University of Science and Technology, with the approval number: 2021-0241 and Jiangyan Hospital Affiliated to Nanjing University of Chinese Medicine, with the approval number: 2021-016. All patients (or their proxies/legal guardians) provided informed consent to participate in the study. The study was registered as a clinical trial (ChiCTR2100053027). Full study protocol can be accessed from https://www.chictr.org.cn.

Data availability statement

Data included in article/supp. material/referenced in article.

CRediT authorship contribution statement

Taotao Liu: Writing – original draft, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Yaocong Duan:** Writing – original draft, Software, Methodology, Formal analysis, Data curation. **Yanchun Li:**

Table 4
Using GPT-4.0 to predict baseline data and prognosis of patients with and without endotracheal intubation.

	GPT-4.0 ≤ 3 n = 46	GPT-4.0 ≥ 4 n = 25	P
Age, mean (SD)	68.26 ± 15.35	69.24 ± 15.58	0.800
Male, n (%)	30 (65.22 %)	15 (60.00 %)	0.663
BMI, mean (SD)	21.71 ± 3.65	22.15 ± 4.11	0.662
Severe pneumonia, n (%)	15 (32.61 %)	14 (56.00 %)	0.055
Type 1 respiratory failure, n (%)	16 (34.78 %)	8 (32.00 %)	0.813
Sepsis, n (%)	6 (13.04 %)	4 (16.00 %)	0.732
Comorbidities			
COPD, n (%)	9 (19.57 %)	2 (8.00 %)	0.198
Other chronic lung diseases, n (%)	7 (15.22 %)	4 (16.00 %)	0.931
Coronary heart disease, n (%)	2 (4.34 %)	1 (4.00 %)	0.945
Heart failure, n (%)	4 (8.70 %)	2 (8.00 %)	0.920
Chronic kidney disease, n (%)	0 (0.00 %)	2 (8.00 %)	0.052
Cerebrovascular disease, n (%)	7 (15.21 %)	6 (24.00 %)	0.361
Active tumor, n (%)	3 (6.52 %)	4 (16.00 %)	0.201
Smoking history, n (%)	20 (43.48 %)	6 (24.00 %)	0.104
Mechanical ventilation within the previous 48 h, n (%)	9 (19.57 %)	4 (16.00 %)	0.711
When starting high flow oxygen therapy			
GCS, median [Q1,Q3]	15.0 [12.0, 15.0]	12.0 [8.5, 14.5]	0.008
Heart rate, min^{-1}, median [Q1,Q3]	101.5 [92.5, 119.3]	110.0 [83.0, 134.0]	0.243
Respiratory rate, min^{-1}, mean (SD)	25.76 ± 7.90	29.48 ± 8.81	0.073
SBP, mmHg, mean (SD)	121.46 ± 24.12	125.04 ± 22.09	0.540
DBP, mmHg, mean (SD)	73.48 ± 16.56	70.04 ± 12.50	0.368
Receiving vasopressor, n (%)	3 (6.52 %)	5 (20.00 %)	0.086
pH, mean (SD)	7.43 ± 0.10	7.38 ± 0.13	0.164
PaO$_2$, mmHg, median[Q1,Q3]	60.0 [52.0, 72.3]	51.0 [41.5, 63.5]	0.014
PaCO$_2$, mmHg, median[Q1,Q3]	38.0 [32.5, 42.5]	32.0 [26.5, 47.0]	0.297
SpO$_2$, mmHg, median[Q1,Q3]	89.5 [84.8, 92.3]	87.0 [82.5, 90.0]	0.096
FiO$_2$, mmHg, median[Q1,Q3]	47.5 [40.0, 55.0]	40.0 [40.0, 47.5]	0.042
Flow, L/min, mean (SD)	38.85 ± 10.06	42.00 ± 7.64	0.177
High flow oxygen therapy at 6 h			
Heart rate, min^{-1}, median[Q1,Q3]	95.0 [86.8, 105.3]	108.0 [89.0, 122.0]	0.021
Respiratory rate, min^{-1}, mean (SD)	20.72 ± 5.06	26.76 ± 8.14	0.002
SBP, mmHg, mean (SD)	123.37 ± 15.65	122.20 ± 23.80	0.826
DBP, mmHg, mean (SD)	73.63 ± 10.92	70.16 ± 11.70	0.217
pH, mean (SD)	7.43 ± 0.70	7.36 ± 0.15	0.025
PaO$_2$, mmHg, median[Q1,Q3]	92.0 [73.5, 125.0]	60.0 [53.0, 67.8]	<0.001
PaCO$_2$, mmHg, median[Q1,Q3]	37.0 [33.0, 45.0]	37.0 [29.3, 51.3]	0.910
SpO$_2$, mmHg, median[Q1,Q3]	97.5 [95.8, 99.0]	92.0 [86.0, 93.0]	<0.001
FiO$_2$, mmHg, median[Q1,Q3]	45.0 [40.0, 55.0]	55.0 [50.0, 77.5]	0.001
Flow, L/min, mean (SD)	38.59 ± 8.48	43.40 ± 5.90	0.014
Outcomes			
LOS in hospital, day, mean (SD)	21.64 ± 21.25	15.20 ± 12.87	0.174
Intubation rate in 48 h, n (%)	3 (6.52 %)	11 (44.00 %)	<0.001
Intubation rate in 28 days, n (%)	7 (15.22 %)	14 (56.00 %)	<0.001
Mortality in 28 days, n (%)	5 (10.87 %)	11 (44.00 %)	<0.001

Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Yingying Hu:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Lingling Su:** Writing – review & editing, Writing – original draft, Visualization, Validation, Formal analysis, Data curation. **Aiping Zhang:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Thanks to Shenyang RMS Medical Tech Company and Mr. Zhong Zhang for providing technical support.

List of abbreviations

HFNC	high-flow nasal cannula
NIV	noninvasive ventilation

ICU	intensive care unit
BMI	body mass index
ABG	arterial blood gas
IQR	interquartile range
AUC	area under the receiver operating characteristic curve
ROX index	ratio of SpO_2/FiO_2 to respiratory rate

References

[1] J.P. Frat, et al., High-flow oxygen through nasal cannula in acute hypoxemic respiratory failure, N. Engl. J. Med. 372 (23) (2015) 2185–2196.

[2] G. Spoletini, et al., Heated humidified high-flow nasal oxygen in adults: mechanisms of action and clinical implications, Chest 148 (1) (2015) 253–261.

[3] G. Hernandez, et al., Effect of postextubation high-flow nasal cannula vs noninvasive ventilation on reintubation and postextubation respiratory failure in high-risk patients A randomized clinical trial supplemental content, JAMA 316 (2016).

[4] K. Nagata, et al., Home high-flow nasal cannula oxygen therapy for stable hypercapnic copd: a randomized clinical trial, Am. J. Respir. Crit. Care Med. 206 (11) (2022) 1326–1335.

[5] J. Li, et al., Awake prone positioning for non-intubated patients with COVID-19-related acute hypoxaemic respiratory failure: a systematic review and meta-analysis, Lancet Respir. Med. 10 (6) (2022) 573–583.

[6] B.J. Kang, et al., Failure of high-flow nasal cannula therapy may delay intubation and increase mortality, Intensive Care Med. 41 (4) (2015) 623–632.

[7] O. Roca, et al., Predicting success of high-flow nasal cannula in pneumonia patients with hypoxemic respiratory failure: the utility of the ROX index, J. Crit. Care 35 (2016) 200–205.

[8] J. Prakash, et al., ROX index as a good predictor of high flow nasal cannula failure in COVID-19 patients with acute hypoxemic respiratory failure: a systematic review and meta-analysis, J. Crit. Care 66 (2021) 102–108.

[9] M.L. Vega, et al., COVID-19 Pneumonia and ROX index: time to set a new threshold for patients admitted outside the ICU, Pulmonology 28 (1) (2022) 13–17.

[10] A. Chandel, et al., High-flow nasal cannula therapy in COVID-19: using the ROX index to predict success, Respir. Care 66 (6) (2021) 909–919.

[11] A. Kansal, et al., Comparison of ROX index (SpO(2)/FIO(2) ratio/respiratory rate) with a modified dynamic index incorporating PaO(2)/FIO(2) ratio and heart rate to predict high flow nasal cannula outcomes among patients with acute respiratory failure: a single centre retrospective study, BMC Pulm. Med. 22 (1) (2022) 350.

[12] T. Liu, Q. Zhao, B. Du, Effects of high-flow oxygen therapy on patients with hypoxemia after extubation and predictors of reintubation: a retrospective study based on the MIMIC-IV database, BMC Pulm. Med. 21 (1) (2021) 160.

[13] G.S. Collins, K.G.M. Moons, Reporting of artificial intelligence prediction models, Lancet 393 (10181) (2019) 1577–1579.

[14] E. Ergin, et al., Can artificial intelligence and robotic nurses replace operating room nurses? The quasi-experimental research, J Robot Surg (2023) 1–9.

[15] M. Areia, et al., Cost-effectiveness of artificial intelligence for screening colonoscopy: a modelling study, Lancet Digit Health 4 (6) (2022) e436–e444.

[16] E.W. Kluge, Artificial intelligence in healthcare: ethical considerations, Healthc Manage Forum 33 (1) (2020) 47–49.

[17] S. Sunarti, et al., Artificial intelligence in healthcare: opportunities and risk for future, Gac. Sanit. 35 (Suppl 1) (2021) S67–s70.