Received: 3 February 2022 | Revised: 10 June 2022 | Accepted: 9 September 2022

DOI: 10.1002/sim.9586

RESEARCH ARTICLE

Statistics in Medicine WILEY

# Conditional probability and ratio-based approaches for mapping the coverage of multi-dose vaccines

**Chigozie Edson Utazi**[1,2] | **Justice Moses K. Aheto**[1] | **Ho Man Theophilus Chan**[1,2] | **Andrew J. Tatem**[1] | **Sujit K. Sahu**[2]

[1]WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton, UK

[2]School of Mathematical Sciences, University of Southampton, Southampton, UK

**Correspondence**
Chigozie Edson Utazi, WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton, SO17 1BJ, UK.
Email: c.e.utazi@soton.ac.uk

**Funding information**
Bill and Melinda Gates Foundation, Grant/Award Number: INV-003287

Many vaccines are often administered in multiple doses to boost their effectiveness. In the case of childhood vaccines, the coverage maps of the doses and the differences between these often constitute an evidence base to guide investments in improving access to vaccination services and health system performance in low and middle-income countries. A major problem often encountered when mapping the coverage of multi-dose vaccines is the need to ensure that the coverage maps decrease monotonically with successive doses. That is, for doses $i$ and $j$, $i < j \Rightarrow p_i(\boldsymbol{s}) \geq p_j(\boldsymbol{s})$, where $p_i(\boldsymbol{s})$ is the coverage of dose $i$ at spatial location $\boldsymbol{s}$. Here, we explore conditional probability (CP) and ratio-based (RB) approaches for mapping $p_i(\boldsymbol{s})$, embedded within a binomial geostatistical modeling framework, to address this problem. The fully Bayesian model is implemented using the INLA and SPDE approaches. Using a simulation study, we find that both approaches perform comparably for out-of-sample estimation under varying point-level sample size distributions. We apply the methodology to map the coverage of the three doses of diphtheria-tetanus-pertussis vaccine using data from the 2018 Nigeria Demographic and Health Survey. The coverage maps produced using both approaches are almost indistinguishable, although the CP approach yielded more precise estimates on average in this application. We also provide estimates of zero-dose children and the dropout rates between the doses. The methodology is straightforward to implement and can be applied to other vaccines and geographical contexts.

**KEYWORDS**

Bayesian inference, binomial geostatistical model, Demographic and Health Surveys, diphtheria-tetanus-pertussis vaccine, vaccination coverage

## 1 | INTRODUCTION

Many international development goals such as the Sustainable Development Goals (SDGs)[1] and the Immunization Agenda 2030[2] recognize the importance of fine-scale (eg, district level) estimates of health and development indicators (HDIs) for program design, monitoring and evaluation in low- and middle-income countries (LMICs). These estimates help reveal programmatically and epidemiologically important geographic inequities in HDIs, which can often

be masked by aggregate national or provincial estimates traditionally produced by most surveys. Thus, the development of model-based approaches for mapping HDIs has been an active area of research over the last two decades. Maps of HDIs are now routinely produced by the Institute for Health Metrics and Evaluation (IHME),[3-5] the Demographic and Health Surveys (DHS) Program,[6-8] WorldPop[9-13] and other research groups.

Typically, the data used for map production come from geolocated household surveys, such as the DHS surveys. Bayesian geostatistical modeling techniques[14-16] which leverage geospatial covariate information, usually obtained from a variety of sources, and the spatial dependence between survey clusters are often employed to predict HDIs at unsampled locations, typically over a $1 \times 1$ km or $5 \times 5$ km grid covering the area of interest. These high-resolution maps are also a means to produce estimates of HDIs at more operationally relevant spatial scales, for example, districts, at which estimates can be less uncertain and more interpretable than at the grid square scale.

Geospatial analysis of indicators of childhood vaccination coverage has gained traction in the past few years,[5,11,12,17-22] giving rise to the need for alternative approaches for producing maps of multi-dose vaccines as often, the maps produced are for single vaccine doses (eg, the first dose of measles-containing vaccine, MCV1) or the methodology employed when mapping the coverage of multi-dose vaccines simultaneously (eg, the first and third doses of diphtheria-tetanus-pertussis, DTP1 and DTP3) does not consider the relationships between the doses and any challenges these may present.[7,23] Unlike single vaccination coverage indicators, any modeling framework employed in mapping the coverage of multi-dose vaccines should guarantee that the coverage maps decrease monotonically with subsequent doses; that is, for vaccine doses $i$ and $j$, $i < j \Rightarrow p_i(\boldsymbol{s}) \geq p_j(\boldsymbol{s})$, where $p_i(\boldsymbol{s})$ is the coverage of dose $i$ at spatial location $\boldsymbol{s}$. This constraint arises from the fact that it is impossible for the coverage of a subsequent dose of a vaccine to be greater than that of a previous dose. Utazi et al[11] used a multivariate modeling framework to map the coverage of three doses of DTP vaccine in five study countries. This framework provides a mechanism for leveraging the interdependence between the vaccine doses, but it does not necessarily guarantee that the modeled estimates satisfy the monotonicity constraint. Mosser et al[19] employed a continuation ratio ordinal regression approach[24] to enforce this constraint and then applied the approach to model the coverage of DTP1-3 across Africa. Their approach involved modeling DTP3 coverage (as a reference indicator), defined as probability of receipt of at least three doses ($p(d \geq 3)$, where $d$ is the number of doses), and two conditional coverage indicators: probability of receipt of two doses given receipt of at most two doses ($p(d = 2|d \leq 2)$) and probability of receipt of one dose given receipt of at most one dose ($p(d = 1|d \leq 1)$). These modeled quantities were then used to estimate some intermediate indicators (the probabilities of vaccination with $0, 1, 2$ or $\geq 3$ doses), from which estimates of DTP1 and DTP3 coverage and other quantities of interest were obtained. Mosser et al[19] noted that the continuation ratio ordinal regression approach was chosen for their work as it allowed the direct modeling of DTP3 which was considered a key indicator in their analysis. However, this approach seems restrictive as it does not enable direct modeling of DTP1 (even in the reverse case of the continuation ratios[25]) which could be a more suitable reference indicator in some contexts. Another limitation of the approach is that fewer data, both in terms of overall sample size and number of sampled locations (compared to one of the approaches explored here), are available to model the conditional coverage quantities owing to their definitions.

Here, we explore a more flexible alternative methodology for mapping multi-dose vaccines featuring two approaches termed the conditional probability (CP) approach and the ratio-based (RB) approach. While the RB approach utilizes more robust point-level data for all modeled indicators unlike the CP approach, both approaches are flexible in terms of choosing either the first or the last dose in the vaccination series as the reference indicator, which is modeled independently. The methodology is embedded within a Bayesian binomial geostatistical modeling framework implemented using the INLA and SPDE approaches. We investigate the effect of varying distributions of point-level sample sizes on the predictive performance of both approaches using a simulation study. We apply the methodology to mapping the coverage of DTP1-3 in Nigeria using data from the 2018 Nigeria Demographic and Health Survey.[26]

## 2 | METHODOLOGY

### 2.1 | The 2018 Nigeria Demographic and Health Survey (NDHS) vaccination coverage data

Georeferenced cluster-level data on the coverage of each of the three doses of diphtheria-tetanus-pertussis vaccine (DTP1-3) were obtained from the 2018 NDHS.[26] The survey was designed to be representative at the national and state levels, and for urban and rural areas. A stratified, two-stage cluster sampling technique was used which involved the selection of clusters (usually enumeration areas) from a national sampling frame in the first stage and households from

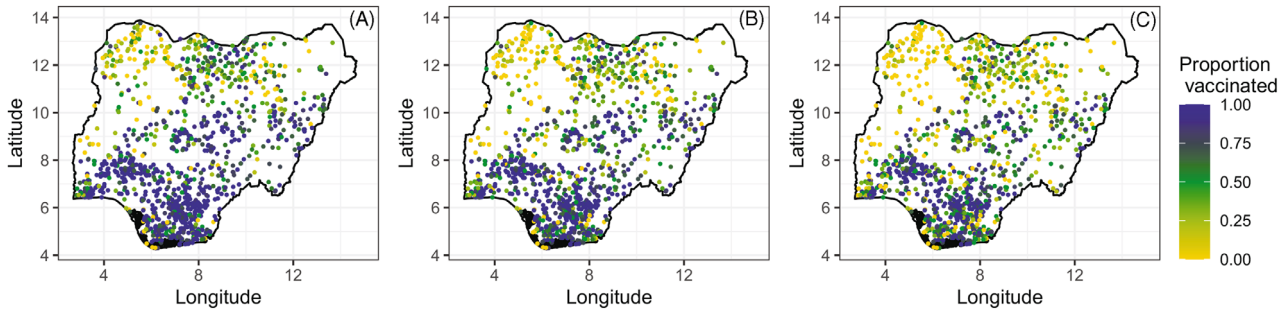


**FIGURE 1** Proportions of children aged 12-23 months who received (A) DTP1, (B) DTP2, and (C) DTP3 vaccinations at the cluster level

within the selected clusters in the second stage. Stratification was achieved by separating the administrative level one areas (ie, the 36 states and the Federal Capital Territory) in the country into urban and rural strata, and samples were drawn independently within each stratum.

For each cluster location, the information extracted were the number of sampled children aged 12-23 months, the numbers who had received each of the vaccine doses as evidenced by their vaccination cards or through caregiver recall, and the displaced geographical (ie, longitude and latitude) coordinates of the cluster. Child records with missing vaccination status (ie, the "don't know cases") were classified as unvaccinated in line with DHS guidelines.[27] In all, the processed data included 6065 children sampled from 1332 clusters, with 3966 (65.4%), 3513 (57.9%), and 3026 (49.9%) reported to have received DTP1-3 vaccinations, respectively. Furthermore, the median number of children sampled per cluster was 4 (IQR: 3-6), and the median numbers of those vaccinated were 3 (IQR: 1-4), 2(IQR: 1-4), and 2 (IQR: 1-3) for each of the respective doses. The empirical cluster-level coverage is displayed in Figure 1, which shows lower coverage levels in the north compared to the southern areas of the country. We also observe that the spatial distribution of the clusters generally aligns with the population distribution in Nigeria, but there are also areas that were under-sampled due to insecurity, for example, the northeastern state of Borno.[26] As in previous work,[12] our analysis did not include clusters where only one child was sampled. We note that cluster-level coverage maps of other modeled indicators discussed in the modeling section are provided in supplementary materials.

## 2.2 | Geospatial and NDHS-derived covariate data, processing and covariate selection

As in previous work,[11,12,18] we assembled some geospatial covariate data known to be either directly linked to coverage or serve as proxies for other unmeasured factors for this study. These include travel time to urban areas, travel time to the nearest health facility (potentially providing routine immunization services), nightlight intensity, and distance to conflict locations as reported in supplementary Table 1. To boost the predictive ability of our models, we obtained additional covariate information from the 2018 NDHS. The geospatial covariates were obtained from various sources and were originally available at different spatial and temporal resolutions. We assembled the most recent data available at the time of analysis, and where applicable, the data were aggregated across multiple years to capture long-term patterns. All geospatial covariate data were processed using ESRI ArcGIS v10.6 to create standardized $1 \times 1$ km gridded covariate layers for our study. Further processing using the geospatial covariates was carried out to extract the corresponding data for each cluster location. Following approaches recommended by Perez-Haydrich et al,[28] we accounted for the displacement of the cluster locations during covariate data extraction by creating 5 km and 2 km buffers around clusters located in rural and urban areas, respectively. We then extracted the mean values of the continuous covariates using all the grid cells falling within the buffer.

The NDHS-derived covariates were first calculated at the cluster level using the definitions provided in supplementary Table 2. We then created $1 \times 1$ km interpolated surfaces of the covariates using kriging interpolation, except the urban-rural covariate. This was implemented using the "krig" function in the *fields* package[29] in R, with the optimal range parameter for each covariate determined using a hold-out cross-validation exercise. Possible range parameters considered were the quartiles of the distances between the clusters in each state. The selected range parameters were mostly the first or the third quartile of the distances. We elected to use kriging interpolation to create the surfaces of these covariates to avoid introducing the problem of circularity (ie, using the same covariates twice) in the analysis. The gridded surface for the urban-rural covariate was created using an approach described in Dong and Wakefield,[20] which utilized gridded

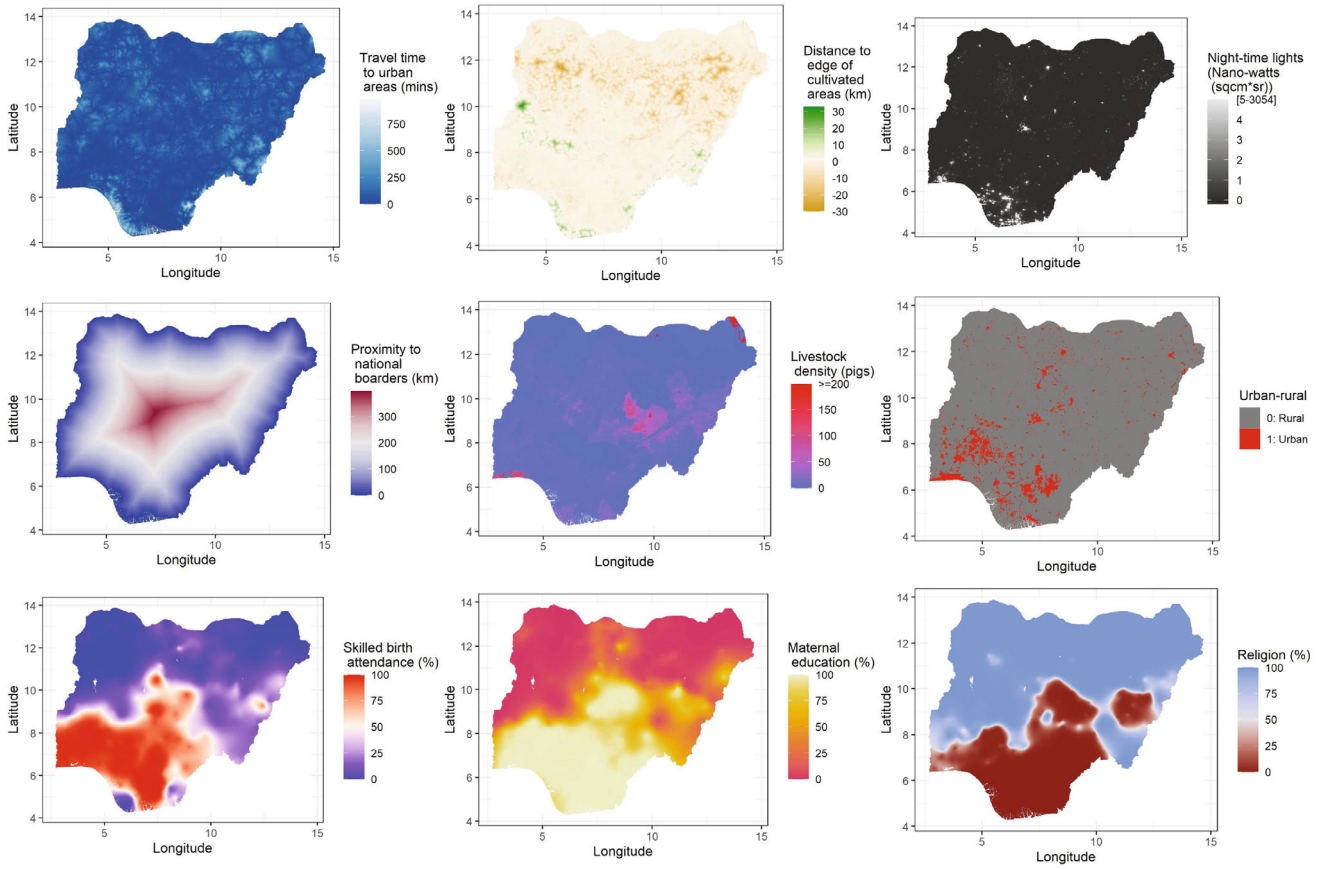


**FIGURE 2** Maps of some geospatial covariates selected for the study

population data from WorldPop[30] and urban population proportion with each administrative level one area within each state obtained from the 2018 NDHS report.[26]

In all, we assembled a total of 23 covariates for the analysis. We selected the best set of covariates for each modeled coverage indicator (see modeling section) in a non-spatial framework, as is standard practice, using the procedure described in Utazi et al.[18] We then created a uniform set of covariates for all modeled indicators for each method investigated here. This resulted in a total of 11 covariates included in the analysis as displayed in Figure 2 and supplementary Figure 3 (see supplementary Tables 1 and 2 for details).

## 2.3 | Population data

To aggregate the grid-level predictions of vaccination coverage to different administrative levels (eg, districts, see modeling section) population data were obtained from WorldPop[30] and processed at $1 \times 1$ km resolution. These were 2018 estimates of numbers of children aged under 5 years, which we used as a proxy for the 12-23 month age group. The data were also used to produce the zero dose estimates, that is, estimates of unvaccinated children, through integration with relevant maps of DTP1 coverage.

## 2.4 | The proposed method

### 2.4.1 | Bayesian binomial geostatistical model

We begin by specifying a geostatistical model for vaccination coverage. For $i = 1, \ldots, m$, let $y(\boldsymbol{s}_i)$ denote the number of children vaccinated at cluster location $\boldsymbol{s}_i$ out of a total of $n(\boldsymbol{s}_i)$ children sampled at the location. We assume that

$Y(\boldsymbol{s}_i)|p(\boldsymbol{s}_i) \sim \text{Binomial}(n(\boldsymbol{s}_i), p(\boldsymbol{s}_i))$, where $p(\boldsymbol{s}_i)$ is the true vaccination coverage (ie, the proportion of children vaccinated) at location $\boldsymbol{s}_i$. Further, $p(\boldsymbol{s}_i)$ is assumed to follow a logistic regression model given by

$$\text{logit}(p(\boldsymbol{s}_i)) = \mathbf{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \omega(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \tag{1}$$

where $\mathbf{x}(\boldsymbol{s}_i)$ is a vector of covariates associated with $\boldsymbol{s}_i$, $\boldsymbol{\beta}$ are the corresponding regression coefficients, $\epsilon(\boldsymbol{s}_i)$ is an independent and identically distributed (iid) Gaussian random effect with variance, $\sigma_\epsilon^2$, used to model non-spatial residual variation, and $\omega(\boldsymbol{s}_i)$ is a Gaussian spatial random effect used to capture residual spatial correlation in the model. That is, $\boldsymbol{\omega} = (\omega(\boldsymbol{s}_1), \ldots, \omega(\boldsymbol{s}_n))' \sim N(0, \Sigma_\omega)$. $\Sigma_\omega$ is assumed to follow the Matérn covariance function[31] given by

$$\Sigma_\omega(\boldsymbol{s}_i, \boldsymbol{s}_j) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_j||)^\nu K_\nu(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_k||),$$

where $||\cdot||$, denotes the Euclidean distance between cluster locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, $\sigma^2 > 0$ is the marginal variance of the spatial process, $\kappa$ is a scaling parameter related to the range $r\left(r = \frac{\sqrt{8\nu}}{\kappa}\right)$-the distance at which spatial correlation is close to 0.1, and $K_\nu$ is the modified Bessel function of the second kind and order $\nu > 0$. Further, for identifiability reasons, we set $\nu = 1$, see Lindgren et al.[32]

As noted previously, applying model (1) to map the coverage of multi-dose vaccines does not guarantee that the monotonic constraint is satisfied for both in- and out-of-sample predictions. We next explore two alternative approaches to tackle this problem.

### 2.4.2 | The conditional probability (CP) approach

This approach relies on conditional probability rules[16,33] to express the interdependencies between the coverage indicators and then exploits these to enforce the monotonic constraint. Throughout, we assume a three-dose vaccination series for simplicity, noting that our approaches can be adapted easily to any number of doses. Let $p_1(\boldsymbol{s}) \geq p_2(\boldsymbol{s}) \geq p_3(\boldsymbol{s})$ denote the respective coverage of the three-dose vaccination series at spatial location $\boldsymbol{s}$. We also refer to these probabilities as the target indicators. All surveyed children at location $\boldsymbol{s}$ can be grouped into four mutually exclusive and completely exhaustive categories, namely zero-dose children, children who received the first dose but not the second dose, children who received the second dose but not the third dose and children who received all three doses. The corresponding probabilities are denoted using $p_{1'}(\boldsymbol{s})$, $p_{1,2'}(\boldsymbol{s})$, $p_{2,3'}(\boldsymbol{s})$, and $p_3(\boldsymbol{s})$, respectively. Thus, $p_{1'}(\boldsymbol{s}) + p_{1,2'}(\boldsymbol{s}) + p_{2,3'}(\boldsymbol{s}) + p_3(\boldsymbol{s}) = 1$. Using conditional probability laws, these probabilities can be further expressed as:

$$\begin{aligned} p_{1,2'}(\boldsymbol{s}) &= p_{2'|1}(\boldsymbol{s}) \times p_1(\boldsymbol{s}), \\ p_{2,3'}(\boldsymbol{s}) &= p_{3'|2}(\boldsymbol{s}) \times p_{2|1}(\boldsymbol{s}) \times p_1(\boldsymbol{s}) = p_{3'|2}(\boldsymbol{s}) \times p_2(\boldsymbol{s}), \\ p_3(\boldsymbol{s}) &= p_{3|2}(\boldsymbol{s}) \times p_{2|1}(\boldsymbol{s}) \times p_1(\boldsymbol{s}) = p_{3|2}(\boldsymbol{s}) \times p_2(\boldsymbol{s}), \end{aligned} \tag{2}$$

where $p_{2'|1}(\boldsymbol{s})$ is the probability of not receiving the second dose given receipt of the first dose, $p_{2|1}(\boldsymbol{s})$ is the probability of receiving the second dose given receipt of the first dose, and so on.

Following the progression from $p_1(\boldsymbol{s})$ to $p_3(\boldsymbol{s})$ in Equation (2), it is apparent that the monotonic condition $p_1(\boldsymbol{s}) \geq p_2(\boldsymbol{s}) \geq p_3(\boldsymbol{s})$ is inherently preserved since $p_1(\boldsymbol{s})$, $p_{2|1}(\boldsymbol{s})$, $p_{3|2}(\boldsymbol{s}) \in [0, 1]$. Hence, it suffices to model these indicators—$p_1(\boldsymbol{s})$, $p_{2|1}(\boldsymbol{s})$, $p_{3|2}(\boldsymbol{s})$—and then enforce the monotonic constraint through using these modeled indicators to derive the remaining target indicators—$p_2(\boldsymbol{s})$ and $p_3(\boldsymbol{s})$. We note that $p_3(\boldsymbol{s})$ could be used in place of $p_1(\boldsymbol{s})$ as the reference indicator if preferable. Also, the indicators: $p_{2|1}(\boldsymbol{s})$ and $p_{3|2}(\boldsymbol{s})$ are different from the conditional probabilities modeled in the continuation ratio ordinal regression approach,[19,24] as these do not include cumulative probabilities.

The corresponding point-level data for the modeled indicators are: $n(\boldsymbol{s})$, $y_1(\boldsymbol{s}); n_1(\boldsymbol{s})$, $y_2(\boldsymbol{s})$ and $n_2(\boldsymbol{s})$, $y_3(\boldsymbol{s})$, for $p_1(\boldsymbol{s})$, $p_{2|1}(\boldsymbol{s})$ and $p_{3|2}(\boldsymbol{s})$, respectively, where $n(\boldsymbol{s})$ is the sample size at location $\boldsymbol{s}$, $y_1(\boldsymbol{s}) = n_1(\boldsymbol{s})$ is the number of surveyed children who were reported to have received at least the first dose, $y_2(\boldsymbol{s}) = n_2(\boldsymbol{s})$ is the number of children who received at least two doses and $y_3(\boldsymbol{s})$ is the number of children who received the third dose. Observe that $n(\boldsymbol{s}) \geq n_1(\boldsymbol{s}) \geq n_2(\boldsymbol{s})$, implying potentially different point-level sample sizes for these indicators. Given that larger sample sizes tend to reduce prediction

error,[12] smaller values of $n_1(\boldsymbol{s})$ and $n_2(\boldsymbol{s})$ could mean that the conditional probabilities—$p_{2|1}(\boldsymbol{s})$ and $p_{3|2}(\boldsymbol{s})$—may not be as well-estimated as $p_1(\boldsymbol{s})$. This is a potential shortcoming of this approach.

### 2.4.3 | The ratio-based (RB) approach

The RB approach aims to address the sample size limitation of the CP approach through modeling the ratios of the target indicators as a strategy to enforce the monotonicity constraint. As with the CP approach, there is the flexibility to determine the modeled indicators from either the beginning ($p_1(\boldsymbol{s})$) or the end of the vaccination series ($p_3(\boldsymbol{s})$) and then construct the modeled indicators as ratios of consecutive doses. In our context, the modeled indicators are:

1. The coverage of the first dose $p_1(\boldsymbol{s})$,
2. The ratio of the coverage of the first and second doses $p_{21}(\boldsymbol{s}) = p_2(\boldsymbol{s}) \times p_1^{-1}(\boldsymbol{s})$, and
3. The ratio of the coverage of the second and third doses $p_{32}(\boldsymbol{s}) = p_3(\boldsymbol{s}) \times p_2^{-1}(\boldsymbol{s})$,

using $p_1(\boldsymbol{s})$ as the reference indicator. From these, $p_2(\boldsymbol{s})$ and $p_3(\boldsymbol{s})$ can be straightforwardly obtained as

$$\begin{aligned} p_2(\boldsymbol{s}) &= p_1(\boldsymbol{s}) \times p_{21}(\boldsymbol{s}) \text{ and} \\ p_3(\boldsymbol{s}) &= p_2(\boldsymbol{s}) \times p_{32}(\boldsymbol{s}), \end{aligned} \tag{3}$$

respectively. Here again, $p_1(\boldsymbol{s})$, $p_{21}(\boldsymbol{s}) \in [0,1]$ and $p_2(\boldsymbol{s}), p_{32}(\boldsymbol{s}) \in [0,1]$ implies that $p_1(\boldsymbol{s}) \geq p_2(\boldsymbol{s}) \geq p_3(\boldsymbol{s})$ is satisfied. To obtain the point-level data for the modeled indicators, let $n(\boldsymbol{s})$ denote the number of children sampled at location $\boldsymbol{s}$ and $y_1(\boldsymbol{s})$ the corresponding number of children who were reported to have received the first dose. Considering that $p_{21}(\boldsymbol{s})$ and $p_{32}(\boldsymbol{s})$ are pseudo indicators, the corresponding pseudo binomial counts can be derived as:

$$\begin{aligned} y_{21}(\boldsymbol{s}) &= n(\boldsymbol{s}) \times p_{21}(\boldsymbol{s}) \text{ and} \\ y_{32}(\boldsymbol{s}) &= n(\boldsymbol{s}) \times p_{32}(\boldsymbol{s}), \end{aligned} \tag{4}$$

respectively. Observe that the point-level sample size $n(\boldsymbol{s})$ is the same for all the modeled indicators. The RB approach is thus unaffected by the potential sample size problem associated with CP approach.

### 2.4.4 | Relationship between the CP and RB approaches

We note that although the CP and RB approaches are different in construction, in our context, the modeled probabilities/indicators are the same under both approaches. Assuming $p_1(\boldsymbol{s})$ to be the reference indicator, it is easy to show that $p_{2|1}(\boldsymbol{s}) = p_{21}(\boldsymbol{s})$ and $p_{3|2}(\boldsymbol{s}) = p_{32}(\boldsymbol{s})$, since the probability of receipt of the first and second doses is the same as the probability of receipt of the second dose, and so on. Thus, as highlighted previously, the differences between both approaches lie in the cluster-level sample sizes associated with the intermediate modeled indicators. Whilst under the RB approach, the cluster-level sample size is $n(\boldsymbol{s})$ for both $p_{21}(\boldsymbol{s})$ and $p_{32}(\boldsymbol{s})$, the sample sizes for $p_{2|1}(\boldsymbol{s})$ and $p_{3|2}(\boldsymbol{s})$ are $n_1(\boldsymbol{s}) = y_1(\boldsymbol{s})$ and $n_2(\boldsymbol{s}) = y_2(\boldsymbol{s})$, respectively, under the CP approach. In the simulation study in Section 3, we investigate the effect of sample sizes on the predictive performance of both approaches.

## 2.5 | Bayesian inference using INLA and SPDE approaches

A fully Bayesian approach was adopted for fitting model (1) for each modeled indicator. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, r, \sigma_\epsilon^2)$ denote the parameters of the model and $\boldsymbol{z}$ all observe data. The joint posterior distribution of the model can be written as:

$$\begin{aligned} \pi(\boldsymbol{\theta}|\boldsymbol{z}) &\propto \prod_{i=1}^{m} \{\text{Binomial}\,(y(\boldsymbol{s}_i); n(\boldsymbol{s}_i), p(\boldsymbol{s}_i), \boldsymbol{\theta}, \boldsymbol{z})\} \times N(\boldsymbol{\omega}; \mathbf{0}, \Sigma_\omega) \times N(\epsilon; \mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}) \times \pi(\boldsymbol{\theta}), \\ &\propto \prod_{i=1}^{m} \left\{p(\boldsymbol{s}_i)^{y(\boldsymbol{s}_i)}(1-p(\boldsymbol{s}_i))^{n(\boldsymbol{s}_i)-y(\boldsymbol{s}_i)}\right\} \times |\Sigma_\omega|^{-\frac{1}{2}}| \ \exp\left(-\frac{1}{2}\boldsymbol{\omega}'\Sigma_\omega^{-1}\boldsymbol{\omega}\right) \times \sigma_\epsilon^{-m} \ \exp\left(-\frac{\sigma_\epsilon^{-2}}{2}\boldsymbol{\epsilon}'\boldsymbol{\epsilon}\right) \times \pi(\boldsymbol{\theta}), \end{aligned} \tag{5}$$

where $\pi(\boldsymbol{\theta})$ is the joint prior distribution on $\boldsymbol{\theta}$. We assigned a $N(0, 10^3\boldsymbol{I})$ prior to the regression parameter, $\boldsymbol{\beta}$. We placed a penalized complexity (PC) prior[34] on $\sigma_\epsilon$ such that $p(\sigma_\epsilon > 3) = 0.01$. Similarly, following Fuglstad et al,[35] a joint PC prior was placed on the covariance parameters of the spatial random effect, $\omega$. These were: $p(r < r_0) = 0.01$ and $p(\sigma > 3) = 0.01$, with $r_0$ chosen to be 5% of the extent of Nigeria in the north-south direction.

The model was implemented using the integrated nested Laplace approximation—stochastic partial differential equation (INLA-SPDE) approach.[32,36] The INLA approach is a faster alternative to the traditional MCMC technique for performing approximate Bayesian inference. The INLA approach produces a numerical approximation of the marginal posterior distributions of each of the unknown quantities in the model. The SPDE approach is particularly required for the estimation of the Gaussian spatial random effect, $\boldsymbol{\omega}$. The approach reduces the computational burden inherent in the estimation of $\Sigma_\omega$ by representing $\boldsymbol{\omega}$ as a Gaussian Markov random field (GMRF)—see Lindgren et al.[32] Further details of the implementation of the INLA-SPDE approach in our work are provided in supplementary materials.

Given the Bayesian context adopted here, the calculation of the modeled estimates of the remaining target indicators from the modeled indicators was implemented using the posterior samples of the modeled indicators and the formulae provided previously.

All analyses were carried out using R[37] and R-INLA package.[38-40]

## 2.6 | Model validation

For each approach, the performance of the fitted models for out-of-sample prediction using the modeled indicators was assessed at the cluster level using a $k$-fold cross-validation scheme, with the folds created as random splits of the $n$ cluster locations. We set $k = 10$ and using the observed ($p(\boldsymbol{s})$) and predicted ($\hat{p}(\boldsymbol{s})$) coverage levels for $m_c$ validation locations, we computed the following model evaluation metrics:

$$\text{Average bias, AvBias} = \frac{1}{m_c}\sum_{i=1}^{m_c}(\hat{p}(\boldsymbol{s}_i) - p(\boldsymbol{s}_i)),$$

$$\text{Root mean square error, RMSE} = \sqrt{\sum_{i}^{m_c}(\hat{p}(\boldsymbol{s}_i) - p(\boldsymbol{s}_i))^2/m_c},$$

and the correlation between observed and predicted values were used to evaluate predictive performance. All three metrics assess the accuracy of the point predictions. The smaller the AvBias (in absolute value) and RMSE, the better the predictions. Conversely, the higher the correlation, the better the predictions. These metrics were calculated and averaged over the cross-validation folds.

Additionally, for the target indicators, the modeled estimates were compared with the direct survey estimates (often considered to be the gold standard[41]) at the state level as in previous work.[12]

## 2.7 | Prediction

For both the CP and RB approaches, predictions using model (1) were first produced at 1 × 1 km resolution for the target indicators (ie, $p_1(\boldsymbol{s}), p_2(\boldsymbol{s})$, and $p_3(\boldsymbol{s})$). Administrative-level predictions using the model were obtained as population-weighted averages taken over all the grid cells falling within each administrative unit. That is, for area $A_i (i = 1, \ldots, m_A$ areas, eg, districts), vaccine dose $k$, and posterior sample $r$,

$$p_k^r(A_i) = \int_{A_i} p_k^r(\boldsymbol{s}) \times q(\boldsymbol{s}) d\boldsymbol{s} \approx \sum_{j=1}^{m_i} p_k^r(\boldsymbol{s_j}) \times q(\boldsymbol{s_j}),$$

where $m_i$ is the number of grid locations with centroids in are $A_i$ and $q(\boldsymbol{s})$ is the proportion of the population of the area at grid location $\boldsymbol{s}$.

To compare the prediction uncertainties associated with the approaches being investigated, we computed the average prediction variance (APV) which is given by

$$\text{APV} = \int_A \text{Var}\{p_k(\boldsymbol{s})|\boldsymbol{z}\}d\boldsymbol{s} \approx \frac{1}{m_p}\sum_{i=1}^{m_p}\text{Var}\{\hat{p}_k^i(\boldsymbol{s})|\boldsymbol{z}\},$$

where $m_p$ is the number of prediction grid cells. Predicted maps with lower APVs are often desirable.

## 3 | SIMULATION STUDY

Here, we describe a simulation study undertaken to investigate the effect of varying point-level sample sizes on the predictive performance of the CP and RB approaches. Using Nigeria as an example geography and the survey clusters from the processed data described in Section 2.1 as the observation locations, data were simulated from model (1) using the following true parameter values: $\sigma^2 = 1,\ r = 2.62,\ \sigma_\epsilon^2 = 1$ and $\boldsymbol{\beta} = (0.5, 0.8, 0.8, 0.2)'$ corresponding to a covariate vector with an intercept term and three variables simulated from $N(0,1)$, Gamma$(1,1)$ and $t(2)$.

We note that the value of $r$ corresponds to the first quartile of the distances between the observation locations. Also, we take the $5 \times 5$ km grid covering the entire country to be the prediction locations for faster computation. Once we have simulated the values of $p_1(\boldsymbol{s})$ for both the observation and prediction locations, we proceeded to simulate those of $p_2(\boldsymbol{s})$ and $p_3(\boldsymbol{s})$ by adding an incremental parameter to the right-hand side of Equation (1) each time to reflect changes in either the regression part of the model or the residual terms. This also ensures that the simulated values of all three indicators satisfy the monotonicity constraint $p_1(\boldsymbol{s}) \geq p_2(\boldsymbol{s}) \geq p_3(\boldsymbol{s})$. Mimicking the patterns in the vaccination coverage data described in Section 2, we set the incremental parameter equal to $-1.3$ for $p_2(\boldsymbol{s})$ and $-2.5$ for $p_3(\boldsymbol{s})$. Next, we assumed the following discrete uniform distributions for the sample sizes at the observation locations: $U\{2,10\}, U\{2,20\}, U\{2,30\}, \ldots, U\{2,80\}$ to reflect varying ranges of sample sizes, although we note that in most DHS surveys, cluster level sample sizes >30 are uncommon. These larger sample sizes are therefore included in the study mainly for illustrative purposes. After obtaining the corresponding counts of successes $y_1(\boldsymbol{s}),\ y_2(\boldsymbol{s})$, and $y_3(\boldsymbol{s})$ through multiplying the sample sizes by the simulated probabilities to preserve the monotonicity constraint, we then proceeded to calculate the additional indicators required for both the CP and RB approaches. This simulation set up resulted in three true $5 \times 5$ km coverage maps for the target indicators: $p_1(\boldsymbol{s}),\ p_2(\boldsymbol{s})$, and $p_3(\boldsymbol{s})$, and a total of 24 data sets, each comprising the same true coverage levels at the observation locations for the target indicators but different sample size distributions. The simulated point and grid level data for the target indicators are displayed in supplementary Figures 4 and 5.

Additionally, we considered a second scenario in which we assumed that the data were not spatially correlated. The data used to investigate the predictive performance of the CP and RB approaches in this case were simulated using the same study design as before, but excluding the spatial random effect, $\boldsymbol{\omega}$, from the model.

For each simulation scenario and modeling approach, we analyzed the simulated data using the Bayesian approaches described in Section 2.5, placing similar prior distributions on all the parameters of the model. We evaluated the predictive performance of both approaches using the metrics described in Section 2.6, all of which were calculated using the true and predicted $5 \times 5$ km maps of $p_1(\boldsymbol{s}),\ p_2(\boldsymbol{s})$, and $p_3(\boldsymbol{s})$ (out-of-sample validation) and the corresponding point-level data for both the modeled indicators and target indicators (in-sample validation) in each case. Additionally, to gain more insights into the predictive performance of the approaches, we calculated the mean absolute error (MAE $= \sum_{i=1}^{m_p}|\hat{p}(\boldsymbol{s}) - p(\boldsymbol{s}_i)|/m_p$) and the actual coverage of the 95% prediction intervals (95% coverage $= 100 \times \sum_{i=1}^{m_p} I(\hat{p}_l(\boldsymbol{s}_i) \leq p(\boldsymbol{s}_i) \leq \hat{p}_u(\boldsymbol{s}_i))/m_p$), where $m_p$ is the number of prediction locations, $p(\boldsymbol{s}_i)$ is the true/simulated coverage at location $\boldsymbol{s}_i$ and $\hat{p}(\boldsymbol{s}_i)$ is the corresponding predicted coverage, $\hat{p}_l(\boldsymbol{s}_i)$ and $\hat{p}_u(\boldsymbol{s}_i)$ are the lower and upper limits of the prediction intervals respectively, and $I(\cdot)$ is an indicator function. The MAE is also used to evaluate the accuracy of the point predictions while the achieved 95% coverage evaluates the accuracy of the uncertainties associated with the predictions. The lower the MAE, the better the prediction. Also, the closer the achieved coverage is to the true value of 95%, the better the predictions.

The results we obtained are displayed in Figure 3 and supplementary Figures 6-9. Figure 3A clearly shows that for in-sample prediction, predictive performance improved as sample sizes increased, and based on nearly all the metrics (except AvBias), the RB approach clearly performed better than the CP approach. This is an indication that the modeled indicators were generally better estimated under the RB approach (see supplementary Figure 6), likely due to the availability of larger sample sizes at the point level for these indicators when using this approach. Also, we observe that for both approaches, the point estimation of $p_1(\boldsymbol{s})$ (this indicator is the same for both approaches in the simulation design,

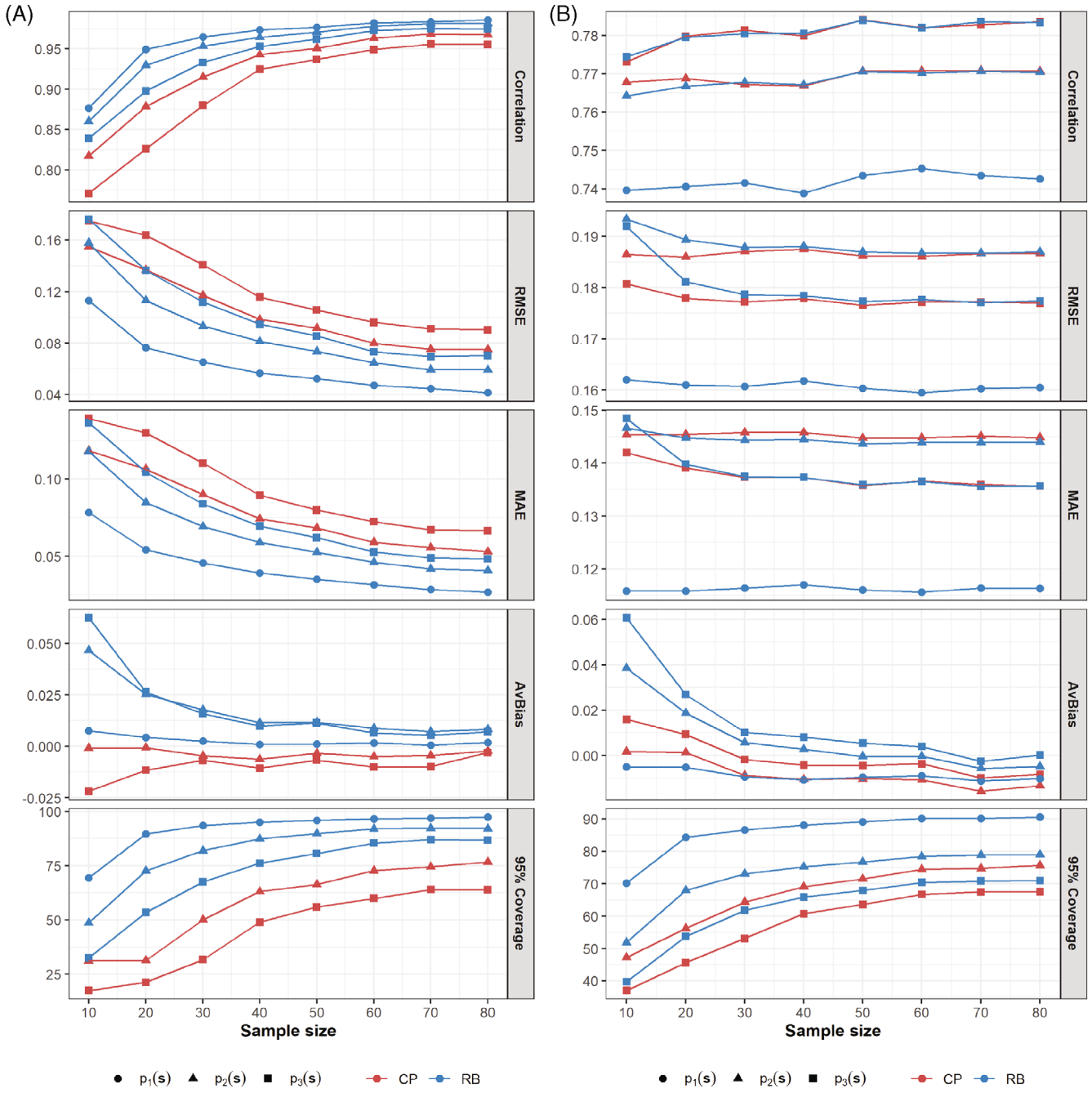


**FIGURE 3** Predictive performance of the conditional probability (CP) and ratio-based (RB) approaches based on different sample size distributions for spatially-correlated point-level data: (A) In-sample prediction of the target indicators; (B) out-of-sample prediction of the target indicators over a 5 × 5 km grid

hence the overlaps in the figures) appears to be consistently better than those of other target indicators. This validates our earlier speculation that the reference indicator—$p_1(\boldsymbol{s})$, which is modeled directly and independently, is likely to be more robustly estimated than other target indicators. Similar patterns were also observed in the in-sample prediction of the modeled indicators under each approach as shown in supplementary Figure 6.

However, for out-of-sample prediction based on the 5 × 5 km grid points, Figure 3B shows that the effect of sample size is negligible when examining the correlation, RMSE and MAE statistics, with both approaches having very similar performances in these instances. Nevertheless, when examining the AvBias and 95% coverage, both approaches exhibit better predictive performance with increasing sample size, particularly for sample sizes ≤50. Also, in terms of 95%

coverage, the RB approach is consistently the better approach for all sample sizes; whereas based on AvBias, the CP approach is better than the RB approach for sample sizes ≤20. The former case is likely an artefact of the larger point level sample sizes for the modeled indicators in the RB approach.

Additionally, owing to the dependence of the sample sizes for $p_{2|1}(\boldsymbol{s})$ and $p_{3|2}(\boldsymbol{s})$ on the values of $p_1(\boldsymbol{s})$ in the CP approach, we investigated the predictive performance of both approaches when $p_1(\boldsymbol{s}) \leq 0.3$, as this is likely to yield very small sample sizes for both conditional probabilities in the CP approach. Interestingly, the results we obtained (see supplementary Figure 7) are very similar to the results reported in Figure 3 for the full range of values of $p_1(\boldsymbol{s})$, except that the AvBias estimates for both approaches are very close in both in-sample and out-of-sample predictions for all sample sizes. Also, we obtained very similar results to those shown in Figure 3 in an additional sensitivity analysis using smaller sample size distributions, that is, discrete uniform $U\{2, 8\}, U\{2, 10\}, U\{2, 12\}, U\{2, 15\}, U\{2, 20\}, \ldots, U\{2, 35\}$ (see supplementary Figure 8).

In general, these results reveal that the effect of sample size on both approaches is more pronounced in in-sample prediction, in which case the RB approach outperformed the CP approach. For out-of-sample prediction, no approach is uniformly the better approach, and the effect of sample size appears to matter for bias and uncertainty estimation only.

With spatially uncorrelated data, the patterns in the results (see, eg, supplementary Figure 9) are similar to those shown in Figure 3, hence we do not discuss these further.

## 4 | RESULTS OF ANALYSIS OF THE 2018 NIGERIA DEMOGRAPHIC AND HEALTH SURVEY (NDHS) VACCINATION COVERAGE DATA

Here, we present the results of application of the proposed approaches to mapping DTP1-3 vaccination coverage using the 2018 NDHS data, including the dropout rates between the doses and estimates of zero-dose children.

With the CP approach, the covariates chosen for model-fitting and prediction were: maternal education, skilled birth attendance, livestock (pigs) density, proximity to national borders, urbanicity (ie, urban/rural), religion, night-time lights, and household wealth. For the RB approach, the first five covariates were also selected in addition to travel time to urban areas, distance to the edge of cultivated areas and distance to conflict areas.

For both approaches, estimates of parameters of the fitted models are reported in supplementary Tables 3 and 4 for the modeled indicators. For the CP approach, maternal education, religion, and skilled birth attendance were significant predictors of DTP1 coverage (ie, $p_1(\boldsymbol{s})$). Maternal education was also a significant predictor of $p_{2|1}(\boldsymbol{s})$, while skilled birth attendance was a significant predictor of $p_{3|2}(\boldsymbol{s})$. For the RB approach, similar patterns were also observed in the relationships between the modeled indicators and the covariates. Maternal education, and skilled birth attendance were significant predictors of $p_1(\boldsymbol{s})$. Maternal education was also the only significant predictor of $p_{21}(\boldsymbol{s})$, while skilled birth attendance and livestock density (pigs) were significant predictors of $p_{32}(\boldsymbol{s})$. Further, education and skilled birth attendance had positive relationships with coverage while religion had a negative relationship with coverage in all cases as expected. Interestingly, livestock (pigs) density also had a positive relationship with coverage. We note that the regression coefficients associated with these significant covariates can be exponentiated to quantify the effect of a unit increase in these covariates on the odds of vaccination. However, this is not of interest here for various reasons including our focus on prediction and potential aggregation bias that could occur with some of the covariates that can be measured at the individual child level. For the CP approach, the estimated spatial ranges were between 115 and 239 km, whereas for the RB approach, these were between 111 and 224 km for the modeled indicators in each case (see supplementary Tables 3 and 4).

Cluster-level out-of-sample model validation results are presented in Table 1. These results show the predictive performance of the fitted models for equivalent modeled indicators under both approaches. Very similar results were obtained for $p_1(\boldsymbol{s})$ with both approaches, even though different sets of covariates were used in the analysis. For the last two indicators, the RB approach had consistently lower AvBias while the CP approach had consistently lower RMSE values. Mixed results were obtained when considering the correlation statistics. Thus, no approach produced consistently better results for these equivalent modeled indicators.

When considering the uncertainties in the modeled $1 \times 1$ km estimates of the target indicators, the APV values show that that the CP approach outperformed the RB approach. Also, the directly modeled indicator, $p_1(\boldsymbol{s})$, had the lowest APV of all the three target indicators in each case, which is a further indication that this indicator was more robustly estimated, the evidence of which is stronger under the RB approach. Subsequent results presented in this work are therefore based on the CP approach, with comparisons with the RB approach included where necessary.

**TABLE 1** Model validation statistics based on a $k$-fold cross-validation exercise and average prediction variance estimates

| Modeled indicators | AvBias | RMSE | Correlation | Target indicators | APV |
|---|---|---|---|---|---|
| Conditional probability approach | | | | | |
| $p_1(\boldsymbol{s})$ | −0.001 | 0.220 | 0.740 | $p_1(\boldsymbol{s})$ | 0.020 |
| $p_{2\|1}(\boldsymbol{s})$ | 0.005 | 0.196 | 0.395 | $p_2(\boldsymbol{s})$ | 0.021 |
| $p_{3\|2}(\boldsymbol{s})$ | 0.005 | 0.219 | 0.322 | $p_3(\boldsymbol{s})$ | 0.023 |
| Ratio-based approach | | | | | |
| $p_1(\boldsymbol{s})$ | −0.001 | 0.220 | 0.739 | $p_1(\boldsymbol{s})$ | 0.022 |
| $p_{21}(\boldsymbol{s})$ | 0.002 | 0.218 | 0.357 | $p_2(\boldsymbol{s})$ | 0.034 |
| $p_{32}(\boldsymbol{s})$ | 0.001 | 0.258 | 0.356 | $p_3(\boldsymbol{s})$ | 0.051 |

Lastly, in supplementary Figure 10, we further validate the estimates of the target indicators using the direct survey estimates at the state level. These plots indicate that there is a strong correspondence (correlation $\geq 0.94$ in each case) between the direct and modeled estimates both when considering the CP and RB approaches. The plots also show that the uncertainties associated with the RB approach were generally wider than those of the CP approach, further corroborating the results presented in Table 1.

## 4.1 | DTP1-3 coverage maps

We first compare $1 \times 1$ km predicted maps of DTP2 and DTP3 coverage produced using the CP and RB approaches in Figure 4. Panels (A) and (B) show slight differences between the predicted maps when covariates were included in the fitted models. These differences became narrower when covariates were excluded from the analysis as shown in panels (C) and (D). Hence, these maps demonstrate that both approaches produced very similar grid level predictions despite being different in construction and implementation.

In Figure 5 we present the coverage maps of all three doses produced using the CP approach. There are substantial heterogeneities in the coverage of each dose, with coverage levels markedly higher in the south compared to the north, particularly the northeastern and northwestern areas. Coverage can also be seen to generally decrease when progressing from DTP1 to DTP3, as expected. The "smooth" predicted maps are most likely an artefact of the kriged DHS covariates (see, eg, Figure 2) which were mostly significant predictors of coverage in the fitted models. The uncertainties associated with these estimates, presented as standard deviations, show that for DTP1, the southern areas where higher coverage levels were estimated had lower uncertainty compared to the north. Some lower coverage areas in the northwest were also predicted with lower uncertainty. Similar patterns are apparent in the coverage maps of DTP2 and DTP3, although there are more areas of higher uncertainty. Generally, areas with lower density of cluster locations (see Figure 1) tend to have higher uncertainty. Also, the patterns in these uncertainty estimates are likely due to the binomial likelihood used in the model—estimates close to the endpoints of the unit interval tend to have higher precision than estimates lying close to the middle of the interval.[18] At the district or local government area (LGA) level (see supplementary Figure 11), significant inequalities in coverage still exist and patterns in coverage are generally similar to those shown in Figure 5.

## 4.2 | Dropout rates and zero-dose estimates

Maps of relative dropout rates between the doses (calculated as $100 \times (\hat{p}_i(\boldsymbol{s}) - \hat{p}_j(\boldsymbol{s}))/\hat{p}_i(\boldsymbol{s}); i < j$) are shown in Figure 6. Clearly, the dropout rates are generally higher between DTP2 and DTP3 (DTP2-3) than between DTP1 and 2 (DTP1-2). Also, lower dropout rates were estimated in areas with higher coverage, while higher dropout rates were estimates in areas with lower coverage. These patterns are more evident when examining the dropout rates between DTP1 and DTP3. This suggests that factors responsible for high dropouts may also influence the likelihood of receipt of DTP1. There are also visible spots of areas of lower dropout rates in urban areas.

Estimates of numbers of children aged under 5 years who had not received any DTP doses, that is, zero-dose children, are displayed at both the district and state levels in Figure 7 (see supplementary Table 5 for details of the district-level

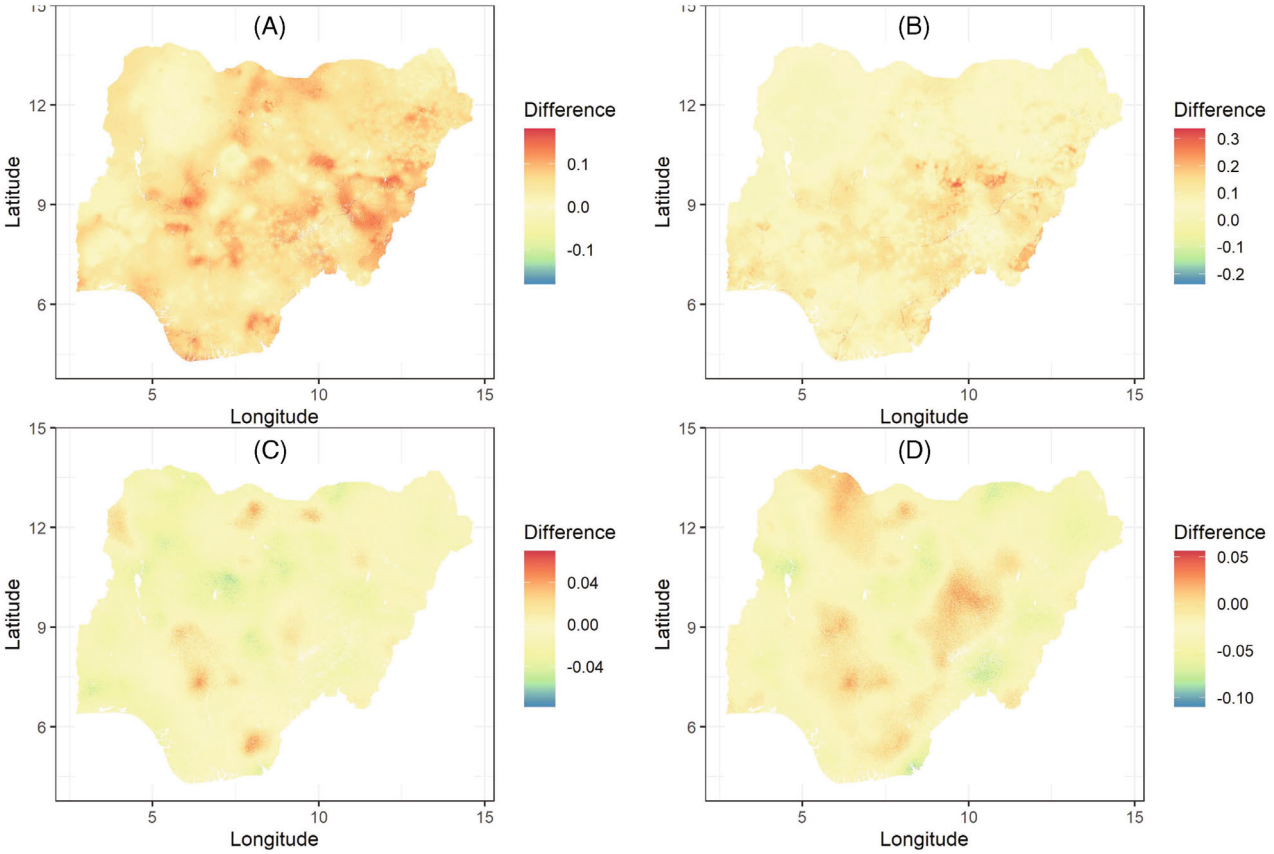


**FIGURE 4** Differences between 1 × 1 km predicted maps of DTP2 (A, C) and DTP3 (B, D) obtained through using the conditional probability and ratio-based approaches when covariates were included (A, B) and excluded (C, D) from the fitted models

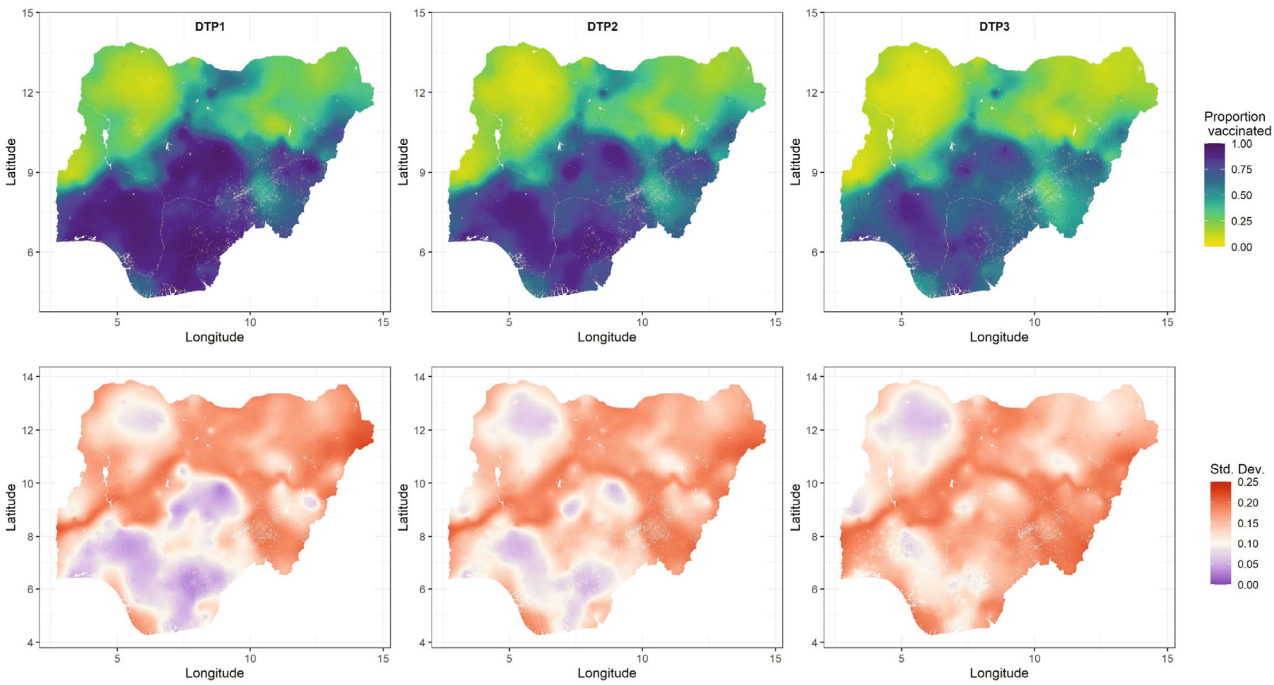


**FIGURE 5** Predicted 1 × 1 km maps of DTP1-3 coverage and associated uncertainties shown as standard deviations

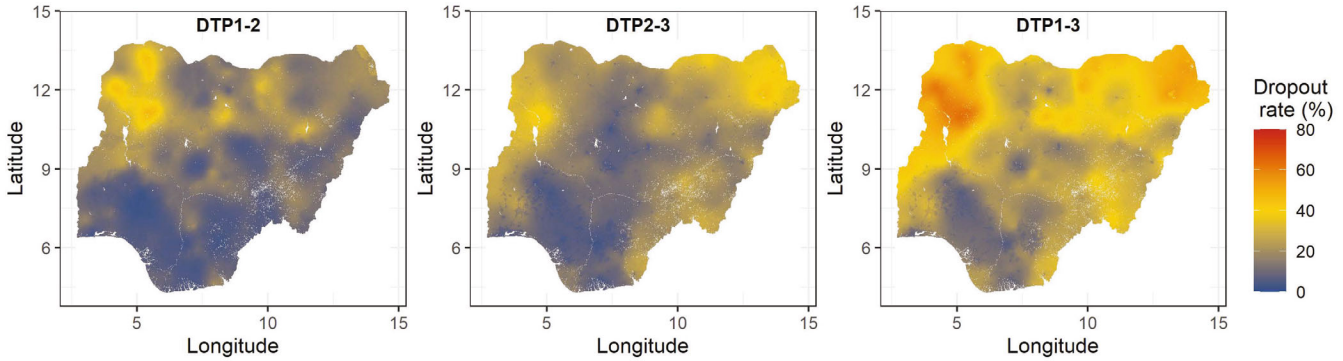


**FIGURE 6** Dropout rates between the doses at 1 × 1 km resolution

estimates). These zero-dose estimates were produced using relevant administrative level coverage estimates and associated uncertainties, and population estimates which were assumed to be fixed. An alternative approach for producing the zero-dose estimates is to use relevant grid level coverage and population estimates and then aggregate the resulting zero-dose estimates to administrative levels of interest. While both approaches can produce very similar zero-dose (point) estimates, the second approach could sometimes yield unreasonably wide uncertainty intervals. Districts with the most unvaccinated children are located in Zamfara (Bungudu, Zurmi, Gusau, Kaura Namoda, Maradun, Maru), Gombe (Yamaltu/Deba), Bauchi (Darazo, Bauchi, Ningi), Kebbi (Wasagu/Danko), Yobe (Fune), Sokoto (Dange-Shuni), Borno (Jere), and Jigawa (Birnin Kudu) states, all of which had at least 50 000 zero-dose children. The uncertainties associated with these estimates (Figure 7B) are generally low (95% CI width < 20 000), apart from some districts in the northeastern and northwestern areas where higher uncertainties were estimated. The patterns in the uncertainty estimates generally reflect the patterns in the uncertainties in the underlying district-level estimates presented in supplementary Figure 11, as expected.

At the state level, Kano, Katsina, Sokoto, Zamfara, Bauchi, Borno, Kebbi, Jigawa, and Niger states had at least 400 000 zero-dose children. These are mostly states where districts with higher estimates of numbers of zero-dose children were located (panel A), and where the poorest coverage levels were estimated. The uncertainties associated with the zero-dose estimates appear to increase as the estimates increase, and these generally show that the numbers of zero-dose children were reasonably well estimated. Furthermore, states where there is an intersection of lower to moderate coverage and higher zero-dose estimate, as shown in panel (D) should be considered as priority areas for improvements in routine immunization coverage.

## 5 | DISCUSSION

In this article, we have explored alternative approaches for mapping the coverage of multi-dose vaccines at fine spatial scales in low- and middle-income settings. Both approaches examined are flexible in terms of using either the first or the last dose in the vaccination series as the reference indicator, which can often be an important consideration. Furthermore, one of the approaches—the RB approach—is not subject to potential sample size restrictions that can be encountered when modeling conditional probabilities which are used to induce the monotonicity constraint in some approaches (eg, the CP approach explored here and the continuation ratio ordinal regression approach used in Mosser et al[19]). We illustrated this using a simulation study in which we found out that the RB approach consistently performed better than the CP approach for in-sample prediction under varying point-level sample size distributions. We also noted that increasing point-level sample sizes had marked positive impact on in-sample prediction using both approaches. However, for out-of-sample prediction, no approach was consistently the better approach. Also, in this case, increases in point-level sample sizes mainly led to improvements in bias and uncertainty estimation. We, however, note that although increasing the point-level sample sizes is desirable in a geostatistical context, in practice, this may need to be balanced against design-based large-area survey analysis considerations, where larger cluster-level sample sizes can be statistically inefficient.[12]

We applied the methodology to map the coverage of DTP1-3 in Nigeria using data from the 2018 NDHS. We modeled DTP1 as the reference indicator due to our interest in producing estimates of zero-dose children. We demonstrated

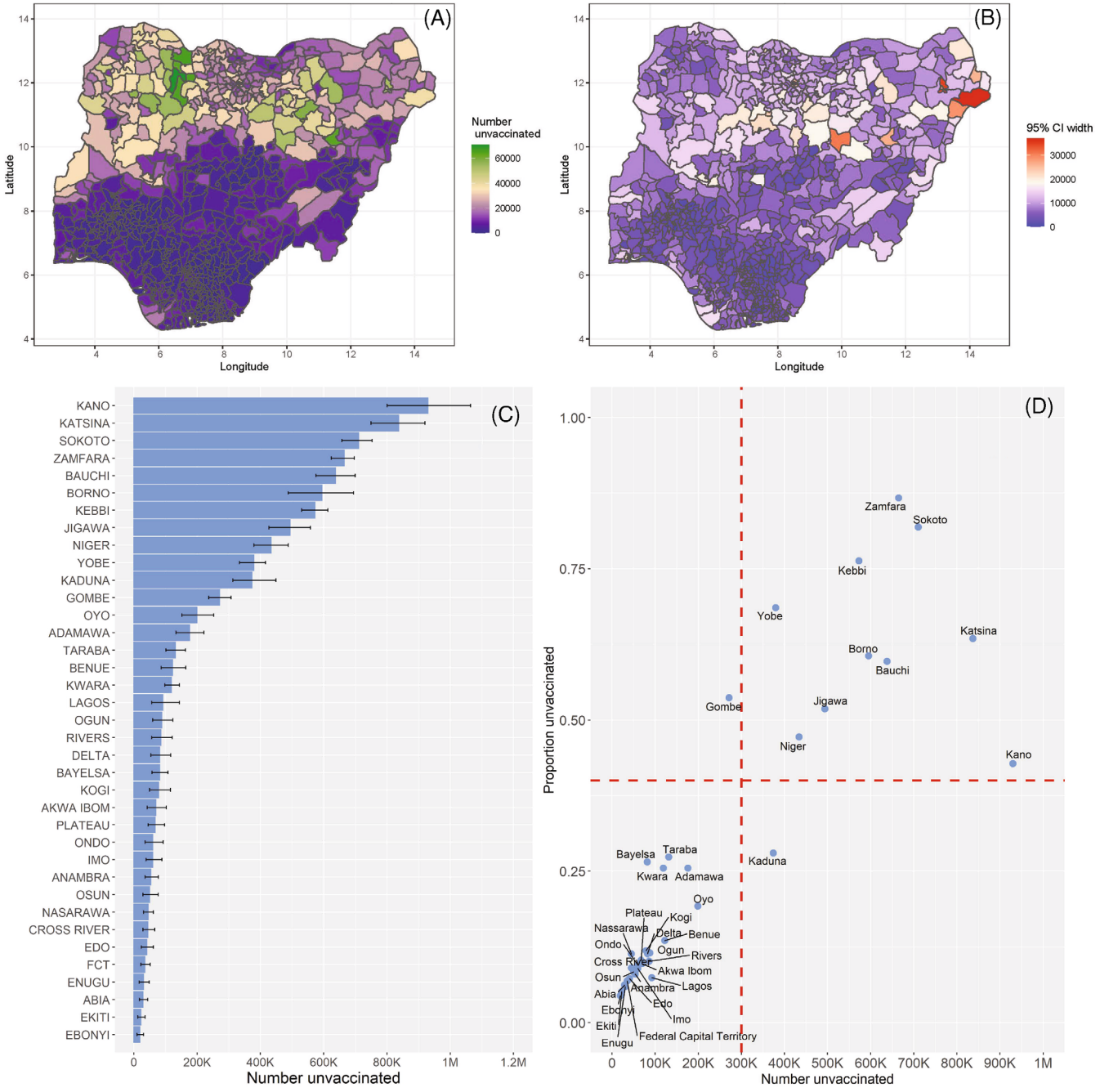


**FIGURE 7** Estimates of numbers of zero-dose children aged under 5 years and associated uncertainties at the district (A, B) and state levels (C). The relationship between proportions of unvaccinated children and corresponding zero-dose estimates at the state level is shown in panel (D). The red dotted lines are used to show different prioritization scenarios

that both approaches yielded very similar results for this application. Our maps of DTP1 and DTP3 coverage—both of which are often used to evaluate access to routine immunization (RI) services and the general performance of RI programs—produced some interesting patterns.[11] These maps revealed substantial heterogeneities in coverage as well as a characteristic north-south divide.[12,20] The northeast, the northwest and parts of the north central zones of the country are the problematic areas where efforts should be targeted to fill coverage and immunity gaps. In addition, the patterns in the dropout rates suggest that areas with lower coverage were more likely to have higher dropout rates, as also noted in a previous study.[11] This is an indication that factors responsible for non-vaccination in these areas are likely responsible for the failure to complete the vaccination series. Aheto et al[42] found these factors to include non-ownership of a health card/document, non-receipt of vitamin A (both are indicators of access to health/vaccination services), poor maternal

education, religion and maternal age (being born to a younger mother), some of which were included as covariates in this study. Thus, any strategies geared towards improving RI coverage in the country should aim to address the inequities emanating from these factors. The process of prioritizing subnational areas for RI improvements often involves an assessment of estimates of DTP vaccine coverage and corresponding numbers of DTP zero-dose children, local measles epidemiology and other concomitant factors such as insecurity and prevalence of other childhood diseases. The coverage maps and zero-dose estimates presented here can serve as a useful input into this process to guide the allocation of resources at the national and subnational levels, as well as being credible alternatives to administrative coverage estimates whose utility is often limited by numerator and denominator issues.[43]

Our work is subject to some limitations. With both approaches, some of the modeled indicators were not as robustly estimated as the reference indicator in our application. This may have been an effect of the little variation in these indicators (supplementary Figures 1 and 2), which also meant that they were more difficult to predict. The data we analyzed included information on vaccination coverage obtained from vaccination cards and through caregiver recall. Although, this increases the data available for modeling, it has the potential to introduce recall bias in the analysis. Grid-level and aggregated predictions of vaccination coverage (including comparisons with direct survey estimates) and corresponding zero-dose estimates can be influenced by the covariates included in the analyses—see, for example, Giorgi et al.[44] Our analyses included both geospatial and NDHS-derived covariates, but our results showed that the latter appeared to have suppressed the geospatial covariates. We were unable to investigate the effect of this outcome and the contributions of both sets of covariates (both separately and combined) to the predictions, but we plan to undertake this elsewhere. Furthermore, we were unable to account for the uncertainties associated with the population estimates used in producing the zero-dose estimates. Accounting for the uncertainties in both the population and coverage estimates simultaneously when producing zero-dose estimates will be better implemented in a joint modeling framework, which will constitute part of future work. In our application, we chose the CP approach because it yielded smaller APV values for the target indicators. While this choice is plausible in a geospatial analysis context in which estimates with less uncertainties are desirable, we note that the APV metric does not evaluate the accuracies of the uncertainties estimated by both approaches. Also, estimates produced in under-sampled areas, for example, conflict areas in Borno state, could be biased if the relationships between the covariates and vaccination coverage in those areas were different from those of other areas where data were collected. Lastly, our methodology and application focused on a snapshot in time. Additional insights can be gained from analyzing trends in coverage over time. In future work, we will consider an extension of the methodology to the spatiotemporal setting.

In conclusion, we consider this work a useful addition to the growing body of methodology for producing maps of vaccination coverage. It is straightforward to apply the methodology to map the coverage of other multi-dose vaccines, for example, the pneumococcal conjugate vaccine and rotavirus vaccine, even as efforts within the global health community are continually targeted towards improving vaccination services, improving access to new vaccines and accelerating progress towards disease elimination.

## ACKNOWLEDGEMENTS

We are grateful to the DHS program for providing the data for this study. We also thank the Bill and Melinda Gates Foundation for funding this study.

## FUNDING INFORMATION

This work was supported by funding from the Bill and Melinda Gates Foundation (Investment ID INV-003287). Chigozie Edson Utazi and Andrew J. Tatem received the Grant. The funder did not play any role in the study design, data collection, analysis and interpretation of data, the report writing, and the decision to submit the manuscript for publication.

## CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

## DATA AVAILABILITY STATEMENT

DHS data supporting this study are publicly available from https://dhsprogram.com/data/available-datasets.cfm. Other data are publicly available via the sources referenced in the methods section. These can also be obtained from the authors upon request. R scripts used for the analyses are available on GitHub (https://github.com/EdsonUtazi/Code-for-multi-dose-vax-paper-v2)

## ORCID

*Chigozie Edson Utazi* https://orcid.org/0000-0002-0534-5310
*Justice Moses K. Aheto* https://orcid.org/0000-0003-1384-2461
*Ho Man Theophilus Chan* https://orcid.org/0000-0001-6821-4206

## REFERENCES

1. United Nation General Assembly. Transforming our world: the 2030 agenda for sustainable development A/RES/70/1; 2015. Resolution adopted by the general assembly on September 25, 2015. http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1. Accessed May 23, 2021.
2. WHO. Immunization agenda 2030: a global strategy to leave no one behind; 2020. Resolution adopted by the general assembly on September 25, 2015. Geneva, Switzerland: World Health Organization. https://www.who.int/immunization/immunization_agenda_2030/en/. Accessed June 25, 2021.
3. Institue for Health Metrics and Evaluation. IHME: measuring what matters; 2020; Institute for Health Metrics and Evaluation. http://www.healthdata.org/. Accessed April 25, 2020.
4. Local Burden of Disease Child Growth Failure Collaborators. Mapping child growth failure across low-and middle-income countries. *Nature*. 2020;577(7789):231.
5. Local Burden of Disease Vaccine Coverage Collaborators. Mapping routine measles vaccination in low-and middle-income countries. *Nature*. 2021;589(7842):415.
6. Gething P, Tatem A, Bird T, Burgert C. Creating spatial interpolation surfaces with DHS data. Project report, University of Southampton; University of Southampton, Southampton SO17 1BJ; 2015.
7. Mayala BK, Dontamsetti T, Fish TD, Croft T. Interpolation of DHS survey data at subnational administrative Level 2. DHS Program, ICF; 2019.
8. ICF. The Demographic and Health Surveys, The DHS Program website; 2021. Funded by USAID. https://dhsprogram.com/. Accessed December 20, 2020.
9. Alegana VA, Atkinson PM, Pezzulo C, et al. Fine resolution mapping of population age-structures for health and development applications. *J Royal Soc Interf*. 2015;12(105):20150073.
10. Bosco C, Alegana V, Bird T, et al. Exploring the high-resolution mapping of gender-disaggregated development indicators. *J Royal Soc Interf*. 2017;14(129):20160825.
11. Utazi CE, Thorley J, Alegana VA, et al. Mapping vaccination coverage to explore the effects of delivery mechanisms and inform vaccination strategies. *Nature Commun*. 2019;10(1):1-10.
12. Utazi CE, Wagai J, Pannell O, et al. Geospatial variation in measles vaccine coverage through routine and campaign strategies in Nigeria: analysis of recent household surveys. *Vaccine*. 2020;38(14):3062-3071.
13. WorldPop. The WorldPop project; 2021; WorldPop. https://www.worldpop.org/. Accessed May 15, 2021.
14. Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. *J Royal Stat Soc Ser C (Appl Stat)*. 1998;47(3):299-350.
15. Banerjee S, Carlin BP, Gelfand AE. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: CRC Press; 2015.
16. Sahu SK. *Bayesian Modelling of Spatio-Temporal Data with R*. 1st ed. Boca Raton: Chapman & Hall/CRC Press; 2022.
17. Takahashi S, Metcalf CJE, Ferrari MJ, Tatem AJ, Lessler J. The geography of measles vaccination in the African Great Lakes region. *Nature Commun*. 2017;8(1):1-9.
18. 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):1583-1591.
19. Mosser JF, Gagne-Maynard W, Rao PC, et al. Mapping diphtheria-pertussis-tetanus vaccine coverage in Africa, 2000–2016: a spatial and temporal modelling study. *Lancet*. 2019;393(10183):1843-1855.
20. Dong TQ, Wakefield J. Modeling and presentation of vaccination coverage estimates using data from household surveys. *Vaccine*. 2021;39(18):2584-2594.
21. Fuglstad GA, Li ZR, Wakefield J. The two cultures for prevalence mapping: small area estimation and spatial statistics. arXiv preprint arXiv:2110.09576, 2021.
22. Utazi CE, Nilsen K, Pannell O, Dotse-Gborgbortsi W, Tatem AJ. District-level estimation of vaccination coverage: discrete vs continuous spatial models. *Stat Med*. 2021;40(9):2197-2211.
23. ICF. Spatial data repository, the demographic and health surveys program. Modeled surfaces; 2020; Funded by the United States Agency for International Development (USAID). https://spatialdata.dhsprogram.com/modeled-surfaces/. Accessed December 20, 2020.
24. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Ordinal logistic regression models. *Applied Logistic Regression*. Vol 398. Hoboken: John Wiley & Sons; 2013.
25. Xing L, Haiyan B. Forward and backward continuation ratio models for ordinal response variables. *J Modern Appl Stat Methods*. 2019;18(2):eP3043. doi:10.22237/jmasm/1604190180
26. National population commission and ICF. Nigeria demographic and health survey 2018 - Final report; 2019; Abuja, Nigeria, and Rockville, Maryland, National Population Commission and ICF. https://dhsprogram.com/publications/publication-fr359-dhs-final-reports.cfm. Accessed March 16, 2021.
27. Croft TN, Marshall AM, Allen CK. Guide to DHS statistics; 2018; Rockville, Maryland, ICF. https://www.dhsprogram.com/pubs/pdf/DHSG1/Guide_to_DHS_Statistics_DHS-7.pdf. Accessed May 27, 2022.

28. Perez-Heydrich C, Warren JL, Emch ME. Guidelines on the use of DHS GPS data; 2013; Spatial Analysis Reports No. 8. Calverton, Maryland, ICF International. https://dhsprogram.com/publications/publication-SAR8-Spatial-Analysis-Reports.cfm. Accessed April 18, 2021.
29. Nychka D, Furrer R, Paige J, Sain S. Fields: tools for spatial data. R package version 11.6; 2017. https://github.com/NCAR/Fields
30. Tatem AJ. WorldPop, open data for spatial demography. *Sci Data*. 2017;4(1):1-4.
31. Matérn B. *Spatial variation*. 2nd ed. Berlin: Springer-Verlag; 1960 Reprinted 1986.
32. Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J Royal Stat Soc Ser B (Stat Methodol)*. 2011;73(4):423-498.
33. Bandyopadhyay PS, Forster MR. Philosophy of Statistics: An Introduction. In: Bandyopadhyay PS, Forster MR, eds. *Philosophy of statistics*. Elsevier: North Holland; 2011:1-50.
34. Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH. Penalising model component complexity: a principled, practical approach to constructing priors. *Stat Sci*. 2017;32(1):1-28.
35. Fuglstad GA, Simpson D, Lindgren F, Rue H. Constructing priors that penalize the complexity of Gaussian random fields. *J Am Stat Assoc*. 2019;114(525):445-452.
36. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J Royal Stat Soc Ser B (Stat Methodol)*. 2009;71(2):319-392.
37. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2021.
38. Martino S, Rue H. R package: INLA; 2009; Department of Mathematical Sciences; NTNU, Norway.
39. Bivand R, Gómez-Rubio V, Rue H. Spatial data analysis with R-INLA with some extensions; 2015; American Statistical Association.
40. Lindgren F, Rue H. Bayesian spatial modelling with R-INLA. *J Stat Softw*. 2015;63(1):1-25.
41. Paige J, Fuglstad GA, Riebler A, Wakefield J. Design-and model-based approaches to small-area estimation in a low and middle income country context: comparisons and recommendations. arXiv preprint arXiv:1910.06512, 2019.
42. Aheto JMK, Pannell O, Dotse-Gborgbortsi W, et al. Multilevel analysis of predictors of multiple indicators of childhood vaccination in Nigeria. *PLoS One*. 2022;17(5):e0269066. doi:10.1371/journal.pone.0269066
43. Cutts FT, Claquin P, Danovaro-Holliday MC, Rhoda DA. Monitoring vaccination coverage: defining the role of surveys. *Vaccine*. 2016;34(35):4103-4109.
44. Giorgi E, Fronterrè C, Macharia PM, Alegana VA, Snow RW, Diggle PJ. Model building and assessment of the impact of covariates for disease prevalence mapping in low-resource settings: to explain and to predict. *J Royal Soc Interf*. 2021;18(179):20210104.


## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Utazi CE, Aheto JMK, Chan HMT, Tatem AJ, Sahu SK. Conditional probability and ratio-based approaches for mapping the coverage of multi-dose vaccines. *Statistics in Medicine*. 2022;1-17. doi: 10.1002/sim.9586