Nursing, AI, and Clinical Research: 2024 Insights

May 30, 2025

Andrew Phillips PhD, RN, FAMIA

Conflict of Interest

Have no real or apparent conflicts of interest to report.

Learning Objectives

- How will AI potentially impact nursing practice

Methods – “Phillips” Study

- Step 1 – Identify the Research Question
- Step 2 – Search for Relevant Literature
- Step 3 – Literature Selection/Review
  - Sub Step – Take Nap
  - Sub Step – Caffeine
- Step 4 – Summarizing, and reporting the results
- Step 5 – Consultation – This is you guys

Step 1: Research Question

- What trends and themes emerge from a survey of the published literature in the area of AI in nursing during the past year

Step 2: Identify Relevant Literature

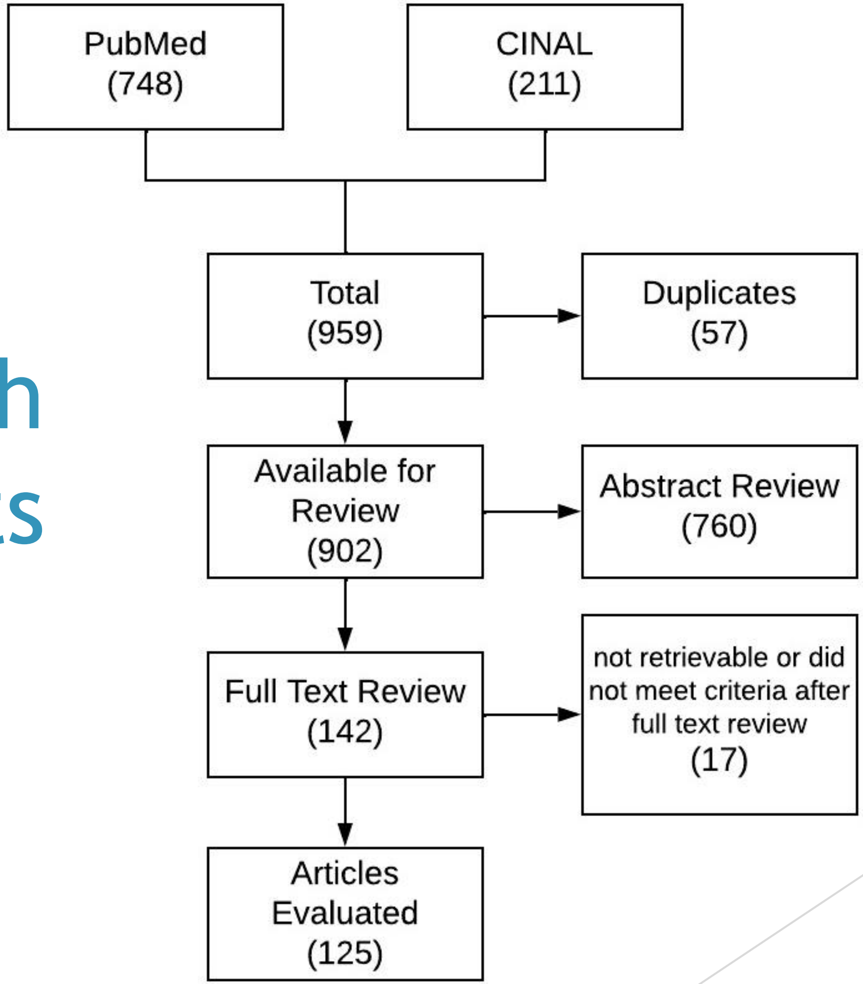
- Search Strategy
  - Databases: PubMed and CINAHL
  - Search terms
- (“artificial intelligence” OR “AI” OR “large language models” OR “machine learning”) AND healthcare AND nurs*
- Publication Dates 3/1/2024 – 3/31/2025

Step 3: Study Selection

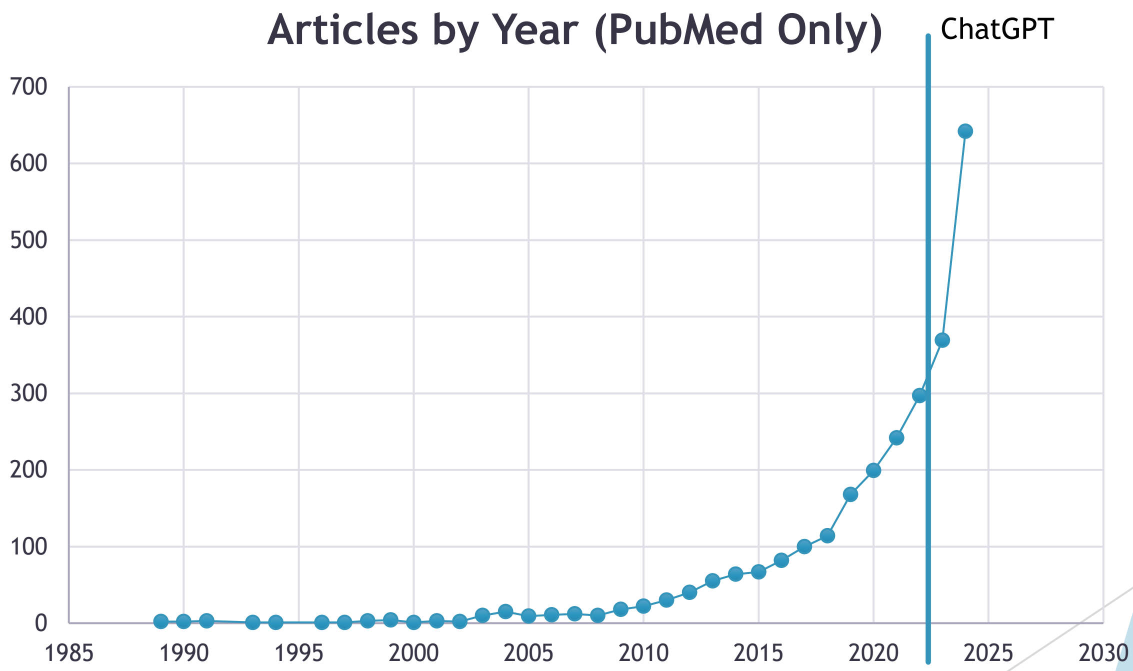
Inclusion and Exclusion Criteria

- Inclusion criteria: healthcare delivery, relevant to nursing, mentions nursing in the abstract
- Exclusions: Articles without an AI informatics focus

Search Results



Search Results



Step 4 – Collating, summarizing, and reporting the results

Themes Identified

1. AI Literacy MUST be a part of Nursing Education
   - Increasing use of tools
   - Understanding ethical limitations
2. AI as Opportunity or Threat
3. Attitudes towards AI
4. Readiness
5. Too much AI – Alert Fatigue
6. Erode clinical thinking skills/Enhance Clinical Thinking Skills
7. Ever increasing amounts of data
8. Patient communications
   - Patient portals (providers only)
   - Mental health/support
9. Robots – mentioned frequently but few studies
   - Companion
   - VS monitoring
   - Medication dispensing
   - Blood draws
   - Disinfection and sterilization
   - Patient transport
10. Ethics
    - Security
    - Privacy
    - Errors
    - Generalizability
    - Over Reliance
    - Algorithmic Bias
    - Generalizability
    - TRUST
11. What wasn't emphasized
    - Greater details on how model was trained
12. Still, most of the literature was some form of literature review
13. Social Determinant of Health
14. Relieving workload
15. Risk prediction models galore
16. Much more I did not understand this year

Representative/Interesting Citations

Loughran, E., Kane, M., Wyatt, T. H., Kerley, A., Lowe, S., & Li, X. (2024). Using large language models to address health literacy in mHealth: case report. CIN: Computers, Informatics, Nursing, 10-1097.

Wagner, L., Jourdan, S., Mayer, L., Müller, C., Bernhard, L., Kolb, S., ... & Wilhelm, D. (2024). Robotic scrub nurse to anticipate surgical instruments based on real-time laparoscopic video analysis. Communications Medicine, 4(1), 156.

Wan, P., Huang, Z., Tang, W., Nie, Y., Pei, D., Deng, S., ... & Long, E. (2024). Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. Nature Medicine, 30(10), 2878-2885.

Wang, T., Mu, J., Chen, J., & Lin, C. C. (2024). Comparing ChatGPT and clinical nurses’ performances on tracheostomy care: A cross-sectional study. International Journal of Nursing Studies Advances, 6, 100181.

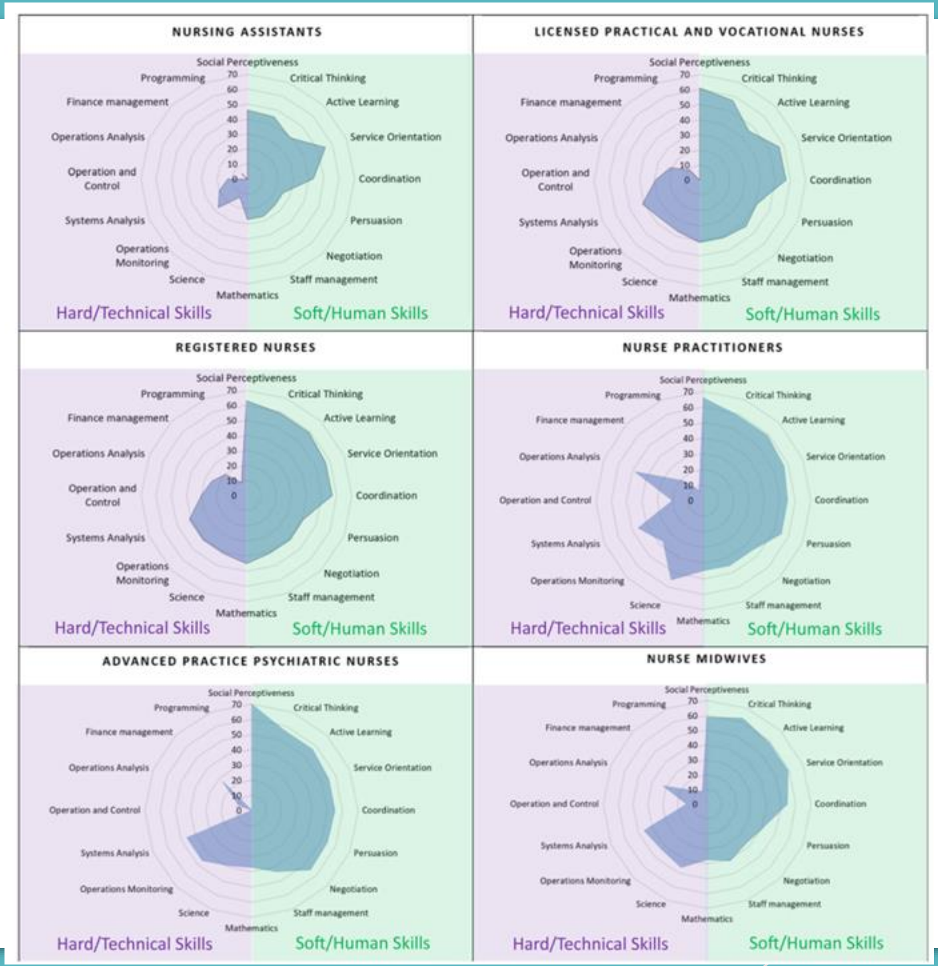
Yakusheva, O., Bouvier, M. J., & Hagopian, C. O. (2025). How Artificial Intelligence is altering the nursing workforce. Nursing Outlook, 73(1), 102300.

Zaboli, A., Brigo, F., Sibilio, S., Mian, M., & Turcato, G. (2024). Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage?. The American Journal of Emergency Medicine, 79, 44-47.

Zolnoori, M., Vergez, S., Xu, Z., Esmaeili, E., Zolnour, A., Anne Briggs, K., ... & McDonald, M. V. (2024). Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare. JAMIA open, 7(4), ooae130.

Yakusheva, O., Bouvier, M. J., & Hagopian, C. O. (2025). How Artificial Intelligence is altering the nursing workforce. Nursing Outlook, 73(1), 102300.

- This paper focuses on the implications of Artificial Intelligence (AI) for the nursing workforce, examining both the **opportunities presented by AI in relieving nurses of routine tasks and enabling better patient care**, and the potential challenges it poses. The discussion highlights the freeing of nurses' time from administrative duties, allowing for more patient interaction and professional development, while also acknowledging concerns about job displacement. Ethically integrating AI into patient care and the need for nurses' **proactive engagement with AI—including involvement in its development and integration in nursing education—are emphasized**. Finally, the paper asserts the necessity for nurses to become active participants in AI's evolution within health care to ensure the enhancement of patient care and the advancement of nursing roles


NURSING ASSISTANTS
LICENSED PRACTICAL AND VOCATIONAL NURSES
REGISTERED NURSES
NURSE PRACTITIONERS
ADVANCED PRACTICE PSYCHIATRIC NURSES
NURSE MIDWIVES
Social Perceptiveness
Critical Thinking
Active Learning
Service Orientation
Coordination
Persuasion
Negotiation
Staff management
Mathematics
Science
Operations Monitoring
Systems Analysis
Operation and Control
Operations Analysis
Finance management
Programming
0
10
20
30
40
50
60
70
Hard/Technical Skills
Soft/Human Skills

Zaboli, A., Brigo, F., Sibilio, S., Mian, M., & Turcato, G. (2024). Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage?. The American Journal of Emergency Medicine, 79, 44-47.

- Introduction: Chat-GPT is rapidly emerging as a promising and potentially revolutionary tool in medicine. One of its possible applications is the stratification of patients according to the severity of clinical conditions and prognosis during the triage evaluation in the emergency department (ED).
- Methods: **Using a randomly selected sample of 30 vignettes recreated from real clinical cases, we compared the concordance in risk stratification of ED patients between healthcare personnel and Chat-GPT.** The concordance was assessed with Cohen's kappa, and the performance was evaluated with the area under the receiver operating characteristic curve (AUROC) curves. Among the outcomes, we considered mortality within 72 h, the need for hospitalization, and the presence of a severe or time-dependent condition.
- Results: The concordance in triage code assignment between triage nurses and Chat-GPT was 0.278 (unweighted Cohen's kappa; 95% confidence intervals: 0.231–0.388). For all outcomes, the ROC values were higher for the triage nurses**. The most relevant difference was found in 72-h mortality, where triage nurses showed an AUROC of 0.910 (0.757–1.000) compared to only 0.669 (0.153–1.000) for Chat-GPT.**
- Conclusions**: The current level of Chat-GPT reliability is insufficient to make it a valid substitute for the expertise of triage nurses in prioritizing ED patients. Further developments are required to enhance the safety and effectiveness of AI for risk stratification of ED patients.**

Table 1
Example of a vignette given to triage nurses and inserted within Chat-GPT for code assignment.

Case	Gender, age and access	Medical history	Vital parameters	MTS standard
Case 1	43-year-old woman accessing the emergency department with a private car	She has been reporting palpitations with a sensation of a knot in the throat for 15 min. She does not report other pathologies. She is very agitated but does not report chest pain. Onset of palpitations while she was at the supermarket.	RR: 22 HR: 170 BP: 190/100 Temp: 36.1 SpO2: 100 GCS: 15 NRS: 1	Red Orange Yellow Green Blue

Table 3
Comparisons of the performance of HP-MTS and AI-MTS for the classification of severe or time-dependent conditions, and for the prediction of 72-h mortality and hospitalization.

Outcome	AUROC	95%CI	*p*-value
Severe condition			0.802
HP-MTS	0.805	0.645–0.965	
Chat-GPT-MTS	0.781	0.621–0.941	
Time-dependent condition			0.133
HP-MTS	0.953	0.897–1.000	
Chat-GPT-MTS	0.810	0.620–1.000	
Death at 72 h			0.210
HP-MTS	0.910	0.757–1.000	
Chat-GPT-MTS	0.669	0.153–1.000	
Hospital admission			0.958
HP-MTS	0.823	0.641–1.000	
Chat-GPT-MTS	0.818	0.652–0.984	

Legend: AUROC: area under the receiver operating characteristic curve; CI: confidence intervals.

Wang, T., Mu, J., Chen, J., & Lin, C. C. (2024). Comparing ChatGPT and clinical nurses' performances on tracheostomy care: A cross-sectional study. International Journal of Nursing Studies Advances, 6, 100181.

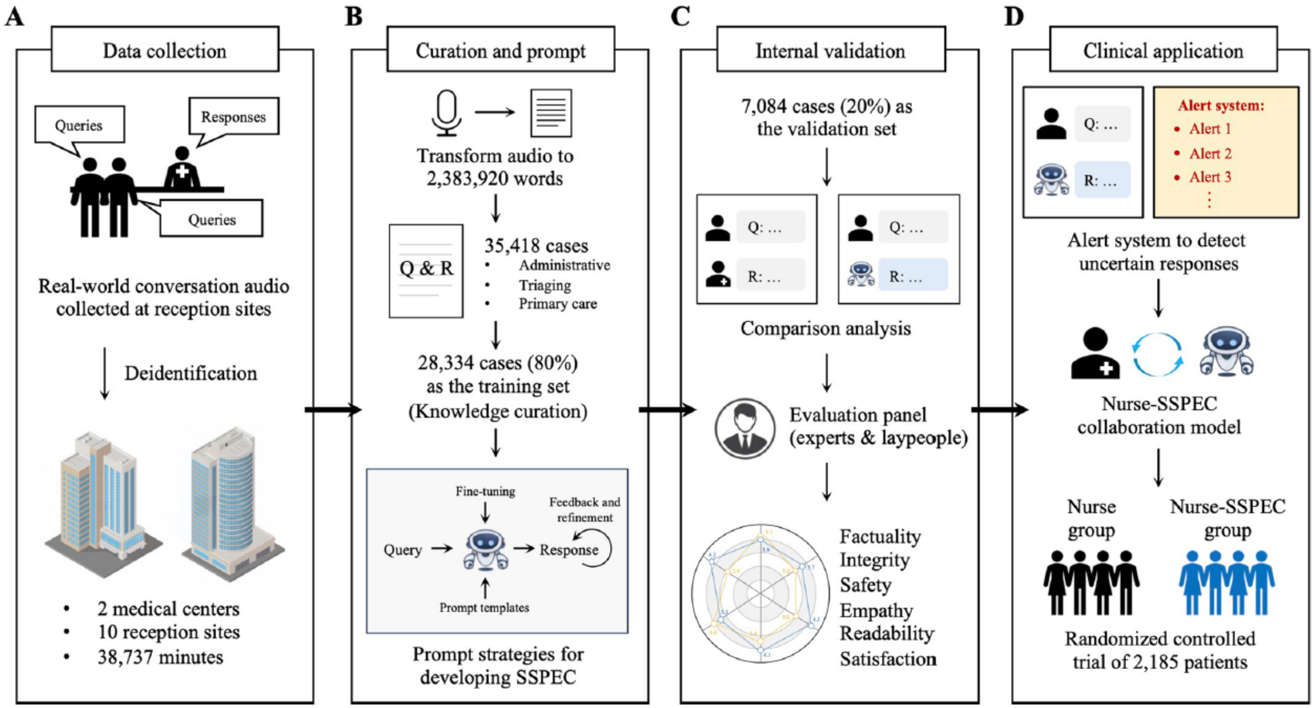
- Background: The release of ChatGPT for general use in 2023 by OpenAI has significantly expanded the possible applications of generative artificial intelligence in the healthcare sector, particularly in terms of information retrieval by patients, medical and nursing students, and healthcare personnel.
- **Objective: To compare the performance of ChatGPT-3.5 and ChatGPT-4.0 to clinical nurses on answering questions about tracheostomy care, as well as to determine whether using different prompts to pre-define the scope of the ChatGPT affects the accuracy of their responses.**
- Design: Cross-sectional study.
- Setting: The data collected from the ChatGPT was collected using the ChatGPT-3.5 and 4.0 using access provided by the University of Hong Kong. The data from the clinical nurses working in mainland China was collected using the Qualtrics survey program.
- Participants: No participants were needed for collecting the ChatGPT responses. A total of 272 clinical nurses, with 98.5 % of them working in tertiary care hospitals in mainland China, were recruited using a snowball sampling approach.
- Method: **We used 43 tracheostomy care-related questions in a multiple-choice format to evaluate the performance of ChatGPT-3.5, ChatGPT-4.0,** and clinical nurses. ChatGPT-3.5 and GPT-4.0 were both queried three times with the same questions by different prompts: no prompt, patient-friendly prompt, and act-as-nurse prompt. All responses were independently graded by two qualified otorhinolaryngology nurses on a 3-point accuracy scale (correct, partially correct, and incorrect). The Chi-squared test and Fisher exact test with post-hoc Bonferroni adjustment were used to assess the differences in performance between the three groups, as well as the differences in accuracy between different prompts.
- Results: **ChatGPT-4.0 showed significantly higher accuracy, with 64.3 % of responses rated as 'correct', compared to 60.5 % in ChatGPT-3.5 and 36.7 % in clinical nurses ($X^2 = 74.192$, $p < .001$).** Except for the 'care for the tracheostomy stoma and surrounding skin' domain ($X^2 = 6.227$, $p = .156$), scores from ChatGPT-3.5 and -4.0 were significantly better than nurses' on domains related to airway humidification, cuff management, tracheostomy tube care, suction techniques, and management of complications. Overall, ChatGPT-4.0 consistently performed well in all domains, achieving over 50 % accuracy in each domain. Alterations to the prompt had no impact on the performance of ChatGPT-3.5 or -4.0.
- Conclusion: ChatGPT may serve as a complementary medical information tool for patients and physicians to improve knowledge in tracheostomy care.

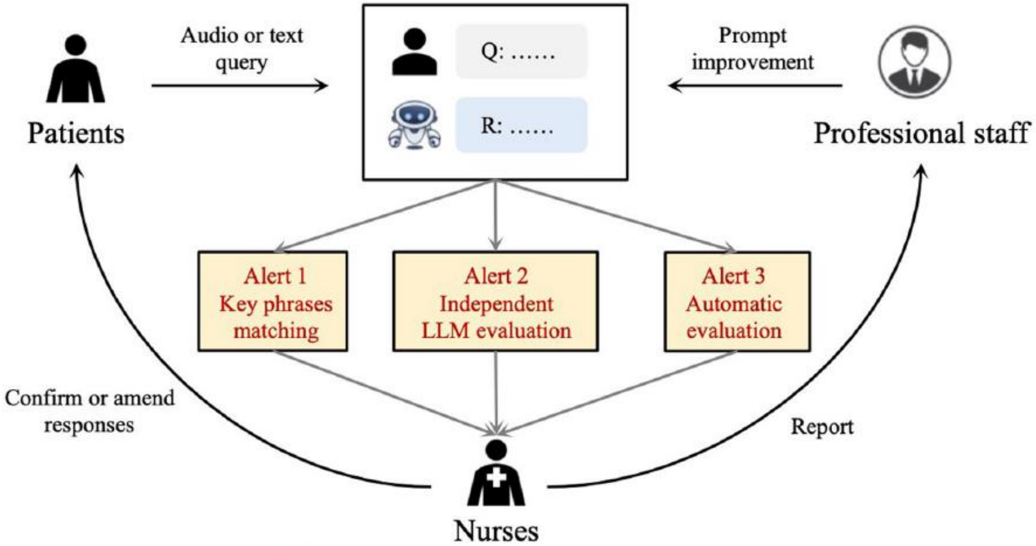
Description of the Tracheostomy Care Practice Measurement (TCP-43).

Domains	Number of Question	Description of the Question
Airway humidification – 8 items	1	Indicators for humidification
	5	Scenario-based questions on how to choose the nebulizing humidification solution
	1	When to evaluate the effectiveness of the humidification
	1	Indicators for effective humidification
Cuff management – 8 items	1	Assessment tool for cuff management
	1	Frequency of cuff assessment
	1	Range of cuff pressure
	1	Frequency of deflating the cuff
	1	Length of the cuff deflated period
	2	Scenario-based questions on whether to suction before deflating the cuff
	1	Selection of cuff or cuffless tracheostomy tube for patients with an established airway
Management of tracheostomy tube – 7 items,	1	Frequency of tracheostomy tube patency assessment
	1	How to clean a plastic inner cannula
	1	How to clean a metal inner cannula
	1	Frequency of assessing the stabilization of the tracheostomy tube
	1	How to stabilize a tracheostomy tube
	1	Preparation before, during and after tracheostomy tube change
	1	Preparation before and after tracheostomy tube removal
Care for the tracheostomy stoma and surrounding skin – 4 items,	1	Frequency of stoma and surrounding skin assessment
	1	Indicators for stoma and surrounding skin
	1	How to clean the stoma and surrounding skin
	1	Nursing interventions for maintaining the stoma and surrounding skin
Suction technique – 9 items, and	1	Whether to assess patients' need for suctioning
	1	When should the patients be suctioned

Wan, P., Huang, Z., Tang, W., Nie, Y., Pei, D., Deng, S., ... & Long, E. (2024). Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. Nature Medicine, 30(10), 2878-2885.

- Reception is an essential process for patients seeking medical care and a critical component influencing the healthcare experience. However, current communication systems rely mainly on human efforts, which are both labor and knowledge intensive. A promising alternative is to leverage the capabilities of large language models (LLMs) to assist the communication in medical center reception sites. Here we curated a unique dataset comprising 35,418 cases of real-world conversation audio corpus between outpatients and receptionist nurses from 10 reception sites across two medical centers, to develop a site-specific prompt engineering chatbot (SSPEC**). The SSPEC efficiently resolved patient queries, with a higher proportion of queries addressed in fewer rounds of queries and responses (Q&Rs; 68.0% ≤2 rounds) compared with nurse-led sessions (50.5% ≤2 rounds) (P = 0.009)** across administrative, triaging and primary care concerns. We then established a nurse-SSPEC collaboration model, overseeing the uncertainties encountered during the real-world deployment. In a single-center randomized controlled trial involving 2,164 participants, the primary endpoint indicated that **the nurse-SSPEC collaboration model received higher satisfaction feedback from patients (3.91 ± 0.90 versus 3.39 ± 1.15 in the nurse group, P < 0.001)**. Key secondary outcomes indicated reduced rate of repeated Q&R (3.2% versus 14.4% in the nurse group, P < 0.001) and reduced negative emotions during visits (2.4% versus 7.8% in the nurse group, P < 0.001) and enhanced response quality in terms of integrity (4.37 ± 0.95 versus 3.42 ± 1.22 in the nurse group, P < 0.001), empathy (4.14 ± 0.98 versus 3.27 ± 1.22 in the nurse group, P < 0.001) and readability (3.86 ± 0.95 versus 3.71 ± 1.07 in the nurse group, P = 0.006). Overall, our study supports the feasibility of integrating LLMs into the daily hospital workflow and introduces a paradigm for improving communication that benefits both patients and nurses.

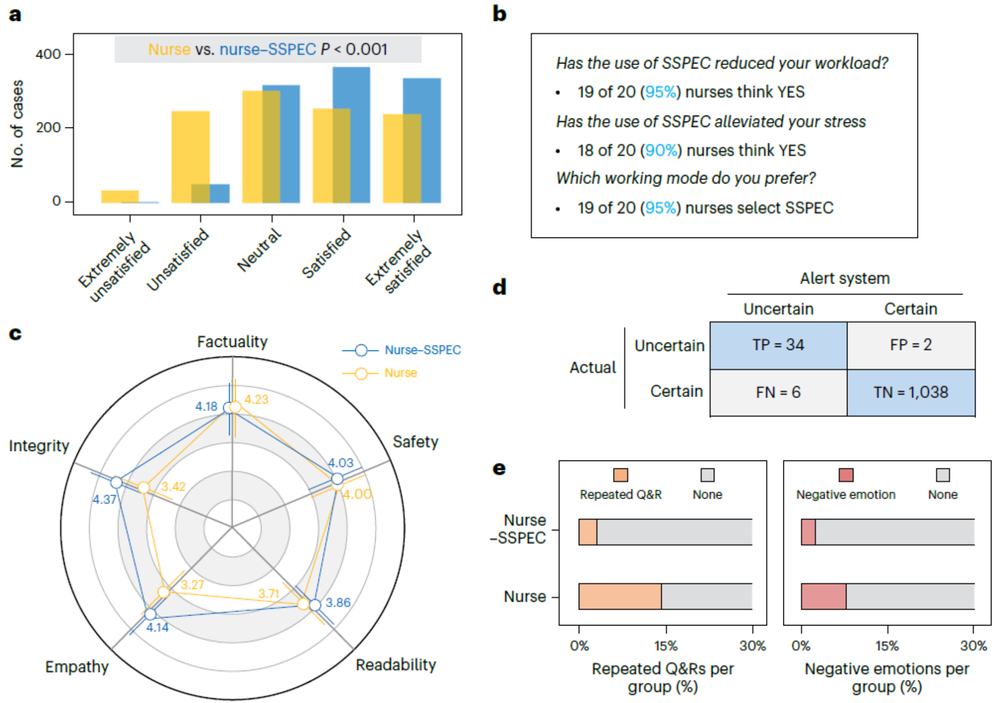

A
Data collection
Queries
Responses
Queries
Real-world conversation audio collected at reception sites
Deidentification
• 2 medical centers
• 10 reception sites
• 38,737 minutes
B
Curation and prompt
Transform audio to 2,383,920 words
Q & R
35,418 cases
• Administrative
• Triaging
• Primary care
28,334 cases (80%) as the training set (Knowledge curation)
Fine-tuning
Feedback and refinement
Query
Response
Prompt templates
Prompt strategies for developing SSPEC
C
Internal validation
7,084 cases (20%) as the validation set
Q: ...
R: ...
Q: ...
R: ...
Comparison analysis
Evaluation panel (experts & laypeople)
Factuality
Integrity
Safety
Empathy
Readability
Satisfaction
D
Clinical application
Q: ...
R: ...
Alert system:
• Alert 1
• Alert 2
• Alert 3
Alert system to detect uncertain responses
Nurse-SSPEC collaboration model
Nurse group
Nurse-SSPEC group
Randomized controlled trial of 2,185 patients



Extended Data Fig. 3 | Nurse-SSPEC collaboration model involving the alert system in mitigating uncertainty. Upon patient arrival at the reception site, their queries are recorded audibly and automatically transformed into text. To address uncertain or potentially harmful responses generated by SSPEC, an alert system has been implemented. This system triggers an alert to the nurses if any 'signals of uncertainty' are detected, through key-phrases matching, independent LLM evaluation, or automatic evaluation. This alert prompts immediate nurses review or modification of the response. Furthermore, a dedicated team reviews all patient-SSPEC conversations to continually refine the prompting.

Patient: Why hasn't me been called? It's been half an hour.
Nurse: Your name was called for several times.
Patient: Was it called? It didn't show up on the screen.
Nurse: It was, and you missed your turn. When your name is called, you should go in. Why are you still sitting here? I heard it being called.
Patient: When was it called?
Nurse: At 10:49, I heard it.
Patient: Isn't it 10:49 now?
Nurse: Yes, it was just called.

Patient: Hello, how do I get a certificate for a sleep disorder?
Nurse: Sleep disorder? For such a young child? What kind of certificate can this hospital provide? We can issue a medical record for you.


a
Nurse vs. nurse–SSPEC P < 0.001
No. of cases
400
200
0
Extremely unsatisfied
Unsatisfied
Neutral
Satisfied
Extremely satisfied
b
Has the use of SSPEC reduced your workload?
19 of 20 (95%) nurses think YES
Has the use of SSPEC alleviated your stress
18 of 20 (90%) nurses think YES
Which working mode do you prefer?
19 of 20 (95%) nurses select SSPEC
c
Nurse–SSPEC
Nurse
Factuality
Safety
Readability
Empathy
Integrity
4.18
4.23
4.03
4.00
3.86
3.71
4.14
3.27
4.37
3.42
d
Alert system
Uncertain
Certain
Actual
Uncertain
TP = 34
FP = 2
Certain
FN = 6
TN = 1,038
e
Repeated Q&R
None
Negative emotion
None
Nurse –SSPEC
Nurse
0%
15%
30%
Repeated Q&Rs per group (%)
Negative emotions per group (%)

Zolnoori, M., Vergez, S., Xu, Z., Esmaeili, E., Zolnour, A., Anne Briggs, K., ... & McDonald, M. V. (2024). Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare. JAMIA open, 7(4), ooae130.

- **Objectives:** As artificial intelligence evolves, integrating speech processing into home healthcare (HHC) workflows is increasingly feasible. Audio-recorded communications enhance risk identification models, with automatic speech recognition (ASR) systems as a key component. This study evaluates the transcription accuracy and equity of 4 ASR systems—Amazon Web Services (AWS) General, AWS Medical, Whisper, and Wave2Vec—in transcribing patient-nurse communication in US HHC, **focusing on their ability in accurate transcription of speech from Black and White English-speaking patients.**

- **Materials and Methods:** **We analyzed audio recordings of patient-nurse encounters from 35 patients (16 Black and 19 White)** in a New York City-based HHC service. Overall, 860 utterances were available for study, including 475 drawn from Black patients and 385 from White patients. Automatic speech recognition performance was measured using word error rate (WER), benchmarked against a manual gold standard. Disparities were assessed by comparing ASR performance across racial groups using the linguistic inquiry and word count (LIWC) tool, focusing on 10 linguistic dimensions, as well as specific speech elements including repetition, filler words, and proper nouns (medical and nonmedical terms).

- **Results:** The average age of participants was 67.8 years (SD = 14.4). Communication lasted an average of 15 minutes (range: 11-21 minutes) with a median of 1186 words per patient. Of 860 total utterances, 475 were from Black patients and 385 from White patients. Amazon Web Services General had the highest accuracy, with a median WER of 39%. However**, all systems showed reduced accuracy for Black patients,** with significant discrepancies in LIWC dimensions such as “Affect,” “Social,” and “Drives.” Amazon Web Services Medical performed best for medical terms, though all systems have difficulties with filler words, repetition, and nonmedical terms, with AWS General showing the lowest error rates at 65%, 64%, and 53%, respectively.

- **Discussion:** While AWS systems demonstrated superior accuracy, significant disparities by race highlight the **need for more diverse training datasets and improved dialect sensitivity**. Addressing these disparities is critical for ensuring equitable ASR performance in HHC settings and enhancing risk prediction models through audio-recorded communication.

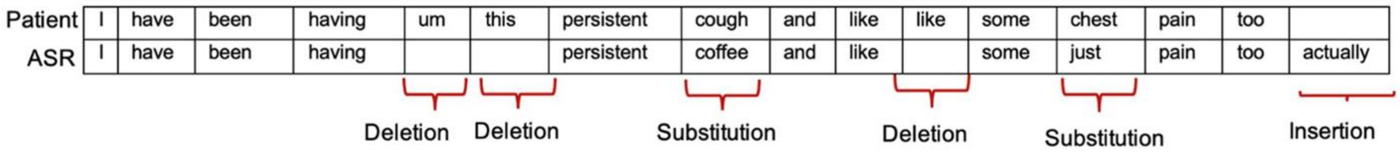
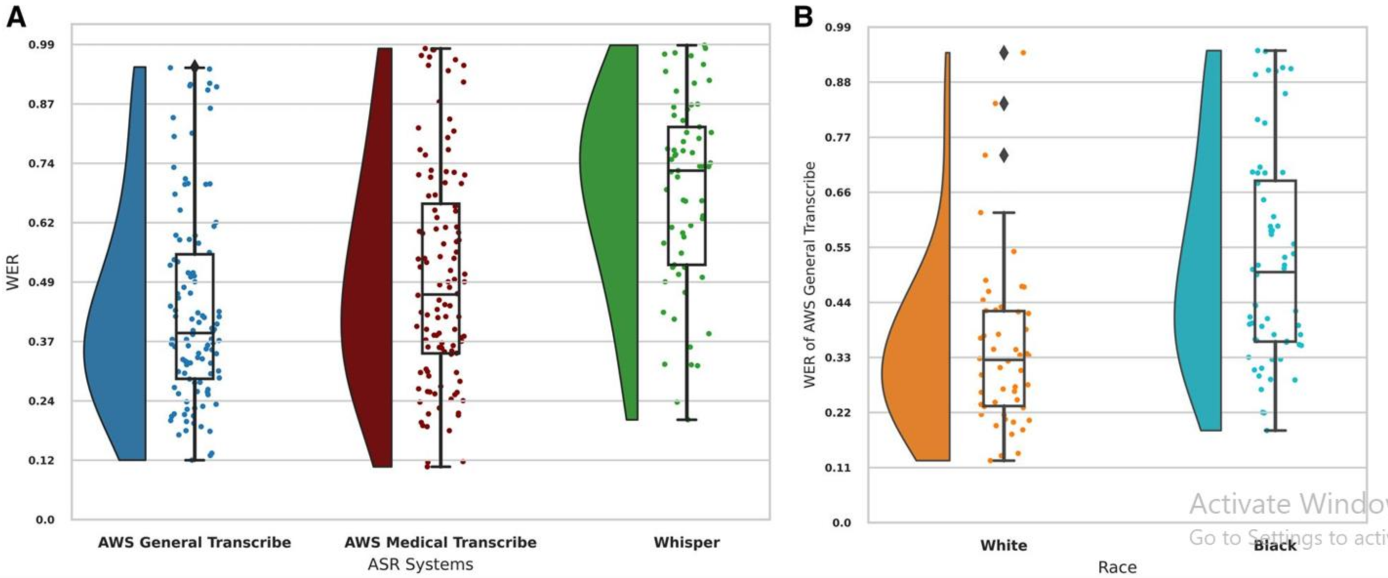


Figure 1. Visualization of WER components: substitutions, insertions, and deletions. Abbreviation: WER, word error rate.



Loughran, E., Kane, M., Wyatt, T. H., Kerley, A., Lowe, S., & Li, X. (2024). Using large language models to address health literacy in mHealth: case report. CIN: Computers, Informatics, Nursing, 10-1097.

- **Abstract**: The innate complexity of medical topics often makes it challenging to produce educational content for the public. Although there are resources available to help authors appraise the complexity of their content, there are woefully few resources available to help authors reduce that complexity after it occurs. **In this case study, we evaluate using ChatGPT to reduce the complex language used in health-related educational materials.** ChatGPT adapted content from the SmartSHOTS mobile application, which is geared toward caregivers of children aged 0 to 24 months. SmartSHOTS helps reduce barriers and improve adherence to vaccination schedules. ChatGPT reduced complex sentence structure and rewrote content to align with a third-grade reading level. Furthermore**, using ChatGPT to edit content already written removes the potential for unnoticed, artificial intelligence–produced inaccuracies**. As an editorial tool, ChatGPT was effective, efficient, and free to use. This article discusses the potential of ChatGPT as an effective, time-efficient, and open-source method for editing health-related educational materials to reflect a comprehendible reading level.

ChatGPT Prompt and Response

Prompt

Rewrite this passage—“A vaccine is a substance that is used to stimulate the body's immune system to protect it from disease. It can be given in three methods: by mouth, like the Rotavirus vaccine, through an intramuscular injection, like the flu shot, and intranasally, like the FluMist vaccine that is sprayed into a patient's nose. Vaccines are also called vaccinations, immunizations, or shots”—while including all core information but making it understandable to a 3rd-grade reading level

Response

A vaccine is a special thing that helps our body stay strong and not get sick. There are three ways to get a vaccine: by swallowing it, like the Rotavirus vaccine, by getting a shot in your muscles, like the flu shot, or by having it sprayed into your nose, like the FluMist vaccine. Some people also call vaccines by other names, like vaccinations, immunizations, or shots. So, a vaccine is like a **superhero** that helps our body fight off sickness!

make it understandable to adults with a 3rd-grade reading ability

Wagner, L., Jourdan, S., Mayer, L., Müller, C., Bernhard, L., Kolb, S., ... & Wilhelm, D. (2024). Robotic scrub nurse to anticipate surgical instruments based on real-time laparoscopic video analysis. Communications Medicine, 4(1), 156.

- Background: Machine learning and robotics technologies are increasingly being used in the healthcare domain to improve the quality and efficiency of surgeries and to address challenges such as staff shortages. Robotic scrub nurses in particular offer great **potential to address staff shortages** by assuming nursing tasks such as the handover of surgical instruments.
- Methods: We introduce a robotic scrub nurse system designed to enhance the quality of surgeries and efficiency of surgical workflows by predicting and delivering the required surgical instruments based on real-time laparoscopic video analysis. We propose a three-stage deep learning architecture consisting of a single frame-, temporal multi frame-, and informed model to anticipate surgical instruments. **The anticipation model was trained on a total of 62 laparoscopic cholecystectomies.**
- Results: Here, we show that our prediction system **can accurately anticipate 71.54% of the surgical instruments required during laparoscopic cholecystectomies in advance**, facilitating a smoother surgical workflow and reducing the need for verbal communication. As the instruments in the left working trocar are changed less frequently and according to a standardized procedure, the prediction system works particularly well for this trocar.
- Conclusions: The robotic scrub nurse thus acts as a mind reader and helps to mitigate staff shortages by taking over a great share of the workload during surgeries while additionally enabling an enhanced process standardization.

Fig. 2 | Deployment of the developed RSN in a simulated surgery at the University Hospital rechts der Isar of the Technical University of Munich. The RSN predicts an instrument change and passes the next required instrument to the surgeon without verbal communication. This allows the surgeon to focus on the right endoscopic screen while real-time AI assistance is provided on the left endoscopic screen.

Table 1 | Evaluation metrics for the two different temporal model types: MS-TCN and LTContext

Model	wAP	wAR	wAA	wAF1
MS-TCN	74.51 ± 1.86	69.59 ± 1.45	87.81 ± 0.79	70.29 ± 1.65
LTContext	77.15 ± 1.93	70.31 ± 0.96	88.32 ± 0.64	71.54 ± 1.07

The evaluation metrics are computed class-wise, for each instrument and working trocar individually, and then averaged over all surgeries in the test set, yielding the weighted-averaged precision (wAP), recall (wAR), accuracy (wAA), and F1 score (wAF1). The weighted-averaged metrics are calculated by taking the mean of all per-class scores while considering the number of actual occurrences of each class in the test data set. The averaged metrics over tenfold are reported (%) with the corresponding standard deviation (±).

Step 6 – Consultation

Feedback and Professional Input…