Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Next-Year Preventive Chemotherapy Delivery Shortfalls for Four Neglected Tropical Diseases in West Africa: A Retrospective Longitudinal Prediction Study

Mumini Idowu Adarabioyo

adarabioyomi@abuad.edu.ng

Afe Babalola University

Juliet Nnenda Olayinka

Afe Babalola University

Research Article

Keywords: neglected tropical diseases, preventive chemotherapy, mass drug administration, prediction, temporal validation, calibration, decision curve analysis, West Africa

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11516211/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Whether longitudinal implementation histories improve next-year preventive chemotherapy warnings beyond coverage, previous failure and contextual information remains uncertain. We examined incremental discrimination, calibration and decision value across evaluation periods.

Methods

Routine records for four neglected tropical diseases in eight West African countries were organized into consecutive annual transitions. Development used 2015–2022 outcomes and validation used 2023–2024. Ghana's unavailable 2025 unit-level outcomes were excluded from definitive temporal testing, retaining its earlier data. The final test comprised seven countries. Five reproducible logistic specifications compared coverage and previous-failure benchmarks with contextual and longitudinal extensions. These comparisons were specified after examination of the original outcomes; original saved models were supplementary. Confidence intervals used 1,000 paired implementation-unit bootstrap resamples stratified by country.

Results

The definitive test contained 1,839 transitions and 851 failures (46.28%). The context-aware benchmark achieved an area under the receiver operating characteristic curve of 0.8179 (95% confidence interval 0.8014–0.8333), precision–recall area of 0.7797 (0.7499–0.8084), and Brier score of 0.1796. The extended longitudinal model achieved 0.7999, 0.7471 and 0.1812, respectively. Its net-benefit increment at the retained threshold of 0.40 was 0.00634 (0.00017–0.01266), but its advantage was absent in 2023–2024. Both models overpredicted aggregate 2025 risk. No specification dominated across thresholds and periods.

Conclusions

Longitudinal information did not consistently improve prediction beyond a context-aware benchmark. Small threshold-specific gains should be considered alongside weaker risk ranking, temporal instability and source completeness. Programme early-warning research requires strong operational comparators and explicit verification that zero-valued records represent observed outcomes rather than unavailable reporting.

Trial registration

Not applicable. This retrospective study did not assign health interventions to human participants.

Background

Preventive chemotherapy is a central strategy for controlling several neglected tropical diseases through repeated treatment of eligible populations. Its implementation depends on sustained delivery across successive cycles. Routine coverage records describe completed activities, whereas programme planning also requires an assessment of where delivery may fall short in the next cycle. WHO guidance emphasizes the quality of reported treatment data and the need to examine coverage estimates critically [1]. ESPEN provides access to implementation information submitted through country reporting systems [2].

The question addressed here is whether historical implementation information adds useful warning beyond simple operational knowledge. A model may appear valuable when compared with coverage alone, yet add little to a benchmark that also includes previous failure and country and disease context. Better ranking of risk also need not imply better probability estimates or more useful prioritization at a specified threshold. These distinctions matter when prediction is proposed to guide preventive chemotherapy support.

The study evaluates next-year recorded delivery shortfalls across four diseases and eight countries, using reproducible operational benchmarks and longitudinal extensions, with original archived models evaluated as supplementary evidence. The specific objectives are to quantify incremental predictive value, examine calibration and threshold-dependent decision benefit, and assess their stability across evaluation periods and national settings. The contribution concerns the empirical conditions under which longitudinal information is useful, rather than novelty of the individual algorithms. Transparent reporting and risk-of-bias considerations draw on prediction-model guidance, without treating this programme-level study as an individual clinical prediction application [3, 4].

Related work

Longitudinal coverage and treatment histories

Existing research has already examined treatment performance longitudinally. Fuller and colleagues analysed 3,880 lymphatic filariasis campaigns across ten West African countries over fourteen years, using hierarchical regression to examine contextual determinants of coverage [5]. Their study establishes a clear precedent for regional longitudinal implementation analysis. The present study therefore does not claim novelty for using repeated observations, modelling coverage or comparing countries.

Related research addresses different levels of the delivery process. Maddren and colleagues reviewed individual longitudinal compliance with mass drug administration and documented limited individually linked evidence and inconsistent terminology [6]. In Nigeria, Imhansoloeva and colleagues evaluated a DHIS2 platform for real-time treatment reporting, illustrating how digital monitoring can support decisions

during campaigns [7]. Programme delivery continuity at implementation-unit level differs from individual receipt or swallowing of medicines, and next-year prediction differs from monitoring activities already underway.

Hardwick and colleagues developed probability models of individual adherence conditional on past behaviour and applied them to TUMIKIA trial data in Kenya [8]. Their analysis demonstrates that repeated treatment behaviour already has a substantive modelling literature. Its individual-level adherence endpoint and transmission simulations differ from prediction of implementation-unit delivery shortfalls. Aggregate treatment continuity in the present study cannot recover individual adherence patterns.

Means and colleagues examined six community-wide deworming rounds delivered during 2018–2020 in Benin, India and Malawi [9]. They analysed cluster-level coverage using generalized estimating equations and individual non-treatment using mixed-effects models. This work provides evidence on repeated-round delivery and contextual variation. The present study instead evaluates whether annual histories improve later-year alerts beyond routinely available coverage, previous failure and country and disease indicators.

Routine reporting and infectious disease monitoring

Namara and colleagues used ESPEN treatment records and Uganda DHIS2 case reports to examine MDA coverage, populations at risk and reported schistosomiasis and soil-transmitted helminth cases [10]. Their BMC Infectious Diseases study used Mann–Kendall trend analyses. It establishes a directly relevant precedent for routine-data research on preventive chemotherapy in this journal. The present endpoint concerns next-year recorded delivery shortfalls; infection incidence and intervention impact are not estimated.

Zoerhoff and colleagues compared routinely reported and survey-estimated coverage for 214 MDA campaigns in 15 countries during 2008–2017 [11]. Agreement on whether the disease-specific threshold was reached was 72% in African campaigns and 52% in Asian campaigns. Reporting disagreement differs from the unavailable Ghana 2025 outcomes identified here, but both require caution when interpreting administrative indicators. Predictions of recorded delivery shortfalls should not be equated with independently verified treatment receipt.

Specific contribution and unresolved question

Taken together, these studies establish longitudinal coverage analysis, dependence on treatment history, contextual heterogeneity, digital monitoring and reporting uncertainty. The unresolved question examined here is narrower: within this multicountry four-disease dataset, do histories improve next-year risk ranking, probability accuracy and decision benefit after current coverage, previous failure and country and disease have been included? The contribution is a reproducible temporal comparison that retains unfavourable findings. It does not rest on a claim to be the first longitudinal or multicountry MDA study, and the targeted literature search does not establish that no comparable prediction study exists.

Methods

Study design and setting

This was a retrospective longitudinal prediction study using annual preventive chemotherapy records for lymphatic filariasis, onchocerciasis, schistosomiasis and soil-transmitted helminthiases in Benin, Burkina Faso, Côte d'Ivoire, Ghana, Liberia, Nigeria, Sierra Leone and Togo. The harmonized panel covered 2014–2025. Each observation linked information in predictor year t to an implementation outcome in year t + 1. The unit of analysis was an implementation unit and disease transition; multiple transitions could arise from the same implementation unit.

Two analysis histories are distinguished. The original archives designate 2025 as a locked test excluded from feature development, tuning, calibration development and threshold selection. The present revision audited those outputs and conducted additional analyses after the 2025 results had been examined. These additions are post hoc and cannot be described as a new untouched confirmatory test. Archived statements about model freezing do not independently establish predictions issued before real-world outcomes occurred.

Data source and analytical cohort

The recovered files contain a harmonized panel of 63,672 records and an original modelling dataset of 32,166 transitions across 1,420 implementation units. Excluding 316 Ghana transitions with unavailable target-year 2025 outcomes leaves a definitive analytical dataset of 31,850 transitions. Original files are preserved unchanged; exclusions are supplied separately. Archived cleaning rules exclude duplicate Côte d'Ivoire copies nested in the Ghana source directory, retain the single implementation-unit year record for lymphatic filariasis and onchocerciasis, and retain targetPop equal to Total for schistosomiasis and soil-transmitted helminthiases. These source rules were documented in the archive; the present audit verified the recovered panel and modelling files rather than reconstructing every original export.

The source_file field identifies 32 country–disease export filenames. Each contains the date string 20260912 and the period 2014–2025. These are filename-based provenance clues rather than independently authenticated download timestamps. Original access timestamps and a formal source release identifier were not retained. The current official download page and four disease dictionaries were verified on 16 October 2026. The complete filename inventory and current dictionary snapshots accompany the analysis. The recovered modelling file has SHA-256 1ce86f294eaeeb346201ea6f00a0acb53bd17661877bd7a2c5100cec53374910, matching the original archived audit. This identifies the analytical snapshot, not a publisher-assigned version. The eight-country sample is not assumed to represent all West Africa; its original selection rationale remains undocumented.

Outcome definition and eligibility

The binary outcome was one minus the recorded target-year effectiveness indicator. Only consecutive annual transitions with target-year population requiring treatment greater than zero were included. The audit found no outcome mismatches, nonconsecutive transitions, duplicate analytical keys or retained targets with nonpositive population requiring treatment. All saved 2025 predictions matched the corresponding modelling records and outcomes.

The endpoint represents a recorded implementation effectiveness shortfall. It does not measure individual drug efficacy, infection persistence or failure to eliminate transmission. Target-year eligibility defines the evaluation population; it should not be interpreted as proof that an operational prediction population could have been identified in advance.

The harmonized effective field matched effMda exactly for lymphatic filariasis and onchocerciasis, and effPc exactly for schistosomiasis and soil-transmitted helminthiases, across all 63,672 panel records. Among eligible records, effectiveness also reconstructed exactly from the unrounded ratio of people treated to people targeted, using thresholds of 65%, 80%, 75% and 75%, respectively. Five apparent mismatches based on displayed coverage disappeared when the unrounded ratio was used. The official disease-specific ESPEN dictionaries independently confirm these programme-coverage thresholds and define the denominator as the population targeted for treatment [12, 13, 14, 15]. Epidemiological coverage instead uses the population requiring treatment. These distinct fields must not be interchanged. Availability of target-year eligibility at the intended prediction date remains unverified.

Longitudinal predictors

Predictors were grouped within implementation unit and disease. One-, two- and three-year lags matched preceding calendar years in the recovered panel. Coverage history comprised lagged coverage, current-minus-previous-year change, and the mean and population standard deviation over year t and the preceding two years. Rolling summaries used at least one available year. These windows ended before the target outcome. Historical coverage fields used raw values; the current capped-coverage field limited values above 100%. Capping current coverage therefore did not cap every historical coverage predictor.

Previous failure was the eligible failure indicator in predictor year t. Recent failure count summed eligible failures over the current and preceding two years; an entirely missing failure window remained missing. Programme Treatment Delivery Continuity was the count of intervention-eligible years with recorded delivery divided by the count of eligible years in the same window. It was undefined when no year was eligible. This aggregate delivery construct does not identify whether the same individuals received or swallowed treatment repeatedly.

Cross Programme Failure Burden was represented by failures among other eligible disease interventions divided by the number of other eligible disease interventions within the same unit and predictor year. The ratio was undefined when no other intervention was eligible. Both the number of other eligible interventions and the number of failures among them were independently reconstructed by summing same-unit, same-year disease records and excluding the focal intervention; both matched the stored

values exactly. The burden ratio also matched its reconstructed numerator and denominator. All 34 implemented lag, rolling-window and ratio checks returned zero mismatches in the recovered harmonized records.

Temporal partitioning

Development comprised target years 2015–2022, validation comprised 2023–2024, and the original held-out temporal test comprised 2025. Ghana target-year 2025 records were excluded because unit-level outcomes were unavailable, leaving 1,839 test transitions in seven countries. All Ghana records in development and 2023–2024 validation were retained. This source-driven restriction was made after the original results had been examined. No prediction, fitted model, calibration function or threshold was changed because of this exclusion. The same implementation unit could contribute to different chronological partitions. This design evaluates later outcomes in the observed regional system and does not establish generalization to previously unseen implementation units or countries. Repeated-unit dependence was addressed in the additional uncertainty analysis.

Primary benchmark and longitudinal models

The primary analysis comprises five reproducible logistic specifications fitted after review of the original results. This primary designation was adopted during manuscript revision because the original complete fitting pipeline was unavailable. It does not imply prospective prespecification. B1 used current capped coverage; B2 used previous failure; B3 combined coverage and previous failure with country and disease. L1 added coverage lags 1 and 2, the three-year coverage mean and standard deviation, annual coverage change, recent failure count, eligible-year count and treatment continuity. L2 additionally included other-disease failure rate and number of other eligible diseases. These specifications are new comparisons and are not exact reproductions of the original M4.

Each used L2-penalized logistic regression with C = 1, the lbfgs solver, a maximum of 3,000 iterations and no class weighting or hyperparameter search. Numeric predictors were median-imputed with missingness indicators and standardized; categorical predictors were fully one-hot encoded. Preprocessing and model fitting used development data only. No recalibration was applied. The existing 0.40 threshold was retained for comparison without searching for a more favourable threshold. Scripts and recorded software versions accompany the revision.

Performance calibration and decision benefit

Discrimination was assessed using the area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC), implemented as average precision. Probability accuracy was assessed using Brier score and logarithmic loss. AUPRC was interpreted alongside outcome prevalence. Calibration-in-the-large was estimated by logistic intercept fitting with the prediction logit as an offset. Calibration slope and joint intercept were estimated by logistic regression of outcome on prediction logit. Quantile-binned calibration plots were descriptive.

Net benefit at risk threshold p was calculated as the true-positive proportion minus the false-positive proportion multiplied by p/(1 − p). The evaluated action is prioritization for implementation support. No beneficial effect of such support was estimated from these data. Additional decision curves used an exploratory grid from 0.10 to 0.80 in increments of 0.05, alongside prioritizing all and none. This grid is not an endorsed operational range. The choice of a practical threshold requires information about resources and the relative consequences of missed shortfalls and unnecessary alerts.

Uncertainty and subgroup assessment

Additional uncertainty estimates used 1,000 bootstrap resamples with seed 20261016. Implementation units were sampled with replacement within countries, retaining all disease observations from each selected unit. The same resample was applied to every model in paired comparisons. Percentile 95% confidence intervals describe fixed-prediction performance conditional on the seven countries in the definitive test. Models were not refitted, so these intervals do not include model-development uncertainty or uncertainty from sampling new countries. Decision-curve plots do not include simultaneous bands. Country and disease subgroup summaries include sample size and prevalence; AUROC is undefined for a subgroup with one outcome class. Subgroup evaluation is distinct from country-exclusion experiments.

Artificial intelligence assisted work

ChatGPT was used to assist with editorial revision, repository inspection and preparation of additional analysis scripts during the manuscript rework. The analyses were executed computationally. The authors are responsible for verifying the code, results, interpretation and manuscript. [AUTHOR VERIFICATION REQUIRED: confirm the precise scope, software/service details and dates of AI assistance, and review the final analyses before submission.]

Results

Definitive cohort and exclusion

As summarized in Table 1, the final dataset contains 31,850 transitions. Development and validation retain 25,498 and 4,513 transitions from eight countries. The definitive 2025 test includes 1,839 transitions from Benin, Burkina Faso, Côte d’Ivoire, Liberia, Nigeria, Sierra Leone and Togo, with 851 failures (46.28%). Ghana’s 316 target-year 2025 transitions are excluded; its earlier observations are retained. The pooled final failure prevalence is 51.34%.

The exclusion addresses source completeness rather than model performance. The recovered Ghana 2025 panel has zero treated and targeted counts throughout, while WHO subsequently reported 219,080 LF treatments in two Ghana implementation units [16]. Official ESPEN release notices identify 2025 treatment releases for the seven retained countries [17]. National totals cannot replace unit-level outcomes, and zero-filled exports cannot establish that every eligible intervention failed. Original eight-country results remain in the historical archive and are not the definitive test estimates.

Table 1
Definitive analytical partitions

Partition	Countries	Target years	Transitions	Failures	Failure percent
Development	8	2015–2022	25,498	13,550	53.14
Validation	8	2023–2024	4,513	1,950	43.21
Definitive test	7	2025	1,839	851	46.28

Period specific implementation history

In development, subsequent failure occurred in 27.43% of records without current eligible failure and 73.33% with current failure. In the definitive test, corresponding values were 28.42% among 1,045 records and 72.13% among 732 records. Current failure was undefined for 62 test records, of which 26 subsequently failed. These associations are predictive and do not identify causal implementation mechanisms.

Treatment continuity groups were defined as low at most one-third, moderate above one-third through two-thirds, and high above two-thirds. Development failure rates were 82.38%, 59.25% and 32.23%, respectively. Corresponding definitive 2025 rates were 63.34% among 311 low-continuity records, 60.79% among 732 moderate-continuity records, and 25.92% among 787 high-continuity records. Continuity was undefined for nine test records. The development and test comparisons therefore use nonoverlapping outcome periods.

Incremental value beyond operational benchmarks

Table 2 compares primary models. In the primary post hoc analyses, B3 combined current capped coverage, previous failure, country and disease. It achieved higher definitive test discrimination than either reconstructed longitudinal specification. In 2023–2024 the longitudinal models had higher AUPRC, so their apparent incremental advantage did not persist in the later period. All primary models use unrecalibrated probabilities from development-only fitting. Supplementary Table S1 reports the original saved recalibrated models separately.

Table 2
Primary benchmark and longitudinal model comparison

Model	Validation AUPRC	Test AUROC	Test AUPRC	Test Brier	Test net benefit at 0.40
B1	0.5225	0.7087	0.6268	0.2189	0.2159
B2	0.5084	0.7122	0.6304	0.2060	0.2142
B3	0.6291	0.8179	0.7797	0.1796	0.2606
L1	0.6458	0.7973	0.7398	0.1820	0.2704
L2	0.6477	0.7999	0.7471	0.1812	0.2670

B3 definitive test AUROC was 0.8179 (0.8014–0.8333), AUPRC was 0.7797 (0.7499–0.8084) and Brier score was 0.1796 (0.1707–0.1890). Corresponding L2 estimates were 0.7999 (0.7823–0.8154), 0.7471 (0.7159–0.7745) and 0.1812 (0.1728–0.1899), respectively. Parentheses contain paired implementation-unit bootstrap 95% confidence intervals.

L2 minus B3 produced an AUROC difference of − 0.01803 (95% CI − 0.02467 to − 0.01231) and an AUPRC difference of − 0.03261 (− 0.04723 to − 0.01854). L2 net benefit at 0.40 was higher by 0.00634 (0.00017–0.01266). This is approximately 0.63 additional net true-positive equivalents per 100 evaluated records under that threshold weighting, not an observed gain in treatment coverage. L1 had a larger net-benefit increment over B3 of 0.00979 (0.00390–0.01570), although its discrimination was also lower. L2 did not establish a net-benefit advantage over L1: the difference was − 0.00344 (− 0.00919 to 0.00199).

Calibration across evaluation periods

Calibration estimates are reported in Table 3 and illustrated in Fig. 1. All three context-aware reconstructed models overpredicted aggregate risk in 2023–2024, when failure prevalence was 43.21%. Calibration slopes were approximately 0.53. Definitive 2025 slopes were closer to one, but B3, L1 and L2 still overpredicted aggregate failure risk. Thus, the near equality of predicted and observed risk previously seen in the eight-country sample did not persist after excluding Ghana's unavailable outcomes.

Table 3
Primary model calibration in the definitive test

Model	Validation slope	Test slope	Test mean risk	Test observed risk
B3	0.5297	0.8268	0.5002	0.4628
L1	0.5375	0.8394	0.4833	0.4628
L2	0.5364	0.8587	0.4985	0.4628

Decision value and local variation

As shown in Fig. 2, decision curves varied by threshold. At 0.40, definitive 2025 net benefits were 0.2606 for B3, 0.2704 for L1 and 0.2670 for L2. In validation they were 0.1762, 0.1803 and 0.1749, respectively. Decision curves crossed over the exploratory grid, and no specification dominated across thresholds and both periods. A decision advantage at one retained threshold therefore does not establish a generally preferable model.

National heterogeneity remained substantial. B3 test AUROC ranged from 0.3177 in Burkina Faso to 0.9640 in Sierra Leone, with small country samples limiting precision. L2 AUROC was 0.3975 in Burkina Faso and 0.5077 in Togo. Nigeria contributed 1,190 of the 1,839 definitive test records, approximately 64.7%. Country and disease sample sizes, prevalence and performance are provided in the accompanying tables. These subgroup evaluations assess regional models in observed settings and do not constitute validation in unseen countries.

Discussion

Principal findings

The performance comparisons (Table 2) and calibration results (Table 3; Fig. 1) indicate that the definitive seven-country analysis shows that longitudinal information can improve prediction relative to coverage alone while failing to improve on a stronger operational benchmark. Coverage, previous failure, country and disease together ranked 2025 outcomes more accurately than the reconstructed historical models. The longitudinal specifications yielded small net-benefit increments at 0.40, but these gains did not establish a consistent advantage across evaluation periods and thresholds. Supplementary comparisons of the original saved models are reported separately because their complete fitting pipeline could not be recovered. Model choice therefore depends on the aspect of performance being assessed.

The source-driven Ghana exclusion is scientifically consequential. The updated test estimates are lower than the archived eight-country discrimination values, and probability calibration changes substantially. The difference demonstrates that an internally consistent binary field can still be an unverified outcome when data reporting is incomplete. Source maturity must be assessed before zero-coded effectiveness is interpreted as observed implementation failure.

Contribution to infectious disease implementation research

The literature gap concerns incremental operational value rather than novelty of longitudinal modelling. Fuller and colleagues already examined contextual determinants of lymphatic filariasis coverage across fourteen years and ten countries [5]. The present analysis asks whether additional histories improve next-year warnings after a practical context-aware comparator is introduced, and whether any improvement survives temporal evaluation. Its evidence places limits on a general superiority claim and identifies the need to evaluate risk ranking, calibration and decision benefit separately.

Treatment continuity was associated with subsequent implementation shortfalls, but the aggregate measure does not identify individuals repeatedly missed by treatment. That distinction matters because individual compliance has a separate evidence base and requires linked individual observations [6]. Likewise, adding cross-disease variables in L2 involved both a burden ratio and the number of other eligible interventions. Any increment over L1 cannot be attributed to the ratio alone, and the decision comparison did not establish an advantage of L2 over L1 at the retained threshold.

Strong comparators and temporal stability

A coverage-only benchmark omits previous failure and contextual information that programme managers may already use. Improvements over that benchmark can overstate the practical necessity of an elaborate historical representation. The reconstructed B3 comparison provides a more demanding test of incremental value.

In 2023–2024, longitudinal models had stronger precision–recall performance than B3; in the definitive 2025 cohort, B3 performed better. This reversal argues for repeated temporal evaluation rather than selection based on one favourable validation result. The restriction from eight countries in validation to seven in testing also changes sample composition, so the observed differences cannot be attributed entirely to time. No threshold was optimized after Ghana exclusion.

Calibration and decision support

As illustrated in Fig. 1, the primary models have definitive test calibration slopes below one and overpredict average risk. Supplementary M4 probabilities instead underpredict aggregate risk and have a calibration slope above two. Their probabilities therefore require additional evaluation even when discrimination is acceptable. Recalibration should be treated as a dated model-updating procedure, with independent assessment in the setting where it is intended for use.

The decision curves (Fig. 2) show that the net-benefit increments for historical models are modest. They represent trade-offs between correct alerts and unnecessary prioritization under a specified false-positive weighting. They do not demonstrate that responding to alerts will increase treatment coverage or reduce infection. The response could involve supervision, logistics review or other implementation support, but its effectiveness and resource requirements were not measured. Digital reporting experience in Nigeria illustrates opportunities to integrate data into programme action [7]; predictive prioritization requires separate evaluation of the action triggered by an alert.

Outcome verification and geographic interpretation

The official dictionaries clarify that programme coverage uses the population targeted for treatment, whereas epidemiological coverage uses the population requiring treatment [12, 13, 14, 15]. Failure to meet a programme-coverage threshold therefore does not directly measure population-level transmission control or progress towards elimination. The distinction between a confirmed zero-delivery year and unavailable unit-level reporting is equally important. Ghana 2025 was excluded because the latter could not be established, not because its removal improved performance.

Nigeria contributes nearly two-thirds of the definitive test records. Poor national discrimination in some smaller settings remains visible despite acceptable pooled metrics. The country-specific results therefore support local assessment before any application of a regional risk model. They should not be presented as evidence of performance in an unseen country. Additional externally evaluated settings and later verified reporting cycles are needed.

Strengths and limitations

Strengths include strict consecutive-year linkage, independent reconstruction of outcome mappings and longitudinal predictors, explicit verification against official ESPEN dictionaries, paired implementation-unit uncertainty estimates, and preservation of original results separately from the corrected test cohort. The revision also tests stronger operational benchmarks and retains unfavourable comparisons.

The analysis remains dependent on routine reported counts, population denominators and national reporting processes. The Ghana finding demonstrates why completeness of numeric fields cannot substitute for verified outcome availability. Release notices for the seven retained countries do not authenticate every individual record. Original raw-export files and download timestamps were not fully recovered, although the modelling snapshot checksum matches the archived audit. Target-year treatment eligibility remains a retrospective inclusion criterion whose availability at an operational prediction date has not been established. The original rationale for selecting these eight development countries is not documented, limiting regional representativeness claims.

Original fitted objects and complete training specifications remain unavailable. Original raw probability summaries from the eight-country cohort cannot be transferred to the definitive seven-country test. The reconstructed benchmarks use independently specified development-only models and are post hoc analyses, not exact recovery of the original M4. Exclusion and revised evaluation occurred after outcomes had been examined; this is a correction for source availability and does not create a new untouched validation experiment. Any apparent incremental value needs confirmation in another outcome period.

Bootstrap intervals condition on fixed predictions and the observed countries, exclude model-refitting uncertainty and do not adjust for multiple comparisons. Subgroup estimates are imprecise in smaller samples; calibration summaries lack interval estimates; and only one definitive later test year is available. Historical coverage remains uncapped in several predictors, despite capping the current coverage variable. The study is predictive and cannot establish causal implementation mechanisms, individual adherence, transmission reduction or health benefits of deployment.

Conclusions

After excluding unavailable Ghana 2025 unit-level outcomes, the definitive seven-country analysis does not support uniform superiority of longitudinal implementation histories over a context-aware benchmark. Simpler information produced stronger discrimination, while historical models showed modest threshold-specific decision gains. The evidence supports careful benchmarking, source verification and separate evaluation of calibration and decision value. Further evaluation with verified later outcomes and a defined programme response is required before operational use.

Abbreviations

AUPRC
area under the precision–recall curve
AUROC
area under the receiver operating characteristic curve
Brier
mean squared probability error
CPFB
Cross Programme Failure Burden

DHIS2
District Health Information Software 2
ESPEN
Expanded Special Project for Elimination of Neglected Tropical Diseases
LPVS
Longitudinal Programme Vulnerability State
MDA
mass drug administration
NTD
neglected tropical disease
PTDC
Programme Treatment Delivery Continuity
WHO
World Health Organization.

Declarations

Ethics approval and consent to participate

This study analysed publicly accessible, aggregated secondary programme implementation data from WHO ESPEN. No participants were recruited or contacted, and the study used no personally identifiable individual-level data or human tissue. The study was not submitted to an ethics committee or institutional review board; therefore, no formal ethics approval or waiver was obtained." Please advise whether your journal requires an institutional determination concerning the need for ethics review for this aggregate secondary-data analysis.

Consent for publication

Not applicable. No identifiable individual information is presented.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no external funding.

Author Contribution

MIA contributed to conceptualization, methodology, data curation, formal analysis, investigation, validation, visualization, software, original drafting, revision, and project administration. JNO confirmed final review and approval of all revised analyses and manuscript content.

Acknowledgements

The author acknowledges WHO Regional Office for Africa and ESPEN for access to implementation data resources. The interpretations are those of the author and do not represent an official WHO, ESPEN or national programme position.

Availability of data and materials

The study used aggregated implementation-unit treatment records from the Expanded Special Project for Elimination of Neglected Tropical Diseases (ESPEN), World Health Organization Regional Office for Africa. Records covered lymphatic filariasis, onchocerciasis, schistosomiasis and soil-transmitted helminthiases during 2014–2025 in Benin, Burkina Faso, Côte d'Ivoire, Ghana, Liberia, Nigeria, Sierra Leone and Togo [2]. Ghana target-year 2025 observations were excluded from the definitive test because unit-level outcomes were unavailable; its earlier observations were retained. The following official links identify the source and supporting documentation.

Data access and implementation-unit export tool: https://espen.afro.who.int/maps-data/data-query-tools/download-data

Data query tools and reporting context: https://espen.afro.who.int/maps-data/data-query-tools

Data release history: https://espen.afro.who.int/maps-data/data-updates

Lymphatic filariasis data dictionary:
https://espen.afro.who.int/sites/default/files/content/document/Lymphatic_Filariasis_EN_FR_PT.pdf

Onchocerciasis data dictionary:
https://espen.afro.who.int/sites/default/files/content/document/Onchocerciasis_EN_FR_PT.pdf

Schistosomiasis data dictionary:
https://espen.afro.who.int/sites/default/files/content/document/Schistosomiasis_EN_FR_PT.pdf

Soil-transmitted helminthiases data dictionary:
https://espen.afro.who.int/sites/default/files/content/document/STH_EN_FR_PT.pdf

These source pages and dictionaries were verified on 16 October 2026. This verification date is not the original download date, which was not independently retained. Source filenames and the analytical

snapshot checksum are documented in Methods and the reproducibility package. Because the portal may be updated, a new export may differ from the archived analytical snapshot.

No unrestricted redistribution licence for the treatment CSVs was established from the reviewed source pages. Author-generated analysis scripts, specifications, exclusions and audit outputs accompany the working reproducibility package. [COMPLETE BEFORE SUBMISSION: insert the permanent public code repository URL and identifier for the definitive analysis version if deposited.]

References

1. World Health Organization. Preventive chemotherapy tools for improving the quality of reported data and information: a field manual for implementation. Geneva: WHO; 2019. https://www.who.int/publications/i/item/9789241516464.
2. World Health Organization Regional Office for Africa. ESPEN data query tools. https://espen.afro.who.int/maps-data/data-query-tools/download-data. Source page verified 16 October 2026; original export access date not retained.
3. Collins GS, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.
4. Moons KGM, et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. https://doi.org/10.1136/bmj-2024-082505.
5. Fuller BB, Harris V, Parker C, Martinez A, Toubali E, Ebene BC, et al. Contextual determinants of mass drug administration performance: Modelling fourteen years of lymphatic filariasis treatments in West Africa. PLoS Negl Trop Dis. 2023;17(2):e0011146. https://doi.org/10.1371/journal.pntd.0011146.
6. Maddren R, Phillips A, Rayment Gomez S, Forbes K, Collyer BS, Kura K, et al. Individual longitudinal compliance to neglected tropical disease mass drug administration programmes, a systematic review. PLoS Negl Trop Dis. 2023;17(7):e0010853. https://doi.org/10.1371/journal.pntd.0010853.
7. Imhansoloeva M, Nwosu C, Osinoiki O, Isiyaku S, Anyaike C, Amodu-Agbi P, et al. Lessons from piloting and scaling a real-time DHIS2 based treatment reporting tool for mass drug administration in Nigeria. PLoS Negl Trop Dis. 2025;19(10):e0013601. https://doi.org/10.1371/journal.pntd.0013601.
8. Hardwick RJ, Truscott JE, Oswald WE, Werkman M, Halliday KE, Pullan RL, et al. Individual adherence to mass drug administration in neglected tropical disease control: A probability model conditional on past behaviour. PLoS Negl Trop Dis. 2021;15(1):e0009112. https://doi.org/10.1371/journal.pntd.0009112.
9. Means AR, Ásbjörnsdóttir KH, Sharrock KC, et al. Coverage of community-wide mass drug administration platforms for soil-transmitted helminths in Benin, India, and Malawi: findings from the DeWorm3 project. Infect Dis Poverty. 2024;13:72. https://doi.org/10.1186/s40249-024-01241-0.

10. Namara BG, Kwesiga B, Migisha R, Bulage L, Tinkitina B, Ario AR, et al. Trends of Coverage of Mass Drug Administration, Population at Risk and Reported Cases of Schistosomiasis and Soil-Transmitted Helminths, Uganda, 2013–2023. BMC Infect Dis. 2025;25:742. https://doi.org/10.1186/s12879-025-11154-4.
11. Zoerhoff KL, Mbabazi PS, Gass K, Kraemer J, Fuller BB, Blair L, et al. How well do coverage surveys and programmatically reported mass drug administration coverage match? Results from 214 mass drug administration campaigns in 15 countries, 2008–2017. BMJ Glob Health. 2023;8:e011193. https://doi.org/10.1136/bmjgh-2022-011193.
12. World Health Organization Regional Office for Africa. Lymphatic Filariasis IU Level Data Dictionary EN FR PT. https://espen.afro.who.int/sites/default/files/content/document/Lymphatic_Filariasis_EN_FR_PT.pdf. Accessed 16 October 2026.
13. World Health Organization Regional Office for Africa. Onchocerciasis IU Level Data Dictionary EN FR PT. https://espen.afro.who.int/sites/default/files/content/document/Onchocerciasis_EN_FR_PT.pdf. Accessed 16 October 2026.
14. World Health Organization Regional Office for Africa. Schistosomiasis IU Level Data Dictionary EN FR PT. https://espen.afro.who.int/sites/default/files/content/document/Schistosomiasis_EN_FR_PT.pdf. Accessed 16 October 2026.
15. World Health Organization Regional Office for Africa. STH IU Level Data Dictionary EN FR PT. https://espen.afro.who.int/sites/default/files/content/document/STH_EN_FR_PT.pdf. Accessed 16 October 2026.
16. World Health Organization. Global programme to eliminate lymphatic filariasis progress report 2025. Weekly Epidemiological Record. 2026;101(39). Table 3. https://www.who.int/publications/journals/weekly-epidemiological-record/wer101-39. Accessed 16 October 2026.
17. World Health Organization Regional Office for Africa. ESPEN Data Updates. https://espen.afro.who.int/maps-data/data-updates. Accessed 16 October 2026.

Figures

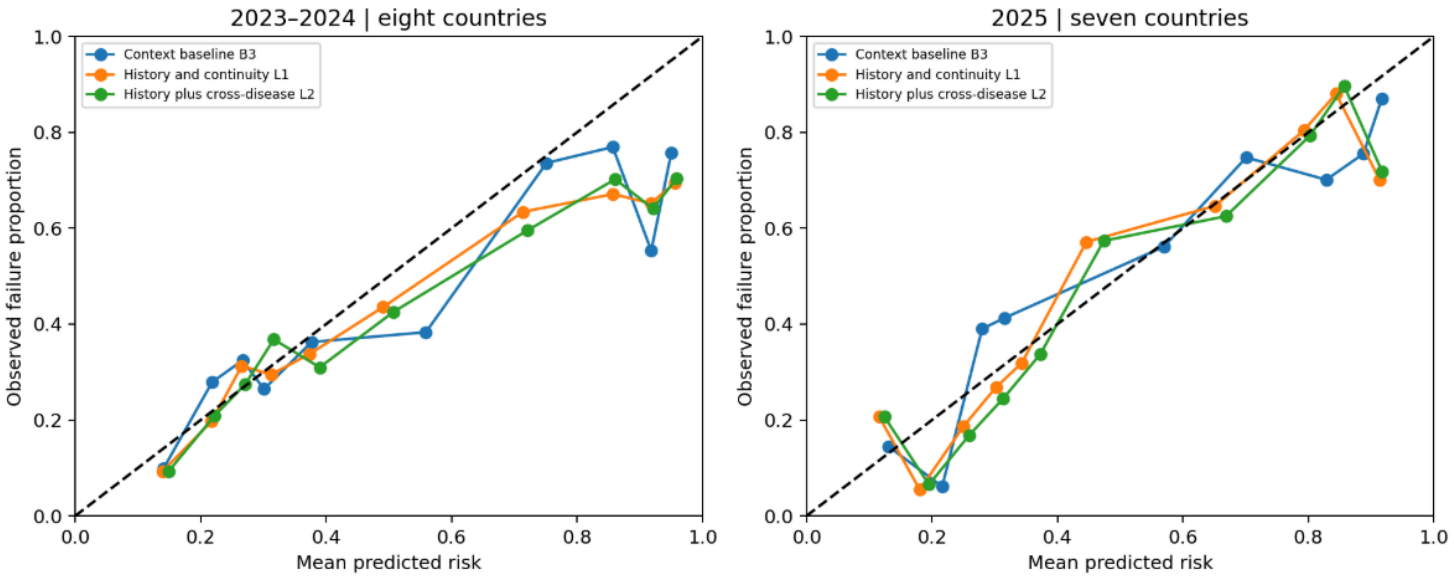

Figure 1

Quantile-binned calibration for eight-country validation in 2023–2024 and seven-country definitive testing in 2025. The diagonal represents ideal calibration. Binned estimates are descriptive and uncertainty is not displayed.

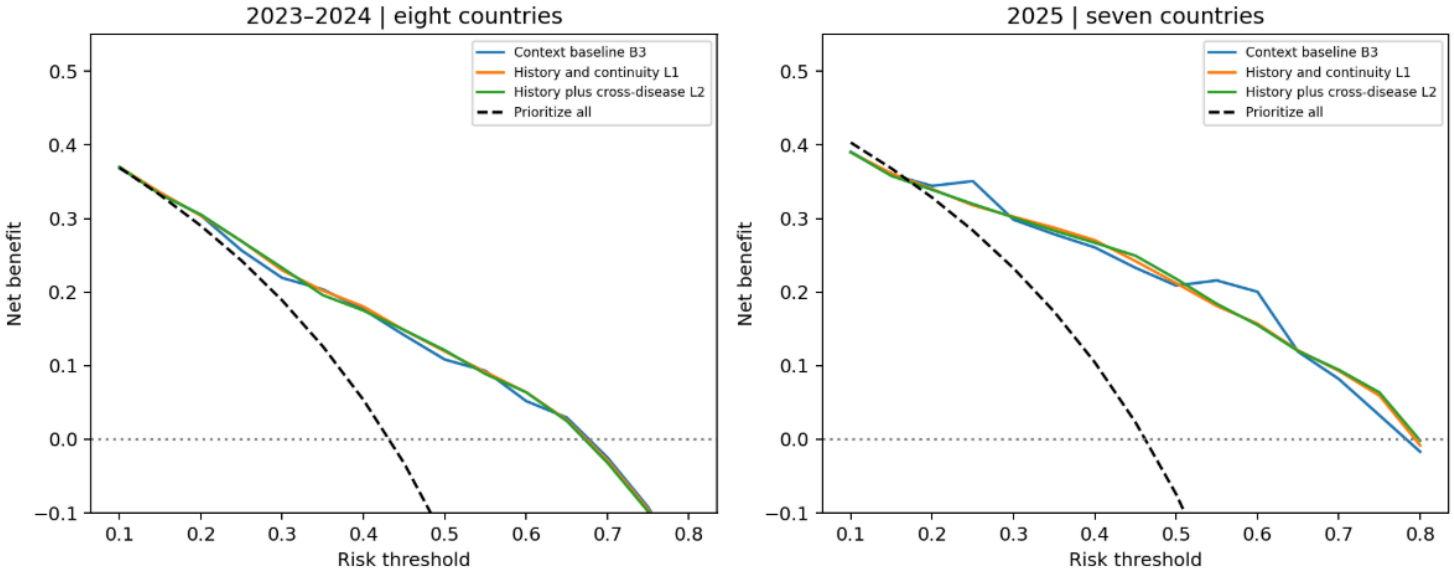

Figure 2

Descriptive decision curves for the unchanged eight-country validation cohort and the seven-country definitive test. Prioritizing none has net benefit zero. The grid is exploratory and is not an operationally endorsed threshold range; no simultaneous bands are shown.