Statistical Analysis Plan; Can Feedback from a Large Language Model Improve Health Care Quality?

Jason Abaluck[*], Robert Pless[†], Nirmal Ravi[‡], Anja Sautmann[§], Aaron Schwartz[¶]

April 29, 2025

1 Introduction

The quality of healthcare in low- and middle-income countries is notoriously low (Banerjee et al., 2023). In Nigeria, only a small proportion of primary care patients are seen by a Medical Officer instead of a mid-level provider such as a Community Health Extension Worker or Community Health Officer. Access to qualified medical doctors is a barrier to good health care (Okeke, 2023).

This project tests whether Large Language Models (LLMs) can improve patient care in Nigerian primary care clinics by giving customized and instant feedback to the provider in natural language. An LLM-based tool integrated into an electronic patient record management system provides "second opinions" to community health extension workers (CHEWs) at two clinics in Nigeria. These second opinions are intended to mirror what a reviewing physician might advise the CHEWs after seeing or hearing their initial report on a patient.

2 Experimental Design

For our main analysis, we use a within-patient comparison of two patient notes created by the CHEW; one during the initial patient consultation, and one after the LLM feedback was received. The patient is also seen by a fully trained medical officer who is in charge of patient care. The MO conducts a blinded review of the CHEW's patient notes to measures changes in the CHEW's care as a result of the LLM feedback. Our data comes from the information captured in the electronic medical record (EMR) of the patient and from survey data collected from CHEWs, reviewing MOs, and a panel of reviewing Medical Doctors (see below).

As part of the research protocol, we also conduct screening tests for malaria, anemia, and UTI (urine analysis). Eligibility for each test is determined using a set of symptoms explicitly checked by the enumerator, and patient demographics (see Appendix A). Tests that were already ordered by the MO and carried out are not repeated (based on EMR information). The criteria were designed as broad as possible, with the goal to test any patient who should be tested or would be recommended for testing in typical clinic practice by at least some providers. Note that CHEWs as part of this study are not able to order any medical tests for ethical reasons. In daily practice, they have a limited set of tests available, primarily rapid diagnostic tests (MOs can order all medical tests that CHEWs can order plus others).

2.1 Sample size

We work with a private clinic and the project is grant funded. The clinic manages the budget directly and projects that funds are sufficient to collect valid data from approximately 500 patients. The intervention is carried out by 20 CHEWs who are rotating through two clinics. Data collection will be conducted until the grant budget is exhausted.

[*] Yale School of Management, Yale University

[†] Department of Computer Science, George Washington University

[‡] EHA Clinics Nigeria

[§] Development Research Group, The World Bank, asautmannworldbank.org

[¶] Department of Medical Ethics and Health Policy and Department of Medicine, University of Pennsylvania School of Medicine

2.2 Patient EMR data

All patient data is saved in the EMR. At various points, this data is collated into SOAP notes (structured text files describing subjective and objective signs and symptoms, the provider's assessment (diagnosis), and the treatment plan).

The SOAP notes that are saved are:

- "Unassisted" SOAP note created on initial patient consultation and submitted to LLM: contains care plan conditional on results of requested medical tests, randomly labeled SOAP A or B
- "Assisted" SOAP note created after LLM feedback was received: conditional care plan, randomly labeled B or A
- Conditional MO SOAP note: conditional care plan, created by MO prior to receiving laboratory test results
- Patient discharge SOAP note: patient care plan created by MO after receiving laboratory test results and communicated to patient
- Final MO SOAP note: patient care plan created after reviewing study medical tests (malaria, anemia, UTI).

In addition, we have the values of individual EMR fields that make up these SOAP notes, e.g. test results, diagnoses, prescriptions, etc., including patient demographics and vital signs.

2.3 Survey Data

We collect various forms of survey data; the key information about the quality of care provided by the CHEW (either without or with LLM assistance) comes from the MO's evaluation of the CHEW's SOAP notes for every patient. We also collect CHEW feedback on the LLM's advice, for each patient and overall; and a survey for a panel of MDs who conduct a second evaluation of the CHEW notes, review the LLM feedback, and assess the MO's care plan.

2.3.1 The MO's SOAP note evaluation

The evaluation focuses on treatment errors in the SOAP note that could cause some form of (medical) harm to the patient.

For each SOAP note, the MO first takes stock of qualitative differences between the CHEW's and the MO's plan, then assesses the type of harms for the patient from any errors, and finally rates the SOAP note on a scale for "healthy time lost". The MO is blinded to whether the SOAP note is LLM-assisted or unassisted and rates SOAP Note A before seeing SOAP Note B.

Deviations from the MO's SOAP The MO is asked to assess for each SOAP note whether medical tests ordered were necessary or clinically useful vs. unlikely to be useful, whether there are missing or incorrect/unnecessary diagnoses, and whether there are missing or incorrect/unnecessary treatment plan elements.

Types of harm incurred The MO is asked to assess any short-term harm (additional symptoms or discomfort for some period), and any long-term serious harm (risk of impairment, death etc.) from the treatment plan in the SOAP note.

Measuring Healthy Time Lost in DALY The MO also provides an overall rating that is intended to reflect the "healthy time lost" from any errors in treatment in the SOAP note. For each assessment and plan constructed by a CHEW (with or without LLM advice), an MO will assess the expected magnitude of healthy life that would be lost if the CHEW plan were implemented instead of the MO's plan. Healthy time is measured in units of disability-adjusted life year (DALYs), which reflect both length and quality of life. Losing one DALY is equivalent to losing a year of life in perfect health, or two years of life experienced in an unhealthy state with a utility weight of 0.5, etc.

MOs estimate DALY losses with the aid of a DALY scale. The DALY scale consists of a set of cutoff points that were calibrated using benchmarks of expected DALY loss for clinical mistakes relevant to local medical practice. Because the DALY consequences of permanent harm (death or disability) are higher for children than for adults, and because different clinical mistakes are relevant for children and adults, we provide a separate DALY scale for the evaluation of CHEW notes for patients who are children (under 18) and patients who are adults.

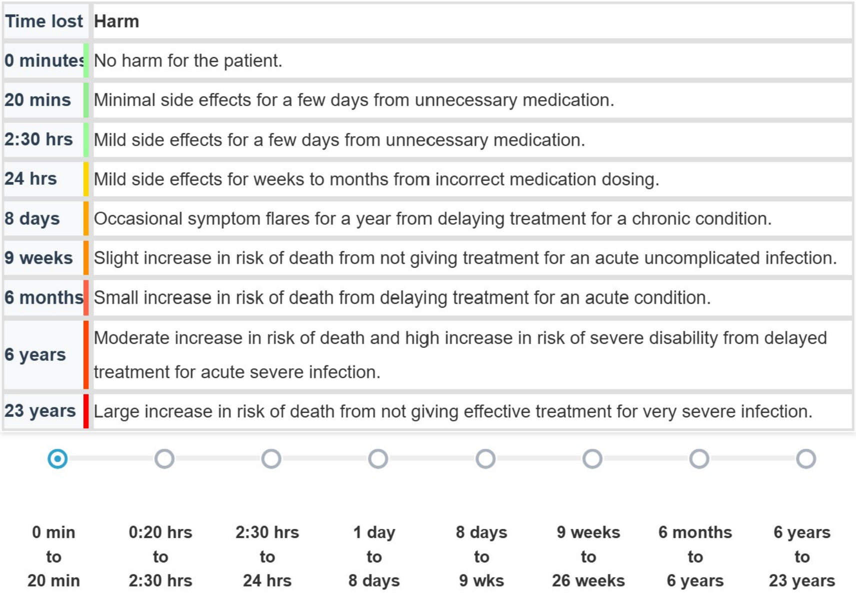
DALY benchmarks are derived from previously published DALY or QALY estimates from the medical research literature, and from novel calculations of expected harms, which employ prior estimates of the utility weights of various health states, of the effects of various medical interventions, and of life expectancy in Nigeria (both absolute and quality adjusted).

To create a consistent scale, specific DALY benchmarks are associated with cutoff points that are described in broad clinical terms, capturing the general nature of the harms.

MOs will first be presented with the following survey prompt:

> "Consider the overall health harms from the treatment errors in SOAP Note [A/B] that you just described (temporary symptoms and risk of longterm or permanent harm).
>
> The scale below shows levels of harm that average [adults/children] experience due to different medical errors, ordered from least to most severe.
>
> Harm is expressed in terms of 'healthy time lost'.
>
> *For example, unpleasant symptoms for 6 days reduce quality of life by 1/3 for those days, resulting in the equivalent of 2 days of life lost.*
>
> Please select the appropriate interval of the harm scale for SOAP Note [A/B]."

Below this prompt, the scale of different levels of harm is shown in table form, as in figure 1.



Time lost	Harm
0 minutes	No harm for the patient.
20 mins	Minimal side effects for a few days from unnecessary medication.
2:30 hrs	Mild side effects for a few days from unnecessary medication.
24 hrs	Mild side effects for weeks to months from incorrect medication dosing.
8 days	Occasional symptom flares for a year from delaying treatment for a chronic condition.
9 weeks	Slight increase in risk of death from not giving treatment for an acute uncomplicated infection.
6 months	Small increase in risk of death from delaying treatment for an acute condition.
6 years	Moderate increase in risk of death and high increase in risk of severe disability from delayed treatment for acute severe infection.
23 years	Large increase in risk of death from not giving effective treatment for very severe infection.

◉	○	○	○	○	○	○	○
0 min to 20 min	0:20 hrs to 2:30 hrs	2:30 hrs to 24 hrs	1 day to 8 days	8 days to 9 wks	9 weeks to 26 weeks	6 months to 6 years	6 years to 23 years

Figure 1: Initial Rank Ordering (Adult Scale)

The MO uses the multiple-choice response option in the bottom to select an interval on the scale.

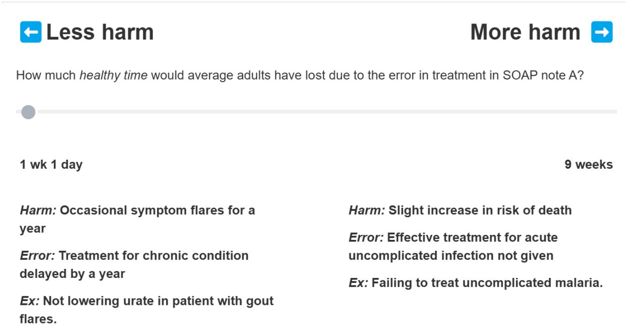
On the next screen, they see a detailed description of the two scale end-points and use a slider to choose the estimated level of harm between those two points, see Figure 2 for an example.

The DALY scale has 9 benchmarks from 0 minutes to 23 years (28 years for children), generating 8 "bins" each. They are described in detail in Appendix C.

2.4 CHEW Feedback

After completing the evaluation of each patient, CHEWs choose on a scale from strongly disagree to strongly agree for three statements:

> "In my opinion, the LLM feedback helped me improve the documentation of the patient's case."
>
> "In my opinion, the LLM feedback helped me provide better healthcare for the patient.



How much *healthy time* would average adults have lost due to the error in treatment in SOAP note A?

Harm: Occasional symptom flares for a year

Error: Treatment for chronic condition delayed by a year

Ex: Not lowering urate in patient with gout flares.

Harm: Slight increase in risk of death

Error: Effective treatment for acute uncomplicated infection not given

Ex: Failing to treat uncomplicated malaria.

Figure 2: Example of Healthy Time Loss Quantification

> *Like avoiding an unneccessary test, prompting me to provide better advice to the patient, or prompting me to choose a better medication*"
>
> "In my opinion, the LLM feedback contained a medical error."
>
> (If yes: "Please briefly describe the error.")

In addition, CHEWs respond to a more general feedback questionnaire when they rotate out of the study.

2.5 MD Panel

A panel of three Medical Doctors (MDs) will review the patient data (CHEW notes, MO notes, and LLM feedback).

Flagging MO error. In a first step, they will review the MO notes only and record whether there is any error in the diagnosis or treatment proposed in the conditional note or in the final note. If an error is identified the MDs will rate the error by severity to distinguish medical mistakes from differences in opinion about a patient who is not present.

SOAP note rating. Next, the MDs will carry out an independent evaluation of the assisted and unassisted SOAP notes as described in the sections "Types of harm incurred" and ""Measuring Healthy Time Lost" (section 2.3.1 above).

LLM review. Finally, the MDs will also review the LLM feedback and answer the following questions:

> "Did the CHEW follow all, some, or none of the LLM recommendations?"
>
> If some or none: "Imagine the CHEW had followed all the recommendations of the LLM. Would the resulting treatment plan be an improvement over their assisted note?" (Yes/no)
>
> If yes: "Please explain."
>
> "Did the LLM make any mistakes?" (Yes/no)
>
> If yes: "Was any aspect of the CHEW's assisted treatment plan worse than the unassisted plan because the CHEW followed the LLM's erroneous recommendation?"
>
> If yes: "Please explain."

3 Primary Outcomes

Our study will have four primary outcomes, each assessing differences between the CHEW's assisted and unassisted note.

1. An indicator for an error in the treatment plan in the SOAP note (with the potential for harm)
2. An indicator for an error in the treatment plan that causes a loss of at least X quality-adjusted life days (where X is the highest benchmark on the appropriate DALY scale so that at least 5% of patients have an error that large in the unassisted SOAP note)
3. An indicator for the better treatment plan

4. Consistency of treatment with the "standard of care" based on our screening tests and medical guidelines.

We describe the construction of each metric in detail below.

3.1 Estimation

The primary analysis estimates the effect of LLM assistance on the four outcomes above, all broad indicators for the overall quality of care provided. Every outcome is measured both for assisted and unassisted SOAP notes. We will estimate

$$Y_{it} = a_i + bT_k + cX_{ik} + \epsilon_{ik}$$

where i is the patient (or patient-test combo in 3.5) and $k = \{A, B\}$ is the SOAP note. Y_{ik} is the outcome, a_i is a patient (patient-test) fixed effect, T_k is an indicator for "assisted" and X_{ik} are optional controls for order effects and unblinding (see next section).

We will calculate robust standard errors. In addition, we will separately report p-values using the Benjamini Hochberg procedure for controlling false discovery rates at $p = 0.05$. This will lead to a conservative estimate of the power of our statistical tests. We will also separately report means, standard deviation, and number of observations of each variable for the assisted and unassisted notes.

3.2 Any Treatment Error

During SOAP note evaluation, the MO is asked to indicate whether the treatment plan for the patient contains any error, conditional on the MO's own diagnosis.

The introductory text (here for SOAP Note A) is:

> Treatment in SOAP Note A:
>
> Please evaluate whether the treatment in SOAP Note A is appropriate for this patient's condition. Please base this on *your own diagnosis*, not the CHEW's diagnosis in SOAP Note A.
>
> The treatment includes medications as well as other care instructions, such as information on home care or referrals.

This is followed by the question:

> Is the treatment plan for the patient in SOAP Note A completely appropriate given your *own* diagnosis (accounting for conditional treatments based on medical tests)?
>
> *Answer "No" if the patient should receive different medical care given your diagnosis. This can include both minor differences (for example, the patient should be advised to rest) and major errors (for example, the patient should receive a completely different set of medications).*
>
> (Answer options: yes/no/unsure)

If the answer is yes, the MO is asked to confirm once more. The same question is asked for SOAP Note B. We will code this question as an indicator variable that equals 1 if the treatment in the SOAP note contains an error and 0 otherwise.

3.3 Severe Treatment Error

The distribution of treatment error harm ratings is expected to be heavily skewed, with a large proportion of zeros.

We define severe errors as any errors (as above) that generate a harm rating at or above the 95th percentile of harm on the unassisted scale (pooling child and adult scales).

3.4 An indicator for the better treatment plan

Based on the DALY rating of SOAP Note A vs. B (counting instances with no errors as 0 DALY loss) we will define an indicator that is 1 if the SOAP Note has the better treatment plan (lower DALY loss). This indicator is 0 for both notes if the notes are judged to be the same according to this question, asked after the MO rated Note A:

> Are there any *meaningful* differences in the treatment plans of SOAP Note A and B?

Only if the answer is "yes" or "unsure", the review of SOAP Note B proceeds.

3.5 Treatment Misallocation Based on Objective Conditions

All at-risk patients in our study will receive malaria, anemia and UTI screening. We define eligibility for testing in Appendix A and the treatment guidelines for each condition in Appendix B.

We will construct a dataset with one observation for each (patient, screening test, note), up to six per patient. We exclude (patient, test) data pairs where the test was not conducted due to our demographic criteria (only women/girls 7 years of age or older are tested for UTI, only adults are tested for anemia). We include (patient,test) combinations where our symptom screen determines the patient is eligible but does not need to be tested and label them as zeros.

We construct an indicator of treatment misallocation that records whether a patient was incorrectly treated for a condition based on the test result or lack of symptoms. The variable is coded as 1 if the patient tested positive and either received inappropriate or no treatment according to the guidelines above. It is also coded as 1 if the patient tested negative or was not tested based on the symptom screen but received treatment for the condition. The variable is only coded as zero if the patient tested negative and was correctly not treated for the corresponding condition, or if they tested positive and received the correct treatment.

Note that some treatments can be given for other conditions or without a screening test, such as antibiotics or iron supplements. In cases where the test was negative but a relevant treatment was received, an MD will conduct a blinded review and code the observation as either a match or mismatch. We will use the "clinical indication," a diagnosis code that is recorded for each individual prescription drug, as auxiliary data for this exercise.

In cases where the provider created a treatment plan that conditions on the result of medical testing we will interpret this plan based on the actual test results. For example, if the CHEW recommends a malaria test and the plan contains the correct malaria treatment conditional on a positive test result and no malaria treatment when the test is negative, the indicator will equal 0 for both a positive and a negative test result.

4 Secondary Outcomes

Our study will also estimate effects on a range of secondary outcomes. Unless otherwise indicated we will use the same analysis approach as above.

4.1 Alignment of Diagnosis and Treatment Between CHEWs and MOs

Separately from the MO's subjective evaluation of the CHEW notes, we will assess to what degree the CHEW's and MO's conditional notes agree with each other.

For each medication in the CHEW's treatment plan, there is a "clinical indication" (the diagnosis associated with the drug) along with an indicator that specifies if a given prescription is conditional on a medical test result. We will assess three indicators of a match:

- any match of the contents of the "clinical indication" field across medications;
- any match of the contents of the "medication" field across indications, including whether the medication is conditional on a test or not;
- a match of both medication and indication (and test conditionality).

Since CHEWs often use less specific diagnosis categories, and medications can be given as different formulations or prescribed with a brand name vs. a generic names but with the same medical properties, we will sort the diagnoses of MOs and CHEWs by ICD10 code, along with the associated treatment choices (but without indicating the SOAP note type that contained the information) and ask an MD to create families of diagnoses and associated groups of medications that are appropriate for the family of diagnoses and provide the same or similar treatment. Note that the same treatment may be associated with multiple diagnosis families (e.g. a broad spectrum antibiotic). For the first two items above, we will code as a match if the diagnoses (treatments) appear in the same family of diagnoses (treatments). For the last item, we will code as a match only if the diagnoses appear in the same family and the treatments appear in the same associated treatment family.

4.2 Objective Misallocation of Treatment – Alternative Specifications

As auxiliary outcomes, we will run the same treatment misallocation regression as in Section 3.5 separately for each possible test (with only patient fixed effects).

In addition, we will assess the contribution of different types of non-compliance with test results by coding patient-test-note level indicators for

- Misallocation due to overprescription: a condition is treated that the patient is confirmed not to have
- Misallocation due to undertreatment: a condition the patient is confirmed to have is not treated
- Misallocation due to incorrect dosing or drug choice: a condition the patient has is treated but the dosing or medication chosen is inappropriate.

Note that the misallocation indicator in Section 3.5 is the union of all three of these cases. We will test each of these three as dependent variables in the same regression as the misallocation indicator.

4.3 Relationship of QALY Loss to Severity of Patient Condition

In patients with only mild illnesses, the scope for QALY loss from mistakes may be limited relative to patients with more severe illnesses.

To assess this, we will regress QALY loss on indicators for mild, moderate, and severe illnesses (as assessed by the MO) each interacted with the assisted note indicator, controlling for patient fixed effects. We will show the results graphically.

4.4 Medical Testing Decisions

We will assess the potential misallocation of medical testing in two ways. First, we will measure it based on the screening test outcome. For each test type, we will create an indicator that is coded as 1 if the CHEW recommends conducting a test that turns out to be negative, and a second indicator that is 1 if the CHEW neglects to request a test that turns out to be positive.

Second, we will also construct a (patient,test,note) level indicator that measures whether the CHEW and MO requested the same or a comparable medical test (e.g. the CHEW requested a malaria RDT whereas the MO requested a malaria bloodsmear).

Finally, we will combine these indicators and count as a mismatch if and only if either: i) a test was not requested by the CHEW but was positive, or ii) the test was requested by the cHEW but the result was negative and no equivalent test was ordered by the MO. This accounts for the possibility that good clinicians might sometimes order tests which are more likely to be negative "just in case".

4.5 Subjective MO Review Outcomes

We will report treatment effects on all MO evaluation measures listed in section 2.3.1.

4.6 Average and Distribution of DALY lost

We use our DALY measure of harm indirectly in several primary outcomes (probability of error and severe error, which note is the better note). As a secondary outcome, we will measure the effect of LLM assistance on average DALY (this is not our primary outcome because we are concerned about power being limited due to a few outliers).

Additionally, we will show the full distribution of DALY ratings for the assisted and unassisted notes.

4.7 3rd Party MD Review of CHEW notes and MO notes

We will reassess our primary outcomes using the MD DALY ratings and subjective harm measures (section 2.5).

4.8 MO Review Only for Subset Approved by MDs

On the one hand, the reviewing MOs have the best information about the patient, but on the other, they may also make medical errors. Therefore we will also consider replicating our primary outcomes looking only at patients for whom MDs do not identify a serious error in the diagnosis or treatment proposed in the "final" MO note (see section 2.5). We will only drop patients for this secondary outcome if at least two MDs agree on the error identified in the MO's note.

4.9 Triage Decision

One question is whether LLM assistance leads CHEWs to make better triage decisions. We will conduct two analyses to assess this.

First, for each (patient, note), we will construct an indicator for whether the CHEW triage decision (an intent to triage indicated in the SOAP note) and the MO suggested triage decision align. We will then regress this on whether the note is assisted, controlling for patient fixed effects.

Second, we will reconsider our primary outcomes excluding patients for whom the MO indicates that triage is recommended. This checks whether CHEWs specifically improve care for patients they should be treating directly.

5 Validation and Auxiliary Analysis

5.1 Blinding

We will conduct checks to control for the possibility that imperfect blinding to the intervention (i.e. MO knowledge whether a patient note was "LLM assisted" or not) is impacting our results.

- We measure each MO's beliefs about whether a given patient note was LLM-assisted with a survey question. We will analyze whether MO beliefs about whether a note was LLM assisted are predicted by true LLM assistance status.
- As a robustness check, we will control for MO beliefs about LLM assistance and for any order effects (i.e. identity of note A vs. B) in the main analysis.
- Ex post, if we find that MOs can reliably predict whether a note was LLM assisted, we will seek to identify whether there are specific outcome-irrelevant features of notes which aid in this prediction, and may run additional robustness checks on a subset of patients where this feature was absent, so blinding is better maintained.

Bias in the difference in evaluation between notes A and B can arise if the MO (i) correctly guesses which note is assisted by the LLM, and (ii) systematically rates notes more highly if they are assisted by the LLM, but *not* due to higher quality of care/true lower harm to the patient.

For the first note that each MO sees (note A), we elicit a "raw" evaluation score prior to seeing note B. This score is then amended into an "adjusted" final evaluation (the reason for the update is the potential for learning about note A, given the cognitively demanding task of the MO: for example, only when shown the highlighted differences between the two notes may the MO detect an error in note A.) In an ideal world, the "adjusted" rating simply reflects the MO's evaluation given full information. This is the measure we use in most of our analysis. However, the "raw" score provides an opportunity for another blinding check, under the assumption that blinding is more problematic for note B than note A.

If this is the case, a cleaner but noisier estimate of our main effect can be obtained by using an across-patient comparison of the "raw" scores of assisted and unassisted SOAP notes A. This regression will be identical to our baseline specification, but omitting the patient fixed effects.

5.2 Data Quality Indicators

Quality of MOs. We will construct objective treatment and testing decision quality indicators for MOs in an analogous way to the measures in sections 3.5 and 4.4 based on the screening test results.

We will also report the share of cases in which MOs update their own treatment plan after reviewing SOAP Note A, SOAP Note B, and the screening test results.

Finally, we will report the assessment of error in the MO's SOAP note by the MDs.

Quality of DALY measures. We will verify the direction of DALY rating change between Note A and B with the response to the question

> All things considered, do you prefer the treatment plan in SOAP Note A or in SOAP Note B?

As a secondary outcome, we will make a scatterplot of QALY loss against whether the MO classifies a patient as mild, moderate, severe, pooling both notes. What we expect to see is that the scope for QALY losses generally occur in the most severe patients.

More generally, we will analyze the response of the DALY ratings of MOs and MDs to subjective and objective indicators of error: indicators for deviations from the MO's SOAP and types of harm incurred (section 2.3.1) and objective treatment misallocation (3.5).

We will also analyze the agreement between MO and MD DALY ratings relative to the agreement on the various harm measures between the three MDs.

5.3 Other Analyses

CHEW Feedback We will report the subjective feedback from CHEWs about the use of the LLM (section 2.4).

We will also report the MD's assessment of the LLM feedback and the degree to which CHEWs adhered to LLM recommendations (section 2.5).

Time trends Finally, we will assess whether CHEW notes appear observationally to change over time in ways that lead to closer ratings between assisted and unassisted notes. We will do this overall, and for specific discrepancies identified (e.g. "undertesting of malaria in the unassisted notes").

5.4 Updates 4/29/2025

We have several updates to the pre-analysis plan – these updates are made prior to analyzing the data by treatment status.

On 2/25, we discovered a randomization error in the order of the notes as presented to the MO (the unassisted note was always presented first, rather than in random order). Due to this error, we are throwing out all observations from prior to 2/25 in any analyses that use MO generated outcomes. Analyses that rely on outcomes not generated by the MO (e.g. objective test data) will still use all data, including the data gathered prior to 2/25.

As a result of these changes, we are extending the timeline of the initial study. The current proposed end date is May 30th.

A Criteria for Screening Tests

We apply the following demographic and symptom screening criteria for the medical screening tests.

Malaria RDT: All demographics are eligible. Malaria testing is offered if the patient reports a fever in the last 24 hours.

Anemia/PCV: All patients 18 years and older.

The test is offered for patients reporting feeling tired or with low energy for 4 weeks or more, feeling like their heart is racing for 4 weeks or more, signs of bleeding for 4 weeks or more (vaginal bleeding, blood in stool or in urine), feeling dizzy, difficulty breathing, currently pregnant, and delivered a baby within the past six weeks.

Urine dipstick analysis: Only female patients 7 years or older.

Urine dipstick analysis is offered for patients reporting pain while urinating, blood in urine, frequent urge to urinate, abnormal discharge from the genital area, and burning feeling while urinating.

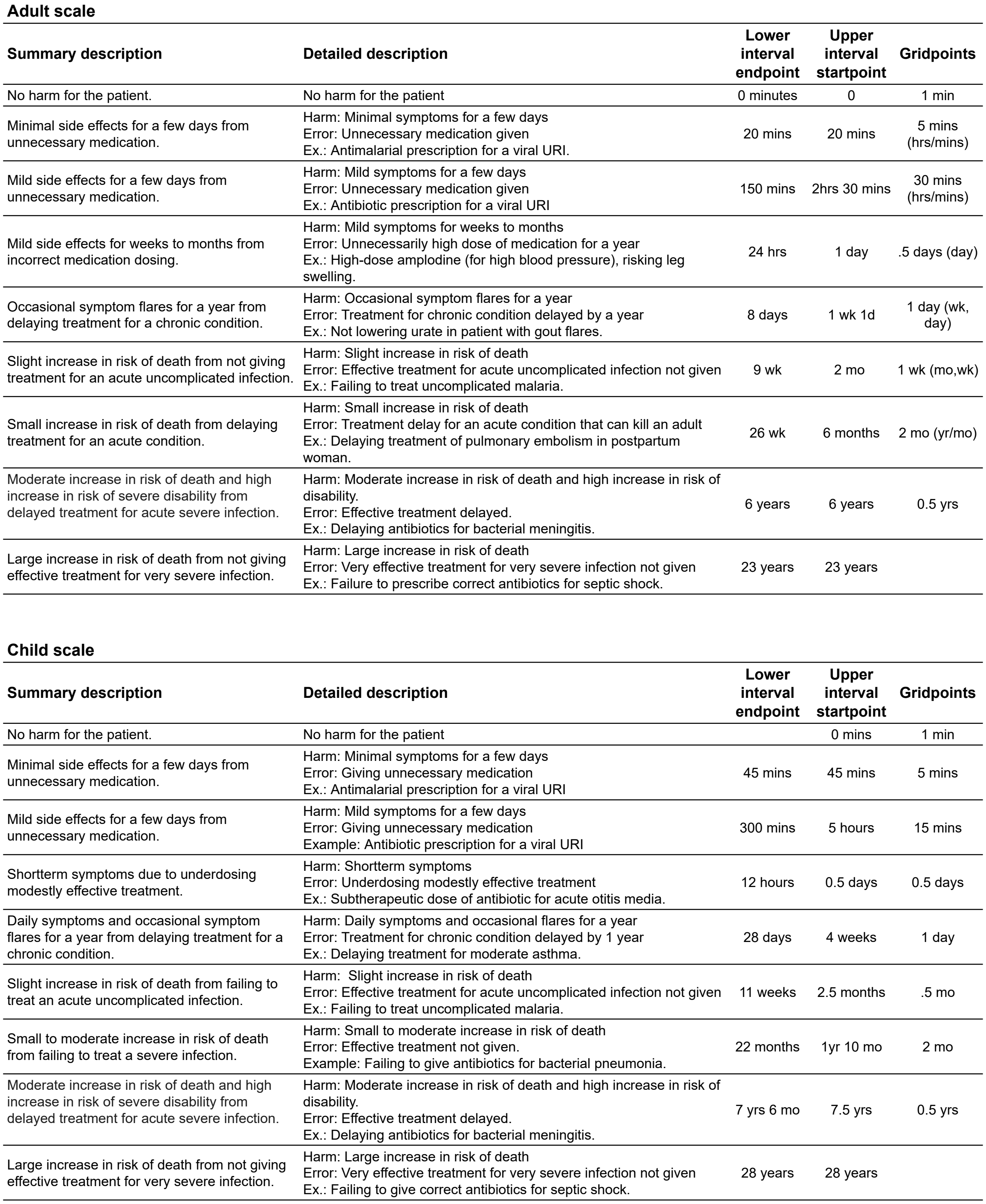
B Treatment guidelines for malaria, anemia and UTI

- For malaria, conditional on testing positive (and meeting the symptom screen), the essential treatment guidelines are:
  Adults: Artemether-Lumefantrine 80 mg/480 mg by mouth twice daily for 3 days
  Children 5-14 Kg: Artemether-Lumefantrine 20 mg/120 mg by mouth twice daily for 3 days
  Children 15-24 Kg: Artemether-Lumefantrine 40 mg/240 mg by mouth twice daily for 3 days
  Children 25-34 Kg: Artemether-Lumefantrine 60 mg/360 mg by mouth twice daily for 3 days
  Severe malaria: Artemether 3.2 mg/Kg IM loading dose once followed by 1.6 mg/kg IM maintenance dose once daily for 3 days
- For anemia, conditional on testing positive (and meeting the symptom screen), the essential treatment guidelines are:
  Adults: Ferrous Sulfate 200 mg by mouth once daily for one month
  Children: Ferrous Sulfate 125mg/ml syrup 3-6mg/kg/day for one month
  Pregnant women: Ferrous Sulfate 200 mg by mouth twice or thrice daily for one month
  Severe anemia: Refer for transfusion
- For UTI (uncomplicated), conditional on testing positive (and meeting the symptom screen), the essential treatment guidelines are:
  Adults: first-line treatment Nitrofurantoin 100 mg by mouth twice daily for 7 days
  OR
  second-line treatment Cotrimoxazole 960 mg by mouth twice daily for 5 days
  Children: Cotrimoxazole 8-10 mg/Kg/day by mouth divided twice daily for 5 days

Pregnant women in third trimester only:
Amoxicillin 500 mg by mouth every 8 hours for 7 days
OR
Amoxicillin/Clavulanate 500/125 mg by mouth twice daily for 3-7 days
OR
Cephalexin 500 mg by mouth twice daily for 7 days

C Benchmarks for Healthy Time Lost

The DALY scale has 9 benchmarks from 0 minutes to 23 years (28 years for children), generating 8 “bins” each. In detail, the benchmarks are shown in Table 1.



Adult scale

Summary description	Detailed description	Lower interval endpoint	Upper interval startpoint	Gridpoints
No harm for the patient.	No harm for the patient	0 minutes	0	1 min
Minimal side effects for a few days from unnecessary medication.	Harm: Minimal symptoms for a few days Error: Unnecessary medication given Ex.: Antimalarial prescription for a viral URI.	20 mins	20 mins	5 mins (hrs/mins)
Mild side effects for a few days from unnecessary medication.	Harm: Mild symptoms for a few days Error: Unnecessary medication given Ex.: Antibiotic prescription for a viral URI	150 mins	2hrs 30 mins	30 mins (hrs/mins)
Mild side effects for weeks to months from incorrect medication dosing.	Harm: Mild symptoms for weeks to months Error: Unnecessarily high dose of medication for a year Ex.: High-dose amplodine (for high blood pressure), risking leg swelling.	24 hrs	1 day	.5 days (day)
Occasional symptom flares for a year from delaying treatment for a chronic condition.	Harm: Occasional symptom flares for a year Error: Treatment for chronic condition delayed by a year Ex.: Not lowering urate in patient with gout flares.	8 days	1 wk 1d	1 day (wk, day)
Slight increase in risk of death from not giving treatment for an acute uncomplicated infection.	Harm: Slight increase in risk of death Error: Effective treatment for acute uncomplicated infection not given Ex.: Failing to treat uncomplicated malaria.	9 wk	2 mo	1 wk (mo,wk)
Small increase in risk of death from delaying treatment for an acute condition.	Harm: Small increase in risk of death Error: Treatment delay for an acute condition that can kill an adult Ex.: Delaying treatment of pulmonary embolism in postpartum woman.	26 wk	6 months	2 mo (yr/mo)
Moderate increase in risk of death and high increase in risk of severe disability from delayed treatment for acute severe infection.	Harm: Moderate increase in risk of death and high increase in risk of disability. Error: Effective treatment delayed. Ex.: Delaying antibiotics for bacterial meningitis.	6 years	6 years	0.5 yrs
Large increase in risk of death from not giving effective treatment for very severe infection.	Harm: Large increase in risk of death Error: Very effective treatment for very severe infection not given Ex.: Failure to prescribe correct antibiotics for septic shock.	23 years	23 years	

Child scale

Summary description	Detailed description	Lower interval endpoint	Upper interval startpoint	Gridpoints
No harm for the patient.	No harm for the patient		0 mins	1 min
Minimal side effects for a few days from unnecessary medication.	Harm: Minimal symptoms for a few days Error: Giving unnecessary medication Ex.: Antimalarial prescription for a viral URI	45 mins	45 mins	5 mins
Mild side effects for a few days from unnecessary medication.	Harm: Mild symptoms for a few days Error: Giving unnecessary medication Example: Antibiotic prescription for a viral URI	300 mins	5 hours	15 mins
Shortterm symptoms due to underdosing modestly effective treatment.	Harm: Shortterm symptoms Error: Underdosing modestly effective treatment Ex.: Subtherapeutic dose of antibiotic for acute otitis media.	12 hours	0.5 days	0.5 days
Daily symptoms and occasional symptom flares for a year from delaying treatment for a chronic condition.	Harm: Daily symptoms and occasional flares for a year Error: Treatment for chronic condition delayed by 1 year Ex.: Delaying treatment for moderate asthma.	28 days	4 weeks	1 day
Slight increase in risk of death from failing to treat an acute uncomplicated infection.	Harm: Slight increase in risk of death Error: Effective treatment for acute uncomplicated infection not given Ex.: Failing to treat uncomplicated malaria.	11 weeks	2.5 months	.5 mo
Small to moderate increase in risk of death from failing to treat a severe infection.	Harm: Small to moderate increase in risk of death Error: Effective treatment not given. Example: Failing to give antibiotics for bacterial pneumonia.	22 months	1yr 10 mo	2 mo
Moderate increase in risk of death and high increase in risk of severe disability from delayed treatment for acute severe infection.	Harm: Moderate increase in risk of death and high increase in risk of disability. Error: Effective treatment delayed. Ex.: Delaying antibiotics for bacterial meningitis.	7 yrs 6 mo	7.5 yrs	0.5 yrs
Large increase in risk of death from not giving effective treatment for very severe infection.	Harm: Large increase in risk of death Error: Very effective treatment for very severe infection not given Ex.: Failing to give correct antibiotics for septic shock.	28 years	28 years	

Table 1: Adult and child scale harm benchmarks.