

From quality measurement to improvement action: A stakeholder-integrated mixed-methods implementation analysis of adult outpatient care in Kenya

Ashford Mutwiri Gideon¹

Rose Olayo²

Wilberforce Cholo³

¹mutwiriashford@gmail.com (+254 721 330 557)

¹<https://orcid.org/0009-0006-4515-3846>

^{1,2,3} Masinde Muliro University of Science and Technology, Kenya

<https://doi.org/10.51867/ajernet.7.4.90>

ABSTRACT

Quality measurement has limited improvement value when patient, provider and manager evidence is collected without an accountable route to action. This study examined how a Stakeholder-Integrated Quality Measurement-to-Improvement Framework (QMIF) linked role-specific outpatient quality evidence to short-cycle improvement in Kenyan Level IV facilities. A convergent mixed-methods implementation analysis was embedded in a 26-facility non-randomised, non-equivalent pretest-posttest quasi-experiment in Kisumu and Nyeri Counties (13 interventions; 13 comparisons). Baseline was collected on 10-15 April 2026, structured QMIF exposure ran from 15 April to 2 July 2026, and endline was collected on 3-5 July 2026. The longitudinal panel comprised 339 participants (260 patients, 54 providers and 25 managers) and the qualitative strand comprised 84 interviews. Theoretical interpretation was informed by Knowledge-to-Action (KTA), the Consolidated Framework for Implementation Research (CFIR) and RE-AIM. Mixed-effects ANCOVA with a facility random intercept provided baseline-adjusted inferential context for the primary perceived-quality outcome; difference-in-differences (DiD) and Hedges g were supportive. Adjusted effects favoured QMIF for patients (+21.4 percentage-of-maximum-possible [POMP] points; 95% CI 17.8-25.0), providers (+20.4; 16.3-24.6) and managers (+16.1; 11.9-20.4); all $p < .001$. Corresponding Hedges g values were 1.64, 1.87 and 2.13. Knowledge-attitude-practice (KAP) and reported quality-improvement (QI) action were retained as descriptive implementation indicators: intervention-arm endline KAP was 84.3%, 84.8% and 83.7% for patients, providers and managers, while QI-action POMP was 63.0%, 63.8% and 61.7%, respectively. Framework Method analysis and mixed-methods joint displays suggested a recurrent pathway of measurement, feedback, prioritisation, PDSA action and remeasurement. Sixteen of 18 stakeholder-domain combinations met the study's pragmatic criterion, while provider reliability (3.98/6) and manager appropriateness (3.96/6) remained boundary cases; medicines, diagnostics, staffing, equipment and procurement remained structural constraints. The combined evidence supports a plausible implementation pathway, but formal mediation, moderation and fidelity-outcome analyses were not performed. The study concludes that stakeholder-integrated measurement can support short-cycle outpatient QI while unresolved structural constraints require escalation; it recommends replication with stronger fidelity measurement, objective service indicators, external psychometric validation and longer follow-up.

Keywords: Implementation Science; Kenya; Outpatient Care; Mixed Methods; Quality Improvement; Stakeholder Participation

I. INTRODUCTION

High-quality health systems require more than service contact: care must be technically competent, safe, responsive, reliable and capable of learning from performance information (Kruk et al., 2018; World Health Organization [WHO], 2025). Yet measurement and improvement often remain separate activities. Patient-experience measures can reveal access, communication, dignity and continuity problems, while provider and managerial assessments reveal clinical workflow, feasibility, governance and readiness constraints. Measurement alone does not determine which gaps should be prioritised, who should act, whether action changed care, or which problems require escalation beyond the facility.

This implementation gap is particularly important in chronic outpatient care, where patients repeatedly traverse reception, consultation, laboratory, pharmacy, referral and follow-up processes. Sub-Saharan African evidence shows that patient-centred-care measurement remains geographically concentrated and that psychometric assessment is often incomplete (Okeny et al., 2026). In Kenya, patient expectations can exceed perceived service quality and deficits in reliability and empathy remain evident (Karume et al., 2025). National quality systems provide standards, certification, supervision and facility assessment, including the Kenya Quality Model for Health and national facility assessment processes (Ministry of Health, Kenya, 2020, 2024). The practical question is therefore how role-specific quality

evidence can be converted into accountable, short-cycle action without erasing differences between patient experience, provider practice and managerial governance.

The Stakeholder-Integrated Quality Measurement-to-Improvement Framework (QMIF) was developed to address this measurement-to-action problem. It uses six shared domains-accessibility, appropriateness, environment, responsiveness, empathy and reliability-but preserves stakeholder-specific meaning. Patient, provider and manager findings are returned as feedback, jointly prioritised, translated into Plan-Do-Study-Act (PDSA) actions and remeasured. The present article focuses on implementation mechanisms and boundary conditions. At the reviewer's request, it also summarises the parent study's cluster-aware primary effectiveness estimates so that the implementation findings are interpreted against appropriate inferential evidence rather than descriptive percentages alone.

1.1 Statement of the Problem

Kenya has established quality-of-care standards and assessment systems, but persistent implementation gaps can remain when periodic measurement is weakly connected to rapid feedback, local ownership, corrective action and remeasurement. Kenyan longitudinal evidence has shown that repeated measurement and feedback can improve quality and management indicators while fixed inputs such as equipment change more slowly (Marx et al., 2018). More recent Kenyan work has also linked management commitment to service delivery in Level IV hospitals and highlighted the contribution of supplier management, inventory control, policy compliance, coordination and governance to health-supply resilience (Okibo & Wagude, 2026; Owuor et al., 2026).

A second problem is stakeholder fragmentation. Patient, provider and manager perspectives are often measured through separate instruments and accountability structures. Patient involvement can become symbolic when feedback is collected without a visible response, while provider-led QI may under-represent lived experience and managerial constraints (Bergerum et al., 2020; Bombard et al., 2018). Conversely, pooling unequal stakeholder groups into one score can conceal meaningful disagreement about where the care pathway is failing. Evidence is therefore needed on how complementary but non-interchangeable quality signals can be connected operationally in the same improvement cycle and on which contextual constraints remain outside facility-level control.

1.2 Research Objectives

- i. To describe stakeholder-specific KAP and reported QI-action signals associated with implementation;
- ii. explain how patient, provider and manager evidence moved through feedback, joint prioritisation, PDSA action and remeasurement;
- iii. To identify cross-stakeholder convergence, complementarity, useful disagreement and structural constraints that bounded improvement.

II. LITERATURE REVIEW

2.1 Theoretical Review

QMIF was interpreted using three complementary frameworks. Knowledge-to-Action (KTA) specifies movement from evidence to action through problem identification, contextual adaptation, barrier assessment, intervention selection, monitoring and evaluation (Graham et al., 2006). In QMIF, baseline quality signals were synthesised into six domains, returned as facility feedback, prioritised and linked to PDSA action before remeasurement.

The updated Consolidated Framework for Implementation Research (CFIR) was used to interpret contextual determinants rather than as a quantified mediator. Inner-setting factors included leadership, workflow, communication and available resources, while outer-setting factors included county financing, supply systems and referral constraints (Damschroder et al., 2022). RE-AIM bounded claims from a short intervention: reach, short-term effectiveness, adoption and implementation could be described, but maintenance could not be established without longer follow-up (Glasgow et al., 2019). This separation of mechanism, context and maintenance is also consistent with contemporary guidance for complex interventions (Skivington et al., 2021).

2.2 Empirical Review

Evidence from Kenya and the region supports multi-perspective measurement. Karume et al. (2025) documented a negative expectation-perception gap at a Kenyan national referral hospital, with reliability and empathy among the larger deficits. Abuosi (2015) found systematic differences between patient and provider/manager quality ratings in Ghana, demonstrating that disagreement can itself contain useful diagnostic information. Marx et al. (2018) showed that repeated measurement in Kenyan facilities can improve quality and management indicators while fixed equipment-related indicators change more slowly.

Measurement quality is itself a concern. Okeny et al. (2026) found that psychometric evaluation of patient-centred-care instruments in sub-Saharan Africa was frequently incomplete, reinforcing the need to report how context-specific instruments are developed, scored and tested. Kenyan primary-care research also shows that access, continuity,

comprehensiveness, coordination and person-centredness require contextual measurement rather than a single generic quality score (Mohamoud & Mash, 2022). The empirical gap addressed here is therefore not simply a shortage of quality measures; it is limited evidence on how patient, provider and manager signals can be connected to action, how the measurement architecture performs, and where locally modifiable processes meet system-dependent constraints.

III. METHODOLOGY

3.1 Study Design, Setting and Dates

A convergent mixed-methods implementation analysis was embedded in a facility-level, non-randomised, non-equivalent pretest-posttest quasi-experiment with a longitudinal participant panel. The study involved 26 Level IV outpatient facilities in Kisumu and Nyeri Counties: 13 intervention and 13 comparison facilities. Twenty-two facilities were public and four were faith-based/private. Facility allocation used county-stratified geographic-administrative separation rather than randomisation or formal matching; Kisumu contributed eight intervention and eight comparison facilities and Nyeri five per arm. Baseline data were collected from 10-15 April 2026. Baseline records were verified and locked before structured QMIF content was released on 15 April. The intervention ran from 15 April to 2 July 2026, and endline data were collected from 3-5 July 2026. Reporting was informed by TREND and SQUIRE 2.0 (Des Jarlais et al., 2004; Ogrinc et al., 2016).

3.2 Target Population, Eligibility, Sample Size and Recruitment

The target population comprised adults receiving longitudinal chronic outpatient care, frontline healthcare providers and facility managers. Patients were eligible if they were aged 18 years or older, were receiving outpatient care for a chronic communicable or non-communicable condition, could provide informed consent, and had documented repeat attendance at the same facility during the preceding three months. Patients younger than 18 years, without the required chronic-care continuity, unable or unwilling to consent, or admitted as inpatients during the study period were excluded. Eligible providers directly delivered outpatient/chronic care, had at least one year of service and consented. Eligible managers had responsibility for outpatient organisation, supervision, QI, readiness or governance, had at least one year of service and consented.

Patients were recruited using systematic clinic-flow recruitment with interval $k=2$: from a fixed starting position, every second eligible attendee was approached (positions 1, 3, 5, 7, and so forth). This is described as systematic clinic-flow recruitment rather than random-start systematic sampling. Providers were recruited from eligible staff frames, and managers were approached by near-census/complete enumeration of the defined managerial position because of their small number and governance role. Qualitative patient interviews used criterion-based maximum variation across county, arm, facility and clinic experience; provider and manager interviews used purposive key-informant sampling.

Sample-size planning fixed 26 facility clusters (13 per arm), used two-sided $\alpha=.05$, 80% power, a planning intraclass correlation coefficient (ICC) of .02 and a target standardised effect of approximately $d=.35$. The fixed-cluster calculation required 340 analysable participants (about 13.1 per facility), inflated by 10% for anticipated non-completion to a recruitment target of 378. This was an aggregate recruitment-planning calculation, not a pooled patient-provider-manager inferential sample; stakeholder-specific effects were estimated separately. Because the design involved a limited number of clusters, cluster-aware modelling and emphasis on effect magnitude and confidence intervals were used in interpretation (Leyrat et al., 2018).

Table 1

Quantitative sample accountability and analytic retention

Stakeholder	Planned recruit	Screened / captured	Excluded	Incomplete	Lost to follow-up	Final analytic n (I/C)	Retention among eligible screened*
Patients	274	280	7	5	8	260 (123/137)	95.2%
Providers	78	66	8	3	1	54 (30/24)	93.1%
Managers	26 (complete enumeration)	32	5	1	1	25 (13/12)	92.6%
Role unclassified	N/A	4	0	4	0	0	N/A
Total	378	382	20	13	10	339 (166/173)	93.6%

*Retention among eligible screened = final analytic / (screened minus exclusions). This is an analytic-retention measure, not a population survey response rate. I/C = intervention/comparison.

3.3 QMIF Intervention and Fidelity Monitoring

QMIF lasted 12 weeks and linked role-specific measurement to action. The first two weeks comprised orientation, role-specific measurement and feedback; eight weeks comprised serial PDSA/action cycles; and the final two weeks comprised evaluation, post-intervention KAP assessment, remeasurement and sustainment review. The intervention used a 12-item six-domain quality instrument: accessibility (delays and barriers), appropriateness (comprehensive and individualised care), environment (risk of harm and functional amenities), responsiveness (complaints handling and follow-up), empathy (compassion and respect), and reliability (consistency and perceived best-care outcomes). Patients supplied experiential signals, providers interpreted clinical/workflow feasibility, and managers interpreted governance, readiness and escalation needs.

Intervention facilities received rapid facility-level feedback, selected a limited number of priorities jointly, assigned actions and used PDSA cycles for eight weeks before remeasurement. A standard action week planned one 60-minute facilitated huddle, five approximately 10-minute micro-action/log entries and one 20-minute review (about 130 minutes distributed across the planned dose-days). Comparison facilities continued routine Kenya Quality Model for Health-aligned QI without the structured QMIF sequence before endline.

Fidelity was monitored prospectively using dated attendance/minute records, competency and PDSA worksheets, action/process logs, adaptation/deviation records, REDCap timestamps and data locks. Field supervisors reviewed delivery, reach, dose, adaptations and deviations during implementation, with an explicit fidelity audit in the final intervention period before endline. All 26 facilities remained represented in the analytic dataset and no suspected or confirmed cross-arm contamination event was recorded. However, the implementation records were not collapsed into a validated composite fidelity scale or a prespecified facility-level adherence percentage. Accordingly, this article reports documentary fidelity evidence and protocol dose descriptively and does not claim a quantified fidelity-outcome relationship.

3.4 Measures, Instrument Development, Scoring and Psychometric Performance

Primary data sources were baseline and endline stakeholder questionnaires, patient in-depth interviews, provider and manager key-informant interviews, facility feedback records, and implementation/PDSA documentation. Secondary sources were peer-reviewed literature, national policy/quality documents and methodological guidance used only for contextual interpretation; no secondary source contributed participant outcome data.

Instrument development followed the QMIF conceptual model. The 12 perceived-quality items were explicitly mapped to the six prespecified domains, reviewed against established patient-experience, service-quality and facility-readiness constructs, reviewed by supervisors/subject experts, and pilot-tested before the main study. A pilot in two non-study Level IV facilities included 60 responses (38 patients and 22 providers/managers) and tested recruitment flow, consent, REDCap logic, item interpretation, administration burden, internal consistency and operational feasibility. Pilot records were excluded from effectiveness analyses. The pilot yielded Cronbach alpha=.74 for baseline perceived quality, .93 for endline perceived quality and .93 for the QI-action composite after refinement.

In the main sample, the perceived-quality instrument comprised twelve six-point items. Delays, barriers and risk-of-harm items were reverse-scored using 7-response so that higher values consistently indicated better perceived quality. The overall score was the mean of 12 items (1-6), equivalent to a 12-72 sum; each domain was the mean of its two items. Percentage of maximum possible (POMP) was calculated as $100 \times (\text{observed score} - \text{scale minimum}) / (\text{scale maximum} - \text{scale minimum})$ (Cohen et al., 1999). A study-defined 60% POMP benchmark corresponded to a mean of 4/6 or sum of 48/72. It was a contextual pragmatic benchmark, not a validated minimal clinically important difference or clinical cut-off.

The 12-item reported QI-action index used the same 1-6 response range and was summarised as a mean/sum and POMP. The 10-item KAP process score also used six-point responses; negatively keyed items 2, 3, 4 and 10 were reverse-scored, producing a total range of 10-60. KAP was expressed as percentage of the maximum possible score ($\text{total}/60 \times 100$); $\geq 80\%$ was an a priori descriptive attainment benchmark. Because KAP intentionally combined knowledge, attitude and practice content, it was treated as a multidimensional process score rather than assumed to be a unidimensional psychometric scale. Neither the 80% KAP nor 60% QI benchmark is presented as a validated clinical threshold.

Table 2*Measurement, scoring and psychometric evidence*

Measure	Structure and scoring	Reliability / validation evidence	Interpretive limit
Perceived outpatient quality	12 items; 6 domains; 1-6 mean (or 12-72 sum); delay, barrier and harm items reverse-scored; POMP = $100 \times (\text{score-min})/(\text{max-min})$.	Pilot alpha .74 baseline and .93 endline; main-study alpha .80 baseline and .94 endline. Content mapping, expert/supervisor review and non-study pilot.	60% POMP (4/6; 48/72) is a pragmatic contextual benchmark, not an MCID or clinical cut-off.
Reported action	12 items; 1-6 mean (or 12-72 sum); converted to POMP.	Pilot alpha .93; main-study alpha .94 baseline and .95 endline.	Implementation/process indicator; not a mediator proven to cause the primary outcome.
KAP process score	10 items; six-point responses; items 2, 3, 4 and 10 reverse-scored; total 10-60; percent of maximum = $\text{total}/60 \times 100$.	Designed as heterogeneous knowledge-attitude-practice content; no unidimensional alpha claim is made.	$\geq 80\%$ is an a priori descriptive benchmark, not a validated QMIF clinical cut-off.

3.5 Quantitative Analysis

The parent study's primary estimand for each stakeholder group was the baseline-adjusted endline intervention-comparison difference in perceived quality. Mixed-effects ANCOVA included study arm, baseline outcome, post-secondary education and county, with a facility random intercept. A mixed-effects DiD model using arm-by-post interaction was supportive. Because there was only one pre-intervention wave, parallel trends could not be tested; DiD was therefore corroborative rather than causal proof. Domain ANCOVAs used Holm adjustment within each six-domain family. Effect estimates, 95% confidence intervals, Hedges g , POMP and cluster-aware benchmark attainment were interpreted together (Hedges et al., 2023).

KAP and QI-action values in this implementation article are process indicators and were prespecified for descriptive implementation/readiness interpretation rather than as primary causal endpoints. No retrospective significance tests were introduced for these percentages solely in response to peer review. Instead, cluster-aware inferential evidence is supplied from the parent primary perceived-quality outcome, which directly addresses whether the intervention and comparison arms differed after baseline adjustment. This distinction avoids presenting post hoc participant-level tests as if they had been prespecified.

3.6 Qualitative Analysis and Mixed-Methods Integration

The qualitative corpus comprised 84 interviews: 62 patient in-depth interviews, eight provider key-informant interviews and 14 manager key-informant interviews. Transcripts were analysed using the Framework Method (Gale et al., 2013): familiarisation, line-by-line coding, development of a hybrid deductive-inductive analytic framework, application to transcripts, charting into case-by-code matrices and interpretation. Purposive sampling sought information adequacy across all six domains and stakeholder roles and continued until no materially new objective-level explanation was emerging; this was an information-rich implementation sample rather than a claim of statistical representativeness.

The coding tree retained six deductive parent domains with inductive subcodes: accessibility (waits/queues, congestion, navigation, affordability, financing-system downtime and medicine access); appropriateness (history/diagnosis, explanations, guidelines/SOPs, tests, referral and organisation); environment (cleanliness, privacy, crowding, equipment readiness and movement); responsiveness (listening/explanations, patient voice, complaints/feedback, follow-up, completion and PDSA/data loops); empathy (dignity/courtesy, communication, non-discrimination, harsh speech and relational warmth); and reliability (medicine/diagnostic availability, correct results, essential services, turnaround and dependable completion). Negative and structurally constrained cases were retained rather than coded away.

The study did not generate a formal independent double-coding coefficient such as Cohen's kappa; none is therefore reported. Credibility relied on an explicit codebook, source-linked case-by-code matrices, analytic memos, retention of negative/deviant cases and checking selected quotations against the source transcripts/translations. A formal reflexivity diary was not prospectively specified, so researcher involvement in the intervention evaluation and participants' awareness of study context remain potential sources of interpretive and social-desirability bias.

Quantitative and qualitative strands were integrated after separate stakeholder-specific analyses using joint-display matrices (Fetters et al., 2013). Convergence was recorded when quantitative direction and qualitative accounts supported the same inference; complementarity/expansion when qualitative evidence explained how or why a quantitative pattern arose; and divergence/useful disagreement when a near-threshold, negative or role-specific finding contradicted or bounded the broader pattern. Patients, providers and managers were not pooled because their roles, sample sizes and construct meanings differed.

3.7 Ethical Considerations

The study received Masinde Muliro University of Science and Technology study approval (MMU/COR: 509099; 6 March 2026), Institutional Ethics Review Committee approval (MMU/COR: 40312 Vol 6(01); 9 March 2026), National Commission for Science, Technology and Innovation research licence 633928 (8 April 2026), and county approvals from Kisumu (CGK/COII/10/VOL.I/41; 13 April 2026) and Nyeri (CGN/HEALTH/HRM/5/VOL.X; 20 April 2026). All participants provided informed consent before study procedures. Individual responses were coded and were not shared with facility leadership.

IV. FINDINGS & DISCUSSION

4.1 Participant Flow and Baseline Context

Of 382 screened/captured records, 20 were excluded, 13 were incomplete and 10 participants were lost to follow-up, leaving 339 complete linked records: 260 patients, 54 providers and 25 managers (Table 1). All 26 facilities were represented in the overall analytic dataset. Measured baseline imbalance was present, including patient baseline perceived quality (44.72 versus 42.08/72; standardised mean difference +0.55), provider baseline KAP/QI activity and manager initial expectations. Consequently, unadjusted endline comparisons were not treated as primary evidence; baseline-adjusted mixed-effects estimates were used.

4.2 Quantitative Outcomes and Implementation Signals

For patients, intervention-arm perceived-quality POMP increased from 54.5% to 67.5%, while comparison POMP decreased from 50.1% to 46.2%. The baseline-adjusted ANCOVA effect was +21.4 POMP percentage points (95% CI 17.8-25.0; $p < .001$), the supportive DiD was +16.9 (14.6-19.2; $p < .001$), and Hedges g was 1.64 (1.36-1.93). The cluster-aware absolute difference in attaining the 60% benchmark was +82.0 percentage points (68.3-95.8; $p < .001$). For providers, intervention POMP increased from 51.7% to 67.6%, while comparison POMP decreased from 54.9% to 47.0%. The adjusted effect was +20.4 POMP points (16.3-24.6; $p < .001$), supportive DiD +23.8 (17.4-30.2; $p < .001$), and Hedges g 1.87 (1.23-2.52). The cluster-aware benchmark-attainment difference was +90.0 percentage points (73.3-100.0; $p < .001$).

For managers, intervention POMP increased from 53.7% to 64.6% and comparison POMP decreased from 56.8% to 48.8%. The adjusted effect was +16.1 POMP points (11.9-20.4; $p < .001$), supportive DiD +19.0 (12.2-25.7; $p < .001$), and Hedges g 2.13 (1.13-3.13). The benchmark-attainment difference was +92.3 percentage points (76.9-100.0; $p < .001$). Because 25 managers represented 25 facilities, respondent-level and facility-level variation were nearly coincident and a separate manager ICC was not interpreted as a stable clustering parameter.

The adjusted outcome models retained meaningful facility-level clustering for patients (ICC approximately .36) and providers (approximately .32), substantially above the .02 planning assumption. This empirical clustering supports facility-aware analysis and also reinforces caution with participant-level inference from only 26 clusters. Accordingly, effect magnitude and confidence intervals are emphasised alongside p -values (Leyrat et al., 2018). Sixteen of 18 stakeholder-domain combinations met the study's pragmatic criterion; the exceptions were provider reliability (3.98/6) and manager appropriateness (3.96/6).

Table 3

Cluster-aware quantitative summary and descriptive implementation indicators

Stakeholder	Primary ANCOVA, POMP pp (95% CI)	Supportive DiD, POMP pp (95% CI)	Hedges g (95% CI)	Empirical outcome ICC	Endline KAP I/C	Endline QI-action POMP I/C
Patients	+21.4 (17.8-25.0), $p < .001$	+16.9 (14.6-19.2), $p < .001$	1.64 (1.36-1.93)	~.36	84.3% / 60.3%	63.0% / 39.2%
Providers	+20.4 (16.3-24.6), $p < .001$	+23.8 (17.4-30.2), $p < .001$	1.87 (1.23-2.52)	~.32	84.8% / 58.8%	63.8% / 39.5%
Managers	+16.1 (11.9-20.4), $p < .001$	+19.0 (12.2-25.7), $p < .001$	2.13 (1.13-3.13)	Not separately interpretable*	83.7% / 57.4%	61.7% / 39.0%

I/C = intervention/comparison; pp = percentage points. KAP and QI-action columns are descriptive implementation indicators and were not retrospectively converted into primary causal tests. *Twenty-five managers were distributed across 25 facilities, making respondent and facility levels nearly coincident.

The descriptive process indicators were directionally coherent with the primary outcome but should not be over-interpreted. At endline, intervention-arm KAP exceeded the prespecified 80% benchmark in all three groups (84.3% patients, 84.8% providers and 83.7% managers), whereas comparison values were 60.3%, 58.8% and 57.4%. Reported QI-action POMP was 63.0%, 63.8% and 61.7% in intervention participants versus 39.2%, 39.5% and 39.0% in

comparison participants. These values are consistent with activation of knowledge and action processes, but they do not prove that training, KAP change or QI activity mediated the adjusted quality effects.

4.3 Qualitative Explanations and Mixed-Methods Integration

Patient accounts located change in service flow, follow-up, respectful communication and dignity while also preserving negative cases. One patient reported, "We are treated respectfully and with dignity" (Patient IDI001-N; translated from Kikuyu), whereas another described "*Many delays and congestion*" (Patient IDI002-N; translated from Kikuyu). This combination of positive and constrained accounts prevented the analysis from treating improvement as uniform.

Providers translated the shared domains into history-taking, guideline use, prescribing, referral, audit, continuing professional learning, teamwork and turnaround-time review. A provider stated, "*We should be providing the essential services and reducing turnaround*" (Provider KII003-K). The near-threshold provider reliability score of 3.98/6 was therefore interpreted as useful disagreement: process change could occur while medicines, diagnostics, records, staffing and referral continuity still limited dependable completion of care.

Managers interpreted the same domains through capacity, staffing, service scope, infection prevention, equipment, medicines, complaint closure, supervision and audit. A manager noted, "*We have a QI team in the facility*" (Manager KII007-K), illustrating the governance infrastructure through which feedback could be assigned and monitored. Manager appropriateness remained at 3.96/6, however, showing that a facility could identify and escalate a diagnostic, staffing or referral gap without being able to resolve it within a short PDSA cycle.

Table 4

Mixed-methods joint display: quantitative signal, qualitative explanation and integration

Stakeholder	Quantitative signal	Qualitative evidence / mechanism	Integration judgement	Boundary condition
Patients	ANCOVA +21.4 POMP pp (95% CI 17.8-25.0); 6/6 domains met.	"We are treated respectfully and with dignity" contrasted with "Many delays and congestion"; feedback became meaningful when patients encountered changed processes.	Convergence with expansion: adjusted improvement plus experience-level explanation, with retained negative cases.	Queues, affordability, medicine/diagnostic availability and follow-up continuity.
Providers	ANCOVA +20.4 POMP pp (16.3-24.6); 5/6 domains met; reliability 3.98/6.	"We should be providing the essential services and reducing turnaround"; reports linked data to audit, teamwork and workflow/PDSA change.	Complementarity and useful disagreement: process activation coexisted with sub-threshold reliability.	Staffing, medicines, diagnostics, records and referral continuity.
Managers	ANCOVA +16.1 POMP pp (11.9-20.4); 5/6 domains met; appropriateness 3.96/6.	"We have a QI team in the facility"; managers described assigning, monitoring and escalating actions.	Complementarity and useful disagreement: stronger governance did not guarantee resolution of resource-dependent gaps.	Budgets, procurement, diagnostics, equipment, skill mix and referral capacity.

Across the three stakeholder groups, the joint displays identified a repeated sequence of measurement, feedback, prioritisation, PDSA action and remeasurement. This is best interpreted as a plausible implementation pathway rather than demonstrated causal mediation. The quantitative primary effects establish an intervention-associated difference in perceived quality after baseline adjustment; the qualitative findings explain how participants understood the change process; and the descriptive KAP/QI patterns provide supporting process evidence. Formal mediation or moderation models were not fitted, so the pathway remains a mechanistic hypothesis for future testing.

4.3.1 Fidelity, Structural Constraints and Boundary Conditions

Documentary fidelity evidence showed that implementation was actively monitored through attendance/minute records, PDSA worksheets, action logs, adaptation/deviation records, REDCap timestamps and supervisory review. No suspected or confirmed contamination event was recorded and all facilities remained represented at analysis. However, the absence of a validated composite fidelity score means that variation in delivered dose cannot be quantified consistently across facilities and cannot be tested as an explanation for variation in outcome. This limitation is material: a complex intervention can appear ineffective because of poor delivery or effective because high-performing sites deliver more completely, and the present study cannot separate those possibilities empirically.

The most consistent boundary was a process-structure gradient. Relational and workflow changes-respectful communication, feedback, audit, queue review, teamwork and action assignment-were more locally modifiable within 12 weeks. Medicines, diagnostics, staffing, equipment, procurement, financing and referral capacity depended more

heavily on wider systems. This pattern is consistent with Marx et al. (2018), who observed that management and quality indicators could improve while fixed equipment changed more slowly, and with broader CQI evidence identifying resources and organisational context as important implementation determinants (Endalamaw et al., 2024). Rather than treating unresolved structural gaps as frontline PDSA failure, QMIF therefore makes escalation an explicit endpoint.

4.4 Interpretation in the Wider Literature

The findings support the value of preserving stakeholder role specificity. Abuosi (2015) similarly found that patient and provider/manager ratings of quality can differ substantially; such disagreement should not automatically be averaged away. In QMIF, a delay represented a lived burden to a patient, a workflow problem to a provider and a capacity signal to a manager. Integration therefore occurred at the level of coordinated action and interpretation rather than by pooling unequal observations into one inferential score.

The results also strengthen the measurement-to-action argument. Patient involvement literature warns that participation can become tokenistic when voice is collected but not linked to decisions or visible response (Bergerum et al., 2020; Bombard et al., 2018). Here, qualitative accounts located value in the return of local data, joint prioritisation, assigned responsibility and remeasurement. KTA provides a coherent conceptual map for that sequence, while CFIR helps distinguish modifiable inner-setting processes from resource and policy constraints, and RE-AIM prevents 12-week implementation from being mislabeled as maintenance.

V. CONCLUSION & RECOMMENDATIONS

5.1 Conclusions

In this controlled mixed-methods evaluation, QMIF exposure was associated with large baseline-adjusted short-term improvements in stakeholder-perceived outpatient quality across patients, providers and managers. The inferential evidence is supported by concordant DiD estimates and large standardised effect magnitudes, while the qualitative and descriptive implementation evidence suggests a plausible measurement-feedback-prioritisation-PDSA-remeasurement pathway. That pathway should not be interpreted as proven mediation or moderation: the study did not formally test whether KAP, QI activity or fidelity caused the observed quality effects.

The most transferable finding is therefore not a claim that a fixed 12-item scorecard itself produces improvement. It is the implementation architecture of preserving role-specific evidence, returning it rapidly, comparing convergence and disagreement, assigning locally feasible actions, remeasuring, and escalating resource-dependent gaps. Persistent medicines, diagnostics, staffing, equipment, procurement and referral constraints demonstrate that short-cycle facility QI must be paired with system-level escalation. Provider reliability and manager appropriateness should be viewed as informative boundary cases rather than failures to fit an overall positive narrative.

5.2 Recommendations

Facility and county quality teams should pilot recurrent QMIF-style cycles within existing Kenyan quality structures: collect role-specific quality signals, return findings rapidly, jointly prioritise a small set of actions, assign responsibility, test changes using PDSA, remeasure and explicitly escalate unresolved structural gaps. Future evaluations should use cluster-randomised or stepped-wedge designs where feasible, collect multiple pre-intervention observations, and add objective indicators such as waiting time, medicine availability, diagnostic completion, referral completion and complaint-closure time.

Future measurement work should externally validate the six-domain instrument using factor-analytic and measurement-invariance methods before cross-stakeholder pooling is considered. Implementation studies should prospectively quantify fidelity with a validated or clearly prespecified dose/adherence framework, use independent qualitative coding or auditable consensus coding, document reflexivity, and test whether feedback exposure, fidelity, KAP change or QI action formally mediate or moderate outcomes. Longer follow-up of at least 6-12 months is needed to assess maintenance and institutionalisation.

Declaration of Interest

The authors declare that they do not have any known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

- Abuosi, A. A. (2015). Patients versus healthcare providers' perceptions of quality of care: Establishing the gaps for policy action. *Clinical Governance: An International Journal*, 20(4), 170–182. <https://doi.org/10.1108/CGIJ-03-2015-0010>
- Bergerum, C., Engström, A. K., Thor, J., & Wolmesjö, M. (2020). Patient involvement in quality improvement—a “tug of war” or a dialogue in a learning process to improve healthcare? *BMC Health Services Research*, 20, Article 1115. <https://doi.org/10.1186/s12913-020-05970-4>
- Bombard, Y., Baker, G. R., Orlando, E., et al. (2018). Engaging patients to improve quality of care: A systematic review. *Implementation Science*, 13, Article 98. <https://doi.org/10.1186/s13012-018-0784-z>
- Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, 34(3), 315–346. https://doi.org/10.1207/S15327906MBR3403_2
- Damschroder, L. J., Reardon, C. M., Opra Widerquist, M. A., & Lowery, J. (2022). The updated Consolidated Framework for Implementation Research based on user feedback. *Implementation Science*, 17, Article 75. <https://doi.org/10.1186/s13012-022-01245-0>
- Des Jarlais, D. C., Lyles, C., Crepaz, N., & TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health*, 94(3), 361–366. <https://doi.org/10.2105/AJPH.94.3.361>
- Donabedian, A. (1988). The quality of care: How can it be assessed? *JAMA*, 260(12), 1743–1748. <https://doi.org/10.1001/jama.1988.03410120089033>
- Endalamaw, A., Khatri, R. B., Mengistu, T. S., Erku, D., Wolka, E., Zewdie, A., & Assefa, Y. (2024). A scoping review of continuous quality improvement in healthcare system: Conceptualization, models and tools, barriers and facilitators, and impact. *BMC Health Services Research*, 24, Article 487. <https://doi.org/10.1186/s12913-024-10828-0>
- Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—principles and practices. *Health Services Research*, 48(6 Pt 2), 2134–2156. <https://doi.org/10.1111/1475-6773.12117>
- Gale, N. K., Heath, G., Cameron, E., Rashid, S., & Redwood, S. (2013). Using the Framework Method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology*, 13, Article 117. <https://doi.org/10.1186/1471-2288-13-117>
- Glasgow, R. E., Harden, S. M., Gaglio, B., et al. (2019). RE-AIM planning and evaluation framework: Adapting to new science and practice with a 20-year review. *Frontiers in Public Health*, 7, Article 64. <https://doi.org/10.3389/fpubh.2019.00064>
- Graham, I. D., Logan, J., Harrison, M. B., et al. (2006). Lost in knowledge translation: Time for a map? *Journal of Continuing Education in the Health Professions*, 26(1), 13–24. <https://doi.org/10.1002/chp.47>
- Hedges, L. V., Tipton, E., Zeinullahi, R., & Diaz, K. G. (2023). Effect sizes in ANCOVA and difference-in-differences designs. *British Journal of Mathematical and Statistical Psychology*, 76(2), 259–282. <https://doi.org/10.1111/bmsp.12296>
- Karume, A. K., Nyongesa, K., Okutoyi, L., & Kinuthia, J. (2025). Patient's expectations and perceptions on quality of care: An evaluation using SERVQUAL gap in Kenya. *PLOS ONE*, 20(3), Article e0315910. <https://doi.org/10.1371/journal.pone.0315910>
- Kruk, M. E., Gage, A. D., Arsenault, C., et al. (2018). High-quality health systems in the Sustainable Development Goals era: Time for a revolution. *The Lancet Global Health*, 6(11), e1196–e1252. [https://doi.org/10.1016/S2214-109X\(18\)30386-3](https://doi.org/10.1016/S2214-109X(18)30386-3)
- Leyrat, C., Morgan, K. E., Leurent, B., & Kahan, B. C. (2018). Cluster randomized trials with a small number of clusters: Which analyses should be used? *International Journal of Epidemiology*, 47(1), 321–331. <https://doi.org/10.1093/ije/dyx169>
- Marx, M., Nitschke, C., Nafula, M., Nangami, M., Brodowski, M., Marx, I., Prytherch, H., Kandie, C., Omogi, I., Paul-Fariborz, F., & Szecsenyi, J. (2018). If you can't measure it—you can't change it: A longitudinal study on improving quality of care in hospitals and health centers in rural Kenya. *BMC Health Services Research*, 18, Article 246. <https://doi.org/10.1186/s12913-018-3052-7>
- Ministry of Health, Kenya. (2020). *Quality of care certification manual for the Kenyan health sector*. <https://medbox.org/document/quality-of-care-certification-manual-for-the-kenyan-health-sector>
- Ministry of Health, Kenya. (2024). *2024 Kenya health facility assessment: Quality of care and human resources for health report*. World Health Organization Regional Office for Africa. <https://www.afro.who.int/countries/kenya/publication/024-kenya-health-facility-assessment-quality-care-and-human-resources-health-report>
- Mohamoud, G., & Mash, R. (2022). The quality of primary care performance in private sector facilities in Nairobi, Kenya: A cross-sectional descriptive survey. *BMC Primary Care*, 23, Article 120. <https://doi.org/10.1186/s12875-022-01700-3>



- Ogrinc, G., Davies, L., Goodman, D., Batalden, P., Davidoff, F., & Stevens, D. (2016). SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): Revised publication guidelines from a detailed consensus process. *BMJ Quality & Safety*, 25(12), 986–992. <https://doi.org/10.1136/bmjqs-2015-004411>
- Okeny, P. K., Pittalis, C., Brugha, R., & Gajewski, J. (2026). Patient-centred care measurement in sub-Saharan Africa: A scoping review of the evidence (1990–2024). *International Health*, 18(4), 570–590. <https://doi.org/10.1093/inthealth/ihaf139>
- Okibo, S. S., & Wagude, J. (2026). Influence of management commitment to occupational safety and health-OSH on service delivery in level 4 public hospitals in Kisii County, Kenya. *African Journal of Empirical Research*, 7(2), 287–299. <https://doi.org/10.51867/ajernet.7.2.27>
- Owuor, V. A., Gachanja, P. M., & Nzau, M. (2026). Enhancing health security system preparedness: The impact of operational practices on medical supply chain resilience in Kenya. *African Journal of Empirical Research*, 7(3), 661–672. <https://doi.org/10.51867/ajernet.7.3.55>
- Skivington, K., Matthews, L., Simpson, S. A., Craig, P., Baird, J., Blazeby, J. M., et al. (2021). A new framework for developing and evaluating complex interventions: Update of Medical Research Council guidance. *BMJ*, 374, Article n2061. <https://doi.org/10.1136/bmj.n2061>
- Taylor, M. J., McNicholas, C., Nicolay, C., Darzi, A., Bell, D., & Reed, J. E. (2014). Systematic review of the application of the plan–do–study–act method to improve quality in healthcare. *BMJ Quality & Safety*, 23(4), 290–298. <https://doi.org/10.1136/bmjqs-2013-001862>
- World Health Organization. (2025, May 19). *Quality health services*. <https://www.who.int/news-room/fact-sheets/detail/quality-health-services>