Correcting Survey Bias in Childhood Immunization Coverage Estimates Using Bayesian Post-Stratification: A Simulation-Based Study

1. Background of the Study

Reliable measurement of childhood immunization coverage is central to public health planning. Immunization surveys inform national strategies, track global progress toward Sustainable Development Goals, and guide resource allocation. Yet, survey-based estimates are vulnerable to nonresponse bias, coverage error, and measurement error-issues particularly pronounced in low- and middle-income countries where logistical constraints and uneven access affect data quality.

In Nigeria, for instance, the Demographic and Health Survey (NDHS) and Multiple Indicator Cluster Survey (MICS) are the main sources of vaccination statistics, but both face substantial underrepresentation of rural and less-educated households. These biases can produce misleading estimates of national immunization coverage and obscure inequities across regions or social groups.

Traditional weighting or post-stratification methods are often used to correct survey imbalances. However, these classical approaches depend on sufficient representation within each subgroup (e.g., age, education, or region) and can perform poorly when data are sparse. As survey nonresponse grows and populations become more heterogeneous, these techniques may no longer produce reliable estimates of true population coverage.

Recent advances in Bayesian hierarchical modeling offer promising alternatives. Through Multilevel Regression and Post-Stratification (MRP), survey estimates can be adjusted even when certain demographic cells are small or underrepresented. The Bayesian framework allows for *partial pooling*—borrowing information across similar groups—to stabilize estimates and quantify uncertainty more effectively. This study applies a Bayesian post-stratification approach to a simulated population that mirrors the social and demographic structure of Nigeria, demonstrating how modern statistical tools can improve survey-based estimation of childhood immunization coverage.

2. Objectives of the Study

The primary objectives of this study are to:

1. Simulate a realistic population reflecting key demographic dimensions affecting childhood immunization coverage.

2. Compare three estimation methods–**Naïve estimation**, **Classical post-stratification**, and **Bayesian Multilevel Regression with Post-Stratification (MRP)** -- in recovering the true population coverage.

3. Evaluate the ability of Bayesian post-stratification to reduce bias and improve the accuracy of survey-based estimates.

3. Methodology and Data

3.1 Population Simulation

A synthetic population of 10 million individuals was generated to represent a national demographic structure similar to Nigeria's. The population was defined by four categorical variables commonly associated with health behavior disparities:

- **Age group**: 18–29, 30–44, 45–64, 65+
- **Sex**: Male, Female
- **Education level**: No formal/primary (HS or less), Some college, College and above
- **Region**: Northeast, Northwest, South, and West

Each unique combination of these characteristics forms a *demographic cell*. Randomly assigned population counts reflect realistic distributional differences across cells.

A "true" immunization uptake probability (`true_p`) was generated for each cell using a logistic model:

logit(p)=−1.2+Age effect+Sex effect+Education effect+Regional effect+ε\text{logit}(p) = -1.2 + \text{Age effect} + \text{Sex effect} + \text{Education effect} + \text{Regional effect} + \varepsilonlogit(p)=−1.2+Age effect+Sex effect+Education effect+Regional effect+ε

where the effects were chosen to reflect known behavioral patterns—higher vaccination likelihood among educated mothers, females, and urban or southern regions. Random noise (ε~N(0,0.15)\varepsilon \sim N(0, 0.15)ε~N(0,0.15)) was added to create heterogeneity.

The true national immunization coverage was then calculated as a weighted average of these probabilities across all cells.

3.2 Biased Survey Simulation

To simulate real-world data collection challenges, a biased survey sample of approximately 2,500 individuals was drawn from the population. Selection probabilities were designed to

oversample educated and urban populations (groups typically more responsive to surveys and health outreach programs) and undersample low-education, rural groups.

Additionally, response probability was correlated with the true probability of vaccination, representing a realistic form of **nonresponse bias**: households less likely to vaccinate their children were also less likely to participate in the survey.

The resulting sample thus reflected typical distortions found in real immunization surveys, where certain demographic segments are systematically underrepresented.

Each respondent's immunization status (**Y**) was then simulated as a binary outcome:

Yi~Bernoulli(pi)Y_i \sim \text{Bernoulli}(p_i)Yi~Bernoulli(pi)

where pip_ipi is the true probability of vaccination for the individual's demographic cell.

3.3 Estimation Methods

Three estimation techniques were applied to compare performance:

1. **Naïve Estimator**
   The simple mean of the survey responses (Y¯\bar{Y}Y¯) without any adjustment for sampling bias.

2. **Classical Post-Stratification (PS)**
   Estimated coverage was computed by weighting cell-level survey means according to their true population proportions. While effective under adequate cell representation, this method struggles with sparse data.

3. **Bayesian Multilevel Regression and Post-Stratification (MRP)**
   A hierarchical logistic regression model was fitted using PyMC:
   logit(pi)=α+aage[i]+asex[i]+aeduc[i]+aregion[i]\text{logit}(p_i) = \alpha + a_{\text{age}[i]} + a_{\text{sex}[i]} + a_{\text{educ}[i]} + a_{\text{region}[i]}logit(pi)=α+aage[i]+asex[i]+aeduc[i]+aregion[i]
   Random intercepts for each demographic variable allowed for **partial pooling**, where information from similar cells improved estimates for smaller groups. Posterior predictions for each population cell were then weighted by their population counts to derive a Bayesian-adjusted national estimate with a 95% credible interval.

4. Results and Analysis

The comparison of estimators yielded the following results:

Estimator	**Estimate**
True Population (ground truth)	0.2836

Naïve (sample)	0.3602
Classical Post-Stratification	0.2452
Bayesian MRP	0.3309

The **naïve sample estimate** (0.36) overstated true immunization coverage due to overrepresentation of more educated respondents.
The **classical post-stratification** adjusted for known margins but underestimated the true value (0.25), partly due to sparse or empty cells with small sample sizes.
In contrast, the **Bayesian MRP** estimate (0.33) was much closer to the true population mean (0.28), capturing the general pattern without over- or under-correction.

The Bayesian credible interval (approximately 0.31–0.35) also encompassed the true value, providing a realistic quantification of uncertainty. Visual comparison confirmed the Bayesian model's ability to shrink extreme group estimates toward plausible averages, reducing volatility and improving interpretability.

These results reinforce findings from applied survey literature: Bayesian hierarchical methods can outperform classical techniques when sample sizes are small, unbalanced, or when population heterogeneity is high.

5. Conclusion and Implications

This study demonstrates that Bayesian post-stratification (MRP) provides a more reliable and interpretable approach for estimating immunization coverage when survey data are biased or incomplete. By combining demographic modeling with Bayesian inference, MRP effectively "borrows strength" across related groups, stabilizing estimates even under sparse data conditions.

In practical terms, this approach could enhance the precision of vaccination coverage estimates in national surveys such as DHS and MICS, particularly in settings where survey nonresponse or fieldwork constraints hinder representativeness. The method's ability to integrate auxiliary data (e.g., census or administrative records) also makes it valuable for data fusion and small-area estimation, which are increasingly important in global health monitoring.

Future research could extend this framework to:

- Integrate real DHS or administrative data for validation,
- Model spatial correlations between regions,
- Explore Bayesian causal inference techniques to understand drivers of low vaccine uptake.

By combining methodological rigor with real-world relevance, this project highlights the potential of Bayesian methods to improve public health data accuracy and support evidence-based policy in developing contexts.