
Large Language Model Recommendations for Empiric Antibiotics Versus Clinician Prescribing: A Non-Interventional Paired Retrospective Antimicrobial Stewardship Analysis

[Ninel Iacobus Antonie](#)*, [Vlad Alexandru Ionescu](#), [Gina Gheorghe](#), [Crista-Loredana Tiuca](#),
[Camelia Cristina Diaconu](#)

Posted Date: 10 March 2026

doi: 10.20944/preprints202603.0720.v1

Keywords: antimicrobial stewardship; antimicrobial resistance; large language model; paired retrospective study; OpenAI API; artificial intelligence



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Large Language Model Recommendations for Empiric Antibiotics Versus Clinician Prescribing: A Non-Interventional Paired Retrospective Antimicrobial Stewardship Analysis

Ninel Iacobus Antonie ^{1,2,*}, Vlad Alexandru Ionescu ^{1,2}, Gina Gheorghe ^{1,2},
Crista-Loredana Tiuca ^{1,2} and Camelia Cristina Diaconu ^{1,2,3}

¹ Faculty of Medicine, University of Medicine and Pharmacy Carol Davila Bucharest, 050474 Bucharest, Romania

² Internal Medicine Department, Clinical Emergency Hospital of Bucharest, 105402 Bucharest, Romania

³ Academy of Romanian Scientists, 050045 Bucharest, Romania

* Correspondence: ninel-iacobus.antonie@drd.umfcd.ro

Abstract

Background/Objectives: Antimicrobial resistance (AMR) remains a major global health threat, strengthening the case for antimicrobial stewardship that limits unnecessary broad-spectrum empiric therapy while preserving timely coverage in severe infection. Large language models (LLMs) are being explored for decision support, but require rigorous offline evaluation before any clinical implementation. **Methods:** Single-center retrospective paired evaluation at Clinical Emergency Hospital of Bucharest (Internal Medicine, 2020–2024). The unit of analysis was the admission (N = 493), with paired 24 h empiric regimens (clinician-prescribed vs post hoc LLM-recommended via OpenAI API; not visible to clinicians; no influence on care). Local laboratory-derived epidemiology was precomputed from microbiology exports and provided as structured prompt context to approximate information parity with clinicians' implicit local ecology knowledge. Primary (prespecified) endpoint: any contextual guardrail violation (unjustified carbapenem/antipseudomonal/anti-MRSA under prespecified structured severity/MDR-risk rules), exact McNemar. Key secondary (prespecified): Δ contextual guardrail penalty (LLM – Clin), sign test and Wilcoxon signed-rank (ties reported). Ethics committee approval was obtained. **Results:** Guardrail violations occurred in 17.0% of clinician regimens vs 4.9% of LLM regimens (paired RD –12.2%; matched OR 0.216, 95% CI 0.127–0.367; McNemar exact $p = 1.60 \times 10^{-10}$). Δ penalty had median 0 with 398/493 ties; among non-ties, improvements ($\Delta < 0$) exceeded adverse shifts (79 vs 16; sign-test $p = 3.47 \times 10^{-11}$). **Conclusions:** In this offline, non-interventional paired evaluation, LLM regimens were associated with fewer prespecified contextual guardrail violations compared to clinician empiric regimens under a rule-based stewardship benchmarking framework. These endpoints strictly quantify concordance with stewardship constraints rather than patient outcomes, necessitating cautious interpretation of secondary and subset analyses. Ultimately, reproducible guardrail-based benchmarking may support subsequent prospective, safety-governed evaluations.

Keywords: antimicrobial stewardship; antimicrobial resistance; large language model; paired retrospective study; OpenAI API; artificial intelligence

1. Introduction

Antimicrobial resistance (AMR) remains one of the most urgent global public health threats: bacterial AMR was estimated to be directly responsible for ~1.27 million deaths in 2019 and associated with ~4.95 million deaths worldwide in the same year [1]. Looking forward, the UK government–

commissioned research projected up to 10 million deaths annually by 2050 in a worst-case scenario without effective action [2], while newer global forecasting work estimates around 1.91 million deaths attributable to AMR (and 8.22 million associated deaths) in 2050, underscoring a sustained and aging-driven burden trajectory [3]. This escalating crisis is a core rationale for antimicrobial stewardship programs, which aim to optimize antimicrobial use, particularly by limiting unnecessary broad-spectrum empiric exposure, while preserving timely, appropriate coverage when severe infection is suspected. Antimicrobial resistance can also be understood as an evolutionary response to antibiotic selection pressure, whereby microorganisms adapt through genetic variation and horizontal gene transfer to survive antimicrobial exposure, thus facing an informational challenge in adaptive survival under antimicrobial pressure [4,5]. In parallel, modern inpatient care confronts a complementary yet unresolved informational challenge: integrating rapidly expanding and heterogeneous clinical data, such as patient history, comorbidities, severity indicators, prior antibiotic exposure, and local epidemiology, into timely and precise empiric antimicrobial decision-making at the bedside [6,7]. Within this framing, large language models (LLMs) and other artificial intelligence (AI)-enabled decision-support approaches can be viewed as attempts to improve clinical information synthesis and decision consistency, while also introducing well-recognized safety risks such as hallucinated or unsupported recommendations that necessitate rigorous offline evaluation prior to any clinical deployment [8,9]. In this study, LLM recommendations were generated post hoc (offline), were not visible to clinicians, and had no influence on patient care. Rather than expanding empiric spectra indiscriminately, such tools may have potential to accelerate context-sensitive decision-making, supporting a better balance between urgency and precision in settings where delays and overly broad coverage both carry downstream consequences (e.g., resistance selection, adverse events, and excess costs) [10–12].

Because empiric prescribing decisions are often made under uncertainty, stewardship evaluation benefits from focusing on a small set of “high-leverage” choices that disproportionately shape ecological impact and downstream resistance selection. In routine acute-care practice, three such decisions are the inclusion of (I) carbapenems, (II) antipseudomonal broad-spectrum β -lactams, and (III) anti-MRSA (Methicillin-resistant *Staphylococcus aureus*) agents. Many agents in these empiric categories fall into higher-stewardship priority groups (e.g., World Health Organization AWaRe “Access/Watch/Reserve” classes) and are consistently targeted by antimicrobial stewardship programs for indication review, risk-stratified initiation, and early de-escalation when structured risk factors are absent [13–15]. Carbapenems, in particular, are central to carbapenem-sparing strategies intended to reduce selective pressure for carbapenem resistance [16]. Anti-MRSA therapy is another common empiric add-on in hospitalized patients, yet multiple stewardship approaches (including rapid diagnostic-supported pathways) emphasize discontinuation when MRSA risk is low, to minimize unnecessary exposure and adverse effects [17]. Finally, antipseudomonal therapy is widely recognized as a major driver of broad-spectrum pressure; stewardship guidance repeatedly stresses reserving such coverage for patients with convincing severity or MDR-risk signals and revisiting empiric breadth once early clinical and microbiological data accrue [13,18,19]. In this study, we therefore operationalized these high-impact stewardship choices as contextual guardrails, pre-specified “safety rails” that label broad-spectrum elements as justified only when objective, structured indicators of severity and MDR/MRSA risk are present.

Empiric antibiotic selection is not solely syndrome-driven but also contingent on local resistance ecology; therefore, facility-specific antibiograms and their syndrome-oriented extensions are routinely used to operationalize “what is likely to work here and now” [20]. In particular, the weighted-incidence syndromic combination antibiogram (WISCA) extends conventional antibiograms by integrating syndrome-specific pathogen incidence with susceptibility data to estimate the expected coverage of candidate empiric regimens (including combinations), and has been proposed as a pragmatic decision aid for empiric therapy [21–23].

Despite growing interest in LLMs for clinical decision support, evaluation frameworks for early-stage clinical AI emphasize transparent reporting, safety assessment, and non-interventional study

designs prior to any real-world deployment [24,25]. This is especially salient in antimicrobial stewardship, where outputs may be clinically plausible yet unsafe or unjustified in specific contexts, and where empiric broad-spectrum escalation decisions are high-impact and heterogeneous under uncertainty [8,9,12].

LLM outputs can be sensitive to prompt context and may exhibit non-determinism and confidently stated but unsupported content (“hallucinations”), raising reproducibility and safety concerns in high-stakes prescribing decisions [8,9,26]. Reporting guidance for LLM-based clinical studies further reinforces the need for reproducible pipelines, clear definition of the evaluation setting, and explicit separation between model development and outcome assessment [24,25].

We therefore conducted a single-center, retrospective offline paired evaluation comparing clinician-prescribed empiric antibiotic regimens with post hoc LLM recommendations within the first 24 h of admission, using pre-specified structured contextual stewardship endpoints. Against this background, we treated local epidemiology as an *information parity* component: in routine practice, clinicians acquire implicit knowledge of local ecology through lived experience and institutional feedback loops, whereas an LLM does not. In this context we precomputed local laboratory-derived epidemiology from microbiology exports and injected it as static structured context into the LLM prompt to approximate institutional ecology awareness, while explicitly separating this module from any endpoint definitions and avoiding any tuning to outcomes or study metrics [24,25].

The primary (prespecified) outcome was any contextual guardrail violation (paired binary), and the key secondary (prespecified) outcome was the paired difference in contextual guardrail penalty (LLM – Clin). All additional analyses were considered secondary, supplementary, or exploratory and interpreted with explicit multiplicity and selection-bias caution, without inference about clinical outcomes.

2. Results

2.1. Cohort Flow

Among 14,879 Internal Medicine admissions at the Clinical Emergency Hospital of Bucharest (2020–2024), 6,734 were identified using an infection-suspected admission diagnosis filter. To ensure feasibility of manual chart abstraction while maintaining balanced temporal coverage, we prespecified a target sample of 100 eligible admissions per year (total N = 500) as a pragmatic design choice. A total of 841 candidate admissions underwent manual chart review; during this process, prespecified exclusion criteria were operationalized at the record level (e.g., non-infectious admission, no systemic antibiotics within 24 h, missing medication documentation), with decision counts summarized in Supplementary Table S1. Of these, 500 admissions (100/year) met eligibility and were enrolled for detailed data abstraction. After applying the prespecified deduplication rule (first eligible admission per patient), 7 admissions were removed, yielding the final analytic cohort of 493 paired admissions. The remaining 341 screened admissions were not enrolled due to prespecified reasons (non-infectious admission N = 45; no systemic antibiotics within 24 h N = 262; missing medication documentation N = 34). See Figure 1.

Baseline cohort characteristics are summarized in Table 1. The median age was 72 years (IQR 64–82), 250/493 (50.7%) were female, and the median length of stay was 9 days (IQR 5–15). Community-onset infection predominated (459/493, 93.1%), and community-acquired pneumonia was the most common index syndrome (311/493, 63.1%). Endpoint mapping quality control identified no unmapped antibiotic agents in the analytic cohort.

Table 1. Baseline characteristics of the paired admission cohort (N = 493).

Characteristic	Total (N = 493)
Demographics	
Age, years, median (IQR)	72 (64–82)

Female sex, n (%)	250 (50.7%)
Length of stay, days, median (IQR)	9 (5–15)
Study period	
Cohort year 2020, n (%)	98 (19.9%)
Cohort year 2021, n (%)	100 (20.3%)
Cohort year 2022, n (%)	100 (20.3%)
Cohort year 2023, n (%)	98 (19.9%)
Cohort year 2024, n (%)	97 (19.7%)
Acquisition setting	
Community-onset, n (%)	459 (93.1%)
Healthcare-associated, n (%)	34 (6.9%)
Index syndrome (top categories)	
Community-acquired pneumonia, n (%)	311 (63.1%)
Urinary tract infection - unspecified site, n (%)	56 (11.4%)
Bloodstream infection / sepsis, n (%)	50 (10.1%)
COPD infectious exacerbation, n (%)	32 (6.5%)
Urinary tract infection – pyelonephritis, n (%)	29 (5.9%)
Skin and soft-tissue infection, n (%)	7 (1.4%)
Other syndromes, n (%)	8 (1.6%)
Severity / support within 24 h	
Sepsis documented, n (%)	183 (37.1%)
Septic shock documented, n (%)	129 (26.2%)
Respiratory failure documented, n (%)	374 (75.9%)
Mechanical ventilation, n (%)	147 (29.8%)
Vasopressors, n (%)	127 (25.8%)
ICU transfer, n (%)	141 (28.6%)
Prior exposure / MDR-risk proxies	
Antibiotics in prior 90 days, n (%)	75 (15.2%)
Hospitalization in prior 90 days, n (%)	70 (14.2%)
Long-term care facility resident, n (%)	32 (6.5%)
Prior MRSA colonization/infection, n (%)	6 (1.2%)
Prior ESBL/CRE/VRE history, n (%)	22 (4.5%)
Home antibiotics before admission: Yes, n (%)	65 (13.2%)
Home antibiotics before admission: No, n (%)	428 (86.8%)
Comorbidities	
Hypertension, n (%)	374 (75.9%)
Diabetes mellitus, n (%)	160 (32.5%)
COPD, n (%)	94 (19.1%)
Chronic kidney disease, n (%)	137 (27.8%)
Congestive heart failure, n (%)	241 (48.9%)
Atrial fibrillation, n (%)	180 (36.5%)
Prior stroke, n (%)	98 (19.9%)

Cirrhosis, n (%)	27 (5.5%)
Malignancy, n (%)	90 (18.3%)
Immunosuppression, n (%)	29 (5.9%)

Values are n (%) unless otherwise stated. Severity/support flags refer to the first 24 h after admission. All years reflect the final counts after applying the deduplication rule (first admission per patient). Abbreviations: IQR, interquartile range; LOS, length of stay; ICU, intensive care unit; CAP, community-acquired pneumonia; SSTI, skin and soft tissue infection; UTI, urinary tract infection; MDR, multidrug-resistant.

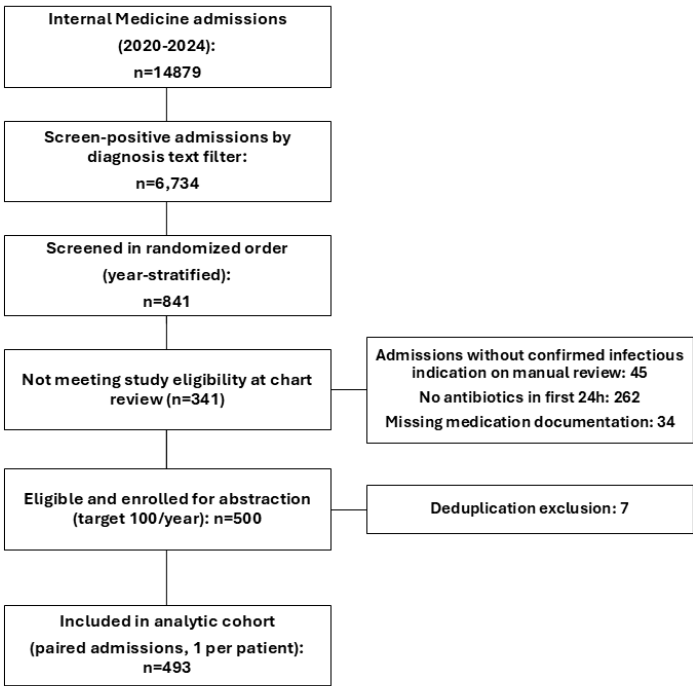


Figure 1. STROBE flow diagram of cohort selection (admissions as the unit of analysis). Numbers refer to admissions. Of 841 chart-reviewed admissions, 341 did not meet eligibility at chart review (admissions without confirmed infectious indication on manual review N = 45; no systemic antibiotics within 24 h N = 262; missing medication documentation N = 34). Of 500 enrolled admissions, 7 were removed by the prespecified deduplication rule (first eligible admission per patient), yielding a final paired analytic cohort of N = 493 admissions.

The final paired analytic cohort included **493 admissions** contributing matched clinician and LLM empiric regimens within the first 24 h. The predefined **72 h continued-therapy** subset included **323 admissions**, and the **microbiology-evaluable paired subset** included **158 admissions** (Table 2).

Table 2. Primary, key secondary, and selected secondary endpoints (paired, first 24 h window).

Panel	Endpoint	N	Clinici an	LLM	Discord ant pairs	Effect	P value	Median 95% CI (bootstra p)
-------	----------	---	---------------	-----	----------------------	--------	---------	-------------------------------------

A (prespecified)	Primary: Any contextual guardrail violation (composite)	49 3	84/493 (17.0%)	24/49 3 (4.9%)	Clin = 1/LLM = 0: 76; Clin = 0/LLM = 1: 16	Matched OR 0.216 (95% CI 0.127–0.367) RD (LLM–Clin) –0.122	p = 1.60 × 10 ^{–10}
B (prespecified)	Key secondary: contextual guardrail penalty (LLM–Clin)	49 Δ 3	—	—	Δ < 0 (lower): 79; Δ > 0 (higher): 16; Δ = 0: 398	Median 0 (IQR 0–0) Mean –0.219 (SD 0.789) Wilcoxon Sign	Median 0 (IQR 0–0); 95% CI 0–0 p = 9.48 × 10 ^{–11} p = 3.47 × 10 ^{–11}
C (Secondary; multiplicity caution)	Any broad-spectrum class used (carb OR APS OR anti-MRSA)	49 3	289/493 (58.6%)	195/4 93 (39.6%)	Clin = 1/LLM = 0: 143; Clin = 0/LLM = 1: 49	Matched OR 0.345 (95% CI 0.250–0.477) RD (LLM–Clin) –0.191	p = 7.26 × 10 ^{–12}
C (Secondary; multiplicity caution)	Carbapenem contextual violation	49 3	39/493 (7.9%)	9/493 (1.8%)	Clin = 1/LLM = 0: 36; Clin = 0/LLM = 1: 6	Matched OR 0.178 (95% CI 0.077–0.410) RD (LLM–Clin) –0.061	p = 2.83 × 10 ^{–6}
C (Secondary; multiplicity caution)	Antipseudomonal contextual violation	49 3	31/493 (6.3%)	2/493 (0.4%)	Clin = 1/LLM = 0: 30; Clin = 0/LLM = 1: 1	Matched OR 0.049 (95% CI 0.010–0.253) RD (LLM–Clin) –0.059	p = 2.98 × 10 ^{–8}
C (Secondary; multiplicity caution)	Anti-MRSA contextual violation	49 3	36/493 (7.3%)	17/49 3	Clin = 1/LLM = 0: 32;	Matched OR 0.415 (95% CI	p = 6.61 × 10 ^{–3}

multiplicity caution)		(3.4%)	Clin = 0/LLM = 1: 13	= 0.220–0.784)	RD (LLM–Clin) –0.039
C (Secondary; multiplicity caution)	Δ empiric antibiotic cost, 24 h (EUR; LLM–Clin)	49 – 3	–	$\Delta < 0$ (lower): 268; $\Delta > 0$ (higher): 164; $\Delta = 0$: 61	Median – 1.43 (EUR) (IQR – 9.80–0.57) Mean –4.11 (EUR) (SD 14.35)
					Wilcoxon Sign p = 2.80 × 10 ^{–13} ; p = 6.41 × 10 ^{–7}
					Median 95% CI (bootstrap p): –4.09 to –0.07

Notes: Unit of analysis is the admission; all comparisons are paired within admission (clinician vs LLM; N = 493). For paired binary endpoints, we report arm-specific counts (%), discordant pairs (Clinician = 1/LLM = 0 and Clinician = 0/LLM = 1), matched odds ratios (0.5 continuity correction, as prespecified) with Wald 95% CIs on the log scale, risk differences (LLM–clinician), and exact McNemar p-values. For paired deltas, Δ denotes LLM – clinician; negative Δ indicates a lower value in the LLM arm (e.g., lower contextual penalty or lower empiric cost). Panel A is the prespecified primary endpoint; Panel B is the prespecified key secondary endpoint; Panel C endpoints are secondary and should be interpreted with multiplicity caution. Holm-adjusted p-values for the prespecified paired-delta family are provided in Supplementary Table S12. Broad-spectrum class exposure is defined as any carbapenem, antipseudomonal β -lactam, or anti-MRSA agent used during the first 24 h empiric management window. Abbreviations: LLM, large language model; OR, odds ratio; CI, confidence interval; RD, risk difference; IQR, interquartile range; SD, standard deviation; MRSA, methicillin-resistant *Staphylococcus aureus*; EUR, euros; DDD, defined daily dose.

2.2. Primary Endpoint

In the paired cohort (N = 493), the primary prespecified endpoint: any contextual guardrail violation, occurred in 17.0% of clinician regimens versus 4.9% of LLM-recommended regimens (definitions, antibiotic mappings and weights in Supplementary Table S3). Discordant pairs favored the LLM arm (Clin = 1/LLM = 0: 76 vs Clin = 0/LLM = 1: 16; 2×2: n00 = 393, n01 = 16, n10 = 76, n11 = 8), corresponding to a matched OR 0.216 (95% CI 0.127–0.367) and exact McNemar p = 1.60 × 10^{–10} (Table 2, panel A, Figure 2). These analyses were performed post hoc/offline; LLM recommendations were not visible to clinicians and did not influence care.

2.3. Key Secondary Endpoint

For the **key secondary prespecified endpoint**, the paired **Δ contextual guardrail penalty (LLM–Clin)** had a **median of 0 (IQR 0–0)** and **mean –0.219 (SD 0.789)**. The penalty was lower in the LLM arm in **79 admissions**, higher in **16**, and tied in **398** (non-ties: 95). The paired distribution differed from zero by **Wilcoxon signed-rank p = 9.48 × 10^{–11}** and **sign test p = 3.47 × 10^{–11}** (Table 2, Panel B, Figure 3).

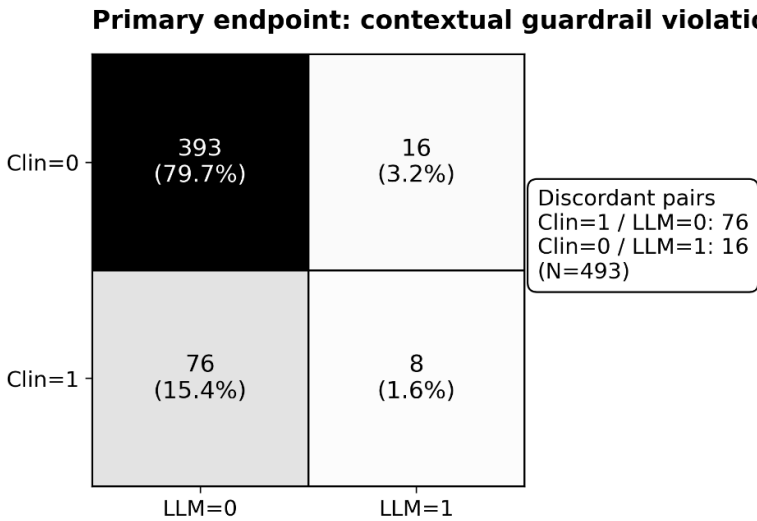


Figure 2. Primary endpoint (paired 2×2). Paired admission-level comparison (N = 493) of any contextual guardrail violation within the first 24 h: clinician regimen (rows) vs LLM regimen (columns). Cells show counts and percentage of total admissions (N = 493). Effect estimates and p-values are reported in Table 2. Abbreviations: LLM, large language model; Clin, clinician.

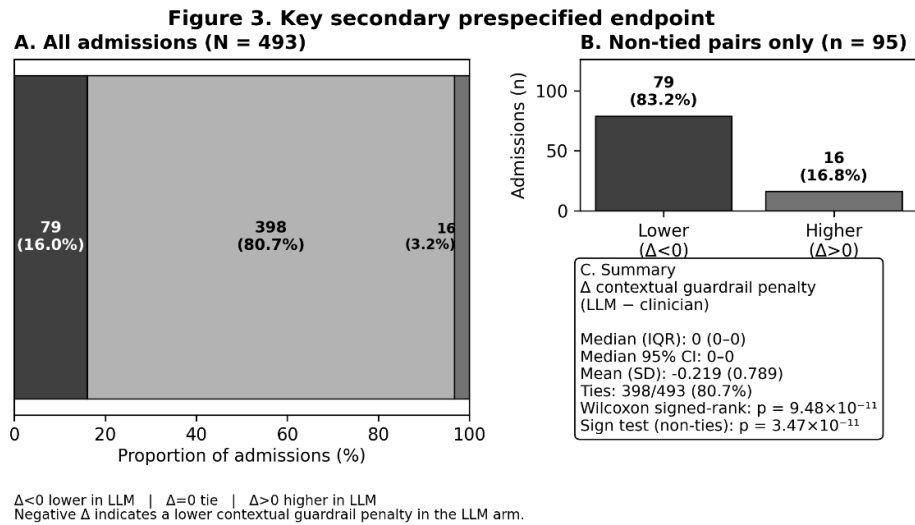


Figure 3. Key secondary prespecified endpoint: distribution of paired Δ contextual guardrail penalty (LLM – clinician) across admissions (N = 493). (A) All admissions, showing the proportion with lower ($\Delta < 0$), tied ($\Delta = 0$), and higher ($\Delta > 0$) penalty in the LLM arm. (B) Non-tied pairs only (N = 95), highlighting the direction of discordance. Summary statistics report median (IQR), mean (SD), the number of ties, and paired tests (Wilcoxon signed-rank and sign test). Negative Δ indicates a lower contextual guardrail penalty in the LLM arm. Abbreviations: LLM, large language model; OR, odds ratio; CI, confidence interval; IQR, interquartile range; SD, standard deviation.

2.4. Secondary Endpoints (Multiplicity Caution)

2.4.1. Contextual Guardrail Components

In secondary component analyses (paired; multiplicity caution), discordant patterns generally favored fewer contextual violations in the LLM arm. Carbapenem contextual violations had 36 clinician-only vs 6 LLM-only discordant pairs (matched OR 0.178, 95% CI 0.077–0.410; exact McNemar $p = 2.83 \times 10^{-6}$). Antipseudomonal contextual violations had 30 clinician-only vs 1 LLM-

only discordant pairs (matched OR 0.049, 95% CI 0.010–0.253; exact McNemar $p = 2.98 \times 10^{-8}$). Anti-MRSA contextual violations had 32 clinician-only vs 13 LLM-only discordant pairs (matched OR 0.415, 95% CI 0.220–0.784; exact McNemar $p = 6.61 \times 10^{-3}$) (Table 2, Panel C; Figure 4).

Consistent with these component-level findings, overall broad-spectrum class exposure, defined as use of at least one carbapenem, antipseudomonal agent, or anti-MRSA agent within the first 24 h, was lower in the LLM arm. In the paired cohort ($N = 493$), any broad-spectrum class exposure occurred in 58.6% of clinician regimens versus 39.6% of LLM regimens, supporting the interpretation that LLM recommendations were, under this prespecified rule-based benchmarking framework, less likely to trigger broad-spectrum stewardship guardrails during early empiric management (Table 2, Panel C).

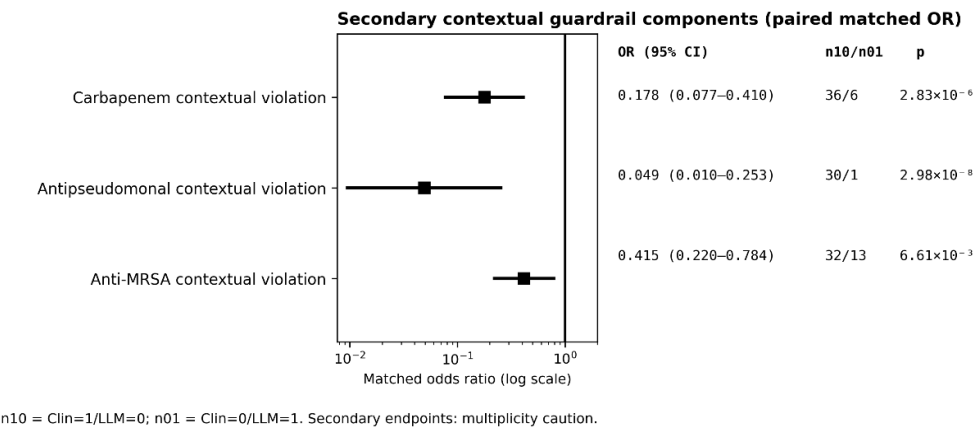


Figure 4. Secondary contextual guardrail components: matched odds ratios (paired analyses; $N = 493$). Matched ORs are computed from discordant pairs ($n10 = \text{clinician} = 1/\text{LLM} = 0$; $n01 = \text{clinician} = 0/\text{LLM} = 1$) with 0.5 continuity correction and Wald 95% confidence intervals on the log scale. P-values are from the exact McNemar test. Values <1 indicate fewer violations in the LLM arm. Secondary endpoints should be interpreted with multiplicity caution. Abbreviations: LLM, large language model; OR, odds ratio; CI, confidence interval.

2.4.2. Costs (Secondary)

Total empiric antibiotic costs over the first 24 h were lower in the LLM arm when aggregated across the cohort. Summing admission-level empiric antibiotic costs yielded 4,689.02 EUR for clinician regimens versus 2,591.83 EUR for LLM regimens (EUR, using the study’s fixed price mapping and exchange-rate assumptions). These totals are presented as descriptive cohort-level context and should be interpreted alongside the paired distributional analyses reported below. For more details, see Supplementary Data File S2 (data_file_s2_costing).

At the admission level, the 24 h empiric antibiotic cost delta (LLM – clinician) had a median of –1.43 EUR (IQR –9.80 EUR to 0.57 EUR) (paired Wilcoxon $p = 2.80 \times 10^{-13}$; $N = 493$; sign test $p = 6.41 \times 10^{-7}$; $N = 493$). In the predefined 72 h continued-therapy subset ($N = 323$), the 72 h cost delta had a median of –9.68 EUR (IQR –23.92 EUR to 1.70 EUR) (paired Wilcoxon $p = 2.34 \times 10^{-10}$). These are process-level economic comparisons of recommended versus prescribed empiric regimens under the study’s costing assumptions and do not imply clinical outcome impact (Figure 5).

2.4.3. Microbiology-Evaluable Subset

In the microbiology-evaluable paired subset ($N = 158$, with 335/493 not eligible for microbiological evaluation), active coverage against the index organism differed between arms (2×2 : $n00 = 38$, $n01 = 23$, $n10 = 10$, $n11 = 87$), with a matched OR 2.24 (95% CI 1.08–4.63) and exact McNemar $p = 0.0351$. Culture acquisition and positivity are post-baseline processes and may correlate with severity/trajectory; therefore, this subset analysis does not estimate causal effectiveness. Because

evaluability depends on culture availability and interpretable susceptibility, these findings should be interpreted as subset-level process benchmarking rather than cohort-wide effectiveness (Figure 6).

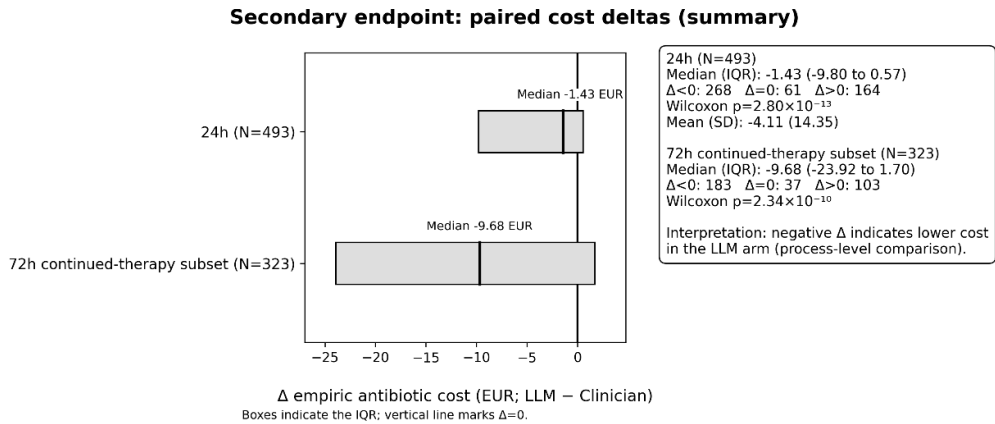


Figure 5. Secondary endpoint: paired empiric antibiotic cost deltas (Δ cost, EUR; LLM – clinician). The 24 h analysis includes all paired admissions (N = 493). The 72 h analysis is restricted to a prespecified continued-therapy subset in which empiric therapy initiated by the clinician was continued for ≥72 h (N = 323). Boxes indicate the interquartile range with the median; the vertical reference line marks Δ = 0. Negative Δ indicates lower empiric antibiotic cost in the LLM arm (process-level comparison). Abbreviations: LLM, large language model; EUR, euros; IQR, interquartile range.

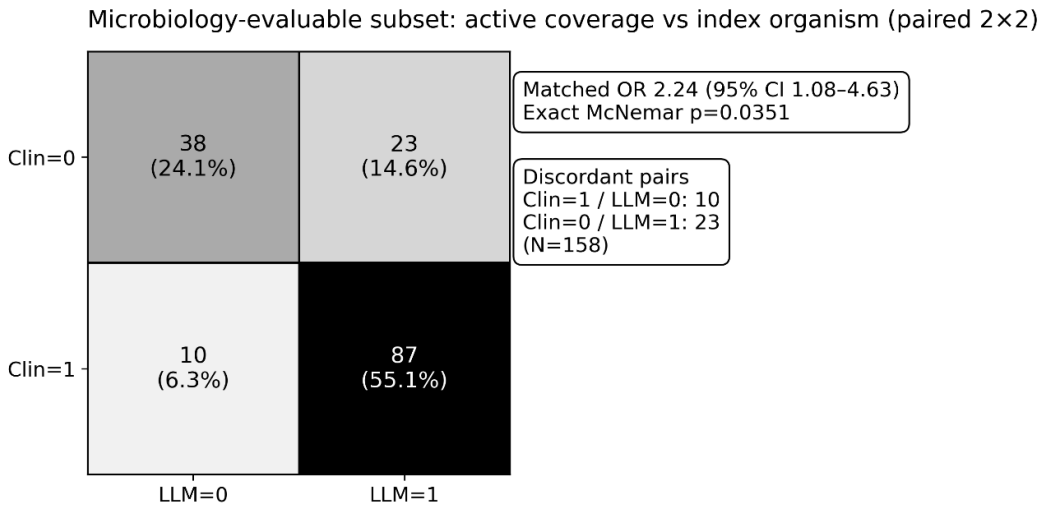


Figure 6. Microbiology-evaluable subset: paired active coverage against the index organism (N = 158). 0 = inactive coverage; 1 = active coverage against the index organism (per prespecified definition). Each cell shows the count (and percentage) of admissions classified as inactive/active for the clinician regimen (rows) versus the LLM regimen (columns), under the study’s prespecified definition of “active coverage” against the index organism. Discordant pairs (clinician active/LLM inactive vs clinician inactive/LLM active) are shown with the matched OR and exact McNemar p-value. Because evaluability depends on culture availability and interpretable susceptibility, this analysis reflects subset-level process benchmarking rather than cohort-wide effectiveness. Abbreviations: LLM, large language model; OR, odds ratio; CI, confidence interval.

2.4.4. Concordance (Supplementary Framing)

Regimen concordance between clinicians and the LLM was limited: 57/493 (11.6%) were identical as an antibiotic set, 137/493 (27.8%) shared the same primary agent, 148/493 (30.0%) had any antibiotic overlap, and 345/493 (70.0%) had no overlap (Table 3).

Table 3. Regimen concordance between clinician and LLM regimens (N = 493). Overlap defined as ≥1 identical antibiotic agent present in both regimen sets (molecule-level mapping).

Metric	n	%
Exact identical regimen set	57	11.6%
Same primary agent (first-listed antibiotic)	137	27.8%
Any overlap (≥1 shared agent)	148	30.0%
No overlap (0 shared agents)	345	70.0%

2.4.5. Supplementary QC Note (NO_ANTIBIOTIC)

The LLM recommended NO_ANTIBIOTIC in 5/493 admissions. None met the structured SEVERE or MDR_RISK triggers used in the contextual guardrail framework. Case-level details are provided in Supplementary Table S4. Per study definition, all five admissions had received empiric antibiotics from the clinician, whereas the LLM recommended withholding antibacterial therapy.

2.4.6. Exploratory Modeling (Supplement Only)

Exploratory complete-case models (N = 493) were used to assess predictors of a nonzero Δ contextual guardrail penalty (LLM–Clin ≠ 0; events = 95) using logistic regression with robust (HC3) standard errors. Given the retrospective observational design, potential sparsity in some covariate levels, and the exploratory intent, these analyses are presented in the Supplementary Materials only and interpreted cautiously (Table S5, Table S5a; model-level complete-case QC in Table S6a).

3. Discussion

3.1. Principal Findings

In this single-center retrospective offline/post hoc paired evaluation of 493 admissions, LLM-generated empiric regimens had markedly fewer contextual guardrail violations than clinician regimens within the first 24 h (4.9% vs 17.0%, paired RD −12.2%; matched OR 0.216; exact McNemar $p = 1.60 \times 10^{-10}$). The reduction was directionally consistent across guardrail components (antipseudomonal, carbapenem, anti-MRSA), and coincided with lower broad-spectrum class use (58.6% vs 39.6%) and a modest reduction in 24 h empiric antibiotic costs (median Δ −1.43 EUR, IQR −9.80 EUR to 0.57 EUR; bootstrap 95% CI −4.09 EUR to −0.07 EUR), while overall early exposure was similar (median Δ DDD/24 h 0 with substantial ties). These patterns align with the core aims of antimicrobial stewardship: optimizing initial agent selection, minimizing unnecessary broad-spectrum escalation, and improving resource utilization, while recognizing that such process endpoints are not direct measures of clinical effectiveness [27–29].

From a stewardship perspective, the signal evoked from the data is clinically plausible: unnecessary broad-spectrum exposure is a well-established driver of avoidable harms (notably *C. difficile* risk) and selection pressure for resistant organisms, and stewardship interventions have been associated with reductions in both CDI and resistant organism infection/colonization in hospital settings [27,29]. Broad-spectrum “intensity” measures have also shown dose–response relationships with hospital-associated CDI risk, reinforcing the biological rationale for prioritizing spectrum-aware empiric choices when patient risk factors do not mandate escalation [30].

3.2. Interpretation in Antimicrobial Stewardship Terms

These results should be interpreted as **process-level antimicrobial stewardship benchmarking**, not as evidence of clinical effectiveness or safety in patient outcomes. In this study, the primary endpoint operationalizes contextual appropriateness constraints (“local stewardship guardrails”), i.e., whether empiric regimens cross prespecified escalation thresholds (antipseudomonal, carbapenem, anti-MRSA) in the absence of documented high-risk features, thereby capturing a proximal marker of prescribing quality rather than downstream outcomes. Consistent with this

framing, the LLM arm showed fewer contextual guardrail violations (4.9% vs 17.0%), fewer discordant escalations, and lower broad-spectrum class use (58.6% vs 39.6%), alongside a modest reduction in 24 h costs (median Δ -1.43 EUR), while overall early exposure (DDD/24 h) remained centered at 0—suggesting that the main effect is **spectrum selection** rather than “less antibiotic” per se [15].

Nuancing the comparison is essential: the clinician arm reflects **real-world empiric prescribing under diagnostic uncertainty**, evolving information, and local organizational constraints, where clinicians may rationally “over-cover” to avoid the well-described harms of **inappropriate empiric therapy** in severe bacterial infections. Meta-analyses across serious infections (e.g., pneumonia/BSI/sepsis) have repeatedly associated inappropriate initial empiric antibiotics with higher mortality, which provides a plausible clinical rationale for risk-averse escalation when pathogen and susceptibility data are unavailable [31,32]. Conversely, stewardship frameworks (including WHO’s AWARe system) explicitly emphasize minimizing unnecessary Watch/Reserve exposure to reduce resistance selection pressure, providing a principled basis for guardrails that discourage broad-spectrum escalation when not clearly indicated [15]. Taken together, these endpoints and findings align with the study’s central stewardship trade-off: **maximizing the probability of appropriate early coverage** while **minimizing avoidable broad-spectrum exposure**, but they do so as an auditable prescribing-process benchmark rather than an outcomes trial.

3.3. Why Might the LLM Look Better on Guardrails?

One plausible explanation is simply that the LLM is being evaluated against a **rule-defined stewardship construct** (the contextual guardrails defined above), and LLM outputs tend to “snap” toward **guideline-like defaults** when high-risk features are not explicit in the structured context. In other words, a language model prompted with a standardized admission summary may preferentially select narrower empiric options unless it sees clear triggers for escalation, which increases concordance with prespecified guardrails. This is conceptually similar to what has been observed with more conventional antimicrobial **clinical decision support systems (CDSS)**: across heterogeneous settings, CDSS interventions tend to show their most consistent benefits on process endpoints (appropriateness, reduced broad-spectrum prescribing) rather than hard clinical outcomes [20,33,34].

Findings from this study highlight that the primary endpoint difference is driven mainly by discordant pairs: far more admissions fell into the “Clinician = 1 / LLM = 0” cell than the reverse, consistent with fewer broad-spectrum escalations being triggered by the LLM under the same recorded admission context. That pattern is compatible with a “default-to-less-escalation” behavior in the LLM arm, again, not necessarily “better care,” but greater alignment with how the guardrails are encoded in the evaluation framework.

The clinician arm, by contrast, reflects real-world empiric prescribing where decisions are made under time pressure, incomplete information, and institutional constraints, and where the perceived downside of under-treatment can dominate. Sepsis-related anxieties and escalation norms can legitimately push practice toward broader empiricism, even when documentation of risk modifiers is imperfect [35]. At the clinician-behavior level, multiple strands of work describe how uncertainty, anticipated regret, and action bias can tilt antibiotic decisions “just in case,” which is directionally consistent with why more clinician regimens may cross broad-spectrum guardrails in routine care [36,37].

3.4. Microbiology-Evaluable Subset: What It Means and What It Does Not Mean

In the microbiology-evaluable paired subset (N = 158, with 335/493 **not eligible for microbiological evaluation**), empiric **active coverage against the index organism** differed between arms (2×2: n00 = 38, n01 = 23, n10 = 10, n11 = 87), corresponding to a matched OR **2.24** (95% CI **1.08–4.63**) with exact McNemar **p = 0.0351**. These data suggest that, among admissions where an organism

and interpretable susceptibility were available, the LLM-generated regimen more often met *in vitro* activity criteria than the clinician regimen within the empiric window.

However, microbiology “evaluability” is not a baseline attribute of the cohort; it depends on **culture acquisition**, organism recovery, and interpretable susceptibility, processes that occur during care and are influenced by clinical severity, diagnostic strategy, prior antibiotic exposure, and workflow factors. Conditioning inference on such a post-baseline filter can induce **selection (collider-stratification) bias**, meaning that associations within the evaluable subset may not reflect cohort-wide performance and should not be interpreted as causal evidence of superior effectiveness [38,39]. In this context, the most defensible interpretation is that the microbiology subset provides **subset-level process benchmarking** (i.e., an auditable “coverage check” when microbiology is available) and is best viewed as **hypothesis-generating**, rather than as a definitive estimate of empiric effectiveness across all admissions.

To strengthen transparency, we treated microbiology evaluability as a selection process. We therefore (i) compared baseline characteristics and severity proxies between microbiology-evaluable and non-evaluable admissions (Table S2), and (ii) interpreted microbiology-subset findings as supportive evidence of prescribing-process alignment rather than cohort-wide effectiveness. Future prospective shadow-mode (offline, non-interventional) validation could predefine culture-based endpoints and consider methods that address evaluability mechanisms (e.g., inverse-probability weighting), while prioritizing patient-centered outcomes.

3.5. Economic and Exposure Endpoints (Cost and DDD)

Although the **median** per-admission 24 h empiric antibiotic cost difference was modest (LLM–Clin **–1.43 EUR**), cohort-level totals illustrate the potential budget impact of spectrum selection. Across all **493** admissions, 24 h anti-infective acquisition costs were **EUR 4,689.02** in the clinician arm versus **EUR 2,591.83** in the LLM arm (absolute difference **EUR 2,097.19, –44.7%**), corresponding to a **mean** reduction of **EUR 4.25 per admission**. In the prespecified **72 h continued-therapy subset** (N = 323), totals were **EUR 8,676.20** versus **EUR 4,636.98** (difference **EUR 4,039.22, –46.6%**), corresponding to **EUR 12.50** per admission in that subset; when scaled to the full cohort this corresponds to **EUR 8.19 per admission** (N = 493). These figures reflect **drug acquisition costs only** (excluding administration, monitoring, and downstream consequences), yet they are directionally consistent with the broader health-economic literature indicating that stewardship interventions can reduce antibiotic expenditures and may avert costly complications linked to broad-spectrum exposure (including hospital-associated *Clostridioides difficile* infection) [29,40,41].

In contrast, **DDD/24 h** differences were small (median Δ 0, with substantial ties), suggesting that the stewardship signal is driven primarily by **spectrum choice** rather than a reduction in overall antibiotic intensity (Table S12). This interpretation aligns with WHO’s **AWaRe** framework, which encourages minimizing unnecessary Watch/Reserve exposure as a stewardship objective while maintaining adequate early therapy when clinically warranted.

3.6. Comparison with Prior Work

The recent literature on AI support for antibiotic decisions spans two partly separate traditions: (i) **rule-based or ML-driven clinical decision support systems (CDSS)** embedded in stewardship programs, and (ii) **generative LLM evaluations** that typically test narrative recommendations against guidelines or expert ratings. For CDSS, the best-supported effects tend to be on **process outcomes** (improved guideline concordance or reduced unnecessary broad-spectrum prescribing) while evidence for consistent patient-outcome benefits is mixed and highly context-dependent [34]. In parallel, LLM-focused work has expanded quickly but remains heterogeneous in design and endpoints, with recurring concerns about standardization, evaluation bias, and the gap between “plausible-sounding” recommendations and bedside usability [42]. Prior research positioned chatbots as potentially useful adjuncts to stewardship, while emphasizing the need for governance,

validation, and human oversight, concerns that are directly addressed by the shadow-mode, offline nature of the current study [10–12].

Most empirical LLM studies in infection-related decision-making have relied on **vignettes or guideline-question formats**, reporting accuracy, completeness, or safety scoring rather than audited, admission-level paired comparisons. For example, LLMs have been assessed for antibiotic advice in general practice scenarios and for guideline compliance in pneumonia-related questions, with conclusions typically framed as “promising but not reliable without oversight” [43]. Beyond vignettes, a key step forward has been evaluation on **real clinical notes**: Williams et al. tested LLM recommendations on emergency department documentation for several clinical tasks, including antibiotic prescription status, highlighting both capability and the need for careful evaluation design [44]. Within stewardship-adjacent workflows, recent work has also explored more constrained tasks (e.g., antimicrobial classification at scale, or structured treatment recommendations anchored to rapid diagnostic outputs), where LLM performance can be strong but still depends on local framing and safeguards [45]. Collectively, this body of work supports a cautious interpretation: LLMs may reproduce guideline-like defaults and can assist with standardized outputs, yet they require domain constraints, quality control, and evaluation against clinically meaningful processes.

Against that backdrop, this study’s contribution is the shift from “Is the answer correct in a vignette?” to “How do LLM recommendations compare to the clinician’s regimen **for the same admission**, under a prespecified stewardship lens?” Compared with prior LLM evaluations that focus on vignette accuracy or unpaired comparisons, **this work adds a paired admission-level offline-mode benchmark with locally defined guardrails and auditable reproducibility artifacts**. The paired design (N = 493 admissions) reduces confounding by case-mix, while the endpoint definition: contextual guardrail violations and penalties, makes explicit the stewardship trade-off between adequate empiric coverage and avoiding unnecessary escalation. This also directly addresses a recurrent critique in the LLM-healthcare literature: that “accuracy” metrics are not interchangeable with operational safety, and that evaluation must be tied to the intended clinical process and context [42].

Two additional features further distinguish the approach implemented in this research. First, the study embeds decision-making in **local constraints** (formulary, local epidemiology priors/guardrails), which is often missing from generic LLM demonstrations and is a major determinant of stewardship appropriateness. Second, the analysis is unusually **audit-ready**: locked cohorts, prespecified endpoints, QC checks, and reproducible artifacts allow reviewers to interrogate robustness rather than infer it. That combination: paired shadow-mode evaluation, locally meaningful guardrails, and reproducibility engineering, puts the research closer to the methodological standard expected in higher-tier venues than many early LLM vignette papers, while still appropriately positioning the evidence as process benchmarking rather than outcomes validation [34].

3.7. Strengths

Several methodological strengths support the interpretability and reproducibility of this work. First, the paired admission-level design (clinician vs LLM regimen for the same admission; N = 493) reduces confounding by case-mix and makes treatment contrasts less sensitive to secular changes in practice or patient mix across the 2020–2024 sampling frame. Second, the study uses pre-specified, operational guardrail endpoints that reflect core stewardship priorities: *avoiding unnecessary escalation to antipseudomonal, carbapenem, and anti-MRSA therapy in the absence of documented high-risk features*, thereby translating stewardship principles into auditable, admission-level metrics. Third, the analysis pipeline was built for audibility: a locked cohort, structured inputs/outputs, reproducible scripts, file hashing/manifesting, and explicit QC checks (including complete endpoint mapping QC in the final dataset) reduce the risk of undisclosed analytic flexibility and facilitate independent verification. Finally, conclusions are supported by complementary endpoints that triangulate the same stewardship signal from different angles: binary violation rates, ordinal/penalty deltas with many

ties, broad-spectrum class use composites and components, acquisition-cost deltas at 24 h (and a prespecified 72 h continued-therapy subset), and DDD/24 h, providing convergent evidence about spectrum selection behavior rather than reliance on a single metric.

3.8. Limitations

This study has important limitations that should be considered when interpreting the findings. As an observational post hoc analysis, it does not establish clinical effectiveness or safety in patient outcomes, nor does it support causal claims about what would have happened had the LLM recommendations been implemented. The clinician regimen reflects real practice, whereas the LLM regimen is a counterfactual comparator generated offline; differences therefore quantify process concordance with prespecified stewardship guardrails, not treatment effects on morbidity or mortality [46].

Generalizability is constrained by context and study size. This was a single-center study from a Romanian tertiary emergency hospital, meaning absolute rates (e.g., violation frequencies, cost deltas) may depend heavily on local formulary constraints, unit prices, prescribing culture, stewardship norms, and resistance ecology (a locally derived epidemiology prior). Additionally, the analytic cohort was derived from an infection-suspected sampling frame and therefore does not represent all Internal Medicine admissions. The sample size was determined by the prespecified chart-review capacity and year-stratified sampling plan; precision for paired effect estimates is appropriately reflected in the reported confidence intervals. Accordingly, external validity is uncertain; multicenter replication and prospective shadow-mode evaluation are needed before drawing conclusions about broader applicability. That said, the evaluation framework: paired admission-level benchmarking using transparent guardrails, is portable, even if guardrail definitions and cost inputs require local re-specification.

Retrospective EHR data introduce measurement error and documentation bias. Manual abstraction was performed by a single investigator; while supervised and QC-checked, residual abstraction error and documentation bias may persist and could affect syndrome labels, severity proxies, and guardrail triggers. Syndrome labels, severity proxies, and “risk modifier” flags are susceptible to misclassification and incomplete documentation, which can influence guardrail adjudication. For structured variables defining guardrail triggers (e.g., severe sepsis criteria, comorbidities), the absence of documentation in the electronic health record was pragmatically treated as the absence of the condition, reflecting standard clinical prescribing constraints under uncertainty. In particular, guardrails that trigger escalation only when high-risk features are documented can create asymmetry: a clinician may have acted on bedside cues or evolving clinical trajectory that were not captured in the structured record, whereas the LLM is necessarily limited to recorded variables and cannot replicate physical examination, real-time reassessment, or dynamic response to therapy. This limitation cuts both ways: documentation gaps can make clinician choices appear “non-concordant” with guardrails, while also restricting the LLM from safely escalating when unrecorded risk is present.

Furthermore, because some severity/support markers may be recorded after initial antibiotic initiation within the 24 h window, information parity at the exact decision timestamp cannot be fully guaranteed. This potential “look-ahead” bias may skew estimates in favor of the LLM and reinforces that results quantify rule-based concordance within a 24 h management frame, not causal clinical effectiveness.

Multiplicity and endpoint hierarchy require explicit framing. Although the primary endpoint and key secondary endpoint were prespecified, additional secondary endpoints (components, composites, cost, DDD, subsets) increase the risk of false-positive findings if over-interpreted. We therefore state explicitly which endpoints are prespecified (primary + key secondary) versus supportive/exploratory, and for the prespecified paired-delta family we report Holm-adjusted p-values and interpret them cautiously [47].

Subset analyses are particularly vulnerable to selection effects. The microbiology-evaluable subset (N = 158; 335/493 not eligible for microbiological evaluation) is defined by post-baseline processes: culture acquisition, organism recovery, and interpretable susceptibility, that correlate with severity, diagnostic workup, and prior antibiotics. Conditioning on evaluability can induce collider/selection bias, so differences in coverage within this subset should be interpreted as subset-level process benchmarking and hypothesis-generating, not cohort-wide effectiveness [39,48].

Economic estimates are partial and setting-dependent. The cost analysis reflects drug acquisition costs (based on local unit pricing assumptions) within 24 h (and a prespecified 72 h continued-therapy subset), and does not capture administration costs, monitoring, adverse events, downstream complications, or opportunity costs. Any extrapolation to institutional budgets should be presented as scenario analysis with explicit assumptions rather than as observed savings.

Finally, because LLM outputs are model/version and prompt-context dependent, performance may vary across systems, time, and governance constraints; any clinical translation would require prospective validation, human oversight, and safety monitoring under an approved protocol.

3.9. Implications and Next Steps

These findings support further prospective validation under governance safeguards, with outcome-focused endpoints and stewardship oversight. A pragmatic next step is a **prospective shadow-mode pilot** in which LLM recommendations are generated in real time but remain **hidden from clinicians**, allowing robust monitoring of model stability, data-flow integrity, and safety signals without influencing care. This phase can also predefine and validate operational endpoints (e.g., guardrail concordance, spectrum class use, time-to-appropriate therapy when microbiology is available) and confirm the feasibility of capturing key covariates reliably (severity proxies, source control, allergy history, renal dosing constraints) under routine workflow (Figure 7).

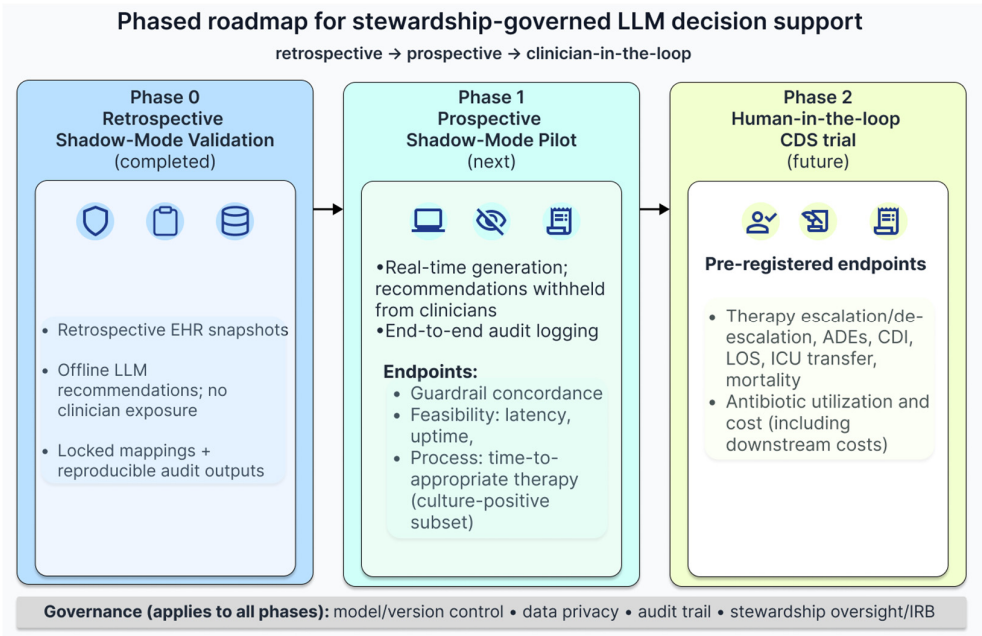


Figure 7. Phased roadmap for stewardship-governed translation of LLM-based empiric antibiotic decision support. Phase 0 corresponds to the present retrospective study, implementing admission-level shadow-mode benchmarking with locally defined guardrails and cost/DDD endpoints, without clinician exposure. Phase 1 illustrates a prospective shadow-mode pilot in which real-time LLM recommendations remain withheld from clinicians, while audit logging, drift monitoring, and prespecified process endpoints are evaluated. Phase 2 depicts a subsequent human-in-the-loop clinical decision support (CDS) trial with stewardship oversight and pre-registered outcome-focused endpoints. Governance requirements, including model/version control, data

privacy, audit trail, and stewardship/IRB oversight, apply across all phases. The diagram is conceptual and does not imply that clinical outcome improvement was assessed in the retrospective phase. Abbreviations: LLM, large language model; CDS, clinical decision support; IRB, institutional review board; DDD, defined daily dose.

If shadow-mode performance remains acceptable, the subsequent step would be a **human-in-the-loop stewardship deployment** rather than autonomous decision-making: the model functions as a **second opinion** that presents one or more empiric options with explicit rationale and **guardrail-aware warnings** (e.g., “anti-MRSA not indicated without X/Y/Z”; “carbapenem reserved unless ESBL risk features present”), while the final decision remains with the clinician and/or stewardship team. This design aligns with established antimicrobial stewardship principles and helps mitigate known risks of LLMs (hallucination, overconfidence, context omission) through structured constraints, audit logging, and escalation pathways to infectious diseases or stewardship review when recommendations deviate from policy.

Critically, prospective evaluation should be **pre-registered** with a clear endpoint hierarchy and analysis plan to minimize analytic flexibility and to support credible inference. Where feasible, a stepped-wedge or cluster-randomized design could compare clinician-only care versus CDS-supported care while tracking both **process endpoints** (appropriateness, broad-spectrum use, de-escalation timing) and **patient-centered outcomes** (clinical deterioration, ICU transfer, length of stay, adverse drug events, CDI, mortality), alongside economic endpoints that capture downstream costs rather than drug acquisition alone. Finally, governance should include explicit model version control, periodic recalibration against local epidemiology, and data protection measures to ensure that any translation to clinical practice remains safe, compliant, and clinically acceptable.

4. Materials and Methods

4.1. Study Design, Setting, and Timeframe

We conducted a **single-center, retrospective, post hoc paired evaluation** of empiric antibiotic regimens prescribed during the first 24 h of admission in an Internal Medicine department at **Clinical Emergency Hospital of Bucharest (Spitalul Clinic de Urgență București; commonly known as “Floreasca”)** across **2020–2024**. Recommendations from a large language model were generated **offline after care was delivered**, were **not visible to clinicians**, and **did not influence** clinical management (non-interventional evaluation). Reporting followed the STROBE framework for observational studies [46].

Unit of analysis was the **admission** (crt_id: a de-identified study admission identifier), with **paired regimens** per admission: (i) the clinician-prescribed empiric regimen and (ii) the post hoc LLM-recommended regimen, each defined over the **first 24 h**.

4.2. Participants: Source Population, Sampling, and Flow

4.2.1. Source Population Sampling Frame

The source population comprised **14,879** Internal Medicine admissions at Clinical Emergency Hospital of Bucharest between **2020 and 2024**. Using an infection-suspected filter applied to admission diagnoses, **6,734** candidate admissions were identified for potential eligibility screening. The infection-suspected screening frame was constructed using free-text admission diagnoses. An LLM-assisted categorization was used only as a clerical aid to pre-sort diagnosis strings into broad syndromic buckets (e.g., UTI, LRTI/pneumonia, IAI, SSTI, sepsis syndromes, COVID, other infections, non-infectious). Importantly, this automated pre-sorting was not used to determine eligibility. All chart-reviewed candidate admissions underwent manual verification of the admission documentation by the investigator, and final classification as infectious vs non-infectious (and all inclusion/exclusion decisions) was made manually during chart review; therefore, no admission entered the analytic cohort based solely on an automated label. Any residual misclassification at the

screening-frame stage would affect only which cases were prioritized for chart review, not the paired endpoint computation within the final manually verified analytic cohort.

4.2.2. Random Sampling and Manual Screening (Year-Stratified)

Within each calendar-year stratum, candidate admissions were randomly ordered using a spreadsheet random number generator (Google Sheets RAND()) and then manually screened in that randomized order. The target sample size of 500 admissions (100 per year) was chosen pragmatically on the basis of manual chart-review capacity while preserving balanced year-stratified sampling across 2020–2024. No formal a priori power calculation was used for cohort construction; statistical precision is therefore conveyed by the reported confidence intervals around the paired effect estimates. The randomized ordering was retained as a locked snapshot (exported and archived) before detailed chart abstraction to preserve auditability of the screening sequence. A total of 841 candidate admissions underwent manual chart review. Of these, 500 were enrolled (100 per year) for detailed data abstraction. After applying a prespecified deduplication rule (one admission per patient within the sampled frame, defined as the first eligible admission), the final analytic cohort comprised 493 paired admissions.

4.2.3. Eligibility Assessment and Reasons for Non-Inclusion

During chart-review screening, non-enrollment was categorized using prespecified chart-based reasons: non-infectious admission; no systemic antibiotics within 24 h; and missing medication documentation (reported in the STROBE flow diagram and Supplementary Table S1). Separately, after enrollment of 500 eligible admissions (100/year), a prespecified deduplication step (one admission per patient within the sampled frame, defined as the first eligible admission) removed 7 admissions, yielding the final analytic cohort of 493 paired admissions.

A de-identified screening log (screen_id, year, randomized order value, inclusion decision, and categorized non-inclusion reason) is provided as Supplementary Data File S1.

4.2.4. Bias and Representativeness

To reduce selection bias and ensure temporal coverage, screening was performed in a year-stratified random order with a fixed annual quota (100 admissions/year). The study was retrospective and single-center, and the analytic cohort represents infection-suspected admissions meeting eligibility criteria and requiring empiric antibiotics within 24 h; findings are therefore intended as process-level benchmarking and may not generalize to other institutions with different case-mix or resistance ecology.

4.3. Data Sources and Manual EHR Abstraction

4.3.1. Data Sources and Measurement

Clinical variables were abstracted from the hospital electronic health record (Hipocrate, Romania) and associated laboratory systems into a structured analytic dataset (data.xlsx) using a pre-specified data dictionary with structured fields whenever feasible. Abstraction captured (as available) demographics, comorbidities, severity proxies within 24 h, healthcare exposure/MDR-risk variables, prior microbiology flags, and antibiotic agents initiated within the first 24 h.

No formal inter-rater reliability exercise was prospectively documented; however, abstraction decisions were informally supervised by senior clinicians and were complemented by rule-based consistency checks within the analysis pipeline. To mitigate single-abstractor bias, we used a prespecified data dictionary, locked variable definitions prior to extraction, and implemented automated QC flags (range checks, completeness checks, and antibiotic-code mapping audits) that were reviewed prior to endpoint computation. Quality control (QC) procedures included (1) completeness checks for all variables contributing to primary and secondary endpoints; (2) range and plausibility checks for numeric fields (e.g., dose, duration); (3) automated detection of unmapped or

deprecated antibiotic codes using a locked reference dictionary; and (4) generation of QC flags that informed predefined exclusion criteria for stewardship-related analyses.

Missingness for all key structured variables (baseline covariates and all fields required for endpoint computation) is summarized in Supplementary Table S6 (n missing and % missing per variable). Model-level missingness and complete-case QC for the exploratory regression analyses are provided in Supplementary Table S6a. Primary and key secondary endpoints were computed without imputation; analyses were performed as complete-case with respect to the variables required for each endpoint module, and denominators are reported per analysis.

4.4. Local Epidemiology Module (Laboratory Exports → Structured Prompt Context)

Because empiric antibiotic selection is influenced by **facility-specific resistance ecology**, we constructed a **local epidemiology context module** from hospital microbiology and susceptibility exports (laboratory information system). This module summarizes local pathogen/resistance patterns relevant to empiric therapy and is conceptually aligned with syndrome-oriented extensions of antibiograms, including **weighted-incidence syndromic combination antibiograms (WISCA)**, which integrate syndrome-specific organism incidence with susceptibility to estimate expected coverage of candidate regimens (including combinations) [49].

To support **information parity**, the epidemiology module was injected into the LLM prompt as **static structured context**, reflecting the pragmatic notion that clinicians acquire local ecology knowledge through institutional feedback loops whereas a general-purpose LLM does not. Specifically, the epidemiology module was provided as a structured JSON/text block representing pathogen frequencies and specific regimen susceptibilities (WISCA) for the given ward and year, enabling the LLM to ground its choices in empirical local data. Clinicians were assumed to have access to local antibiograms and experiential feedback in routine care; the injected context was intended to approximate this routinely available institutional knowledge for the model. This epidemiology module was computed **independently of the 493-admission analytic cohort** and was **not used to define outcomes and not tuned to study endpoints or metrics** (anti-tuning separation).

Because the local epidemiology module was derived from annual ward-level laboratory exports, admissions occurring early in a given calendar year may have been exposed to contextual data reflecting isolates identified later in that same year. The module was used exclusively as static supportive context to approximate institutional ecology awareness and was not used to define endpoints or tuned to study metrics.

4.5. LLM Recommendation Generation (OpenAI API) and Reproducibility/Auditability

Time zero was defined as hospital admission. The clinician regimen was defined as the set of systemic antibiotics initiated within the first 24 h after admission (the 0–24 h empiric management window). Clinician empiric regimens were derived from the pharmacy dispensing request records for the initial empiric order set. To ensure comparability with the single-time-point LLM recommendation, we verified that agents included in a multi-drug regimen were initiated concurrently (i.e., no sequential add-ons occurred within the 0–24 h window in the curated cohort). To mirror this definition, the LLM prompt was constructed from de-identified structured variables documented during the same 0–24 h window (e.g., admission labs and prespecified severity/support flags), and the model was asked to recommend an empiric regimen for that window under a shadow-mode, non-interventional design. Because both arms were defined over the same 0–24 h interval, this evaluation benchmarks early empiric management as a stewardship-oriented process measure rather than a strict admission-time (minute-0) bedside decision.

Building on this clinical framework, LLM recommendations were generated via the **OpenAI API** [50] using the Responses endpoint (model identifier returned by the API at execution time: gpt-5.2). We used a fixed prompt template populated with the aforementioned structured patient context and the local epidemiology module. Recommendations were produced post hoc and did not influence care. Requests were executed between 2026-02-07T13:23:14Z and 2026-02-07T13:39:43Z. We

utilized temperature = 0.0 and max_output_tokens = 800, with up to three retry attempts for transient failures. For privacy, we set store=false in API requests. All other generation parameters were left at API defaults (e.g., top_p = 1.0; presence_penalty = 0; frequency_penalty = 0).

Outputs were constrained to a structured format (validated against a strict JSON schema) mapped to a predefined antibiotic dictionary to enable deterministic downstream scoring and auditing. Model calls were audit-logged (request/response metadata, timestamps, attempt counters, and response identifiers) and integrity-protected using cryptographic hashing (SHA-256) with a versioned manifest enabling reconstruction of the reported aggregate results. The public audit bundle provides the prompt template with placeholders only, output constraints, allowed regimen code list, JSON schemas, mapping tables, hashed manifests, and derived summary outputs sufficient to reproduce all reported statistics; raw per-admission prompts and full model responses remain restricted due to privacy and institutional governance. The full prompt template and structured-output specification are provided in Supplementary Appendix 1. Evaluation/reporting practices were informed by early-stage clinical AI guidance and LLM-specific reporting considerations [24,25].

4.6. Outcomes (Contextual Stewardship Guardrails)

4.6.1. Regimen Definition (24 H)

For each admission, the clinician empiric regimen was defined as the set of systemic antibacterial agents initiated concurrently as the initial empiric order set within the first 24 h of admission (0–24 h empiric management window), as verified using pharmacy dispensing request timestamps. The LLM regimen was defined as the antibacterial agents recommended by the LLM for the same 24 h empiric decision point, normalized through the same antibiotic dictionary/mapping.

4.6.2. Primary (Prespecified) Endpoint

The primary endpoint was any contextual guardrail violation (paired binary), defined as the presence of ≥1 unjustified broad-spectrum element within the 24 h empiric regimen under pre-specified structured context rules.

Contextual guardrail violations and penalty. For each admission and each arm (clinician vs LLM), we evaluated three prespecified contextual guardrail components: carbapenem use, antipseudomonal use, and anti-MRSA use. A component violation was defined as antibiotic class use in the absence of prespecified justification proxies (binary rules), yielding component indicators I_k . The contextual guardrail penalty was computed as a weighted sum of component violations:

$$\text{Penalty} = \sum_k w_k I_k,$$

where $k \in \{\text{carb}, \text{APS}, \text{MRSA}\}$, $I_k = 1$ if component k is violated (i.e., the class is used without a prespecified justification proxy) and 0 otherwise. The prespecified weights were $w_{\text{carb}} = 2$, $w_{\text{APS}} = 1$, and $w_{\text{MRSA}} = 1$; equivalently, $\text{Penalty} = 2 \cdot I_{\text{carb}} + 1 \cdot I_{\text{APS}} + 1 \cdot I_{\text{MRSA}}$. The paired delta was defined as $\Delta \text{Penalty} = \text{Penalty}(\text{LLM}) - \text{Penalty}(\text{clinician})$ (negative values favor the LLM arm). Full component definitions (trigger logic, data fields, mappings, and weights) are provided in Supplementary Table S3.

Justification proxies used in guardrails. Severity proxy (SEVERE) was defined using prespecified early clinical indicators (ICU admission/transfer within the first 24 h, septic shock documentation, vasopressor use, mechanical ventilation, or respiratory failure documentation). Multidrug-resistant risk proxy (MDR_RISK) was defined using prespecified structured history/setting fields (prior ESBL/CRE/VRE, systemic antibiotic exposure in the prior 90 days, hospitalization in the prior 90 days, long-term care facility residence, or healthcare-associated acquisition setting). Anti-MRSA justification was defined by documented prior MRSA colonization. Exact Boolean logic and field names are listed in Supplementary Table S3.

4.6.3. Secondary/Supplementary Endpoints (Multiplicity Caution)

Secondary and supplementary analyses were prespecified as non-confirmatory and interpreted with caution regarding multiplicity. These included:

- empiric antibiotic cost differences at 24 h (N = 493) and at 72 hours only in admissions where clinician empiric therapy was continued for ≥ 3 days (N = 323);
- Δ DDD/24 h (N = 493) calculated using the WHO ATC/DDD methodology [51];
- microbiology-evaluable coverage analyses (paired; N = 158; descriptive due to selection);
- regimen concordance measures;
- AWARe class distributions (supplementary; interpretation-sensitive) based on the WHO AWARe framework [15];
- exploratory multivariable models (complete-case; N = 493), reported as exploratory only

Microbiology-evaluable admissions were defined as admissions with ≥ 1 clinical culture collected within the first 24 h of admission (collection timestamp) that yielded an index organism with interpretable antimicrobial susceptibility results, and that could be plausibly linked to the presumed infection syndrome used for empiric therapy assessment. Paired evaluability required that active-coverage classification be possible in both arms; admissions not meeting these criteria in either arm were coded as non-evaluable (not missing) and excluded from the paired coverage analysis. Because culture collection/positivity and susceptibility availability are post-baseline clinical processes correlated with severity and clinical trajectory, microbiology-evaluable analyses are interpreted as subset-level exploratory process benchmarking and are inherently susceptible to selection bias.

AWARe-derived delta metrics (supplementary). For regimen-level AWARe summarization, we derived two paired metrics: Δ AWARe mean score and Δ AWARe max score (LLM-clinician), computed from the regimen's agent-level AWARe mapping (see Table S10a–S10c). These metrics are reported as supplementary paired-delta outcomes and included in the prespecified paired-delta multiplicity family (Table S12).

4.7. Statistical Analysis

Analyses followed a pre-specified hierarchy: (1) primary endpoint tested first; (2) key secondary endpoint interpreted as prespecified conditional on primary significance; all other endpoints were treated as secondary/supplementary/exploratory with explicit multiplicity caution. Missing data were handled using a complete-case approach for each analysis; denominators therefore reflect admissions with non-missing outcome data in both paired arms (and, where applicable, membership in prespecified subsets). No imputation was performed. Regarding the paired primary and key secondary endpoints, all included admissions had complete endpoint data (N = 493). For endpoints defined on restricted subsets (e.g., 72 h continuation, microbiology-evaluable subset), analyses were performed only in admissions meeting the predefined evaluability criteria, with denominators reported explicitly.

4.7.1. Primary Endpoint (Paired Binary)

For the paired binary primary endpoint, we used the exact McNemar test based on discordant pairs. We report the paired 2x2 table (n00, n01, n10, n11), the paired risk difference (RD, LLM–Clin), and a matched odds ratio estimated from discordant pairs using a continuity correction: $OR = (n01 + 0.5) / (n10 + 0.5)$. A 95% confidence interval was computed using a Wald interval on the log scale: $\log(OR) \pm 1.96 \times \sqrt{(1 / (n01 + 0.5) + 1 / (n10 + 0.5))}$.

4.7.2. Key Secondary Endpoint (Paired Integer Deltas with Many Ties)

For Δ contextual guardrail penalty (LLM–Clin), we report median and IQR, the number of ties, and compare paired deltas using the Wilcoxon signed-rank test and a sign test on non-tied pairs

(appropriate when ties are frequent) [52]. We additionally computed a bootstrap 95% confidence interval for the median Δ by resampling admissions with replacement ($n_boot = 5000$, $seed = 7$) [53].

4.7.3. Secondary Analyses and Exploratory Modeling

Secondary paired continuous outcomes (e.g., cost deltas, $\Delta DDD/24$ h) were analyzed using paired nonparametric methods (Wilcoxon signed-rank), reporting denominators per module; when ties were frequent, we additionally report a sign test on non-tied pairs (prespecified for Δ contextual guardrail penalty and reported as a sensitivity analysis for other paired-delta endpoints); Wilcoxon signed-rank tests use the standard handling of zero differences per the implementation. DDD calculations followed WHO ATC/DDD guidance [51]. To provide transparent multiplicity control across the prespecified paired-delta family ($m = 6$), we report Holm step-down adjusted p-values in Supplementary Table S12 for: Δ contextual guardrail penalty (key secondary), Δ empiric antibiotic cost at 24 h, Δ empiric antibiotic cost at 72 h (continued-therapy subset), Δ AWARe mean score, Δ AWARe max score (derived from regimen-level AWARe mapping; see Table S10), and $\Delta DDD/24$ h (all two-sided). Unadjusted p-values are reported alongside each endpoint (Table 2 for main-text endpoints; Supplementary tables for supplementary endpoints). Inference for the prespecified endpoint hierarchy followed the prespecified strategy (primary tested first; key secondary interpreted conditional on primary). Holm-adjusted p-values are reported for the prespecified paired-delta family for transparency but are not used to re-define the primary/key-secondary hierarchy. Where an outcome is defined on a prespecified subset (e.g., 72 h cost), the corresponding p-value is computed on that subset denominator and then included in the Holm family. Microbiology-evaluable coverage analyses were treated as subset-level exploratory process benchmarking due to selection into the evaluable subset. The microbiology-evaluable paired subset was defined as admissions with an identified index organism and interpretable susceptibility allowing classification of active empiric coverage in both arms; all other admissions were classified as non-evaluable. Because culture acquisition/positivity are post-baseline processes correlated with severity and clinical trajectory, analyses in this subset are interpreted as process benchmarking rather than causal effectiveness. Selection-bias QC comparing evaluable vs non-evaluable admissions is reported in Supplementary Table S2.

Exploratory multivariable models were conducted as complete-case analyses ($N = 493$) and reported as exploratory only, with attention to instability/separation risks typical in sparse paired events.

For structured variables defining guardrail triggers (e.g., severe sepsis criteria, comorbidities), the absence of documentation in the electronic health record was pragmatically treated as the absence of the condition, reflecting standard clinical prescribing constraints under uncertainty.

Empiric antibiotic costs were estimated as drug acquisition costs only, using a fixed unit-price mapping applied to standardized regimen codes; administration, monitoring, and hospitalization costs were not included. Unit prices (including price year/source) and the EUR conversion assumption are specified in Supplementary Data File S2 (“data_file_s2_costing”); costs therefore reflect drug acquisition only, without administration/monitoring or downstream costs.

4.7.4. Software and Reproducibility

All analyses were conducted using a fully scripted and version-controlled pipeline implemented in Python (version 3.14.x), with pandas for data manipulation, NumPy and SciPy for numerical operations, and statsmodels for statistical modeling. Analysis scripts, data dictionaries, antibiotic code-mapping tables, JSON schemas, and non-identifiable audit artifacts (including file manifests and cryptographic checksums) are publicly available in a versioned repository (Zenodo: DOI 10.5281/zenodo.18731938).

4.8. Ethics, Data Governance, and Availability Statements

The study protocol was reviewed and approved by the Ethics Committee of the Clinical Emergency Hospital of Bucharest, approval no. 5352, issued on 02 July 2025, prior to data analysis. The study involved no direct patient contact and no modification of standard clinical care. Data processing and storage followed applicable data protection requirements (GDPR) [54].

All analyses were conducted using a fully scripted and auditable pipeline. A complete reproducibility package including analysis code, data dictionaries, code-mapping tables, JSON schemas, derived aggregate tables, and audit artifacts (e.g., hashes and file manifests), is publicly available at DOI 10.5281/zenodo.18731938.

Patient-level EHR-derived data and raw per-admission LLM prompt/response payloads are not publicly shared due to privacy, legal (GDPR), and institutional governance constraints. Access to de-identified patient-level data or to a controlled analysis environment may be considered upon reasonable request, subject to institutional approvals, a data-sharing agreement, and applicable data-protection regulations.

Regarding the LLM input privacy safeguards, no direct identifiers were transmitted to the API. All admissions were processed using pseudonymous study identifiers. Prompts contained only de-identified, structured clinical variables (e.g., age, sex, coded comorbidities, syndrome labels, and selected admission laboratory values). We strictly excluded direct identifiers (names, national identifiers, addresses), exact calendar dates, and free-text clinical notes; temporal information was provided only in relative or coarse form where needed for clinical interpretation. Data were processed in a controlled environment with access restrictions, API requests were sent with store = false, and all payloads were retained locally in an auditable log. Consistent with the vendor’s policy for business/API services, inputs and outputs are not used for model training by default, unless an organization explicitly opts in; retention behavior (e.g., abuse-monitoring logs and application-state retention) follows the platform’s documented data controls [55].

5. Conclusions

In this single-center retrospective **shadow-mode**, paired evaluation of **493 admissions**, LLM-generated empiric antibiotic regimens were more concordant with locally defined **contextual stewardship guardrails** than clinician regimens, with fewer broad-spectrum escalations and modestly lower **24 h** acquisition costs. These findings should be interpreted as **process-level stewardship benchmarking** rather than evidence of clinical effectiveness or safety. Prospective **shadow-mode** validation followed by a **human-in-the-loop** stewardship trial with pre-registered endpoints and governance safeguards is warranted to determine impact on patient-centered outcomes. No patient-centered outcomes were evaluated, and no causal conclusions about clinical benefit or harm can be drawn from this study.

Supplementary Materials: The following supporting information is provided: Preprints.org, Table S1 (Cohort flow / inclusion–exclusion decision counts); Table S2 (Microbiology-evaluable subset: selection-bias QC, evaluable vs non-evaluable); Table S3 (Contextual guardrail rule specification, including trigger logic, variable mapping, and penalty weights); Table S4 (NO_ANTIBIOTIC cases: case-level QC summary); Table S5 (Predictors of nonzero Δ guardrail penalty; multivariable logit with robust SE); Table S5a (Predictors of LLM improvement in contextual guardrails; multivariable logit with robust SE); Table S6 (Missingness map for baseline covariates and endpoint-defining variables; n missing and % missing per variable; includes not-applicable/nonevaluable where subset-defined); Table S6a (Missingness and complete-case QC); Table S7 (Regimen concordance summary); Table S8 (Microbiology-evaluable composite: active coverage AND no contextual violation); Table S9 (Contextual guardrail component endpoints); Table S10a–S10c (AWaRe binaries, transitions, and unmapped QC); Table S11 (Low-risk subgroup analyses); data_file_S1 (De-identified screening log for chart-reviewed candidate admissions; includes decision codes and year); Table S12 (Holm step-down adjustment across the prespecified paired-delta family, $m = 6$; includes the key secondary Δ penalty plus five additional paired-delta outcomes; two-sided tests); data_file_s2_costing (Costing assumptions - drug acquisition only);

supplementary_appendix_1_prompt_template (full prompt template - placeholders only - output constraints, allowed regimen code list, and strict response format).

Author Contributions: Conceptualization, N.I.A.; Methodology, N.I.A.; Software, N.I.A.; Validation, N.I.A., C.-L.T., G.G., V.A.I., and C.C.D.; Formal Analysis, N.I.A.; Investigation, N.I.A.; Resources, C.C.D.; Data Curation, N.I.A.; Writing—Original Draft Preparation, N.I.A.; Writing—Review & Editing, all authors; Visualization, N.I.A.; Supervision, C.C.D.; Project Administration, C.C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by the authors.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Clinical Emergency Hospital of Bucharest (approval no. 5352; 02 July 2025). The study involved no direct patient contact and no modification of standard clinical care.

Informed Consent Statement: Patient consent was waived due to the retrospective nature of the study and the use of de-identified data, in accordance with Ethics Committee/IRB guidance.

Data Availability Statement: Derived, non-identifiable audit artifacts (analysis code, schemas, mapping tables, derived summary tables, and file manifests with SHA-256 checksums) are available at DOI 10.5281/zenodo.18731938. With the exception of a limited de-identified case-level QC table for the five NO_ANTIBIOTIC admissions (Supplementary Table S4), individual-level EHR-derived data and per-admission LLM payloads remain restricted due to institutional governance and GDPR.

Reproducibility / Code availability statement: The public audit bundle includes the fully scripted statistical analysis pipeline, antibiotic code-mapping tables, epidemiology/QC scripts, and derived analysis outputs required to reproduce the reported summary statistics from the restricted dataset. The bundle also contains file manifests and cryptographic SHA-256 checksums enabling integrity verification. The LLM request/response logs and patient-level prompt inputs are not publicly shared because they contain admission-level derived clinical data.

Acknowledgments: During the preparation of this manuscript, the authors used OpenAI ChatGPT for language editing. Python (3.14.x) was used for statistical analyses and for the generation of Figure 6. Figure 7 was prepared manually by the authors. The authors reviewed and edited all outputs and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Definition
AI	Artificial intelligence
AMR	Antimicrobial resistance
AMS	Antimicrobial stewardship
APS	Antipseudomonal broad-spectrum β -lactam
ASP	Antimicrobial stewardship program
AWaRe	Access, Watch, Reserve (WHO antibiotic classification)
CDI	<i>Clostridioides difficile</i> infection
CI	Confidence interval
DDD	Defined daily dose
EHR	Electronic health record
ICU	Intensive care unit

IQR	Interquartile range
LLM	Large language model
MDR	Multidrug-resistant
MRSA	Methicillin-resistant <i>Staphylococcus aureus</i>
ESBL	Extended-spectrum beta-lactamase
CRE	Carbapenem-resistant <i>Enterobacterales</i>
VRE	Vancomycin-resistant <i>enterococci</i>
RD	Risk difference

References

1. Murray, C.J.L.; Ikuta, K.S.; Sharara, F.; Swetschinski, L.; Aguilar, G.R.; Gray, A.; Han, C.; Bisignano, C.; Rao, P.; Wool, E.; et al. Global Burden of Bacterial Antimicrobial Resistance in 2019: A Systematic Analysis. *The Lancet* **2022**, *399*, 629–655, doi:10.1016/S0140-6736(21)02724-0.

2. O'Neill, J. (2016) Tackling Drug-Resistant Infections Globally Final Report and Recommendations. Review on Antimicrobial Resistance. Wellcome Trust and HM Government. Available online: https://amr-review.org/sites/default/files/160525_Final%20paper_with%20cover.pdf (accessed on 18 February 2026).

3. Naghavi, M.; Vollset, S.E.; Ikuta, K.S.; Swetschinski, L.R.; Gray, A.P.; Wool, E.E.; Aguilar, G.R.; Mestrovic, T.; Smith, G.; Han, C.; et al. Global Burden of Bacterial Antimicrobial Resistance 1990–2021: A Systematic Analysis with Forecasts to 2050. *The Lancet* **2024**, *404*, 1199–1226, doi:10.1016/S0140-6736(24)01867-1.

4. Uddin, T.M.; Chakraborty, A.J.; Khusro, A.; Zidan, B.R.M.; Mitra, S.; Emran, T.B.; Dhama, K.; Ripon, Md.K.H.; Gajdacs, M.; Sahibzada, M.U.K.; et al. Antibiotic Resistance in Microbes: History, Mechanisms, Therapeutic Strategies and Future Prospects. *J. Infect. Public Health* **2021**, *14*, 1750–1766, doi:10.1016/j.jiph.2021.10.020.

5. Oliveira, M.; Antunes, W.; Mota, S.; Madureira-Carvalho, Á.; Dinis-Oliveira, R.J.; Silva, D.D. da An Overview of the Recent Advances in Antimicrobial Resistance. *Microorganisms* **2024**, *12*, doi:10.3390/microorganisms12091920.

6. Kanthali, M.; Bhagwat, G.; Pathak, A.; Purohit, M. Antibiotic Stewardship through Clinical Data Digitization: Perceived Opportunities and Obstructions by Medical Doctors from Semi-Urban Setting in Central India. *Front. Digit. Health* **2025**, *7*, doi:10.3389/fdgth.2025.1652086.

7. Yoon, Y.K.; Kwon, K.T.; Jeong, S.J.; Moon, C.; Kim, B.; Kiem, S.; Kim, H.; Heo, E.; Kim, S.-W. Guidelines on Implementing Antimicrobial Stewardship Programs in Korea. *Infect. Chemother.* **2021**, *53*, 617–659, doi:10.3947/ic.2021.0098.

8. Amin, S.U.; Guizani, M.; Hossain, M.S. Advances, Evaluation, and Explainability of Large Language Models in Healthcare: A Systematic Review. *ACM Trans Multimed. Comput Commun Appl* **2026**, *22*, 60:1-60:32, doi:10.1145/3786334.

9. Artsi, Y.; Sorin, V.; Glicksberg, B.S.; Korfiatis, P.; Freeman, R.; Nadkarni, G.N.; Klang, E. Challenges of Implementing LLMs in Clinical Practice: Perspectives. *J. Clin. Med.* **2025**, *14*, doi:10.3390/jcm14176169.

10. Pinto, A.; Pennisi, F.; Ricciardi, G.E.; Signorelli, C.; Gianfredi, V. Evaluating the Impact of Artificial Intelligence in Antimicrobial Stewardship: A Comparative Meta-Analysis with Traditional Risk Scoring Systems. *Infect. Dis. Now* **2025**, *55*, 105090, doi:10.1016/j.idnow.2025.105090.

11. Antonie, N.I.; Gheorghe, G.; Ionescu, V.A.; Tiucă, L.-C.; Diaconu, C.C. The Role of ChatGPT and AI Chatbots in Optimizing Antibiotic Therapy: A Comprehensive Narrative Review. *Antibiotics* **2025**, *14*, doi:10.3390/antibiotics14010060.

12. Al Mazrouei, N.; Ahmed Elnour, A.; Badi, S.; Alsulami, F.T.; Awadallah Mohamed Saeed, A.; Awad Al-Kubaisi, K.; Menon, V.; Yousif Khidir, I.; Ismail, M.; Osman Mahagoub, M.M.; et al. The Impact of Artificial Intelligence on the Prescribing, Selection, Resistance, and Stewardship of Antimicrobials: A Scoping Review. *BMC Infect. Dis.* **2026**, *26*, 222, doi:10.1186/s12879-025-12320-4.

13. Giamarellou, H.; Galani, L.; Karavasilis, T.; Ioannidis, K.; Karaikos, I. Antimicrobial Stewardship in the Hospital Setting: A Narrative Review. *Antibiotics* **2023**, *12*, doi:10.3390/antibiotics12101557.

14. Rapti, V.; Poulakou, G.; Mousouli, A.; Kakasis, A.; Pagoni, S.; Pechlivanidou, E.; Masgala, A.; Sympardi, S.; Apostolopoulos, V.; Giannopoulos, C.; et al. Assessment of De-Escalation of Empirical Antimicrobial Therapy in Medical Wards with Recognized Prevalence of Multi-Drug-Resistant Pathogens: A Multicenter Prospective Cohort Study in Non-ICU Patients with Microbiologically Documented Infection. *Antibiotics* **2024**, *13*, doi:10.3390/antibiotics13090812.
15. WHO AWaRe System for Antimicrobial Stewardship Available online: <https://www.who.int/teams/surveillance-prevention-control-AMR/control-and-response-strategies/AWaRe> (accessed on 18 February 2026).
16. Karaikos, I.; Giamarellou, H. Carbapenem-Sparing Strategies for ESBL Producers: When and How. *Antibiotics* **2020**, *9*, doi:10.3390/antibiotics9020061.
17. Parente, D.M.; Cunha, C.B.; Mylonakis, E.; Timbrook, T.T. The Clinical Utility of Methicillin-Resistant *Staphylococcus Aureus* (MRSA) Nasal Screening to Rule Out MRSA Pneumonia: A Diagnostic Meta-Analysis With Antimicrobial Stewardship Implications. *Clin. Infect. Dis.* **2018**, *67*, 1–7, doi:10.1093/cid/ciy024.
18. Metlay, J.P.; Waterer, G.W.; Long, A.C.; Anzueto, A.; Brozek, J.; Crothers, K.; Cooley, L.A.; Dean, N.C.; Fine, M.J.; Flanders, S.A.; et al. Diagnosis and Treatment of Adults with Community-Acquired Pneumonia. An Official Clinical Practice Guideline of the American Thoracic Society and Infectious Diseases Society of America. *Am. J. Respir. Crit. Care Med.* **2019**, *200*, e45–e67, doi:10.1164/rccm.201908-1581ST.
19. Trautner, B.W.; Cortés-Penfield, N.W.; Gupta, K.; Hirsch, E.B.; Horstman, M.; Moran, G.J.; Colgan, R.; O'Horo, J.C.; Ashraf, M.S.; Connolly, S.; et al. Clinical Practice Guideline by Infectious Diseases Society of America (IDSA): 2025 Guideline on Management and Treatment of Complicated Urinary Tract Infections: Selection of Antibiotic Therapy for Complicated UTI. *Clin. Infect. Dis.* **2025**, ciaf460, doi:10.1093/cid/ciaf460.
20. Rawson, T.M.; Moore, L.S.P.; Hernandez, B.; Charani, E.; Castro-Sanchez, E.; Herrero, P.; Hayhoe, B.; Hope, W.; Georgiou, P.; Holmes, A.H. A Systematic Review of Clinical Decision Support Systems for Antimicrobial Management: Are We Failing to Investigate These Interventions Appropriately? *Clin. Microbiol. Infect.* **2017**, *23*, 524–532, doi:10.1016/j.cmi.2017.02.028.
21. Dhaliwal, M.; Elligsen, M.; Lam, P.W.; Daneman, N. Weighted-Incidence Syndromic Combination Antibigram (WISCA) to Guide Antibiotic Regimens for Empiric Treatment of Prosthetic Joint Infections: A Retrospective Cohort Study. *CMI Commun.* **2026**, *3*, 105170, doi:10.1016/j.cmicom.2025.105170.
22. Cook, A.; Sharland, M.; Yau, Y.; Bielicki, J. Improving Empiric Antibiotic Prescribing in Pediatric Bloodstream Infections: A Potential Application of Weighted-Incidence Syndromic Combination Antibigrams (WISCA). *Expert Rev. Anti Infect. Ther.* **2022**, *20*, 445–456, doi:10.1080/14787210.2021.1967145.
23. Hebert, C.; Ridgway, J.; Vekhter, B.; Brown, E.C.; Weber, S.G.; Robicsek, A. Demonstration of the Weighted-Incidence Syndromic Combination Antibigram: An Empiric Prescribing Decision Aid. *Infect. Control Hosp. Epidemiol.* **2012**, *33*, 381–388, doi:10.1086/664768.
24. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting Guideline for the Early-Stage Clinical Evaluation of Decision Support Systems Driven by Artificial Intelligence: DECIDE-AI. *Nat. Med.* **2022**, *28*, 924–933, doi:10.1038/s41591-022-01772-9.
25. Gallifant, J.; Afshar, M.; Ameen, S.; Aphinyanaphongs, Y.; Chen, S.; Cacciamani, G.; Demner-Fushman, D.; Dligach, D.; Daneshjou, R.; Fernandes, C.; et al. The TRIPOD-LLM Reporting Guideline for Studies Using Large Language Models. *Nat. Med.* **2025**, *31*, 60–69, doi:10.1038/s41591-024-03425-5.
26. Asgari, E.; Montaña-Brown, N.; Dubois, M.; Khalil, S.; Balloch, J.; Yeung, J.A.; Pimenta, D. A Framework to Assess Clinical Safety and Hallucination Rates of LLMs for Medical Text Summarisation. *Npj Digit. Med.* **2025**, *8*, 274, doi:10.1038/s41746-025-01670-7.
27. Barlam, T.F.; Cosgrove, S.E.; Abbo, L.M.; MacDougall, C.; Schuetz, A.N.; Septimus, E.J.; Srinivasan, A.; Dellit, T.H.; Falck-Ytter, Y.T.; Fishman, N.O.; 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.* **2016**, *62*, e51–e77, doi:10.1093/cid/ciw118.
28. Schoffelen, T.; Papan, C.; Carrara, E.; Eljaaly, K.; Paul, M.; Keuleyan, E.; Martin Quirós, A.; Peiffer-Smadja, N.; Palos, C.; May, L.; et al. European Society of Clinical Microbiology and Infectious Diseases Guidelines

- for Antimicrobial Stewardship in Emergency Departments (Endorsed by European Association of Hospital Pharmacists). *Clin. Microbiol. Infect.* **2024**, *30*, 1384–1407, doi:10.1016/j.cmi.2024.05.014.
29. Baur, D.; Gladstone, B.P.; Burkert, F.; Carrara, E.; Foschi, F.; Döbele, S.; Tacconelli, E. Effect of Antibiotic Stewardship on the Incidence of Infection and Colonisation with Antibiotic-Resistant Bacteria and Clostridium Difficile Infection: A Systematic Review and Meta-Analysis. *Lancet Infect. Dis.* **2017**, *17*, 990–1001, doi:10.1016/S1473-3099(17)30325-0.
 30. Ray, M.J.; Strnad, L.C.; Tucker, K.J.; Furuno, J.P.; Lofgren, E.T.; McCracken, C.M.; Park, H.; Gerber, J.S.; McGregor, J.C. Influence of Antibiotic Exposure Intensity on the Risk of Clostridioides Difficile Infection. *Clin. Infect. Dis.* **2024**, *79*, 1129–1135, doi:10.1093/cid/ciae259.
 31. Bassetti, M.; Rello, J.; Blasi, F.; Goossens, H.; Sotgiu, G.; Tavošchi, L.; Zasowski, E.J.; Arber, M.R.; McCool, R.; Patterson, J.V.; et al. Systematic Review of the Impact of Appropriate versus Inappropriate Initial Antibiotic Therapy on Outcomes of Patients with Severe Bacterial Infections. *Int. J. Antimicrob. Agents* **2020**, *56*, 106184, doi:10.1016/j.ijantimicag.2020.106184.
 32. Paul, M.; Shani, V.; Muchtar, E.; Kariv, G.; Robenshtok, E.; Leibovici, L. Systematic Review and Meta-Analysis of the Efficacy of Appropriate Empiric Antibiotic Therapy for Sepsis. *Antimicrob. Agents Chemother.* **2010**, *54*, 4851–4863, doi:10.1128/aac.00627-10.
 33. Bosetti, D.; Grant, R.; Catho, G. Computerized Decision Support for Antimicrobial Prescribing: What Every Antibiotic Steward Should Know. *Antimicrob. Steward. Healthc. Epidemiol.* **2025**, *5*, e210, doi:10.1017/ash.2025.10091.
 34. Hatton, C.; Quarton, S.; Livesey, A.; Alenazi, B.A.; Jeff, C.; Sapey, E. Impact of Clinical Decision Support Software on Empirical Antibiotic Prescribing and Patient Outcomes: A Systematic Review and Meta-Analysis. *BMJ Open* **2025**, *15*, e099100, doi:10.1136/bmjopen-2025-099100.
 35. Fitzpatrick, F.; Tarrant, C.; Hamilton, V.; Kiernan, F.M.; Jenkins, D.; Krockow, E.M. Sepsis and Antimicrobial Stewardship: Two Sides of the Same Coin. *BMJ Qual. Saf.* **2019**, *28*, 758–761, doi:10.1136/bmjqs-2019-009445.
 36. Mcleod, M.; Campbell, A.; Hayhoe, B.; Borek, A.J.; Tonkin-Crine, S.; Moore, M.V.; Butler, C.C.; Walker, A.S.; Holmes, A.; Wong, G. How, Why and When Are Delayed (Back-up) Antibiotic Prescriptions Used in Primary Care? A Realist Review Integrating Concepts of Uncertainty in Healthcare. *BMC Public Health* **2024**, *24*, 2820, doi:10.1186/s12889-024-20248-8.
 37. Richards, A.R.; Linder, J.A. Behavioral Economics and Ambulatory Antibiotic Stewardship: A Narrative Review. *Clin. Ther.* **2021**, *43*, 1654–1667, doi:10.1016/j.clinthera.2021.08.004.
 38. Schisterman, E.F.; Cole, S.R.; Platt, R.W. Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies. *Epidemiol. Camb. Mass* **2009**, *20*, 488–495, doi:10.1097/EDE.0b013e3181a819a1.
 39. Hernán, M.A.; Hernández-Díaz, S.; Robins, J.M. A Structural Approach to Selection Bias. *Epidemiology* **2004**, *15*, 615, doi:10.1097/01.ede.0000135174.63482.43.
 40. Nathwani, D.; Varghese, D.; Stephens, J.; Ansari, W.; Martin, S.; Charbonneau, C. Value of Hospital Antimicrobial Stewardship Programs [ASPs]: A Systematic Review. *Antimicrob. Resist. Infect. Control* **2019**, *8*, 35, doi:10.1186/s13756-019-0471-0.
 41. Malone, D.C.; Armstrong, E.P.; Gratie, D.; Pham, S.V.; Amin, A. A Systematic Review of Real-World Healthcare Resource Use and Costs of Clostridioides Difficile Infections. *Antimicrob. Steward. Healthc. Epidemiol.* **2023**, *3*, e17, doi:10.1017/ash.2022.369.
 42. Giacobbe, D.R.; Marelli, C.; La Manna, B.; Padua, D.; Malva, A.; Guastavino, S.; Signori, A.; Mora, S.; Rosso, N.; Campi, C.; et al. Advantages and Limitations of Large Language Models for Antibiotic Prescribing and Antimicrobial Stewardship. *Npj Antimicrob. Resist.* **2025**, *3*, 14, doi:10.1038/s44259-025-00084-5.
 43. Ngoc Nguyen, O.; Amin, D.; Bennett, J.; Hetlevik, Ø.; Malik, S.; Tout, A.; Vornhagen, H.; Vellinga, A. GP or ChatGPT? Ability of Large Language Models (LLMs) to Support General Practitioners When Prescribing Antibiotics. *J. Antimicrob. Chemother.* **2025**, *80*, 1324–1330, doi:10.1093/jac/dkaf077.
 44. Williams, C.Y.K.; Miao, B.Y.; Kornblith, A.E.; Butte, A.J. Evaluating the Use of Large Language Models to Provide Clinical Recommendations in the Emergency Department. *Nat. Commun.* **2024**, *15*, 8236, doi:10.1038/s41467-024-52415-1.

45. Vo, T.; Dahal, K.; Klepser, M.; Pontefract, B.; Caniff, K.E.; Sohn, M. Evaluation of Large Language Models for Antimicrobial Classification: Implications for Antimicrobial Stewardship Programs. *Antimicrob. Steward. Healthc. Epidemiol.* **2025**, *5*, e324, doi:10.1017/ash.2025.10235.
46. von Elm, E.; Altman, D.G.; Egger, M.; Pocock, S.J.; Gøtzsche, P.C.; Vandenbroucke, J.P. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *The Lancet* **2007**, *370*, 1453–1457, doi:10.1016/S0140-6736(07)61602-X.
47. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
48. Cole, S.R.; Platt, R.W.; Schisterman, E.F.; Chu, H.; Westreich, D.; Richardson, D.; Poole, C. Illustrating Bias Due to Conditioning on a Collider. *Int. J. Epidemiol.* **2010**, *39*, 417–420, doi:10.1093/ije/dyp334.
49. Randhawa, V.; Sarwar, S.; Walker, S.; Elligsen, M.; Palmay, L.; Daneman, N. Weighted-Incidence Syndromic Combination Antibigrams to Guide Empiric Treatment of Critical Care Infections: A Retrospective Cohort Study. *Crit. Care* **2014**, *18*, R112, doi:10.1186/cc13901.
50. API Platform Available online: <https://openai.com/api/> (accessed on 18 February 2026).
51. ATCDDD - Guidelines Available online: https://atcddd.fhi.no/atc_ddd_index_and_guidelines/guidelines/ (accessed on 18 February 2026).
52. Dixon, W.J.; Mood, A.M. The Statistical Sign Test. *J. Am. Stat. Assoc.* **1946**, *41*, 557–566, doi:10.1080/01621459.1946.10501898.
53. Efron, B.; Tibshirani, R. *An Introduction to the Bootstrap*; Monographs on statistics and applied probability; Nachdr.; Chapman & Hall: Boca Raton, Fla., 1998; ISBN 978-0-412-04231-7.
54. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Off. J. Eur. Union 2016, L 119, 1–88.
55. Data Controls in the OpenAI Platform Available online: <https://developers.openai.com/api/docs/guides/your-data/> (accessed on 20 February 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.