Pre-Analysis Plan

Effectiveness of AI-Assisted Antibiotic Prescribing: A Randomized Controlled Trial

Ihsan Ayyub Qazi, Muhammad Hamad Alizai, Ayesha Ali, Muhammad Ammar Faisal, Muhammad Abdullah, Danesh Moradigaravand, Sara Iftikhar, Ge Zhou, Hatib Zia, Mehwish Lali, Nadya Majeed

Lahore University of Management Sciences (LUMS); King Abdullah University of Science & Technology (KAUST)

September 16, 2026

Introduction

Inappropriate antibiotic prescribing remains one of the most consequential drivers of antimicrobial resistance globally, with particular urgency in low- and middle-income countries (Otaigbe & Elikwu, 2023), where regulatory enforcement, antibiogram coverage, and specialist infectious-disease support are limited. In Pakistan specifically, an outpatient audit reports high antibiotic prescribing rates, frequent selection of broad-spectrum agents over guideline-concordant first-line options, and substantial rates of clinically unnecessary prescribing (Khawaja et al., 2026).

Recent advances in Large Language Models (LLMs) have shown promise in supporting clinical decision-making (Qazi, Ali, Khawaja, & others, 2026). LLMs have been explored as tools for synthesizing antibiotic recommendations and flagging spectrum mismatches (Giacobbe et al., 2025). Yet LLMs may produce confident but incorrect recommendations, and their use may reduce the quality of physicians' own clinical reasoning through over-reliance (Qazi, Ali, Khawaja, Akhtar, et al., 2026).

Whether this promise translates into measurably better antibiotic prescribing behavior among practicing physicians, measured across the full spectrum of selection, dosing, and clinical reasoning, has not been rigorously evaluated in a randomized, controlled setting.

This randomized controlled trial (RCT) aims to evaluate the impact of providing LLM access to physicians on antibiotic prescribing appropriateness, using a composite outcome instrument, the Antibiotic Prescribing Appropriateness Score (APAS), that simultaneously evaluates antibiotic selection, dosing, clinical reasoning, and management planning. Participants will be randomly assigned to one of two groups: the intervention group will have access to an LLM alongside conventional resources, while the control group will use conventional resources only (e.g. UpToDate, PubMed and Google Search with AI features disabled). Both groups will respond to a set of clinical vignettes covering common infectious disease scenarios requiring antibiotic prescribing decisions.

Research Question

To what extent does access to an LLM affect APAS among licensed physicians in Pakistan?

Hypotheses

Primary Hypothesis (H1): Physicians in the intervention group (LLM access plus conventional resources) will demonstrate significantly higher APAS scores compared to the control group (conventional resources only).

Secondary Hypotheses:

- **H2:** Physicians in the intervention group will select more appropriate antibiotics and demonstrate better spectrum-matching compared to the control group.
- **H3:** Physicians with less clinical experience will derive greater APAS benefit from LLM access compared to more experienced clinicians.
- **H4:** Physicians in the intervention arm will not demonstrate significantly greater time per vignette compared to the control arm.

Methods

Study Design

This study will be an assessor-blind, randomized controlled trial with two arms:

- **Intervention Arm:** Participants will have access to ChatGPT in addition to conventional diagnostic and reference resources (UpToDate, PubMed, Google search with AI features disabled). They will evaluate the same set of clinical vignettes covering common infectious disease scenarios requiring antibiotic prescribing decisions, with vignette order randomized independently for each participant. All participant interactions with the LLM, including prompts submitted and outputs received, will be recorded.
- **Control Arm:** Participants will use conventional resources only (UpToDate, PubMed, Google Search with AI features disabled). They will evaluate the same set of clinical vignettes presented in a randomized order.

Randomization: Participants will be randomly assigned to either the intervention or control arm in a 1:1 ratio using a computer-generated random number sequence.

Blinding: Outcome assessors (those scoring the APAS responses) will be blinded to participants' group assignments. Assessors will evaluate vignette responses without knowledge of which arm the responding physician was in. To reduce the risk of inadvertent unblinding via stylistic cues associated with LLM-assisted responses, common AI markers (e.g., em-dashes, markdown formatting symbols such as asterisks and bullet points) will be removed from the responses prior to assessor review. Blinding integrity will additionally be assessed empirically: for a random subset of scored responses, assessors will be asked to guess the participant's group assignment after they have scored a response, and the guess accuracy will be reported against the chance baseline of 50%.

Measuring Prescribing Appropriateness: The primary measure of prescribing appropriateness will be the APAS, a four-domain rubric scored out of 13 points per vignette

and expressed as a percentage. The adjusted mean difference in vignette-level APAS between the intervention and control arms is the primary outcome of our study.

Participants

Inclusion Criteria

- Completed Bachelor of Medicine, Bachelor of Surgery (MBBS) or equivalent degree. The equivalent degree of MBBS in the US and Canada is Doctor of Medicine (MD).
- Full or Provisionally Registered Medical Practitioners with the Pakistan Medical and Dental Council (PMDC).

Exclusion Criteria

- Any other Registered Medical Practitioner (Full or Provisional) with PMDC not holding an MBBS or equivalent degree (e.g., practitioners with a Bachelor of Dental Surgery (BDS)).

Sample Size and Power Analysis

Target Sample Size: 50 participants, with 25 participants assigned to each study arm.

Power Calculation: Consistent with two prior similar trials (Goh et al., 2024; Qazi et al., 2026), we use the same effect size (8 percentage points), standard deviation (16.2 percentage points), and ICC range (0.05-0.15) as those studies, along with 4 completed vignettes per participant and a two-sided α of 0.05. Power was calculated using a linear mixed-effects model with crossed random intercepts for participant and vignette, applying the 0.05 - 0.15 ICC range to each random-effect level. Under this specification, a sample of 200 completed cases (50 participants × 4 vignettes) provides at least 80% power to detect the specified 8-percentage-point difference.

Clinical Vignettes

Participants will be presented with a set of clinical vignettes covering common infectious disease scenarios requiring antibiotic prescribing decisions. These vignettes will:

- Be sourced and modified based on clinically realistic cases relevant to the target population.
- Cover a range of infectious disease presentations encountered in family medicine, general practice, internal medicine, and emergency medicine.
- Avoid overly rare or highly specialized conditions.

Vignette Development and Validation:

All vignettes will be developed by an independent expert panel, with attention to:

- Clinical accuracy, guideline concordance, and realism.
- Appropriateness of difficulty level for the target population.

- Clarity and completeness of information.

Each vignette will be accompanied by an expert-validated grading rubric specifying expected answers for antibiotic selection and spectrum; dosing, route, and duration; clinical reasoning; and management planning.

Vignette Completion:

- Each participant is expected to complete 8-12 vignettes.
- The order of vignettes will be randomized for each participant to control for order effects.

Data Collection

Session Protocol

- Participants will complete the study under proctored conditions (either in person or via a secure video conferencing platform with screen sharing). For control-arm participants, proctors will monitor screen activity throughout the session to verify that no LLM tools (including standalone LLM interfaces, browser extensions, or AI-generated search summaries) are accessed. Participants will also be instructed not to use personal devices or resources outside the monitored study environment while completing the vignettes.
- Participants will complete a baseline questionnaire collecting demographic information (gender, age, specialty, years of experience) and prior experience with LLMs.
- Participants will be given instructions specific to their assigned group (intervention or control).
- Participants will work through clinical vignettes one by one.
- For each vignette, participants will be asked to complete a standardized within-vignette response template including the following fields:
    - Recommended medication, including drug name, dose, route, and duration.
    - Clinical reasoning, including the working diagnosis (or syndrome) with the specific supporting clinical, exam, and laboratory findings.
    - Management plan, including admission decision, referrals/consults, further investigations, stepdown or IV-to-oral switch criteria where applicable, patient counseling, and follow-up.
- Participants will record their responses using a standardized electronic form.
- Time spent on each vignette will be automatically recorded.

Outcome Measures

Primary Estimand

The primary estimand is the adjusted mean difference in vignette-level APAS between the intervention and control arms, estimated from a linear mixed-effects model with crossed random intercepts for participant and vignette, adjusting for the pre-specified covariates (years of practice post-MBBS, gender, prior LLM experience). Participant-level mean APAS is reported descriptively only and is not the basis of inference.

Primary Outcome

Antibiotic Prescribing Appropriateness Score (APAS): The primary outcome will be a composite score (expressed as a percentage) based on an expert-validated rubric that evaluates physician responses across four domains for each clinical vignette:

1. **Antibiotic Selection & Spectrum** (0-6 points)
    - 6: Optimal antibiotic selection at the correct spectrum level, or correct and well-justified decision to withhold; WHO AWaRe-concordant and consistent with local resistance data.
    - 4: Correct drug choice with minor spectrum issues, or acceptable second-line agent; resistance data partially integrated.
    - 2: Correct antibiotic class but suboptimal spectrum or AWaRe category not considered.
    - 0: Wrong drug, inappropriate spectrum, unjustified broad-spectrum use, or antibiotic prescribed when clearly not indicated.
2. **Dosing, Route & Duration** (0-3 points)
    - 3: Correct dose, route, and duration; appropriate patient-specific adjustments (renal function, weight, age) where indicated.
    - 1.5: Partially correct (e.g., correct drug and dose but inappropriate route or duration; adjustment absent when warranted).
    - 0: Incorrect dose, route, or duration; or dosing information absent when an antibiotic is recommended.
3. **Clinical Reasoning** (0-2 points)
    - 2: Comprehensive, guideline-anchored reasoning that explicitly justifies both the antibiotic selection and the dosing, route, and duration decisions; integrates clinical findings, local resistance data, patient-specific risk factors, and relevant microbiological considerations.
    - 1: Sound reasoning with partial integration of guidelines or resistance data; justification addresses selection or dosing decisions but not both in full.

    - 0.5: Basic reasoning present but incomplete; vague or missing reference to resistance data, guidelines, or dosing rationale.
    - 0: No reasoning provided or fundamentally flawed clinical logic.
4. **Management Plan** (0-2 points)
    - 2: Explicit forward-looking plan including follow-up timeline, de-escalation or IV-to-oral switch criteria where applicable, safety-netting advice, and reassessment criteria, distinct from the initial prescribing decision itself.
    - 1: Partial management plan; some follow-up, de-escalation, or reassessment elements present but incomplete.
    - 0: No follow-up plan, de-escalation strategy, or reassessment criteria documented.

The total score for each vignette is the sum of scores across the four domains (maximum 13 points). APAS is expressed as a percentage:

$$APAS = \frac{total\ score}{13} \times 100\%$$

The primary endpoint is vignette-level APAS, analyzed in a linear mixed-effects model with crossed random intercepts for participant and vignette. Responses will be independently evaluated by three qualified raters blinded to participant identity and treatment assignment. The arithmetic mean of the three raters' total scores will constitute the vignette-level APAS used in the primary analysis.

Instrument validation: Inter-rater reliability will be assessed using Krippendorff's α and intraclass correlation coefficients, reported both overall and separately by domain, with a target of α ≥ 0.80. In cases of significant disagreement between raters, the raters will discuss the response until consensus is reached, which may lead to refinement of the scoring rubric. Internal consistency of APAS across the vignettes will be assessed using Cronbach's α.

Secondary Outcomes

1. **Antibiotic Selection & Spectrum Subscore:** The Antibiotic Selection & Spectrum domain of APAS (0-6 points per vignette, expressed as a percentage), scored using the same rubric as the primary outcome. Used to test H2.
2. **Time per Vignette:** The difference between study arms in time spent per vignette will be estimated as a secondary outcome. The magnitude and precision of the between-arm difference will be reported. Used to test H4.

Data Analysis Plan

Descriptive Statistics

- Baseline characteristics of the participants in each group will be summarized using means and standard deviations for continuous variables (e.g., age, years of

experience post-MBBS) and frequencies and percentages for categorical variables (e.g., gender, prior LLM experience).

- Descriptive statistics will be used to summarize APAS scores, time spent per vignette, and antibiotic prescribing rates for each group.

Primary Outcome Analysis

A linear mixed-effects model will be used to estimate the adjusted mean difference in vignette-level APAS between the intervention and control groups, consistent with the Primary Estimand defined above.

Model Structure:

- **Dependent Variable:** APAS (percentage).
- **Fixed Effect:** Treatment group (Intervention vs. Control).
- **Random Effects:**
    - Participant ID (to account for within-participant correlation across vignettes).
    - Vignette ID (to account for potential variability in vignette difficulty).
- **Covariates:**
    - Gender
    - Prior LLM usage frequency
    - Years of practice post-MBBS

Primary Comparison: The estimated treatment coefficient from the model above, reported with a 95% confidence interval.

Significance Threshold: A two-sided p-value ≤ 0.05 will be considered statistically significant.

Secondary Outcome Analyses

1. **Antibiotic Selection & Spectrum Subscore:** The Antibiotic Selection & Spectrum subscore will be analyzed using a linear mixed-effects model with the same random-effects and covariate structure as the primary analysis, using the subscore (percentage) as the dependent variable in place of total APAS. This analysis tests H2, evaluating whether LLM access improves antibiotic selection and spectrum-matching specifically, independent of the other three APAS domains.
2. **Time per Vignette:** Time per vignette will be analyzed using a linear mixed-effects model with the same random-effects and covariate structure as the primary analysis. Time per vignette is expected to be right-skewed; a log transformation will be applied if diagnostic checks indicate this is needed. Results will be reported as adjusted times in seconds with 95% confidence intervals. This analysis will estimate whether

and by how much LLM access changes completion time. A non-significant treatment coefficient supports H4, indicating that prescribing quality gains are achievable without a time-efficiency penalty.

Subgroup Analyses

Subgroup analyses will be performed to explore whether the effect of the intervention on APAS differs by physician experience level (years since MBBS, tests H3), gender, and prior LLM usage experience. Interaction terms between treatment group and each subgroup variable will be tested within the primary mixed-effects model.

Multiple Comparisons

The primary hypothesis (H1) is tested without adjustment for multiplicity, as APAS is the single pre-specified primary endpoint. The three secondary hypotheses (H2: Antibiotic Selection & Spectrum Subscore; H3: Physician-Experience Interaction; H4: Time per Vignette) will be jointly corrected for multiple comparisons using the Holm-Bonferroni procedure at a family-wise $\alpha = 0.05$.

Sensitivity Analyses

The following sensitivity analyses will be conducted to assess the robustness of the primary finding: residuals from the primary mixed-effects model will be assessed for normality (QQ plot, Shapiro-Wilk test), heteroscedasticity (residuals-vs-fitted plot), and influential observations (participant-level case-deletion diagnostics); if assumptions are materially violated, the primary analysis will be repeated using bootstrap confidence intervals or a robust mixed-effects model as a sensitivity check.

Exploratory Analyses

Standalone LLM Performance: The LLM will be queried independently on each vignette, and its responses will be scored by the expert panel using the same APAS rubric. Standalone performance will be compared against both study arms as a reference benchmark. Published benchmarks for LLM antibiotic prescribing vary by model and setting: Claude scored highest among six models evaluated on general-practice antibiotic vignettes across four countries (mean 80%, rising to 87% on US-standardized cases) (Nguyen et al., 2025), whereas Claude 3 Opus was among the lowest-performing of fourteen models evaluated on antibiogram-embedded hospital vignettes, where ChatGPT-o1 scored highest (71.7%) (De Vito et al., 2025). This variability underscores the value of benchmarking the specific model and vignette set used in the present trial rather than relying on prior literature alone.

Missing Data

- Primary analysis will be based on complete cases only.
- The pattern and extent of any missing data will be described.

- If a substantial amount of missing data is present (>5% for any key variable), multiple imputation will be considered, and the primary analysis will be repeated including incomplete vignettes to evaluate the impact on results.

Statistical Software

All statistical analyses will be performed using R (version 4.3.2 or later) with appropriate packages (e.g., lme4, ordinal, mice) or equivalent software in Python or Stata.

Ethical Considerations

This study poses minimal risk to participants. No real patient data are used and no clinical decisions affecting actual patients are made during the study. Participants will provide informed consent prior to enrollment and may withdraw at any time without penalty. Individual scores will not be shared; results will be reported only in aggregate or de-identified form. Participants will receive compensation of PKR 10,000 upon completion of the study. The Institutional Review Board has approved this study at the Lahore University of Management Sciences (Protocol Number: LUMS-IRB-0492/08062026/IAQ-FWA-00030383).

Conclusion

This pre-analysis plan outlines a randomized controlled trial evaluating the impact of large language model access on antibiotic prescribing appropriateness among licensed physicians in Pakistan. By adhering to this plan, we aim to minimize bias, enhance transparency, and produce high-quality evidence that can inform the integration of LLMs into antimicrobial stewardship workflows in low- and middle-income settings. The findings will contribute to a better understanding of the potential benefits and limitations of using LLMs as clinical decision support tools for antibiotic prescribing, ultimately aiming to improve patient care and support responsible antibiotic use in a context where antimicrobial resistance poses a substantial and growing threat.

References

De Vito, A., Geremia, N., Bavaro, D. F., Seo, S. K., Laracy, J., Mazzitelli, M., Marino, A., Maraolo, A. E., Russo, A., Colpani, A., Bartoletti, M., Cattelan, A. M., Mussini, C., Parisi, S. G., Vaira, L. A., Nunnari, G., & Madeddu, G. (2025). Comparing large language models for antibiotic prescribing in different clinical scenarios: Which performs better? *Clinical Microbiology and Infection, 31*(8), 1336-1342. https://doi.org/10.1016/j.cmi.2025.03.002

Giacobbe, D. R., Marelli, C., La Manna, B., Padua, D., Malva, A., Guastavino, S., Signori, A., Mora, S., Rosso, N., Campi, C., Piana, M., Murgia, Y., Giacomini, M., & Bassetti, M. (2025). Advantages and limitations of large language models for antibiotic prescribing and antimicrobial stewardship. *Npj Antimicrobials and Resistance*, *3*(1), 14. https://doi.org/10.1038/s44259-025-00084-5

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized

clinical trial. *JAMA Network Open*, *7*(10), e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

Khawaja, T., Kajova, M., & Kantele, A. (2026). Improving antibiotic prescribing in LMICs: Insights from an outpatient clinic in Pakistan. *New Microbes and New Infections*, *69*, 101680. https://doi.org/10.1016/j.nmni.2025.101680

Nguyen, O. N., Amin, D., Bennett, J., Hetlevik, Ø., Malik, S., Tout, A., Vornhagen, H., & Vellinga, A. (2025). GP or ChatGPT? Ability of large language models (LLMs) to support general practitioners when prescribing antibiotics. *Journal of Antimicrobial Chemotherapy*, *80*(5), 1324-1330. https://doi.org/10.1093/jac/dkaf077

Otaigbe, I. I., & Elikwu, C. J. (2023). Drivers of inappropriate antibiotic use in low- and middle-income countries. *JAC-Antimicrobial Resistance*, *5*(3), dlad062. https://doi.org/10.1093/jacamr/dlad062

Qazi, I. A., Ali, A., Khawaja, A. U., et al. (2026). Large language model diagnostic assistance for physicians in a lower-middle-income country: A randomized controlled trial. *Nature Health*, *1*, 198-205. https://doi.org/10.1038/s44360-025-00007-8

Qazi, I. A., Ali, A., Khawaja, A. U., Akhtar, M. J., Sheikh, A. Z., & Alizai, M. H. (2026). Automation bias in large language models-assisted diagnostic reasoning among physicians trained in AI literacy - a randomized clinical trial. *NEJM AI*, *3*(5), AIoa2501001. https://doi.org/10.1056/AIoa2501001