Journal of Infection and Public Health 18 (2025) 102627

Contents lists available at ScienceDirect

Journal of Infection and Public Health

journal homepage: www.elsevier.com/locate/jiph

Original article

Evaluation of ChatGPT-4 for the detection of surgical site infections from electronic health records after colorectal surgery: A pilot diagnostic accuracy study

Josep M. Badia [a,b,*,1], Daniel Casanova-Portoles [a,b,2], Estela Membrilla [c,3], Carles Rubiés [d], Miquel Pujol [e,f,g,4], Joan Sancho [c,5]

[a] Department of Surgery, Hospital General de Granollers, Granollers, Spain
[b] Universitat Internacional de Catalunya. Sant Cugat del Vallès, Barcelona, Spain
[c] Department of Surgery, Hospital del Mar, Barcelona, Spain
[d] Department of Digital Transformation, Hospital General de Granollers, Granollers, Spain
[e] VINCat Program, Servei Català de la Salut, Catalonia, Spain
[f] Centro de Investigación Biomédica en Red de Enfermedades Infecciosas (CIBERINFEC), Instituto de Salud Carlos III, Madrid, Spain. VINCat Program, Barcelona, Catalonia, Spain
[g] Department of Infectious Diseases, Hospital Universitari de Bellvitge - IDIBELL. L'Hospitalet de Llobregat, Spain

ARTICLE INFO

Article history:
Received 21 October 2024
Received in revised form 29 November 2024
Accepted 16 December 2024

Keywords:
Surgical site infection
Diagnosis
Accuracy
Sensitivity and specificity
Artificial intelligence
ChatGPT
Natural language processing
NLP
Large Language Model
LLM
OpenAI

ABSTRACT

Background: Surveillance of surgical site infection (SSI) relies on manual methods that are time-consuming and prone to subjectivity. This study evaluates the diagnostic accuracy of ChatGPT for detecting SSI from electronic health records after colorectal surgery via comparison with the results of a nationwide surveillance programme.
Methods: This pilot, retrospective, multicentre analysis included 122 patients who underwent colorectal surgery. Patient records were reviewed by both manual surveillance and ChatGPT, which was tasked with identifying SSI and categorizing them as superficial, deep, or organ-space infections. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated. Receiver operating characteristic (ROC) curve analysis determined the model's diagnostic performance.
Results: ChatGPT achieved a sensitivity of 100 %, correctly identifying all SSIs detected by manual methods. The specificity was 54 %, indicating the presence of false positives. The PPV was 67 %, and the NPV was 100 %. The area under the ROC curve was 0.77, indicating good overall accuracy for distinguishing between SSI and non-SSI cases. Minor differences in outcomes were observed between colon and rectal surgeries, as well as between the hospitals participating in the study.
Conclusions: ChatGPT shows high sensitivity and good overall accuracy for detecting SSI. It appears to be a useful tool for initial screenings and for reducing manual review workload. The moderate specificity suggests a need for further refinement to reduce the rate of false positives. The integration of ChatGPT alongside electronic medical records, antibiotic consumption and imaging data results for real-time analysis may further improve the surveillance of SSI. ClinicalTrials.gov Identifier: NCT06556017.

1. Introduction

Epidemiological surveillance of healthcare-associated infections (HAIs) is one of the eight core components of the World Health Organization (WHO)'s Infection Prevention and Control Programmes [1]. These programmes, which include surveillance for surgical site infection (SSI), have proven to be effective in all types of surgery and in a variety of settings [2,3].

The data recorded on a systematic basis by these SSI surveillance programmes are then used to evaluate the quality of care provided,

* Correspondence to: Department of Surgery, Hospital General de Granollers, Av Francesc Ribas 1, Granollers, Barcelona 08402, Spain.
E-mail addresses: jmbadiaperez@gmail.com (J.M. Badia), dcasanova@fphag.org (D. Casanova-Portoles), estelabe@gmail.com (E. Membrilla), crubies@fphag.org (C. Rubiés), mpujol@bellvitgehospital.cat (M. Pujol), jsancho@parcdesalutmar.cat (J. Sancho).
[1] https://orcid.org/0000-0003-2928-5233
[2] https://orcid.org/0000-0001-9837-7444
[3] https://orcid.org/0000-0002-1392-040X
[4] https://orcid.org/0000-0002-6475-6208
[5] https://orcid.org/0000-0003-4603-0363

https://doi.org/10.1016/j.jiph.2024.102627

assess the impact of SSI prevention measures, and facilitate benchmarking across hospitals or health systems. Currently, SSI surveillance is a labour-intensive and extremely costly task that is carried out manually by the infection control teams (ICT), with the result that its application is not universal but is limited to high-risk interventions [4–6]. An Australian study showed that the percentage of time spent on HAI surveillance activities by ICT members was 36.0 % of their contract [7]. Conceivably, a transition to automated surveillance leveraging the possibilities offered by algorithms and artificial intelligence (AI), based on big data analysis, natural language processing (NLP) and machine learning, might introduce notable improvements [8,9]. In recent years, it has been reported that semi-automated screening models based on algorithms can detect deep and organ/space SSI efficiently, without the need for manual screening of each individual case, although their ability to detect superficial SSI is limited [9–11]. A Dutch algorithm achieved a 63.4 % reduction in the number of records needing a full manual review [10].

The wider application of AI in healthcare, in the form of NLP models such as ChatGPT (OpenIA), has been investigated in a number of contexts such as the analysis of electronic health records (EHR), the provision of assistance in clinical or radiological decision-making, and the delivery of patient education [12–18]. To the best of our knowledge, the application of ChatGPT for the surveillance of SSI has not been described in peer-reviewed medical literature.

It is hypothesised that SSI surveillance supported by NLP systems could lead to a reduction in the number of medical records requiring full manual review and to a considerable reduction in the workload of ICTs.

The objective of this pilot study is to evaluate the performance of ChatGPT in assessing overall SSI and their classification into three anatomical levels using EHR data from a cohort of elective colorectal surgery patients previously screened by a national healthcare-associated infection surveillance system.

2. Methods

2.1. Study

Retrospective, multicentre study comparing the diagnostic efficacy of ChatGPT with manual surveillance for the detection of SSI in patients undergoing elective colorectal surgery.

2.2. Setting, data sources and definitions

Patient records came from the prospective database of the surveillance system for healthcare-associated infections in Catalonia (VINCat) at two hospitals. The structure and results of the programme have been described in detail elsewhere [19]. In brief, since 2008, VINCat has been a well-established, nationwide, audited program monitoring HAIs in 71 public and private hospitals across Catalonia, Spain. For elective colorectal surgery surveillance, ICTs in each hospital have conducted prospective surveillance using a standardized manual methodology to ensure comprehensive data collection. This includes a mandatory minimum follow-up of 30 days post-surgery, electronic reviews of medical records to identify readmissions, emergency department visits, or visits to other healthcare facilities, as well as the collection of microbiological and radiological data generated during this period.

The programme followed cases of elective Class 2 and 3 wounds and used the definitions of the Centers for Disease Control and Prevention-National Health Safety Network (CDC-NHSN)[20,21]. Accordingly, SSI was defined as any infection arising at the surgical site within 30 days after surgery and were categorised as superficial incisional (S-SSI), deep incisional (D-SSI) and organ-space (O/S-SSI). The EHR was defined as the comprehensive set of patient data, including demographics, medical history, diagnoses, medications, imaging studies, laboratory test results, treatment plans, and clinical notes. Conversely, narrative clinical notes written by health professionals were referred to as Narrative Clinical notes within the EHR (EHR-NC).

During the period from which the patients were selected, the overall SSI rate in VINCat hospitals was 7.0 % for colon surgery and 12.0 % for rectal surgery [22]. Since this is a pilot study focused on evaluating the performance of ChatGPT with a limited number of cases, it was essential to ensure a balanced representation of both colon and rectal surgery patients due to their differing risk profiles and infection rates. Additionally, an equal number of cases with and without known SSI were included to comprehensively assess both scenarios. The sample size was therefore calculated assuming a SSI prevalence of 50 %. To further evaluate the generalizability of the approach, the study was conducted across two hospitals with different levels of complexity.

Considering a confidence level of 95 % with a margin of error of 5 % and an expected level of agreement of 90 %, the sample size for this pilot study was set at 100 patients, divided equally between colon surgery and rectal surgery, and between the two hospitals. Patients were enrolled consecutively, starting retroactively from December 2023; fifty per cent had known SSI after standardized manual assessment and 50 % did not.

The GPT-4-turbo model by OpenAI was accessed via the ChatGPT Plus subscription in August 2024 to generate automated responses in this study [23]. GPT-4-turbo is an optimized variant of OpenAI's GPT-4 language model, designed to be faster, and more cost-effective in terms of computational resources, while maintaining high quality among the responses generated.

2.3. Intervention

ChatGPT was used as an aid to define the research prompt for detecting SSI cases. The process of developing the final prompt is presented in the supplementary material.

The SSI assessment of the procedures included in the VINCat surveillance was taken as the gold standard for comparison with the ChatGPT results. The chatbot was asked whether the answers should be defined as 'suspected' or 'certain' SSI. It was also asked to provide information regarding the anatomical level of the infection, namely whether it should be classified as S-SSI, D-SSI or O/E-SSI. For the purposes of statistical analysis, the responses classified as 'suspicion' and 'certainty' were grouped into a single category. The accuracy of ChatGPT for detecting SSI cases both overall and at each of its anatomical levels was compared with the already known results obtained with the VINCat surveillance system.

2.4. Measurement

For the SSI diagnosis of the AI chatbot, only the unmodified EHR-NC of the selected patients written by the healthcare professionals were taken. These were compared with the results of the VINCat surveillance, which used the full content of the EHR, including post-discharge secondary care data. All notes for 60 days after surgery were collected in an attempt to capture any comments that might give an idea of a suspected SSI event after discharge. Texts written by the medical, nursing, physiotherapy and psychology teams were included. The texts in the EHR-NCs were written in a mixture of Catalan and Spanish, the languages used in the hospitals in the study setting, which are both recognized by ChatGPT. Laboratory, microbiology or imaging test results and drug prescriptions that were not transcribed in the EHR were not used. The selected text was transferred to a separate Microsoft Word document for each patient in order to anonymize the professionals' comments.

To evaluate each case in ChatGPT, the prompt was entered into the system first, followed by the text of the previously selected EHR-NC. As the responses generated by ChatGPT are dependent on the context of the ongoing discourse, to prevent the model from being influenced by responses from other clinical cases a new ChatGPT session was initiated for each patient included. To account for variation in responses, each patient was tested on two occasions, with a different user administering the test in each instance. The prompts remained consistent across users.

The average time taken by one of the investigators for each of the assessments was calculated. For this purpose, the time taken for the first 10 cases was discarded and the time taken for the whole process, from entering the EHR-NC to obtaining and reporting the ChatGPT response in a Word document for each patient, was recorded.

2.5. Ethical issues

The anonymity and confidentiality of patient data and of the healthcare professionals who recorded the EHR-NC were maintained throughout the research process, including access to records, data coding and archiving of information. Prior to the input of any information into the chatbot, the data of each patient were anonymized and all personally identifiable information was removed in accordance with data privacy regulations. Confidential patient information was protected in accordance with European standards and was approved by the Research Ethics Committee of the Hospital General de Granollers (code no. 82–2024). The project was registered with the ClinicalTrials.gov Identifier: NCT06556017, and is reported in accordance with the STARD2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies [24].

2.6. Statistical methods

The SSI assessment conducted by the VINCat surveillance system was used as the gold standard for comparison with the results generated by ChatGPT. The accuracy of ChatGPT in detecting SSI cases, including its performance at various anatomical levels, was compared to the established outcomes obtained with the VINCat system.

Specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV), and overall predictive value were calculated for the ChatGPT assessment using two-by-two contingency tables. Receiver Operating Characteristic (ROC) curve analysis was performed to evaluate ChatGPT's diagnostic performance, with the area under the ROC curve (AUC) calculated to quantify its overall accuracy. Sensitivity and specificity were assessed across various thresholds, with the optimal cut-off point determined using Youden's Index.

Additionally, inter-method agreement for SSI diagnosis was evaluated using Cohen's Kappa coefficient, with a Kappa value greater than 0.75 considered indicative of excellent agreement. Differences between the two methods' assessments were analysed using McNemar's test for paired nominal data.

A p-value of < 0.05 was considered statistically significant. All data analyses were conducted using SPSS Statistics version 28.0 (IBM Corp., Armonk, NY, USA).

3. Results

The study enrolled 122 patients undergoing colorectal surgery who had previously undergone scrutinization under the surveillance system. Patients' demographics are shown in Table 1. The concordance between the results obtained by the two evaluators for each case was 100 % at both hospitals. The average time spent per patient assessment was six minutes.

Table 1
Demographics of the patients included in the study.

	Overall	Hospital 1 (n = 56)	Hospital 2 (n = 66)	P
Age, years (SD)	70.7 (12)	72.6 (9)	69.1 (15)	0.132
Sex, M/F	85/37	42/14	43/23	0.323
Colon surgery	63 (52)	28 (44)	35 (56)	0.856
Rectal surgery	59 (48)	28 (46)	31 (53)	

SD: standard deviation. Values are n (%) unless otherwise indicated.

3.1. General analysis

The ChatGPT assessment achieved a sensitivity of 100 % (95 % c.i. 93.9 %, 100 %), and a specificity of 53.9 % (95 % c.i. 41.8 %, 65.7 %). The positive and negative predictive values were 67.1 % (95 % c.i. 56.7 %, 76 %), and 100 % (95 % c.i. 89.8 %, 100 %), respectively. The ROC curve for the ChatGPT system showed an AUC of 0.77 (95 % c.i. 0.69, 0.85). This value suggests that ChatGPT has good overall accuracy in distinguishing between cases with and without SSIs (Table 2).

The diagnostic concordance analysis revealed a Cohen's Kappa of 0.501.

3.2. Analysis by type of surgery

ChatGPT showed a sensitivity of 100 % (95 % c.i. 89.3 %, 100 %), and a specificity of 65 % (95 % c.i. 46.9 %, 78.9 %), for colon surgery. The PPV was 74 % (95 % c.i. 59.8 %, 85.1 %), while the NPV remained at 100 % (95 % c.i. 83.9 %, 100 %). In rectal surgery, ChatGPT also demonstrated a high sensitivity of 100 % (95 % c.i. 87.5 %, 100 %), but the specificity was lower, at only 44 % (95 % c.i. 28.2 %, 60.7 %). The PPV was 60 % (95 % c.i. 45.5 %, 73.0 %), while the NPV was 100 % (95 % c.i. 78.5 %, 100 %).

3.3. Analysis by centre

In the analysis of Hospital 1 records, ChatGPT showed a sensitivity of 100 % (95 % c.i. 87.9 %, 100 %), with a specificity of 82.1 % (95 % c.i. 64.4 %, 92.1 %). The PPV was 84.9 % (95 % c.i. 69.1 %, 93.3 %), and the NPV was 100 % (95 % c.i. 85.7 %, 100 %). These results highlight that ChatGPT performs significantly well at this centre in terms of specificity and the ability to avoid false positives.

At Hospital 2, ChatGPT also showed a sensitivity of 100 % (95 % c.i. 89.0 %, 100 %), but had a lower a specificity, at 31.4 % (95 % c.i. 18.6 %, 48 %). The PPV was 56.4 % (95 % c.i. 43.3 %, 68.6 %), and the NPV remained at 100 % (95 % c.i. 74.1 %, 100 %). These results indicate that, while ChatGPT was highly effective in detecting all true infections at this centre, the high rate of false positives significantly limits its specificity.

3.4. Analysis by level of infection

Table 3 shows the comparison of VINCat and ChatGPT diagnoses both overall and according to SSI level. All seven S-SSIs diagnosed by the manual VINCat review were correctly detected by ChatGPT.

4. Discussion

This pilot study presents first-of-its-kind evidence of the potential of advanced AI tools, specifically ChatGPT, for identifying SSI after elective colorectal surgery based on patients' EHR-NCs. The findings indicate that ChatGPT performs well in detecting SSI, exhibiting high sensitivity, acceptable specificity, and an AUC of 0.77. These results define it a promising tool for semi-automated surveillance systems, in which high sensitivity is of paramount importance in order to guarantee that patients requiring manual review are accurately identified.

Table 2
Sensitivity, specificity, accuracy and positive and negative predictive values of the ChatGPT assessment.

		Sensitivity	Specificity	PPV	NPV	Accuracy
Overall		100	53.9	67.1	100	76.23
Centre	Hospital 1	100	82.1	84.9	100	91.1
	Hospital 2	100	31.4	56.4	100	63.6
Type of Surgery	Colon	100	64.5	74.4	100	82.5
	Rectum	100	43.8	60	100	69.5

PPV: positive predictive value; NPV: negative predictive value.

Table 3
Diagnostic outcomes for surgical site infections (SSI) as determined by ChatGPT in comparison with the VINCAT diagnosis. The table categorizes patients into four groups based on the type of SSI identified: No SSI, Superficial Incisional SSI (S-SSI), Deep Incisional SSI (D-SSI), and Organ/Space SSI (O/S-SSI). Each cell indicates the number of cases where ChatGPT's diagnosis aligns or diverges from the VINCAT reference diagnosis.

		VINCAT diagnosis				Total
		No SSI	S-SSI	D-SS	O/S-SSI	
ChatGPT diagnosis	No SSI	34	0	0	0	34
	S-SSI	7	3	1	3	14
	D-SSI	2	3	3	2	10
	O/S-SSI	20	1	0	43	64
Total		63	7	4	48	122

SSI: surgical site infection; S-SSI: superficial incisional site infection; D-SSI: deep incisional site infection; O/S-SSI: organ/space surgical site infection.

The results suggest that, although ChatGPT is effective in detecting all true infections in rectal surgeries, the proportion of false positives is higher than in colon surgeries. This lower specificity indicates that ChatGPT tends to over-predict infections in rectal surgeries, possibly due to differences in clinical presentation or data quality. In fact, the majority of suspected SSI detected by the chatbot were infections from another source, such as central line-associated bloodstream infection or urinary tract infections. The higher specificity at Hospital 1 suggests that, while ChatGPT is particularly well suited for detecting infection, the predictive ability of the model may be influenced by the quality or quantity of data entered into the EHR-NCs at each facility.

In 2019, a Dutch study investigated an algorithm for surveillance of deep SSIs after colorectal surgery (aggregating D-SSI and O/S-SSI) based on clinical variables. The final model included five variables: postoperative length of stay, wound class, readmission, reoperation and 30-day mortality, and achieved a specificity of 68.7 % and a sensitivity of 98.5 %, with an AUC of 0.950. The positive and negative predictive values were 21.5 % and 99.8 % respectively [10]. The same group validated the algorithmic methodology through a retrospective study conducted at three European hospitals, focusing also on D-SSI and O/S-SSI across different types of surgery. The sensitivity of the standardized algorithm ranged from 82 % to 100 % for orthopaedic surgery, from 67 % to 100 % for cardiac surgery, and from 84 % to 100 % for colon surgery. The implementation of the algorithm led to a 72 %-98 % reduction in the workload for ICT teams [9]. In another study, a Bayesian network coupled with NLP demonstrated high accuracy for detecting “clinically important” SSI after colorectal surgery, with an AUC of 0.827 [25].

Despite being based on a limited patient sample and employing a distinct methodology under semi-experimental conditions, the findings of the present study are comparable to those of algorithm-based models, demonstrating similar sensitivity and specificity along with a higher PPV. Nevertheless, the observed AUC indicates a need for further optimisation. This underscores the need to extend the study to include a larger cohort of patients and to explore the potential integration of additional data beyond that provided by the EHR to enhance model performance.

Furthermore, it appears that ChatGPT is able to identify infections that are less “clinically important” such as S-SSIs, which are likely to be missed by the above algorithms. S-SSI typically lack the features that those algorithms rely on, as they often do not require antibiotic treatment, do not prompt imaging studies, and do not cause significant changes in the clinical course; in contrast, they are often recorded in the EHR-NCs by nurses and surgeons, which the chatbot has effectively identified and catalogued.

In this analysis of ChatGPT's iterative query functionality, the main objective was to achieve the highest possible sensitivity for detecting a high proportion of patients with SSI. The consequence of this high sensitivity was a relatively high rate of false positives due to moderate specificity. This shortcoming was accepted at the onset of the project because the aim was to obtain a semi-automated model able to identify patients with a high probability of SSI whose records would then be reviewed manually. While the moderate false positive rate reduces the model's overall efficacy, it still performed better than a universal chart review of all operated patients. Due to its high sensitivity, this model may prove valuable for quality improvement and benchmarking purposes.

ChatGPT was chosen from among the machine learning systems currently available because of its wide diffusion and availability in our environment. Its version GPT-4o was selected for its significant improvements over ChatGPT-3.5, especially in handling complex language tasks and contextual analysis. According to its developers, GPT-4 demonstrates superior natural language understanding, with a more nuanced grasp of medical terminology and complex sentence structures. This enables it to more accurately interpret EHR-NCs and detect subtle signs of conditions like SSI. While ChatGPT-3.5 performed well with simpler tasks, GPT-4 shows enhanced contextual awareness, reduced error rates, and has greater responsiveness to prompts. This allows for more customized queries and increases its effectiveness in addressing specific medical tasks, such as identifying infection risks, a known limitation of GPT-3.5. In addition, a study revealed that GPT-4 surpassed both resident physicians and its predecessor, GPT-3.5, in diagnostic accuracy when compared to the discharge diagnosis gold standard in emergency medicine [12].

In the referenced study on time allocation for HAI surveillance [7], 56 % was dedicated to data collection, 27 % to monitoring compliance with infection control measures, and 17 % to communicating HAI data to clinicians and management. While the time investment in the present study was estimated by only one investigator, it may still indicate a significant reduction in the workload for ICTs. The effectiveness of ChatGPT surveillance, combined with its ease of implementation and the minimal time needed to analyse each case, suggests that it may allow the extension of SSI surveillance to currently unmonitored surgical procedures and achieve full surveillance of cases already being monitored. This may require specific adjustments based on clinical context and hospital environment in order to optimize diagnostic accuracy, incorporating additional clinical features, integrating the drug prescription programme and the results of imaging studies and microbiology, or improving data quality to increase the model's specificity.

It is important to highlight that VINCat's manual surveillance utilized all available information within the EHR, including data from the 'Catalan Shared Medical Record,' which encompasses details of all post-discharge interactions with the healthcare system. In contrast, the results generated by ChatGPT were based exclusively on

EHR-CN data, without the inclusion of supplementary in-hospital or post-discharge information sources.

As ChatGPT and other similar models continue to evolve, their performance in detecting SSI may improve, particularly with the integration of laboratory, microbiology, and imaging data. This could position these models as increasingly valuable tools in modern healthcare practices. In comparison to conventional techniques, integrating ChatGPT with EHR systems presents a number of significant benefits, such as the capacity for real-time availability and the potential to standardize the detection process, thereby reducing the likelihood of human error and variability in clinical assessments.

Furthermore, it is essential to ensure that ChatGPT-based surveillance models meet the requirements for safe and effective use in healthcare, while maintaining patient and professional confidentiality and taking into account ethical and regulatory considerations so as to guarantee the provision of equitable and effective healthcare.

4.1. Limitations

The primary limitation of this study is the small sample size, which stems from its design being focused solely on testing the efficacy of NLP in detecting SSI through the analysis of EHR clinical notes. This approach was adopted due to the lack of prior research addressing this specific topic. To mitigate this limitation, the study was conducted under semi-experimental conditions, employing a comparative cohort with an SSI rate of 50 %.

As SSI detection was based solely on the HER-CN, the results may be influenced by variations in the quality and terminology of the data input. Ambiguous or incomplete descriptions of symptoms may have affected the accuracy of the assessments. Additionally, not all medical and nursing professionals routinely transcribe laboratory, microbiology, or imaging results in the notes. Furthermore, post-discharge interactions with health services, typically documented by infection control teams through shared electronic health records across all hospitals within the health system, were not recorded in this study. Despite these limitations, the study has several strengths: it utilizes robust data from the VINCat infection surveillance system and focuses on two types of surgery with high infection and complication rates.

4.2. Conclusions

The results indicate that this AI-based model has excellent sensitivity and a high negative predictive value in detecting SSI, making it an effective tool for ruling out infections when none are present. The AUC result further confirms that ChatGPT has good discriminative ability for predicting the presence or absence of SSI compared to the gold standard VINCat.

Future research expanding on this pilot study will focus on refining the model to improve specificity without compromising sensitivity, and on increasing its clinical applicability and efficacy in a variety of healthcare settings. In addition, exploring the integration of more nuanced clinical data and patient-specific factors could improve the robustness and reliability of ChatGPT for detecting SSI.

Ethics

Confidential patient information was protected in accordance with European standards and was approved by the Research Ethics Committee of the Hospital General de Granollers (code no. 82–2024).

Author Declarations

All named authors have seen and agreed to the submitted version of the paper; all people included in the acknowledgements section have agreed to be included.

All the material is original, unpublished and has not been submitted elsewhere.

Registration

The project was registered with the ClinicalTrials.gov Identifier: NCT06556017.

Reporting

The study is reported in accordance with the STARD2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies.

Funding

This study has received no external funding.

Data availability

All data will be made available on request.

Declaration of Competing Interest

All authors declare no conflict of interest relevant to this article.

References

[1] WHO 2022. Global report on infection prevention and control Global report on infection prevention and control. Geneva: World Health Organization; 2022.
[2] Abbas M, Tartari E, Allegranzi B, Pittet D, Harbarth S. The effect of participating in a surgical site infection (SSI) surveillance network on the time trend of SSI rates: a systematic review. Infect Control Hosp Epidemiol 2017 Nov;38(11):1364–6. doi: 10.1017/ice.2017.186. Epub 2017 Aug 24. PMID: 28836491.
[3] Wolfhagen N, Boldingh QJJ, Boermeester MA, De Jonge SW. Perioperative care bundles for the prevention of surgical-site infections: meta-analysis. Br J Surg 2022 Sep 9;109(10):933–42. doi: 10.1093/bjs/znac196. PMID: 35766252; PMCID: PMC10364698.
[4] Tartari E, Tomczyk S, Pires D, Zayed B, Coutinho Rehse AP, Kariyo P, Stempliuk V, Zingg W, Pittet D, Allegranzi B. Implementation of the infection prevention and control core components at the national level: a global situational analysis. J Hosp Infect 2021 Feb;108:94–103. doi: 10.1016/j.jhin.2020.11.025. Epub 2020 Nov 30. PMID: 33271215; PMCID: PMC7884929.
[5] Abbas M, de Kraker MEA, Aghayev E, Astagneau P, Aupee M, Behnke M, et al. Impact of participation in a surgical site infection surveillance network: results from a large international cohort study. J Hosp Infect 2019 Jul;102(3):267–76. doi: 10.1016/j.jhin.2018.12.003. Epub 2018 Dec 7. PMID: 30529703.
[6] van Mourik MSM, van Rooden SM, Abbas M, Aspevall O, Astagneau P, Bonten MJM, et al. PRAISE: providing a roadmap for automated infection surveillance in Europe. Clin Microbiol Infect 2021 Jul;27(1):S3–19. doi: 10.1016/j.cmi.2021.02.028. PMID: 34217466.
[7] Mitchell BG, Hall L, Halton K, MacBeth D, Gardner A. Time spent by infection control professionals undertaking healthcare associated infection surveillance: A multi-centred cross sectional study. Infect Dis Health 2016;21:36–40.
[8] Van Mourik MSM, Perencevich EN, Gastmeier P, Bonten MJM. Designing Surveillance of Healthcare-Associated Infections in the Era of Automation and Reporting Mandates. Clin Infect Dis 2018 Mar 5;66(6):970–6. doi: 10.1093/cid/cix835. PMID: 29514241.
[9] Van Rooden SM, Tacconelli E, Pujol M, Gomila A, Kluytmans JAJW, Romme J, et al. A framework to develop semiautomated surveillance of surgical site infections: An international multicenter study. (Available from:). *Infect Control Hosp Epidemiol* [Internet] 2019;41:194–201. https://doi.org/10.1017/ice.2019.321
[10] Mulder T, Kluytmans-Van Den Bergh MFQ, Van Mourik MSM, Romme J, Crolla RMPH, Bonten MJM, et al. A diagnostic algorithm for the surveillance of deep surgical site infections after colorectal surgery. Infect Control Hosp Epidemiol 2019;40:574–8.
[11] Verberk Msc JDM, Van Rooden SM, Koek MBG, Hetem DJ, Smilde AE, Bril WS, et al. Validation of an algorithm for semiautomated surveillance to detect deep surgical site infections after primary total hip or knee arthroplasty-A multicenter study. (Available from:). *Infect Control Hosp Epidemiol* [Internet] 2021;42:69–74. https://doi.org/10.1017/ice.2020.377

[12] Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis. *J Med Internet Res* J Med Internet Res 2024 Jul 8;26:e56110. https://doi.org/10.2196/56110. PMID: 38976865; PMCID: PMC11263899.

[13] Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. 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. PMID: 37606976; PMCID: PMC10481210.

[14] Zandi R, Fahey JD, Drakopoulos M, Bryan JM, Dong S, Bryar PJ, et al. Exploring Diagnostic Precision and Triage Proficiency: A Comparative Study of GPT-4 and Bard in Addressing Common Ophthalmic Complaints. Bioeng (Basel) 2024 Jan 26;11(2):120. doi: 10.3390/bioengineering11020120. PMID: 38391606; PMCID: PMC10886029.

[15] Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. Lancet Digit Health 2024 Aug;6(8):e555–61. doi: 10.1016/S2589-7500(24)00097-9. PMID: 39059888.

[16] Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. J Am Coll Radio 2023 Oct;20(10):990–7. doi: 10.1016/j.jacr.2023.05.003. Epub 2023 Jun 21. PMID: 37356806; PMCID: PMC10733745.

[17] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198. doi: 10.1371/journal.pdig.0000198. PMID: 36812645; PMCID: PMC9931230.

[18] Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. NPJ Digit Med 2021 Jan 8;4(1):5. doi: 10.1038/s41746-020-00376-2. PMID: 33420381; PMCID: PMC7794558.

[19] Badia JM, Arroyo-Garcia N, Vázquez A, Almendral A, Gomila-Grange A, Fraccalvieri D, et al. Leveraging a nationwide infection surveillance program to implement a colorectal surgical site infection reduction bundle: a pragmatic, prospective, and multicenter cohort study. Int J Surg 2023 Apr 1;109(4):737–51. doi: 10.1097/JS9.0000000000000277. PMID: 36917127; PMCID: PMC10389383.

[20] National Healthcare Safety Network, Surgical Site Infection (SSI) Event: National Healthcare Safety Network., (2023). ⟨https://www.cdc.gov/nhsn/PDFs/pscManual/9pscSSIcurrent.pdf?agree=yes&next=Accept⟩. (accessed December 16, 2022).

[21] Horan TC, Andrus M, Dudeck MA. CDC/NHSN surveillance definition of health care-associated infection and criteria for specific types of infections in the acute care setting. Am J Infect Control 2008 Jun;36(5):309–32. https://doi.org/10.1016/j.ajic.2008.03.002

[22] Flores-Yelamos M, Gomila-Grange A, Badia J, Almendral A. Comparison of two bundles for reducing surgical site infection in colorectal surgery: Multicentre cohort study. BJS Open 2024 Jul 2;8(4):zrae080. doi: 10.1093/bjsopen/zrae080. PMID: 39107075; PMCID: PMC11303006.

[23] Schulman J., Zoph B., Kim C. Introducing ChatGPT. OpenAI [Internet]. 2022 [cited 2024 Aug 13]. Available from: ⟨https://openai.com/index/chatgpt/⟩.

[24] Bossuyt, Reitsma PM, Bruns DE JB, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015 Oct 28;351:h5527. doi: 10.1136/bmj.h5527. PMID: 26511519; PMCID: PMC4623764.

[25] Sohn S, Larson DW, Habermann EB, Naessens JM, Alabbad JY, Liu H. Detection of clinically important colorectal surgical site infection using Bayesian network. J Surg Res 2017 Mar;209:168–73. doi: 10.1016/j.jss.2016.09.058. Epub 2016 Oct 5. PMID: 28032554; PMCID: PMC5391146.