NEJM AI

NEJM AI 2024;1(11)
DOI: 10.1056/AIcs2400420

CASE STUDY

Large Language Models for More Efficient Reporting of Hospital Quality Measures

Aaron Boussina, Ph.D.,[1] Rishivardhan Krishnamoorthy, M.S.,[1] Kimberly Quintero, R.N., M.S.,[2] Shreyansh Joshi, Gabriel Wardi, M.D.,[1,3,4] Hayden Pour, M.S.,[1] Nicholas Hilbert, R.N., M.S.N.,[2] Atul Malhotra, M.D.,[3] Michael Hogarth, M.D.,[1] Amy M. Sitapati, M.D.,[1] Chad VanDenBerg, M.P.H.,[2] Karandeep Singh, M.D., M.M.Sc.,[5] Christopher A. Longhurst, M.D., M.S.,[3] and Shamim Nemati, Ph.D.[1]

Received: April 25, 2024; Revised: August 9, 2025; Accepted: August 16, 2024; Published: October 21, 2024

Abstract

Hospital quality measures are a vital component of a learning health system, yet they can be costly to report, statistically underpowered, and inconsistent due to poor interrater reliability. Large language models (LLMs) have recently demonstrated impressive performance on health care–related tasks and offer a promising way to provide accurate abstraction of complete charts at scale. To evaluate this approach, we deployed an LLM-based system that ingests Fast Healthcare Interoperability Resources data and outputs a completed Severe Sepsis and Septic Shock Management Bundle (SEP-1) abstraction. We tested the system on a sample of 100 manual SEP-1 abstractions that University of California San Diego Health reported to the Centers for Medicare & Medicaid Services in 2022. The LLM system achieved agreement with manual abstractors on the measure category assignment in 90 of the abstractions (90%; κ=0.82; 95% confidence interval, 0.71 to 0.92). Expert review of the 10 discordant cases identified four that were mistakes introduced by manual abstraction. This pilot study suggests that LLMs using interoperable electronic health record data may perform accurate abstractions for complex quality measures. (Funded by the National Institute of Allergy and Infectious Diseases [1R42AI177108-1] and others.)

Introduction

In 2022, quality reporting at a single U.S. acute care hospital was estimated to cost more than US$5 million and require more than 100,000 person-hours of work.[1] Moreover, among all U.S. physician practices, quality reporting has been estimated to cost more than US$15 billion and require 785 hours per physician annually.[2] Yet, despite such massive financial and reporting burdens, quality measures are often assessed on a small denominator of patients, which limits statistical validity and can lead to delays in both measurement and improvement.[3-6] These limitations were manifested in March 2020 when, during the Covid-19 pandemic, the Centers for Medicare & Medicaid Services (CMS) granted health care organizations relief from quality reporting "so the healthcare delivery system can direct its time and resources toward caring for patients."[4,7,8]

The author affiliations are listed at the end of the article.

Dr. Boussina can be contacted at aboussina@health.ucsd.edu.

NEJM AI is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from ai.nejm.org on November 1, 2024. For personal use only.
No other uses without permission. Copyright © 2024 Massachusetts Medical Society. All rights reserved.

The Severe Sepsis and Septic Shock Management Bundle (SEP-1) measure from CMS is a microcosm of the challenges involved in hospital quality reporting.[9] Previously a pay-for-reporting program, the SEP-1 measure will be included in the Hospital Value-Based Purchasing Program, starting in 2026.[10] This addition has been met with opposition from various professional societies, including the Infectious Diseases Society of America and the American College of Emergency Physicians, due, in part, to the measure's reporting burden and abstraction variability.[3,11] Indeed, SEP-1 is an "all-or-nothing" composite measure requiring a complex, 63-step abstraction process that is completed through manual chart review.[12,13]

At University of California San Diego Health (UCSDH), abstraction involves an initial determination by nonclinical analysts from an external vendor, followed by a review from nurses on the quality team, and then by a final physician review. CMS requires monthly sampling of at least 20 patients who meet the measure's inclusion criteria (e.g., inpatient, *International Classification of Diseases*, Tenth Revision, Clinical Modification [ICD-10-CM] principal or other diagnosis code of sepsis, severe sepsis, or septic shock).[9,14]

Recent work has demonstrated that large language models (LLMs) can achieve impressive performance on medical-related tasks, including human-level performance on standardized medical tests, even without task-specific fine-tuning.[15–17] Quality measurement is a complex task that entails a unique set of challenges based on both the medical knowledge required to answer questions as well as the need to parse the temporal nature of the clinical course of diagnosis and treatment. In this work, we investigate whether LLMs using interoperable electronic health record (EHR) data can enable the accurate automated abstraction of complex quality measures. We use the SEP-1 measure as a case study due to its well-studied complexity.

Methods

STUDY DESIGN AND COHORT

We developed and deployed an interoperable LLM system and tested it on a convenience sample of all manual SEP-1 abstractions at UCSDH that were reported to CMS from January to May 2022. The sample represented 100 cases across two hospitals from three abstractors. The abstractors were nonclinical specialists from a single vendor who were trained on a standard operating procedure for SEP-1 abstraction.

Our primary outcome was the measure of agreement on category assignment (pass, fail, or out of measure) between the LLM system and the current standard inclusive of human abstractors. We tested agreement using Cohen's kappa with a two-sided test.[18] Disagreements between the LLM system and human abstractors were adjudicated by a board-certified emergency medicine and critical care physician who chairs the UCSDH Sepsis Committee; the disagreements are reported separately.

We performed three independent trials of the LLM system to evaluate consistency. A random 10% of cases on which the LLM and human abstractors agreed were evaluated for interrater reliability by the same physician expert. We additionally determined compliance rates that were system-generated and system-reported (to CMS), as well as their 95% confidence intervals using the Clopper–Pearson exact method. A P value of less than 0.05 was considered significant for all analyses. All statistical analyses were performed using Python version 3.11, the SciPy package version 1.10.1, and the statsmodels package version 0.13.5.[19,20] UCSDH Institutional Review Board approval was obtained with a waiver of informed consent (805726).

SYSTEM DESIGN

Our system architecture is shown in Figure 1. Data are retrieved in Fast Healthcare Interoperability Resources (FHIR) version R4 format.[21] Structured data are retrieved from the Patient, Observation, ServiceRequest, Consent, Flag, and MedicationRequest FHIR resource types. Unstructured notes are gathered from the DocumentReference and Binary resources. Medication administration information is not available in FHIR R4 and is retrieved from a proprietary Epic application programming interface (API).

The 63-step SEP-1 process flowchart was translated into Python and hosted on a cloud-based virtual machine within a Health Insurance Portability and Accountability Act (HIPAA)–compliant virtual private cloud (VPC).[22] The system proceeds through the measure and queries an LLM by performing retrieval-augmented generation (RAG) on a patient's clinical notes. It uses CMS guidelines as prompt instructions when it reaches a step that requires information from unstructured data.[23] The system leverages the Sepsis Consensus Toolkit (Fig. 1), a set of utilities developed for this case study, to establish the presence of clinical criteria such as systemic inflammatory response syndrome and the presence of organ failure from structured FHIR data. These criteria are then combined with

NEJM AI is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from ai.nejm.org on November 1, 2024. For personal use only.
No other uses without permission. Copyright © 2024 Massachusetts Medical Society. All rights reserved.

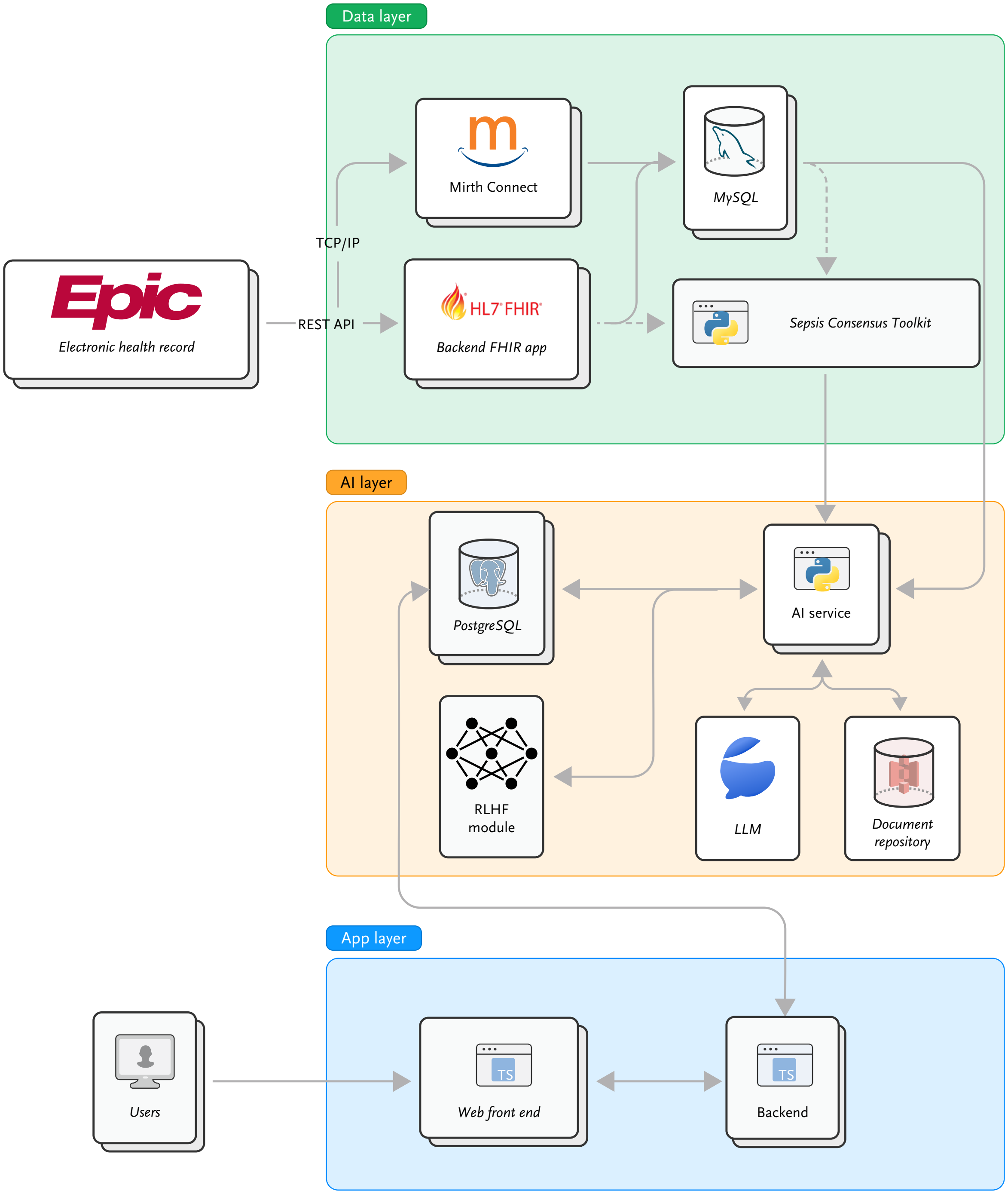
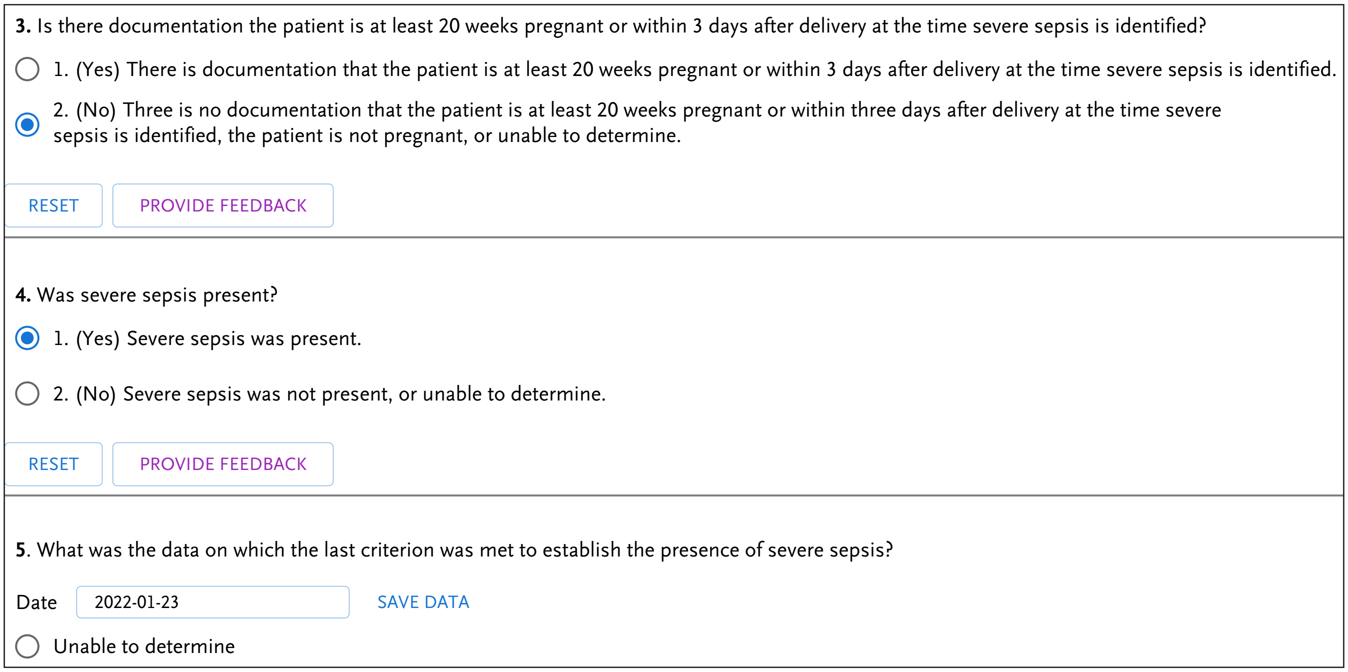


Figure 1. System Architecture for Automation of Hospital Quality Measures.

The data layer (green) enables the collection of electronic health record data through Fast Healthcare Interoperability Resources (FHIR) and the computation of clinical criteria. Mirth Connect stores all encounters from admission-discharge-transfer messages. The backend FHIR application then queries encounter data and stores it in MySQL. The Sepsis Consensus Toolkit applies standard rule-based criteria to the structured data to identify systemic inflammatory response syndrome and organ failure events. The artificial intelligence layer (orange) manages the large language model for abstraction. The app layer (blue) services the completed abstractions and collects human feedback. AI denotes artificial intelligence; API, application programming interface; FHIR, Fast Healthcare Interoperability Resources; LLM, large language model; REST, representational state transfer; RLHF, reinforcement learning from human feedback; and TCP/IP, Transmission Control Protocol/Internet Protocol.

NEJM AI is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from ai.nejm.org on November 1, 2024. For personal use only.
No other uses without permission. Copyright © 2024 Massachusetts Medical Society. All rights reserved.



3. Is there documentation the patient is at least 20 weeks pregnant or within 3 days after delivery at the time severe sepsis is identified?

○ 1. (Yes) There is documentation that the patient is at least 20 weeks pregnant or within 3 days after delivery at the time severe sepsis is identified.

● 2. (No) Three is no documentation that the patient is at least 20 weeks pregnant or within three days after delivery at the time severe sepsis is identified, the patient is not pregnant, or unable to determine.

4. Was severe sepsis present?

● 1. (Yes) Severe sepsis was present.

○ 2. (No) Severe sepsis was not present, or unable to determine.

5. What was the data on which the last criterion was met to establish the presence of severe sepsis?

Date 2022-01-23

○ Unable to determine

Figure 2. Web Application Front End for the System.

Shown is sample output for the Severe Sepsis and Septic Shock Management Bundle (SEP-1) measure. The measure data elements are preloaded from a database in the artificial intelligence layer. The user can change the element, which creates a human feedback record.

LLM responses to select the appropriate allowable value for each SEP-1 data element.

The final output from the system is a completed SEP-1 abstraction including the measure category assignment. We provide this result to users through a web application (Fig. 2). Users can change data elements within the application's front-end interface, which triggers creation of a back-end "human feedback" record.

LLM IMPLEMENTATION

LLM inference was performed using the open-source, general-purpose SOLAR 10.7B model with 8-bit quantization and a context length of 8092 tokens.[24,25] We selected this model because it could be hosted on a single 24-gigabyte graphics processing unit in a HIPAA–compliant environment and because it has relatively strong performance on standard benchmarks for its size.[26] No additional fine-tuning or prompt tuning was performed, and all data remained within the VPC. We utilized chain-of-thought and few-shot prompting strategies with a temperature of 0.1.[27,28] This temperature is lower than the default value and was chosen to improve the reproducibility of the system. LLM outputs were cast to JavaScript Object Notation, and invalid outputs were regenerated. The prompt template, all prompts, and few-shot examples are detailed in Notes S1 and S2 in the Supplementary Appendix.

RAG was performed on the clinical notes by chunking the text into 1000-character segments with 50 characters of overlap, embedding the chunks and query with the Instructor model, calculating the cosine similarity between the query and the embeddings, and inserting the top six most similar chunks into the prompt.[29] The relevant code is available at https://github.com/aboussina/quallm.

Results

Table 1 summarizes patient characteristics and SEP-1 measure results based on standard reporting for the study cohort, and Table 2 shows LLM system agreement with standard reporting inclusive of manual abstraction. We observed that the LLM system generated identical measure category assignments across all three trials and achieved agreement with manual abstractors on measure category assignment for 90 of 100 abstractions (90%; κ=0.82; 95% confidence interval [CI], 0.71 to 0.92). The physician adjudication of the 10 discordant cases is described in Table 3. In 4 of the 10 cases, the reviewing physician concluded

NEJM AI is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from ai.nejm.org on November 1, 2024. For personal use only.
No other uses without permission. Copyright © 2024 Massachusetts Medical Society. All rights reserved.

Table 1. Study Cohort Demographics and SEP-1 Measure Results.*

Patient Characteristics	SEP-1 Abstraction Cohort (n=100)
Age — median years (IQR)	66.5 (53.50–74.25)
Sex — no. of patients (%)	
Female	37 (37)
Male	63 (63)
Race — no. of patients (%)†	
White	49 (49)
Black or African American	8 (8)
American Indian or Alaska Native	0 (0)
Asian or Pacific Islander	10 (10)
Other race or multiracial	33 (33)
Major measure elements — no./total no. of patients (%)	
Blood culture collection	41/42 (97.6)
Initial lactate level collection	36/39 (92.3)
Broad spectrum or other antibiotic administration — no./total no. of patients (%)	45/50 (90)
Repeat lactate level collection	22/24 (91.7)
Crystalloid fluid administration	5/17 (29.4)
Measure category assignment — no. of patients (%)	
B (not in measure population)	62 (62)
D (in measure population)	22 (22)
E (in numerator population)	16 (16)
Compliance rate — no./total no. of patients (%)	16/38 (42.1)

* Of 100 cases, 62 did not qualify for the measure (not in measure population; B) per the SEP-1 guidelines. There were 22 patients in the measure population (D) and 16 numerator-compliant patients in the numerator population (E) for an overall compliance rate of 16 of 38 (42.1%; 95% confidence interval, 28.6% to 61.7%). Blood cultures were collected within the specified time frame in 41 of 42 cases (97.6%); timely antibiotics were administered in 45 of 50 cases (90.0%); and, when initial hypotension was present, 30 ml/kg of crystalloid fluids were initiated and completely infused in 5 of 17 cases (29.4%). IQR denotes interquartile range; and SEP-1, Severe Sepsis and Septic Shock Management Bundle.

† Race was reported by the participants.

that the LLM system was more accurate than the human abstractor.

Agreement by measure category is detailed in Note S3. The LLM system classified 19 cases as numerator compliant and 20 as noncompliant, which together make up the denominator and resulted in a compliance rate of 19 of 39 (48.7%; 95% CI, 32.4% to 65.2%). The system also classified 61 cases as out of measure. Of the random 10% of cases in which there was agreement between the LLM and human abstractors, our physician expert found a Cohen's kappa of 1.0. An example abstraction from the LLM system is shown in Note S4. Example errors from the LLM are shown in Note S5.

Discussion

Prior work has advocated reducing the number of quality measures or transitioning to simpler electronic clinical quality measures, which are approaches that, historically, have presented challenges in matching robust performance.[30] This study offers an alternative: relief from reporting burden through better tools that appropriately capture case complexity and provide timely feedback. To that end, we have demonstrated that LLMs using interoperable EHR data may accurately perform abstraction of the SEP-1 quality measure and, furthermore, that open-source LLMs running on consumer-grade hardware may be sufficiently capable. To our knowledge, this represents the first work to explore the capabilities of LLMs for hospital quality reporting. The SEP-1 measure is one of the most complex quality measures, which makes it a suitable stress test for quality measurement in general. The availability of previously reported abstractions and the ability to collect user feedback from our system interface also offer opportunities for improved performance through supervised fine-tuning and reinforcement learning from human feedback.[31,32]

NEJM AI is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from ai.nejm.org on November 1, 2024. For personal use only.
No other uses without permission. Copyright © 2024 Massachusetts Medical Society. All rights reserved.

Table 2. Large Language Model System Agreement with Manual Abstraction for SEP-1.[*]

SEP-1 Question	No. (%) of Abstractions Where System Answer Resulted in Agreement with Manual Category Assignment[†]	Data Element Distribution (no.; %) from Manual Abstraction
Was the patient received as a transfer from an inpatient, outpatient, or emergency/observation department of an outside hospital or from an ambulatory surgery center?[‡]	99/100 (99)	Y (10; 10)
		N (90; 90)
During this hospital stay, was the patient enrolled in a clinical trial in which patients with the same condition as the measure set were being studied?[‡]	100/100 (100)	Y (0; 0)
		N (100; 100)
Is there documentation the patient is at least 20 weeks pregnant or within 3 days after delivery at the time severe sepsis is identified?[‡]	100/100 (100)	Y (0; 0)
		N (100; 100)
Was severe sepsis present?[‡]	98/100 (98)	Y (49; 53.8)
		N (42; 46.2)
When was the last criterion met to establish the presence of severe sepsis?[‡]	97/100 (97)	—
Is there documentation that the patient or authorized patient advocate refused either a blood draw, IV fluid administration, or IV antibiotic administration within the specified time frame?[‡]	100/100 (100)	Y (3; 5.4)
		N (53; 94.6)
Is there physician/APN/PA documentation of comfort measures only, palliative care, or another inclusion term before or within 6 hours after the presentation of severe sepsis?[‡]	99/100 (99)	Y (5; 8.9)
		N (51; 91.1)
What was the patient's discharge disposition on the day of discharge?[‡]	100/100 (100)	Home (62; 62)
		Hospice — home (2; 2)
		Hospice — health care facility (3; 3)
		Acute care facility (2; 2)
		Other health care facility (16; 16)
		Expired (10; 10)
		Left AMA (5; 5)
		Not documented or unable to determine (0; 0)
Was a broad-spectrum or other antibiotic administered within the specified time frame?	100/100 (100)	Y (45; 90)
		N (5; 10)
What was the earliest datetime on which an antibiotic was started within the specified time frame?	100/100 (100)	—
Was a blood culture collected within the specified time frame?	100/100 (100)	Y (41; 97.6)
		N (1; 2.4)
When was the blood culture collected?	100/100 (100)	—
Is there documentation supporting an acceptable delay in collecting a blood culture?[‡]	99/100 (99)	Y (1; 33.3)
		N (2; 66.6)
Was an initial lactate level drawn within the specified time frame?	100/100 (100)	Y (36; 92.3)
		N (3; 7.7)
What was the datetime on which the initial lactate level was drawn?	100/100 (100)	—
What was the initial lactate level result? (≤2 mmol/l, >2 mmol/l and <4 mmol/l, or ≥4 mmol/l)	100/100 (100)	≤2 mmol/l (12; 33.3)
		>2 mmol/l and <4 mmol/l (17; 47.2)

(*Continued*)

NEJM AI is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from ai.nejm.org on November 1, 2024. For personal use only.
No other uses without permission. Copyright © 2024 Massachusetts Medical Society. All rights reserved.

Table 2. (*Continued*) Large Language Model System Agreement with Manual Abstraction for SEP-1.

SEP-1 Question	No. (%) of Abstractions where System Answer Resulted in Agreement with Manual Category Assignment†	Data Element Distribution (no.; %) from Manual Abstraction
		≥4 mmol/l (7; 19.4)
Was a repeat lactate level drawn within the specified time frame?	100/100 (100)	Y (22; 91.7) N (2; 8.3)
What was the earliest datetime on which the repeat lactate level was drawn?	100/100 (100)	—
Was initial hypotension present within the specified time frame?	100/100 (100)	Y (12; 35.3) N (22; 64.7)
When was initial hypotension present 6 hours prior to or within 6 hours following severe sepsis presentation date and time?	100/100 (100)	—
Were crystalloid fluids initiated within the specified time frame and completely infused based on the target ordered volume?	99/100 (99)	Y (5; 29.4) N (12; 70.6)
What was the earliest datetime on which crystalloid fluids were initiated within the specified time frame?	100/100 (100)	—
Is there documentation of the presence of septic shock?‡	100/100 (100)	Y (7; 26.9) N (19; 73.1)
What was the datetime on which the last criterion was met to establish the presence of septic shock?‡	100/100 (100)	—
Is there documentation that the patient or authorized patient advocate refused either a blood draw, IV fluid administration, or vasopressor administration before or within 6 hours after the septic shock presentation time?‡	100/100 (100)	Y (0; 0) N (7; 100)
Is there physician/APN/PA documentation of comfort measures only, palliative care, or another inclusion term before or within 6 hours after the presentation of septic shock?‡	100/100 (100)	Y (0; 0) N (7; 100)
Was persistent hypotension or new onset of hypotension present within 1 hour of when the target ordered volume of crystalloid fluids was completely infused?	99/100 (99)	Y (1; 16.7) N (5; 83.3)
Was an IV or intraosseous vasopressor administered within the specified time frame?	100/100 (100)	Y (1; 100) N (0; 0)
What was the datetime on which an IV or intraosseous vasopressor was administered within the specified time frame?	100/100 (100)	—
Was a repeat volume status and tissue perfusion assessment performed within the specified time frame?‡	100/100 (100)	Y (3; 100) N (0; 0)
What datetime was a repeat volume status and tissue perfusion assessment performed?‡	100/100 (100)	—
Final measure category assignment	90/100 (90)§	B (62; 62) D (22; 22) E (16; 16)

* Each row represents a SEP-1 process step and the rows are ordered from the beginning to the end of the process flow. AMA denotes against medical advice; APN, advanced practice nurse; IV, intravenous; LLM, large language model; PA, physician assistant; and SEP-1, Severe Sepsis and Septic Shock Management Bundle.

† The numerator is the number of cases where a response generated by the LLM system did not result in disagreement of the measure category assignment. The denominator is the total number of abstractions.

‡ Indicates that an LLM query was utilized for the question.

§ Because SEP-1 is an all-or-nothing measure, a mismatch of even a single criterion can result in a different final measure category assignment.

NEJM AI is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from ai.nejm.org on November 1, 2024. For personal use only.
No other uses without permission. Copyright © 2024 Massachusetts Medical Society. All rights reserved.

Table 3. Physician Review of Discrepant Cases.*

Root Cause of Disagreement	Number of Cases	Physician Adjudication
The patient had chronic kidney disease with a creatinine baseline of (2–3) mg/dL. The system identified a creatinine value elevated >0.5 above baseline as evidence of organ failure. The abstractor did not identify this as a sign of organ failure.	1	LLM system is more accurate
INR organ failure missed by abstractor.†	1	LLM system is more accurate
Difference in documentation of infection time.	2	LLM system is more accurate
	1	Abstractor is more accurate
LLM is too sensitive to the presence of palliative care.	1	Abstractor is more accurate
LLM is too sensitive to an acceptable delay in blood cultures.	1	Abstractor is more accurate
Missing data fields (e.g., arterial line blood pressure measurements).†	3	Abstractor is more accurate

* INR denotes international normalized ratio; and LLM, large language model.
† Fields were incorporated into the system after physician adjudication.

This approach is promising because evaluating a measure across a cohort of patients can be easily scaled beyond standard sampling for robust statistical findings. Within our study cohort, only 38 patients were included in the measure after 5 months of reporting. This data sample is insufficient to identify meaningful quality improvement opportunities. However, this LLM system affords a feasible approach that may enable SEP-1 abstraction to scale to every patient with an encounter during a reporting period.

Equally importantly, these findings can be generated shortly after patient discharge, which can shorten the time necessary to incorporate process improvements within a learning health system. Timely auditing and feedback have been shown to improve measure compliance.[33,34] Unfortunately, quality measures are often prepared at either a monthly or quarterly resolution. For SEP-1 reporting at UCSDH, only cases from 2 months prior are prepared in a given month, which precludes the use of the measure to proactively target systemic issues.

Artificial intelligence for quality reporting also offers a promising avenue to reduce the variability inherent in human chart review. The National Quality Forum (NQF) takes the position that a performance measure cannot be scientifically acceptable if its data elements have poor interrater reliability.[35] The NQF recommends that measure developers avoid data elements with a kappa statistic lower than 0.41. Yet, Rhee et al. demonstrated that the SEP-1 pass rate had a kappa of 0.39 across three reviewers at three hospitals and that abstractors agreed on time zero in only 36% of cases.[5] In this study, we observed a few examples of human error that could contribute to poor reliability (Table 3). In two cases, clear documentation of suspected infection was overlooked by reviewers, resulting in a different time zero. In one case, the presence of organ failure due to an international normalized ratio value greater than 1.5 was missed. We also observed clear errors and hallucinations by the LLM, resulting in incorrect abstractions (Note S5). In one case, the LLM inappropriately conflated palliative radiation therapy with comfort measures only. In another, the LLM inferred, on the basis of insufficient evidence, that an infection was being treated prior to the presentation of severe sepsis. With improved grounding and alignment, the ability to apply the same criteria, prompts, and model to a consistent set of interoperable data elements holds promise for improving intrasystem and intersystem reliability.

Our study has several limitations. First, our study cohort is a small convenience sample of only 100 cases across two hospitals. This sample represents 5 months of abstraction in our health system and highlights the limited scope of manual review in the current state. Second, while we used interoperable data standards wherever possible, the system was reliant on a proprietary API for medication administration information and would require modification to support other EHR vendors. Third, we did not explore performance across different LLMs. However, since we used a midsize, general-purpose LLM that was not fine-tuned on our data or adapted to the medical domain, we expect that our results are generalizable. Finally, although we provided a web application front end for our system, we did not explore the human–computer interaction. Future work is needed to evaluate whether human abstractors equipped with this tool can achieve greater accuracy, reliability, and efficiency. Future work will also evaluate whether automated abstraction generation can save clinician reviewer time by presenting clear evidence and enabling rapid rework.

Ultimately, the evolution of quality metrics through the adoption of interoperability standards and artificial intelligence offers a promising avenue to alleviate the workload associated with manual chart reviews, thereby reallocating precious time to health care quality initiatives.

NEJM AI is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from ai.nejm.org on November 1, 2024. For personal use only.
No other uses without permission. Copyright © 2024 Massachusetts Medical Society. All rights reserved.

Funding

This work has been supported by the National Institute of Allergy and Infectious Diseases (1R42AI177108 to Dr. Boussina and Dr. Nemati), the National Library of Medicine (2T15LM011271-11 to Dr. Boussina), and the National Institute of General Medical Sciences (R35GM143121 to Dr. Nemati and K23GM146092 to Dr. Wardi). The opinions or assertions contained herein are the private ones of the authors and are not to be construed as official or reflecting the views of the National Institutes of Health or any other agency of the U.S. government. The authors would also like to acknowledge the support provided by the Joan & Irwin Jacobs Center for Health Innovation.

Disclosures

Author disclosures are available at ai.nejm.org.

Author Affiliations

[1] Division of Biomedical Informatics, University of California, San Diego, San Diego

[2] Department of Quality, University of California, San Diego, San Diego

[3] Department of Medicine, University of California, San Diego, San Diego

[4] Department of Emergency Medicine, University of California, San Diego, San Diego

[5] Joan & Irwin Jacobs Center for Health Innovation, University of California, San Diego, San Diego

References

1. Saraswathula A, Merck SJ, Bai G, et al. The volume and cost of quality metric reporting. JAMA 2023;329:1840-1847. DOI: 10.1001/jama.2023.7271.

2. Casalino LP, Gans D, Weber R, et al. US physician practices spend more than $15.4 billion annually to report quality measures. Health Aff (Millwood) 2016;35:401-406. DOI: 10.1377/hlthaff.2015.1258.

3. Ash AS, Fienberg SF, Louis TA, Normand S-L, Stukel TA, Utts J. Statistical issues in assessing hospital performance. 2012. (http://hdl.handle.net/20.500.14038/46655).

4. Austin JM, Kachalia A. The state of health care quality measurement in the era of COVID-19: the importance of doing better. JAMA 2020;324:333-334. DOI: 10.1001/jama.2020.11461.

5. Rhee C, Brown SR, Jones TM, et al. Variability in determining sepsis time zero and bundle compliance rates for the Centers for Medicare & Medicaid Services SEP-1 measure. Infect Control Hosp Epidemiol 2018;39:994-996. DOI: 10.1017/ice.2018.134.

6. Bauer SR, Gonet JA, Rosario RF, Griffiths LA, Kingery T, Reddy AJ. Inter-rater agreement for abstraction of the early management bundle, severe sepsis/septic shock (SEP-1) quality measure in a multi-hospital health system. Jt Comm J Qual Patient Saf 2019;45:108-111. DOI: 10.1016/j.jcjq.2018.10.002.

7. Centers for Medicare & Medicaid Services. CMS announces relief for clinicians, providers, hospitals and facilities participating in quality reporting programs in response to COVID-19. March 22, 2020 (https://www.cms.gov/newsroom/press-releases/cms-announces-relief-clinicians-providers-hospitals-and-facilities-participating-quality-reporting).

8. Rosenbaum L. Reassessing quality assessment—the flawed system for fixing a flawed system. N Engl J Med 2022;386:1663-1667. DOI: 10.1056/NEJMms2200976.

9. Centers for Medicare & Medicaid Services. Quality net—inpatient hospitals specifications manual. April 16, 2024 (https://qualitynet.cms.gov/inpatient/specifications-manuals).

10. Centers for Medicare & Medicaid Services. Medicare program: proposed hospital inpatient prospective payment systems for acute care hospitals and the long-term care hospital prospective payment system and policy changes and fiscal year 2024 rates. Fed Regist 2023;88:27193. (https://www.federalregister.gov/documents/2023/08/28/2023-16252/medicare-program-hospital-inpatient-prospective-payment-systems-for-acute-care-hospitals-and-the).

11. Rhee C, Strich JR, Chiotos K, et al. Improving sepsis outcomes in the era of pay-for-performance and electronic quality measures: a joint IDSA/ACEP/PIDS/SHEA/SHM/SIDP position paper. Clin Infect Dis 2024;78:505-513. DOI: 10.1093/cid/ciad447.

12. Gesten F, Evans L. SEP-1—taking the measure of a measure. JAMA Netw Open 2021;4:e2138823. DOI: 10.1001/jamanetworkopen.2021.38823.

13. Aaronson EL, Filbin MR, Brown DFM, Tobin K, Mort EA. New mandated Centers for Medicare & Medicaid Services requirements for sepsis reporting: caution from the field. J Emerg Med 2017;52:109-116. DOI: 10.1016/j.jemermed.2016.08.009.

14. Centers for Medicare & Medicaid Services. QualityNet - Hospital quality reporting important dates and deadlines. April 23, 2024 (https://www.qualityreportingcenter.com/globalassets/2024/01/iqr/19.-iqr-important-dates-deadlines_february-2024508.pdf).

15. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. March 20, 2023 (https://arxiv.org/abs/2303.13375). Preprint.

16. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620:172-180. DOI: 10.1038/s41586-023-06291-2.

17. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. Nature 2023;619:357-362. DOI: 10.1038/s41586-023-06160-y.

18. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276-282. DOI: 10.11613/BM.2012.031.

19. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261-272. DOI: 10.1038/s41592-019-0686-2.

20. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. SciPy 2010;7:1. DOI: 10.25080/MAJORA-92BF1922-011.

21. HL7 FHIR Release 4. April 17, 2024 (https://hl7.org/fhir/R4/).

NEJM AI is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from ai.nejm.org on November 1, 2024. For personal use only.
No other uses without permission. Copyright © 2024 Massachusetts Medical Society. All rights reserved.

22. Boussina A, Shashikumar S, Amrollahi F, Pour H, Hogarth M, Nemati S. Development & deployment of a real-time healthcare predictive analytics platform. April 11, 2023 (https://www.medrxiv.org/content/10.1101/2023.04.10.23288373v1). Preprint.

23. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst 2020;33:9459-9474.

24. Kim, D, Park C, Kim S, et al. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. December 23, 2023 (https://doi.org/10.48550/arXiv.2312.15166). Preprint.

25. Dettmers T, Zettlemoyer L. The case for 4-bit precision: k-bit inference scaling laws. International Conference on Machine Learning. PMLR 2023:7750-7774. DOI: 10.48550/arXiv.2212.09720.

26. Hugging Face. Open LLM Leaderboard. April 18, 2024 (https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard).

27. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst 2022;35:24824-24837.

28. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877-1901.

29. Su H, Shi W, Kasai J, et al. One embedder, any task: Instruction-finetuned text embeddings. December 19, 2022 (https://doi.org/10.48550/arXiv.2212.09741). Preprint.

30. Ahmad FS, Rasmussen LV, Persell SD, et al. Challenges to electronic clinical quality measurement using third-party platforms in primary care practices: the healthy hearts in the heartland experience. JAMIA Open 2019;2:423-428. DOI: 10.1093/jamiaopen/ooz038.

31. J Wei, M. Bosma, V. Y. Zhao, et al. Finetuned language models are zero-shot learners. September 3, 2021 (https://doi.org/10.48550/arXiv.2109.01652). Preprint.

32. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 2022;35:27730-27744.

33. Borgert M, Binnekade J, Paulus F, Goossens A, Vroom M, Dongelmans D Timely individual audit and feedback significantly improves transfusion bundle compliance-a comparative study. Int J Qual Health Care 2016;28:601-607. DOI: 10.1093/intqhc/mzw071.

34. Clark RC, Carter KF, Jackson J, Hodges D. Audit and feedback: a quality improvement study to increase pneumococcal vaccination rates. J Nurs Care Qual 2018;33:291-296. DOI: 10.1097/NCQ.0000000000000289.

35. Glance LG, Maddox KJ, Johnson K, et al. National Quality Forum guidelines for evaluating the scientific acceptability of risk-adjusted clinical outcome measures: a report from the National Quality Forum Scientific Methods Panel. Ann Surg 2020;271:1048-1055. DOI: 10.1097/SLA.0000000000003592.

NEJM AI is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from ai.nejm.org on November 1, 2024. For personal use only.
No other uses without permission. Copyright © 2024 Massachusetts Medical Society. All rights reserved.