Can LLMs Improve Healthcare Delivery?

Evidence from Physician Review and Objective Testing*

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

November 1, 2025

Abstract

We deployed a large language model (LLM) decision support using GPT4 for health workers at two outpatient clinics in Nigeria. For each patient, health workers drafted care plans that were optionally revised after LLM feedback. We compared unassisted and assisted plans using (i) blinded randomized assessments by on-site physicians who assessed and treated the same patients and (ii) results from laboratory tests for common conditions. Academic physicians performed blinded retrospective reviews of a subset of notes. Providers reported high satisfaction with LLM feedback, and retrospective academic reviewers rated LLM-assisted plans more favorably. However, on-site physicians observed little to no improvement in diagnostic alignment or treatment decisions. Objective testing showed mixed effects of LLM-assistance, with reduced overtesting for malaria but increased overtesting for urinary tract infection and anemia. This highlights a gap between chart-based reviews and real-world clinical relevance that may be especially important in evaluating the effectiveness of LLM-based interventions.

*The authors gratefully acknowledge the research assistance of Eduardo Arancón Simal, Nina (Zi Wei) Low, Tedi Totojani, Grady McPeak, Kylie Kim, Owen Hart, and Lisa Timm. Fatimah Howeidy, Attat Yohannes, Adham Hallal for project support and Ogochukwu Maduabum for modifications to the electronic medical record. Drs. Fatimah Tsiga-Ahmed, Rabiu Jalo and Aminatu Kwaku reviewed the patient notes. This research was funded by the Gates Foundation Grand Challenges Initiative (INV-081875) and a World Bank Research Support Budget grant. Dr. Schwartz received additional support from the University of Pennsylvania Matt Slap Research Award and the Department of Veterans Affairs Career Development Award initiative (IK2HX003634). This trial was pre-registered in the AEA RCT registry under trial number AEARCTR-0015226 and approved by the Yale University IRB (protocol ID 2000035990) and the Kano State health research ethics committee. The findings, interpretations, and conclusions expressed in this article are entirely those of the authors and do not represent the views of the World Bank and its affiliated organizations, of the Executive Directors of the World Bank or the governments they represent, or of the U.S. government or the Department of Veterans Affairs.

[†]Yale School of Management, Yale University and the National Bureau of Economic Research

[‡]Department of Computer Science, George Washington University

[§]eHealth Africa EHA Clinics Nigeria and Department of Emergency Medicine, George Washington University

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

[‖]Department of Medical Ethics and Health Policy and Division of General Internal Medicine, University of Pennsylvania Perelman School of Medicine, and Crescenz VA Medical Center

1 Introduction

Large language models (LLMs) have the potential to improve health provider decision-making without requiring substantial infrastructure to deploy (Pressman et al., 2024). LLMs have shown diagnostic performance comparable to general physicians on written tests and in simulated patient encounters such as vignettes or model patients (Takita et al., 2025; Huang et al., 2024; Tu et al., 2025). A smaller number of studies have tested LLMs with real patient information, or LLM decision support in realistic clinical environment (Goh et al., 2024; Cadamuro et al., 2023; Gaber et al., 2025; Hager et al., 2024). Evaluation of LLM performance is typically retrospective, comparing the LLM's conclusions with the decisions of the treating physician (Shmilovitch et al., 2025; Yıldırım et al., 2025). We are only aware of one peer-reviewed and one unpublished study that implement forms of LLM support in a live clinic setting (Korom et al., 2025; Wan et al., 2024). In both, the evaluation of quality of care with and without LLM is done via retrospective review, which by its nature focuses on the internal consistency of the patient record and cannot account for information that the treating provider may have missed or decided not to record.

We build on the existing literature by evaluating LLM-assisted care using (blinded) review by physicians who independently assessed the same patient, as well as objective medical test results. We use both approaches (and others) to evaluate a prototype intervention in two outpatient clinics in Kano, Nigeria, in which an LLM gives providers an instant "second opinion" on their care plans. This intervention models lightweight approaches that could be broadly deployed, and mirrors the informal ways providers are already consulting LLMs (Blease et al., 2024). Health workers (here Community Health Extension Workers, or CHEWs) drafted initial, unassisted care plans and then revised them after receiving LLM feedback, producing paired unassisted and assisted notes. All patients were subsequently examined by on-site physicians (Medical Officers, or MOs), who developed their own care plans, treated the patients (to avoid any negative impacts of LLM use), and then evaluated both health worker plans given their direct, personal knowledge of the patient in a blinded manner.

Our design allows us to compare unassisted and LLM-assisted care plans using four complementary lenses: (1) subjective ratings from on-site physicians who saw the same patients; (2) concordance between the health worker and on-site physician care plans; (3) agreement of testing and treatment decisions with screening based on laboratory tests for the three most commonly tested conditions, malaria, urinary tract infection, and anemia; and (4) blinded retrospective assessment of a subsample of patients by academic attending physicians (MDs) who reviewed the written notes produced by all providers who saw the patient.

In all four domains, health workers' care plans *prior* to LLM assistance show substantial deficits. We find that LLM assistance led health workers to meaningfully revise diagnoses (41% of notes), test ordering

(33% of notes), and prescribing decisions (54% of notes). The health workers themselves overwhelmingly reported that they found the LLM feedback helpful. Attending physicians also rated LLM-assisted notes more favorably in retrospective review. However, on-site physicians who had examined the same patients observed little or no improvement in diagnostic alignment or treatment decisions, and assisted plans did not resemble physician plans more closely. Comparison with screening tests revealed mixed effects: the LLM led the health workers to remove negative tests for malaria, but added them for urinary tract infection and anemia, and did not encourage testing for cases later found to be positive for the condition. There was subsequently no improvement in the appropriateness of treatment given underlying conditions.

When analyzing the LLM feedback itself, we find that the LLM makes 3.75 recommendations per patient, and of those, 53% would make the health worker care plan more similar to the physician care plan. Despite following only about one third of all recommendations, health workers were only slightly more likely than chance to follow recommendations that improved physician concordance; in other words, the "human in the loop" does not serve to filter out the less helpful AI advice. When presented with the symptoms noted by the medical officer, LLM feedback improves, but only enough to close a small fraction of the gap between health worker and physician plans.

These findings reveal a notable divergence: LLM feedback improves plan documentation according to retrospective academic reviewers, yet fails to produce consistent improvements according to physicians with firsthand knowledge of the patient, or according to laboratory results. These findings suggest a potential disconnect between retrospective evaluation and locally observed clinical performance.

2 Study Design

2.1 Experimental Design

Setting. This study was conducted in two private clinics in the city of Kano in Kano State, operated by EHA Clinics Nigeria. One clinic is a private community clinic staffed primarily by Community Health Extension Workers (CHEW), the other is a full-service private clinic where all patients are seen by Medical Officers (MO). Community Health Extension Workers in Nigeria undergo a three-year licensed diploma program. Their training focuses on community outreach, basic preventive and curative care, maternal health, and referrals to higher-level providers (Ajisegiri et al., 2023; Dotimi et al., 2024). While they generally work under supervision of a nurse or MO, CHEWs may be de facto the primary provider in a public clinic or health outpost in rural or underserved areas (Okoroafor et al., 2022). In contrast, Medical Officers have completed training that includes at minimum a four-year medical bachelor degree and internship; their scope of practice is that of a full physician. In a large randomized trial, Okeke (2023) showed that simply deploying one additional MO to a public clinic in Nigeria significantly improved quality of care and patient

outcomes.

EHA Clinics provides training for CHEWs and MOs in their clinical protocols, which include rules to prevent the overuse of antimalarials and antibiotics. EHA Clinics also uses a custom electronic medical record (EMR) system to record patient information and track medical test orders and prescriptions. For this study, the information recorded in the EMR was used in the background to collate patient information collected by the provider into a SOAP note ((S)ubjective-(O)bjective-(A)ssessment-(P)lan). The main clinic has a pharmacy and an in-house laboratory for routine laboratory tests, which is also used by the community clinic.

Patient flow. Patients were enrolled and informed consent was obtained before being accompanied through the steps of the study by dedicated research staff. In our study, a patient's first clinician encounter was with a CHEW, who conducted the patient consultation and prepared a "conditional" SOAP note that specified the provisional diagnosis and prescriptions conditional on medical test results. CHEWs at EHA Clinics are able to order and interpret results of some basic lab tests.

The health worker submitted the "unassisted" SOAP note to the LLM indirectly through the EMR for feedback. A sample of a CHEW SOAP note, the LLM prompt, and sample LLM feedback are shown in Appendix A. Upon receiving the feedback from the LLM in the EMR, the health worker prepared an updated "assisted" (conditional) SOAP note.

It is important to note that health workers did not actually order any medical tests or diagnose and treat patients in this study for ethical reasons. They were trained to conduct a patient consultation that is as close to their daily practice as possible, but to refrain from any diagnosis or remarks on the patient's health to the patient. They were also trained to view the LLM feedback as potentially flawed and to only adopt advice that they agreed with.

Patients next saw the physician (MO) on site who conducted an independent patient evaluation. Patients were advised upon enrollment to treat the two consultations as independent and repeat information to the physician even if they had shared it with the health worker previously. The physician prepared their own conditional SOAP note and ordered any laboratory testing needed to diagnose the patient. Test results that were available immediately were incorporated into the physician's patient record and included in the discharge SOAP note. We interpret this discharge note as illustrating a typical care plan that would have been recommended had the patient visited a physician at these clinics outside of the study.

At this stage of the study, patients were re-approached by research staff who offered additional laboratory (point of care) testing for three common conditions – malaria, anemia, and urinary tract infection – based on a set of demographic criteria and symptom checks (see below and Appendix C.2 for details). The purpose

Table 1: Study Samples: All Patients and Fully Randomized Sample.

	Full sample Mean (SD)	Blinded sample Mean (SD)
Share Female	0.64 (0.48)	0.66 (0.47)
Share Age 0-4	0.06 (0.24)	0.07 (0.25)
Share Age 5-14	0.11 (0.31)	0.13 (0.33)
Share Age 15-45	0.72 (0.45)	0.71 (0.45)
Share community clinic	0.49 (0.50)	0.48 (0.50)
Share health worker would refer to physician	0.77 (0.42)	0.77 (0.42)
Share physician rated seriously ill	0.33 (0.47)	0.24 (0.42)
Share unassisted note is note A	0.62 (0.49)	0.49 (0.50)
N (Patients)	660	491
N (Health workers)	21	21
N (Physicians)	8	6

This table reports demographic and clinical characteristics for two samples. The full sample includes all patient visits from January 30 to May 26, 2025, representing the complete study period. The fully randomized sample restricts to visits from February 25 to May 26, 2025, when the order of health worker notes presented to physicians was correctly randomized. All statistics report means with standard deviations in parentheses. "Share unassisted note is note A" indicates the proportion of cases where the unassisted health worker note was randomly assigned to be presented first to the physician.

of this testing step was to collect additional objective data beyond the results of physician-ordered tests so that the study could better evaluate the quality of health worker care plans. Patients were unaware up to this point that they could be offered these three tests, and the test results were not made available at the time of the clinic visit. Rather, blood and urine samples were collected and patients were discharged, with any supplemental testing conducted in the course of the next 1-2 days. This design was chosen to avoid contamination of the physician's testing decisions. As the final step, typically in the evening or on the next day, the physician reviewed the full patient record including any additional test results from the research staff, updated their own SOAP note (and flagged patients with actionable findings for follow-up), and then carried out a blinded evaluation of the unassisted and assisted health worker notes, labeled randomly SOAP A or SOAP B.

Sample. Study funding was projected to cover approximately 500 patients in the main sample. After extensive piloting, the LLM-assistance intervention was initially rolled out on January 30, 2025, and study patient encounters were conducted through May 26. Due to an error in the EMR system, randomization of care plan labels was initially not implemented correctly, and physicians always reviewed the LLM-unassisted note first (labeled SOAP A). This error was corrected by February 25. Thus, while we include the full sample of 660 patient visits in analyses of objective outcome measures (i.e. direct comparisons of care plans, or comparison of care plans to medical test results, etc..), we use only the subsample of 491 patient encounters after February 25 in any analyses of local physicians' evaluation of care worker notes (this was preregistered

at the time when the error was discovered).

Table 1 gives an overview of the two samples. This population extends beyond the intended scope of a health worker's practice, but reflects patients these health workers may treat in understaffed public clinics. The demographic profile is broad, with a significant share of adolescents and adult men who are not a focus of community health interventions; 24% of patients (33% in the larger sample) are judged by the physician to be seriously ill; and in 77% of cases the health worker states that outside of the study they would refer this patient to a physician. It is worth noting that EHA Clinics have higher staffing levels than typical public clinics in Nigeria. For example, a physician is available through telemedicine at all times from the community clinic. Physicians are also the main providers at the full-service clinic. These factors likely influences patient composition.

The second-to-last panel verifies the balance in randomized assignment in the sample after February 25. The intervention was carried out by 21 health workers who rotated in and out of the study and each saw between 25 and 40 patients; and there were a total of 8 reviewing physicians.

2.2 Data Sources

We use two main sources of data; patient EMR data and survey data from health workers (CHEWs), physicians (MOs), and third-party attending physicians (MDs).

2.2.1 Patient encounter data in the EMR

All patient data entered by the intake nurse, the health worker and physician, and the lab technician are saved in the EMR. As described above, a total of five SOAP notes were created for each patient, the LLM-unassisted and LLM-assisted notes by the health worker, the physician's conditional note (prior to obtaining any lab test results), the discharge note (after seeing the results of lab tests ordered by the physician), and the final physician note which includes any additional test results for malaria, UTI and anemia collected by the study staff. In addition to the text files containing the structured SOAP notes, we collected the values of the individual EMR entry fields within the SOAP notes, such as test results, diagnoses, prescriptions, vital signs, and basic demographic information.

The subjective information in the SOAP notes includes the self-reported chief complaint, history of present illness, any past medical history, current medications, allergies, and relevant family or social history. Objective data include vital signs, physical examination findings, and lab test results. The assessment and plan contain the diagnosis and differential diagnosis, testing or treatment the clinician is pursuing, including non-prescription advice (e.g., "patient counseled to drink plenty of fluids"), referrals to other clinicians or sites, or warning signs that would warrant another clinic visit.

For each recommended medication order entered into the EMR, the provider must supply detailed dosing

information along with a clinical indication, i.e., the related diagnosis. For this study, clinicians could also optionally choose to attach a condition to the prescription based on medical tests ordered. For example, the health worker might use the lab testing section of the EMR to request a malaria test, and then in the prescription section choose a malaria drug that is only given on condition of a positive test result. In standard practice, this conditioning would not be explicitly recorded, but rather a prescription would be entered only if the malaria diagnosis was confirmed by a positive test.

Medical test results. For all medical tests conducted, we summarized the physician's interpretation of the test as "normal" or "abnormal". As described earlier, we conducted broad testing among our patient sample for three common conditions – malaria, urinary tract infection (UTI), and anemia. Research staff checked a broad list of symptoms intended to identify anyone who should plausibly be tested for one of these three conditions. Patients were then consented for any tests that were not already ordered by the physician. We restricted testing to the most relevant demographic groups; all symptomatic patients for malaria, symptomatic female patients age 7 or older for UTI, and symptomatic patients over 18 for anemia. These test results provide "ground truth" data on the patient's true disease status. Appendix C.2 provides further detail.

2.2.2 Physician and MD Evaluation of Care Plans

Key measures of the quality of the health worker's care plans are blinded evaluations by 1) the physician who treated the same patient and 2) a panel of three attending Medical Doctors (MDs). These MDs, who work at Bayero University and Aminu Kano Teaching Hospital in Kano, have completed four years of post-graduate training in family and community medicine. They were provided retrospective access to both health worker notes (blinded to LLM-assistance status) and the physician note.

The evaluation was done via a detailed survey in conjunction with (for the physicians) a custom review mask implemented directly in the EMR. The physicians first reviewed any lab test results ordered by the research staff (malaria, UTI, anemia), and updated their own SOAP note to a final SOAP note, and then began review with SOAP note A (randomly selected to be either the unassisted or assisted health worker note). At this stage, they were unable to see SOAP note B.

The physician 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 physician first takes stock of qualitative differences between the health worker's plan and their own plan, then assesses the type of harms that would arise for the patient from any errors identified, and finally rates the SOAP note on a scale for "healthy time lost".

Measuring Healthy Time Lost in DALY. We constructed an anchored scale to measure how much healthy time would have been lost if a health worker plan had been implemented instead of the physician's

plan. The scale translates specific types of error into 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.

The anchored 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. DALY benchmarks were derived from previously published DALY or QALY estimates from the medical research literature (Salomon et al., 2015), 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. Because the DALY consequences of death or disability are higher and different clinical mistakes are relevant for children, we constructed a separate DALY scale for the evaluation of patients who are children (under 18) and patients who are adults.

MOs first chose one interval of the harm scale, and then on the next screen saw a detailed description of the two scale end-points and used a slider to choose the estimated level of harm between those two points. The MDs used the same scale for their evaluation. A detailed description is in Appendix C.4.

2.2.3 Health Worker Survey Questions on the LLM Feedback

We asked health workers to give Yes/No answers to three pulse feedback questions after every patient seen; in addition, health workers responded to a more general feedback questionnaire when they rotated out of the study. These were intended to elicit whether they perceived LLM assistance as useful and accurate.

2.3 Empirical Specification

Our primary analyses obtain the effect of LLM assistance from the within-patient difference between the unassisted and assisted notes. Every outcome is measured for both SOAP notes, and in most cases we estimate

$$Y_{ik} = a_i + \beta k + cA_{ik} + \epsilon_{ik}$$

where i is the patient and $k = 1$ for the assisted note and zero otherwise, Y_{ik} is the outcome, and a_i is a patient fixed effect. For outcomes that involve a subjective rating by the physician, A_{ik} is an indicator for whether SOAP note k was shown to the physician first (to control for order effects). We cluster all regressions at the patient level.

Table 2: Assessment of the LLM feedback by health workers; pulse survey (after each patient) and exit survey (at the end of the health worker's study rotation).

	Mean (SD)	N
Panel A: Health Worker Note Changes from LLM Suggestions		
Fraction of notes with medical test change	0.329 (0.47)	660
Fraction of notes with Rx change	0.539 (0.50)	660
Fraction of notes with diagnostic code change	0.409 (0.49)	660
Panel B: Health Worker Pulse Survey Feedback		
Fraction of notes where health workers report 'LLM improves documentation'	0.952 (0.21)	642
Fraction of notes where health workers report 'LLM helps provide better care'	0.950 (0.22)	642
Fraction of notes where health workers report 'LLM made an error'	0.005 (0.07)	642
Panel C: Exit Survey Feedback		
LLM feedback is useful and justifies additional time	0.895 (0.32)	19
I feel more confident in my decisions with LLM support	0.947 (0.23)	19
I trust the LLM's advice	0.895 (0.32)	19
LLM has improved my patient record taking	1.000 (0.00)	18
I would like to use LLM in the future	1.000 (0.00)	19
LLM's advice is too vague	0.444 (0.51)	18
LLM's advice is too wordy and long	0.444 (0.51)	18
LLM is difficult to use	0.105 (0.32)	19
LLM contained errors	0.556 (0.51)	18

This table presents summary statistics on community health worker responses to AI-assistance with clinical documentation. Panel A reports the fraction of clinical notes where health workers modified medical tests, prescriptions, or diagnostic codes following LLM suggestions, using all assessments in the study. Panel B summarizes health worker pulse survey responses regarding LLM utility and accuracy; some observations are missing due to data loss. Panel C presents exit survey responses on overall LLM experience and satisfaction from health workers. Coded as 1 if health worker indicated agreement, 0 if disagree, and missing if health worker responded "Don't know/Dont want to answer."

3 Results

3.1 Health worker response to LLM feedback

Health workers made significant changes to their care plans in the assisted SOAP note and generally received LLM feedback positively. In paired comparisons of unassisted and assisted care plans, 33% of pairs differed in the medical test ordered, 41% differed in the clinical indications (diagnoses) recorded, and 54% differed in the prescriptions or drugs recommended (Panel A).

For 95% of patient encounters, the health worker reported that the LLM improved patient documentation and helped provide better patient care (Table 2, Panel B). In our exit surveys, 90-100% of health workers agreed with positive statements about the LLM, such as "I would like to use the LLM in the future" (Panel C). At the same time, 44% of the health workers reported that the LLM responses were too wordy or vague, and 56% agreed that the LLM sometimes made errors. Note that the health care workers rarely reported LLM errors during study encounters (0.5% of encounters), possibly because of the added documentation required to explain reported errors.

3.2 Physician and MD Evaluations of Unassisted and Assisted Care Plans

Based on the health worker's responses alone, one might think that the LLM served as a helpful nudge; health workers were able to discern good advice from bad and followed it selectively. However, the evaluation by the physicians who saw the same patients gives a more mixed picture of the value of the LLM feedback.

We begin by analyzing what errors physicians identified in the health worker notes. To do so, we use a set of questions on whether different care plan elements were incorrect.

Physician Evaluation of Errors in Treatment Plan Elements. Table 3 shows the effect of the LLM on identified errors. The outcomes in the first six columns are indicator variables for whether the physician thought that there were missing or incorrect/unnecessary diagnoses, medications, or care instructions. We also code the number of medical tests the health worker ordered that the physician judged to be "unlikely to be clinically useful."[1]

Physicians found many problems with the health workers' care plans, as shown by the average outcomes for unassisted notes. Approximately 40% of plans are missing key elements like diagnoses that should have been treated, or medications or care instructions patients should have received. In addition to these omissions, 19% of unassisted care plans include incorrect diagnoses, 21% include incorrect or unnecessary medications, and 4% include incorrect or unnecessary care instructions. There are 0.08 medical tests per patient that are not clinically useful.

We find no effect of LLM feedback on diagnoses, incorrect/unnecessary medications or incorrect/unnecessary care instructions, but LLM assistance did reduce the share of notes with missing medications or care instructions by 2.5 percentage points and 2.2 percentage points respectively. These effects are quantitatively small relative to the baseline error rate, but statistically significant. On the other hand, with LLM assistance, the number of tests judged to be not useful increased by 2.2 percentage points, or 29%.

Physician Evaluation of Patient Harm. We now turn to whether care plans with LLM-assistance were judged as more or less harmful to patients. Table 4 shows treatment effects of LLM-assistance on various measures of harm, as graded by the physician who evaluated and treated the same patient as the health care worker.

The first column outcome is a binary indicator denoting that the care plan was *not* completely appropriate for the patient. This indicator captures all errors regardless of the magnitude of expected harm to the patient. Column 2 captures whether the physician indicated that the health worker's treatment plan could cause

[1] We did not record whether the physician thought any medical tests were missing. CHEWs are authorized to request only a small number of tests that can be 1:1 compared with the MO, and we do so in section 3.4. The question on medical tests asked instead whether a given test was "necessary", "not necessary but clinically useful" or "unlikely to be clinically useful". This was based on findings from prior pilots that CHEWs tend to use laboratory testing more than MOs, and allows for the possibility that CHEWs need to use medical tests for diagnosis where MOs use their greater diagnostic skills.

Table 3: Physician Evaluation of Specific Treatment Plan Elements

	Diagnosis		Medication		Care Instructions		Medical Tests
	Missing	Incorrect/ Unnecessary	Missing	Incorrect/ Unnecessary	Missing	Incorrect/ Unnecessary	# Not Clinically Useful
Effect of assisted note	0.000	0.002	-0.025**	0.004	-0.022**	-0.002	0.022**
	(0.009)	(0.012)	(0.012)	(0.013)	(0.011)	(0.007)	(0.010)
Order effect	0.000	0.016	0.021*	0.013	0.000	-0.007	-0.009
	(0.009)	(0.012)	(0.012)	(0.013)	(0.011)	(0.007)	(0.010)
Avg. unassisted note	0.426***	0.186***	0.383***	0.208***	0.411***	0.045***	0.080***
	(0.006)	(0.009)	(0.009)	(0.008)	(0.008)	(0.004)	(0.007)
R-squared	.	0.004	0.015	0.003	0.009	0.002	0.013
N (Patients)	450	446	448	450	450	450	450

Notes: This table presents results from ordinary least squares regressions with patient fixed effects. The dependent variables in columns 1-6 are indicators for missing or incorrect/unnecessary items in the SOAP note. The outcome in column 7 is the number of medical tests in the SOAP note that the physician deems unlikely to be clinically useful. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

temporary or short-term harm to the patient, in the form of additional symptoms or missing usual daily activities. The outcome in column 3 is an indicator for whether there was a risk of long-term or permanent harm such as disability or death. Column 4 captures whether the physician judged the care plan to cause strictly less overall harm than the health worker's alternative care plan. This binary outcome was coded as zero in the 51% of cases when the physician judged the care plans as producing equal expected harms. Finally, columns 5 and 6 are constructed from the expected DALY loss estimates the physicians provided for each SOAP note using our two-step harm scale described earlier. Column 5 is an indicator for exceeding the 95th percentile of DALY loss among the unassisted notes, pooling child and adult scales (so the baseline share of the variable is by construction approximately 5%). Column 6 is an indicator for the care plan with the strictly lower DALY loss. This variable differs from the "Less harm" indicator because the DALY scale was collected via a visual scale interface with discrete selectable "notches" so that very small differences in harm could not be recorded. Asterisks denote the pre-specified primary outcomes.

While these results are directionally consistent with LLM assistance reducing patient harm, the effect magnitudes for most outcomes are not statistically significant. The "Less harm" and "Lower loss" variables show that in 14-26% of patients, the physician strictly prefers the *unassisted* over the assisted note, and the reverse is only true in a marginally higher share of patients; with many unassisted and assisted notes judged as roughly equivalent. The only meaningful change is the 1.1pp reduction in care plans exceeding the 95th percentile of DALY loss. This effect is significant at the 10% level. These findings suggest that LLM benefits may be concentrated among a small proportion of encounters with the potential for large errors. There appears to be little positive impact for the majority of patients. This result is striking because physicians state that 64% of the unassisted care plans to contain an error, implying ample room for improvement via care plan revisions (Column 1).

Table 4: Physician Assessment of Changes in Quality of Care.

	Qualitative error/harm				DALY loss		
	Any error?*	Short-term harm	Long-term harm	Less harm	Severe loss*	Lower loss*	DALY hours
Effect of assisted note	-0.009	-0.018	-0.009	0.028	-0.011*	0.036	-147.496
	(0.016)	(0.017)	(0.005)	(0.033)	(0.006)	(0.026)	(353.375)
Order effect	0.027*	0.018	-0.009	-0.072**	-0.002	-0.019	24.001
	(0.016)	(0.017)	(0.005)	(0.033)	(0.006)	(0.026)	(353.375)
Avg. unassisted note	0.642***	0.524***	0.086***	0.263***	0.054***	0.140***	1017.686***
	(0.012)	(0.012)	(0.004)	(0.023)	(0.004)	(0.018)	(37.968)
R-squared	0.01	0.00	0.01	0.01	0.01	0.01	0.00
N (Patients)	450	450	450	448	450	450	450

This table presents fixed effects regression estimates of the impact of AI-assisted clinical notes on various measures of medical assessment quality. The dependent variables include binary indicators for treatment errors and harm assessments (columns 1-4), as well as DALY (disability-adjusted life year) loss metrics (columns 5-6). The sample is restricted to assessments evaluated after February 25, when presentation order was randomized to control for order effect. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Since the estimates suggest sizable order effects in some of the variables, with a harsher judgment on the note randomly chosen to be presented first (SOAP A), we repeat the estimates including only SOAP A and comparing assisted and unassisted notes in the cross section, shown in Appendix Table B.1. The results vary somewhat across outcomes but do not fundamentally alter the findings in terms of the magnitude of quality of care improvements by the LLM. We also repeated this analysis with short-term and long-term harm assessed by another LLM (GPT-5) rather than by the treating physician; findings are similar, with small but insignificant harm reduction, albeit a much higher baseline rate of harm and no order effects (Appendix Table B.2). Finally, noting that physicians are somewhat better than chance at identifying the assisted SOAP note after seeing both notes (Appendix Table B.6), we assess whether issues with incomplete blinding could have caused bias in our results, for example because physicians oppose the use of LLMs in patient care. Physicians tended to rate a care plan more favorably when they believed the plan to be LLM-assisted. However, the estimated effects of LLM assistance remained nearly unchanged when regression models were modified to include a binary variable denoting the physician's guess regarding which note was LLM-assisted (Table B.3 in the Appendix).

MD Evaluation of Harm. To supplement the treating physicians' ratings, third-party MDs re-rated care plans for 59 patients in double review. For each study encounter, two MDs were initially presented with physician notes and all medical test results. MDs were then asked whether the physician made any serious errors; they reported such errors for only 2 encounters. They then evaluated the health worker notes in random order, in parallel fashion to the evaluations by the treating physician. The 59 patients were selected for MD review based on either a high loss rating from the physician in the unassisted note or a strong increase or decline in the loss rating. Our main interest is not generalizing from this (selected) subsample to the full sample, but understanding how MDs and physicians compare in their evaluations within this pivotal subsample. Table B.5 in the Appendix documents that there is substantial disagreement between

Table 5: MD and Physician Assessment of Changes in Quality of Care in the Same Sample

	Qualitative error/harm				DALY loss	
	Any error?*	Short-term harm	Long-term harm	Less harm	Severe loss*	Lower loss*
Panel A: MD ratings						
Effect of assisted note	-0.158***	-0.125**	-0.084**	0.254***	-0.007	0.135***
	(0.057)	(0.055)	(0.035)	(0.078)	(0.018)	(0.043)
Order effect	-0.068	-0.036	-0.013	0.058	-0.025	0.010
	(0.057)	(0.055)	(0.035)	(0.078)	(0.018)	(0.043)
Avg. unassisted note	0.634***	0.500***	0.201***	0.126**	0.063***	0.292***
	(0.041)	(0.039)	(0.024)	(0.051)	(0.016)	(0.035)
R-squared	0.12	0.07	0.09	0.12	0.03	0.08
N (MD-patient pair)	118	118	118	117	118	118
Panel B: Physician ratings (MD sample)						
Effect of assisted note	-0.036	-0.050	-0.065	-0.071	-0.016	-0.005
	(0.048)	(0.062)	(0.040)	(0.119)	(0.016)	(0.116)
Order effect	0.036	-0.014	-0.065	0.071	-0.016	0.102
	(0.048)	(0.062)	(0.040)	(0.119)	(0.016)	(0.116)
Avg. unassisted note	0.932***	0.905***	0.471***	0.407***	0.160***	0.341***
	(0.033)	(0.040)	(0.031)	(0.084)	(0.016)	(0.082)
R-squared	0.02	0.01	0.09	0.01	0.03	0.01
N (Patients)	59	59	59	59	59	59

This table presents fixed effects regression results comparing Medical Doctor (MD) and Medical Officer (MO) assessments of care quality for the same sample of 59 patients. These patients were selected for MD review based on either a high DALY loss rating from the physician in the unassisted note or a substantial change in the physician's loss rating between unassisted and assisted notes. Each patient was independently reviewed by two MDs and one physician, with each reviewer evaluating both the unassisted and assisted SOAP notes. Panel A shows MD ratings across six outcome measures: qualitative errors/harms (any error, short-term harm, long-term harm) and DALY loss outcomes (less harm, severe loss, lower loss), based on 118 MD-patient pairs (59 patients $\times$ 2 MDs per patient). Panel B presents physician ratings for the same 59 patients. The regressions use reviewer-patient pair fixed effects to compare assisted versus unassisted notes within the same patient. The reported N refers to the number of reviewer-patient pairs. Standard errors are clustered at the patient level for MD ratings (Panel A) and robust for physician ratings (Panel B) and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

the physicians and MDs on the presence of short- and long-term harm as well as preference between the two notes.

Table 5 shows treatment effects for the MD ratings in Panel A and the physician ratings in the same sample of patients in Panel B. Because two MDs rated each note, standard errors were clustered at the patient level. However, results are very similar if each MD ratings are analyzed separately (not shown). Physician ratings in Panel B confirm this subsample has much larger baseline rates of error and harm than the full sample, as well as larger – but still statistically insignificant – estimated effects of LLM-assistance. By comparison, the MD ratings indicate a much stronger preference for the assisted SOAP note, with significant reductions in error and harm rates, and a much higher share of assisted notes rated as causing "Less harm" and "Lower loss" than the unassisted note. A manual review suggests that the differences are driven by patients with uncontrolled hypertension. The LLM seems to change the health worker's treatment plan for hypertensive patients, but the MDs seem to ascribe greater health gains than the physicians to the additional

Table 6: Under- and Over-Testing for Malaria, UTI and Anemia.

	Malaria		UTI		Anemia	
	Over (1)	Under (2)	Over (3)	Under (4)	Over (5)	Under (6)
Panel A						
Effect of assisted note	-0.037***	0.002	0.028***	-0.003	0.045***	-0.002
	(0.011)	(0.002)	(0.010)	(0.003)	(0.012)	(0.002)
Avg. unassigned note	0.327***	0.003***	0.062***	0.037***	0.073***	0.100***
	(0.006)	(0.001)	(0.005)	(0.001)	(0.006)	(0.001)
R squared	0.019	0.002	0.020	0.003	0.032	0.002
N (Patients)	575	575	354	354	449	449
Panel B						
Avg. assisted note	0.290	0.005	0.090	0.034	0.118	0.098
Avg. physician	0.224	0.010	0.031	0.034	0.036	0.100
Absolute testing rate (unassisted)	0.552		0.080		0.105	
Absolute testing rate (assisted)	0.506		0.120		0.170	
Absolute testing rate (physician)	0.223		0.045		0.045	
Share abnormal test	0.026		0.042		0.109	

The independent variables for overtesting are coded as 1 whenever the test was ordered by the health worker but the patient exhibited no related symptoms in the symptom screen or tested negative. The independent variables for undertesting are coded as 1 whenever the test was not ordered by the health worker but the patient had a positive test result either in our screening tests or in a test ordered by the physician.
Patient fixed effects included. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

care that the LLM recommends. This divergence between the MD and treating physician's assessment might reflect superior knowledge about future risks on the part of more highly trained MDs. Alternatively, the treating physicians might ascribe lower health returns to the LLM-induced treatment plan updates based on direct interactions with the patient that inform expectations regarding the long-term success of more aggressive treatment. check. I found 3 assessments from each inclusion criteria to see if MDs picked up on it.

3.3 Testing and Treatment Misallocation

A key measure of quality is whether the LLM improved the allocation of medical care, in the sense that appropriate treatments for a given condition are matched to patients with that specific condition. Such efficient allocation usually requires correct identification of a patient's condition, i.e. diagnosis. We use the results from our expanded testing for malaria, UTI, and anemia to determine whether the LLM improved the allocation of medical testing and, ultimately, treatment, for these conditions.

Effects on Overtesting and Undertesting. Panel A of Table 6 shows the effect of the LLM on over-testing and under-testing for the three conditions, where over-testing is defined as the health worker ordering a test for a patient with no symptoms or a confirmed negative test result, and under-testing is defined as not ordering a test for a patient with a positive test result.[2] Panel B also shows average over- and under-testing

[2]While false negatives are rare, medical tests can produce false positives. We here assume that the reported symptoms qualifying patients for testing, combined with the test result, provide sufficient likelihood of the condition to warrant treatment.

and the absolute testing rates in the health worker and physician SOAP notes, and the total share of patients with symptoms and an abnormal test for each condition, i.e. the assumed share of patients who should have received treatment in this sample.

A first observation is that physicians are generally more efficient in the use of medical tests: despite lower absolute testing rates for all conditions and especially for malaria, their under-testing rates are almost identical to those of the health workers (and consequently, their overtesting rates are significantly lower). The average of the unassisted note shows that health workers overtest in 32.7% of patients for malaria and only 2.6% of patients have positive tests.[3] There is very little under-testing for malaria. Conversely, while 4.2% of urine test results are consistent with UTI and 10.9% of blood tests are consistent with anemia, both health workers and physicians fail to order tests for most of these cases. Unfortunately, while the LLM leads to the addition of a large number of tests – 8% of patients for UTI, and 6.5% for anemia – almost all the additional testing is over-testing, in other words, the LLM does not identify the right patients for testing.

We conduct a variety of robustness checks. First, a share of patients who were eligible for medical tests based on their symptoms did not consent to the test and had to be excluded from the analysis. This may bias us towards finding more overtesting and missing reductions in under-testing given that these patients exhibited symptoms. In Appendix Table B.8 we therefore restrict the sample to patients who were actually tested (i.e. exhibited symptoms and qualified for our screening tests, or were tested by the physician) as a proxy bound on effect sizes. The effects are qualitatively similar and there is no significant reduction in over- or undertesting.

Second, negative test results are part of an efficient testing strategy and are not sufficient evidence to indicate poor decision-making. To address this issue, we define alternative outcome variables that use the physician's decision to test as a benchmark, especially for overtesting (see appendix B.7). The results are remarkably similar : both the LLM-induced increase in anemia and UTI tests and the reduction in malaria testing are in patients that the physicians did not test, and so overtesting is increased for the former but reduced for the latter.

Testing Cost Effectiveness. Another way to address testing efficiency is to ask whether LLM-induced changes to test utilization pass a cost-benefit analysis: an increase in testing for UTI and anemia that identifies even a small number of additional positive cases could be worthwhile if tests are cheap relative to the value of identifying positive patients. Conversely, identifying one fewer malaria-positive patient could be beneficial if it saves many negative tests. A back-of-the-envelope cost-effectiveness analysis based on the marginal detection rates in our sample suggests that the LLM has a net positive effect on malaria testing if

[3] The study took place in the dry season when malaria rates are low.

Table 7: Treatment Misallocation for Malaria, UTI and Anemia.

	Overall*	Malaria	UTI	Anemia
Effect of assisted note	-0.004	-0.002	-0.006	-0.007
	(0.004)	(0.006)	(0.004)	(0.007)
Avg. unassisted note	0.074***	0.030***	0.060***	0.144***
	(0.002)	(0.003)	(0.002)	(0.004)
R squared	0.001	0.000	0.006	0.002
N (Patients)	1,416	604	358	454

Misallocation is coded as 1 when there is an intent to treat not conditioned on a positive test but the test is negative; or when the incorrect drug or dosage was prescribed and the test is positive. When the test result is missing, misallocation is coded as 1 when there is an unconditional incorrect drug or dosage, and 0 when a conditional plan specifies correct drug and dosage and otherwise as missing. Overall misallocation is prespecified in the pre-registration plan and at the patient-test level.

we believe that it is justified to avoid 30 additional RDTs in exchange for missing one malaria case. For urine and blood testing, LLM-induced additional tests are worthwhile if it is desirable to conduct 15 additional UTI tests and 27 anemia tests in exchange for detecting one additional case each (see Appendix Table B.4).

Misallocation of Treatment. Ultimately, what matters for quality of care is whether patients receive the right treatment. Misallocation of treatment is a persistent problem in many low-resource settings (e.g. (Das et al., 2016; Sulis et al., 2020; Wagner et al., 2024)). Undertreatment may occur if a given condition is not detected or if the provider gives the wrong drug. Overtreatment for a condition may occur if the provider gives presumptive treatment without a test, or treats despite a negative test result.[4] Overtreatment can simply be a result of misdiagnosis, but may also reflect patient demand or provider incentives to up-sell (Currie et al., 2014; Lopez et al., 2022).

While we cannot identify misallocation of treatment for all diseases and all care plans, we can ask whether treatment for malaria, UTI and anemia was correctly received (or not received) in our screened sample. As illustrated in Table 7, there were no statistically significant reductions in treatment misallocation associated with LLM assistance. It is worth keeping mind that we may underestimate the incidence of overtreatment in our setting, because health workers knew they were under scrutiny and at the same time were not actually making decisions about patient care and so were not subject to patient pressure or a motivation to increase clinic revenue. Accordingly, most health workers created conditional treatment plans that they would only give treatment with a confirmatory test result. Note also that the study period did not overlap with malaria season, removing a primary potential cause of undertreatment among the conditions we screened. Our

[4]We define misallocation at the level of the condition. At the level of the patient, misdiagnosis means that overtreatment for one condition often coincides with undertreatment for another. Note also that empiric therapy can be the correct approach when diagnostic tools are unavailable or there is a life-threatening emergency. This would not be the case for patients in this study.

results partly confirm this intuition; we find low baseline rates of misallocation, and, in line with the results on medical testing, we do not find that the LLM feedback improves these rates significantly.

3.4 Concordance Between Health Worker and Physician Diagnosis and Treatment

Yet another approach to assessing quality of care is to assess whether LLMs make the care plans of health workers more like those of treating physicians. Given physicians' longer training and broader scope of practice, it is reasonable to believe that physician care plans are higher-quality care than those of community health workers. Thus, if LLMs induced better alignment between health workers and physicians with respect to specific care plan elements, it would be suggestive of LLM benefits.

In order to assess objective matches in diagnosis and treatment elements, we constructed categories that grouped similar diagnostic codes and prescriptions. We then created variables measuring discordance between physician and health worker care plans with respect to these grouped elements, These discordance measures are defined as the count of items in the health worker note that were not included in the physician note (overprescription/overdiagnosis) and the number of items in the physician note not included in the health worker note (underprescription/underdiagnosis).[5]

In Table 8 we report indicators for any over- and under-prescription and any over- and under-diagnosis to facilitate comparison with the subjective assessments of missing vs. unnecessary items by the physician in Table 3. The results show high rates of mismatch and small improvements across the different indicators, with only the effect on under-prescription marginally significant. In Appendix Table B.9 we also report the effects on the count variables.

4 Mechanisms: adoption of LLM suggestions and discretion in data gathering

Overall, our results paint a mixed picture. Both the objective results on concordance and the physician evaluations suggest at best small benefits of LLM assistance, and substantial variability.

Why did the LLM feedback not move the health worker's care plans closer to that of the treating physician? In this section, we directly analyze the LLM feedback health workers received, as well as the role of health worker discretion in adopting LLM suggestions and the role of incomplete data elicitation from patients by health workers.

We first check whether the LLM provided actionable recommendations that *could* have led to more concordance with the physician. We ask GPT-5 to identify the specific recommendations the LLM made

[5]Two PIs reviewed all diagnoses in the study and all treatments for these diagnoses blinded to their origin SOAP note and created the groupings. This approach was intended to account for different forms of giving the same medication (e.g. as a syrup or a pill) and different brand names, along with idiosyncratic differences in the use of the ICD-10 based diagnosis options, which were unlikely to be clinically important. We use the conditional physician SOAP notes for comparability. For the diagnoses, we use either the "clinical indication" fields or the main "diagnosis" field in the EMR if there are no prescription drugs in the SOAP note.

Table 8: Health Worker-Physician Objective Discordance

	Prescriptions		Diagnoses	
	Any Over	Any Under	Any Over	Any Under
Effect of assisted note	-0.011	-0.014*	-0.011	-0.002
	(0.010)	(0.008)	(0.009)	(0.007)
Avg. unassisted note	0.812***	0.656***	0.645***	0.730***
	(0.005)	(0.004)	(0.004)	(0.003)
R-squared	0.002	0.005	0.002	0.000
N (Patients)	660	660	660	660

This table presents fixed-effects regressions examining the effect of AI assistance on discordance between health workers and physicians "Any Over" is a dummy for whether there were any items documented by the health worker but not by the physician, "Any Under" indicates any items documented by the physician but not by the health worker. All notes in the study are included in the sample. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 9: LLM Recommendations: Alignment with Physician and Health Worker Follow-Through

	Any		Rx		Labtest		Advice		Subj/Obj
Variable	Mean (SD)	%	Mean	%	Mean	%	Mean	%	Mean
Changes	3.757 (1.645)		0.973	26%	0.921	25%	0.635	17%	1.161
More similar to physician	1.980 (1.633)	53%	0.488	50%	0.369	40%	0.362	57%	0.733
Less similar to physician	1.324 (1.305)	35%	0.429	44%	0.486	53%	0.157	25%	0.219
Direction unclear	0.453 (0.877)	12%	0.056	6%	0.065	7%	0.117	18%	0.210
Changes followed	1.269 (1.189)	34%	0.430	44%	0.410	45%	0.161	25%	0.251
More similar to physician	0.739 (0.995)	58%	0.252	59%	0.178	43%	0.122	75%	0.178
Less similar to physician	0.444 (0.720)	35%	0.150	35%	0.219	53%	0.018	11%	0.050
Direction unclear	0.087 (0.340)	7%	0.027	6%	0.014	3%	0.021	13%	0.023
N (Patients)	658								

This table reports counts of AI-generated recommendations to modify health worker clinical notes. Recommendations classified as "more similar to physician" would close the gap between the health worker note and the physician's independent assessment, while those"less similar to physician" would widen the gap. "Followed" indicates the health worker implemented the recommendation.

for the health worker, label them by whether the LLM suggests changing a part of the treatment plan (e.g., "Consider an h. pylori test") or keeping it the same (e.g., "Continue the antibiotic treatment plan"), and classify which part of the treatment plan they refer to: prescriptions, medical (lab) tests, or advice.

Next, we ask whether the LLM recommendation would make the health worker note *more similar* to that of the physician – meaning whether following it would completely or partially reduce any differences between health worker and physician (e.g., add a missing physician item, remove an incorrect health worker item) – or *less similar* – e.g., introduce a new test/drug not in the physician note, or alter a shared correct element to diverge.

The top panel of Table 9 shows the average counts of recommendations per patient that included some change, broken down by whether those changes encourage a note to be more or less similar to the physician.

The LLM makes a lot of suggestions – 3.8 per patient – despite the prompt instructions to be concise and focus on the most important issues. 53% of these recommendations would make the health worker note more similar to the physician. Another 35% would make the notes less similar and 12% of the recommendations do neither; these might suggest for example to replace a medication not in the physician note with another one not in the physician note.

We next conduct two pieces of analysis to understand the role of the health worker better. First, we let GPT-5 label each recommendation as "followed" or "ignored" based on the changes from the unassisted to the assisted health worker note.[6] Table 9 shows that the health workers only followed 34% of the LLM's recommendations. They were on average slightly more likely to follow good advice than bad: 58% of the followed changes increased concordance with the physician, while 42% had an unclear effect or decreased concordance. The share of (followed) changes that made the notes "more similar" varied across categories from 43% for lab tests to 75% for advice. The health workers were in all categories slightly better than chance in identifying and following the "more similar" rather than the "less similar" and "unclear" recommendations, but the "human in the loop" is far from a perfect safeguard against the 47% of LLM recommendations that are unhelpful (measured against the physician as the benchmark). The health workers had the best ratio of more to less similar followed suggestions in the "Advice" category; on-site physicians in their subjective reviews may have weighted good advice less heavily than appropriate prescriptions or treatment.

Second, we want to test the hypothesis that the unassisted SOAP note does not contain sufficient relevant patient information for the LLM to provide useful feedback (e.g., by flagging inconsistencies or unaddressed signs and symptoms). To do so we construct a synthetic SOAP note composed of the Assessment and Plan from the health worker note, but the Subjective and Objective components of the physician note. In other words, we combine the physician's description of the patient's appearance, signs, and symptoms with the health worker's lab tests ordered, prescriptions and advice, and diagnosis. We then regenerate the LLM feedback with the original prompt the LLM used in the experiment and repeat the first part of the exercise above to measure the extent to which LLM recommendations based on the *synthetic note* would make the health worker's treatment plan more similar to the treating physician's. If the physician records more detailed information than the health worker, it might enable the LLM to identify inconsistencies in the health worker's treatment plan and improve the alignment of the LLM's recommendations with the physician. To test this, we run a fixed effect regression with the original and the synthetic SOAP notes where the outcome Y_{ik} is the number of LLM recommendations in a given category ("more similar" or "less similar", by component) and the independent variable T_k is an indicator for the synthetic note.

[6]The first steps above were done within a single prompt that evaluated the health worker's unassisted note, the LLM feedback, and the physician's note. This last step was completed with a prompt that evaluated the structured list of annotated recommendations and the health worker's assisted and unassisted notes.

Table 10: Synthetic Note Analysis: Effect of Including Physician's Patient Observations on Quality of LLM Feedback.

	Any Change				
	Any	Rx	Lab Test	Advice	Rx+Lab
Effect of Physician's S-O	0.124	-0.012	0.031	0.024	0.018
	(0.089)	(0.045)	(0.037)	(0.044)	(0.058)
Baseline (Health Worker's S-O)	3.756***	0.973***	0.921***	0.635***	1.894***
	(0.064)	(0.034)	(0.029)	(0.033)	(0.044)
	More Similar to Physician				
	Any	Rx	Lab Test	Advice	Rx+Lab
Effect of Physician's S-O	0.007	0.030	0.096***	-0.009	0.126**
	(0.081)	(0.037)	(0.030)	(0.034)	(0.051)
Baseline (Health Worker's S-O)	1.980***	0.488***	0.369***	0.362***	0.857***
	(0.064)	(0.028)	(0.023)	(0.027)	(0.039)
	Less Similar to Physician				
	Any	Rx	Lab Test	Advice	Rx+Lab
Effect of Physician's S-O	-0.078	-0.020	-0.068**	0.006	-0.088*
	(0.062)	(0.033)	(0.030)	(0.023)	(0.045)
Baseline (Health Worker's S-O)	1.324***	0.429***	0.486***	0.156***	0.915***
	(0.051)	(0.025)	(0.024)	(0.017)	(0.036)
N (Patients)	659	659	659	659	659

This table reports the impact of providing the physician's subjective-objective (S-O) information to the LLM compared to providing only the health worker's S-O information, estimated using patient fixed effects regressions. Panel A shows the total number of recommendation changes by category. Panel B presents changes that made the LLM's assessment more similar to the physician's final assessment. Panel C shows changes that made the LLM's assessment less similar to the physician's final assessment. Each column represents different types of recommendations: any change, prescriptions (Rx), lab tests, advice, and the sum of Rx and lab test changes. Standard errors are clustered at the patient level. Statistical significance: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

Table 10 shows the change in the number of recommendations, in total and split into "more similar" and "less similar." We see a decrease of "less similar" recommendations with the synthetic note by -0.078 and a small shift from "less" to "more" similar advice in the Rx category (both not statistically significant). The synthetic note significantly shifts recommendations on lab tests from the "less similar" to the "more similar" category. This is suggestive evidence that the subjective and objective information the health worker records partly contributes to leading the LLM "down the wrong path." However, even with the treating physician's S-O portion of the note, the LLM continues to make on average more than one recommendation per patient that would move the health worker and physician further apart. We cannot distinguish whether this is because the physician's S-O records on the patient are also incomplete, or because the LLM interpretation of the S-O contents is different from the physicians.

5 Conclusion

We study the use of LLMs for real-time clinical decision support in a low resource setting. We find that health workers value the LLM feedback and respond with many changes to their care plans, and in a subsample of patients, care quality seems to improve according to MDs who review a record of the patient interaction. Our main contribution is an analysis of care plan reviews by physicians who evaluated and treated the patient during the same clinic encounter, as well as medical test results for three commonly treated health conditions.

These new methods of measurement suggest a more mixed picture. Objective testing suggests that LLMs reduce overtesting for malaria but increase testing for anemia and UTIs without identifying any of the undetected abnormal cases. The treating physicians also mostly do not judge the LLM-assisted care plans as improved, and objective measures of concordance between health worker and physician show almost no change. The discrepancy between the treating physicians and reviewing MDs may reflect differences in expertise, but could also be due to better information or a different assessment of the health consequences of the changed treatment plan on the part of the treating physicians. In the subsample reviewed by the MDs, a major benefit of the assisted notes was better identification of hypertensive patients; local physicians might believe that addressing the elevated blood pressure did not contribute to solving the patient's acute problem, was not a patient priority, or would not be effective.

Exploring mechanisms, we find that LLMs make many suggestions for changes that do not move the health worker's care plan closer to the physician, and while the health workers follow only about one third of the LLM recommendations, they are only slightly more likely to follow advice that improves alignment with physicians. There is suggestive evidence that some of the problem can be attributed to insufficient patient information recorded by the health workers, but even when supplying the same signs and symptoms the treating physicians recorded, the LLM still makes many recommendations that would make the care provided by the health workers less similar to that provided by the physician.

There are two main takeaways from our results, one practical and one methodological. Current generation LLMs seem capable of modestly improving some aspects of healthcare delivery, although the benefits are mixed when assessed using objective tests or by physicians who evaluated and treated the patient. While we do not observe evidence that LLMs harm patient care, the absence of consistently measured and large beneficial effects suggests that deployment of LLMs in clinical decision support might not (yet) be an urgent policy priority.

Methodologically, our results highlight the importance of assessing care through multiple means. Comparison with medical guidelines or 3rd party retrospective review may deliver a very different picture from

an evaluation based on objective test results or comparison with the care decisions by physicians who see and treat the same patient.

References

Ajisegiri, W. S., S. Abimbola, A. G. Tesema, O. O. Odusanya, D. Peiris, and R. Joshi (2023). "we just have to help": Community health workers' informal task-shifting and task-sharing practices for hypertension and diabetes care in nigeria. *Frontiers in Public Health 11*, 1038062.

Blease, C. R., C. Locher, J. Gaab, M. Hägglund, and K. D. Mandl (2024). Generative artificial intelligence in primary care: an online survey of uk general practitioners. *BMJ Health & Care Informatics 31*(1), e101102.

Cadamuro, J., F. Cabitza, Z. Debeljak, S. D. Bruyne, G. Frans, S. M. Perez, H. Ozdemir, A. Tolios, A. Carobene, and A. Padoan (2023). Potentials and pitfalls of chatgpt and natural-language artificial intelligence models for the understanding of laboratory medicine test results. an assessment by the european federation of clinical chemistry and laboratory medicine (eflm) working group on artificial intelligence (wg-ai). *Clinical Chemistry and Laboratory Medicine (CCLM) 61*(7), 1158–1166.

Currie, J., W. Lin, and J. Meng (2014). Addressing antibiotic abuse in China: An experimental audit study. *Journal of Development Economics 110*, 39–51.

Das, J., A. Holla, A. Mohpal, and K. Muralidharan (2016, December). Quality and accountability in health care delivery: Audit-study evidence from primary care in india. *American Economic Review 106*(12), 3765–99.

Dotimi, D. A., Y. Wamanyi, and J. U. Silas (2024, jan). Community health practice: the nigerian model. *International Journal of Community Medicine and Public Health 11*(1), 560–565.

Gaber, F., M. Shaik, F. Allega, A. J. Bilecz, F. Busch, K. Goon, V. Franke, and A. Akalin (2025, May). Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine 8*(1), 263.

Goh, E., R. Gallo, J. Hom, E. Strong, Y. Weng, H. Kerman, J. A. Cool, Z. Kanjee, A. S. Parsons, N. Ahuja, E. Horvitz, D. Yang, A. Milstein, A. P. J. Olson, A. Rodman, and J. H. Chen (2024, October). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open 7*(10), e2440969.

Hager, P., F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis, and D. Rueckert (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine 30*, 2613–2622.

Huang, T., P. H. Hsieh, and Y. Chang (2024). Performance comparison of junior residents and ChatGPT in the objective structured clinical examination (OSCE) for medical history taking and documentation of medical records: Development and usability study. *JMIR Medical Education 10*(e59902).

Korom, R., S. Kiptinness, N. Adan, K. Said, C. Ithuli, O. Rotich, B. Kimani, I. King'ori, S. Kamau,

E. Atemba, et al. (2025). Ai-based clinical decision support for primary care: A real-world study. *arXiv preprint arXiv:2507.16947*.

Lopez, C., A. Sautmann, and S. Schaner (2022, January). Does patient demand contribute to the overuse of prescription drugs? *American Economic Journal: Applied Economics 14*(1), 225–60.

Okeke, E. N. (2023, March). When a Doctor Falls from the Sky: The Impact of Easing Doctor Supply Constraints on Mortality. *American Economic Review 113*(3), 585–627.

Okoroafor, S. C., A. Ahmat, M. Osubor, J. Nyoni, J. Bassey, and W. Alemu (2022). Assessing the staffing needs for primary health care centers in cross river state, nigeria: a workload indicators of staffing needs study. *Human Resources for Health 19*(Suppl 1), 108.

Pressman, S. M., S. Borna, C. A. Gomez-Cabello, S. A. Haider, C. R. Haider, and A. J. Forte (2024). Clinical and surgical applications of large language models: A systematic review. *Journal of Clinical Medicine 13*(11), 3041.

Salomon, J. A., J. A. Haagsma, A. Davis, C. Maertens de Noordhout, S. Polinder, A. H. Havelaar, A. Cassini, B. Devleesschauwer, M. Kretzschmar, N. Speybroeck, C. J. L. Murray, and T. Vos (2015, nov). Disability weights for the global burden of disease 2013 study. *The Lancet Global Health 3*(11), e712–e723.

Shmilovitch, A. H., M. Katson, M. Cohen-Shelly, S. Peretz, D. Aran, and S. Shelly (2025, March). GPT-4 as a Clinical Decision Support Tool in Ischemic Stroke Management: Evaluation Study. *JMIR AI 4*(1), e60391.

Sulis, G., B. Daniels, A. Kwan, S. Gandra, A. Daftary, J. Das, and M. Pai (2020, September). Antibiotic overuse in the primary health care setting: a secondary data analysis of standardised patient studies from India, China and Kenya. *BMJ Global Health* (9), e003393.

Takita, H., D. Kabata, S. L. Walston, H. Tatekawa, K. Saito, Y. Tsujimoto, Y. Miki, and D. Ueda (2025, 10). A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. *NPJ Digital Medicine 8*(1), 1–13.

Tu, T. et al. (2025). Towards conversational diagnostic artificial intelligence. *Nature 642*, 442–450.

Wagner, Z., M. Mohanan, R. Zutshi, A. Mukherji, and N. Sood (2024). What drives poor quality of care for child diarrhea? Experimental evidence from India. *Science 383*(6683), eadj9986.

Wan, P., Z. Huang, W. Tang, Y. Nie, D. Pei, S. Deng, J. Chen, Y. Zhou, H. Duan, Q. Chen, and E. Long (2024, October). Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nature Medicine 30*(10), 2878–2885.

Yıldırım, C., A. Aykut, E. Günsoy, and M. V. Öncül (2025, October). Evaluating GPT-4o for emergency disposition of complex respiratory cases with pulmonology consultation: a diagnostic accuracy study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine 33*(1), 159.

Supplemental Materials

A Sample SOAP notes and LLM prompt and feedback

An example CHEW note as extracted from the EMR is here.

```
Mar 12 2025
ASSESSMENT
Diagnosis
Headache
Differential Diagnosis
Plasmodium falciparum malaria; Typhoid fever
PLAN
Treatment plan
Advice on adequate rest and sleep
Advice on adequate fluid intake
Advice on proper diet
Advice on using mosquito net
Avoid eating spicy food
Review with the lab test result
Exhaustive list of Investigations Ordered
Stool MCS (done)
Malaria Parasite (Smear +RDT) (done)
Prescriptions conditional on test results
1g Emcap (Paracetamol 500mg) Caplets X10 3 times per day 3 days
1capsule Afrab (Ciprofloxacin 500mg) Caplet x 10 Twice daily 3 days
if typhoid fever test result is positive
1capsule Amatem Soft gel(Artemether + Lumenfantrine 20/120mg) Capsulex24 3 times per day 5 days
if malaria test result is positive
SUBJECTIVE
AGE
34y
Gender
female
Chief Complaint
Fever and headache for 3
History of presenting Illness (HPI)
A patient of 34 years old come to the clinic with the complain of fever and headache for 3 days, with heart burn, she said the headache is radiating to her eyes, and
she rated the pain as 4 in pain scale assessment, patient also said the fever usually started at night, but she has relieved when she sleep. And also she suffering
from heart and has relieve when take anti-acid, no abdominal pain, no nausea or vomiting, no dizziness, no loss of appetite, she is not on any medication and has
known allergy.
Medical History
No any past medical history
Gynaecology Summary
She had history of 2 CS and 2 pregnancies, the 2 children are alive
OBJECTIVE
Review of Systems
Neurological system: she has fever
Nervous system: headache
Respiratory system: normal
Digestive system: normal
cardiovascular system: heart burn
Physical Examination
General condition: stable, I'll looking, not pale, no pallor, not dehydrated, no fetal edema
Vital Signs
Blood pressure (131/87 mm[Hg])
Temperature (37.7°C)
Respiratory rate (18/min)
Heart rate (87/min)
Oxygen saturation (98%)
Anthropometry
Height (143 cm)
Weight (58 kg)
BMI (28.4 kg/m2)
```

The following is the prompt. A CHEW note (like the one above) is appended to this prompt and sent to the LLM:

```
I want you to act as an attending physician giving immediate, specific, and actionable feedback to a Community Health Worker (CHEW) based on the SOAP note they
created.  Your response will guide the CHEW, who is currently with the patient, so focus on concise, high-priority advice.

### Key Instructions:
1. **SOAP Note Context**: This SOAP note is fro an acute care clinic in Kano Nigeria. It is difficult here to measure vital signs accurately, include data from
today's clinic visit as well as historical patient-reported data if available (e.g., recent fever or symptoms reported earlier).
2. **CHEW Capabilities**: Keep in mind the CHEW's limited physical exam skills. Suggest only basic inspection, palpation, or vital sign checks.
3. **Clinic Constraints**: Recommendations must align with the clinic's capabilities, focusing on the following:
- **Available Testing**: Malaria RDT, Typhoid RDT, PCV, Pregnancy Test, Urine Dipstick, H. Pylori, HIV Confirmatory Test (ELFA).
- **No Imaging**: Avoid suggestions for imaging, ultrasound, or other unavailable diagnostics.
- **Avoid Low-Value Care**: Recommend tests only when they are critical to the diagnosis or treatment.
- **Return visits are rare**: The goal is to treat the patient today and not require return visits.
4. **Disease-Specific Treatment Recommendations**: Follow these clinic-specific guidelines:
- **Malaria**:
- Only test if fever cannot be explained by other causes (e.g., upper respiratory infection, UTI).
- Treat only if malaria RDT is positive. Use oral ACT for 3 days or injectable ACT if oral therapy isn't possible.
- **Typhoid**:
- Test only if fever lasts >=3 days and is accompanied by abdominal pain, diarrhea, or constipation.
- Treat positive cases with oral Ciprofloxacin or Azithromycin for 5 days.
- **H. Pylori**:
- Test if gastritis symptoms persist >3 weeks or recur after antacid/PPI treatment.
- Treat positive cases with triple therapy.
- **Anemia**:
- Test if fatigue, exertional shortness of breath, dizziness, or lack of energy is chronic.
- Treat positive cases with oral iron for 3 months.
- **Hypertension**: Address only if associated with signs of severe cardiac conditions, as it is a low priority.
5. **Feedback Scope**: Consider giving advice in the following categories:
- **Feedback on diagnosis**: Note inconsistencies, overlooked possibilities, or any inaccurate reasoning.
- **Feedback on treatment plan**: Identify errors, such as inappropriate antibiotics, or suggest improvements.
- **Assessment of risk**: Highlight potential for serious conditions based on symptoms.
- **Test recommendations**: Suggest needed tests and flag unnecessary ones.
- **Additional Diagnostic Questions: Suggest questions the CHEW could ask to differentiate between diagnosis and treatment options, and include specific instructions
of how to interpret different answers.
6. **Brevity**: Your feedback must be clear, simple, concise, and actionable. The ideal response starts with 2 or 3 short, direct, actionable sentences.  If helpful
for a CHEW, a simple additional paragraph can be used to support or explain the actions for each sentence.

### Patient Record:
[CHEW NOTE HERE]
```

The LLM response is received and forwarded to the CHEW in real time. The LLM response for this note was:

```
**Immediate Actions:**
1. **Recheck and Manage Blood Pressure:** Immediately remeasure the patient's blood pressure to confirm the elevated reading (163/107 mmHg). If persistently high,
initiate or adjust antihypertensive therapy before starting NSAIDs.
2. **Assess Joint for Specific Signs:** Perform a basic inspection of the right leg joint for swelling, redness, or limited range of motion to better define the joint
derangement diagnosis.
3. **Evaluate NSAID Use:** Given her hypertension, consider the risks of prescribing Naproxen. If NSAIDs are necessary for pain management, ensure close monitoring of
her blood pressure.

**Additional Guidance:**
High blood pressure needs urgent attention to prevent complications. NSAIDs  like Naproxen can increase blood pressure, so it's crucial to manage her hypertension
effectively before and while using these medications. Clarifying the exact nature of the joint issue will help in providing targeted treatment and reduce the need for
return visits
```

B Additional Figures and Tables

Table B.1: Physician Assessment of Quality of Care, SOAP Note A only.

	Qualitative error/harm				DALY loss	
	Any error?*	Short-term harm	Long-term harm	Less harm	Severe loss*	Lower loss*
Effect of assisted note	-0.016	0.039	-0.043*	0.018	-0.024	0.052
	(0.045)	(0.047)	(0.025)	(0.038)	(0.020)	(0.033)
Avg. unassisted note	0.673***	0.514***	0.095***	0.196***	0.059***	0.114***
	(0.032)	(0.034)	(0.020)	(0.027)	(0.016)	(0.021)
R-squared	0.00	0.00	0.01	0.00	0.00	0.01
N (cases)	450	450	450	448	450	450

Table B.2: LLM Assessment of Changes in the Quality of Care

	Qualitative error/harm	
	Short-term harm	Long-term harm
Effect of assisted note	-0.011	-0.020
	(0.011)	(0.026)
Order effect	0.007	0.020
	(0.011)	(0.026)
Avg. unassisted note	0.959***	0.572***
	(0.007)	(0.018)
R-squared	0.00	0.00
N (Patients)	450	376

Table B.3: Physician Assessment of Changes in Quality of Care, Controlling for Assisted Note Guess

	Qualitative error/harm				DALY loss	
	Any error?*	Short-term harm	Long-term harm	Less harm	Severe loss*	Lower loss*
Effect of assisted note	-0.004	-0.010	-0.007	0.006	-0.009*	0.020
	(0.016)	(0.017)	(0.005)	(0.032)	(0.005)	(0.026)
Physician guessed assisted	-0.038**	-0.054***	-0.015**	0.158***	-0.012*	0.109***
	(0.018)	(0.020)	(0.006)	(0.037)	(0.006)	(0.030)
Order effect	0.024	0.014	-0.010*	-0.060*	-0.003	-0.011
	(0.016)	(0.017)	(0.006)	(0.032)	(0.006)	(0.025)
Avg. unassisted note	0.658***	0.546***	0.092***	0.199***	0.059***	0.097***
	(0.013)	(0.015)	(0.006)	(0.027)	(0.006)	(0.021)
R-squared	0.02	0.02	0.03	0.06	0.02	0.04
N (Patients)	450	450	450	448	450	450

This table presents fixed effects regression estimates of the impact of AI-assisted clinical notes on various measures of medical assessment quality. The dependent variables include binary indicators for treatment errors and harm assessments (columns 1-4), as well as DALY (disability-adjusted life year) loss metrics (columns 5-6). The sample is restricted to assessments evaluated after February 25, when presentation order was randomized to control for order effect. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.4: Marginal Test Detection Rate, by Testing Condition and Eligibility

	Malaria		UTI		PCV	
	Demo	Symp	Demo	Symp	Demo	Symp
Total patients with test result	575	354	354	123	449	321
Positive tests	15	13	15	14	49	47
Overall positive rate (%)	2.61%	3.67%	4.24%	11.38%	10.91%	14.64%
Unassisted (Baseline)						
Patients with test ordered	364	273	31	18	48	35
Positive tests	13	11	2	2	4	4
Detection rate (%)	3.57%	4.03%	6.45%	11.11%	8.33%	11.43%
Assisted (with LLM)						
Patients with test ordered	334	255	46	24	75	53
Positive tests	12	10	3	3	5	5
Detection rate (%)	3.59%	3.92%	6.52%	12.50%	6.67%	9.43%
Net Marginal Effect						
Change in positive tests	-1	-1	1	1	1	1
Change in tests ordered	-30	-18	15	6	27	18
Marginal detection rate (%)	3.33%	5.56%	6.67%	16.67%	3.70%	5.56%

Sample restricted to demographically eligible patients (meeting age/gender criteria). "Demo" denotes the sample of all patients eligible for screening; "Symp" denotes the subsample who were eligible for lab testing based on symptoms. Detection rate is the share of ordered tests that are positive. Marginal detection rate equals the change in positive tests divided by the change in tests ordered (assisted minus unassisted). When tests are removed (negative change in tests ordered), the marginal detection rate reflects the positive rate among removed tests. Missing test results are imputed as negative for non-symptomatic patients.

Table B.5: MD-Physician Rating Concordance

	κ (SE)		Agreement
	MD A-Physician	MD B-Physician	with ≥1 MD (%)
Short-term harm	0.055 (0.057)	0.025 (0.051)	56.8
Long-term harm	0.277 (0.073)	0.321 (0.076)	72.0
Note preference	0.196 (0.090)	0.291 (0.087)	76.3
Severe DALY loss	0.345 (0.070)	0.545 (0.082)	91.5
Lower DALY loss	0.260 (0.068)	0.157 (0.050)	72.0
N (Patients)	59		

This table presents inter-rater reliability statistics comparing physician evaluations with MD evaluation of selected health worker SOAP notes. Each SOAP note was independently evaluated by one physician (MO) and two randomly selected attending physicians (MDs) from a pool of three (labeled MD A and MD B). Cohen's kappa (κ) measures agreement between the physician and each physician evaluator, with standard errors in parentheses. The "Agreement with $\geq$ 1 MD" column shows the percentage of cases where the physician's rating matched at least one of the two MDs. The outcome variables are coded as binary indicators.

Table B.6: Blinding Validation: Physician Identification of Assisted SOAP Note

Panel A: Distribution of all responses				
	Correct (%)	Incorrect (%)	Unsure (%)	*N*
All patients	49.5	35.6	14.9	618
Fully randomized sample	50.4	36.2	13.3	450
Panel B: Excluding unsure responses				
	Correct (%)	95% CI	*p*-value	*N*
All patients	58.2	[53.9, 62.3]	0.000	526
Fully randomized sample	58.2	[53.3, 63.0]	0.001	390

Physicians evaluated pairs of SOAP notes and were asked to identify which note was LLM-assisted. The answer options included "unsure" (to allow for abstention e.g. in the case of identical notes). The full sample includes all assessments with complete data. The fully randomized sample restricts to assessments after February 25, 2025. 95% confidence intervals use the Wilson score method. *p*-values are from two-tailed binomial tests against 50% (chance performance).

Table B.7: Under- and Over-Testing for Malaria, UTI, and Anemia Relative to Physician Testing Decisions

	Malaria		Urine Analysis		Anemia	
	Over (1)	Under (2)	Over (3)	Under (4)	Over (5)	Under (6)
Effect of assisted note	-0.044***	0.002	0.033***	-0.006	0.058***	-0.008**
	(0.011)	(0.007)	(0.009)	(0.004)	(0.011)	(0.003)
Avg. unassisted note	0.358***	0.029***	0.062***	0.027***	0.095***	0.036***
	(0.005)	(0.003)	(0.005)	(0.002)	(0.005)	(0.002)
R squared	0.025	0.000	0.020	0.004	0.042	0.008
Patients	660	660	660	660	660	660
Share tested by MO	0.223	0.223	0.045	0.045	0.045	0.045

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.8: Under- and Over-Testing for Malaria, UTI, and Anemia (Tested Sample Only)

	Malaria		UTI		Anemia	
	Over (1)	Under (2)	Over (3)	Under (4)	Over (5)	Under (6)
Panel A						
Effect of assisted note	-0.034***	0.003	0.016	-0.008	0.043***	-0.003
	(0.013)	(0.003)	(0.016)	(0.008)	(0.014)	(0.003)
Avg. unassigned note	0.346***	0.005***	0.102***	0.102***	0.073***	0.138***
	(0.006)	(0.001)	(0.008)	(0.004)	(0.007)	(0.002)
R squared	0.018	0.003	0.008	0.008	0.030	0.003
N (Patients)	382	382	127	127	327	327
Panel B						
Avg. assisted note	0.312	0.008	0.118	0.094	0.116	0.135
Avg. physician	0.327	0.016	0.087	0.094	0.049	0.138
Absolute testing rate (unassisted)	0.552		0.080		0.105	
Absolute testing rate (assisted)	0.506		0.120		0.170	
Absolute testing rate (physician)	0.223		0.045		0.045	
Share abnormal test	0.039		0.118		0.150	

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.9: Health Worker-Physician Objective Discordance

	Prescriptions		Diagnoses	
	Over	Under	Over	Under
Effect of assisted note	0.047**	-0.012	-0.002	0.014
	(0.023)	(0.014)	(0.013)	(0.010)
Avg. unassisted note	1.683***	1.058***	0.802***	0.974***
	(0.012)	(0.007)	(0.007)	(0.005)
R-squared	0.006	0.001	0.000	0.003
N (Patients)	660	660	660	660

This table presents fixed effects regressions examining the effect of AI assistance on discordance between health workers and physicians. The dependent variables measure the number of mismatched items per patient assessment. "Over" indicates items documented by the health worker but not by the physician, while "Under" indicates items documented by the physician but not by the health worker. All models include assessment fixed effects and report robust standard errors. All notes in the study are included in the sample. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C Construction of Outcome Measures

C.1 Subjective Physician Evaluation

Here we describe in detail how the various outcome variables are defined and constructed.

Any Treatment Error. During SOAP note evaluation, the physician indicates whether the note contains any error. 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 health worker'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)

We code this question as an indicator variable that equals 1 if the answer is "no". If the answer is "unsure" the variable is treated as missing.

Deviations from the physician's SOAP. The physician is also asked to assess for each SOAP note whether medical tests ordered were necessary or clinically useful vs. unlikely to be clinically useful. We code a variable that counts the number of tests that were unlikely to be clinically useful.

In addition, there are variables that ask whether there are missing or incorrect/unnecessary diagnoses, missing or incorrect/unnecessary medications, and missing or incorrect/ unnecessary care instructions. We code 6 indicator variables for any of these mistakes occurring in the SOAP note.

Short-term harm. To obtain a subjective assessment of patient harm, we ask the physician

> Now suppose the patient were to receive the treatment prescribed in SOAP Note (A) instead of your treatment. Please characterize what the _consequences for the patient would be based on your *own* diagnosis.

> What would be the short-term (temporary) consequences, such as _additional symptoms or discomfort, that the patient would experience?

The answer options for this question are

- Severe: at least 7 additional days of being unable to carry out usual daily activities
- Moderately severe: 3-6 additional days of being unable to carry out usual daily activities
- Moderate: 1-2 additional days of being unable to carry out usual daily activities
- Mild: 2 or more additional days of discomfort, but no effect on usual daily activities
- None or minimal

Long-term harm. This question is followed by

> Suppose the patient received the treatment in SOAP Note (A) instead of your treatment. Would this increase _the risk of serious long-term (permanent) harm for the patient, including death?

with answer options yes, no, and unsure. We code an indicator variable that is 1 if the answer is yes and 0 otherwise. We also asked a follow-up question on the likelihood of such harm, but we do not show the results, because all harm is either unchanged between notes A and B or one of the notes has no harm, so this question does not convey additional information.

C.2 Protocols for Malaria, UTI and Anemia Screening

Within budget constraints, we aimed to screen all relevant patients for malaria, UTI and anemia. We first specified a set of demographic inclusion criteria for participation in screening. All patients who fulfilled these criteria were considered part of the screening sample.

The screening itself consisted of a symptom screen followed by confirmation with a laboratory test. We coded the final result of the screening as either "normal" or "abnormal". Patients with an abnormal screen for the condition should receive treatment.

The symptom screen was designed to identify all patients who would plausibly be treated for the condition if they had a positive lab test. Patients who did not exhibit any relevant symptoms and did not have a test result from a physician-ordered test were coded as having a "normal" screening result. All patients who exhibited at least one relevant symptom and for whom the test was not already ordered by the physician were asked for their consent to conduct a laboratory test. For any patients for whom a test result was available (either test ordered by the physician, or passed the symptom screen and consented to testing), the screening was coded as "abnormal" with a positive test result and "normal with a negative test.

We applied the following demographic and symptom screening criteria:

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

Anemia/PCV test: 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.

C.3 Misallocation of Treatment for Malaria, Anemia and UTI

In order to determine whether a patient received the correct treatment for the three screening conditions, we proceeded in multiple steps.

First, we determined using the clinical indication, test condition attached to a prescription, or drug itself whether the SOAP note indicated an intent to treat for the condition.

Second, we used the treatment guidelines below to determine whether the provider specified the correct treatment for the condition they intended to treat.

Finally, we used information in the conditioning field for each prescription to determine whether they intended to treat the patient only with a positive laboratory test result or unconditionally.

We then coded as "overtreatment" any case in which the provider intended to treat for a condition, did not specify that the treatment was conditional on a test result, and the screening test or a physician-test was normal/negative. This approach accounts for giving the patient a drug intended for the condition in question but choosing a incorrect dosage or treatment option. For example, some health workers prescribe vitamins for anemia but choose a product that does not contain iron; if the patient's PCV test is normal, this is overtreatment because it involves an unnecessary medication.

We code as "undertreatment" any case in which the provider did not intend to treat for the condition, or intended to treat but prescribed an incorrect drug. This accounts for cases where the provider meant to treat but gave something incorrect. For example, a vitamin prescription that does not contain iron is undertreatment for anemia because it is missing the relevant ingredient to address the condition.

Note that we define misallocation at the level of the condition, not the level of the patient. A patient who receives the wrong treatment because they are misdiagnosed is overtreated for one condition and undertreated for another. We only measure misallocation for three conditions and so we cannot fully assess misallocation at the patient level (i.e., "was this patient treated for the condition they actually have?").

The treatment guidelines we applied are:

- For malaria:

  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:

  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):

  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.4 DALY Harm Scales

To measure healthy time lost translated to DALY, we used a multi-step survey procedure combined with extensive training about how to evaluate harms.

MOs were first guided through a process of detailing qualitative differences between the health worker's note and their own, and then asked to specify the temporary and permanent harms from implementing the health worker's treatment plan rather than their own.

The physicians were 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).

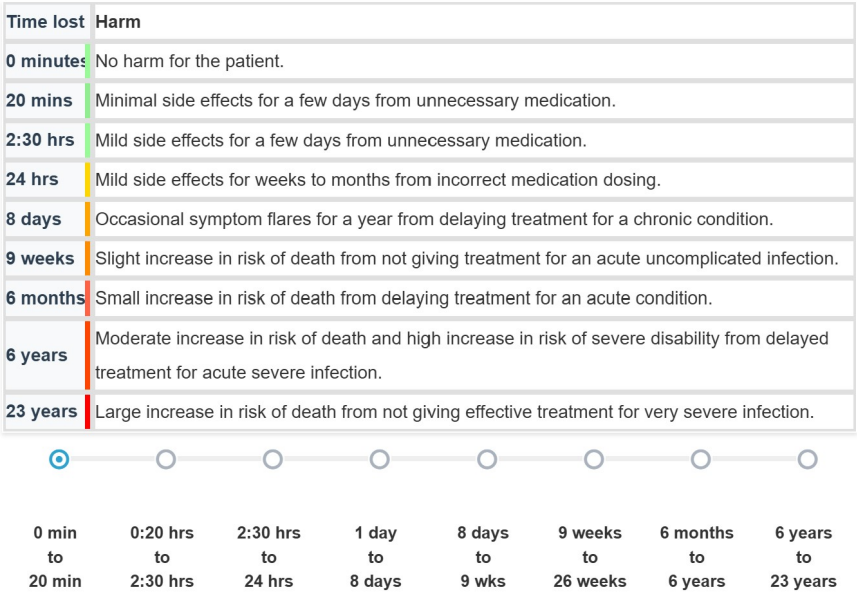
The scale below shows levels of harm that average [adults/children] experience due to different medical errors, ordered from least to most severe.

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 B.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.

Figure B.1: Initial Rank Ordering (Adult Scale)

The physician 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 B.2 for an example.

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 B.10.

Severe 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 at or above the 95th percentile of harm on the unassisted scale (pooling child and adult scales).

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 B.10: Adult and child scale harm benchmarks.

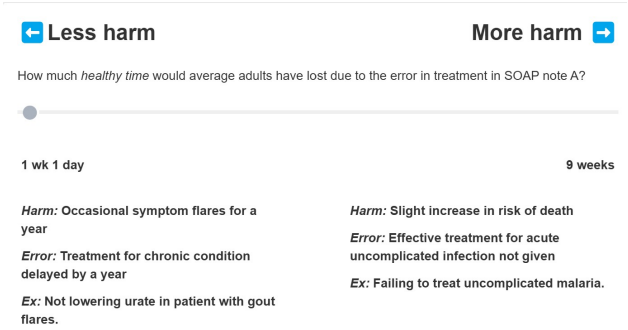


Figure B.2: Example of Healthy Time Loss Quantification

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 strictly 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 physician 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.