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 health-service exclusion before it becomes visible: an explainable, spatially validated machine-learning study integrating NDHS 2023–24, GRID3 and OpenStreetMap data in Nigeria

Akingbola Sunday Omokanye
asomokanye@oauife.edu.ng
Obafemi Awolowo University
Jacob Wale Mobolaji
Obafemi Awolowo University
Ada Daisy Oguizu
Rivers State University
Olayinka Isiaka Baruwa
Obafemi Awolowo University

Research Article

Keywords: health services accessibility, universal health coverage, machine learning, explainable artificial intelligence, SHAP, geospatial analysis, Demographic and Health Surveys, Nigeria

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11678723/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

Health-service exclusion in Nigeria is usually detected only after periodic household surveys reveal low coverage, and surveys cannot produce direct estimates for the wards and local government areas (LGAs) where primary health care is planned. We aimed to predict multi-service exclusion from geospatial data available for every location, identify its drivers and quantify the excluded population that surveys do not reach.

Methods

We linked 1,270 clusters of the 2023–24 Nigeria Demographic and Health Survey with GRID3 health-facility, population and settlement data, OpenStreetMap-derived travel times, night-time lights and conflict data. A high-exclusion cluster had mean non-receipt of four or more antenatal visits, facility delivery and a first DTP-containing vaccine of at least 50%. Penalised spline logistic regression, random forest, gradient boosting and XGBoost were compared under nested state-blocked and leave-one-zone-out validation. Predictions were explained with SHAP values grouped by Levesque access domain, and zone-benchmarked risk was mapped for every populated 1-km cell, ward and LGA.

Results

In total, 425 clusters (33.5%) were highly excluded, from 3.6% in the South East to 74.5% in the North West. Elastic-net logistic regression performed best under spatial validation (AUC 0.868, 95% CI 0.844–0.888; calibration slope 0.97), outperformed the three machine-learning algorithms (Holm-adjusted $p \leq 0.022$) and retained an AUC of 0.845 when whole zones were withheld. Facility capacity accounted for 64.5% of the SHAP explanation, led by the private share of facilities and population per facility; travel-time and road features added no discrimination (ΔAUC − 0.003). After benchmarking, 2,534 of 9,014 wards (28.1%) and 231 of 774 LGAs were at high or very high risk, home to 104.7 million people; 88.1 million (84.2%) lived in wards without a survey cluster, 60.0 million of them in the North West. Model-based and direct state estimates agreed moderately (concordance correlation coefficient 0.66, 95% CI 0.52–0.76).

Conclusions

Exclusion from maternal and child health services can be anticipated from open geospatial data, chiefly from facility capacity relative to population. Benchmarked ward-level risk maps reveal a large excluded population that national surveys do not reach and can guide primary health care investment and outreach before exclusion appears in the next survey.

Background

Universal health coverage (UHC) commits countries to ensuring that everyone can use the essential health services they need, yet an estimated 4.5 billion people were not fully covered by essential health services in 2021 [1]. Nigeria, Africa's most populous country, shows both the scale and the unevenness of this gap. The Lancet Nigeria Commission documented persistent sub-national inequities in service coverage and health outcomes [2]. In the 2023–24 Nigeria Demographic and Health Survey (NDHS), only 43% of live births in the two years before the survey took place in a health facility, 52% of women with a recent birth reported four or more antenatal care (ANC) visits, and 31% of children aged 12–23 months had received no vaccinations [3]. Facility delivery ranged from 22% in the North West zone to 87% in the South East, and from 9% in Kebbi State to 97% in Imo State [3].

These figures are averages over very large populations. The NDHS sample is designed to give representative estimates for the nation, for urban and rural areas, for the six geopolitical zones and for the 36 states and the Federal Capital Territory [3]. Primary health care, however, is planned and delivered at the level of local government areas (LGAs) and wards. Most of

Nigeria's 774 LGAs and several thousand wards contain few or no survey clusters, so direct estimates are unavailable at precisely the scale where decisions about facility siting, outreach and staffing are made. Routine facility data cannot fill this gap, because they count people who reach services rather than those who do not. Exclusion is therefore detected late and indirectly, usually when the next survey round shows that coverage has stagnated. The problem is sharpest where exclusion is likely to be greatest: 20 of the 1,400 clusters selected for the 2023–24 NDHS could not be surveyed because of insecurity [3], so some of the most excluded communities are missing from the very data used to measure exclusion.

Two bodies of work have tried to see beneath these averages. Geographic accessibility studies model travel time from where people live to the nearest facility using road networks and friction surfaces, and they consistently show large populations beyond reasonable reach of care [4–6]. Model-based mapping studies take a second route: they link geolocated survey clusters to gridded covariates and predict coverage at fine resolution, using Bayesian geostatistical models, machine learning or hybrids of both [7–10]. In Nigeria, Kawakatsu and colleagues used generalised additive models and gradient boosting to map ten maternal and child health services at 1–5 km resolution from surveys conducted between 2003 and 2018 [9]. Utazi and colleagues compared geostatistical, machine-learning and hybrid approaches for mapping vaccination coverage from NDHS data and found that geostatistical models performed best, while some machine-learning methods over-smoothed the surface and could mask coverage gaps [10].

Four gaps remain. First, existing maps are built to predict rather than to explain. They show where coverage is low but not whether this reflects too few facilities, poor physical access or the wider context, which is the information a planner needs to choose between building, equipping, connecting or reaching out. Second, most mapping studies validate models with random cross-validation. When observations are spatially autocorrelated, random folds place near neighbours in both training and test sets, which overstates accuracy and says little about performance in unsampled regions [11, 12]. Third, the supply side has usually been represented by public-sector facility inventories [13] and global friction surfaces, rather than the national multi-ownership facility registry and road network now available through GRID3 and OpenStreetMap. Fourth, outcomes are modelled one service at a time, whereas the communities of greatest concern are those that miss contact across the whole continuum of care. In addition, no published model uses the 2023–24 NDHS, the first round since the COVID-19 pandemic. More broadly, machine-learning models in global health are often criticised as opaque [14]; SHapley Additive exPlanations (SHAP), which decompose each prediction into additive contributions from its inputs, offer a principled response [15, 16] but have rarely been linked to an explicit theory of access.

This study addresses these gaps. Using the 2023–24 NDHS linked to GRID3, OpenStreetMap and contextual data, we developed and spatially validated explainable machine-learning models that predict multi-service health-service exclusion from geospatial data alone, so that predictions can be made for every location in Nigeria, including places that surveys do not reach. Our objectives were to:

1. construct a cluster-level analytic dataset linking NDHS-measured multi-service exclusion to GRID3 service-capacity, OpenStreetMap accessibility and contextual covariates;
2. compare the predictive performance of random forest, gradient boosting and XGBoost against a penalised logistic regression benchmark under spatially blocked validation;
3. identify, using SHAP values aggregated by access domain, the predictors that contribute most to predicted exclusion and how their contributions vary across the risk gradient; and
4. produce ward- and LGA-level risk maps, quantify the population at high predicted risk in areas with no survey coverage, and assess agreement between model-based and direct survey estimates.

We tested four null hypotheses: ($H0_1$) there is no difference in discrimination between the four algorithms; ($H0_2$) geographic-accessibility features do not improve discrimination beyond capacity and contextual features; ($H0_3$) there is no association between the population-to-facility ratio and exclusion risk; and ($H0_4$) there is no concordance between model-aggregated and direct survey estimates of exclusion at state level.

Methods

Study design and reporting

This was a cross-sectional, ecological, geospatial prediction-modelling study using secondary data. The unit of analysis was the NDHS survey cluster for model development and the 1-km grid cell, ward and LGA for prediction. Reporting follows the TRIPOD + AI statement for prediction models developed with regression or machine-learning methods [17]; the completed checklist is provided in Additional file 2. The analysis followed the study protocol, in which the outcome, predictors, algorithms, validation schemes and hypotheses were specified before model fitting.

Conceptual framework

The study is anchored in two theories. Levesque, Harris and Russell's patient-centred framework conceptualises access as the interface between five dimensions of service supply and five corresponding abilities of the population [18], extending earlier work that defined access as the fit between clients and the health system [19]. We used it to define the outcome and to organise predictors into domains: **availability and accommodation** (facility capacity, from GRID3), **ability to reach** (physical accessibility, from OpenStreetMap), **ability to pay** (an area-level economic proxy, from night-time lights) and a **contextual domain** of settlement pattern and insecurity. Shapley value theory, which allocates a joint outcome among contributors according to their average marginal contribution [20], underpins the SHAP method used to attribute each prediction to its inputs [15]. Because SHAP values are additive, the contribution of each access domain can be obtained by summing the values of its features, which links the model's explanations back to the framework. The acceptability and appropriateness dimensions of the framework cannot be measured from geospatial data and are not modelled; we return to this in the Discussion.

Data sources

Nigeria Demographic and Health Survey 2023–24. The NDHS used a stratified two-stage cluster design with 74 strata (urban and rural areas of each state and the Federal Capital Territory). A total of 1,400 clusters (701 urban, 699 rural) were selected with probability proportional to size and 30 households were selected in each; fieldwork ran from 1 December 2023 to 7 May 2024, and 1,380 clusters were completed after 20 were dropped because of insecurity [3]. We used the women's (IR), children's (KR) and household (HR) recode files and the cluster GPS dataset (GE). To protect confidentiality, cluster coordinates are randomly displaced by up to 2 km in urban areas and up to 5 km in rural areas, with 1% of rural clusters displaced by up to 10 km [21].

Geospatial data. Health-facility locations, ownership and level of care were taken from the GRID3 Nigeria health-facilities dataset [22]; gridded population estimates and settlement extents from GRID3 [23, 24]; the road network and major water bodies from OpenStreetMap via a dated Geofabrik extract [25]; state and LGA boundaries from GRID3 and ward boundaries from the GRID3 operational wards dataset (version 4.0); annual night-time light radiance from the VIIRS annual composite for 2023 [26]; and geolocated conflict events from the Armed Conflict Location and Event Data project (ACLED) [27]. Table 1 lists the version, date and resolution of each source.

Table 1
Data sources used in the study

Source	Version / date accessed	Type and resolution	Variables derived	Licence
NDHS 2023–24 (IR, KR, HR, GE files)	Recode 8B (IR, KR, HR); GE 8A	Household survey; displaced cluster coordinates	Outcome components; survey weights; urban/rural	DHS data-use agreement
GRID3 Nigeria Health Facilities	v2.0 (November 2024); 51,012 facilities	Point locations	Facility counts, distances, ratios, ownership, level	CC BY 4.0
GRID3 gridded population estimates	v3.0 (August 2025)	Raster, ~ 100 m	Population density; denominators; aggregation weights	CC BY 4.0
GRID3 settlement extents	v4.1 (October 2026); 2,546,560 polygons	Polygons by settlement type	Settlement composition of buffers and cells	CC BY 4.0
OpenStreetMap (Geofabrik extract)	14 September 2026; 1,791,268 road segments	Road lines by class; water polygons	Road density; distance to major road; friction surface	ODbL
Administrative boundaries	GRID3 state and LGA boundaries (December 2020); GRID3 operational wards v4.0	State, LGA and ward polygons	Aggregation units; survey-visibility status	CC BY 4.0
VIIRS annual night-time lights	VNL v2.2, 2023 annual average (masked)	Raster, ~ 500 m	Mean radiance (ability-to-pay proxy)	CC BY 4.0
ACLED	Accessed 28 September 2026; political-violence events, 1 January 2022 to 7 May 2024 (9,515 events)	Point events	Conflict events and fatalities within 10 km	ACLED terms of use

Outcome: multi-service health-service exclusion

Exclusion was measured directly from NDHS service-use records using three tracer services spanning the continuum of maternal and child care: (i) fewer than four ANC visits for the most recent live birth covered by the survey's maternity module (live births in the two years preceding the survey), among women aged 15–49 ("don't know" responses counted as fewer than four, following DHS convention); (ii) delivery of that birth outside a health facility; and (iii) zero-dose status, defined as a child aged 12–35 months who had received no dose of a DTP-containing vaccine according to card or maternal recall. Zero-dose status is the marker used by the Immunization Agenda 2030 to identify communities missed by routine services [28]. We used the 12–35-month age band to increase the number of eligible children per cluster, and examined the standard 12–23-month band in sensitivity analysis.

For each cluster we calculated the proportion not receiving each service among those eligible for it, and defined the cluster exclusion score as the unweighted mean of the three proportions. Scores were calculated for clusters with at least one eligible person for at least two of the three services and at least five eligible persons in total. The primary outcome was binary: a **high-exclusion cluster** was one with an exclusion score of 0.50 or more, meaning that, on average across the continuum, most eligible people in the cluster did not receive the service. Each cluster was given equal weight in model training, so that the model reproduces the cluster-level prevalence of high exclusion; weighting by the number of eligible persons and by survey design weights was examined in sensitivity analyses. Because the outcome is built only from service-use records and none of its components is used as a predictor, the models predict an independently observed quantity rather than reproducing a composite constructed from their own inputs.

Predictors

Candidate predictors were restricted to variables that can be calculated for any location in Nigeria, which is what allows the model to make predictions where there is no survey data. For the same reason, we did not use individual or household characteristics from the NDHS, administrative identifiers or geographic coordinates as predictors in the main models. Predictors were specified in advance and grouped into the access domains described above (Table 2).

Table 2
Candidate predictors grouped by access domain

Domain (Levesque dimension)	Predictor	Operational definition	Source
Availability and accommodation	Facility density	Number of facilities (all; primary) within 5 km	GRID3
	Higher-level facility density	Number of secondary or tertiary facilities within 10 km	GRID3
	Distance to nearest facility	Straight-line distance (km) to nearest facility of any level, and of secondary/tertiary level	GRID3
	Population per facility	Population within 10 km divided by number of facilities within 10 km (catchment ratio; zero-facility buffers coded with a 0.5 continuity correction and flagged)	GRID3
	Private share	Proportion of facilities within 10 km that are privately owned	GRID3
Ability to reach	Motorised travel time	Modelled minutes to nearest primary facility and to nearest secondary/tertiary facility	OSM + GRID3
	Walking travel time	Modelled minutes to nearest facility on foot	OSM + GRID3
	Road density	Kilometres of road per km² within 5 km (all classes; trunk, primary and secondary only)	OSM
	Distance to major road	Kilometres to nearest trunk, primary or secondary road	OSM
Ability to pay (area proxy)	Night-time lights	Mean log radiance within the extraction buffer	VIIRS
Context: settlement	Population density	Persons per km² within the extraction buffer	GRID3
	Settlement composition	Share of buffer population in built-up areas, small settlement areas and hamlets	GRID3
Context: insecurity	Conflict exposure	Number of conflict events and fatalities within 10 km, 1 January 2022 to end of fieldwork	ACLED

Travel times were derived from an accumulated-cost surface (see Covariate extraction). Settlement categories follow the GRID3 settlement-extent classification.

Covariate extraction and travel-time modelling

All layers were projected to a common equal-area coordinate reference system. For continuous raster covariates, values at each cluster were averaged within a buffer equal to the maximum displacement distance (2 km for urban and 5 km for rural clusters), following DHS guidance for displaced coordinates [29]. Counts and ratios were calculated within fixed 5-km and 10-km buffers, which are larger than most displacement distances and therefore less sensitive to them.

Travel time was modelled with an accumulated-cost approach at 500-m resolution, chosen so that national surfaces could be computed on a standard 8-GB computer. OpenStreetMap roads were rasterised by class and assigned speeds (motorway and trunk 70 km/h; primary 60; secondary 50; tertiary 40; unclassified and residential 30; tracks and service roads 15), with off-road movement at 5 km/h and major water bodies treated as barriers. The walking scenario applied 5 km/h on roads and tracks and

3 km/h off-road. Minimum travel time from every cell to the nearest facility of each level was computed with the minimum-cost-path routine (MCP_Geometric) of scikit-image, which accounts for diagonal movement. Road density was approximated from the road network rasterised at 500 m. Speed assumptions are listed in Additional file 1: Table S3 and were varied in sensitivity analysis by increasing major-road speeds by 25% while reducing minor-road speeds by 25%, and the reverse.

Prediction grid. All 1-km grid cells with an estimated population of at least one person formed the prediction surface. Each cell received the same predictors as the clusters, extracted from its centroid with a 2-km buffer where the cell lay in a built-up area and a 5-km buffer elsewhere, mirroring the extraction rule for urban and rural clusters.

Sample size

The adequacy of the available sample was assessed against the three criteria of Riley and colleagues for binary prediction models [30] (expected shrinkage of at least 0.90, optimism in Nagelkerke R^2 of at most 0.05, and precise estimation of overall risk), calculated from the published formulas with the observed outcome prevalence and a conservative anticipated Cox–Snell R^2 of 15% of its maximum. Calculations were made for the 21 features used by the tree-based models and for the larger number of parameters in the spline logistic regression. The minimum required sample was 1,642 clusters for the 21-feature tree-based models and 9,459 for the 121-parameter spline logistic regression, against 1,270 available (Additional file 1: Table S6); we therefore relied on penalisation and on calibration under spatial validation to detect overfitting.

Model development

Four algorithms were developed: (i) logistic regression with cubic B-splines (four knots placed at quantiles) for continuous predictors and an elastic-net penalty, as a transparent benchmark, given evidence that machine learning often offers little advantage over well-specified regression [31]; (ii) random forest [32]; (iii) gradient boosting machines [33]; and (iv) XGBoost [34]. A fifth, hybrid model added spatial residual kriging to the best tree-based model, in the spirit of regression-kriging with random forests [35]. This allows spatial structure not captured by the covariates to be recovered, and addresses the concern that tree ensembles over-smooth coverage surfaces [10].

Hyperparameters were tuned by grid search within each outer training set using inner grouped cross-validation, optimising log-loss as a proper scoring rule; the grids are given in Additional file 1: Table S2. Each cluster was weighted equally in training and counted once in performance measures. The final model was prespecified as the model with the highest spatially validated AUC among those with a calibration slope between 0.80 and 1.25, so that a model with good discrimination but over-dispersed predictions would not be used for mapping. Models were fitted in Python with scikit-learn [36], xgboost and shap, called from Stata 17 through its Python integration. Random seeds were fixed and package versions recorded.

Validation and performance measures

Three validation schemes were applied, each with tuning nested inside the training folds:

- **Random 10-fold cross-validation**, reported only to quantify the optimism that arises when spatial dependence is ignored;
- **Spatially blocked 10-fold cross-validation** (primary analysis), with the 37 states as blocking units so that all clusters from a state fall in the same fold [11, 12]; and
- **Leave-one-zone-out validation**, in which each of the six geopolitical zones is predicted from a model trained on the other five, as a stringent test of transportability to unsampled regions.

Performance was summarised from pooled out-of-fold predictions. Discrimination was measured by the area under the receiver operating characteristic curve (AUC), with 95% confidence intervals from 1,000 bootstrap resamples of clusters. Calibration was assessed with calibration-in-the-large, the calibration slope and flexible calibration curves [37], and overall accuracy with the Brier score and scaled Brier score. Sensitivity, specificity, positive and negative predictive values, F1-score and Cohen's kappa were reported at a threshold of 0.50 and at the Youden-optimal threshold. Decision-curve analysis [38]

estimated the net benefit of using the model to target areas for outreach across threshold probabilities from 0.20 to 0.80. Residual spatial autocorrelation in out-of-fold residuals was tested with Moran's I [39] using inverse-distance weights with a 50-km cut-off.

Hypothesis tests

$H0_1$ was tested by comparing the AUCs of the four algorithms under spatially blocked validation with DeLong's test for correlated ROC curves [40], with Holm adjustment for multiple comparisons. **$H0_2$** was tested with nested feature sets fitted with the best-performing of the four base algorithms: M1 (contextual and ability-to-pay features), M2 (M1 plus availability and accommodation) and M3 (M2 plus ability to reach). The change in AUC from M2 to M3 was tested with DeLong's test, and the change in Brier score was reported with a bootstrap confidence interval. **$H0_3$** was assessed in two ways: from the shape and direction of the SHAP dependence of predicted risk on population per facility, and from a fractional logistic regression of the continuous exclusion score on all predictors, with standard errors clustered by state, reporting the odds ratio per interquartile-range increase. **$H0_4$** was tested by comparing, for each state, the population-weighted mean predicted risk with the direct survey-weighted proportion of the population living in high-exclusion clusters. To keep this comparison out-of-sample, each state's grid cells were predicted from a model trained without that state's clusters. Agreement was summarised with Lin's concordance correlation coefficient [41] and the proportion of states whose predicted value fell within the 95% confidence interval of the direct estimate.

Explainability

The final model, refitted on all clusters, was explained with SHAP values [15]: TreeSHAP [16] for tree-based models and, for the penalised regression, a model-agnostic Shapley explainer on the probability scale with a background sample of 100 clusters. We report global importance as the mean absolute SHAP value of each feature, and domain-level importance as the mean absolute value of the summed SHAP values of the features in each access domain. Dependence plots show how predicted risk changes with travel time, facility density and population per facility, and, had a tree-based model been selected, SHAP interaction values between the ability-to-reach and availability domains would have tested whether poor access and low capacity combine more than additively. The stability of feature rankings across outer folds was summarised with Spearman's rank correlation. Because SHAP values can be distorted when features are correlated [42], we also report grouped permutation importance by domain. SHAP values describe how the model uses its inputs; they are associational and are not interpreted causally.

Risk mapping and survey visibility

The final model predicted the probability of high exclusion for every populated 1-km cell. Uncertainty was quantified from 200 bootstrap refits of the final model on resampled clusters, giving 95% intervals and the probability that each cell's risk exceeded 0.50; where the hybrid model was selected, its kriged residual surface was added as a fixed component. Cell predictions were aggregated to wards and LGAs as population-weighted means and grouped into four tiers: Low (< 0.25), Moderate (0.25 to < 0.50), High (0.50 to < 0.75) and Very High (≥ 0.75). A ward or LGA was classed as **survey-visible** if it contained at least one NDHS cluster and **survey-invisible** otherwise. We report the number of High and Very High units, and the population living in them, by zone and visibility status. Because DHS coordinates are displaced, some clusters will be assigned to a neighbouring unit; this affects the visibility classification near boundaries but not the predictions themselves. In line with the DHS geographic data policy, no cluster locations are shown on any map.

Benchmarking. Because a model calibrated at survey clusters may not be calibrated across the whole population surface, the 1-km predictions were benchmarked to direct survey estimates [44]. For each zone, a single logit-scale shift was found such that the population-weighted mean of the shifted cell predictions equalled the survey-weighted direct estimate of the proportion of the zone's population living in high-exclusion clusters; the same shift was applied to the bootstrap interval limits. Benchmarking preserves the model's ranking of cells within each zone. Benchmarked predictions are used for the primary maps and Table 6, and model-only predictions are reported in Additional file 1: Table S7 and Additional file 1: Figure S1.

Sensitivity analyses

Prespecified sensitivity analyses varied the outcome threshold (0.40 and 0.60); modelled each tracer service separately; modelled the continuous exclusion score with regression; restricted zero-dose to children aged 12–23 months; excluded clusters with fewer than ten eligible persons; trained with weights equal to the number of eligible persons and with survey design weights; varied relative road speeds as described above; replaced the OpenStreetMap-derived travel times with those from the global friction surface of Weiss and colleagues [5]; used 10-km extraction buffers; and fitted a model restricted to NDHS, GRID3 and OpenStreetMap inputs, without night-time lights and conflict data.

Software and reproducibility

The entire workflow, from the raw NDHS, GRID3, OpenStreetMap, VIIRS and ACLED files to every table and figure, is contained in a single Stata 17 master do-file. Survey-design estimation and the DeLong and fractional-logit analyses run in Stata; geospatial extraction and model fitting run in Python 3.11.7 blocks stored inside the same do-file and executed from it (geopandas, rasterio, scikit-image, scikit-learn, xgboost, shap, pykrige, libpysal, esda). Package versions are written to the output workbook. The code is available from the corresponding author on reasonable request; NDHS data can be obtained from the DHS Program on request.

Ethical considerations

The 2023–24 NDHS protocol was approved by the ICF Institutional Review Board and the National Health Research Ethics Committee of Nigeria, and all participants gave informed consent [3]. This study is a secondary analysis of publicly available, anonymised data released by the DHS Program with permission; no identifiable information was accessed and no additional ethical approval was required. This study was conducted in accordance with the Declaration of Helsinki. All geospatial inputs are openly licensed and are cited and attributed.

Results

Study sample and distribution of exclusion

All 1,380 completed NDHS clusters had valid coordinates. Of these, 1,270 met the eligibility rule and had complete predictor data and formed the analytic sample (Fig. 1). In total, 425 clusters (33.5%) were classified as high exclusion. The proportion varied sharply by zone, from 3.6% in the South East, 6.5% in the South West and 14.4% in the South South to 34.4% in the North Central, 48.5% in the North East and 74.5% in the North West, where the median cluster exclusion score was 0.68.

High- and low-exclusion clusters differed most in facility capacity and settlement pattern (Table 3). High-exclusion clusters had a median of 4 facilities within 5 km compared with 24 in low-exclusion clusters, and a median of 0 versus 11 secondary or tertiary facilities within 10 km. The private share of nearby facilities was 0.00 versus 0.30 (standardised mean difference [SMD] − 1.33), and the median population per facility was 6,370 versus 2,966. High-exclusion clusters were also further from secondary or tertiary care (median 11.2 versus 1.9 km; SMD 1.08), had lower road density (SMD − 1.20), lower night-time light radiance (SMD − 1.08) and a much smaller share of their population living in built-up areas (SMD − 1.37).

Table 3
Characteristics of clusters by exclusion status

Domain	Characteristic, median (IQR)	High exclusion (n = 425)	Low exclusion (n = 845)	SMD
Availability and accommodation	Facilities within 5 km	4.00 (2.00 to 8.00)	24.00 (10.00 to 61.00)	-0.86
Availability and accommodation	Primary facilities within 5 km	4.00 (2.00 to 7.00)	19.00 (9.00 to 46.00)	-0.88
Availability and accommodation	Secondary/tertiary facilities within 10 km	0.00 (0.00 to 2.00)	11.00 (2.00 to 37.00)	-0.8
Availability and accommodation	Distance to nearest facility (km)	1.72 (0.86 to 2.79)	0.73 (0.37 to 1.35)	0.74
Availability and accommodation	Distance to nearest secondary/tertiary facility (km)	11.24 (5.19 to 19.39)	1.88 (0.90 to 4.84)	1.08
Availability and accommodation	Population per facility within 10 km	6369.70 (3866.62 to 10474.66)	2966.40 (1960.73 to 4548.36)	0.6
Availability and accommodation	No facility within 10 km (flag)	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.00)	0.17
Availability and accommodation	Private share of facilities within 10 km	0.00 (0.00 to 0.13)	0.30 (0.16 to 0.46)	-1.33
Ability to reach	Motorised travel time to nearest primary facility (min)	8.85 (5.74 to 14.13)	2.36 (1.17 to 5.23)	0.79
Ability to reach	Motorised travel time to nearest secondary/tertiary facility (min)	26.76 (14.06 to 41.91)	4.24 (2.09 to 11.01)	0.81
Ability to reach	Walking time to nearest facility (min)	32.80 (22.03 to 46.64)	12.96 (7.61 to 21.92)	0.9
Ability to reach	Road density within 5 km (km/km2)	0.68 (0.43 to 1.07)	1.52 (0.94 to 1.86)	-1.2
Ability to reach	Major-road density within 5 km (km/km2)	0.14 (0.00 to 0.20)	0.30 (0.16 to 0.44)	-1.06
Ability to reach	Distance to nearest major road (km)	2.66 (1.81 to 6.68)	0.99 (0.60 to 2.00)	0.61
Ability to pay	Night-time light radiance (log)	0.00 (0.00 to 0.05)	0.52 (0.01 to 1.50)	-1.08
Context: settlement	Population density (persons/km2)	376.89 (201.37 to 633.60)	1253.15 (491.66 to 2392.27)	-0.85
Context: settlement	Share of population in built-up areas	0.00 (0.00 to 0.27)	0.85 (0.40 to 0.98)	-1.37
Context: settlement	Share of population in small settlements	0.15 (0.06 to 0.23)	0.00 (0.00 to 0.10)	0.7
Context: settlement	Share of population in hamlets	0.02 (0.01 to 0.04)	0.00 (0.00 to 0.01)	0.65
Context: insecurity	Conflict events within 10 km	1.00 (0.00 to 5.00)	10.00 (2.00 to 29.00)	-0.74
Context: insecurity	Conflict fatalities within 10 km	0.00 (0.00 to 6.00)	8.00 (1.00 to 26.00)	-0.14

IQR, interquartile range; SMD, standardised mean difference (difference in means divided by the pooled standard deviation).

Model performance ($H0_1$)

Under the primary, state-blocked spatial cross-validation, elastic-net logistic regression achieved the highest discrimination, with an AUC of 0.868 (95% CI 0.844 to 0.888), a Brier score of 0.134 (scaled Brier 0.397), a calibration slope of 0.97 and

calibration-in-the-large of 0.01 (Table 4; Fig. 2). The three tree-based algorithms performed similarly to one another (AUC 0.858 to 0.859). Discrimination differed significantly across the four algorithms (DeLong $\chi^2 = 14.53$, $p = 0.002$), so $H0_1$ was rejected: after Holm adjustment, logistic regression outperformed random forest ($p = 0.022$), gradient boosting ($p = 0.003$) and XGBoost ($p = 0.004$), whereas the tree-based models did not differ from one another (all adjusted $p = 1.00$).

Random cross-validation overstated performance for every algorithm, by 0.016 to 0.023 in AUC for the four base models. Performance remained good when whole zones were withheld (logistic regression AUC 0.845, 95% CI 0.821 to 0.867), ranging from 0.690 when the South East, where only 3.6% of clusters were high exclusion, was predicted from the other zones, to 0.865 for the North Central zone. Adding residual kriging to the best tree-based model (random forest) increased the spatially validated AUC to 0.883 (DeLong $p = 0.003$ versus random forest alone) but produced over-dispersed predictions (calibration slope 0.53), so under the prespecified selection rule the calibrated logistic regression was retained as the final model. At the Youden-optimal threshold of 0.365, the final model had a sensitivity of 0.819, specificity of 0.808, positive predictive value of 0.682 and negative predictive value of 0.899 (Additional file 1: Table S4). Decision-curve analysis showed a positive net benefit, greater than that of targeting all areas or none, across threshold probabilities from 0.20 to 0.79 (Fig. 3). Out-of-fold residuals remained spatially autocorrelated (Moran's I 0.234, $p = 0.001$), less so for the hybrid model (0.125).

Table 4
Predictive performance by algorithm and validation scheme

Model	Validation	AUC (95% CI)	Brier	Scaled Brier	Calibration slope	CITL	Sensitivity	Specificity	F1	Kappa
Elastic-net logistic regression	Random 10-fold	0.884 (0.863 to 0.903)	0.126	0.434	1.01	-0.01	0.729	0.882	0.743	0.616
Elastic-net logistic regression	Spatial (state-blocked) 10-fold	0.868 (0.844 to 0.888)	0.134	0.397	0.97	0.01	0.704	0.882	0.726	0.594
Elastic-net logistic regression	Leave-one-zone-out	0.845 (0.821 to 0.867)	0.149	0.332	0.88	0.1	0.607	0.886	0.662	0.515
Random forest	Random 10-fold	0.881 (0.859 to 0.899)	0.128	0.426	1.02	-0.01	0.751	0.879	0.754	0.631
Random forest	Spatial (state-blocked) 10-fold	0.859 (0.835 to 0.881)	0.139	0.374	0.96	0.01	0.685	0.871	0.705	0.564
Random forest	Leave-one-zone-out	0.835 (0.810 to 0.858)	0.154	0.31	0.91	0.07	0.598	0.872	0.645	0.488
Gradient boosting	Random 10-fold	0.877 (0.855 to 0.896)	0.128	0.423	0.99	0	0.734	0.875	0.74	0.611
Gradient boosting	Spatial (state-blocked) 10-fold	0.858 (0.833 to 0.879)	0.138	0.381	0.99	0.02	0.68	0.876	0.706	0.566
Gradient boosting	Leave-one-zone-out	0.826 (0.800 to 0.851)	0.153	0.313	0.88	0.09	0.619	0.866	0.657	0.499
XGBoost	Random 10-fold	0.881 (0.859 to 0.900)	0.127	0.431	0.93	0	0.753	0.877	0.754	0.63
XGBoost	Spatial (state-blocked) 10-fold	0.858 (0.833 to 0.880)	0.138	0.382	0.86	0.05	0.694	0.871	0.712	0.572
XGBoost	Leave-one-zone-out	0.834 (0.809 to 0.857)	0.153	0.311	0.78	0.12	0.626	0.865	0.661	0.504
Random forest + residual kriging	Random 10-fold	0.917 (0.901 to 0.933)	0.105	0.526	0.7	-0.1	0.8	0.896	0.797	0.695

Model	Validation	AUC (95% CI)	Brier	Scaled Brier	Calibration slope	CITL	Sensitivity	Specificity	F1	Kappa
Random forest + residual kriging	Spatial (state-blocked) 10-fold	0.883 (0.862 to 0.905)	0.125	0.439	0.53	-0.02	0.727	0.892	0.749	0.629
Random forest + residual kriging	Leave-one-zone-out	0.877 (0.856 to 0.896)	0.135	0.394	0.52	0.13	0.675	0.88	0.706	0.568

CITL, calibration-in-the-large. Classification metrics at a threshold of 0.50; results at the Youden-optimal threshold are in Additional file 1: Table S4. The final model was selected as the model with the highest spatially validated AUC among those with a calibration slope between 0.80 and 1.25.

Added value of geographic accessibility ($H0_2$)

A model containing only settlement, insecurity and night-time light features reached an AUC of 0.810 (Table 5). Adding the availability and accommodation domain raised the AUC by 0.061 ($p < 0.001$) and reduced the Brier score by 0.031 (95% CI 0.024 to 0.039). Adding the ability-to-reach domain on top of this did not improve discrimination: the AUC changed by − 0.003 and the Brier score by + 0.001 (95% CI − 0.0004 to 0.0026). Accessibility features therefore added no discrimination beyond facility capacity and context, and $H0_2$ was not rejected.

Table 5
Nested feature-set comparison using elastic-net logistic regression (state-blocked spatial cross-validation)

Feature set	AUC (95% CI)	Delta AUC vs previous (DeLong p)	Brier	Delta Brier (95% CI)
M1: context + ability to pay	0.810 (0.785 to 0.834)	—	0.164	—
M2: M1 + availability and accommodation	0.871 (0.848 to 0.891)	+ 0.061 (p = 0.0000)	0.133	-0.0309 (-0.0388 to -0.0239)
M3: M2 + ability to reach	0.868 (0.844 to 0.888)	-0.003 (p = 0.0355)	0.134	0.0011 (-0.0004 to 0.0026)

ΔAUC tested with DeLong's test; ΔBrier with 1,000 bootstrap resamples. The small, statistically significant change from M2 to M3 is a decrease in AUC.

Drivers of predicted exclusion (Objective 3 and $H0_3$)

The availability and accommodation domain accounted for 64.5% of the total mean absolute SHAP value of the final model, followed by ability to reach (13.8%), ability to pay (8.1%), insecurity (7.1%) and settlement pattern (6.6%) (Fig. 4). Grouped permutation confirmed this ordering: permuting the availability domain reduced the out-of-fold AUC by 0.285 on average, against 0.011 for ability to reach and less than 0.006 for each remaining domain. The leading individual features were the private share of facilities within 10 km (18.1% of total SHAP), population per facility (12.0%), the number of facilities within 5 km (6.0%), night-time light radiance (5.8%) and conflict fatalities within 10 km (5.7%). Feature rankings were stable across spatial folds (mean pairwise Spearman ρ 0.78, minimum 0.61). The contribution of the availability domain was largest where predicted risk was highest (mean absolute SHAP 0.342 among clusters with predicted risk ≥ 0.50, against 0.067 among those with predicted risk between 0.25 and 0.49).

SHAP dependence plots showed that predicted risk rose steadily with population per facility, crossing from a negative to a positive contribution at roughly 4,000 persons per facility, and fell steeply as the number of facilities within 5 km rose from zero to about ten, with little further change beyond that (Fig. 5). Motorised travel time to the nearest primary facility made only small

contributions (absolute SHAP values below 0.03). Because the final model is a penalised regression, SHAP interaction values, which require a tree-based model, were not estimated.

In the fractional logistic regression of the continuous exclusion score, population per facility was not independently associated with exclusion once the other features were included (odds ratio per interquartile-range increase of 3,753 persons per facility 1.036, 95% CI 0.967 to 1.110; p = 0.313), so $H0_3$ was not rejected. Other capacity measures were associated with lower exclusion: the private share of nearby facilities (coefficient − 1.82, p = 0.001), the number of primary facilities within 5 km (− 0.018 per facility, p = 0.002) and the number of secondary or tertiary facilities within 10 km (− 0.006 per facility, p = 0.042). Exclusion was also lower where more of the population lived in built-up areas (p < 0.001), and higher with population density (p < 0.001) and conflict fatalities (p = 0.046).

Figure 4. (a) SHAP summary plot for the 15 most influential features of the final model (probability scale).

Risk maps and survey-invisible exclusion (Objective 4)

Before benchmarking, the population-weighted mean predicted risk across all populated 1-km cells was 0.538, above the direct survey estimate. The discrepancy was concentrated in the South South (model 0.382 versus direct 0.123), South West (0.281 versus 0.060), South East (0.127 versus 0.022) and North East (0.708 versus 0.446), whereas the model agreed closely with the direct estimate in the North West (0.682 versus 0.711) (Additional file 1: Table S7). After within-zone benchmarking, the national population-weighted mean was 0.414.

On the benchmarked surface, 2,534 of 9,014 wards (28.1%) and 231 of 774 LGAs (29.8%) were classified as High or Very High risk, home to an estimated 104.7 million people (Table 6; Fig. 6). Of these wards, 2,239 (88.4%) contained no NDHS cluster, and they accounted for 88.1 million people, or 84.2% of the population living in High or Very High risk wards. The burden was heavily concentrated in the north: the North West contained 1,633 High or Very High wards and 160 of its 186 LGAs, with 60.0 million people, followed by the North East (23.5 million) and the North Central zone (20.6 million). No ward in the South East or South West, and 25 wards in the South South, were classified as High or Very High risk. Prediction intervals were widest in the North East and along parts of the southern coastline (Fig. 6b). Using the model-only surface, 3,845 wards and 145.9 million people would have been classified as High or Very High risk (Additional file 1: Table S7; Additional file 1: Figure S1).

Table 6
Wards and LGAs at High or Very High predicted risk of exclusion and their survey visibility, by zone (benchmarked predictions)

Zone	LGAs (n)	LGAs High/Very High	of which survey-invisible	Wards (n)	Wards High/Very High	of which survey-invisible	Population in High/Very High wards (millions)	of which survey-invisible (millions)	Mean predicted risk
North Central	121	36	3	1,484	427	373	20.6	16.57	0.445
North East	112	34	6	1,306	449	393	23.46	19.03	0.447
North West	186	160	29	1,972	1,633	1,450	60	51.93	0.711
South East	95	0	0	1,378	0	0	0	0	0.022
South South	123	1	1	1,366	25	23	0.6	0.54	0.123
South West	137	0	0	1,508	0	0	0	0	0.061
Nigeria	774	231	39	9,014	2,534	2,239	104.66	88.08	0.414

Survey-invisible: no NDHS 2023–24 cluster located in the unit. Predictions benchmarked within each zone to the direct survey estimate (Additional file 1: Table S7). Mean predicted risk is population weighted.

Agreement with direct survey estimates ($H0_4$)

Leave-one-state-out predictions agreed moderately well with direct state estimates (Lin's CCC 0.656, 95% CI 0.515 to 0.758), so $H0_4$ was rejected (Fig. 7). The model-based estimate lay within the 95% confidence interval of the direct estimate in 12 of the 37 states. The model overestimated exclusion in several southern states where no surveyed cluster was highly excluded (for example, Edo 0.43 and Cross River 0.35 against direct estimates of 0) and in Borno (0.73 versus 0.27) and Adamawa (0.66 versus 0.24), and underestimated it in the north-western states with the highest direct estimates, including Kebbi (0.66 versus 0.99), Zamfara (0.72 versus 0.96) and Sokoto (0.62 versus 0.89).

Sensitivity analyses

Conclusions were robust across the 14 sensitivity analyses that were run (Additional file 1: Table S5). Spatially validated AUCs ranged from 0.761 when only zero-dose status defined the outcome to 0.877 with 10-km extraction buffers, and calibration slopes ranged from 0.85 to 0.99. Changing the outcome threshold to 0.40 or 0.60 (AUC 0.874 and 0.861), restricting zero-dose status to children aged 12–23 months (0.849), excluding clusters with fewer than ten eligible persons (0.852), weighting training by eligible count (0.863) or by survey design (0.860), and altering relative road speeds (0.868 in both directions) changed discrimination by at most 0.019. A model without night-time lights and conflict data performed almost identically (0.865). Modelling the continuous exclusion score explained 49.2% of its variance. The prespecified comparison with friction-surface travel times was not performed because that input was not obtained. The results of the four hypothesis tests are summarised in Table 7.

Table 7
Summary of hypothesis tests

Hypothesis	Test	Result	Decision
$H0_1$ No difference in discrimination between the four algorithms	DeLong test (overall; pairwise Holm-adjusted in sheet H01_delong_pairwise)	chi2 = 14.53, p = 0.0023	Reject H0
$H0_2$ Accessibility features add no discrimination	DeLong test, M2 vs M3; Delta Brier	chi2 = 4.42, p = 0.0355	Do not reject H0 (no improvement)
$H0_3$ No association of population per facility with exclusion	Fractional logit (SE clustered by state); SHAP dependence	OR per IQR = 1.036 (0.967 to 1.110), p = 0.3131	Do not reject H0
$H0_4$ No concordance between model-based and direct state estimates	Lin's concordance correlation coefficient (bootstrap CI)	CCC = 0.656 (0.515 to 0.758); 12/37 states within direct 95% CI	Reject H0

Discussion

Principal findings

Using only data that exist for every location in Nigeria, a penalised spline logistic regression predicted which NDHS clusters were highly excluded from maternal and child health services with good discrimination and close calibration under state-blocked spatial validation, and it retained most of its accuracy when whole geopolitical zones were withheld. The model's predictions were driven overwhelmingly by facility capacity, in particular the private share of nearby facilities and the number of people per facility, while modelled travel time added nothing once capacity was known. Applied to every populated square kilometre and benchmarked to zone-level survey estimates, the model identified 2,534 wards and about 105 million people at High or Very High risk of exclusion. Most of this population, 88 million people, lives in wards where the 2023–24 survey placed no cluster, and more than half of it lives in the North West.

Comparison with previous studies

Our finding that a well-specified penalised regression outperformed random forest, gradient boosting and XGBoost is consistent with systematic evidence that machine learning rarely improves on logistic regression for tabular prediction problems [31], and with Utazi and colleagues' observation that some machine-learning approaches to coverage mapping over-smooth the predicted surface [10]. Kawakatsu and colleagues mapped maternal and child health services in Nigeria with gradient boosting and generalised additive models from earlier surveys [9]; our results extend that work by using a multi-service outcome, the 2023–24 survey, the national multi-ownership facility registry and spatial rather than random validation, which overstated discrimination by up to 0.023 in our data. Residual kriging recovered spatial structure that the covariates missed and improved discrimination, but at the cost of calibration, a trade-off that is rarely reported when hybrid models are selected on discrimination alone.

The limited contribution of travel time contrasts with the prominence of geographic accessibility in the literature [4–6]. One explanation is that, at the cluster scale, facility density and travel time carry much of the same information, and capacity measures additionally reflect whether there are enough facilities for the population they serve. Another is that the private share of facilities, the single strongest predictor, captures more than physical supply: it marks local markets that can sustain fee-charging providers, and so partly reflects economic and social conditions that the night-time light proxy measures only crudely. This feature should therefore be read as a marker of under-served areas rather than as evidence that private provision itself prevents exclusion.

Two findings concern transport of the model across space. First, the model was well calibrated at survey clusters but, before benchmarking, overestimated the population share at high risk in the South South, South West and North East. The likely cause is covariate shift: many southern rural cells have few nearby facilities and few private providers and so resemble highly excluded northern clusters, although surveys show low exclusion there. Benchmarking predictions to direct survey estimates, as is standard in small-area estimation [44], removes this bias at zone level while retaining the model's ranking of wards within each zone. Second, the remaining discrepancies with state estimates were systematic: the model overestimated exclusion in Borno and Adamawa, where surveyed clusters lie mainly in accessible areas with humanitarian services and insecure areas were excluded from fieldwork, and underestimated it in the north-western states with the highest direct estimates. Direct estimates are themselves imprecise at state level, and the comparison should be read with this in mind.

Implications for policy and practice

The domain decomposition points to where action is most likely to reduce exclusion. Because capacity, not distance, distinguished high- from low-exclusion areas, the first priority in high-risk wards is to increase the number and functionality of primary health care facilities relative to the population they serve, rather than to rely on transport or road investment alone. The concentration of high-risk wards in the North West, and the fact that most of them have never been sampled by a national survey, argue for using model-based ward risk as a routine input to state annual operational plans and to the targeting of outreach, immunisation catch-up and community-based maternal care. Because the predictors are refreshed regularly by GRID3, OpenStreetMap and conflict monitoring, the model can be re-run between survey rounds to flag wards whose risk is rising before the next survey reveals it.

Strengths and limitations

This study has several strengths. It uses only predictors available everywhere in Nigeria, the most recent national survey and facility registry, nested state-blocked and leave-one-zone-out validation, a calibration-aware model-selection rule, decision-curve analysis, explanations organised by an explicit theory of access, and benchmarking of the final maps to survey estimates. The entire analysis is reproducible from a single script.

This study has some limitations. First, the analysis is ecological: it predicts the risk that an area is highly excluded, not the risk for any individual. Second, displacement of cluster coordinates introduces error into extracted covariates, which tends to

weaken associations and blur predicted surfaces. Third, GRID3 records the existence and type of facilities but not whether they are staffed, equipped or open, so capacity is only partly measured, and the acceptability, appropriateness and cost dimensions of access cannot be captured from geospatial data. Fourth, the available sample fell short of the Riley criteria (1,642 clusters required for 21 features, and more for the spline parameterisation, against 1,270 available). Strong penalisation and calibration slopes close to one under spatial validation suggest that overfitting was limited, but the shortfall should be noted. Fifth, residual spatial autocorrelation remained, indicating structure that the covariates do not capture. Sixth, 20 clusters were not surveyed because of insecurity, so predictions for the most insecure areas are extrapolations, and benchmarking to zone estimates assumes that the surveyed population represents each zone. Seventh, wards are classed as survey-visible according to displaced coordinates, which may misclassify wards near boundaries. Finally, differences between gridded population products can change estimates of the population affected [43].

Conclusions

Health-service exclusion in Nigeria can be predicted with good accuracy from openly available geospatial data, and the prediction rests mainly on facility capacity relative to population rather than on travel distance. Most people living in high-risk wards are in places that national surveys do not reach. Benchmarked, ward-level risk maps of this kind can help planners direct primary health care investment and outreach to excluded communities before the next survey makes their exclusion visible.

Abbreviations

ACLED
Armed Conflict Location and Event Data
AUC
area under the receiver operating characteristic curve
CCC
concordance correlation coefficient
CI
confidence interval
CITL
calibration-in-the-large
DHS
Demographic and Health Surveys
DTP
diphtheria–tetanus–pertussis
GRID3
Geo-Referenced Infrastructure and Demographic Data for Development
IQR
interquartile range
LGA
local government area
NDHS
Nigeria Demographic and Health Survey
OSM
OpenStreetMap
SHAP
SHapley Additive exPlanations
SMD
standardised mean difference
TRIPOD + AI

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, updated for artificial intelligence
VIIRS
Visible Infrared Imaging Radiometer Suite.

Declarations

Ethics approval and consent to participate

The 2023–24 NDHS protocol was reviewed and approved by the ICF Institutional Review Board and the National Health Research Ethics Committee of Nigeria, and all participants gave informed consent before interview. This study is a secondary analysis of publicly available, anonymised data released by the DHS Program with permission; no identifiable information was accessed, and no additional ethical approval was required. This study was conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Availability of data and materials

The NDHS datasets are available from the DHS Program (https://dhsprogram.com) on approval of a request. GRID3, OpenStreetMap, VIIRS and ACLED data are publicly available from the sources listed in Table 1. The analysis do-file, the ward-level and LGA-level predictions and the output workbook are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Authors' contributions

ASO conceived the study, designed the methodology, curated the data, wrote the analysis code, performed the formal analysis, produced the visualisations and wrote the original draft. JWM contributed to the survey-data methodology and validation and reviewed and edited the manuscript. ADO supervised the study, contributed to validation and reviewed and edited the manuscript. OIB contributed to the statistical methodology and validation and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank the DHS Program and ICF for access to the NDHS data, and the GRID3 programme, OpenStreetMap contributors, the Earth Observation Group and ACLED for their open data.

Declaration of AI use

AI tools were used to help structure the text, edit the language and generate the analysis code; the authors reviewed and edited all content, ran and checked all analyses, and take full responsibility for the manuscript.

References

1. World Health Organization, World Bank. Tracking universal health coverage: 2023 global monitoring report. Geneva: World Health Organization; 2023.
2. Abubakar I, Dalglish SL, Angell B, et al. The Lancet Nigeria Commission: investing in health and the future of the nation. Lancet. 2022;399(10330):1155–200. 10.1016/S0140-6736(21)02488-0.
3. Federal Ministry of Health and Social Welfare of Nigeria (FMoHSW). National Population Commission (NPC) [Nigeria], ICF. Nigeria Demographic and Health Survey 2023–24: Key Indicators Report. Abuja, Nigeria, and Rockville. Maryland, USA: NPC and ICF; 2024.
4. Ouma PO, Maina J, Thuranira PN, et al. Access to emergency hospital care provided by the public sector in sub-Saharan Africa in 2015: a geocoded inventory and spatial analysis. Lancet Glob Health. 2018;6(3):e342–50.
5. Weiss DJ, Nelson A, Vargas-Ruiz CA, et al. Global maps of travel time to healthcare facilities. Nat Med. 2020;26(12):1835–8.
6. Banke-Thomas A, Wong KLM, Olubodun T, et al. Geographical accessibility to functional emergency obstetric care facilities in urban Nigeria using closer-to-reality travel time estimates: a population-based spatial analysis. Lancet Glob Health. 2024;12(5):e848–58.
7. Utazi CE, Thorley J, Alegana VA, et al. High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. Vaccine. 2018;36(12):1583–91. 10.1016/j.vaccine.2018.02.020.
8. Bosco C, Alegana V, Bird T, et al. Exploring the high-resolution mapping of gender-disaggregated development indicators. J R Soc Interface. 2017;14(129):20160825. 10.1098/rsif.2016.0825.
9. Kawakatsu Y, Mosser JF, Adolph C, et al. High-resolution mapping of essential maternal and child health service coverage in Nigeria: a machine learning approach. BMJ Open. 2024;14(6):e080135. 10.1136/bmjopen-2023-080135.
10. Utazi CE, Yankey O, Chaudhuri S, et al. Geostatistical and machine learning approaches for high-resolution mapping of vaccination coverage. Spat Spatiotemporal Epidemiol. 2025;54:100744. 10.1016/j.sste.2025.100744.
11. Roberts DR, Bahn V, Ciuti S, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913–29.
12. Ploton P, Mortier F, Réjou-Méchain M, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat Commun. 2020;11:4540.
13. Maina J, Ouma PO, Macharia PM, et al. A spatial database of health facilities managed by the public health sector in sub Saharan Africa. Sci Data. 2019;6:134. 10.1038/s41597-019-0142-2.
14. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Glob Health. 2018;3(4):e000798.
15. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74.
16. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67.
17. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
18. Levesque JF, Harris MF, Russell G. Patient-centred access to health care: conceptualising access at the interface of health systems and populations. Int J Equity Health. 2013;12:18.
19. Penchansky R, Thomas JW. The concept of access: definition and relationship to consumer satisfaction. Med Care. 1981;19(2):127–40.
20. Shapley LS. A value for n-person games. In: Kuhn HW, Tucker AW, editors. Contributions to the Theory of Games II. Princeton: Princeton University Press; 1953. pp. 307–17.
21. Burgert CR, Colston J, Roy T, Zachary B. Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys. DHS Spatial Analysis Reports No. 7. Calverton. Maryland, USA: ICF International; 2013.
22. GRID3. GRID3 NGA – Health Facilities v2.0. Geo-Referenced Infrastructure and Demographic Data for Development. 2024. Available from: https://data.grid3.org

23. GRID3. GRID3 NGA – Gridded Population Estimates v3.0. 2025. Available from: https://data.grid3.org
24. GRID3. GRID3 NGA – Settlement Extents v4.1. 2026. Available from: https://data.grid3.org
25. OpenStreetMap contributors. Nigeria extract, 14 September 2026. Geofabrik GmbH. Available from: https://download.geofabrik.de/africa/nigeria.html
26. Elvidge CD, Zhizhin M, Ghosh T, Hsu FC, Taneja J. Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. Remote Sens. 2021;13(5):922.
27. Raleigh C, Linke A, Hegre H, Karlsen J. Introducing ACLED: an armed conflict location and event dataset. J Peace Res. 2010;47(5):651–60.
28. World Health Organization. Immunization Agenda 2030: a global strategy to leave no one behind. Geneva: World Health Organization; 2020.
29. Perez-Haydrich C, Warren JL, Burgert CR, Emch ME. Guidelines on the use of DHS GPS data. DHS Spatial Analysis Reports No. 8. Calverton. Maryland, USA: ICF International; 2013.
30. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.
31. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22.
32. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
33. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
34. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. pp. 785–94. 10.1145/2939672.2939785
35. Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ. 2018;6:e5518.
36. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
37. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
38. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26(6):565–74.
39. Moran PAP. Notes on continuous stochastic phenomena. Biometrika. 1950;37(1–2):17–23.
40. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
41. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255–68.
42. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artif Intell. 2021;298:103502.
43. Hierink F, Boo G, Macharia PM, et al. Differences between gridded population data impact measures of geographic access to healthcare in sub-Saharan Africa. Commun Med (Lond). 2022;2:117.
44. Wakefield J, Okonek T, Pedersen J. Small area estimation for disease prevalence mapping. Int Stat Rev. 2020;88(2):398–418.

Figures

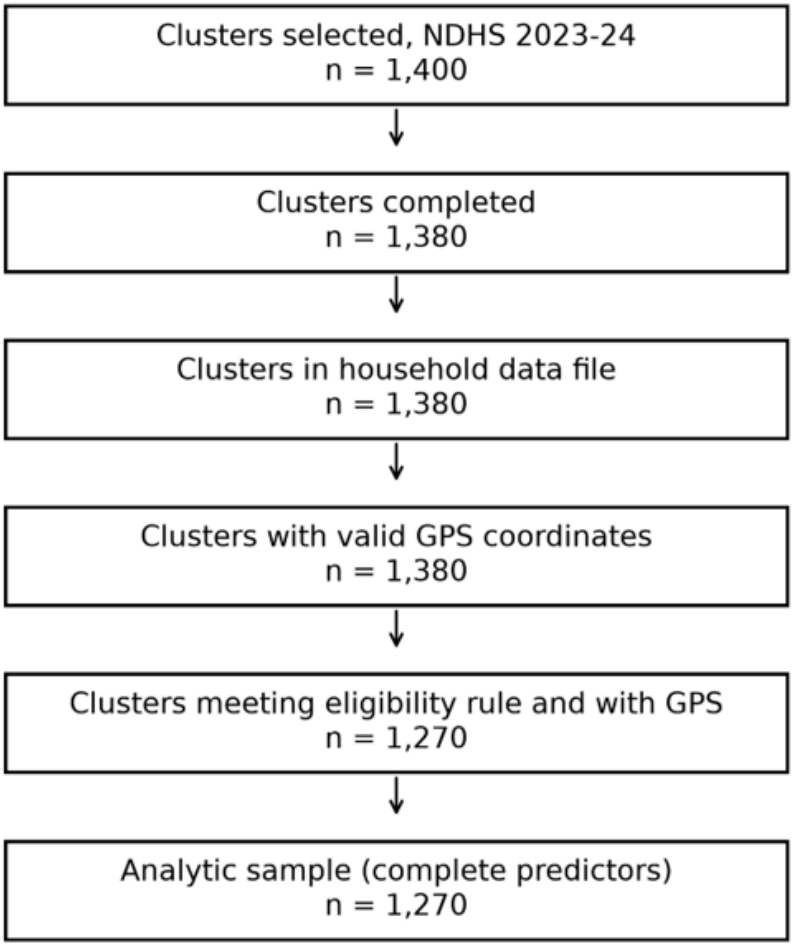

Figure 1

Flow of NDHS 2023–24 clusters into the analytic sample.

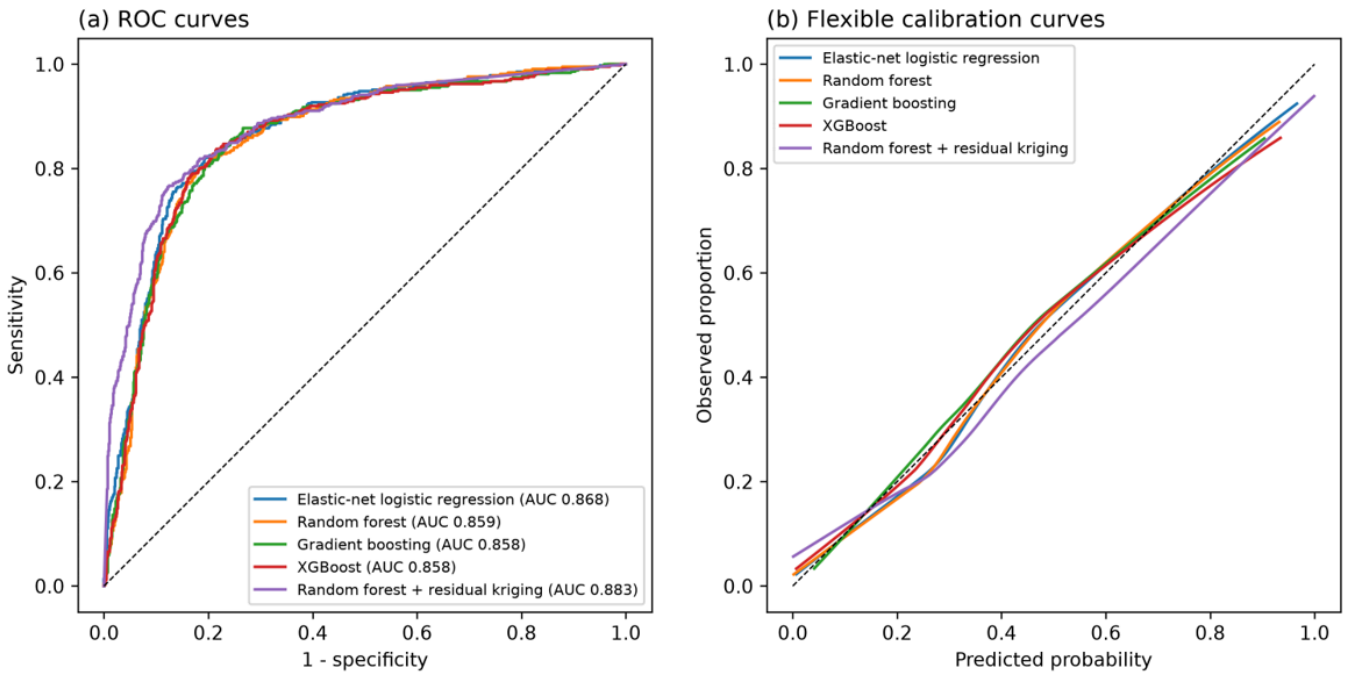

Figure 2

(a) Receiver operating characteristic curves and (b) flexible calibration curves under state-blocked spatial cross-validation.

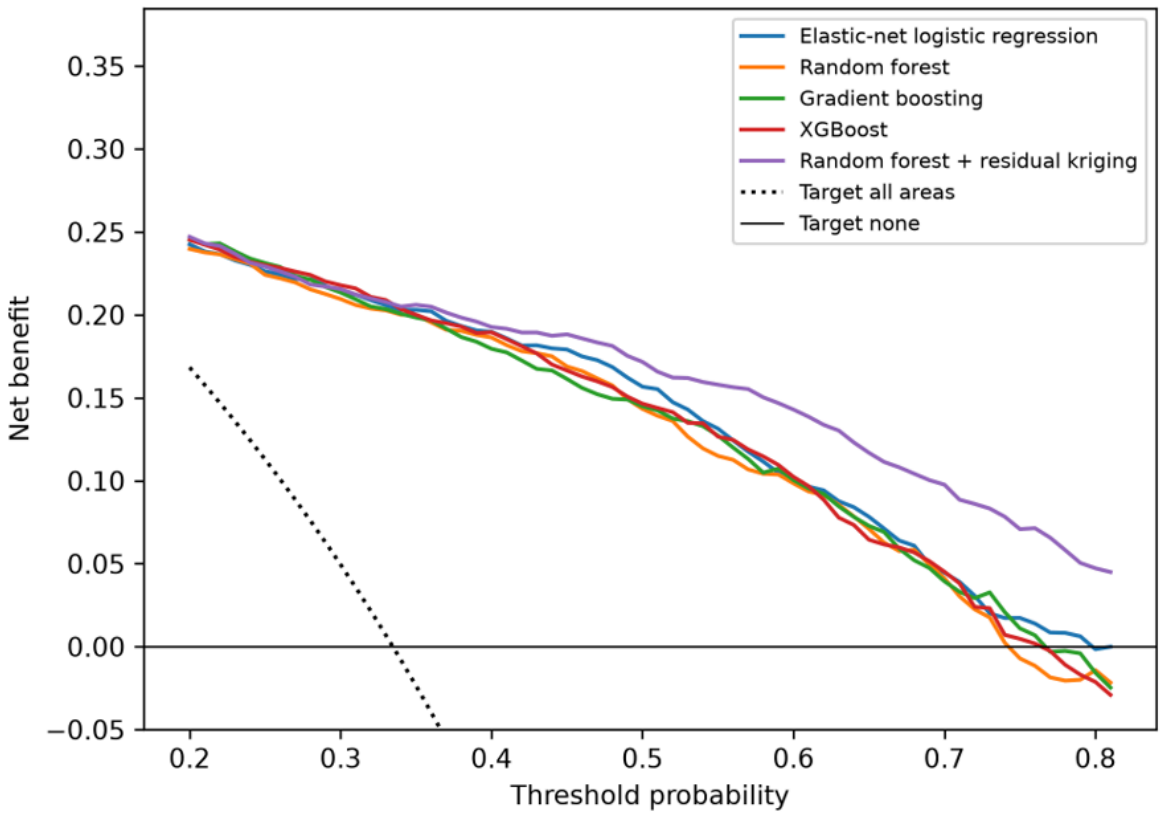

Figure 3

Decision curves for targeting areas for outreach under state-blocked spatial cross-validation.

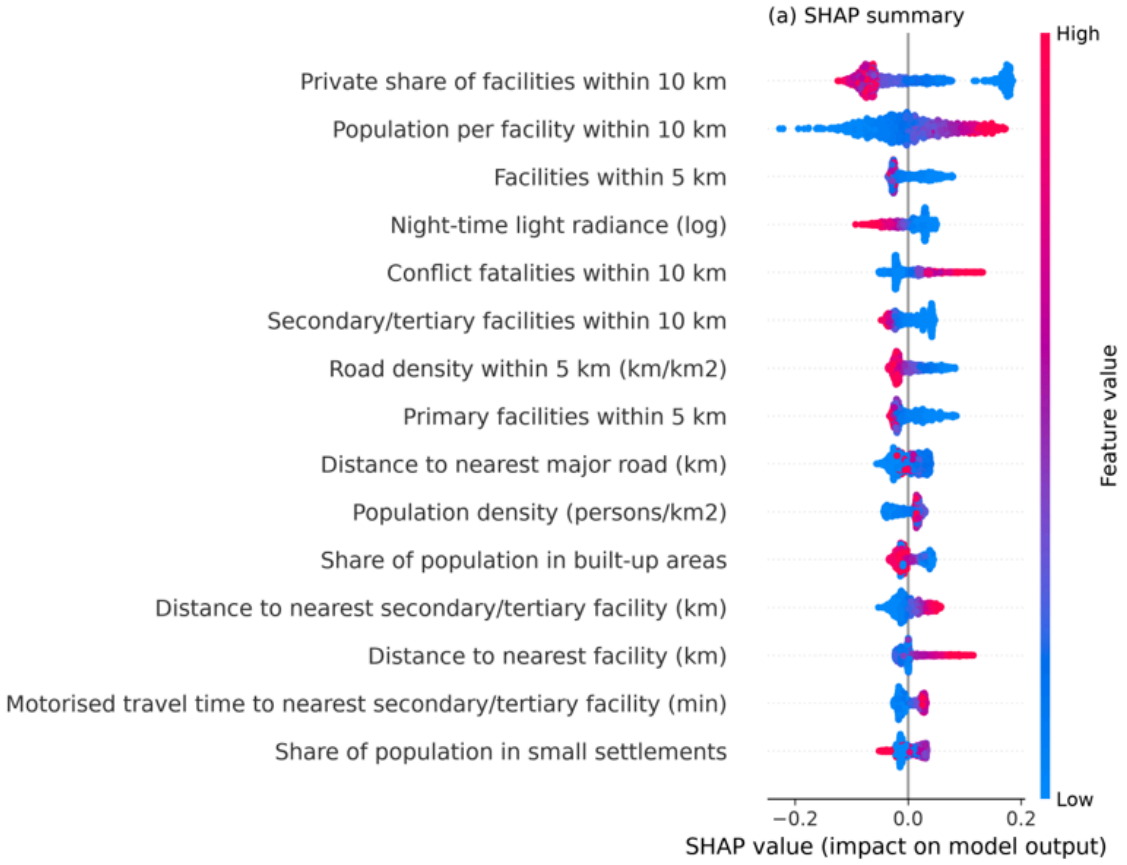

Figure 4. (a)

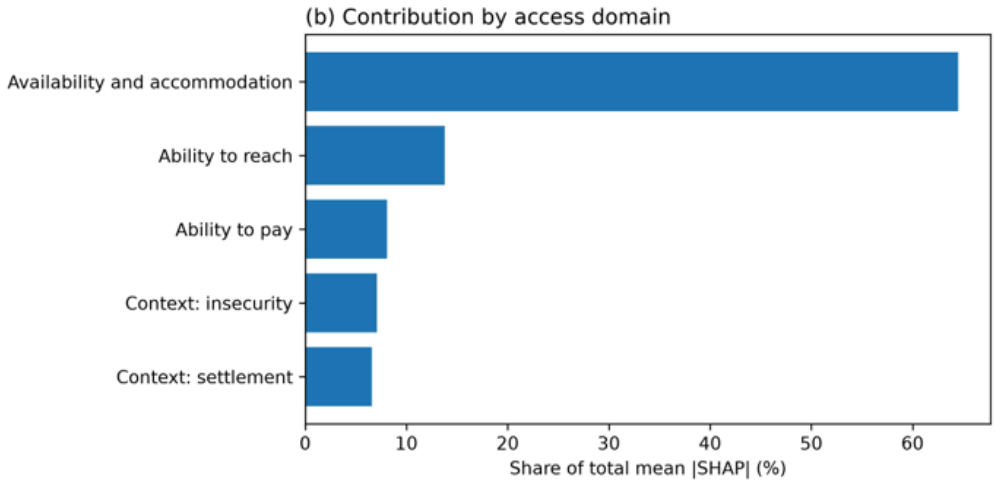

Figure 4 (b)

Figure 4

(a) SHAP summary plot for the 15 most influential features of the final model (probability scale).

(b) Share of total mean absolute SHAP value by access domain.

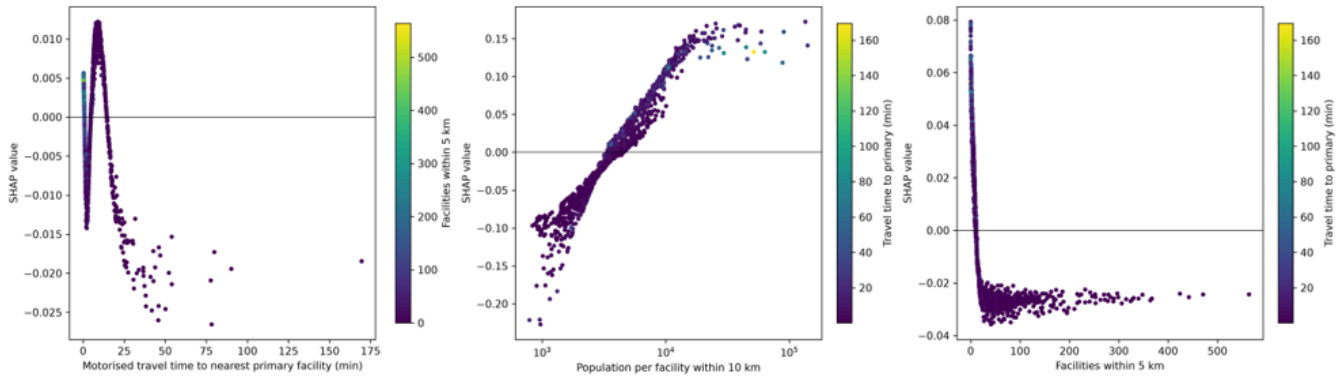

Figure 5

SHAP dependence of predicted risk on motorised travel time to the nearest primary facility, population per facility within 10 km and number of facilities within 5 km; colour shows a second feature.

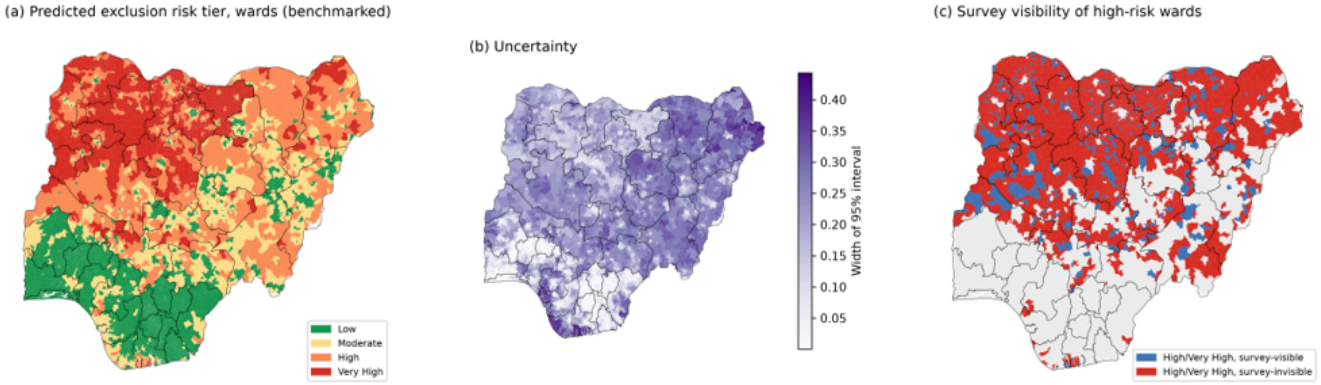

Figure 6

(a) Ward-level predicted risk tier after zone benchmarking; (b) width of the 95% bootstrap prediction interval; (c) High and Very High risk wards with and without an NDHS cluster.

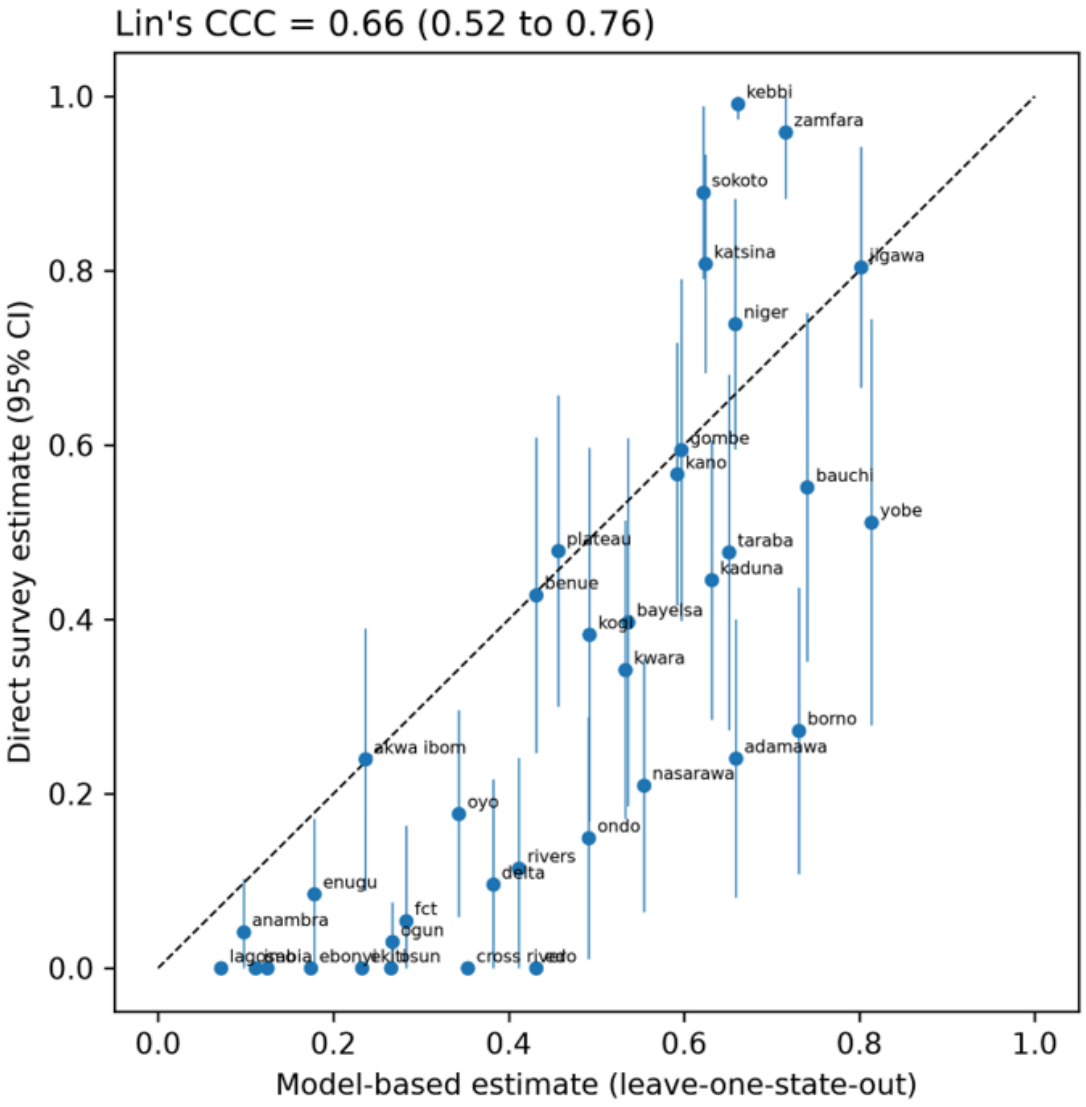

Figure 7

Model-based (leave-one-state-out) versus direct survey estimates of the proportion of the population living in high-exclusion clusters, by state.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- IJHGAdditionalfile1HealthServiceExclusion.docx
- IJHGAdditionalfile2TRIPODAIchecklist1.docx