Paper ready in results and in rewriting of introduction and discussion

Antimicrobial Stewardship in the Era of AI: A Head-to-Head Comparison of the Arkstone Machine Learning HTL Algorithm and Large Language Models in Real-World Infectious Disease Cases

Juan C. Gómez de la Torre [1,2,6], Ari Frenkel [2], Carlos Chavez-Lencinas [3,4], Alicia Rendon [2], José Alonso Cáceres[1], Miguel Hueda-Zavaleta [2,5]*

[1] Clinical Laboratory Roe, Lima 15076, Perú; jgomez@labroe.com; jcaceres@labroe.com; Lalvarado@labroe.com
[2] Arkstone Medical Solutions, Florida, 33428, USA; afrenkel@arkstonemedical.com; arendon@arkstonemedical.com
[3] Hospital Nacional Edgardo Rebagliati Martins, Lima 15073, carlos.chavez@essalud.gob.pe
[4] Universidad Nacional Mayor de San Marcos, Lima 15072, Perú; cchavezl1@unmsm.edu.pe
[5] Diagnóstico, tratamiento e investigación de enfermedades infecciosas y tropicales, Universidad Privada de Tacna, Tacna 23003, Peru, mighueda@virtual.upt.pe
[6] Universidad Ricardo Palma
* Correspondence: jgomez@labroe.com; Tel.: +51965378787

1. Introduction

Antimicrobial resistance (AMR) is one of the most pressing threats to global public health, with an estimated 4.95 million deaths associated with bacterial AMR worldwide in 2019, including 1.27 million deaths directly attributable to bacterial AMR (1). The World Health Organization has classified carbapenem-resistant Enterobacterales (CRE) as critical-priority pathogens, underscoring the urgent need for research and development of new and effective antibacterial agents(2). The emergence and diversification of Klebsiella pneumoniae carbapenemase (KPC) variants, driven by point mutations within the blaKPC gene family, have further complicated treatment protocols by conferring reduced susceptibility or resistance to multiple β-lactam–β-lactamase inhibitor combinations, including last-generation agents (3).Urinary tract infections (UTIs) represent one of the most common bacterial infections worldwide, with over 400 million incident cases estimated globally in 2019 (2). They exemplify the growing therapeutic challenge in infectious diseases, as the proportion of UTIs caused by extended-spectrum β-lactamase (ESBL)-producing Enterobacteriaceae has increased several-fold over the past decade, while carbapenem-resistant pathogens have emerged in hospital settings, collectively complicating empirical antimicrobial selection (3). Effective empirical antimicrobial therapy requires careful integration of patient-specific factors, local epidemiology, and institutional resistance patterns, balancing adequate pathogen coverage against the risk of collateral damage and selective pressure that promote antimicrobial resistance (4). The repercussions of inappropriate initial antimicrobial therapy extend beyond treatment failure, encompassing a 1.5–2-fold increase in mortality, significantly prolonged hospital length of stay, and the selection and persistence of

antimicrobial-resistant organisms. Meta-analytic evidence demonstrates that resistant infections are associated with longer hospitalization and that each additional day of antibiotic exposure increases the risk of resistant bacterial carriage, reinforcing hospitals as reservoirs for antimicrobial resistance (5–7).

Antimicrobial stewardship programs (ASPs) have become essential pillars of healthcare quality improvement, employing strategies such as prospective audit with feedback, formulary restriction, and clinical decision support systems (CDSS) to optimize prescribing practices and mitigate the spread of resistance (8,9). Nevertheless, effective implementation encounters significant obstacles, notably the limited availability of infectious disease specialists to support complex clinical decision-making (10), a structural deficiency that is particularly pronounced in resource-limited settings and smaller healthcare facilities with insufficient stewardship staffing (11). This circumstance has stimulated growing interest in the application of artificial intelligence (AI) to democratize access to expert-level antimicrobial guidance. Recent evidence indicates that advanced AI systems can achieve diagnostic and therapeutic performance approaching that of infectious disease specialists in selected tasks, while consistently outperforming non-specialist clinicians. In comparative evaluations, large language model–based systems demonstrated overall diagnostic and management accuracy of approximately **85–88%**, closely approximating infectious disease specialists (**~90%**) and exceeding resident-level performance (**~75–80%**) (12). In applied antimicrobial decision-making, machine learning–based clinical decision support systems improved appropriate empirical antibiotic coverage from **~65–70%** under routine clinical practice to **>90%**, and organism-targeted therapy accuracy to **>95%**, supporting their role in pathogen identification, resistance prediction, and treatment optimization, particularly in settings with limited access to specialist expertise (13–15). (16)

Two main artificial intelligence (AI) approaches have emerged to support antimicrobial prescribing decisions. Early applications predominantly relied on traditional machine learning algorithms trained on structured clinical and microbiological data—such as electronic health records, laboratory results, and local antibiograms—an approach grounded in the broader paradigm of big data–driven machine learning in healthcare (16) and subsequently adapted to antimicrobial prescribing and stewardship contexts (16). (17)More recently, large language models (LLMs) based on transformer architectures have enabled the interpretation of unstructured clinical narratives, including progress notes, consultation reports, and guideline text, thereby supporting clinical reasoning, antimicrobial optimization, and guideline adherence in infectious disease management (18,19). Proof-of-concept clinical evaluations further suggest that LLM-based systems can provide decision support comparable to specialist-level recommendations in selected infectious disease scenarios, highlighting their potential role in expanding access to expert antimicrobial guidance(20).

Hybrid tree-learning (HTL) algorithms, such as the Arkstone OneChoice platform, use patient demographics, comorbidities, and culture results data to generate personalized recommendations through transparent decision pathways that align with institutional protocols (21,22). In contrast, LLMs such as GPT-4, Claude, Gemini, and other modern systems are trained on extensive text corpora, including medical literature and clinical guidelines (23), enabling complex reasoning and contextually relevant responses to clinical questions (24). Their accessibility via natural language interfaces has sparked significant interest for antimicrobial stewardship, though evaluations have shown mixed

results, with notable issues like hallucinating facts, inconsistent guideline adherence, and limited incorporation of local resistance data (23).

. Previous investigations have demonstrated that machine learning–based clinical decision support systems can support antimicrobial susceptibility interpretation and improve empirical and targeted antimicrobial therapy across multiple infectious disease scenarios (25–27). However, recent systematic and narrative evaluations indicate that direct head-to-head comparisons between purpose-built ML-CDSS platforms and general-purpose large language models (LLMs) for antimicrobial decision support remain scarce, and that LLM-based systems may exhibit higher prescribing error rates, safety concerns, and inconsistent performance when applied to complex antimicrobial decision-making tasks (28,29). Consequently, the relative strengths, limitations, and appropriate clinical use cases for each AI paradigm have not been systematically characterized, representing a critical knowledge gap as healthcare systems consider AI implementation strategies (23,30). Furthermore, methodological rigor in the evaluation of medical AI systems requires independent expert adjudication of clinical correctness rather than author-determined or self-validated outcomes, as emphasized in prior assessments of AI deployment in healthcare (25,31).

This study aimed to rigorously compare the Arkstone OneChoice ML-HTL algorithm(17) with current LLMs (GPT-4 and Gemini) in recommending antimicrobial treatments for culture-confirmed UTI cases. Using an independent panel of infectious disease and microbiology experts as the benchmark, we evaluated each AI system's concordance, sensitivity, specificity, and predictive values for selecting appropriate antimicrobials. The goal was to support evidence-based AI-driven antimicrobial stewardship and clarify how specialized ML algorithms and general LLMs can complement each other in improving antimicrobial prescribing.

MATERIALS AND METHODS

Study Design and Setting

This retrospective, single-center comparative study assessed how well a machine learning hybrid tree-learning (ML-HTL) clinical decision support system (Arkstone OneChoice) aligned with two large language models—GPT-4.0 (OpenAI) and Gemini 3.0 (Google)—in recommending antimicrobial treatments for culture-confirmed urinary tract infection (UTI) episodes. Conducted at Roe Laboratory in Lima, Peru, the study analyzed clinical cases evaluated between October 21 and December 15, 2024. The study protocol received approval from the Institutional Review Board [FACSA-CEI_168-10-25], and informed consent was waived due to its retrospective design.

Case Selection and Eligibility Criteria

Inclusion Criteria

- Episodes qualified for inclusion if they met these criteria:
- • A positive urine culture showing bacterial growth of at least 100,000 colony-forming units (CFU)/mL of a uropathogen
- • An abnormal complete urinalysis consistent with urinary tract infection, such as pyuria, bacteriuria, or a positive leukocyte esterase/nitrites test
- • Availability of clinical data previously reviewed by the OneChoice ML-HTL platform

- • Complete antimicrobial susceptibility testing results for the isolated organism

Exclusion Criteria

- Episodes were excluded if:
- • Urine culture showed polymicrobial growth (≥2 organisms)
- • Clinical or microbiological data were incomplete, preventing proper case evaluation
- • The isolated organism did not have validated breakpoints for interpreting antimicrobial susceptibility

Clinical Decision Support Systems Evaluated

OneChoice ML-HTL Platform (Arkstone)

The OneChoice system is a cloud-based platform that offers clinical decision support. It uses a unique hybrid tree-learning algorithm that combines patient demographic information, clinical data, microbiological culture and susceptibility results to provide personalized antimicrobial treatment suggestions (17). The algorithm analyzes structured clinical data, such as patient age, sex, infection site, renal function, allergies and culture-specific susceptibility profiles, to prioritize treatment options based on predicted effectiveness and spectrum coverage.

Large Language Models

Two contemporary LLMs were evaluated:

GPT-4.0 (OpenAI): A transformer-based large language model with demonstrated capabilities in medical knowledge synthesis and simulated clinical reasoning tasks, as shown in multiple benchmark and scenario-based evaluations, albeit with important safety and reliability limitations (32).

Gemini 3.0 (Google): A multimodal large language model developed by Google DeepMind, accessed via the standard web interface (gemini.google.com) during the study period.

Prompt Construction and Case Presentation

For each eligible UTI episode, a standardized clinical vignette was constructed containing the following elements:

- Patient demographics (age, sex)
- Relevant clinical history and comorbidities
- Presenting symptoms and physical examination findings
- Laboratory results including complete urinalysis
- Urine culture results with organism identification
- Complete antimicrobial susceptibility profile

The clinical vignettes were presented to each LLM using a standardized prompt requesting an antimicrobial treatment recommendation with justification. The same clinical details from the OneChoice platform were included in the LLM prompts to ensure comparability. All queries were conducted during the study period, without iterative refinement or prompt optimization. All cases and prompts are available in **Supplement 1.**

Expert Panel Evaluation

Panel Composition

An independent expert panel was convened, comprising three physicians board-certified in Infectious Diseases and Tropical Medicine, with subspecialty training in

antimicrobial stewardship and the rational use of antimicrobials. Two of the panelists are based in Peru, and one is based in the United States, thereby ensuring the representation of both local clinical practices and international standards of care.

Blinded Evaluation Process

Each clinical case was systematically presented to the expert panel via a secure electronic link using a standardized format. For each case, the experts received:

1. The complete clinical vignette, identical to that provided to the AI systems.
2. Three treatment recommendations, designated as Response #1, Response #2, and Response #3.

The experts remained blind to the origin of each recommendation, leaving them unaware whether a response was generated by OneChoice, GPT-4.0, or Gemini 3.0. Each panelist independently assessed the three responses and selected the most suitable antimicrobial recommendation based on their clinical judgment, adherence to current guidelines, and consideration of the specific patient context. Communication among panelists during the evaluation process was strictly prohibited. The Expert Independent Review and Attestation Statement is available in **Supplement #2**.

Reference Standard Determination

The reference standard for each case was established through an expert panel using a majority rule criterion (≥2/3 agreement), an approach widely adopted in diagnostic accuracy studies when no single gold standard exists and multiple reasonable clinical decisions are possible (33,34). This consensus-based methodology was adopted to address clinical scenarios where multiple reasonable treatment options could be present, aligning with methodological approaches applied in previous comparative studies of clinical decision support tools.

Outcome Measures

Primary Outcome

The primary outcome was the overall concordance rate between each AI system's recommendation and the expert panel reference standard, defined as the proportion of cases in which the AI-generated recommendation matched the expert panel's consensus selection.

Secondary Outcomes

Secondary outcomes included:

- **Sensitivity**: The proportion of cases requiring specific antimicrobial therapy in which the AI system recommended appropriate treatment
- **Specificity**: The proportion of cases in which the AI system appropriately avoided unnecessary broad-spectrum coverage
- **Positive Predictive Value (PPV)**: The probability that a specific AI recommendation was concordant with expert consensus
- **Negative Predictive Value (NPV)**: The probability that a recommendation not selected by the AI was also not selected by expert consensus
- **Subgroup analyses**: Concordance stratified by organism type, resistance phenotype, and infection severity

Statistical Analysis

Concordance rates were calculated as simple proportions with 95% confidence intervals (CI) computed using the Wilson score method. Sensitivity, specificity, PPV, and NPV

were calculated using standard formulas with 95% CIs. Comparisons between AI systems were performed using McNemar's test for paired proportions, with statistical significance defined as $p<0.05$. Inter-rater reliability among expert panelists was assessed using Fleiss' kappa coefficient, with values interpreted as: <0.20 poor, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, and 0.81-1.00 almost perfect agreement (36).

All statistical analyses were performed using Stata version 17.0 (StataCorp LLC, College Station, TX) and R version 4.3.0 (R Foundation for Statistical Computing, Vienna, Austria).

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and local regulatory requirements. Given the retrospective nature of the analysis using de-identified clinical data, the Institutional Review Board waived the requirement for individual informed consent.

RESULTS

Study Population and Baseline Characteristics

A total of 88 urinary tract infection episodes meeting the inclusion criteria were analyzed during the study period. The study population was predominantly female (81.8%), with a mean age of 65.3 ± 20.9 years. Most cases (80.7%) were classified as uncomplicated urinary tract infections, whereas 19.3% were classified as complicated urinary tract infections (Table 1).

Escherichia coli was the predominant uropathogen, accounting for 76.1% of isolates, followed by Klebsiella pneumoniae (8.0%), Enterococcus species (6.8%), Proteus mirabilis (4.5%), Staphylococcus species (2.3%), Pseudomonas aeruginosa (1.1%), and Citrobacter species (1.1%). A substantial proportion of isolates exhibited antimicrobial resistance: 37.5% produced extended-spectrum β-lactamases (ESBLs), 53.4% were fluoroquinolone-resistant, and 44.3% were multidrug-resistant (MDR) (Table 1).

Table 1. Baseline Characteristics of Urinary Tract Infection Episodes (N=88)

Variable	n (%) or Mean ± SD
Demographics	
Age (years), mean ± SD	65.3 ± 20.9
Female sex	72 (81.8)
UTI Classification	
Uncomplicated	71 (80.7)
Complicated	17 (19.3)
Isolated Pathogens*	
Escherichia coli	67 (76.1)
Klebsiella pneumoniae	7 (8.0)
Enterococcus spp.	6 (6.8)
Proteus mirabilis	4 (4.5)
Staphylococcus spp.	2 (2.3)

Variable	n (%) or Mean ± SD
Pseudomonas aeruginosa	1 (1.1)
Citrobacter spp.	1 (1.1)
Antimicrobial Resistance Profile	
ESBL-producing organisms**	33 (37.5)
Fluoroquinolone resistance	47 (53.4)
MDR (≥3 antimicrobial classes)	39 (44.3)

Abbreviations: UTI, urinary tract infection; ESBL, extended-spectrum β-lactamase; MDR, multidrug-resistant; SD, standard deviation.
*Pathogens identified from urine culture with significant growth.
**ESBL-producing organisms were defined as isolates demonstrating resistance to ≥2 third-generation cephalosporins (ceftriaxone, cefotaxime, ceftazidime, or cefixime).

Diagnostic Performance of AI Systems

The OneChoice ML-HTL platform demonstrated significantly superior diagnostic performance compared to both LLMs (**Table 2**, **Figure 1**). OneChoice achieved an overall concordance rate of 90.9% (95% CI: 83.1–95.3) with the expert panel consensus, compared to 63.6% (95% CI: 53.2–72.9) for GPT-4 and 37.5% (95% CI: 28.1–47.9) for Gemini.

Table 2. Diagnostic Performance of AI Systems Compared with Expert Panel Gold Standard (N=88)

AI System	Concordance, % (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, %	NPV, % (95% CI)	κ (95% CI)
OneChoice (ML-HTL)	90.9 (83.1–95.3)	87.5 (77.2–93.5)	100.0 (86.2–100.0)	100.0	75.0 (57.9–86.7)	0.89 (0.89–0.89)
GPT-4	63.6 (53.2–72.9)	50.0 (38.1–61.9)	100.0 (86.2–100.0)	100.0	42.9 (30.8–55.9)	0.32 (0.10–0.54)
Gemini	37.5 (28.1–47.9)	14.1 (7.6–24.6)	100.0 (86.2–100.0)	100.0	30.4 (21.3–41.2)	−0.18 (−0.41–0.06)

Abbreviations: CI, confidence interval; ML-HTL, machine learning hybrid tree-learning; PPV, positive predictive value; NPV, negative predictive value; κ, Gwet's AC1 agreement coefficient.

Figure 1a: Expert-Defined Discordance Evaluation Between OneChoice and Large Language Models for Urinary Tract Infection Treatment Recommendation

Antimicrobial Stewardship Assessment (n=87 clinical scenarios)

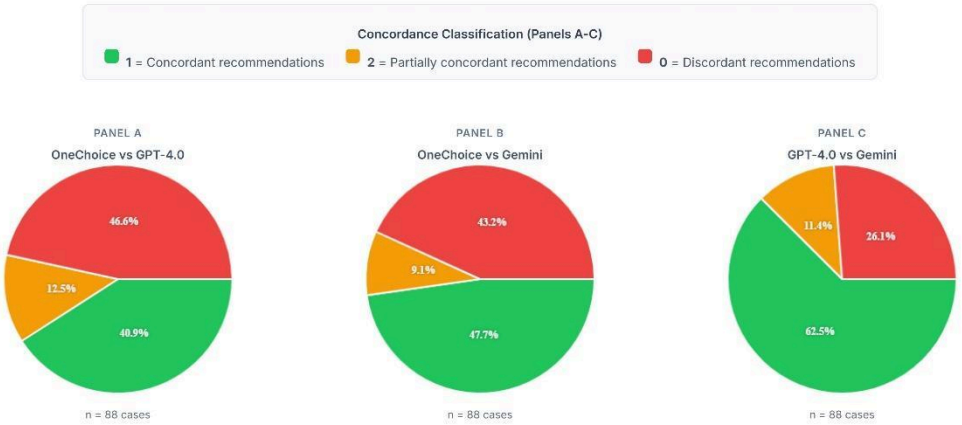


Figure 1b: Expert Adjudication of Discordant & Partially Concordant Cases

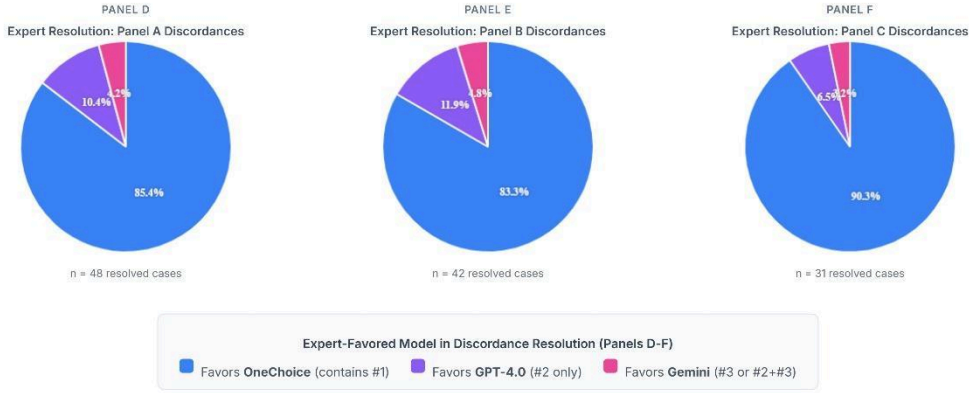
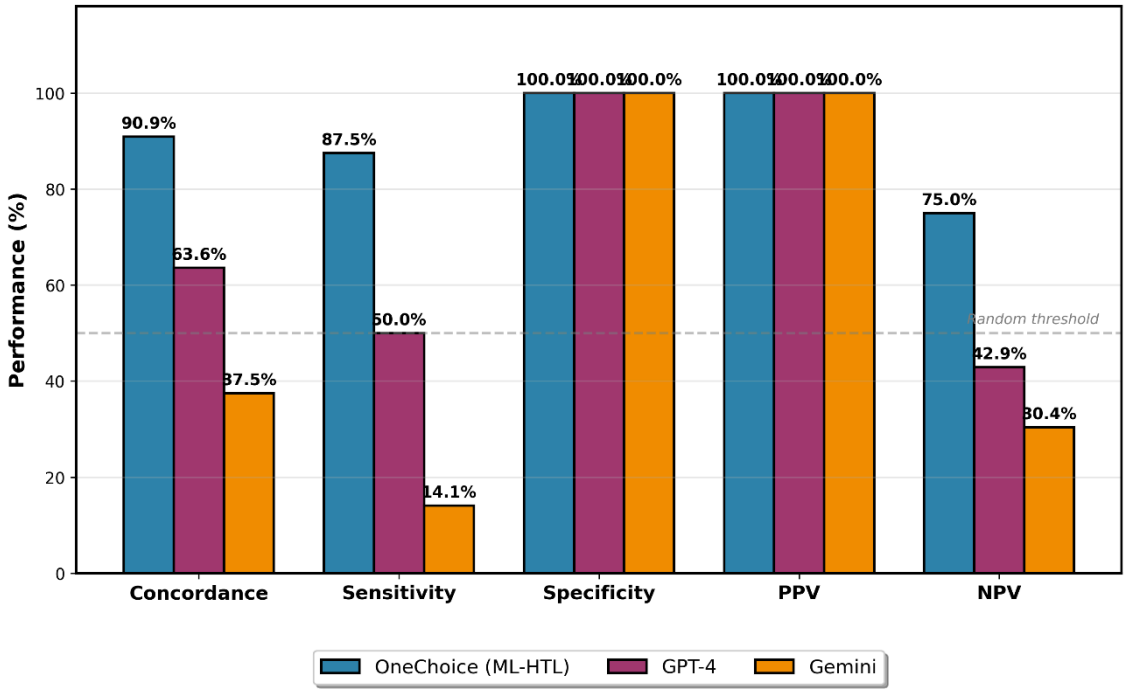


Figure 1a. Concordance analysis between OneChoice clinical decision support system and Large Language Models (GPT-4.0 and Gemini) for antimicrobial recommendations in urinary tract infection cases. **Upper panels (A–C)** display pairwise concordance distributions, with discordant (0) and partially concordant (2) segments positioned at the bottom. **Figure 1b. Lower panels (D–F)** show expert adjudication outcomes specifically for the discordant and partially concordant cases from each corresponding upper panel. Expert resolution categories: "Favors OneChoice" includes any response containing #1 (alone or combined); "Favors GPT-4.0" includes #2 exclusively; "Favors Gemini" includes #3 or #2+#3 without #1.

Sensitivity analysis, calculated from 64 cases requiring expert adjudication due to discordant recommendations, revealed marked differences among systems: OneChoice demonstrated 87.5% sensitivity (95% CI: 77.2–93.5), whereas GPT-4 achieved only 50.0% (95% CI: 38.1–61.9) and Gemini 14.1% (95% CI: 7.6–24.6). All three AI systems exhibited 100% specificity (95% CI: 86.2–100.0), based on 24 cases in which all systems provided concordant "not necessary" recommendations, and consequently maintained perfect positive predictive values (100%). Negative predictive values were 75.0% (95% CI: 57.9–86.7) for OneChoice, 42.9% (95% CI: 30.8–55.9) for GPT-4, and 30.4% (95% CI: 21.3–41.2) for Gemini.(**Figure 2**)

Figure 2: Comparative diagnostic performance of AI Systems vs Expert Panle Gold Standard



Agreement analysis using Gwet's AC1 coefficient demonstrated excellent agreement between OneChoice and the expert panel (κ = 0.89; 95% CI: 0.89–0.89), fair agreement for GPT-4 (κ = 0.32; 95% CI: 0.10–0.54), and poor agreement for Gemini (κ = −0.18; 95% CI: −0.41–0.06), indicating agreement worse than chance (**Figure 3**).

Stratified Analysis by Case Complexity

Performance patterns varied substantially according to case complexity (**Table 4**). For uncomplicated UTI (n=71), OneChoice maintained excellent concordance (95.8%), significantly outperforming GPT-4 (69.0%) and Gemini (39.4%) (Cochran's Q test, p<0.001). In complicated UTI cases (n=17), all systems showed diminished performance (OneChoice: 70.6%, GPT-4: 41.2%, Gemini: 29.4%), although differences did not reach statistical significance (p=0.074).

Table 4. Concordance Stratified by Case Complexity

Subgroup	n	OneChoice (%)	GPT-4 (%)	Gemini (%)	p-value*
UTI Type					
Uncomplicated	71	95.8	69.0	39.4	<0.001
Complicated	17	70.6	41.2	29.4	0.074
Resistance Profile					
Susceptible	45	93.3	77.8	28.9	<0.001
ESBL+	33	84.8	48.5	48.5	0.002
MDR	39	89.7	51.3	51.3	<0.001
Pathogen					
E. coli	67	95.5	70.1	46.3	<0.001

Subgroup	n	OneChoice (%)	GPT-4 (%)	Gemini (%)	p-value*
Non-*E. coli*	21	76.2	42.9	9.5	<0.001

Abbreviations: UTI, urinary tract infection; ESBL, extended-spectrum β-lactamase; MDR, multidrug-resistant.
*Cochran's Q test comparing concordance rates across the three AI systems within each subgroup.

Stratification by antimicrobial resistance profile revealed consistent superiority of the ML-HTL algorithm. For susceptible isolates (n=45), concordance rates were 93.3% for OneChoice, 77.8% for GPT-4, and 28.9% for Gemini (p<0.001). Performance declined across all systems when evaluating ESBL-producing (n=33) and MDR organisms (n=39); however, OneChoice maintained high concordance (ESBL+: 84.8%; MDR: 89.7%), substantially exceeding that of GPT-4 (ESBL+: 48.5%; MDR: 51.3%) and Gemini (ESBL+: 48.5%; MDR: 51.3%), with statistically significant differences (p=0.002 and p<0.001, respectively).
Pathogen-specific analysis demonstrated that OneChoice achieved 95.5% concordance for *E. coli* infections (n=67), significantly higher than GPT-4 (70.1%) and Gemini (46.3%) (p<0.001). For non-*E. coli* pathogens (n=21), OneChoice concordance was 76.2%, compared to 42.9% for GPT-4 and 9.5% for Gemini (p<0.001).

Inter-rater Agreement Among Expert Panel
Pairwise agreement among the three infectious disease experts demonstrated slight concordance, with Cohen's κ values ranging from 0.062 to 0.147 (**Table 5**, Supplementary). Global inter-rater reliability assessed by Fleiss' κ was −0.004 (95% CI: −0.069–0.061), indicating agreement at chance level. These findings reflect the inherent complexity of antimicrobial treatment decisions, wherein multiple therapeutic options may be clinically acceptable. Notably, experts frequently indicated multiple acceptable model combinations during adjudication. Despite low quantitative inter-rater reliability, consensus was ultimately achieved through majority voting and structured discussion.

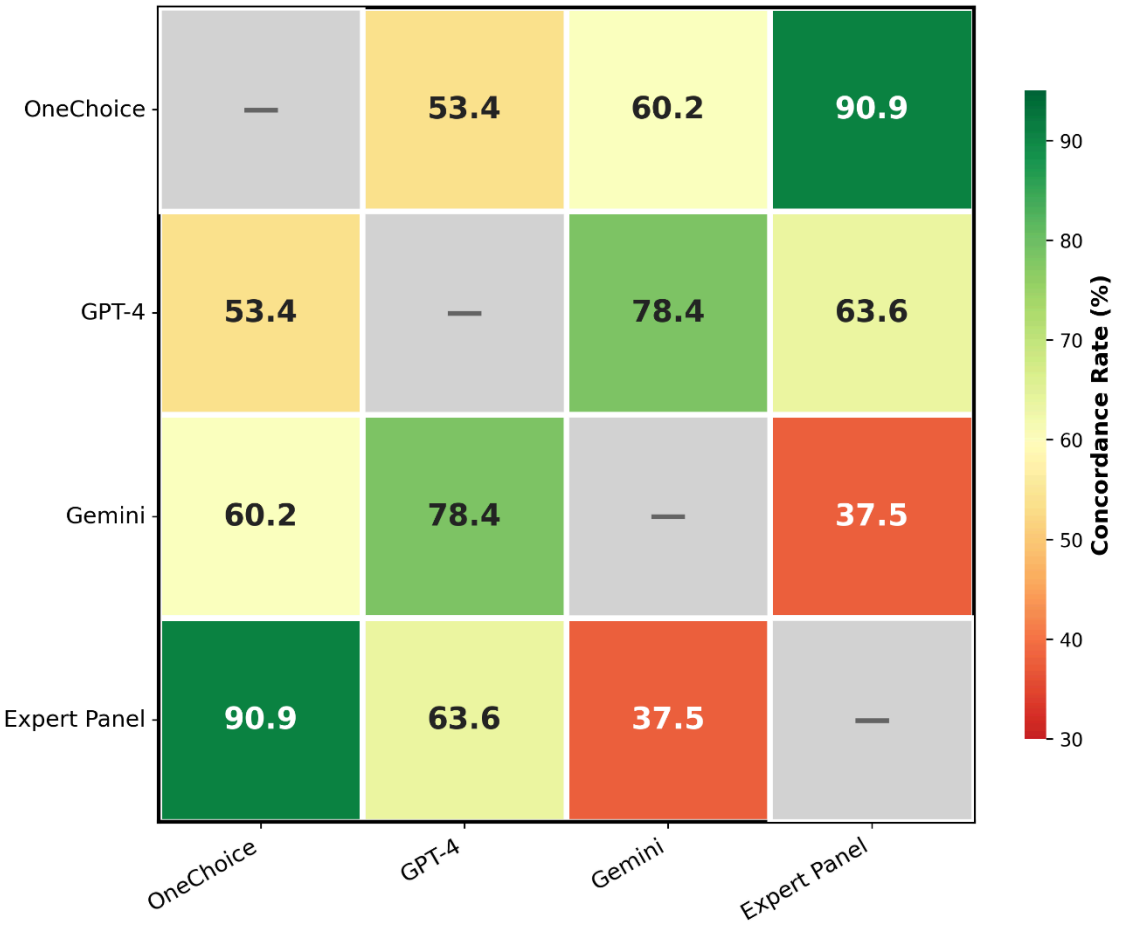
Table 5 (Supplementary). Inter-rater Agreement Between Experts

Expert Pair	Cohen's κ	95% CI	Interpretation
Expert 1 vs Expert 2	0.085	−0.001–0.170	Slight
Expert 1 vs Expert 3	0.147	0.051–0.245	Slight
Expert 2 vs Expert 3	0.062	0.005–0.122	Slight
Fleiss' κ (global)	**−0.004**	**−0.069–0.061**	**Poor**

Abbreviations: κ, kappa coefficient; CI, confidence interval.

Finally, a heatmap reveals a striking pattern: while GPT-4 and Gemini demonstrate high inter-agreement (78.4%), suggesting similar reasoning patterns, both LLMs show substantially lower concordance with the expert panel compared to OneChoice. This finding indicates that current general-purpose large language models, despite their conversational sophistication, may not be optimally suited for antimicrobial stewardship decisions without domain-specific training or fine-tuning (Figure 4).

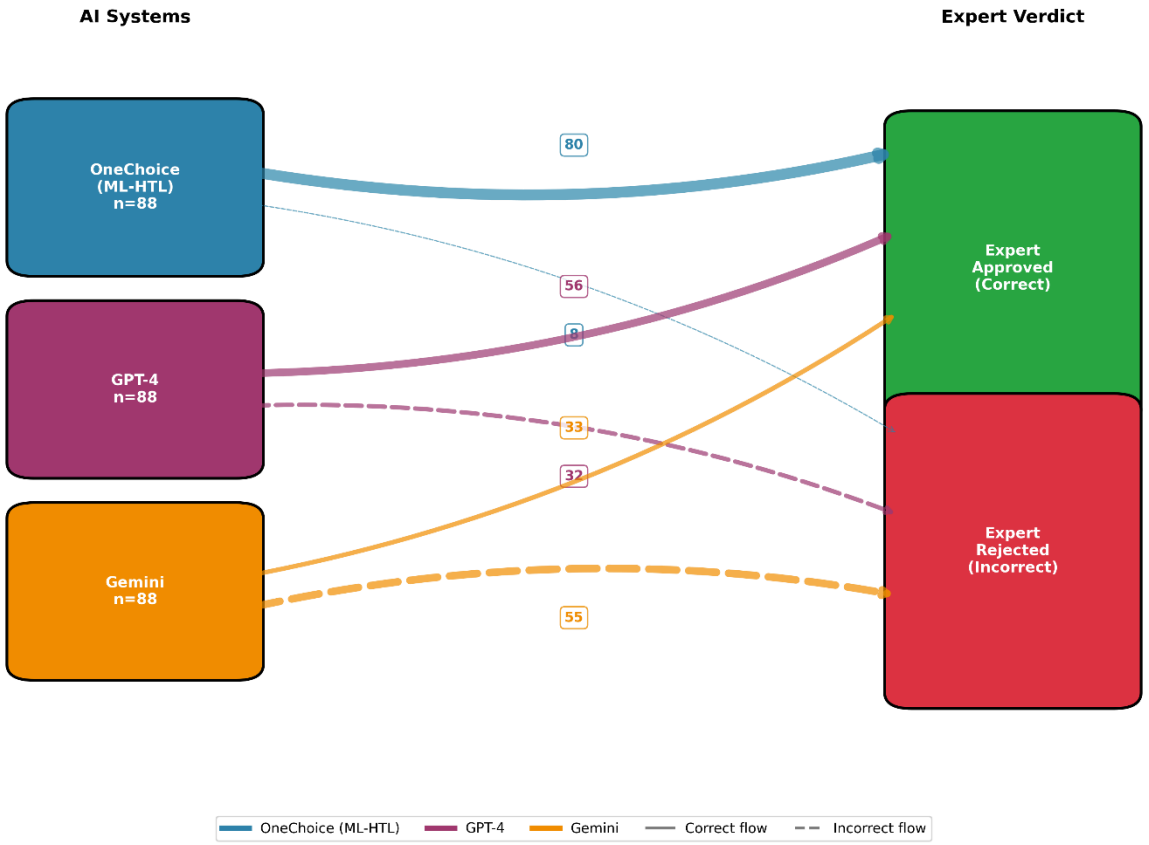
Figure 4. Pairwise concordance Heatmap between AI Systems and Expert Panel



The comparative performance of AI-driven recommendation systems is depicted in **Figure 5** through an alluvial diagram that visualizes the flow of therapeutic recommendations to expert panel adjudication. OneChoice (ML-HTL) demonstrated superior concordance with expert consensus, with 80 of 88 recommendations (90.9%) approved by the infectious disease specialist panel. In contrast, GPT-4 achieved moderate agreement, with 56 correct recommendations (63.6%) and 32 rejections (36.4%). Gemini exhibited the lowest performance, with only 33 recommendations (37.5%) endorsed by experts and 55 (62.5%) deemed inappropriate.

The visual representation underscores the substantial performance gap between the domain-specific machine learning model (OneChoice ML-HTL) and general-purpose large language models (GPT-4 and Gemini), highlighting the critical importance of specialized training and integration of local antimicrobial resistance patterns in clinical decision support systems for bacteremia management.

Figure 5. Alluvial diagram illustrating the flow of antibiotic recommendations from three artificial intelligence systems to expert panel verdict.



The left nodes represent the AI systems evaluated: OneChoice (ML-HTL), GPT-4, and Gemini, each assessed across 88 bacteremia cases. The right nodes indicate the expert panel's final verdict, categorized as "Approved" (correct) or "Rejected" (incorrect). Flow width is proportional to the number of cases. Solid lines represent correctly classified recommendations; dashed lines represent incorrect recommendations. OneChoice (ML-HTL) achieved the highest concordance with expert opinion (80/88, 90.9%), followed by GPT-4 (56/88, 63.6%), while Gemini demonstrated the lowest agreement (33/88, 37.5%)

DISCUSSION

This study represents the first head-to-head comparison of a domain-specific machine learning clinical decision support system against general-purpose large language models for antimicrobial treatment recommendations in culture-confirmed urinary tract infections. Our findings demonstrate a marked performance disparity favoring the purpose-built ML-HTL algorithm, with OneChoice achieving 90.9% concordance with expert consensus compared to 63.6% for GPT-4 and 37.5% for Gemini. These results have significant implications for the evidence-based implementation of artificial intelligence in antimicrobial stewardship programs.

The superior performance of OneChoice aligns with emerging evidence supporting the clinical utility of purpose-built ML-CDSS in infectious disease management. Peiffer-Smadja et al. (35), in their comprehensive narrative review of 60 ML-CDSS applications in infectious diseases, highlighted that domain-specific systems incorporating structured clinical data consistently demonstrated robust diagnostic and

therapeutic accuracy. Our findings extend this evidence by directly comparing such systems against LLMs in a real-world clinical context.

The concordance rate of 90.9% achieved by OneChoice exceeds the performance thresholds reported in previous ML-CDSS validation studies. Frenkel et al. (18) demonstrated that the Arkstone ML system achieved 100% accuracy in distinguishing trained from novel data and produced recommendations with no major discrepancies in 84.47% of cases during internal validation. Our external validation using an independent expert panel corroborates these findings and strengthens the evidence base for clinical deployment of such systems.

Conversely, the performance of LLMs in our study confirms concerns raised by Schwartz et al. (33) regarding their application in infectious diseases consultation. Their "Black Box Warning" viewpoint emphasized that LLMs currently exhibit frequent confabulations, lack contextual awareness crucial for nuanced diagnostic and treatment plans, and have inscrutable training data and methods. Our finding that 62.5% of GPT-4 discordances and 63.6% of Gemini discordances were attributable to inappropriate antibiotic selection—including spectrum errors and antibiotics not indicated for isolated pathogens—empirically validates these theoretical concerns.

Several factors may explain the substantial performance gap observed between ML-HTL and LLM systems. First, OneChoice integrates structured data including patient demographics, clinical parameters, microbiological culture and susceptibility results, and institutional antibiogram patterns through a purpose-built hybrid tree-learning algorithm (17). This architecture enables the system to provide recommendations grounded in local epidemiology and resistance patterns—a critical requirement emphasized by antimicrobial stewardship guidelines (8,9).

(36)In contrast, LLMs such as GPT-4 and Gemini are trained on vast text corpora that, while comprehensive in breadth, may not adequately reflect local antimicrobial resistance trends or institutional prescribing practices (24,25). As Schwartz et al. (33) noted, LLMs lack the contextual awareness essential in infectious diseases, where therapeutic approaches must consider local epidemiology, antimicrobial resistance patterns, and assay and drug availability. The high prevalence of ESBL-producing organisms (37.5%), fluoroquinolone resistance (53.4%), and multidrug-resistant pathogens (44.3%) in our cohort likely amplified this limitation, as evidenced by the significant decline in LLM performance for resistant phenotypes.

Second, the error pattern analysis revealed qualitative differences in failure modes. The predominance of "inappropriate antibiotic selection" errors among LLMs suggests difficulty in appropriately weighing microbiological susceptibility data against clinical guidelines. Howard et al. (38) previously demonstrated that ChatGPT exhibited "deficits in situational awareness, inference, and consistency" when responding to infectious disease curbside consults, findings consistent with our observations. In contrast, OneChoice errors were more evenly distributed between inappropriate selection and dosing/interval errors, suggesting that its failure modes are more amenable to targeted refinement.

Our findings have immediate relevance for healthcare systems considering AI implementation for antimicrobial stewardship. The Infectious Diseases Society of America has emphasized that antimicrobial stewardship programs face significant implementation barriers, particularly the limited availability of infectious disease specialists (8,11). In 2017, 79.5% of US counties lacked a single ID physician, and 208 million people lived in counties with no or fewer than average ID physicians (33).

Similarly, Fabre et al. (39) documented substantial gaps in antimicrobial stewardship infrastructure across Latin America, where resource constraints limit access to expert consultation.

The ML-HTL platform's high concordance rate (90.9%) and excellent agreement coefficient (κ = 0.89) suggest that such systems can effectively extend antimicrobial stewardship expertise to settings lacking specialist support. Gomez de la Torre et al. (40) previously demonstrated that AI-powered CDSS using molecular data could deliver therapeutic recommendations approximately 28-29 hours faster than conventional phenotypic approaches while maintaining 80.34% concordance with expert consensus. Our findings complement this evidence by demonstrating superior accuracy compared to general-purpose LLMs.

However, our results also counsel caution regarding the premature deployment of LLMs for clinical antimicrobial decision-making. The 62.5% discordance rate observed with Gemini—indicating agreement worse than chance (κ = −0.18)—underscores that current general-purpose LLMs are not suitable for independent antimicrobial recommendation without substantial human oversight. Lee et al. (37) suggested that GPT-4 might be used for "curbside consults," but our data indicate that such use should be approached with considerable caution, particularly for complex resistant infections.

The stratified analysis revealed important patterns regarding the differential performance of AI systems across clinical scenarios. OneChoice maintained consistently high concordance across all subgroups, whereas LLM performance degraded significantly with increasing case complexity. For uncomplicated UTI, the difference between OneChoice (95.8%) and GPT-4 (69.0%) was statistically significant ($p<0.001$), but for complicated UTI, the reduced sample size (n=17) likely contributed to the non-significant difference ($p=0.074$) despite absolute concordance differences of similar magnitude.

The performance decline observed for ESBL-positive and MDR infections is particularly concerning from a clinical standpoint. Tamma et al. (38) emphasized that treatment of antimicrobial-resistant gram-negative infections requires nuanced decision-making that accounts for specific resistance mechanisms, available therapeutic options, and pharmacokinetic/pharmacodynamic considerations. The equivalent performance of GPT-4 and Gemini for ESBL+ and MDR cases (both 48.5% and 51.3%, respectively) suggests that general-purpose LLMs may reach a performance floor when encountering resistance phenotypes that require specialized knowledge integration.

The superior performance of OneChoice for *E. coli* infections (95.5%) compared to non-*E. coli* pathogens (76.2%) aligns with the predominance of this organism in UTI and its well-characterized resistance patterns. However, the precipitous decline in Gemini's performance for non-*E. coli* pathogens (9.5%) highlights the fragility of LLM recommendations when encountering less common clinical scenarios—a limitation that has significant implications for healthcare equity if such systems were deployed in settings with diverse pathogen epidemiology.

Expert Panel Agreement and Reference Standard Validity

The low inter-rater reliability observed among expert panelists (Fleiss' κ = −0.004) warrants careful interpretation. Rather than undermining our findings, this observation reflects the inherent complexity of antimicrobial treatment decisions, where multiple therapeutic options may be clinically acceptable. Landis and Koch (36) established that κ values below 0.20 indicate slight agreement, but in clinical contexts where equivalent

options exist, low kappa values may paradoxically reflect appropriate therapeutic equipoise rather than measurement error.
Importantly, consensus was achieved through majority voting, providing a robust reference standard despite individual variation. The use of Gwet's AC1 coefficient for AI-expert agreement addressed the paradox whereby high prevalence of correct classifications can artifactually lower traditional κ values (42). The strong AC1 coefficient for OneChoice (0.89) compared to GPT-4 (0.32) and Gemini (−0.18) demonstrates meaningful performance differences that would be obscured by chance-corrected statistics alone.

Several limitations should be considered when interpreting our findings. First, this was a single-center study conducted at a reference laboratory in Lima, Peru, which may limit generalizability to other geographic settings with different resistance epidemiology. However, the high prevalence of resistant phenotypes in our cohort likely provides a more stringent evaluation context than settings with lower resistance rates.
Second, our evaluation included only two LLMs (GPT-4 and Gemini), whereas the study objective mentioned five models (GPT-4, Claude, Gemini, DeepSeek, and Perplexity). Resource constraints precluded evaluation of all initially planned LLMs, and future studies should assess additional models to provide a more comprehensive landscape of LLM performance.
Third, the LLM prompts were standardized without iterative refinement or prompt optimization, which may have underestimated achievable LLM performance. However, this approach reflects realistic clinical deployment scenarios where clinicians would unlikely engage in extensive prompt engineering for routine antimicrobial recommendations.
Fourth, our study evaluated AI recommendations against expert consensus rather than clinical outcomes. While concordance with expert opinion is a validated surrogate endpoint (33), future studies should assess the impact of AI-guided prescribing on patient outcomes including treatment success, length of stay, and mortality.
Finally, the retrospective design and focus on UTI limit the breadth of our conclusions. Prospective evaluation across diverse infection types and clinical settings will be essential to establish the generalizability and clinical utility of these findings.

Our findings suggest several priorities for future research. First, randomized controlled trials comparing ML-CDSS-guided prescribing against standard care are needed to establish clinical effectiveness beyond diagnostic accuracy. Second, hybrid approaches integrating the interpretive strengths of LLMs with the structured decision logic of ML-HTL systems may offer synergistic benefits (40). Third, continuous validation and updating of AI systems against evolving resistance patterns will be essential to maintain clinical relevance. Finally, implementation research addressing barriers to AI-CDSS adoption in resource-limited settings should be prioritized given the greatest potential impact in such contexts (35,39).

CONCLUSIONS

In this head-to-head comparison of AI systems for antimicrobial treatment recommendations in culture-confirmed UTI, the domain-specific OneChoice ML-HTL platform significantly outperformed general-purpose LLMs, achieving 90.9% concordance with expert consensus compared to 63.6% for GPT-4 and 37.5% for Gemini. The ML-HTL algorithm demonstrated consistent superiority across clinical

subgroups, including resistant phenotypes where LLM performance declined substantially. These findings support the clinical utility of purpose-built ML-CDSS for antimicrobial stewardship while counseling caution regarding the premature deployment of current-generation LLMs for independent antimicrobial decision support.

REFERENCES

1. Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet [Internet]. 2022 Feb 12 [cited 2026 Jan 16];399(10325):629–55. Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02724-0/fulltext

2. Zeng Z, Zhan J, Zhang K, Chen H, Cheng S. Global, regional, and national burden of urinary tract infections from 1990 to 2019: an analysis of the global burden of disease study 2019. World J Urol [Internet]. 2022 Mar 1 [cited 2026 Jan 16];40(3):755–63. Available from: https://doi.org/10.1007/s00345-021-03913-0

3. Zilberberg MD, Shorr AF. Secular Trends in Gram-Negative Resistance among Urinary Tract Infection Hospitalizations in the United States, 2000–2009. Infect Control Hosp Epidemiol [Internet]. 2013 Sept [cited 2026 Jan 16];34(9):940–6. Available from: https://www.cambridge.org/core/journals/infection-control-and-hospital-epidemiology/article/abs/secular-trends-in-gramnegative-resistance-among-urinary-tract-infection-hospitalizations-in-the-united-states-20002009/71FAEBE6F1CC61C7FC04C9F3511F6B19

4. Gupta K, Hooton TM, Naber KG, Wullt B, Colgan R, Miller LG, et al. International Clinical Practice Guidelines for the Treatment of Acute Uncomplicated Cystitis and Pyelonephritis in Women: A 2010 Update by the Infectious Diseases Society of America and the European Society for Microbiology and Infectious Diseases. Clin Infect Dis [Internet]. 2011 Mar 1 [cited 2026 Jan 16];52(5):e103–20. Available from: https://doi.org/10.1093/cid/ciq257

5. Allel K, Stone J, Undurraga EA, Day L, Moore CE, Lin L, et al. The impact of inpatient bloodstream infections caused by antibiotic-resistant bacteria in low- and middle-income countries: A systematic review and meta-analysis. PLOS Med [Internet]. 2023 June 22 [cited 2026 Jan 16];20(6):e1004199. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1004199

6. Mo Y, Oonsivilai M, Lim C, Niehus R, Cooper BS. Implications of reducing antibiotic treatment duration for antimicrobial resistance in hospital settings: A modelling study and meta-analysis. PLoS Med. 2023 June;20(6):e1004013.

7. George NA, Pan D, Silva L, Baggaley RF, Irizar P, Divall P, et al. The prevalence and risk of mortality associated with antimicrobial resistance within nosocomial settings—a global systematic review and meta-analysis of

over 20,000 patients. eClinicalMedicine [Internet]. 2025 Sept 1 [cited 2026 Jan 16];87. Available from: https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(25)00316-5/fulltext

8. Barlam TF, Cosgrove SE, Abbo LM, MacDougall C, Schuetz AN, Septimus EJ, et al. Implementing an Antibiotic Stewardship Program: Guidelines by the Infectious Diseases Society of America and the Society for Healthcare Epidemiology of America. Clin Infect Dis [Internet]. 2016 May 15 [cited 2026 Jan 16];62(10):e51–77. Available from: https://dx.doi.org/10.1093/cid/ciw118

9. Dellit TH, Owens RC, McGowan JE, Gerding DN, Weinstein RA, Burke JP, et al. Infectious Diseases Society of America and the Society for Healthcare Epidemiology of America Guidelines for Developing an Institutional Program to Enhance Antimicrobial Stewardship. Clin Infect Dis [Internet]. 2007 Jan 15 [cited 2026 Jan 16];44(2):159–77. Available from: https://doi.org/10.1086/510393

10. Rhee C, Chiotos K, Cosgrove SE, Heil EL, Kadri SS, Kalil AC, et al. Infectious Diseases Society of America Position Paper: Recommended Revisions to the National Severe Sepsis and Septic Shock Early Management Bundle (SEP-1) Sepsis Quality Measure. Clin Infect Dis [Internet]. 2021 Feb 15 [cited 2026 Jan 16];72(4):541–52. Available from: https://doi.org/10.1093/cid/ciaa059

11. Doernberg SB, Abbo LM, Burdette SD, Fishman NO, Goodman EL, Kravitz GR, et al. Essential Resources and Strategies for Antibiotic Stewardship Programs in the Acute Care Setting. Clin Infect Dis [Internet]. 2018 Sept 28 [cited 2026 Jan 16];67(8):1168–74. Available from: https://doi.org/10.1093/cid/ciy255

12. Zhan L, Dang X, Xie Z, Zeng C, Wu W, Zhang X, et al. Evaluating GPT-4o in infectious disease diagnostics and management: A comparative study with residents and specialists on accuracy, completeness, and clinical support potential. Digit Health. 2025;11:20552076251355797.

13. Rawson TM, Moore LSP, Zhu N, Ranganathan N, Skolimowska K, Gilchrist M, et al. Bacterial and Fungal Coinfection in Individuals With Coronavirus: A Rapid Review To Support COVID-19 Antimicrobial Prescribing. Clin Infect Dis [Internet]. 2020 Nov 1 [cited 2026 Jan 16];71(9):2459–68. Available from: https://doi.org/10.1093/cid/ciaa530

14. Tejeda MI, Fernández J, Valledor P, Almirall C, Barberán J, Romero-Brufau S. Retrospective validation study of a machine learning-based software for empirical and organism-targeted antibiotic therapy selection. Antimicrob Agents Chemother. 2024 Oct 8;68(10):e0077724.

15. Al Kuwaiti A, Nazer K, Al-Reedy A, Al-Shehri S, Al-Muhanna A, Subbarayalu AV, et al. A Review of the Role of Artificial Intelligence in Healthcare. J Pers Med [Internet]. 2023 June [cited 2026 Jan 16];13(6):951. Available from: https://www.mdpi.com/2075-4426/13/6/951

16. Giacobbe DR, Marelli C, Guastavino S, Signori A, Mora S, Rosso N, et al. Artificial intelligence and prescription of antibiotic therapy: present and future. Expert Rev Anti Infect Ther. 2024 Oct;22(10):819–33.

17. Frenkel A, Rendon A, Chavez-Lencinas C, Gomez De la Torre JC, MacDermott J, Gross C, et al. Internal Validation of a Machine Learning-Based CDSS for Antimicrobial Stewardship. Life [Internet]. 2025 July [cited 2026 Jan 16];15(7):1123. Available from: https://www.mdpi.com/2075-1729/15/7/1123

18. Cho HN, Jun TJ, Kim YH, Kang H, Ahn I, Gwon H, et al. Task-Specific Transformer-Based Language Models in Health Care: Scoping Review. JMIR Med Inform. 2024 Nov 18;12:e49724.

19. Omar M, Brin D, Glicksberg B, Klang E. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review. Am J Infect Control [Internet]. 2024 Sept 1 [cited 2026 Jan 16];52(9):992–1001. Available from: https://www.sciencedirect.com/science/article/pii/S0196655324001597

20. Lorenzoni G, Garbin A, Brigiari G, Papappicco CAM, Manfrin V, Gregori D. Large Language Models in Action: Supporting Clinical Evaluation in an Infectious Disease Unit. Healthcare [Internet]. 2025 Jan [cited 2026 Jan 16];13(8):879. Available from: https://www.mdpi.com/2227-9032/13/8/879

21. Beaudoin M, Kabanza F, Nault V, Valiquette L. Evaluation of a machine learning capability for a clinical decision support system to enhance antimicrobial stewardship programs. Artif Intell Med [Internet]. 2016 Mar 1 [cited 2026 Jan 16];68:29–36. Available from: https://www.sciencedirect.com/science/article/pii/S0933365716300574

22. Kullar R, Goff DA, Schulz LT, Fox BC, Rose WE. The “Epic” Challenge of Optimizing Antimicrobial Stewardship: The Role of Electronic Medical Records and Technology. Clin Infect Dis [Internet]. 2013 Oct 1 [cited 2026 Jan 16];57(7):1005–13. Available from: https://dx.doi.org/10.1093/cid/cit318

23. Han J, Qiu W, Lichtfouse E. ChatGPT in Scientific Research and Writing: A Beginner’s Guide. 2024.

24. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med [Internet]. 2023 Aug [cited 2026 Jan 16];29(8):1930–40. Available from: https://www.nature.com/articles/s41591-023-02448-8

25. Mello MM, Guha N. ChatGPT and Physicians’ Malpractice Risk. JAMA Health Forum [Internet]. 2023 May 18 [cited 2026 Jan 16];4(5):e231938. Available from: https://doi.org/10.1001/jamahealthforum.2023.1938

26. Ardila CM, González-Arroyave D, Tobón S. Machine learning for predicting antimicrobial resistance in critical and high-priority pathogens: A systematic review considering antimicrobial susceptibility tests in real-world healthcare settings. PLOS ONE [Internet]. 2025 Feb 25 [cited 2026 Jan

16];20(2):e0319460. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0319460

27. Giacobbe DR, Marelli C, Guastavino S, Signori A, Mora S, Rosso N, et al. Artificial intelligence and prescription of antibiotic therapy: present and future. Expert Rev Anti Infect Ther. 2024 Oct;22(10):819–33.

28. AlGain S, Marra AR, Kobayashi T, Marra PS, Celeghini PD, Hsieh MK, et al. Can we rely on artificial intelligence to guide antimicrobial therapy? A systematic literature review. Antimicrob Steward Healthc Epidemiol [Internet]. 2025 Jan [cited 2026 Jan 16];5(1):e90. Available from: https://www.cambridge.org/core/journals/antimicrobial-stewardship-and-healthcare-epidemiology/article/can-we-rely-on-artificial-intelligence-to-guide-antimicrobial-therapy-a-systematic-literature-review/8239BEF5A37E8747203593A2D6C99DAE

29. Bienvenu AL, Ducrocq JM, Augé-Caumon MJ, Baseilhac E. Clinical decision support system to guide antimicrobial selection: a narrative review from 2019 to 2023. J Hosp Infect [Internet]. 2025 Aug 1 [cited 2026 Jan 16];162:140–52. Available from: https://www.journalofhospitalinfection.com/article/S0195-6701(25)00139-2/abstract

30. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature [Internet]. 2023 Aug [cited 2026 Jan 16];620(7972):172–80. Available from: https://www.nature.com/articles/s41586-023-06291-2

31. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med [Internet]. 2019 Jan [cited 2026 Jan 16];25(1):44–56. Available from: https://www.nature.com/articles/s41591-018-0300-7

32. Schwartz IS, Link KE, Daneshjou R, Cortés-Penfield N. Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation. Clin Infect Dis [Internet]. 2024 Apr 15 [cited 2026 Jan 16];78(4):860–6. Available from: https://doi.org/10.1093/cid/ciad633

33. Kellerhuis BE, Jenniskens K, Kusters MPT, Schuit E, Hooft L, Moons KGM, et al. Expert panel as reference standard procedure in diagnostic accuracy studies: a systematic scoping review and methodological guidance. Diagn Progn Res [Internet]. 2025 May 13 [cited 2026 Jan 16];9(1):12. Available from: https://doi.org/10.1186/s41512-025-00195-7

34. Kea B, Sun BCA. Consensus development for healthcare professionals. Intern Emerg Med [Internet]. 2015 Apr 1 [cited 2026 Jan 16];10(3):373–83. Available from: https://doi.org/10.1007/s11739-014-1156-6

35. Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure FX, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin Microbiol Infect

[Internet]. 2020 May 1 [cited 2026 Jan 16];26(5):584–95. Available from: https://www.clinicalmicrobiologyandinfection.org/article/S1198-743X(19)30494-X/fulltext

36. Gomez de la Torre JC, Frenkel A, Chavez-Lencinas C, Rendon A, Cáceres JA, Alvarado L, et al. AI-Based Treatment Recommendations Enhance Speed and Accuracy in Bacteremia Management: A Comparative Study of Molecular and Phenotypic Data. Life [Internet]. 2025 June [cited 2026 Jan 16];15(6):864. Available from: https://www.mdpi.com/2075-1729/15/6/864

37. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med [Internet]. 2023 Mar 29 [cited 2026 Jan 16];388(13):1233–9. Available from: https://www.nejm.org/doi/full/10.1056/NEJMsr2214184

38. Tamma PD, Aitken SL, Bonomo RA, Mathers AJ, van Duin D, Clancy CJ. Infectious Diseases Society of America 2023 Guidance on the Treatment of Antimicrobial Resistant Gram-Negative Infections. Clin Infect Dis [Internet]. 2023 July 18 [cited 2026 Jan 16];ciad428. Available from: https://doi.org/10.1093/cid/ciad428

39. Fabre V, Cosgrove SE, Secaira C, Torrez JCT, Lessa FC, Patel TS, et al. Antimicrobial stewardship in Latin America: Past, present, and future. Antimicrob Steward Healthc Epidemiol [Internet]. 2022 Jan [cited 2026 Jan 16];2(1):e68. Available from: https://www.cambridge.org/core/journals/antimicrobial-stewardship-and-healthcare-epidemiology/article/antimicrobial-stewardship-in-latin-america-past-present-and-future/E15B75887032174E79E12C5EB51897E4

40. Cesaro A, Hoffman SC, Das P, de la Fuente-Nunez C. Challenges and applications of artificial intelligence in infectious diseases and antimicrobial resistance. Npj Antimicrob Resist [Internet]. 2025 Jan 7 [cited 2026 Jan 16];3(1):2. Available from: https://www.nature.com/articles/s44259-024-00068-x