Infection Control & Hospital Epidemiology (2025), 1–3
doi:10.1017/ice.2025.20

Commentary

Chatting new territory: large language models for infection surveillance from pilot to deployment

Julie T. Wu MD, PhD[1,2], Bradley J. Langford PharmD, MPH[3], Erica S. Shenoy MD, PhD[4,5,6], Evan Carey PhD[7,8] and Westyn Branch-Elliman MD, MMSc[9,10,11]

[1]Department of Medicine, VA Palo Alto Healthcare System, Palo Alto, CA, USA, [2]Department of Medicine, Division of Oncology, Stanford University School of Medicine, Palo Alto, CA, USA, [3]Dalla Lana School of Public Health, University of Toronto, Canada, [4]Infection Control, Mass General Brigham, Boston, MA, USA, [5]Division of Infections Diseases, Massachusetts General Hospital, Boston, MA, US, [6]Harvard Medical School, Boston, MA, USA, [7]Digital Health Office, Veterans Health Administration, National Artificial Intelligence Institute, Washington, DC, USA, [8]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA, [9]Department of Medicine, Section of Infectious Diseases, VA Greater Los Angeles Healthcare System, Los Angeles, CA, USA, [10]Digital Health Office, National Artificial Intelligence Institute, Washington, DC, USA and [11]Department of Medicine, Section of Infectious Diseases, UCLA David Geffen School of Medicine, Los Angeles, CA, USA

Abstract

Rodriguez-Nava *et al*. present a proof-of-concept study evaluating the use of a secure large language model (LLM) approved for healthcare data for retrospective identification of a specific healthcare-associated infection (HAI)—central line-associated bloodstream infections—from real patient data for the purposes of surveillance.[1] This study illustrates a promising direction for how LLMs can, at a minimum, semi-automate or streamline HAI surveillance activities.

(Received 12 January 2025; accepted 13 January 2025)

The authors tested a secure version of ChatGPT integrated within their hospital's electronic health record (EHR).[1] Using only the last two progress notes and blood culture results that triggered a possible CLABSI alert in the infection control module of the EHR, their initial approach achieved a 57.5% agreement with the current partially automated manual review process. The method demonstrated higher sensitivity (80%) than specificity (35%), with substantial time savings—just 5 minutes, compared with the 75 minutes reported for manual review. Error analysis revealed that missing essential information in progress notes fed into the LLM, such as absent culture results, contributed to inaccuracies. When access to additional relevant data was provided (e.g., provision of additional notes), agreement improved to 82.5%, though at the expense of need for review to identify the missing data. They note that additional computational resources, such as higher capacity for data collection, could have improved performance. The authors suggest that LLMs could serve as a preliminary screening tool for CLABSI detection, streamlining the typical surveillance process and minimizing the need for extensive manual review. In essence, the LLMs could scan the medical records for relevant datapoints and flag notes that contain the key data elements that could then be used by infection prevention teams to confirm the positive blood culture met CLABSI criteria as defined by the Centers for Disease Control and Prevention (CDC) National Healthcare Safety Network (NHSN).

This study highlights two current opposing trends in healthcare LLM use: (1) the increasing reliance on the out-of-the-box generalizability and ease of use of LLMs for everyday tasks, including surveillance activities, and (2) the growing complexity of optimizing and standardizing LLM performance for high-stakes clinical applications. The rapid evolution of LLM technology has increased this tension. Tools like ChatGPT, increasingly accessible for healthcare use, for example in the EPIC ChatGPT model[2] and the OpenAI Pilot in the national VA healthcare system,[3] allow clinicians to perform technically complex tasks like summarizing and classifying data from the EHR without requiring formal training. There is precedent for integrating technologies into workflows without added liability or reimbursement concerns, such as dictation software to speed documentation. However, significant gaps remain between proof-of-concept studies and the development of tools ready for broad deployment.

When considering a surveillance tool for broader deployment, one must balance the potential risks of implementation (e.g., inaccuracies) against the need for easy-to-use tools that streamline workflow. By focusing on retrospective CLABSI identification (a retrospective surveillance task), the authors removed immediate and direct patient safety risks and created a lower-stakes test case for investigating LLMs applied to a standardized quality metric. The tool essentially is a highly advanced search engine that allows infection prevention programs to better sift through the "noise" to get to the "signal." However, to achieve the robustness in quality necessary for scale beyond this pilot, their approach requires further optimization. Current performance gaps between pilot and

Corresponding author: Julie T. Wu; Email: JulieWu@stanford.edu

Cite this article: Wu JT, Langford BJ, Shenoy ES, Carey E, Branch-Elliman W. Chatting new territory: large language models for infection surveillance from pilot to deployment. *Infect Control Hosp Epidemiol* 2025. doi: 10.1017/ice.2025.20

deployment can be categorized into three dimensions: (1) automation complexity, (2) input data selection, (3) model availability and diversity, and (4) standardization of user workflows.

The first key consideration is the complexity of automation: the degree to which surveillance activities are already automated significantly influences the feasibility of LLM integration. As noted in Shenoy and Branch-Elliman,[4] there are limited examples of fully automated HAI surveillance. Most are partially automated, and one NHSN-defined HAI, ventilator-associated events (VAE), is mostly automated, relying primarily on structured EHR variables with only one item (lung histopathology) requiring manual review, making it an optimal target for a LLM-assisted automation to extract the relevant pathology data from notes. Transitioning from "mostly" to "partially" automated surveillance categories, the complexity and data requirements for automation increases. In the framework presented by Shenoy and Branch-Elliman, CLABSI is considered "partially" automated.[4]

After considering automation complexity, the second key consideration is the selection of input data, as the conclusions drawn by an LLM rely heavily on the quality and completeness of the information it receives. Missing, incomplete, and inaccessible data (or insufficient computational infrastructure leading to artificial limits on accessible data) can significantly compromise accuracy as demonstrated in this study, where such gaps were identified as the primary source of errors. Scaling use beyond pilots such as the one reported by Rodriguez-Nava et al will require efficient methods to detect and address missing information. This will require technology to support access to full patient records—including scanned records and PDF documents—as well as significant investment in computing infrastructure that is substantial enough to run models on larger datasets. The sheer volume of data involved raises the question of whether an LLM alone is sufficient for automating surveillance workflows. A retrieval-augmented generation architecture could help identify and retrieve relevant context for determining HAI occurrences. Methods for capturing relevant information must also account for diverse clinical scenarios, including the less frequent edge cases likely underrepresented in this study's limited 40-case sample, all of which were collected within a 5-month period, and account for the varying degree of missingness across facilities/populations. Such methods for selecting input data will also need to be optimized for robustness to avoid encouraging data manipulation practices that favorably impact quality metrics. Tackling these challenges to obtain reliable and comprehensive input data is a necessary, foundational step for scalable clinical LLM applications.

The third consideration is the selection of the specific model, as it can significantly impact task performance. Even within the same model, different versions may yield varied results, and updates can sometimes degrade performance in ways that are not predictable or knowable in advance. Moreover, ChatGPT, the model used in the study, is proprietary and unavailable in many healthcare systems. Open-source LLMs may serve as alternatives, but their performance could fall short of ChatGPT, with potential task-specific gaps unaddressed by current LLM quality metrics. To address the issue of model diversity in the larger LLM community, benchmark datasets, which have defined questions with defined correct answers, have been used to compare different models and set the standards for future model development. However, the current medical benchmarks used to assess LLM quality often rely on scripted, USMLE-style multiple-choice questions, where all relevant information is pre-collected, cleaned, and explicitly included, and there is one "correct" answer.[5] In contrast, real-world patient data are incomplete, delayed, and missing in nonrandom ways, and there may be more than one correct answer for a given scenario. This discrepancy accounts for a significant gap in performance between tools that are tested on retrospective datasets and those that are implemented for real-world use. As the diversity of models and the pace of innovation in the LLM space continue to grow, standardized guidelines and task-specific benchmarks are needed for ensuring consistent quality and reliable performance across different models. If these tools are going to be widely adopted for the purpose of interfacility comparisons and quality measurement, these task-specific benchmarks will need to be developed and defined by the NHSN. Creation of acceptable standards for accuracy cannot be the primary responsibility of individual healthcare systems.

The fourth consideration is the need to standardize LLM use among healthcare facility HAI surveillance programs to ensure consistent and reliable outputs. As a support tool, LLM-generated results would require verification. The specificity of 35% reported in this study highlights this particular system's limitations, indicating that full automation is not yet feasible (and may never be). Review by experts in HAI surveillance is necessary to eliminate false positives. However, it is uncertain whether the 80% sensitivity observed is sufficient to confidently exclude low-probability CLABSI cases. Any tool designed to filter records for review must ensure that unreviewed cases reliably represent true negatives. Individual variability complicates this further, as infection prevention specialists may have differing thresholds for acceptable performance and variable awareness of potential shortcomings of the system. Establishing clear standards for high-quality outputs and clear workflows to interact with the LLM would help reduce variability and improve the reliability of LLM applications and increase the likelihood that such tools will be considered acceptable and approved adjuncts to HAI surveillance NHSN.

Despite the above challenges, this study provides an important example of how secure LLMs may be able to streamline surveillance workflows with minimal task-specific customization. In theory, LLMs can be used for three purposes to support surveillance activities: first, to flag relevant information in notes for subsequent manual review; second, to read the notes and assign a preliminary classification; and third, to complete the surveillance activity for direct reporting to the CDC. The authors have demonstrated how this tool may be useful for flagging data elements and notes that contain information germane to CLABSI identification – however, it is unclear what level of accuracy is needed in order to maintain the utility of the surveillance process. Facilities will need to determine what level of accuracy and sensitivity is necessary to inform their own internal efforts. If these tools are ever to be used for more direct reporting to the CDC-potentially without manual review step- federal guidance about accuracy and validation standards will be needed.

In summary, selecting the right applications of the tools that leverage their strengths (quickly reading and summarizing large volumes of data) while minimizing their limitations and their computing infrastructure resourcing needs is critical. Retrospective surveillance activities are fundamentally a good use of these tools. However, addressing key challenges, including consideration of automation complexity, the selection of appropriate input data, computing infrastructure, maintaining consistent performance across the diversity of available LLMs, and developing standardized workflows for human verification, will be needed for transitioning from individual-level pilots to broader, scalable deployment in clinical settings. Ultimately, the model is only as good as the data it can "see"—and given the nature of clinical data, this will be a perennial challenge as we integrate this new technology into surveillance workflows.

Acknowledgements. We are thankful to Lauren Murphy and Leo Meyerovich for useful comments and discussion about the manuscript. They were not compensated.

Financial support. This work is supported by Centers for Disease Control and Prevention Cooperative Agreement CK-22-2203 (ESS) and VA National Artificial Intelligence Institute in the VA Digital Health Office (WBE), and VA VISN21 Early Career Award (JTW).

Competing interests. ESS has received royalties from UpToDate, consulting fees from BARDA/CARBX/Boston University, travel fees from ID Week, SHEA Spring, and ICPIC, leadership or fiduciary roles from HICPAC, Boston Biosafety Commission, Massachusetts Infectious Diseases Society and SHEA, and is supported by grants from CDC, ASPR, and MIT outside this work. WBE is supported by grants from VA Health Services Research and Development Service, VA Cooperative Studies Program outside this work. JW is supported by grants from LCFA/IASLC/BMS outside this work.

References

1. Rodriguez-Nava G, Egoryan G, Goodman KE, Morgan DJ, Salinas JL. Performance of a large language model for identifying central line-associated bloodstream infections (CLABSI) using real clinical notes. *Infect Control Hosp Epidemiol* 2024 Oct 30:1–4.
2. Cool Stuff Now: Epic and Generative AI | Epic. https://www.epic.com/epic/post/cool-stuff-now-epic-and-generative-ai/.
3. VA AI Use Case Inventory. https://department.va.gov/ai/ai-use-case-inventory/.
4. Branch-Elliman W, Sundermann AJ, Wiens J, Shenoy ES. Leveraging electronic data to expand infection detection beyond traditional settings and definitions (Part II/III). *Antimicrob Steward Healthc Epidemiol* 2023;3:e27.
5. Yan LKQ, Niu Q, Li M, *et al.* Large language model benchmarks in medical tasks. Published online December 9, 2024.

https://doi.org/10.1017/ice.2025.20 Published online by Cambridge University Press