



JULY 29, 2026

Assessing the effectiveness of Large Language Model (LLM) -
enabled nurse treatment planning in 2 Indian districts: A
pilot study

ETHICS REVIEW SUBMISSION – PROTOCOL, TOOLS & CONSENT
FORMS

PI: DR. ANUP AGARWAL

Study Investigators: Ms. Sarah Nabia, Dr. Priyank Jain, Dr. Parthsarathi Mukherjee, Dr. Abhijit
Chowdhury, Dr. Tushar Garg, Dr. Yogesh Jain,



Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

I. Background and Rationale

Large language model (LLM) based generative artificial intelligence (AI) is revolutionizing healthcare by enhancing diagnostic accuracy, personalizing treatments, and streamlining clinical decision-making. By offering evidence-based recommendations, AI empowers healthcare providers to develop more effective care plans. This capability is especially promising in low- and middle-income countries (LMICs), where urban–rural disparities in healthcare provider availability hinder access to quality care.

In India—home to 1.43 billion people—a critical urban–rural divide persists. For example, the National Medical Commission reports that India’s overall doctor-population ratio is 1 per 843 persons (*Update on Ratio Of Patients And Doctors Nurses*, 2023), and the Economic Survey (2024-2025) stated that doctor density is 3.8 times higher in urban than in rural areas (Desk, 2025). Using the most recent World Bank rural and urban population distribution, this implies 1 physician per 2,550 rural residents versus 1 per 670 residents in urban India, which is more than a three-fold disparity. Although current initiatives, including availability of health and wellness centers, infrastructure investments, telemedicine, and public–private partnerships in conjunction with longstanding community health worker programs like ASHA aim to mitigate these disparities, significant gaps remain. Leveraging AI tools to empower non-physician healthcare workers (e.g., nurses) to develop patient care plans represents a novel strategy to improve healthcare access and quality in underserved regions.

LLMs have proven to be better than physicians in clinical reasoning and evidence based clinical solution when provided with clinical vignettes or diagnosis, respectively (Goh et al., 2024; Nori et al., 2025; Rao et al., 2023). However, LLMs have not been assessed for their ability to generate a clinical vignette via multiturn conversations when starting from a chief complaint. Some advances have been made in AI systems to conduct diagnostic dialogue with patient actors. For instance, AMIE by Google has shown superior history-taking and diagnostic accuracy in randomized, text-chat consultations with validated patient actors, yet it remains a research prototype that has not been tested in real-world clinical encounters.(Tu et al., 2025). To the best of our knowledge, LLMs have also not been assessed in real-life scenarios in resource-constrained settings.

This research protocol outlines the plan for a within-subject, paired comparison study (a type of crossover study design) to test the effectiveness of nurse - LLM paired solution in clinical reasoning and evidence-based clinical solutions in primary care settings when used for actual patients in a real-life encounter in a resource constrained LMIC context.

The proposed pilot project will evaluate the effectiveness of the intervention for managing Hypertension, Diabetes Mellitus (DM), musculoskeletal pain, breathlessness and fever. According to the ICMR-INDIAB study (2022), the weighted prevalence of DM and hypertension among adults ≥ 20 years was 11.4% and 35.5%, respectively (Anjana et al., 2023). Another nation-wide study (POSEIDON study) estimated that fever (35.5%) was the most common symptom among patients presenting to a primary healthcare provider, more than half (50%) of all patients had respiratory symptoms, almost 10% patients had musculoskeletal symptoms (Salvi et al., 2015). High burden of

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

these conditions and symptoms underscores the urgent need for scalable, high-quality care solutions in primary care settings.

The research team will also study the user (nurse) and client (patient) experience of participating in an LLM empowered solution for bridging gaps in human resource for healthcare. The nurses' experience will be captured through qualitative methods and the patient experience will be assessed via mixed methods.

II. Objectives

1. Assess the effectiveness of LLM-enabled, nurse-led treatment planning for select NCDs and symptoms compared to the standard of care. (primary objective)
2. Measure the attitudes, enablers and barriers of patients towards a nurse + LLM care intervention in primary healthcare settings. (secondary objective I)
3. Describe the attitudes and perceptions of nurses in using LLM tools in their clinical practice. (secondary objective II)

III. Methodology

This is a randomized, crossover study with embedded qualitative interviews for nurses and a cross-sectional exit survey for patients. Prior to patient enrolment, the LLM model and interface selection for the purpose of this study will be undertaken.

A. LLM model selection

Safety, clinical accuracy and adaptability to changing contexts are few of the many critical characteristics necessary for an LLM to be deployed in population-level clinical practice. Although there has been a rapid surge in the number of LLMs that can provide insights into clinical conditions, the reliability, accuracy, and relevance of these model outputs need to be systematically assessed before it can be deployed for programmatic use. HealthBench, by OpenAI is the largest known open-source benchmarking evidence available to compare various LLMs(Arora et al., 2025). HealthBench was developed to evaluate LLMs on real-world healthcare tasks such as clinical decision support, symptom triage, and patient communication. Developed in collaboration with 262 physicians across 60 countries, HealthBench features 5,000 multi-turn, open-ended clinical scenarios assessed in 49 languages using 48,562 rubric-based criteria. It evaluates performance across five critical domains: scientific and clinical reasoning, patient safety, helpfulness, patient-centered care, and equitable communication.

The March 2025 release of HealthBench evaluated 11 themes – which included global health as one. The criteria for assessing cases in the global health theme were:

- i. Adaptation to Resource Constraints
Could the model recognize limited availability of tests, treatments, or specialists and adjust its recommendations accordingly?

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

- ii. Contextual Relevance
Did the model align its responses with local epidemiology, common health conditions, and standard practices in LMICs?
- iii. Cultural and Linguistic Appropriateness
Was the communication culturally sensitive and linguistically accessible across diverse populations?
- iv. Clinical Accuracy and Safety
Were the medical recommendations both accurate and safe, especially in settings where clinical guidelines may differ?
- v. Recognition of Systemic Limitations
Did the model identify when ideal care options might not be feasible and offer reasonable, safe alternatives?

The LLM selection for this study will be anchored in evidence from the HealthBench. o3 by OpenAI (60%), Grok 3 by X (54%) and Gemini 2.4 Pro by Google (52%) achieved similar overall performance scores. These models or others available in another HealthBench update before the study rollout will be vetted by a technology partner and by nurses (users) in a human-centered design workshop for practical considerations (Arora et al., 2025).

B. Interface & usability testing

We will conduct a structured human-centered design (HCD) workshop that integrates clinical simulation and usability evaluation. A HCD workshop is a structured, participatory process that places end users—such as patients, nurses, or physicians—at the center of solution development. It draws on methods from design thinking and behavioral science to understand users' needs, generate ideas, and prototype solutions. Core principles include empathy, co-creation, iteration, and user testing. A group of MD-level physicians will first develop a set of realistic clinical cases on hypertension, diabetes mellitus, fever, breathlessness, and musculoskeletal pain representing common scenarios encountered in the target healthcare setting. These cases will be used to assess how nurses interact with different combinations of LLMs and user interfaces. A group of 10 to 15 practicing nurses will participate in a simulation exercise, where they solve these clinical cases using pre-selected LLM-interface combinations. The interfaces will vary in modality to reflect possible real-world deployments, including audio-only (spoken input/output) and multimodal interface (combining text and audio). Following each case, participants will independently rate their experience using a usability evaluation tool focusing on clarity, ease of use, responsiveness, and confidence in clinical decision-making. Structured debriefs and group discussions will follow to gather qualitative feedback on barriers, preferences, and perceived trustworthiness of the model-interface combinations. Findings from this exercise will inform the selection of the

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

optimal LLM-interface combination for the implementation study and workflow integration strategies.

C. Study design

1. Primary objective:

The intervention (nurse + LLM - patient encounter) and control (licensed physician - patient encounter) will be sequential. Each patient will participate in both the intervention and control arms, at the end of which separate treatment and follow-up plans are generated. This allows for direct comparisons within the same individual; thus, controlling for between-subject variability that may otherwise influence outcomes. Since the same patients are evaluated under both intervention and control, we can pair each nurse - LLM treatment outcome with the corresponding doctor-led outcome for that patient.

Advantages:

- Reduction in variability: Each patient acts as their own control, confounding variables like age, comorbidities and lifestyle factors are inherently balanced.
- Increased statistical power: Paired design typically requires a smaller sample size to detect a given effect compared to parallel group designs because variability between study subjects is minimized making it a practical study design option for implementation science research.

2. Secondary objective I

A cross-sectional exit survey will assess patient experience following LLM-assisted, nurse-led care using a brief tool developed from validated frameworks. Questions are adapted primarily from the WHO Health System Responsiveness domains (e.g., dignity, autonomy, etc), a globally recognized framework assessing dignity, communication, autonomy, and other aspects of patient experience in diverse health systems (Darby et al., 2000). Additional items are adapted from the Patient Satisfaction Questionnaire (PSQ-18), a widely used patient satisfaction tool developed by RAND Corporation (Marshall & Hays, 1994). AI-specific questions on trust and comfort are based on recent literature evaluating patient perceptions of AI in healthcare (Nong & Platt, 2025). The tool will be pilot tested for clarity and cultural relevance to ensure appropriateness for the study population. To complement the quantitative findings from the exit survey, we will conduct in-depth interviews (IDIs) with a small, purposively selected sample of patients. These interviews will provide richer insights into patients' perceptions, trust, and comfort with the AI-assisted nurse consultations, allowing us to explore contextual nuances and explanations that may not be captured through structured survey responses.

3. Secondary objective II

As part of the study, we will conduct IDIs with participating nurses to explore their experiences using the LLM during patient care. These interviews will provide qualitative insights into the usability, perceived accuracy, clinical relevance, and workflow integration of the LLM tool. Nurses will be asked about their comfort level, trust in the

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

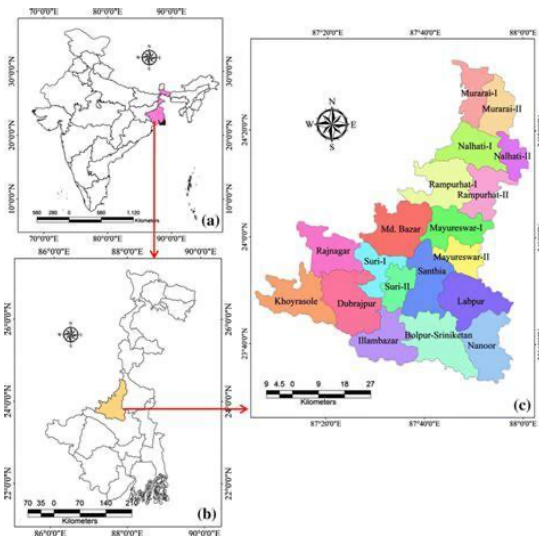
AI-generated recommendations, ease of interaction (e.g., via text or voice), and any challenges encountered during real-time use. Interviews will also explore suggestions for improving the tool's design and implementation. The IDIs will be conducted using a semi-structured guide, audio-recorded with consent, and thematically analyzed to inform future adaptation and scale-up of the intervention.

IV. Study Setting

The study will be implemented in 2 sites in the state of West Bengal (Birbhum and Purulia districts).

Site 1: Nagari Village, Suri-I Block, Birbhum District, West Bengal

Nagari is a rural village in Suri-I block of Birbhum district with a population of 1,959 across 396 households as per the 2011 Census. The population is predominantly Bengali speaking, with 61.9% literacy (63.7% male; 60.3% female) (*Nagari Village Population - Suri - I - Birbhum, West Bengal*, n.d.). The village exhibits a balanced gender ratio (~1,102 females per 1,000 males) and includes substantial Scheduled Caste (49.5%) and Scheduled Tribe (27.8%) communities. Nagari relies on a nearby sub-center for primary care, with higher-tier facilities located ~10 km away in Suri (PHC and district hospital) . Most residents engage in agriculture or informal labor, with basic mobile access. This setting—with limited local clinical capacity and frontline nurse responsibility—is suitable for evaluating the usability and acceptability of nurse-delivered, AI-assisted clinical tools in a resource-constrained rural environment.



Site 2: Purulia District, West Bengal

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Purulia is a predominantly rural district in the western plateau region of West Bengal, with a total population of 2,930,115 as per the 2011 Census, of whom 87.3% reside in rural areas. The district has a balanced gender ratio (sex ratio ~957 females per 1,000 males) and a relatively low population density (468 persons per sq. km). It includes substantial Scheduled Caste (19.4%) and Scheduled Tribe (18.5%) communities, and literacy stands at 64.5%. The rural economy is predominantly agrarian and largely rainfed, with paddy the primary crop and about 73% of agricultural holdings belonging to small and marginal farmers; a large share of households derive their income from manual casual labour. The health system is supported by ASHAs, ANMs, and nurses, who play a vital role in frontline care, and many healthcare interactions still rely on in-person triage and verbal communication. Patients will be enrolled at clinics in rural and remote parts of the district, away from the district headquarters. This setting—with limited local clinical capacity, an agrarian and socioeconomically deprived catchment population, and frontline nurse responsibility—is suitable for evaluating the usability and acceptability of nurse-delivered, AI-assisted clinical tools in a resource-constrained rural environment.



V. Study Population

A. Primary objective

Inclusion criteria:

Criteria	Hypertension	Diabetes Mellitus	Fever	Musculoskeletal pain	Breathlessness
Age	>18 years	>18 years	>18 years	>18 years	>18 years
Duration of condition / symptom	Known diagnosis	Known diagnosis OR HbA1C > 6.4 OR fasting blood			

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

		sugar >126 OR PP >200			
Significance of complaint	If presented as chief complaint or in combination with DM	If presented as chief complaint or in combination with Hypertension	If presented as chief complaint and/or in combination with fever and breathlessness	If presented as chief complaint only and no evidence of fever	If presented as chief complaint only and no evidence of fever

Exclusion criteria: Individuals who don't have the ability to provide informed consent due to lack of cognitive capacity (e.g. dementia, mental retardation, etc.).

Acutely sick patients - Systolic Blood Pressure <90mmhg, SpO2 <92%, Respiratory rate >22/min

Sample size

We powered this single-session, counter-balanced, within-participant study to detect an absolute two-sided difference of $\pm 3.5\%$ in the consultation score between the intervention (nurse + LLM) and control (standard of care consultation by doctor). Assuming a within-subject design variance term derived from Goh et al., an intraclass correlation (ICC) of 0.10, two-sided α of 0.05, 80% power, and 5% dropout, the sample size across both sites for each clinical outcome is 336 (Goh et al., 2024).

Site	Clinical management (Hypertension & Diabetes Mellitus)	Clinical reasoning (Fever, joint pain, breathlessness)	Total
Purulia	168	168	336
Birbhum	168	168	336
Total			672

Recruitment will be block-randomized equally between sites and sequences.

1. Note: Sample size correction after pilot data

A 30-patient internal pilot (15 per outcome) will be analyzed with a linear mixed-effects model that treats intervention, order (first vs second consultation), and site (study location) as fixed effects, and random intercepts will be included for each patient and for each doctor and nurse. We will extract the residual variance and random effect variances

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

and use the *simr* package to re-estimate power for the planned sample size. The research team will inform the IRB if adjustments > 10% to the sample is required.

B. Secondary objective I

Inclusion criteria: Only enrolled patients who can fill out / respond to a tablet-based exit survey in English / Hindi / Bengali.

Sample size: All patients who have participated in both the intervention and control will be asked to complete the exit survey ~ 672 patients. A group of 30 to 40 patients (15 to 20 in each site) who have completed the consultations in both study arms will be conveniently sampled to participate in the patient IDIs until we reach information saturation.

C. Secondary objective II

Inclusion criteria: Nurses who have done at least 10 patient consults using the LLM tool.

Sample size: 6 to 8 nurses who will do the consults across the 2 sites over the life of the project.

VI. Study Procedures

Recruitment: On arrival at the clinic, prospective participants will be screened by a nurse who will measure the blood pressure, blood sugar levels, temperature, and take an oral history of the reason for the visit. This is the existing history-taking procedure at the clinic sites. Patients who meet the inclusion criteria for any of the 2 conditions or 3 symptoms will be offered the opportunity to enroll in the study.

Informed consent: A research assistant will explain the objectives of the study, expectations from the participant – 1. to devote extra time to this visit as they would be seen by 2 providers (doctor and nurse + LLM) and 2. fill out an exit survey, and the incentive available for the participant if they fulfil the requirements of the study. The research assistant will also review other aspects of informed consent with the patient (e.g. confidentiality, data reporting, etc.). Patients who agree to participate in the study will sign an informed consent form. The processes for informed consent will be done in the local language.

Randomization: Upon enrolment, the patient will be assigned to study sequence A or B, as per a computer-generated randomization sequence.

- Sequence A (Nurse + LLM → Doctor): Participants receive the intervention first, followed by control.
- Sequence B (Doctor → Nurse + LLM): Participants receive control first, followed by intervention.

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

VII. Data Collection and Management

Data Collection

Quantitative: Data for quantitative analysis will be collected for primary objective and secondary objective I.

Primary objective: The nurse + LLM – patient conversation will be audio recorded via a voice recording device. After taking the patient history, the nurse will enter the details into the LLM platform and carry out a multi-turn conversation with the LLM tool, seeking answers or clarifications from the patient, as required. Transcripts from the conversation and the LLM’s recommendation will be retrieved. Similarly, the doctor – patient conversations will be audio – recorded and a summary sheet with the differentials, diagnosis and recommendations will be filled out for each patient by the doctor. Outputs for each patient will be assigned a unique patient ID. Each patient will be associated with 4 data sources – 1. Nurse- LLM -patient conversation transcript 2. Nurse-LLM generated prescription 3. doctor-patient conversation transcript and 4. Doctor’s prescription.

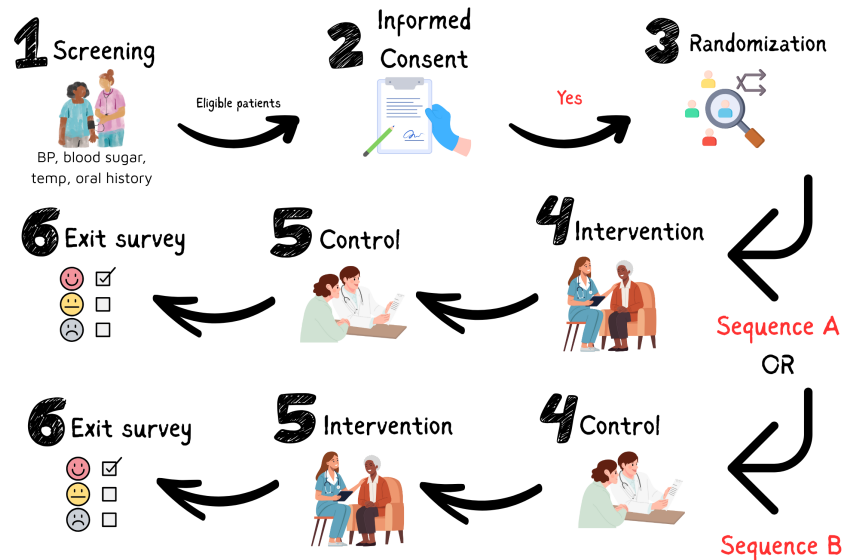
Secondary objective I: Upon completion of the nurse-LLM -patient consult, the participants will be asked to complete a short, 10-question exit survey about their experience and perceptions of the consult. The survey will be administered through a digital platform such as Qualtrics/ Survey Monkey and will be available in both Bengali and Hindi. We will not collect any personal identifiable details such as name or telephone number in the survey. The raw data will be retrieved in the form of .xls or .csv files and stored in a password protected cloud-storage drive.

Note: Participant compensation: All patients enrolled in the study who complete all study activities (1 consult each with the doctor and nurse + LLM, and the patient exit survey) will receive a compensation of INR 250 in acknowledgment of their time spent in participating in the study.

See figure 1 for study flow:

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Study Flow Diagram



Qualitative: In-depth interviews with a select group of patients will be conducted using a semi-structured interview guide. A convenient sample of respondents will be approached for participating in the patient IDIs. The interviewer will meet with the respondents in a semi-private location within 1 to 2 days of the clinic visit to complete the interview. Each patient IDI will last for 30 to 45 minutes. The patient IDIs will be conducted in either Bengali or Hindi, as per the respondent's choice.

Note: An additional INR 250 will be given as compensation to the patients who participate in the IDIs.

The IDIs with nurses will take 45 to 60 minutes and will be conducted in English or the participant's preferred local language (Bengali / Hindi) in a mutually agreed semi-private setting. The guide will focus on exploring questions about the comfort and usability of the AI tool, enablers and barriers of using an AI tool in a patient consult, willingness to integrate the tool in regular practice, and suggestions for improving the user experience. All interviews will be audio recorded and transcribed using software.

Data Management

1. Data collection tools and formats

The study will generate both quantitative and qualitative data across its three objectives. Data sources will include:

- Audio recordings and transcripts of nurse–LLM and doctor–patient consultations.
- Prescriptions and treatment plans generated by the nurse–LLM interface and the attending doctor.

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

- Responses from a structured, digital patient exit survey.
- Transcripts of in-depth interviews (IDIs) with participating patients and nurses.

All quantitative data will be collected using structured tools and stored in spreadsheet formats (.xls or .csv). Audio recordings will be transcribed using secure software, and qualitative interview transcripts will be exported as text documents.

2. **De-identification and confidentiality**

Each patient will be assigned a unique ID to link data across sources while ensuring anonymity. Personally identifiable information (PIIs) will not be recorded in transcripts, prescriptions, or survey responses. Transcription files will be reviewed and cleaned of any inadvertent identifiers. Names and other identifiers will be removed from audio and text data prior to analysis.

3. **Data storage and security**

All data will be stored in a secure, cloud-based storage system with restricted access. Access will be password protected and limited to designated members of the study team. Audio recordings, transcripts, and survey data will be stored in separate folders according to data type and collection site. Regular backups will be scheduled to prevent data loss.

4. **Data access and oversight**

Only authorized research team members will have access to identifiable data during initial processing. De-identified datasets will be shared with data analysts and physicians involved in scoring, with strict adherence to confidentiality agreements. All team members will be trained on data handling and privacy protocols in accordance with institutional ethics guidelines.

5. **Data retention and destruction**

All data will be retained securely for a period of five years following the completion of the study, in line with institutional policy and national research standards. After this period, all data will be permanently deleted from storage systems. Audio recordings will be deleted after transcription and quality control are completed, unless retention is necessary for verification.

6. **Data sharing and publication**

No individual-level data will be shared outside the research team. Findings will be reported in aggregate form, and any direct quotes from nurse interviews will only be included in publications or dissemination materials with explicit permission from the participants. Data sharing with external parties (e.g., journals or funders) will comply with institutional and legal guidelines, always ensuring participant anonymity.

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

VIII. Outcomes and Measures

A. Primary objective

Two primary outcomes will be measured based on participant subgroups. The **management reasoning outcome** will be assessed among patients with a known diagnosis of hypertension or diabetes, while the **diagnostic reasoning outcome** will be evaluated among patients presenting with symptoms such as fever, breathlessness, or musculoskeletal pain. For both outcomes, each patient interaction—whether in the standard of care or intervention arm—will be scored by licensed physicians (evaluators) using pre-defined assessment criteria. The primary outcome is the absolute total score of the consultation transcript on the grading rubric defined in Annexure 1: 0–8 for Disease (clinical management) cases and 0–12 for Symptom (clinical reasoning) cases. For reporting, the total is also expressed on a normalised 0–100% scale for concordance with the originally registered outcome scale; within a sub-study this is a linear rescaling of the total and leaves all model estimates, test statistics, and inference identical. The Disease and Symptom sub-studies are analysed separately; there is no pooled cross-sub-study estimate of the treatment effect.

B. Secondary objective I

The patient exit survey is designed to assess key implementation outcomes of the AI-assisted, nurse-led care model from the patient perspective. The primary outcomes include perceived communication quality, trust and comfort with AI use, and overall acceptability and satisfaction. These will be measured using structured Likert-scale items adapted from validated tools, including the WHO Health System Responsiveness framework and the PSQ-18. Additional items explore willingness to receive similar care in the future, as a proxy for future acceptability. Quantitative responses will be analyzed using descriptive statistics and composite scores for key domains, while qualitative data will undergo thematic analysis to contextualize patient experience and expectations. The qualitative data will explore themes like those in the patient exit survey and will add contextual richness and nuance to the patients' views.

C. Secondary objective II

The in-depth interviews with nurses are designed to assess their experience using an AI-enabled clinical decision support tool in routine patient care. The interviews will explore themes related to the feasibility of using the tool (such as ease of use, impact on workflow, and access to required resources), and its acceptability (including trust in recommendations, comfort during use, and willingness to continue use in the future). Interviews will also examine practical challenges, such as digital literacy, internet connectivity, and preferred modes of interaction (text or voice).

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

IX. Analysis Plan

A. Primary objective

Rubric-based grading:

Transcripts and treatment plans from the nurse and doctor consultations for every patient will be graded using the domain-based rubric in Annexure 1. The rubric defines six domains, each scored on three anchored levels (0 = inadequate, 1 = suboptimal, 2 = optimal): Quality of history; Quality of differential; Accuracy of provisional diagnosis; Accuracy of next steps; Safety; and Comprehensiveness. Symptom (clinical reasoning) cases are scored on all six domains (maximum 12 points). Disease (clinical management) cases are scored on four domains — Quality of history, Accuracy of next steps, Safety, and Comprehensiveness (maximum 8 points).

The Safety domain is anchored as: 0 = life-threatening harm (a missed do-not-miss diagnosis, or a missed appropriate referral or return precaution); 1 = low-to-moderate risk of harm (financial harm from unnecessary testing or medication, psychological distress, or avoidable delay); 2 = safe. This anchored Safety domain replaces the earlier practice of assigning negative points for individual harmful steps; harm is now captured within the Safety domain. This rubric was locked before enrolment of main-trial participants; the earlier condition-specific point schemes above are superseded.

Each transcript is scored by two independent physician graders, blinded to each other's assessments. Grading is performed by a pool of qualified physician graders rather than a single fixed pair; grader characteristics will be reported with the analysis. Patients are assigned to grader pairs by balanced rotation through the available pairs, constructed independently per site, and both transcripts for any one patient (Doctor and Nurse+AI) are always scored by the same grader pair, so that any grader's systematic severity affects both arms equally and cancels in the within-patient contrast. Study-arm labels are masked on the transcript files sent to graders; because the format of an AI-mediated consultation may be identifiable from transcript content, residual detection bias on the subjectively scored outcome cannot be excluded and is acknowledged as the design's principal measurement limitation.

Discordant scores between the two graders are not adjudicated: there is no third arbitrator and no discussion between graders. Both graders' scores enter the analysis directly, and grader-level severity is absorbed by a grader random effect in the primary model. Inter-rater reliability for the rubric will be reported with the results. If a grader completes only one, or neither, of a patient's two transcripts by grading close, that partial work is excluded and both transcripts are regraded by a fresh pair; the analysis set therefore contains only patients both of whose transcripts were scored by the same complete grader pair.

The primary analysis will use the same mixed-effects model on all randomized patients with score for both intervention and SoC consultations. Each sub-study (Disease,

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Symptom) is analysed separately with a pre-specified linear mixed-effects model of the total score on treatment (Nurse+AI vs Doctor), order (first vs second consultation) and site as fixed effects, with random intercepts for patient, transcript (patient-by-arm), and grader — the grader term absorbing grader severity, which matters because discordant scores are not adjudicated. One row per grading is used, so both graders' scores enter the model directly. Provider (doctor and nurse) effects are treated as fixed characteristics of the observed consultations rather than random draws from a clinician population, because the estimand is conditional on the small number of study clinicians. The null hypothesis

$$H_0: \mu_{\{Nurse+AI\}} - \mu_{\{Doctor\}} = 0$$

will be tested with a two-sided 95% confidence interval. We will report the point estimate, CI, P-value, and intraclass correlations. Sensitivity analyses will repeat the model without provider random effects and with robust standard errors. The registered $\pm 3.5\%$ is retained as the design's minimum detectable difference and reported alongside the confidence interval as an interpretive benchmark — equivalent to ± 0.28 points (Disease, /8) and ± 0.42 points (Symptom, /12); it is not an equivalence or non-inferiority margin, and no equivalence conclusion is drawn.

Pre-specified model-selection fallback ladder. If a random-effect variance is estimated at zero or the fit fails to converge, the model descends one rung by a mechanical rule applied once at the pre-specified fit, with no analyst discretion: Rung 1 = the full specification above (always attempted and reported); Rung 2 = drop a zero-estimated random effect other than the transcript term; Rung 3 (floor) = grader-averaged paired model, score $\sim$ treatment + order + site + (1 | patient). The rung used for the primary estimate is reported with it.

Missing-score analyses (pre-specified). A patient may lack a score for one arm through an intercurrent event (failed or aborted consultation, or withdrawal) or a missing measurement (transcript lost or grading not completed). Because such loss may differ by arm, three pre-specified analyses — all run at the grader-averaged patient level — accompany the primary: (1) a composite zero-score refit, setting a missing arm's total to 0 identically in both arms; (2) deterministic bounds, bracketing the estimate by setting missing scores to the scale extremes in each direction; and (3) a tipping-point grid, sweeping imputed totals over a pre-specified grid to report where the primary conclusion would change. If no enrolled patient is missing a score at the analysis cut, all three are reported as not applicable; none replaces the primary. Every enrolled patient is accounted for in a CONSORT-style flow, with a per-arm reason code for any transcript unavailable for analysis.

Sample size (no change). The registered sample size — 336 enrolled participants per outcome (672 total, inclusive of the 5% dropout allowance; ≈ 320 analysable per outcome) — is unchanged by this amendment. The planned pilot-based reassessment of the sample-size assumptions was executed once, before any main-trial data were graded; it confirmed adequate power under the refined rubric and analysis and indicated no change greater than the pre-specified 10% threshold, so the registered sample size stands.

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

B. Secondary objective I

The analysis of the patient exit survey will focus on quantitative outcomes derived from Likert scale Descriptive statistics will be used to report frequencies, percentages, means, and standard deviations for each item. Where appropriate, composite domain scores will be calculated by averaging responses within each domain, and internal consistency will be assessed using Cronbach's alpha. Responses to yes/no and categorical questions (e.g., willingness to receive similar care again) will be summarized using proportions. See the survey tool for more details on the questions. For the patient qualitative data, we will use an abductive approach for coding the transcripts using Dedoose, a collaborative qualitative analysis software(*Abductive Coding in Qualitative Research*, 2025). A codebook will be initially developed based on *a priori* assumptions and will be further refined after inductively coding three transcripts. This finalized codebook will then be applied to all transcripts for coding and analysis.

C. Secondary objective II

The analysis for the IDIs with nurses will follow the same approach as described above for the IDIs with patients. The initial codebook will be based on *a priori* assumptions for the interviews with nurses and iteratively refined through inductive coding.

X. Ethical Considerations

Informed Consent

Written informed consent will be obtained from all participants before enrollment. A trained research assistant will explain the objectives of the study, study procedures, the expectations from the participant and the financial reimbursement to be offered as a gesture for the participant's time. The participant will also be briefed on their right to withdraw at any point in the study, the steps taken to ensure confidentiality of the information collected as part of the study and the risks (which are minimal) of participating. For those who consent, a signature / thumb impression and the date will be recorded in the consent form. The verbal explanation of the consent process and the forms will be available in both Hindi and Bengali – the local languages spoken in the study sites. A separate signed consent will be obtained for the small group of patients selected for the IDIs.

Doctors and nurses will also be asked to consent for their participation in the study. Both forms will elaborate on the purpose and objectives of the study, the expectations from the individuals, confidentiality issues, and their right to voluntary withdraw from the study. For the nurses, the same consent form will also cover their participation in an in-depth interview which will last for 45 to 60 minutes. Signatures will be collected for all consenting doctors and nurses. These consent forms will be in English.

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

XI. Study Risks

This study involves minimal risk to participants. The following potential risks have been identified, along with corresponding mitigation strategies:

1. Breach of confidentiality

Risk: The study involves collection of audio recordings and clinical data, which could lead to unintended disclosure of personal health information.

Mitigation: All data will be de-identified prior to transcription and analysis. Unique participant IDs will be used in place of names or other identifiers. Files will be stored in a password-protected, encrypted cloud-based system with access limited to authorized members of the research team.

2. Discomfort due to audio recording

Risk: Patients, doctors, or nurses may feel uncomfortable being audio recorded during clinical consultations or interviews.

Mitigation: Participation in audio recording is entirely voluntary. During the informed consent process, participants will be clearly informed about the purpose of the recordings, how they will be used, and the steps taken to protect their privacy. Participants may decline recording or withdraw at any time without penalty.

3. Disruption to clinical workflow

Risk: Dual consultations (doctor and nurse + AI) may extend visit time, disrupt routine clinic operations, or increase patient burden.

Mitigation: Study scheduling will be coordinated to avoid peak hours and minimize operational disruption. Participants will be informed of the additional time commitment during consent and compensated appropriately for their time.

4. Emotional or cognitive burden during nurse interviews

Risk: Nurses may experience discomfort or fatigue during the in-depth interviews.

Mitigation: Participation in interviews is voluntary. Interviews will be conducted in a semi-private setting and will be scheduled at convenient times. Nurses may skip questions or terminate the interview at any point without consequence.

5. Over-reliance on AI recommendations

Risk: There is a possibility that nurses may depend too heavily on AI outputs, reducing their use of independent clinical judgment.

Mitigation: All participating nurses will be trained to use the AI tool strictly as a

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

decision-support system. The AI model will be vetted for safety and usability in advance. Clinical supervision will be available during the implementation phase. It will also be reiterated to the patients that recommendations from nurse are for study purposes only - and that they should only follow the doctor's clinical advice.

6. Sequence effects and response bias

Risk: The order in which patients receive consultations (doctor first vs. nurse + LLM first) may influence their experience or survey responses.

Mitigation: Participants will be randomly assigned to one of two sequences. The exit survey will be administered immediately after the nurse + LLM consultation and will focus only on that experience. The analysis will control for consultation order to adjust for sequence effects. Survey language will be neutrally framed to avoid bias.

XI. Timeline

Below is an abridged version of the timeline. See attached MS excel sheet for a detailed workplan by site.

Activities	Aug 2025	Sept	Oct	Nov	Dec	Jan 2026	Feb	Mar	Apr	May
Protocol finalization & IRB submission	X									
Study team recruitment	X	X								
Curriculum development	X									
Training		X	X							
HCD workshop				X						
Pilot testing & primary data collection					X	X	X	X		

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Analysis, report writing									X	X
--------------------------------	--	--	--	--	--	--	--	--	---	---

XII. Dissemination Plan

Findings from this study will be shared with key stakeholders to inform the responsible use of AI-assisted, nurse-led care in primary health settings. Local dissemination will include feedback sessions with participating nurses, clinic staff, and district health officials, where results will be presented in non-technical language and local translations where needed. Summary findings will also be shared with the State Health Department to support future planning and scale-up discussions. Broader dissemination will include presentations at academic and public health forums, and publication in peer-reviewed journals. All results will be reported in aggregate, with no individual- or facility-level identifiers. Participants will not be named in any dissemination materials. Any materials used for training or implementation improvements will be shared openly and credited appropriately. The dissemination process will prioritize respectful engagement with participants and health system partners and will uphold principles of confidentiality, transparency, and equitable knowledge sharing.

XIII. Potential Benefits

While there is no direct clinical benefit guaranteed to participants, this study has the potential to generate significant benefits at the individual, health system, and policy levels. For patients, the study may enhance the quality of care received by enabling more structured and thorough consultations, particularly with AI-supported clinical reasoning. For nurses, the intervention offers an opportunity to build confidence in clinical decision-making with supportive technology, potentially improving job satisfaction and professional autonomy. At the health system level, the findings may inform scalable strategies for addressing provider shortages in rural and resource-limited settings by empowering frontline workers with AI-enabled tools. More broadly, this study will contribute critical real-world evidence on the safety, usability, and acceptability of large language models in primary care delivery in LMIC settings—an area with limited existing research. These insights can guide future investments and policy decisions on responsible AI integration in healthcare.

XIV. Limitations

This study has several limitations.

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

- i. First, as a clinic-based implementation study conducted in selected rural primary care settings, findings may not be generalizable to other regions or higher-level health facilities.
- ii. Second, the evaluation of clinical quality is based on physician-adjudicated scoring of recorded interactions, which, despite using standardized criteria, may introduce some subjectivity.
- iii. Third, while patient satisfaction is appropriately measured through self-report, there remains potential for social desirability bias, even with the use of tablet-based, self-administered surveys. Although this approach reduces interviewer influence and increases privacy, patients may still moderate their responses due to perceived surveillance or limited digital literacy.
- iv. Fourth, the in-depth interviews with nurses are qualitative and based on a small sample, which may not capture the full range of experiences and could be influenced by recall bias.
- v. Lastly, the study captures short-term implementation outcomes and does not assess long-term effectiveness, sustainability, or cost-effectiveness, which will require further research.

References

Abductive Coding in Qualitative Research: A Practical Guide. (2025, March 31). Delve. <https://delvetool.com/blog/abductive-coding>

Anjana, R. M., Unnikrishnan, R., Deepa, M., Pradeepa, R., Tandon, N., Das, A. K., Joshi, S., Bajaj, S., Jabbar, P. K., Das, H. K., Kumar, A., Dhandhanian, V. K., Bhansali, A., Rao, P. V., Desai, A., Kalra, S., Gupta, A., Lakshmy, R., Madhu, S. V., ... Ghosh, S. (2023). Metabolic non-communicable disease health report of India: The ICMR-INDIAB national cross-sectional study (ICMR-INDIAB-17). *The Lancet Diabetes & Endocrinology*, 11(7), 474–489. [https://doi.org/10.1016/S2213-8587\(23\)00119-5](https://doi.org/10.1016/S2213-8587(23)00119-5)

Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimplouras, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., & Singhal, K. (2025). *HealthBench: Evaluating Large Language Models Towards Improved Human Health* (arXiv:2505.08775). arXiv. <https://doi.org/10.48550/arXiv.2505.08775>

Darby, C., Valentine, N., Murray, C. J., & de Silva, A. (2000). *World Health Organization (WHO): Strategy on Measuring Responsiveness*.

Desk, E. (2025, January 31). *Economic survey flags gaps in India's medical education system, calls for reforms*. EdexLive. <https://www.edexlive.com/news/2025/Jan/31/economic-survey-flags-gaps-in-indias-medical-education-system-calls-for-reforms>

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

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>

Marshall, G. N., & Hays, R. D. (1994). *The Patient Satisfaction Questionnaire Short Form (PSQ-18)*. <https://www.rand.org/pubs/papers/P7865.html>

Nagari Village Population—Suri—I - Birbhum, West Bengal. (n.d.). Retrieved August 12, 2025, from <https://www.census2011.co.in/data/village/317069-nagari-west-bengal.html>

Nong, P., & Platt, J. (2025). Patients' Trust in Health Systems to Use Artificial Intelligence. *JAMA Network Open*, 8(2), e2460628. <https://doi.org/10.1001/jamanetworkopen.2024.60628>

Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M. T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M. P., Gross, B., Hames, P., Suleyman, M., King, D., & Horvitz, E. (2025). *Sequential Diagnosis with Language Models* (arXiv:2506.22405). arXiv. <https://doi.org/10.48550/arXiv.2506.22405>

Rao, A., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A. K., Landman, A., Dreyer, K., & Succi, M. D. (2023). Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *Journal of Medical Internet Research*, 25(1), e48659. <https://doi.org/10.2196/48659>

Salvi, S., Apte, K., Madas, S., Barne, M., Chhowala, S., Sethi, T., Aggarwal, K., Agrawal, A., & Gogtay, J. (2015). Symptoms and medical conditions in 204 912 patients visiting primary health-care practitioners in India: A 1-day point prevalence study (the POSEIDON study). *The Lancet Global Health*, 3(12), e776–e784. [https://doi.org/10.1016/S2214-109X\(15\)00152-7](https://doi.org/10.1016/S2214-109X(15)00152-7)

Tu, T., Schaekermann, M., Palepu, A., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Cheng, Y., Vedadi, E., Tomasev, N., Azizi, S., Singhal, K., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S. S., Semturs, C., ... Natarajan, V. (2025). Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067), 442–450. <https://doi.org/10.1038/s41586-025-08866-7>

Update on Ratio Of Patients And Doctors Nurses. (2023, December 12). <https://www.pib.gov.in/www.pib.gov.in/Pressreleaseshare.aspx?PRID=1985423>

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Annexure – 2 – Patient exit survey

Patient Exit Survey – LLM-Assisted Nurse Care

Estimated Completion Time: Under 5 Minutes

This survey is designed to assess your experience with today's visit. Your responses will help us improve care.

Section A: Communication and Understanding

1. The nurse understood my health problem clearly and followed up with questions adequately. (adapted from PSQ18 Q1)

☐ Agree ☐ Neutral ☐ Disagree

Section B: Trust and Comfort with AI Use

2. I was comfortable that the nurse used a computer or phone to assist my care. (Nong & Platt)

☐ Agree ☐ Neutral ☐ Disagree

Section C: Respect and Satisfaction

3. I would be willing to receive care like this again.

☐ Yes ☐ Maybe ☐ No

Annexure 3 – In-depth interview guide for patients

In-Depth Interview Guide: Patient Experience with Nurse + AI Consultation

Estimated Duration: 30–45 minutes

Purpose: To explore in greater depth how patients perceived the nurse-led consultation assisted by an AI tool. This includes their understanding, trust, comfort, and overall acceptability of this new model of care.

Section 1: Introduction

Thank you for agreeing to speak with us today. This interview is about your experience during your recent visit to the clinic. Your responses will help us improve how nurses and technology work together to provide better care. Everything you say will remain confidential, and your name will not be used in any reports. You can choose to skip any question or stop the interview at any time. Do you have any questions before we start?

Section 2: Communication and Understanding

1. What was the reason for your clinic visit today?
2. How well did the nurse seem to understand your problem?
 - a. Did the nurse ask you enough questions?
3. How was the counseling you received from the nurse?
 - a. Were the next steps or advice adequately explained to you?
 - b. Did the nurse answer your questions in a way that you could understand?
4. Did you feel like you had enough time to ask your questions?

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

- a. Were any medical words or terms used that were hard to follow?

Section 3: Comfort and Trust with AI Tool

5. During your consultation, you may have noticed the nurse using a mobile phone or computer. How did you feel about that? Please explain
 - a. Did the nurse tell you why they were using the device?
6. Did the use of the phone/computer make you more confident or more unsure of the clinical counseling you received from the nurse?
 - a. Did you feel that the care or advice was different because a computer tool was being used?
7. In your opinion, did you feel the tool helped or got in the way of the clinical consultation?
 - a. Do you think the nurse's time spent in talking with you was affected because of presence of a phone or computer?
 - b. If it was affected, was it a positive or negative effect? Please elaborate.

Section 4: Respect, Satisfaction, and Acceptability

8. How did you feel you were treated during the visit?
 - a. Did you feel listened to and respected?
 - b. Overall, how satisfied were you with the care you received?
9. Would you feel comfortable coming back for a clinical visit like this (with a nurse + AI tool) again?
10. Would you recommend this kind of care to a friend or family member?

Section 5: Open Feedback and Suggestions

11. What did you like most about your visit today?
12. Was there anything you didn't like or found confusing?
13. How could we improve this type of care in the future?
 - a. If we used this tool in more clinics, what would you want us to keep the same or change?

Section 6: Closing

Thank you very much for your time and feedback. We appreciate your help in making healthcare better for everyone.

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Annexure – 4 – In-depth interview guide for nurses

Title: In-depth interview guide for assessing the perceptions & acceptability of AI tool use by nurses in primary health care settings in rural India.

Icebreaker

1. How long have you been a nurse? Where did you work prior to this project?
2. How many patients have you seen during this project?
3. How has the experience been?

Training

4. You received training on medicine-specific topics and digital literacy prior to the start of the project. What is your opinion on the training material?
 - a. Did you feel enough time was spent on medical literacy? How can it be made better?
 - b. What is your opinion on the depth of teaching done for digital literacy? What changes should be made?
 - c. What was your opinion about the training material ? Was it easy to follow?
Please elaborate.
5. How was the teaching? Were you able to follow the language used?
6. How do you feel about the frequency of classes / sessions? Should it be changed? If yes, what should the change be?
7. The trainings were delivered to you online. How was your experience with online learning?
 - a. Did the sessions allow you sufficient opportunity to ask questions?
8. What suggestions do you have to improve the training experience ?

Feasibility

9. You used an AI tool to see patients. You spoke to both the patient and the AI tool during the consult. How comfortable were you with this process?
 - a. Were you able to use the tool with patients without major interruptions? How did it alter your workflow?
10. Would you rate the AI tool as an enabler or hindrance to your workflow? Please elaborate.

Acceptability

11. How did you feel using the AI tool during patient care?
12. Do you trust the recommendations it gave?
13. If there was an option to permanently integrate a similar tool in your daily work where you could see patients for common conditions and complaints, would you want to continue doing it?

Practical considerations

14. You communicated with the AI tool through text/ voice/ both. How was your experience using these modalities?

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

- a. How can it be improved? Single mode? Voice or text only?
- 15. Did internet connection affect the smoothness of your workflow?
 - a. If yes, how was your work affected?

Learnings

- 16. If you had to give 1 advice to the people who designed the nurse-AI intervention project, what would that be?

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Annexure – 5 – Consent script for patients

Patient Consent Script – Nurse-AI Study

Study Title: Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Site: [Insert site]

Principal Investigator: [Insert Name, Contact Number]

Hello, my name is [Name]. We are conducting a study at this clinic. We are testing a new AI tool that supports nurses in managing common health problems like high blood pressure, diabetes, joint pain, fever, or breathlessness.

If you agree to participate, here's what will happen:

- You will be seen by both a nurse (who will use the AI tool) and a doctor during today's visit.

- This may take a little extra time compared to a regular visit.

- At the end, you'll be asked to complete a short survey about your experience.

Please note: The nurse may give some advice based on the AI tool, but this is only for the study. You will leave the clinic with the doctor's prescription. Please follow only the doctor's advice for your treatment.

There are no medical risks in joining the study. Whether or not you take part, your regular care will not be affected. Your personal information will be kept private. We will use the information from your consultation recordings for research purposes only. This data may also be used for secondary analysis in the future.

As a thank-you, you will receive INR 250 upon completion of your visit and the survey.

Taking part is completely up to you. You can say no or stop at any time.

If you have questions after the study, you may contact the Principal Investigator at [Insert Phone Number], or the local Institutional Review Board (IRB) at [Insert IRB Contact Information].

Do you have any questions?

If you agree, we will now ask for your written consent.

☐ Yes, I agree to participate

☐ No, I do not wish to participate

Participant Name & signature : _____

Date: _____

Annexure – 6 – Consent script for patients (IDI)

Study Title: Evaluating AI-Assisted Nurse-Led Consultations in Primary Care

Introduction

Namaste / Nomoskar. My name is [Interviewer's Name], and I'm part of the research team working on a study to improve how nurses and technology can work together to provide better healthcare.

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

We are inviting you to participate in a one-time interview about your experience during your recent clinic visit, where you received care from a nurse who used a phone or computer (AI tool) to assist in the consultation.

Purpose of the Interview

The goal of this interview is to better understand what patients like you thought and felt about this new way of receiving care. Your feedback will help us improve the quality and experience of such care for others.

What Will Happen

- If you agree, we will talk for about 30 to 45 minutes.
- We will ask you about how well you understood the consultation, how comfortable you felt, and if you would be open to receiving similar care in the future.
- With your permission, the conversation will be audio recorded, so we don't miss anything important you say.

Voluntary Participation

- Your participation is completely voluntary.
- You can refuse to answer any question or stop the interview at any time without giving a reason.
- Saying no will not affect your care in any way.

Privacy and Confidentiality

- Your name or identity will not be shared in any reports.
- Audio recordings will be stored securely and only accessed by the research team.
- If we use any direct quote from your interview in a report or publication, we will only do so with your permission.

Risks and Benefits

There are no direct risks or benefits. But your feedback may help improve how care is provided in clinics like this one.

Incentive

As a small thank you for your time, we will provide INR 250 after the interview is completed.

Do You Have Any Questions?

[Pause for any questions]

If you are willing to participate, we will ask you to sign or give your thumb impression on a consent form. Do you agree to take part in this interview?

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Annexure – 7 – Consent script for nurses

Study Title: Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Principal Investigator: [Insert Full Name, Qualifications, Contact Information]

Study Site:[Insert Health Facility Name and Address]

You are being invited to participate in a research study being conducted by [Name of Institution/NGO]. This study aims to assess the effectiveness of an AI-enabled-nurse treatment care model for managing common chronic conditions and symptoms in a primary care setting in rural India.

Before you decide, please read the following information carefully. You are welcome to ask questions before agreeing to participate.

Your role and responsibilities:

If you agree to participate, you will be expected to:

- Participate in a training program about medical and AI literacy.
- Do a clinical consultation with patients using the AI-supported clinical decision tool.
- Allow your consultations with patients to be audio recorded. These recordings will help researchers study how the tool is used in real clinical interactions.
- Participate in a one-on-one in-depth interview with a member of the research team. This interview will explore your experience using the AI tool, including what worked well and what challenges you faced. The interview will take approximately 45 to 60 minutes and will be scheduled at your convenience.

Your identity and all personal information will be kept strictly confidential. Audio recordings will be securely stored and only accessible to the core research team. All findings will be reported in aggregate or anonymized form. If any direct quotes from your interview are used in research outputs (such as reports, presentations, or publications), they will be included only with your explicit permission.

Your participation is entirely voluntary. You are free to decline or withdraw from the study at any time.

There are no known risks associated with your participation. Your insights and feedback will help improve the design and implementation of digital tools for nurses in clinical settings.

If you have any questions, please contact:

Principal Investigator: [Insert Name, Phone Number, Email]

Institutional Review Board (IRB): [Insert IRB Contact Information]

Consent Statement

☐ I have read and understood the information above. I have had the opportunity to ask questions and all my questions have been answered. I voluntarily agree to participate in this study, including having my consultations audio recorded and participating in an in-depth interview. I understand that any direct quotes from my interview will only be used with my permission.

Nurse's Name: _____

Signature: _____

Date: _____

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Annexure – 8 – Consent script for doctors

Study Title: Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Principal Investigator: [Insert Name, Contact Number]

Site: [Insert Facility Name]

Dear Doctor,

You are being invited to participate in a research study that aims to assess the effectiveness of an AI-enabled-nurse treatment care model for managing common chronic conditions and symptoms in a primary care setting in rural India.

If you agree to participate:

- You will see patients presenting with two chronic conditions (e.g., hypertension and diabetes) and three common symptoms (e.g., fever, breathlessness, and joint pain) as part of your routine consultations.
- After each consultation, you will be asked to complete a brief proforma documenting key clinical details, which is expected to take 5 to 8 minutes.
- The audio of your consultation with each patient will be recorded. These recordings will be used solely for analysis of clinical communication and decision-making.

Your identity will be kept strictly confidential. No audio recordings will be shared outside the research team. The audio recordings will be retained for 5 years, as stipulated by national and local IRB rules and regulations. All data will be used in aggregate form only. No quotes or identifying information will be used in any reports or publications. Your participation is entirely voluntary. You may withdraw from the study at any time without any consequences to your professional standing or responsibilities.

If you have any questions about the study, you may contact the Principal Investigator at [Insert Contact Number] or the local IRB at [insert contact details]

****Consent:****

☐ I have read and understood the purpose of the study and my role in it. I voluntarily agree to participate.

Doctor's Name: _____

Signature: _____

Date: _____

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Annexure 1 (revised) — Grading rubric

This revised rubric replaces the previous condition-specific grader sheets in Annexure 1 (the symptom sheet of 3 differential diagnoses, final diagnosis and 5 next steps, and the disease sheet of review of data, screening for complications, medication-compliance assessment, counselling and treatment plan). The patient-overview header captured for each patient is unchanged: age; unique study ID; nurse ID; doctor ID; time spent in the nurse encounter; time spent in the doctor encounter; study complaint; and any system challenges due to internet or electricity.

Domains used by case type (all domains scored 0 / 1 / 2)

Domain	Disease (HTN, DM)	Symptom (fever, breathlessness, MSK pain)
Quality of history	✓	✓
Quality of differential	—	✓
Accuracy of provisional diagnosis	—	✓
Accuracy of next steps	✓	✓
Safety	✓	✓
Comprehensiveness	✓	✓
Maximum total	8	12

Anchor definitions

Domain	0 — Inadequate	1 — Suboptimal	2 — Optimal / good enough
Quality of history (rapport, respect, extraction of red flags)	Dismisses the chief complaint or red-flag symptoms raised during the interview.	History taking is coherent but narrow: only confirmatory questions asked; discriminating questions for differentials or do-not-miss (DNM) diagnoses omitted.	Asks discriminating questions across multiple diagnoses; elaborates comorbidities/complications; follows up on red-flag symptoms raised by the patient.
Quality of differential (breadth and DNM list) — Symptom only	Extremely narrow: misses locally endemic or life-threatening causes not excluded by the interview.	Narrow or illogical: misses a plausible system, or lists a diagnosis unsupported by the transcript.	Broad and logical: covers multiple systems unless the ground truth strongly limits it.
Accuracy of provisional diagnosis (most likely cause, local prevalence) — Symptom only	Wrong: inadequate to explain the salient clinical features.	Plausible but imprecise (e.g. “infection” rather than UTI), or an alternate diagnosis explaining many features.	Most likely diagnosis: accounts for local prevalence and the transcript evidence.
Accuracy of next steps (plan to reach diagnosis and treat)	Ineffective: medically incorrect test or treatment.	Correct treatment but wrong/not-first-line, or a correct but expensive test where a simpler point-of-care test exists.	Precise, evidence-based, cost-effective; prioritises the most useful local diagnostics.
Safety (harm/waste from over- or under-treatment)	Life-threatening harm: DNM diagnosis not considered; missed appropriate referral or return precaution.	Low-to-moderate risk of harm: financial harm (unnecessary testing/medication), psychological distress, or avoidable delay.	Safe: minimised out-of-pocket waste; covered DNM diagnoses through testing, referral, or return precautions.

Assessing the effectiveness of Large Language Model (LLM) - enabled nurse treatment planning in 2 Indian districts: A pilot study

Domain	0 — Inadequate	1 — Suboptimal	2 — Optimal / good enough
Comprehensiveness (all concerns addressed; patient counselled)	Missing most of: education; counselling; follow-up plan; appropriate treatment and diagnosis.	Some treatment elements present but a key prescription, test, or counselling point missed.	Thorough: disease/treatment education; lifestyle and compliance counselling; follow-up and return precautions; complete treatment and diagnosis.

Grading procedure. Both transcripts for a patient are scored by the same pair of physician graders, independently and blinded to each other, with study-arm labels masked. Each grading records a score of 0, 1, or 2 for every applicable domain; totals are never prorated from partial domains — a grading missing any applicable domain score is treated as an incomplete grading and excluded. Discordant scores are not adjudicated. The condition-specific grader guidance (do-not-miss red flags, expected differentials, priority tests and treatments, and referral triggers) is maintained as a grader instruction sheet available on request.