University of Nairobi

School of Mathematics

ISSN: 2410-1397

Master Project in Biometry

Monitoring Routine Health Indicators from District Health Information System (DHIS2): A Statistical Subsampling Approach

Research Report in Mathematics, Number 12, 2020

Abubaker Kalule

June 2020

Submitted to the School of Mathematics in partial fulfilment for a degree in Masters of Science in Biometry

Master Project in Biometry **University of Nairobi**
June 2020

Monitoring Routine Health Indicators from District Health Information System (DHIS2): A Statistical Subsampling Approach

Research Report in Mathematics, Number 12, 2020

Abubaker Kalule

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the School of Mathematics in partial fulfilment for a degree in Masters of Science in Biometry

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

Routine health facility data is collected using health information systems. In Kenya, it's collected using the District Health Information System (DHIS2). This data is continuously collected and cheaper to obtain compared to surveys. Currently, there has been increased advocacy for using this data by governments and development organizations such as the World Health Organization (WHO). Currently, it is unclear about how much DHIS2 data one needs to estimate indicators. All the studies that have used routine data use all the available reports to obtain estimates. This study proposes a novel sub-sampling approach to the estimation of indicators from routine data. The null-hypothesis the study set out was that smaller subsamples of routine data provide credible estimates.

Data from 1,808 health facilities in Western Kenya is obtained from DHIS2. Information of 5 data elements, the number of DPT1 doses, the number of DPT3 doses, the number of LLITNs distributed to pregnant mothers attending ANC, the number of pregnant women completing at least 4 ANC visits, and the number of pregnant women completing the first ANC visits are used to compute three indicators. The three indicators that were calculated from the 5 data elements are; the coverage of the third dose of pentavalent vaccine (DPT3), the proportion of pregnant women who receive LLINs, and the proportion of pregnant women who completed at least 4 ANC visits. The study then uses both spatial and non-spatial sampling to obtain proportions of data from the entire dataset and compute estimates. The proportions were 90%,80%,70%,60%,50%,40%,30%, and 20%. Spatial sampling was used because of the indicators of interest exhibit some spatial variability. The study then used a z-test to determine whether a significant difference exists between the subsample estimates and the population estimates. We also used power calculations to determine the statistical power each subsample had.

The results from the study indicate that there was no significant difference between the population estimate and sub-sample estimates after using both spatial and non spatial sampling (all p-values > 0.05). This implies that one doesn't need the whole data set to obtain estimates from DHIS2, and the sampling design doesn't matter unless the indicator of interest exhibits some spatial variation. However, based on the confidence intervals, we found that larger samples had narrower confidence intervals, so we recommend sampling above 60%. The power calculation also supported this conclusion. We found that although the power of the subsamples to obtain estimates was generally high (> 70%), it reduced as the sample size reduced.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

Signature Date

ABUBAKER KALULE
Reg No. I56/24856/2019

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

Signature Date

Dr Nelson Owuor Onyango
School of Mathematics
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: onyango@uonbi.ac.ke

Signature Date

Dr Victor Alegana
Population Health Unit (PHU),
KEMRI-Wellcome Trust Research Programmei,
P.O Box 43640 – 00100 Nairobi, Kenya.
E-mail: VAlegana@kemri-wellcome.org

Dedication

I dedicate this project to my friends and family for the support they have given me throughout the master’s program.

Contents

Figures and Tables

Figures

Tables

List of Abbreviations

ANC	Antenatal Care
CIs	Confidence Intervals
DHIS2	District Health Information System 2
DPT1	Pentavalent Vaccine Dose 1
DPT3	Pentavalent Vaccine Dose 3
FBO	Faith-Based Organization
GRTS	Generalized Random Tessellation Stratified sampling design
HF	Health Facility
KDHS	Kenya Demographic Health Survey
LLITNs	Long-Lasting Insecticide Treated Nets
MNH	Maternity and Nursing Home
MoH	Ministry of Health
NGO	Non-Governmental Organization
SSA	Sub-Saharan Africa
SDGs	Sustainable Development Goals
UN	United Nations

Acknowledgments

First, I would like to appreciate The Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB), for the scholarship to study this master at the University of Nairobi. Secondly, I would like to appreciate SSACAB, The Initiative to Develop African Research Leaders (IDeAL) and KEMRI Wellcome Trust Research Program (KWRTP) for the funding and support to conduct this research at KWTRP in Nairobi. Thirdly, I would like to appreciate the team at Population Health Unit (PHU) for the support throughout the time I conducted the research. Lastly, I thank my supervisors, Dr Nelson Owour and Dr Victor Alegana, for the mentorship and support.

“This work was supported through the DELTAS Africa Initiative Grant No. 107754/Z/15/Z-DELTAS Africa SSACAB and additional support from DELTAS Africa Initiative [DEL-15-003]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)’s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa’s Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant No. 107754/Z/15/Z) and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government”

Abubaker Kalule

Nairobi, 2020.

1 Introduction

1.1 Background

All United Nations (UN) member states adopted the Sustainable Development Goals (SDGs) in September 2015. Goal 3 sets out to ensure healthy lives and promote well-being for all at all ages UN General Assembly (2015). To track the progress of each country so that none is left behind, there is a need for quality and reliable data to design health indicators effectively. Also, the collection of quality data is essential for monitoring the performance of each country's health programs, such as immunization programs at the national and subnational levels I. Maina et al. (2017); Nabyonga-Orem (2017).

In practice, household surveys provide a vast amount of data that can be used for estimation of coverage indicators and for monitoring progress towards SDGs Cutts et al. (2013); Hancioglu & Arnold (2013); Khan et al. (2017). In Kenya, for instance, the Kenya Demographic and Health Survey (KDHS) provides enormous amounts of data for assessment of the population and health status of Kenya. The latest KDHS was conducted in 2014 and was powered to give estimates on several demographic and health indicators at the national, regional, and County levels KNBS (2015). While the KDHS can provide County-level estimates for over 100 data indicators, it is not statistically powered to give estimates at lower sub-national levels such as sub-counties and wards. Also, it is limited in terms of frequency I. Maina et al. (2017). The KDHS is expected to be conducted every five years, but in reality, that is usually not the case. For example, the last was done in 2014, and it is not clear if a KDHS will be conducted in 2021. It evident that it will not be done in 2020 either.

The collection of routine health data from health facilities, using Health Management Information Systems (HMIS), is an alternative solution Bhattacharya et al. (2019); Farnham et al. (2020); I. Maina et al. (2017). HMIS is designed for data collection from health facilities to help with planning, decision making, management, policy formulation, program monitoring, and evaluation. This data is captured when patients visit health facilities to seek services and reported monthly into the DHIS2. It, therefore, has several advantages over survey data. For example, since health facilities are meant to serve the whole population, the data collected can be used to obtain estimates of disease indicators at lower administrative levels, such as sub-counties and wards Hemkens et al. (2016). It is also cheaper to collect hence frequently available.

In Kenya, the District Health Information System 2 (DHIS2) was rolled out in 2011 to provide health facilities with a web-based tool to collect and manage their data. Most levels 4, 5, and 6 health facilities record the data directly into the system, while levels 2 and 3 submit their reports to their County or sub-County offices. This data forms the basis for annual health statistics reports and review of the performance of the health system. It is also used by many analysts to study trends of diagnoses, coverage of health interventions, and comparison of spatial differences.

Several studies have assessed the quality and use of routine health data in Sub Saharan Africa (SSA), and they reported a couple of limitations such as the incompleteness of health facility reporting, inconsistencies, presence of outliers, and under-reporting. Some studies have suggested solutions to address the above limitations, for example, using an adjustment factor to adjust for incompleteness I. Maina et al. (2017); Maïga et al. (2019). The World Health Organization (WHO) has also developed toolkits for review of data quality, guidelines for analysis, and the use of health facility data. Still, they only give general guidelines for all stakeholders, and this doesn't take into account country-specific health facility data challenges.

While routine data can provide coverage estimates at both national and sub-national levels, there is a need to ascertain the credibility of these estimates. For instance, it is unclear whether routine data can be used for monitoring all indicators or only those with near-universal coverage like vaccination, antenatal care, and whether it is possible to get these estimates at all administrative units. It is also unclear about how much data one needs to estimate indicators. All the studies that have used routine data use all the available reports to obtain estimates. This study proposes a novel sub-sampling approach to the estimation of indicators from routine data. The null-hypothesis the study set out was that smaller subsamples of routine data provide credible estimates. This implies that one doesn't need 100% of the health facility data, but can obtain a smaller sample and still obtain reliable estimates.

This study uses three indicators for testing the hypothesis through a series of experiments, namely, DPT3 Immunization coverage, the proportion of pregnant women who receive LLINs, and the proportion of ANC clients who complete at least 4 ANC visits. First, eight subsamples are obtained from the whole dataset using spatial and non-spatial spatial sampling methods, and then estimates of the three indicators are derived from the entire dataset. Lastly, a z-test and power calculations are used to determine the sampling threshold below which the estimates cease to be reliable.

1.2 Statement of Problem

Routine health data is collected using the DHIS2 in all health facilities in Kenya. This data is cheaper to obtain compared to conducting surveys, thus frequently available. Currently, there is increased advocacy to use routine health data in monitoring progress towards SDGs by multinational organizations such as WHO and MEASURE Evaluation, as well as governments. Still, the use of this data is limited by concerns about the quality of the data; such as incompleteness and inconsistency. Several studies have been conducted to determine the best approach to the use of routine health data. Solutions that have come up include removing outliers, obtaining denominators from routine data itself, using adjustment factors as proposed by I. Maina et al. (2017), among others.

It is, however, unclear if one needs to use the whole dataset or just a subsample to obtain reliable estimates. This study aims to conduct a series of experimentations to test the hypothesis that subsamples of routine data can provide credible estimates of indicators. We apply the methods to Western Kenya and use three indicators and two spatial units and different sampling methods to generalize the results.

1.3 Objectives

1.3.1 Overall objective

The broad aim of the study is to establish the effect of subsampling on the estimation of indicators from routine health facility data (DHIS2).

1.3.2 Specific objectives

1. To obtain subsamples from the whole data set using spatial and non-spatial sampling at different spatial units (Regional and County levels).
2. To obtain estimates of three indicators from the entire dataset and each subsample.
3. To test whether there is a significant difference between the population estimates and the sub-sample estimates to establish a threshold of sampling.

1.4 Justification of Study

The results of this study will act as guidelines for various stakeholders such as governments, researchers and program managers who need to use health facility data (DHIS2) for research, program monitoring and evaluation, policy formulation, and decision making. We also hope that the results of this study will apply to health facility surveys to inform researchers about how much of the health facilities are needed to answer research questions.

2 Literature Review

2.1 Introduction

This chapter reviews the literature on using routine health facility data for monitoring indicators. It expounds on the benefits of routine data compared to surveys, the problems, and how people have handled them in the past. It concludes by identifying a gap in the use of routine data.

2.2 The organization of Kenya's Health System

Health facilities in Kenya are categorized into six levels, as shown in Table 1 below.

Table 1. Levels of Kenya's health system

Categorization	Facility
Level 1	Community Health Facility
Level 2	Medical Clinic; Dental Clinic; Dispensary; Faith-Based Organizations (FBO); Mobile Clinic; Eye Clinic
Level 3	Basic Health Centre; Comprehensive Health Centre; Medical or Dental Centre; Funeral Home; Nursing Home or Cottage Hospital; Maternity Home
Level 4	County Hospital or Internship Training Centre; County Specialized Hospital
Level 5	County Referral Hospitals
Level 6	National Referral and Teaching Hospitals and Specialized Hospitals

Source (The Kenya Gazette, Vol. CXXII — No. 24, Date: 04-February 2020)

According to The Kenya Gazette (1925), Level 1 is composed of community health facilities, and their primary focus is to ensure that communities do appropriate healthy behaviours and to recognize signs and symptoms that need to be handled by other levels in the health system. Level 2 is composed of dispensaries and clinics. Such clinics include medical, dental, faith-based organizations, mobile and eye clinics. Level 3 is divided into level 3A which is composed of basic health centres and level 3B which is formed of comprehensive health centres; medical or dental centres; funeral homes; nursing homes or cottage hospitals and maternity homes. Level 4 contains county hospitals, internship training centres and county specialized hospitals. Level 5 contains county referral hospitals while level 6 contains national referral, teaching and specialized hospitals.

2.3 The use of routine data for monitoring indicators

Routine data is one that is collected continuously at various time intervals such as daily, weekly, monthly and so on. Because this kind of information is collected continuously, there is a considerable amount of it, which makes it cheaper to use for monitoring indicators and evaluation of programmes (Kane et al., 2000). This data can, for example, be used for malaria morbidity estimation Alegana, Okiro, & Snow (2020), and monitoring maternal and newborn indicators (Bhattacharya et al., 2019; Jordans et al., 2016; I. Maina et al., 2017). The introduction of routine health information systems (RHIS) such as the DHIS2 in low and middle-income countries has improved the collection of regular health facility data. Currently, health facility-based data is the principal source of data for monitoring and evaluation of national health programmes (Githinji et al., 2017). Besides, according to a study conducted by (Ndabarora et al., 2014), improved efficiency in the collection of health data was found to improve health care services delivery.

In Kenya, DHIS2 was rolled out countrywide in 2011 and significantly improved the national health system's capabilities. Health facilities in Kenya can report on all interventions and treatments given to patients in realtime or through their respective sub-counties and county health offices (Manya et al., 2012). In Uganda, a study by Kiberu et al. (2014) found out that the roll-out the DHIS2 led to a significant improvement in timeliness and completeness of reporting of data from health facilities.

The significance of routine data is overlooked because of problems such as incompleteness, inconsistency, non-representativeness and inaccuracy (Wagenaar et al., 2016). In a study done by Maïga et al. (2019) to assess the quality of routine health facility data across 14 countries in Africa, the results suggest that although completeness of reporting is high especially at the national levels, routine data presents specific issues especially at the subnational levels such as outliers, inconsistency overtime between indicators like vaccination and ANC, and challenges related to target populations that are needed to compute coverage indicators.

A lot of research has been done to address these challenges and put routine health facility data into use, for example, WHO developed several toolkits to guide the analysis and use of health facility data that is collected through HMIS. These toolkits cover core indicators, data quality review, and programme specific analysis such as immunization and malaria.

In Kenya, routine data has been used to determine the subnational coverage of maternal health indicators, and adjustment factors were used to solve the problem of incompleteness. Both the numerators and denominators used to formulate indicators were also derived from routine data. For example, the number of pregnant women or the number of children eligible for vaccination was the number of women receiving at least 4 ANC visits and the number of children receiving the 1st dose of pentavalent vaccine respectively I. Maina et al. (2017). We applied the same methods to derive the indicators used in the experimentations in our study. These methods are also in line with the WHO guidelines on analysis of health facility data. For example, according to WHO, the immunization coverage rate of a vaccine is formulated using the number of children who have received the vaccine as the numerator and the expected target population, such as children less than one year of age as the denominator. I. Maina et al. (2017) suggest that the target population size can be estimated from health facility data elements with near-universal coverage such as the first ANC or the first dose of pentavalent vaccine with a few adjustments on the number of children that may not access health facilities and hence not captured in DHIS2.

Routine health facility data has also been used by; I. Maina et al. (2017) to estimate the coverage of routine reporting on malaria parasitological testing in Kenya; Alegana, Khazenzi, et al. (2020) to estimate hospital catchments from in-patient admission records but they all used the whole dataset. Up to now, it is unclear whether using a fraction of the dataset will give the same results as using the entire dataset. No research has been conducted previously to determine the effect of subsampling on the estimation of indicators using routine health facility data. Our study attempts to investigate this novel approach to estimating indicators from routine data through a series of experiments at different spatial units, various sample sizes and sampling designs.

3 Methods

3.1 Introduction

This chapter discusses the materials and methods that were used to answer the study objectives. It discusses the dataset, the experimental design, and the testing of the study hypothesis.

3.2 Location of the study

The area of study was western Kenya (Figure 1) which comprises of eight counties; Bungoma, Busia, Kakamega, Kisumu, Homa Bay, Migori, Siaya, and Vihiga. Western Kenya was of interest in the study because of the recent roll-out of RTS,S, malaria vaccine for children six months of age and above by the Ministry of Health, through the National Vaccines and Immunization Programme. Western Kenya is also characterized by high child mortality, KDHS (2014), high malaria transmission rates Macharia et al. (2018) and high pregnancy-related mortality (Desai et al., 2013).

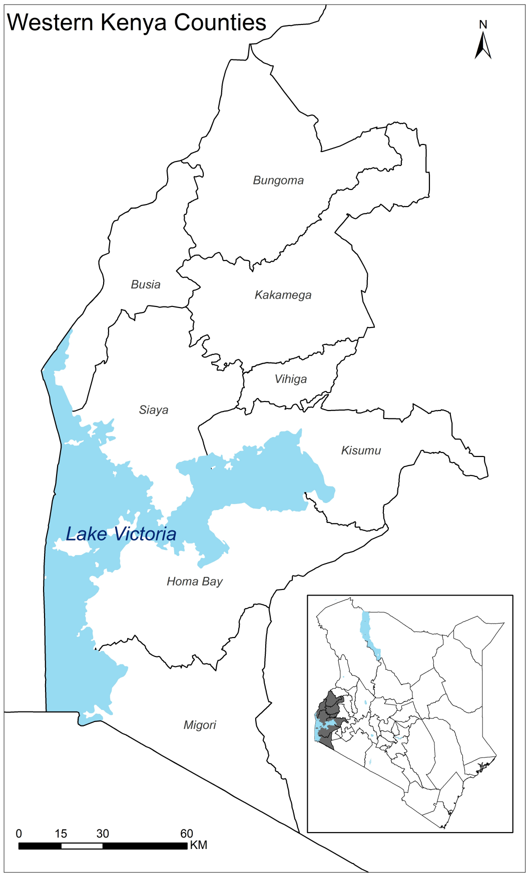

Figure 1. Map to show the location of the eight Counties in Western Kenya

3.3 Data set

3.3.1 Data source

The study used a dataset extracted from the District Health Information System (DHIS2) software. This data contained monthly reports by all western Kenya health facilities on five data elements, namely; the number of DPT1 doses, the number of DPT3 doses, the number of LLITNs distributed to pregnant mothers attending ANC, the number of pregnant women completing at least 4 ANC visits, and the number of pregnant women completing the first ANC visits. It included 1,808 health facilities, which reported between 2015 to 2019, spanning 60 months of reports per health facility.

The key variables in the data were the County where the health facility is located, the level, type, ownership of health facility, and geographical location (latitude, longitude). Most of the health facility coordinates were obtained from the spatial national health facility database for public health sector planning in Sub Saharan Africa (J. Maina et al., 2019). The other coordinates, especially those belonging to private facilities, were obtained through geocoding using google earth pro software. Information on the level, type, and ownership of the health facilities was obtained from the Kenya Master Health Facility List (KMHFL).

3.4 Indicators

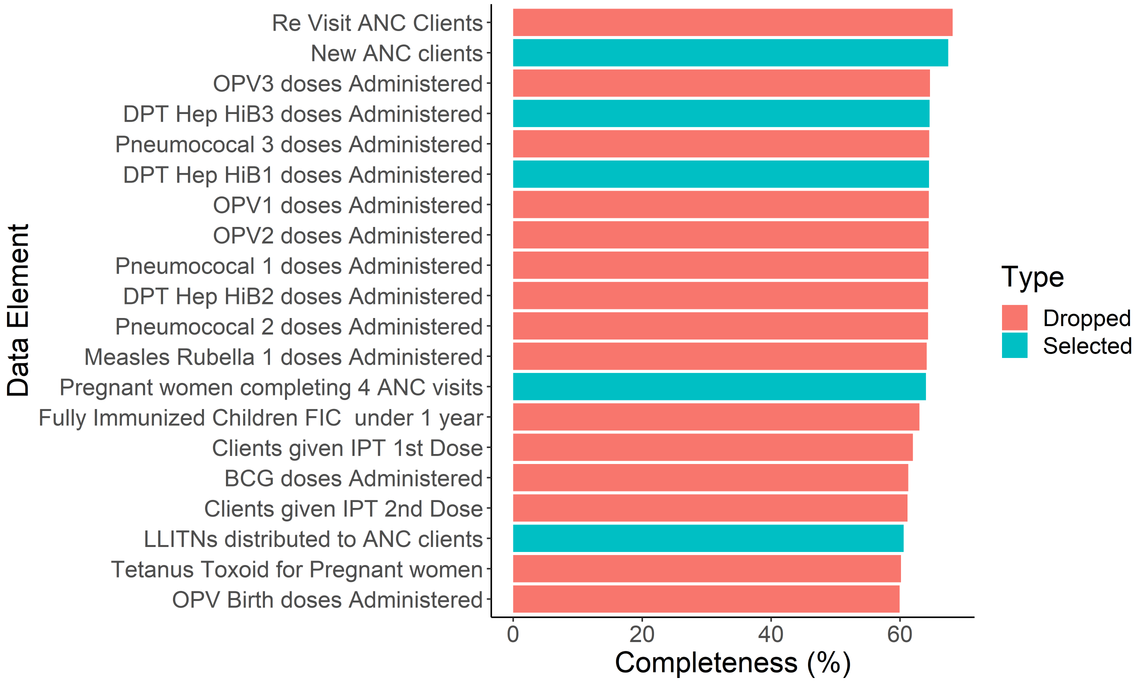

Figure 2. Completeness of key data elements

DHIS2 captures more than 1,000 data elements and 500 indicators. More than 200 immunization, antenatal care (ANC) and malaria data elements were extracted from DHIS2. From these, 20 data elements had a reporting percentage above 50%, and five were selected randomly, namely; the number of DPT1 doses, the number of DPT3 doses, the number of LLITNs distributed to pregnant mothers attending ANC, the number of pregnant women completing at least 4 ANC visits, and the number of pregnant women completing the first ANC visits as shown in Figure 2 above.

The three indicators that were computed from the 5 data elements are; the coverage of the third dose of pentavalent vaccine (DPT3), the proportion of pregnant women who receive LLINs, and the proportion of pregnant women who completed at least 4 ANC visits. The methods of coming up with the numerators and denominators, as shown in the equations below have been mentioned elsewhere (I. Maina et al., 2017).

$$DPT3\ coverage = \frac{Number\,of\,DPT3\,doses}{Number\,of\,DPT1\,doses} \tag{1}$$

$$Coverage\ of\ 4\ ANC\ visits = \frac{Number\,of\,pregnant\,women\,completing\,at\,least\,4\,ANC\,visits}{Number\,of\,pregnant\,women\,completing\,1\,ANC\,visit} \tag{2}$$

$$Proportion\ of\ pregnant\ women\ who\ receive\ LLINs = \frac{Number\,of\,LLINs\,distributed\,at\,HFs\,to\,ANC\,clients}{Total\,number\,of\,ANC\,clients} \tag{3}$$

3.5 Experimental Design

3.5.1 Generalization of results

To generalize the study findings, the methods were applied across three indicators: two spatial units and sampling was done using different sampling methods. In terms of the spatial units, first, experiments were done at the regional level (western Kenya) and the County level (Homa Bay County). The three indicators are already mentioned above. In terms of the sampling methods, we used both spatial and non-spatial sampling, comparing both simple and stratification designs. The stratification was done using the type of health facility and county as the stratification variables.

3.5.2 Sample Sizes

We used spatial and non-spatial sampling to obtain eight (8) subsamples from the whole dataset representing western Kenya. The obtained subsamples were equivalent to 90%; 80%; 70%; 60%; 50%; 40%; 30% and 20% of the whole dataset, equivalent to an actual sample size of 1,627; 1,446; 1,266; 1,085; 904; 723; 542 and 362 health facilities respectively (Table 2). The sampling frame was a list of all health facilities in Kenya (n = 1,808), and the sampling unit was the health facility. The null hypothesis was that the coverage estimates are stable with the different sample sizes.

Table 2. Size of the subsamples drawn from western Kenya

Sample	1	2	3	4	5	6	7	8
Sample Size (%)	90%	80%	70%	60%	50%	40%	30%	20%
Actual Sample Size (n)	1,627	1,446	1,266	1,085	904	723	542	362

HomaBay County had a total of 299 health facilities. From these, eight subsamples were obtained, and the subsequent subsamples drawn are shown in table 3 below.

Table 3. Size of the subsamples drawn from HomaBay County

Sample	1	2	3	4	5	6	7	8
Sample Size (%)	90%	80%	70%	60%	50%	40%	30%	20%
Actual Sample Size (n)	269	239	209	179	150	120	90	60

3.6 Spatial sampling

According to Delmelle (2009) when estimates of interest are characterized by spatial variation, it is necessary to use spatial sampling to obtain the optimal sample locations in a given study area. Spatial sampling is a 2-dimensional sampling which takes into account both the sample size and the location of the sampling units, hence providing a spatially balanced sample of health facilities in Western Kenya. The study used the generalized random tessellation stratified (GRTS) sampling design to conduct spatial/2-dimension sampling. GRTS is one of the spatially balanced design, and it generates an ordered list of sample units over the population of interest that exhibits spatial coverage. This means that the sample units obtained after sampling are not spatially clustered.

More technically, GRTS maps 2-dimensional space into one-dimensional space using a Quadrant-Recursive function. The Quadrant-Recursive function creates an ordered spatial address for each sampling unit, and this ordered spatial address helps preserve the spatial relationships. The heart of the GRTS sample selection method is a function f that maps the unit square $\Im^2 = (0,1] \times (0,1]$ onto the unit interval $\Im = (0,1]$.

To be useful in achieving a spatially balanced sample, f must preserve some proximity relationships, so we need to impose some restrictions on the class of functions to be considered. Mark (1990), in studying discrete two- to one-dimensional maps, defined a property called quadrant recursive, which required that sub-quadrants be mapped onto sets of adjacent points. To define the continuous analogue, let

$$Q^n_{jk} = \left(\frac{j}{2^n}, \frac{j+1}{2^n}\right] \times \left(\frac{k}{2^n}, \frac{k+1}{2^n}\right], \quad j,k = 0,1,\ldots,2^n-1 \tag{4}$$

$$J^n_m = \left(\frac{m}{4^n}, \frac{m+1}{4^n}\right], \quad m = 0,1,\ldots,4^n-1 \tag{5}$$

A function $f : \mathfrak{I}^2 \rightarrow \mathfrak{I}$ is quadrant recursive if, for all $n \geq 0$ there is some $m \in \{0, 1, \ldots, 4^n - 1\}$ such that $f\left(Q^n_{jk}\right) = J^n_m$.

Another aspect of the GRTS sample design is the reverse hierarchical ordering property that ensures that, for a sample of size n, the first n units in the sample will exhibit spatial balance. More generally, any contiguous set of sample units in reverse hierarchical ordered GRTS sample exhibits spatial balance.

A square grid overlay of a region becomes 4 grid cells within that original grid cell, and the subdivision continues with grids nesting within grids. When mapped onto a line, the first quarter of the points come from the first quarter of the original grid, the second quarter of points comes from the second quarter of the original grid, and so on (Figure 3). By selecting points from each quarter of the mapped line, spatially balanced design is guaranteed.

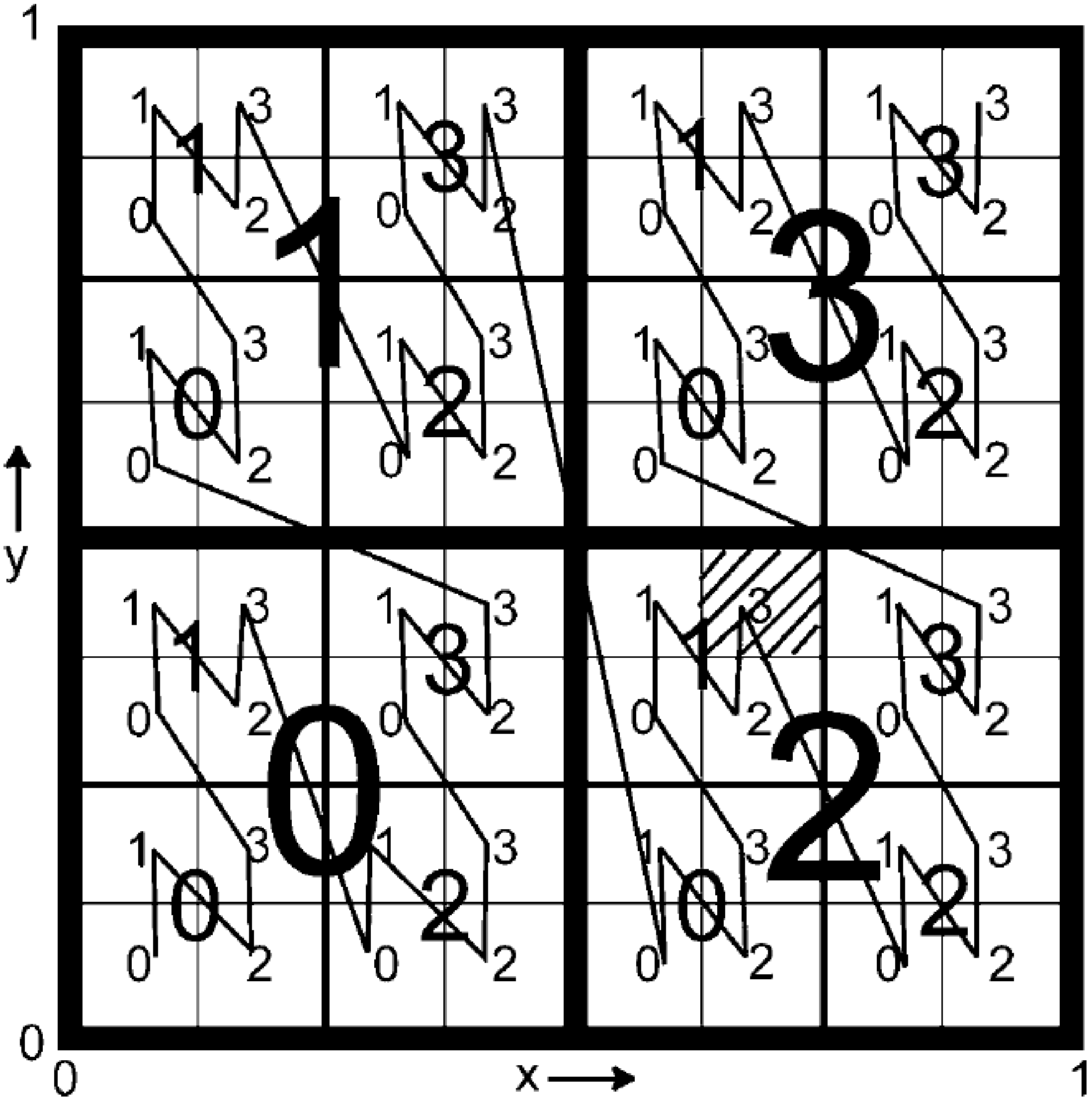

Figure 3. Quadrant-recursive partitioning of a unit square (Stevens Jr & Olsen, 2004)

3.6.1 Testing for spatial variability

To test for the spatial variability of the indicators under study, we used a variogram.

Using data values of the primary variable, an empirical A semivariogram $\widehat{\gamma}(h)$ uses dat values of the primary variable to summarizes the variance of values separated by a particular distance lag (h):

$$\widehat{\gamma}(h) = \frac{1}{2d(h)} \sum_{|s_i - s_j| = h} \left(y\left(\mathbf{s}_i\right) - y\left(\mathbf{s}_j\right)\right)^2 \tag{6}$$

where $d(h)$ is the number of pairs of points for a given distance lag value, and $y(\mathbf{s}_i)$ the observation value at location $(\mathbf{s}_i)$. The semivariogramis has a nugget effect a and a sill σ^2 where $\widehat{\gamma}(h)$ (semivariogram) levels out. The nugget effect reflects the spatial dependence at microscales, caused by measurement errors at distances smaller than sampling distances (Cressie, 2015). Once the lag distance exceeds the range r, there is no spatial dependence between the sample sites anymore. The semivariogram function $\widehat{\gamma}(h)$ becomes constant at a value called the sill σ^2. A model $\widehat{\gamma}(h)$ is fitted to the experimental variogram, for instance, an exponential model:

$$\gamma(h) = \sigma^2 \left(1 - e^{\frac{-3h}{r}}\right) \tag{7}$$

In the presence of a nugget effect a, Eq. (70.2) becomes

$$\gamma(h) = a + \left(\sigma^2 - a\right)\left(1 - e^{\frac{-3h}{r}}\right) \tag{8}$$

3.6.2 Simple Spatial Sampling

Under simple spatial sampling, ***m*** sample points are chosen randomly within the study area ***D***, and each location in ***D*** represented by latitude, and longitude has an equal chance of being sampled Ripley (1981). Each sample unit is chosen uniformly and independently from the other within ***D*** Delmelle (2009); King (1969).

3.6.3 Stratified Spatial Sampling

Under stratified spatial sampling, the study area ***D*** is divided into non-overlapping strata and the sum of sample units drawn from all the strata should be equal to ***m***. The stratification is informed using known covariates Delmelle (2009).

Both simple spatial and stratified sampling was done using the spsurvey package in R software version 3.6.2.

3.7 Test of hypothesis

A z-test was used to test whether there is a significant difference between the population estimate and the subsample estimates at a 95% confidence level. The z-test formula is given below.

$$Z = \frac{P_1 - P_0}{\sqrt{P_0(1-P_0)/N}} 4 \tag{9}$$

Where:

Z is the z-test calculated value

P$_1$ is the estimated proportion for the chosen sample size

P$_0$ is the estimated proportion for 100% sample size

N is the total number of health facilities (100% sample size)

The power calculations were done using the formula below;

$$n = (\frac{Z_{1-\alpha} + Z_{1-\beta}}{ES})^2 \tag{10}$$

$$ES = \frac{P_1 - P_0}{\sqrt{P_0(1-P_0)}} \tag{11}$$

Where:

n is the Sample Size

P$_1$ is the estimated proportion for the chosen sample size

P$_0$ is the estimated proportion for 100% sample size

1-β is the power of the sample size

4 Results

4.1 Introduction

This chapter presents the results of the data analysis procedures. It is divided into descriptive statistics, spatial maps, and results from both spatial and non-spatial sampling. The chapter ends with results from the experimentation at a lower administrative unit, Homa Bay County, using spatial sampling.

4.2 Descriptive Statistics

4.2.1 Summary of the key variables

A total of 1,808 health facilities located in western Kenya were included in the analysis. Kakamega County had the highest number of facilities, 16.7% (n=302), while Vihiga County had the least amount of facilities, 5.92% (n=107). In terms of the level, the majority were level 2 health facilities, 67.7% (n=1,224), followed by level 3 health facilities, 23.34% (n=422). In terms of the type of health facility, most were dispensaries, 44.8% (n=810), followed by clinics, 22.84% (n=413). In terms of ownership, 58.46% (n=1,057), of the health facilities, were owned by the Ministry of Health (Table 4).

Table 4. Summary of key variables

Variable	**n**	**%**
N = 1,808		
County		
Bungoma	237	13.11
Busia	143	7.91
Homa Bay	299	16.54
Kakamega	302	16.70
Kisumu	222	12.28
Migori	266	14.71
Siaya	232	12.83
Vihiga	107	5.92
Health Facility Level		
Level 2	1,224	67.70
Level 3	422	23.34
Level 4	160	8.85
Level 5	2	0.11
Facility Type		
Clinic	413	22.84
Dispensary	810	44.80
Health Centre	337	18.64
Hospital	162	8.96
Maternity & Nursing Home	24	1.33
Medical Centre	62	3.43
Ownership		
MoH	1,057	58.46
FBO	168	9.29
NGO	66	3.65
Private	517	28.60

4.2.2 Key data elements and indicators

The total number of children who received DPT1 between 2015 to 2019 was higher than those who received DPT3 in both western Kenya and Homa Bay County. 1,395,962 and 1,315,362 children were vaccinated with DPT1 and DPT3 in western Kenya, respectively. The DPT3 Immunization coverage was 0.942. In Homa Bay, 169,690, and 157,258 children were vaccinated with DPT1 and DPT3 respectively, bringing the DPT3 Immunization coverage to 0.927.

A similar trend was observed with antenatal care. The total number of pregnant women receiving at least one ANC visit was higher than those receiving 4 ANC visits. In western Kenya, 1,422,578 and 806,446 pregnant women received at least one and at least 4 ANC visits respectively. The proportion of pregnant women who completed at least 4 ANC visits was 0.567. In Homa Bay, the total number of pregnant women who received at least one and at least 4 ANC visits was 174,711 and 89,920 respectively, bringing the proportion of pregnant women who completed at least 4 ANC visits to 0.515. The proportion of pregnant women who received LLINs was 0.955 and 0.963 in western Kenya and Homa Bay, respectively (Table 5).

Table 5. The distribution of key data elements and indicators

Variable	Western Kenya	Homa Bay
Number of DPT1 Doses	1,395,962	169,690
Number of DPT3 Doses	1,315,362	157,258
Number of pregnant women receiving at least 1 ANC visit	1,422,578	174,711
Number of pregnant women receiving at least 4 ANC visit	806,446	89,920
Number of pregnant women receiving LLITNs	1,359,022	168,273
DPT3 Immunization coverage	0.942	0.927
The proportion of pregnant women who completed at least 4 ANC visits	0.567	0.515
The proportion of pregnant women who received LLINs	0.955	0.963

4.3 Maps of health facilities before and after spatial sampling

4.3.1 Distribution of health facilities in Western Kenya

Map A in Figure 4 below shows the distribution of all health facilities in Western Kenya (N = 1,808), B and C show the distribution of 80% (n = 1,446) and 60% (n= 1,085) of health facilities in Western Kenya selected randomly using simple spatial sampling. There is an indication of a reduction in the density of health facilities as the sample size reduces.

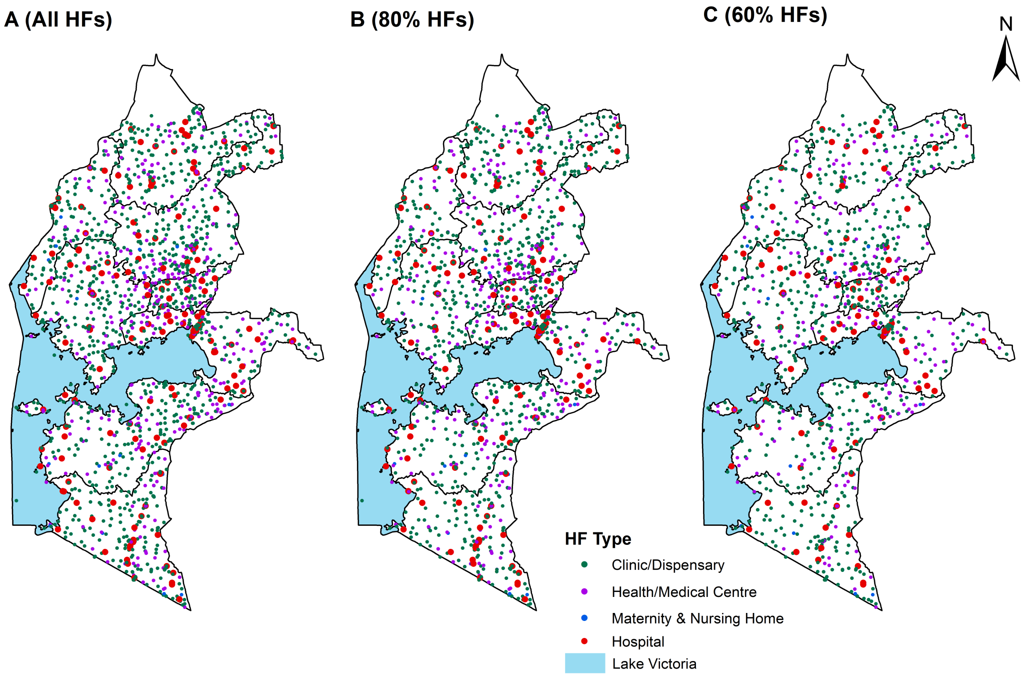

Figure 4. Western Kenya Maps Showing A) Distribution of all health facilities in Western Kenya (N = 1,808); B) Distribution of 80% of the health facilities after spatial sampling (n=1,446) C) Distribution of 60% of the health facilities after spatial sampling (n=1,085)

4.3.2 Distribution of health facilities in Homa Bay County

Map A in Figure 5. below shows the distribution of all health facilities in Homa Bay County (N = 299). Maps B and C show the distribution of 80% (n = 239) and 60% (n = 179) of the health facilities in Homa Bay County, respectively. There is a reduction in the density of health facilities as sample size reduces, as was seen in Figure 4.

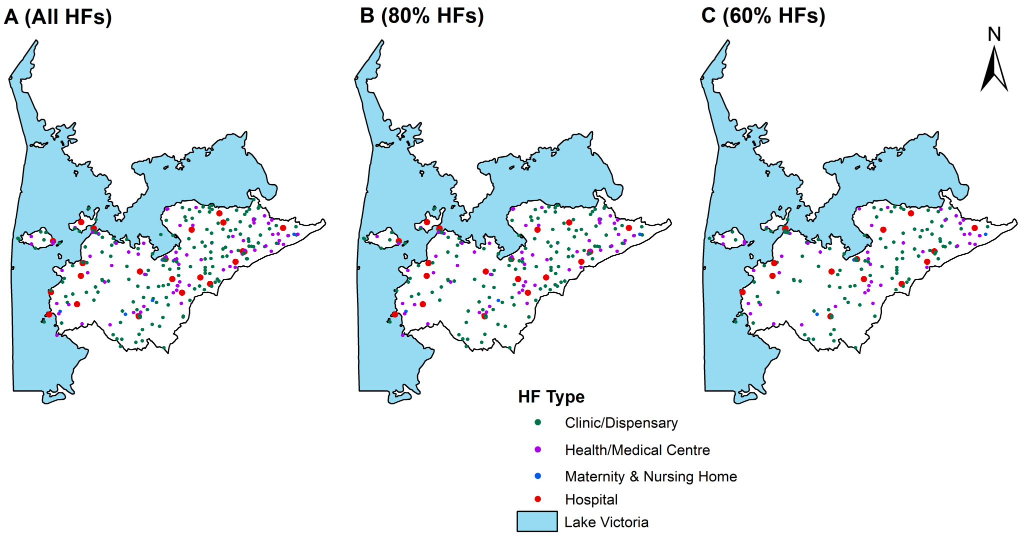

Figure 5. Homa Bay County Maps Showing A) Distribution of all health facilities in Homa Bay (N = 299); B) Distribution of 80% of the health facilities after spatial sampling (n=239) C) Distribution of 60% of the health facilities after spatial sampling (n = 179)

4.3.3 Testing for spatial variability

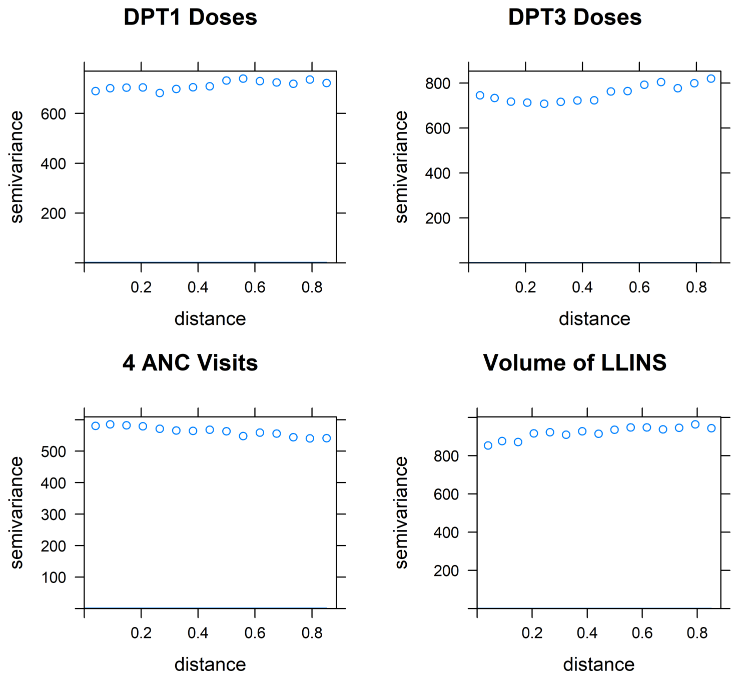

Figure 6. Variograms for the 5 data elements

The results in Figure 6 show that there is some little spatial variation in 3 data elements, that is, the number of DPT1 doses, the number of pregnant women who received at least one ANC visit, and the number of LLINs distributed to pregnant women.

This is a motivation for use of spatial sampling since it accounts for the spatial variability in the outcome of interest compared to non-spatial sampling that samples in a 1-dimension framework.

Two data elements, the number of DPT3 doses and the number of pregnant women who received at least four ANC visits showed little and no spatial variation respectively as shown in Figure 6 above.

4.4 Estimation of DPT3 Coverage in Western Kenya

4.4.1 Simple Random Sampling for Estimation of DPT3 Coverage

Table 6. Estimates of DPT3 coverage in Western Kenya obtained after Simple Random Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1,808	0.942	0.095	0.756	1.128	Ref	Ref	Ref
90%	1,627	0.942	0.100	0.746	1.137	-0.097	0.794	96.9
80%	1,446	0.944	0.108	0.733	1.155	0.292	0.765	95.5
70%	1,266	0.941	0.112	0.721	1.162	-0.185	0.784	96.4
60%	1,085	0.939	0.120	0.705	1.174	-0.508	0.701	94.1
50%	904	0.938	0.130	0.684	1.193	-0.723	0.614	92.6
40%	723	0.935	0.141	0.658	1.212	-1.307	0.340	87.1
30%	542	0.929	0.155	0.624	1.233	-2.422	0.043	73.7
20%	362	0.945	0.217	0.519	1.370	0.416	0.732	96.2

Table 6 show the DPT3 coverage estimates obtained from the subsamples after simple random sampling (SRS), the standard error, confidence intervals of the estimates, p-values from the z-test and the statistical power. The estimates obtained from the 8 subsamples were close to the population estimate of 0.942. The 60% subsample had the smallest estimate of 0.939 while at 20%, the DPT3 coverage estimate was the highest (0.945). A z-test that was done to test the difference between subsample estimates and the population estimate found that there was no statistically significant difference betweeen the population estimate and estimates from all subsamples (all p-values > 0.05). In addition, the statistical power of the subsamples was general high except at the 30% subsample, which had a power of 73.7%. The rest of the subsamples had a power greater than 80%.

4.4.2 Stratified Sampling for Estimation of DPT3 Coverage

Table 7. Estimates of DPT3 coverage in Western Kenya obtained after Stratified Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1,808	0.942	0.095	0.756	1.128	Ref	Ref	Ref
90%	1,628	0.941	0.099	0.747	1.134	-0.276	0.768	95.5
80%	1,450	0.942	0.106	0.735	1.149	-0.054	0.797	97.2
70%	1,263	0.942	0.113	0.720	1.163	-0.111	0.793	96.9
60%	1,086	0.937	0.117	0.707	1.168	-0.875	0.544	90.0
50%	903	0.948	0.142	0.670	1.225	0.991	0.488	89.6
40%	722	0.945	0.154	0.643	1.246	0.452	0.720	95.3
30%	545	0.946	0.180	0.593	1.299	0.763	0.596	93.8
20%	358	0.936	0.201	0.541	1.330	-1.217	0.380	92.2

The of DPT3 coverage estimates obtained from the subsamples drawn using stratified sampling were also found to be similar to the entire dataset estimate of 0.942, as shown in Table 7. The results from the z test also indicate that there was no significant difference between the estimate computed using the entire data set and the estimates computed from subsample or just a proportion of the data set (all p-values > 0.05). The statistical power of the subsamples was generally high, that is, all above 80%. There was however an indication of reduction of the statistical power with decrease in sample size. At 90% sample size, the power was 95.5% while at 20%, it was found to be 92.2%.

4.4.3 Simple Spatial Sampling for Estimation of DPT3 Coverage

Table 8. Estimates of DPT3 coverage in Western Kenya obtained after Simple Spatial Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1,808	0.942	0.095	0.756	1.128	Reference		
90%	1,627	0.941	0.099	0.747	1.136	-0.182	0.785	96.3
80%	1,446	0.943	0.107	0.734	1.152	0.182	0.785	96.4
70%	1,266	0.942	0.114	0.720	1.165	0.000	0.798	97.5
60%	1,085	0.943	0.123	0.702	1.184	0.182	0.785	96.6
50%	904	0.939	0.130	0.684	1.194	-0.546	0.687	94.2
40%	723	0.947	0.158	0.638	1.257	0.910	0.528	91.7
30%	542	0.952	0.190	0.579	1.325	1.819	0.153	83.2
20%	362	0.936	0.201	0.542	1.330	-1.091	0.440	92.9

The study found similar results after Simple Spatial Sampling. The DPT3 Immunization coverage estimates obtained from the subsamples were close with the population estimate of 0.942 (8). The z test also indicated that there was no significant difference between the estimates obtained from the entire data set and those obtained from subsamples (all p-values > 0.05). All the subsamples have statistical power above 80%, but there was an indication of reduction in statistical power as the sample size reduced.

4.4.4 Stratified Spatial Sampling for Estimation of DPT3 Coverage

Table 9. Estimates of DPT3 coverage in Western Kenya obtained after Stratified Spatial Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1,808	0.942	0.095	0.756	1.128	Reference		
90%	1,627	0.942	0.100	0.746	1.138	-0.014	0.798	97.4
80%	1,448	0.942	0.106	0.734	1.150	-0.013	0.798	97.4
70%	1,264	0.941	0.112	0.721	1.160	-0.314	0.759	95.5
60%	1,084	0.946	0.127	0.697	1.196	0.721	0.615	92.0
50%	905	0.939	0.131	0.683	1.195	-0.578	0.675	94.0
40%	724	0.948	0.159	0.637	1.259	1.067	0.452	90.1
30%	544	0.939	0.168	0.609	1.269	-0.572	0.678	95.0
20%	360	0.945	0.218	0.517	1.373	0.500	0.704	95.9

After Stratified Spatial Sampling, the DPT3 Immunization coverage estimates obtained from the subsamples were also similar with the population estimate (Table 9). The z test also indicated that there was no significant difference between the estimates obtained from the entire data set and those obtained from subsamples (all p-values > 0.05). All the subsamples have statistical power above 90%, but there was an indication of reduction in statistical power as the sample size reduced.

4.4.5 Bar charts to visually represent the DPT3 coverage estimates and their respective confidence intervals across the 4 sampling schemes

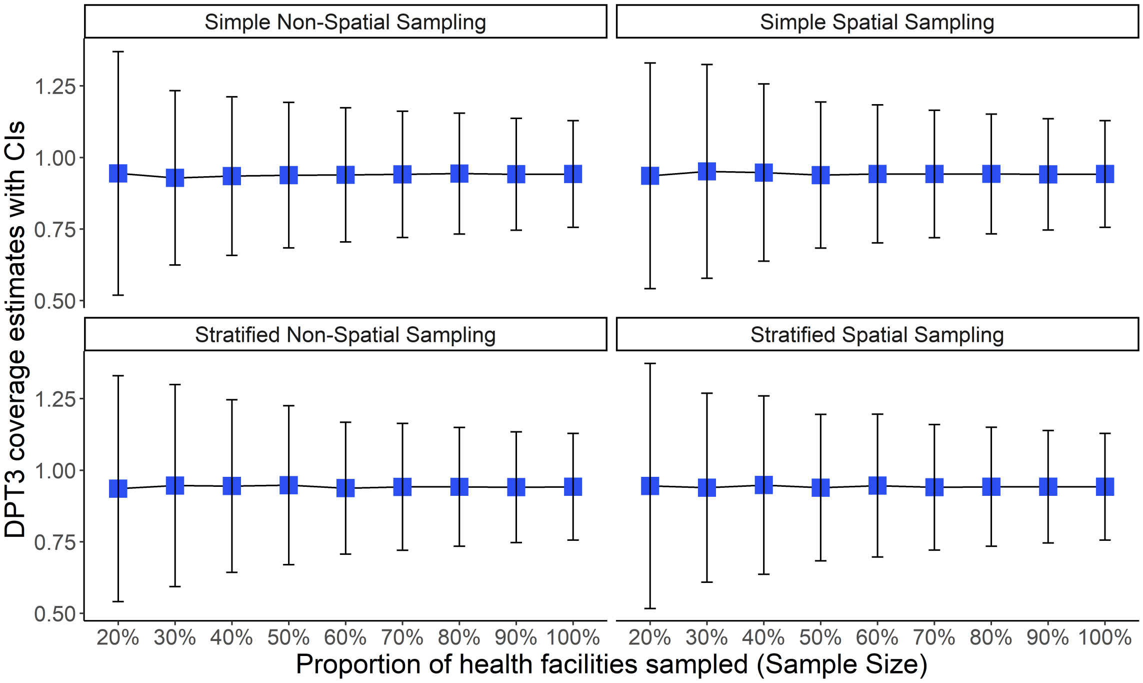

Figure 7. DPT coverage estimates with CIs versus sample size

The visual representation of the estimates and their respective confidence intervals (CIs) across the 4 sampling designs in Figure 7 indicates that the the estimates are similar at different sample sizes, but there is and increment in the width of the CIs as the sample size reduces. In addition, there is an overlap of the CIs which is still an indication of lack of significant difference in the estimates at different sample sizes.

4.5 Estimation of ANC Coverage in Western Kenya

4.5.1 Simple Random Sampling for Estimation of ANC Coverage

Table 10. Estimates of ANC coverage in Western Kenya obtained after Simple Random Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1,808	0.567	0.027	0.514	0.620	Reference		
90%	1,627	0.569	0.028	0.513	0.625	0.184	0.785	96.3
80%	1,446	0.566	0.030	0.507	0.625	-0.077	0.796	97.1
70%	1,266	0.568	0.032	0.505	0.631	0.119	0.792	96.9
60%	1,085	0.561	0.034	0.493	0.628	-0.532	0.693	93.9
50%	904	0.568	0.038	0.493	0.643	0.106	0.793	97.0
40%	723	0.574	0.043	0.489	0.658	0.571	0.678	94.5
30%	542	0.569	0.049	0.472	0.665	0.157	0.788	97.0
20%	362	0.559	0.059	0.443	0.675	-0.675	0.635	95.1

After Simple Random Sampling, the ANC coverage estimates obtained from the subsamples were consistent with the population estimate (Table 10). The z test also indicated that there was no significant difference between the estimates obtained from the entire data set and those obtained from subsamples (all p-values > 0.05). In addition, all the subsamples have statistical power above 90%.

4.5.2 Stratified Sampling for Estimation of ANC Coverage

Table 11. Estimates of ANC coverage in Western Kenya obtained after Stratified Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1,808	0.567	0.027	0.514	0.620	Reference		
90%	1,628	0.568	0.028	0.513	0.624	0.126	0.792	96.7
80%	1,450	0.564	0.030	0.505	0.622	-0.280	0.767	95.6
70%	1,263	0.563	0.032	0.500	0.625	-0.354	0.749	95.2
60%	1,086	0.560	0.034	0.493	0.627	-0.566	0.680	93.6
50%	903	0.572	0.038	0.496	0.647	0.401	0.736	95.3
40%	722	0.562	0.042	0.479	0.644	-0.440	0.724	95.4
30%	545	0.561	0.048	0.466	0.656	-0.497	0.705	95.4
20%	358	0.576	0.062	0.455	0.697	0.802	0.578	94.6

A similar trend was observed after Stratified Sampling. The ANC coverage estimates obtained from the subsamples were consistent with the population estimate (Table 11). The z test also indicated that there was no significant difference between the estimates obtained from the entire data set and those obtained from subsamples (all p-values > 0.05). In addition, all the subsamples have statistical power above 90%.

4.5.3 Simple Spatial Sampling for Estimation of ANC Coverage

Table 12. Estimates of ANC coverage in Western Kenya obtained after Simple Spatial Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1808	0.567	0.027	0.514	0.620	Reference		
90%	1627	0.565	0.028	0.509	0.620	-0.172	0.786	96.4%
80%	1446	0.569	0.030	0.510	0.628	0.172	0.786	96.5%
70%	1266	0.567	0.032	0.504	0.630	0.000	0.798	97.5%
60%	1085	0.564	0.035	0.496	0.632	-0.257	0.772	96.1%
50%	904	0.565	0.038	0.491	0.639	-0.172	0.786	96.7%
40%	723	0.561	0.042	0.479	0.644	-0.515	0.699	94.9%
30%	542	0.561	0.049	0.466	0.657	-0.515	0.699	95.3%
20%	362	0.570	0.060	0.451	0.688	0.257	0.772	96.7%

A similar trend was observed with the proportion of pregnant women who receive at least four (4) Antenatal visits. The estimates from the subsamples were consistent with the population estimate (Table 12), and the confidence intervals (CIs) became wider with smaller sample sizes which implied that there is a reduction in the precision of the estimate as the sample size reduces. Also, there was a huge overlap of the CIs, which implied that there was no significant difference between the estimates of the subsamples and the population estimate. This was again confirmed by the results from the z-test (all p-values > 0.05).

4.5.4 Stratified Spatial Sampling for Estimation of ANC Coverage

Table 13. Estimates of ANC coverage in Western Kenya obtained after Stratified Spatial Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1,808	0.567	0.027	0.514	0.620	Reference		
90%	1,627	0.568	0.028	0.512	0.623	0.055	0.797	97.2
80%	1,448	0.564	0.030	0.505	0.622	-0.263	0.771	95.8
70%	1,264	0.568	0.032	0.504	0.631	0.063	0.796	97.2
60%	1,084	0.567	0.035	0.499	0.635	0.027	0.798	97.4
50%	905	0.557	0.037	0.484	0.630	-0.831	0.565	91.5
40%	724	0.575	0.043	0.490	0.659	0.659	0.642	93.9
30%	544	0.559	0.048	0.464	0.653	-0.704	0.623	94.2
20%	360	0.577	0.062	0.456	0.698	0.881	0.541	94.1

A similar trend was observed after Stratified Spatial Sampling. The ANC coverage estimates obtained from the subsamples were consistent with the population estimate (Table 11). The z test also indicated that there was no significant difference between the estimates obtained from the entire data set and those obtained from subsamples (all p-values > 0.05). In addition, all the subsamples have statistical power above 90%.

4.5.5 Bar charts to visually represent the ANC coverage estimates and their respective confidence intervals across the 4 sampling schemes

Again, the visual representation of the estimates and their respective confidence intervals (CIs) across the 4 sampling designs in Figure 8 below indicates that the the estimates are similar at different sample sizes, but there is and increment in the width of the CIs as the sample size reduces. In addition, there is an overlap of the CIs which is still an indication of lack of significant difference in the estimates at different sample sizes.

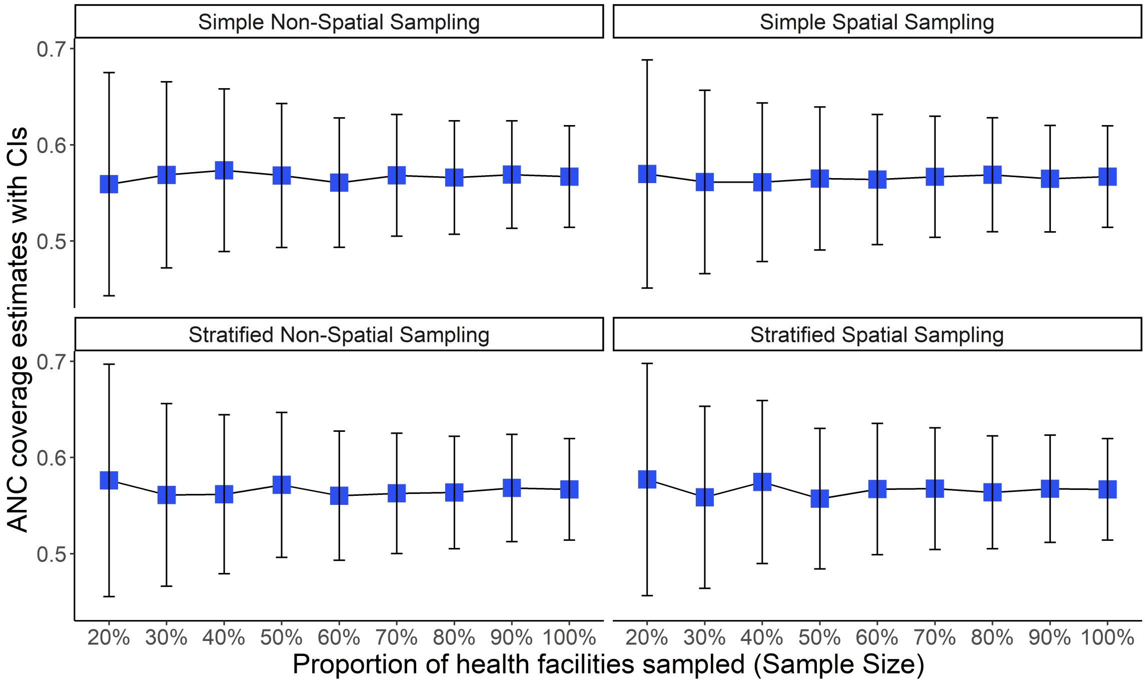

Figure 8. ANC coverage estimates with CIs versus sample size

4.6 Estimation of LLINs Coverage in Western Kenya

4.6.1 Simple Random Sampling for Estimation of LLINs Coverage

Table 14. Estimates of the proportion of pregnant mothers in Western Kenya obtained after Simple Random Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1,808	0.955	0.109	0.742	1.168	Reference		
90%	1,627	0.955	0.114	0.731	1.180	-0.026	0.798	97.4
80%	1,446	0.956	0.122	0.716	1.195	0.081	0.795	97.0
70%	1,266	0.957	0.132	0.698	1.216	0.281	0.767	95.8
60%	1,085	0.959	0.146	0.672	1.245	0.699	0.625	92.2
50%	904	0.948	0.142	0.670	1.227	-1.498	0.260	81.6
40%	723	0.957	0.175	0.614	1.300	0.296	0.764	96.2
30%	542	0.957	0.202	0.561	1.352	0.267	0.770	96.5
20%	362	0.950	0.230	0.500	1.401	-1.025	0.472	93.3

A similar trend was observed after Simple Random Sampling. The LLINs coverage estimates obtained from the subsamples were consistent with the population estimate (Table 14). The z test also indicated that there was no significant difference between the estimates obtained from the entire data set and those obtained from subsamples (all p-values > 0.05). In addition, all the subsamples have statistical power above 90%.

4.6.2 Stratified Sampling for Estimation of LLINs Coverage

A similar trend was observed after Stratified Sampling. The LLINs coverage estimates obtained from the subsamples were consistent with the population estimate (Table 15). The z test also indicated that there was no significant difference between the estimates obtained from the entire data set and those obtained from subsamples (all p-values > 0.05). In addition, all the subsamples have statistical power above 90%.

Table 15. Estimates of the proportion of pregnant mothers in Western Kenya obtained after Stratified Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1,808	0.955	0.109	0.742	1.168	Reference		
90%	1,628	0.957	0.116	0.729	1.184	0.249	0.774	95.8
80%	1,450	0.956	0.122	0.717	1.195	0.065	0.796	97.1
70%	1,263	0.952	0.125	0.707	1.197	-0.703	0.623	91.5
60%	1,086	0.955	0.140	0.680	1.231	0.009	0.798	97.5
50%	903	0.953	0.149	0.660	1.245	-0.536	0.691	94.3
40%	722	0.949	0.161	0.634	1.265	-1.253	0.364	87.9
30%	545	0.959	0.208	0.552	1.367	0.811	0.574	93.5
20%	358	0.973	0.314	0.356	1.589	3.537	0.002	65.0

4.6.3 Simple Spatial Sampling for Estimation of LLINs Coverage

Table 16. Estimates of the proportion of pregnant mothers in Western Kenya obtained after Simple Spatial Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1808	0.955	0.109	0.742	1.168	Reference		
90%	1627	0.956	0.116	0.729	1.184	0.205	0.781	96.1%
80%	1446	0.955	0.121	0.718	1.192	0.000	0.798	97.5%
70%	1266	0.956	0.131	0.700	1.212	0.205	0.781	96.3%
60%	1085	0.949	0.132	0.692	1.207	-1.231	0.374	84.3%
50%	904	0.964	0.173	0.625	1.304	1.846	0.145	74.4%
40%	723	0.945	0.155	0.642	1.248	-2.051	0.097	74.6%
30%	542	0.958	0.206	0.555	1.362	0.615	0.660	94.8%
20%	362	0.971	0.304	0.375	1.566	3.282	0.004*	68.8%

A similar trend was observed after Simple Spatial Sampling. The LLINs coverage estimates obtained from the subsamples were consistent with the population estimate (Table 16). The z test also indicated that there was no significant difference between the estimates obtained from the entire data set and those obtained from subsamples (all p-values > 0.05). In addition, all the subsamples have statistical power above 90%.

4.6.4 Stratified Spatial Sampling for Estimation of LLINs Coverage

Table 17. Estimates of the proportion of pregnant mothers in Western Kenya obtained after Stratified Spatial Sampling

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	1,808	0.955	0.109	0.742	1.168	Reference		
90%	1,627	0.957	0.117	0.728	1.186	0.317	0.759	95.1
80%	1,448	0.952	0.117	0.723	1.181	-0.701	0.624	0.909
70%	1,264	0.954	0.128	0.703	1.204	-0.314	0.759	95.5
60%	1,084	0.949	0.131	0.692	1.206	-1.302	0.342	82.9
50%	905	0.954	0.151	0.658	1.249	-0.336	0.754	95.8
40%	724	0.959	0.180	0.606	1.313	0.812	0.574	92.6
30%	544	0.961	0.214	0.542	1.380	1.238	0.371	90.0
20%	360	0.951	0.231	0.497	1.404	-0.953	0.507	93.8

A similar trend was observed after Stratified Spatial Sampling. The LLINs coverage estimates obtained from the subsamples were consistent with the population estimate (Table 17). The z test also indicated that there was no significant difference between the estimates obtained from the entire data set and those obtained from subsamples (all p-values > 0.05). In addition, all the subsamples have statistical power above 90%.

4.6.5 Bar charts to visually represent the LLINs coverage estimates and their respective confidence intervals across the 4 sampling schemes

Again, the graphical representation of the LLINs estimates and their respective confidence intervals (CIs) across the 4 sampling designs in Figure 9 below indicates that the the estimates are similar at different sample sizes, but there is and increment in the width of the CIs as the sample size reduces. In addition, there is an overlap of the CIs which is still an indication of lack of significant difference in the estimates at different sample sizes.

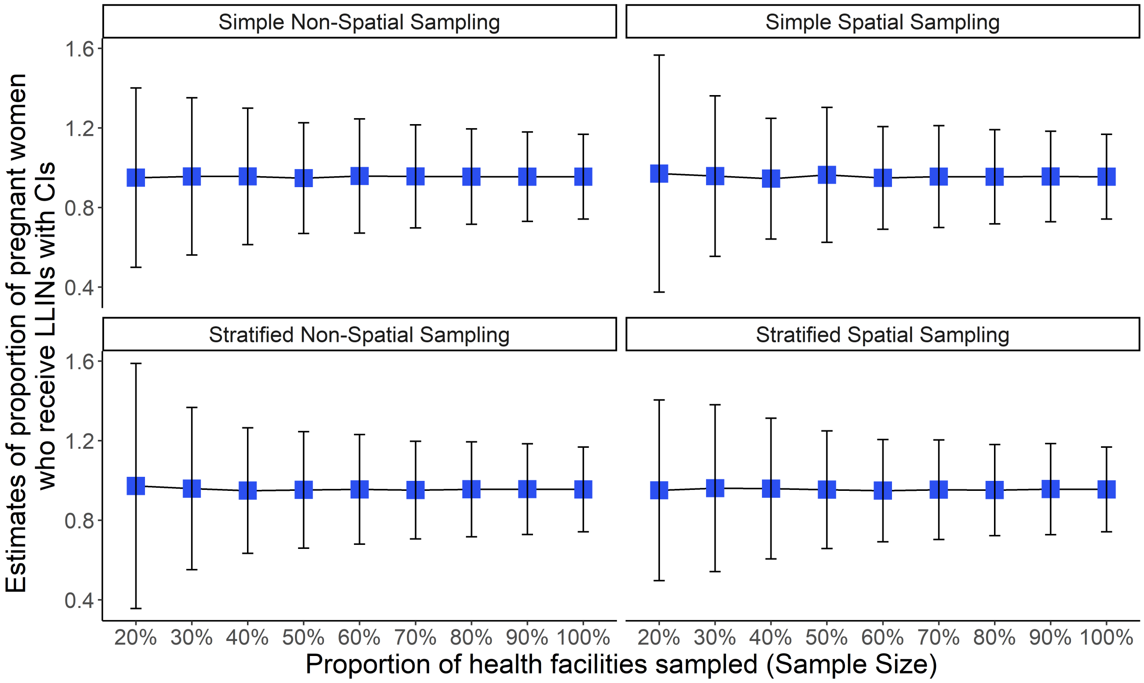

Figure 9. LLINs coverage estimates with CIs versus sample size

4.7 HomaBay County

4.7.1 The proportion of children who receive three (3) doses of pentavalent vaccine (DPT3) in Homa Bay County

The results indicate that there was no significant difference between the population estimate and all the sub-sample estimates of the proportion of children who receive DPT3 in Homa Bay County (all p-values > 0.05) (Table 18). The subsamples also exhibited statistical power above 90%.

Table 18. DPT3 coverage estimates after stratified spatial sampling in Homa Bay county

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	299	0.927	0.206	0.524	1.330	Reference		
90%	269	0.925	0.214	0.505	1.345	-0.108	0.793	96.7%
80%	239	0.925	0.227	0.480	1.370	-0.116	0.793	96.7%
70%	209	0.925	0.243	0.449	1.402	-0.109	0.793	96.8%
60%	179	0.921	0.255	0.421	1.420	-0.398	0.737	95.1%
50%	150	0.923	0.284	0.368	1.479	-0.216	0.779	96.2%
40%	120	0.927	0.324	0.291	1.562	-0.014	0.798	97.5%
30%	90	0.926	0.372	0.196	1.656	-0.060	0.796	97.3%
20%	60	0.934	0.486	-0.018	1.886	0.483	0.710	96.0%

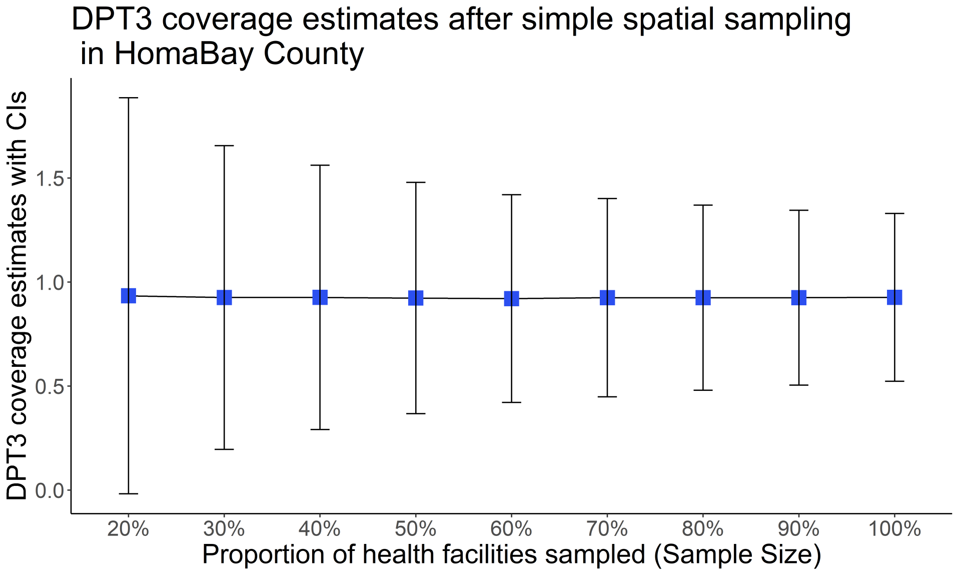

Figure 10. DPT3 coverage estimates with CIs versus sample size after stratified spatial sampling in HomaBay county

A plot of the estimate with their respective CIs reveals that they become wider as the sample size reduces. However, they are overlapping which indicates that there is no significant difference between the population estimate and subsample estimates (Figure 10).

4.7.2 The proportion of pregnant women who receive at least four (4) Antenatal visits in Homa Bay County

The data in table 19 indicates that the sub-sample estimates were less consistent with the population estimate. This could be caused by the small number of health facilities in Homa Bay compared to Western Kenya. There was, however, no significant difference between the population estimate and the sub-sample estimates of the proportion of pregnant women who receive at least 4 ANCs (all P-Values > 0.05).

Table 19. ANC coverage estimates after stratified spatial sampling in Homa Bay county

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	299	0.515	0.060	0.398	0.631	Reference		
90%	269	0.518	0.063	0.394	0.642	0.124	0.792	96.9%
80%	239	0.521	0.067	0.389	0.653	0.216	0.780	96.2%
70%	209	0.505	0.070	0.368	0.642	-0.341	0.753	95.3%
60%	179	0.515	0.077	0.364	0.665	-0.006	0.798	97.5%
50%	150	0.522	0.085	0.355	0.690	0.263	0.771	96.3%
40%	120	0.554	0.102	0.355	0.754	1.372	0.312	86.5%
30%	90	0.498	0.105	0.292	0.703	-0.585	0.672	94.9%
20%	60	0.519	0.134	0.256	0.782	0.152	0.789	97.1%

A similar trend was observed for ANC visits. A plot of the estimate with their respective CIs reveals that they become wider as the sample size reduces. However, they are overlapping which indicates that there is no significant difference between the population estimate and subsample estimates (Figure 11).

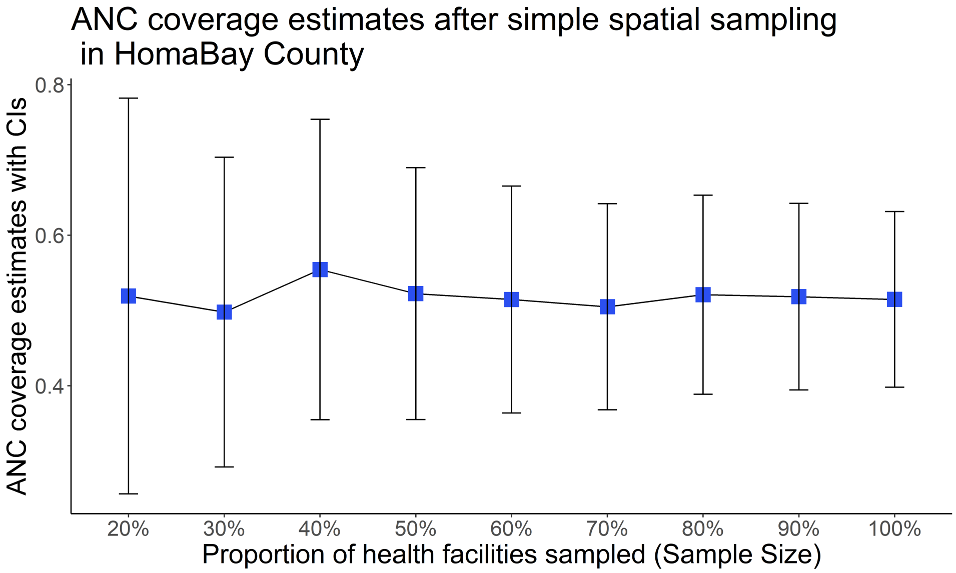

Figure 11. ANC coverage estimates with CIs versus sample size after simple spatial sampling in HomaBay County

4.7.3 The proportion of pregnant women who receive LLITNs (Mosquito Nets) in Homa Bay County

The results in Table 20 below indicate that there is a consistency between the population estimate and the sample estimates pregnant women who receive LLITNs. There was no significant difference between the population estimate and all the sub-sample estimates. (all p-values > 0.05).

Table 20. Estimates of the proportion of pregnant women who receive LLITNs obtained after simple spatial sampling in Homa Bay county

Sample Size (%)	Sample Size (n)	Estimate	Standard Error	LCL	UCL	z-test	p-value	Power (%)
100%	299	0.963	0.296	0.384	1.543	Reference		
90%	269	0.966	0.323	0.333	1.599	0.224	0.778	95.5
80%	239	0.961	0.322	0.330	1.593	-0.173	0.786	96.4
70%	209	0.970	0.391	0.203	1.737	0.601	0.666	92.3
60%	179	0.969	0.416	0.153	1.784	0.512	0.700	93.8
50%	150	0.977	0.534	-0.069	2.023	1.283	0.351	85.4
40%	120	0.968	0.503	-0.017	1.953	0.452	0.720	95.3
30%	90	0.958	0.506	-0.033	1.950	-0.438	0.725	95.6
20%	60	0.963	0.655	-0.320	2.246	-0.054	0.797	97.5

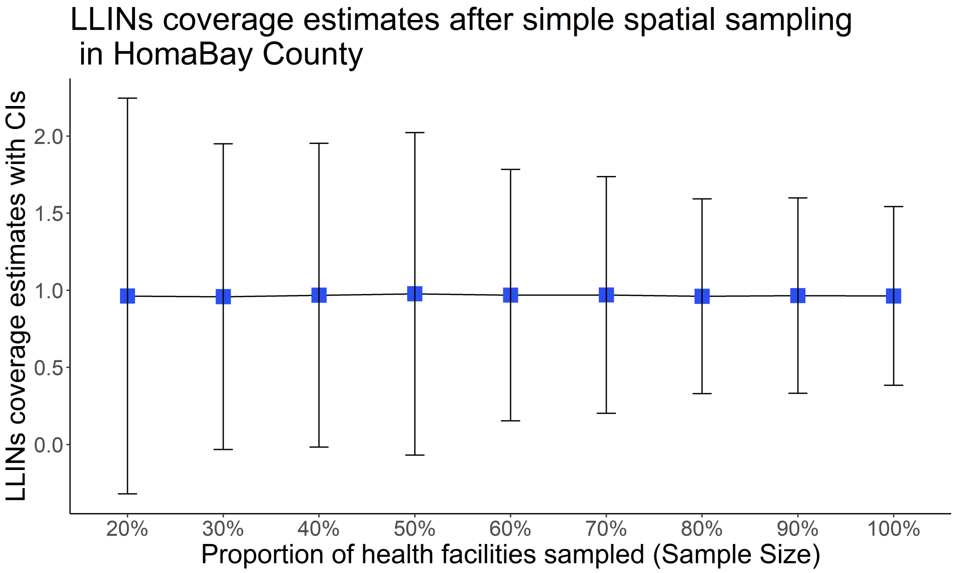

Figure 12. Estimates with CIs of the proportion of pregnant mothers who receive LLINs versus sample size after stratified sampling in HomaBay County

A similar trend was observed for LLINs coverage estimates as well. A plot of the estimate with their respective CIs reveals that they become wider as the sample size reduces. However, they are overlapping which indicates that there is no significant difference between the population estimate and subsample estimates (Figure 12).

5 Conclusion

The purpose of this study was to determine the effect of subsampling on the monitoring indicators using routine health data. This study used a series of experiments to investigate and derive a novel approach to the estimation of indicators from routine health facility data. Through the tests, we found that there was no significant difference between the population estimate and the subsample estimates for all the indicators, across the different sampling designs (p-values > 0.05).

Conventionally, individuals who have used routine data use all the available data from all the health facilities in a given spatial unit, as indicated in the literature review. The findings of this study, however, suggest that a proportion of routine data would be enough to obtain credible estimates that represent the actual picture that would be provided by the entire dataset.

The study found that although the estimates were similar at different sample sizes, the confidence intervals became wider as the sample size reduced. This is expected because the width of the confidence intervals is inversely related to the sample size (Du Prel et al., 2009). The confidence intervals of the estimate were however narrower at 60% and above sample size. This study recommends that a sample size of 60% or higher is used in practice to control the uncertainty of the estimates of routine health indicators. The power calculations also support these recommendations. The study found that all the subsamples across the different sampling designs had a statistical power above 70%. However, the statistical power reduced as the sample size reduced.

The experiments were done at two spatial units to generalize the findings, that is, Western Kenya and HomaBay County. The study found that subsampling can be done at both regional and County levels. In terms of the sampling design, both spatial and non-spatial sampling gave the same results. This study recommends that one can use any of the two sampling design, but care should be taken to account for the spatial variability of the indicator of interest. If the indicator exhibits any spatial variation, one would be better off using spatial sampling (Delmelle, 2009; Stevens Jr & Olsen, 2004).

5.1 Future Research

There is need to establish whether these results can be applied to health facility surveys for sample size determination.

There is also need to determine the effect of subsampling on incompleteness reporting and inconsistencies such as presence of outliers in the DHIS2 data.

5.2 Limitations of the study

Subsampling is not suitable when one is interested in estimation of data elements directly, such as the total number of children immunised in a given region. Its more appropriate for indicators that have numerators and denominators.

Bibliography

Alegana, V. A., Khazenzi, C., Akech, S. O., & Snow, R. W. (2020). Estimating hospital catchments from in-patient admission records: a spatial statistical approach applied to malaria. *Scientific reports*, *10*(1), 1–11.

Alegana, V. A., Okiro, E. A., & Snow, R. W. (2020). Routine data for malaria morbidity estimation in africa: challenges and prospects. *BMC Medicine*, *18*(1), 1–13.

Bhattacharya, A. A., Umar, N., Audu, A., Felix, H., Allen, E., Schellenberg, J. R., & Marchant, T. (2019). Quality of routine facility data for monitoring priority maternal and newborn indicators in dhis2: A case study from gombe state, nigeria. *PLoS One*, *14*(1), e0211265.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Cutts, F. T., Izurieta, H. S., & Rhoda, D. A. (2013). Measuring coverage in mnch: design, implementation, and interpretation challenges associated with tracking vaccination coverage using household surveys. *PLoS Med*, *10*(5), e1001404.

Delmelle, E. (2009). Spatial sampling. *The SAGE handbook of spatial analysis*, *183*, 206.

Desai, M., Phillips-Howard, P. A., Odhiambo, F. O., Katana, A., Ouma, P., Hamel, M. J., … others (2013). An analysis of pregnancy-related mortality in the kemri/cdc health and demographic surveillance system in western kenya. *PloS one*, *8*(7), e68733.

Du Prel, J.-B., Hommel, G., Röhrig, B., & Blettner, M. (2009). Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, *106*(19), 335.

Farnham, A., Utzinger, J., Kulinkina, A. V., & Winkler, M. S. (2020). Using district health information to monitor sustainable development. *Bulletin of the World Health Organization*, *98*(1), 69.

Githinji, S., Oyando, R., Malinga, J., Ejersa, W., Soti, D., Rono, J., … Noor, A. M. (2017). Completeness of malaria indicator data reporting via the district health information software 2 in kenya, 2011–2015. *Malaria journal*, *16*(1), 344.

Hancioglu, A., & Arnold, F. (2013). Measuring coverage in mnch: tracking progress in health for women and children using dhs and mics household surveys. *PLoS Med*, *10*(5), e1001391.

Hemkens, L. G., Contopoulos-Ioannidis, D. G., & Ioannidis, J. P. (2016). Routinely collected data and comparative effectiveness evidence: promises and limitations. *CMAJ*, *188*(8), E158–E164.

Jordans, M. J., Chisholm, D., Semrau, M., Upadhaya, N., Abdulmalik, J., Ahuja, S., … others (2016). Indicators for routine monitoring of effective mental healthcare coverage in low-and middle-income settings: a delphi study. *Health policy and planning*, *31*(8), 1100–1106.

Kane, R., Wellings, K., Free, C., & Goodrich, J. (2000). Uses of routine data sets in the evaluation of health promotion interventions: opportunities and limitations. *Health Education*.

KDHS, K. (2014). *Kenya demographic health survey (2014 kdhs).*

Khan, S. M., Bain, R. E., Lunze, K., Unalan, T., Beshanski-Pedersen, B., Slaymaker, T., … Hancioglu, A. (2017). Optimizing household survey methods to monitor the sustainable development goals targets 6.1 and 6.2 on drinking water, sanitation and hygiene: A mixed-methods field-test in belize. *PloS one*, *12*(12).

Kiberu, V. M., Matovu, J. K., Makumbi, F., Kyozira, C., Mukooyo, E., & Wanyenze, R. K. (2014). Strengthening district-based health reporting through the district health management information software system: the ugandan experience. *BMC medical informatics and decision making*, *14*(1), 1–9.

King, L. J. (1969). *Statistical analysis in geography.* Prentice-Hall Englewood Cliffs, NJ.

KNBS. (2015). Kenya demographic and health survey 2014 kenya national bureau of statistics.

Macharia, P. M., Giorgi, E., Noor, A. M., Waqo, E., Kiptui, R., Okiro, E. A., & Snow, R. W. (2018). Spatio-temporal analysis of plasmodium falciparum prevalence to understand the past and chart the future of malaria control in kenya. *Malaria journal*, *17*(1), 340.

Maïga, A., Jiwani, S. S., Mutua, M. K., Porth, T. A., Taylor, C. M., Asiki, G., … others (2019). Generating statistics from health facility data: the state of routine health information systems in eastern and southern africa. *BMJ global health*, *4*(5), e001849.

Maina, I., Wanjala, P., Soti, D., Kipruto, H., Droti, B., & Boerma, T. (2017). Using health-facility data to assess subnational coverage of maternal and child health indicators, kenya. *Bulletin of the World Health Organization*, *95*(10), 683.

Maina, J., Ouma, P. O., Macharia, P. M., Alegana, V. A., Mitto, B., Fall, I. S., … Okiro, E. A. (2019). A spatial database of health facilities managed by the public health sector in sub saharan africa. *Scientific data*, *6*(1), 1–8.

Manya, A., Braa, J., Overland, L. H., Titlestad, O. H., Mumo, J., & Nzioka, C. (2012). National roll out of district health information software (dhis 2) in kenya, 2011–central server and cloud based infrastructure. In *Ist-africa 2012 conference proceedings* (Vol. 5).

Mark, D. M. (1990). Neighbor-based properties of some orderings of two-dimensional space. *Geographical Analysis*, *22*(2), 145–157.

Nabyonga-Orem, J. (2017). Monitoring sustainable development goal 3: how ready are the health information systems in low-income and middle-income countries? *BMJ global health*, *2*(4), e000433.

Ndabarora, E., Chipps, J. A., & Uys, L. (2014). Systematic review of health data quality management and best practices at community and district levels in lmic. *Information Development*, *30*(2), 103–120.

Ripley, B. D. (1981). *Spatial statistics* (Vol. 575). John Wiley & Sons.

Stevens Jr, D. L., & Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American statistical Association*, *99*(465), 262–278.

The Kenya Gazette, L. (1925). The kenya gazette. *The Land*, *CXXII - No 24.*

UN General Assembly, U. (2015). Transforming our world: the 2030 agenda for sustainable development. *Division for Sustainable Development Goals: New York, NY, USA.*

Wagenaar, B. H., Sherr, K., Fernandes, Q., & Wagenaar, A. C. (2016). Using routine health information systems for well-designed health evaluations in low-and middle-income countries. *Health policy and planning*, *31*(1), 129–135.