Comparison of complete and spatial sampling frames for estimation of the prevalence of hypertension and diabetes mellitus

Vasna Joshua, Kamaraj Pattabi, Yuvaraj Jeyaraman, Prabhdeep Kaur, Tarun Bhatnagar, Suresh Arunachalam, Sabarinathan Ramasamy, Venkateshprabhu Janagaraj, Manoj V. Murhekar

National Institute of Epidemiology (ICMR), Ayapakkam, Chennai, Tamil Nadu, India

Abstract

A complete sampling frame (CSF) is needed for the development of probability sampling structures; utilisation of a spatial sampling frame (SSF) was the objective of the present study. We used two sampling methods, simple random sampling (SRS) and stratified random sampling (STRS), to compare the prevalence estimates delivered by a CSF to that by a SSF when applied to self-reported hypertension and diabetes mellitus in a semi-urban setting and in a rural one. A CSF based on Geodatabase of all households and all individuals was available for our study that focused on adults aged 18-69 years in the two settings. A single digitized shapefile of solely household regions/structures as SSF was developed using Google Earth and employed for the study. The results from the two sampling frames were similar and not significantly different. All 95%CI calculations contained the prevalence rates of the two medical conditions except for one occasion based on STRS and CSF. The SRS based on CSF showed a minimum 95% CI width for diabetes mellitus, whereas SSF showed a minimum 95% CI width for hypertension. The coefficient of variation exceeded 10.0% on six occasions for CSF but only once for SSF, which was found to be as efficient as CSF.

Correspondence: Vasna Joshua, National Institute of Epidemiology (ICMR), R-127, Second Main Road, Tamil Nadu Housing Board, Ayapakkam, Chennai, Tamil Nadu, India.
Fax: 26820464/26820355
E-mail: vasnajoshua@yahoo.com

Key words: Complete sampling frame; spatial sampling frame; hypertension; diabetes mellitus; Tamil Nadu.

Conflict of interest: The Authors declare no conflict of interest.

Acknowledgments: The authors thank all staff, including field staff at Kallur and Ayapakkam, and all study participants, for their contributions.

Consents: Written informed consent was obtained from all participants after explaining the purpose of the study and data collection procedures.

Received for publication: 11 April 2022.
Revision received: 21 September 2022.
Accepted for publication: 21 September 2022.

Geospatial Health 2022; 17:1097
doi:10.4081/gh.2022.1097

Introduction

Sampling is a statistical technique used to make inferences about study populations based on a sample of observations or individuals drawn from a total population. It is essential to choose a representative sample to get a reliable population estimate. Several probability sampling techniques are used to construct the sample: simple random sampling (SRS), stratified random sampling (STRS), etc. The characteristics studied could be a proportion, a mean value (Christakos, 2005; Haining, 2003) or an attribute based on the location (longitude and latitude) of the target population (Rogerson *et al.*, 2004).

Large-scale national surveys, such as the National Family Health Survey (NFHS) and Demographic and Health Surveys (DHS), use samples based on complete sampling frames (CSF) that are time-consuming, requires resources and are expensive. Sampling methods, such as SRS and STRS, are essential for achieving adequate probability and therefore usefulness of the sampling frames applied must be recent or periodically updated and available in a usable format to map the target population accurately (Thomas *et al.*, 2019). Household (HH) data are more difficult to obtain in rural areas than in urban or developed areas since they are scattered or irregularly spaced and sometimes lacking. Hence, estimations based on survey sample designs may have biases sometimes providing misleading results (Eric, 2008; Haining, 2001).

The spatial sampling frame (SSF) is helpful when it is difficult to sample every target population or get a reliable sampling frame. It is widely used in environmental, ecological, mining, geology, and hydrology studies (Cressie, 1991; Haining, 2003; Muller, 1998; Ripley, 1981; Stehman, 1996). Many studies have developed and used a SSFs based on geographical information systems (GIS), the global positioning system (GPS) and satellite imagery for data collection. The use of SSFs is gaining importance as an alternative approach when the population under study is unavailable or unknown, or a complete sampling frame is unavailable.

Various studies based on sampling frames have been carried out. For example, a study among very high-density slum dwellers in Delhi, India (Abhijit *et al.*, 2014), where high-resolution satellite imagery was used to identify housing density and assess slum-dwellers concerning the quality of public services; a sampling frame for a survey based on geo-located dwelling locations and satellite data was developed and applied in nine countries (Improving Health in Slums Collaborative, 2019); a study focussing on urban poor women in six cities in Uttar Pradesh, India employed GIS to construct the primary sampling unit (PSU) of residential HHs eligible for interviews (Speizer *et al.*, 2012); and in Lilongwe, Malawi every household structure within the catchment area was digitized and assigned geographic coordinates using Google Earth satellite imagery (Escamilla *et al.*, 2014). In the latter case, a sampling frame of a list of HHs and a random sample was generated to study the intensity of *Plasmodium falciparum* transmission and the authors concluded that their approach to developing a sampling frame was accurate and that it could have utility beyond morbidity studies.

The availability of a CSF is challenging, time-consuming and expensive, so researchers need to opt for an appropriate sampling frame to capture precise and representative estimates (Haining, 2001; Swacha *et al.*, 2017). Using a SSF appears to be a cost-effective option in resource-poor settings. The use of satellite images and GIS to sample structures made it feasible to rapidly select a representative population sample at a low-cost for a prevalence survey in a rural Guatemalan village (Miller *et al.*, 2020). A system of unique household identifiers devised for household enumeration made re-identification possible in a densely inhabited slum of Maharashtra, India (Thomson *et al*., 2014).

Comparison of sampling methods using CSF and SSF is rarely attempted; moreover, the outcome variable is binary, while the spatial data are in vector format. Hence our objective was to estimate and compare CSF and SSF by two probability sampling methods, SRS and STRS, in a semi-urban and rural study setting. We applied these sampling frames in cross-sectional study design in a semi-urban and a rural area to estimate the prevalence of two self-reported disease conditions: hypertension (HT) and diabetes mellitus (DM). Overall, our ultimate aim of using these two sampling methods was to estimate the population parameters with respect to feasibility, efficiency and generalizability.

Materials and Methods

The master file of the list of all HHs geocoded with the list of all individuals as a CSF with factors like gender, age, presence or absence of disease conditions for the two study settings (semi-urban and rural) are readily available. Ayapakkam, Chennai, is a semi-urban setting, while Kallur Village, Tirunelveli is rural. Despite the known geographic coordinates of the HHs, a single digitized shapefile showing only HH structures as SSF was developed independently using Google Earth. Two settings, developed under the umbrella of the Indian Council of Medical Research, were chosen for this study.

Study settings

Semi-urban

A HH-level cohort for the study of demographic and health surveillance has been developed by the National Institute of Epidemiology (ICMR-NIE) in Ayapakkam, Chennai to document various demographic and health indicators on a longitudinal basis. The study population was surveyed from April 2015 to July 2019 noting its socio-economic profile, morbidity profile and health-seeking behaviour as well as births, deaths, migration and other vital events. All HHs and the key landmarks were mapped using GPS. There were 10,927 HHs with 50,249 individuals in the study area (Figure 1) and the proportions of self-reported HT and DM in Ayapakkam were 6.4% and 7.4%, respectively.

Rural

The Department of Health Research's flagship program has led to the establishment, under a local medical college leadership, of the Model Rural Health Research Unit (MRHRU) in Tirunelveli, which is mentored by the ICMR-NIE (Joshua *et al.*, 2020). One of the key objectives of the MRHRU is to develop area-specific models depending on disease profiles, topography and morbidity patterns to provide better health care services by undertaking relevant research on local health issues identified by state and local health authorities. As a first round, a demographic health database of all 11,006 HHs with 36,289 individuals was developed for Kallur Village, a rural area in Tirunelveli (Figure 2). The study was done between Jan 2016 and Aug 2019 using GPS and digitizing all HHs and key landmarks in this study area. Here, the proportions of self-reported HT and DM were 5.4% and 4.7%, respectively.

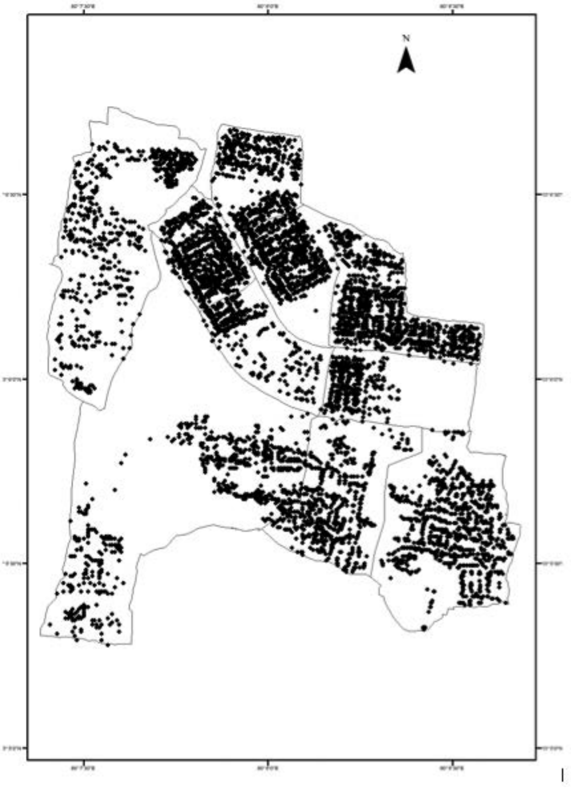


Figure 1. The digitized study area and households of Ayapakkam, Chennai in 2019.

Study population

As per WHO guidelines in estimating the DM and HT prevalence (WHO STEPS Surveillance Manual, 2017), all persons aged 18-69 years were included in both study settings. The population details are presented in Figure 3.

Study design

Probability sampling using CSF

A Geodatabase of all HHs and a listing of all individuals were adopted as CSF were readily available for both study sites. To estimate the prevalence of self-reported HT and DM the two probability sampling methods (SRS and STRS) were applied based on individuals and also on gender.

SRS selection without replacement was used to select the required sample size from the CSF of adults aged 18-69 years that had been stratified by gender for both sites (17,466 males and 18,241 females in Ayapakkam and 10,716 males and 13,017 females in Kallur Village). The required sample sizes resulted in 35,707 individuals for Ayapakkam and 23,733 for Kallur Village.

Spatial sampling using SSF

The shapefile consisted of solely residential regions of HHs, digitized by Google Earth Pro 7.3.2 to generate the SSF. We surveyed the area with a Garmin Handheld Trimble Juno SC and captured each HH location with values rounded to six decimals. Utmost care was taken so that all HHs would lie within the generated shapefile in all possible forms. The result was a single, non-continuous polygon, irregular in shape and size consisting of residential areas excluding non-residential regions (vacant land, parks, watery bodies, commercial complexes, etc.). Figure 4A illustrates this with respect to the digitized shapefile of Ayapakkam super

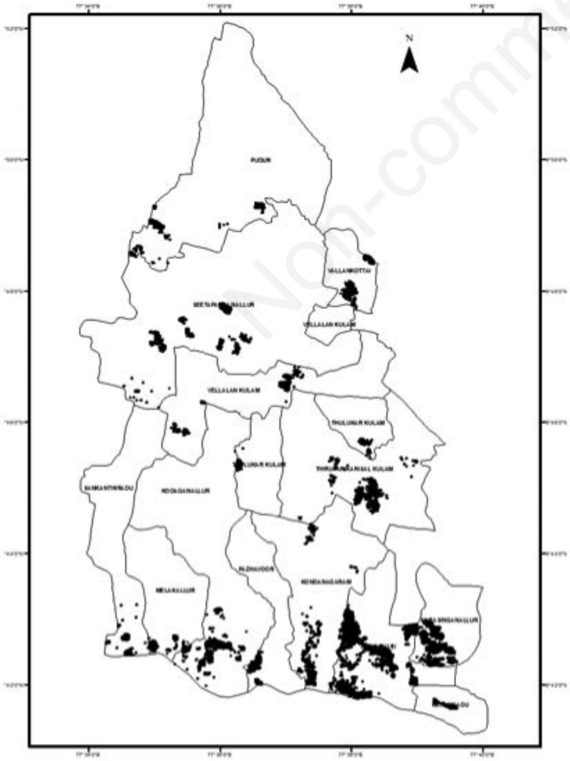


Figure 2. The digitized study area and households of Kallur Village, Tirunelveli in 2019.

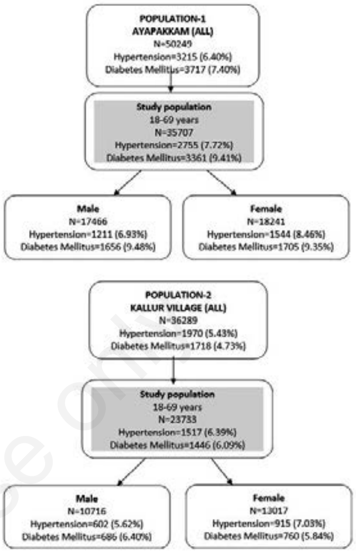

POPULATION-1
AYAPAKKAM (ALL)
N=50249
Hypertension=3215 (6.40%)
Diabetes Mellitus=3717 (7.40%)
Study population
18-69 years
N=35707
Hypertension=2755 (7.72%)
Diabetes Mellitus=3361 (9.41%)
Male
N=17466
Hypertension=1211 (6.93%)
Diabetes Mellitus=1656 (9.48%)
Female
N=18241
Hypertension=1544 (8.46%)
Diabetes Mellitus=1705 (9.35%)
POPULATION-2
KALLUR VILLAGE (ALL)
N=36289
Hypertension=1970 (5.43%)
Diabetes Mellitus=1718 (4.73%)
Study population
18-69 years
N=23733
Hypertension=1517 (6.39%)
Diabetes Mellitus=1446 (6.09%)
Male
N=10716
Hypertension=602 (5.62%)
Diabetes Mellitus=686 (6.40%)
Female
N=13017
Hypertension=915 (7.03%)
Diabetes Mellitus=760 (5.84%)

Figure 3. Population details of Ayapakkam and Kallur (2015-19).

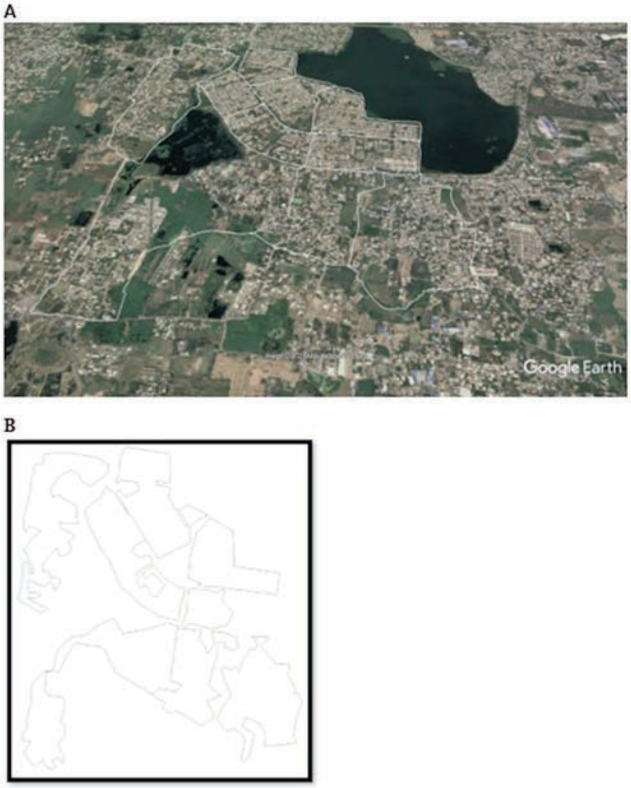


Figure 4. A) Ayapakkam study area with Google Earth view in 2019. B) Using Google Earth, a single shapefile of Ayapakkam with solely digitized household regions (8 polygons).

imposed on Google Earth imagery. Figure 4B shows a single shapefile (union of 8 irregular polygons) of solely HH regions in the study area. In the semi-urban area of Ayapakkam, houses were much closer and had mainly similar structures. Around 15% of the population there lives in multi-storeyed buildings or multiple living units. Around two-thirds of each HH was evenly distributed.

We adopted the same GPS procedure for SSF in Kallur Village, which is shown in Figure 5A. The polygon generated was also irregular and manifold (156 irregular polygons) and combined to form one shapefile (Figure 5B). The HHs in this rural area consisted of small farmhouses, sometimes together in livestock-keeping communities with huts closer to the river. There were occasions when one side of the street belonged to the Kallur study area, while the other side did not and where a few successive houses in the street belonged to the study area, and the other half to a different hamlet. Hence, utmost care was taken to digitize the residential regions, as the range covered polygons varying from only three houses up to one containing nearly 1,000 houses. Probability sampling based on the SRS and STRS methods as used before was adopted.

Spatial SRS

The single shapefile was fed into the R program. The required (unique) random sample of points was generated as latitude and longitude data. The random numbers were all unique geographic locations of the residential region. If the selected coordinates did not fall precisely on a HH or did not contain an individual in the 18-69 years age group, the closest neighbourhood HH was selected (it occurred in less than 5% of the instances in both settings). For points far away from an HH structure, the closest HH east of the point was chosen [for Ayapakkam within 250 feet (76 m) and for Kallur within 600 feet (183 m)]. This exercise was repeated independently for both diseases. The required numbers of HHs were selected first from the SSF of HH and then an individual aged 18-69 years was randomly selected from the selected HH.

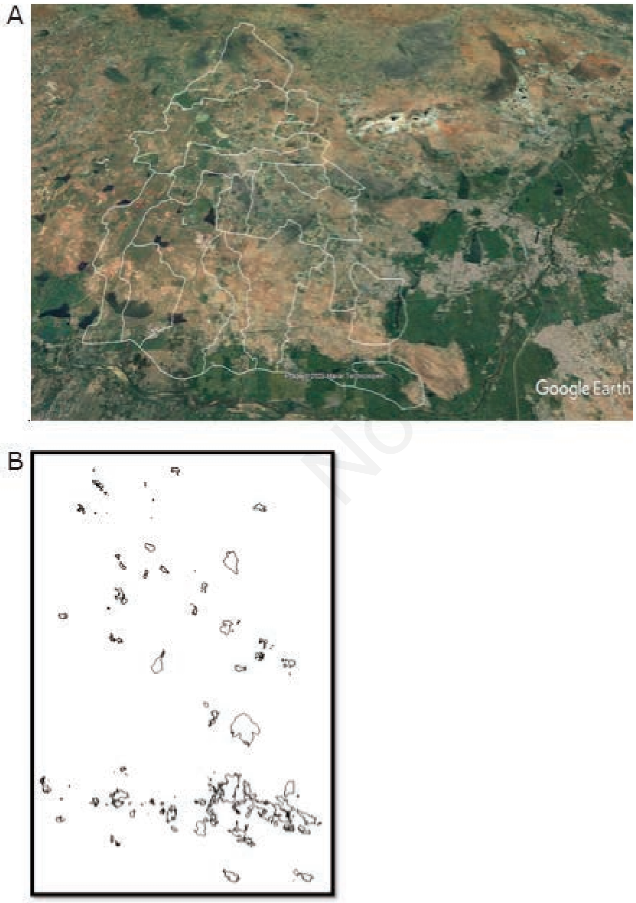


Figure 5. A) Kallur Village area with Google Earth view in 2019; B) A single shapefile of Kallur village with solely digitized household regions (156 polygons) in 2019.

Spatial STRS

As mentioned earlier, this approach was based on a polygon shapefile as sampling frame, assuming that genders in the HH were independent. Hence, two independent SSFs of HHs were generated, one for the males and one for the females. When these two separate layers were overlaid, the entire population was comprised. This exercise was independently repeated for both disease conditions. We chose the required HHs from the SSF of the male stratum and the required sample of HHs from the SSF of the female stratum. For each stratum (male and female) we chose an individual aged 18-69 years at random from the selected HHs.

Sample size in probability sampling

SRS sample size

The total number of individuals aged 18-69 years was taken as the size of the target population. The sample size was calculated using the formula:

$$n = \left[\frac{\text{DEFF}\times \text{N}\times \text{prop}\times(1-\text{prop})}{\frac{d^2}{Z^2_{1-\frac{\alpha}{2}}}\times(\text{N}-1)+\text{prop}\times(1-\text{prop})} \right] \qquad (1)$$

where DEFF is the design effect (1.0 for SRS); N the size of the population; prop the proportion of people with disease; d the limit of accuracy required, defined in terms of a percentage of the estimate (*e.g.*, 10% of prop); and Zα the confidence level factor usually taken to be 1.96, corresponding to 95% Cl; hence $Z_{1-\frac{\alpha}{2}}$ is 1.96 for a two-tail test where α is taken as 5%.

STRS sample size

The sample size formula for this approach (Cochran, 1977) is given by the formula:

$$n = \frac{Z_\alpha^2 \sum_{h=1}^{k} W_h S_h^2}{d^2 + \frac{Z_\alpha^2 \sum_{h=1}^{k} W_h S_h^2}{N}} \qquad (2)$$

where n is the required sample size; h the stratum number; N(h) the population size of h^{th} stratum; prop(h) the proportion of outcome in each stratum; and d the limit of accuracy.

$$N = \sum_{h=1}^{k} N(h);\ W_h = N(h)/N;\ S_h^2 = \text{prop}(h)\ X\ (1-\text{prop}(h)).$$

Sample sizes

For comparison purposes, the same sample size calculated for probability sampling was used for both spatial sampling methods

(CSF and SSF). The following sizes were used: i) For a self-reported HT prevalence in Ayapakkam by SRS of 7.7% among the CSF of 35,707 individuals aged 18-69 years and a relative precision of 20% (i.e. ± 1.5% at 95%CI), the required sample size would be 1,174 individuals. For one by STRS based on gender of 6.9% among the CSF of 17,466 males (8.5% among the CSF of 18,241 females), the required total sample size would be 1,176 with a proportional allocation of 575 males and 601 females; ii) For a self-reported DM prevalence in Ayapakkam by SRS of 9.4% among the CSF of 35,707 individuals aged 18-69 years and a relative precision of 20% (i.e. ± 1.9% at 95%CI), the required sample size would be 884 individuals. For one by STRS based on gender of 9.5% among the CSF of 17,466 males (9.3% among the CSF of 18,241 females), the required total sample size would be 885 with proportional allocation of 433 males and 452 females; iii) For a self-reported HT prevalence in Kallur by SRS of 6.4% among the CSF of 23,733 individuals aged 18-69 years and a relative precision of 20% (i.e. ± 1.3% at 95%CI), the required sample size would be 1,288 individuals. For one by STRS based on gender of 5.6% among the CSF of 10,716 males (7.0% among the CSF of 13,017 females), the required total sample size would be 1,282 with a proportional allocation of 579 males and 703 females; iv) For a self-reported DM prevalence in Kallur by SRS of 6.1% among the CSF of 23,733 individuals aged 18-69 years and a relative precision of 20% (i.e. ± 1.2% at 95%CI), the required sample size would be 1,436 individuals. For one by STRS based on gender of 6.4% among the CSF of 10,716 males (5.8% among the CSF of 13,017 females), the required total sample size would be 1,431 with a proportional allocation of 646 males and 785 females.

Data collection

The ICMR has developed two Geodatabases on population health, one for Ayapakkam, Chennai and one for Kallur Village, Tirunelveli. We used these Geodatabases and generated 100 random samples for our present study comparing CSF and SSF without distorting the order of the dataset. Identical, pretested, structured questionnaires produced in English and Tamil, the local language, were used in both areas. Using personal digital assistants (PDAs) in the form of tablet computers, various sections on household characteristics, socio-economic profile, morbidity, and treatment-seeking behaviour, were collected in both areas.

Point and interval estimates

Our objective was to estimate the prevalence (prop) (in the form of a binomial proportion) at 95%CI from samples based on the above two sampling methods and two sampling frames. For a single sample drawn through sampling methods using CSF and SSF, the point estimate and 95%CI are given as (prop±1.96×SE(prop)) where prop is the proportion and SE(prop) the standard error of the proportion were calculated as per the appropriate probability sampling method. The same estimation procedure was adopted to get the proportion (prop) and 95%CI(prop) even for SSF.
95%CI(prop) = 95%CI of the binomial proportion = *prop±1.96× SE(prop)*

$$= prop \pm 1.96\times\sqrt{\frac{prop\ X\ (1-prop)}{n}} \qquad (3)$$

where n is the sample size; prop the binomial proportion, *i.e.* the prevalence estimate from a single sample; (1 – prop) is 1 – the prevalence estimate from a single sample; and SD(prop) the standard deviation of the prevalence estimate from a single sample $=\sqrt{prop\ X\ (1-prop)}$, whereas SE(prop) is the SE of prevalence estimate from a single sample is SD divided by $\sqrt{n}$.

$$\left(i.e.\, SE(prop) = \frac{SD(prop)}{\sqrt{n}} = \frac{\sqrt{prop\ X\ (1-prop)}}{\sqrt{n}} = \sqrt{\frac{prop\ X\ (1-prop)}{n}}\right) \qquad (4)$$

Since there is a risk that the estimates may not be similar or consistent if the SRS method were adopted only once from CSF and only once from the SSF, we generated 100 independent samples using both frames. All the samples drawn in the study were without replacement. Therefore, each sample unit drawn from the study population had only one chance to be selected. We generated 100 independent samples with 100 different random numbers and the corresponding 100 independent estimates for both sampling methods were calculated. Then we combined all the 100 independent sample estimates to get a 95%CI for the population parameter and then compared the two sampling frames.

Combined proportion method– CSF

By generating 100 independent samples of the same sample size n, we arrived at 100 prop values and 100 (1-prop) values where $prop_i$ stands for the prevalence estimate from the i^{th} independent sample where 1-$prop_i$ stands for the 1-prevalence estimate from the i^{th} independent sample. By taking the average of the 100 $prop_i$'s, we got the combined proportion ($\overline{prop}$) and correspondingly from the average of 100 (1-$prop_i$) values, the combined proportion of non-disease ($\overline{1\text{–}prop}$) (and the equation below gives the SE of $\overline{prop}$:

$$SE\ (\overline{prop}) = \sqrt{\frac{\overline{prop}\times(\overline{1-prop})}{n}} \qquad (5)$$

where the calculated 95%CI of $\overline{prop}$: (the average of the 100 $prop_i$ values) =

$\overline{prop} \pm 1.96\sqrt{\frac{\overline{prop}\times(\overline{1-prop})}{n}}$ and n the same sample size for all the 100 independent samples.

Sampling distribution method - CSF

By treating each of the 100 $prop_i$ values (the sample estimates) as a random variable x_i, we got the average of x_i as $\overline{x}$ and standard deviation (SD) of x_i as SD(x), which is also the SE of x by sampling distribution principle. We then used ($\overline{x}$ ±1.96 SE(x)) as 95%CI. The mean of the 100 independent $prop_i$ values is $\overline{x}$ and the SD can be calculated by the usual formula for mean & standard deviation and 95%CI is $\overline{x}$ ±1.96×SD(x).

Combining the proportion - SSF

By treating each of the 100 $prop_i$ values (the sample estimates) from SSF as 100 independent x_i estimates, we got an average of x_i as $\overline{x}$ and SD of x_i as SD(x). We can then use $\overline{x}$ ±1.96×SD(x)) as 95%CI.

Data analysis

The results are presented in the form of proportions expressed in percentages. The prevalence of self-reported HT and DM among adults aged 18-69 years in the two study areas was estimated based on CSF and SSF. We used STATA SE, version 16·0 software (Stata Corp LLC, Texas, USA) to estimate the proportion of HT and DM for the 100 samples of the two probability sampling methods and for the two sampling frames. Prevalence of self-reported HT and self-reported DM as estimates, along with 95%CI, was calculated. We used four metrics to compare CSF and SSF: i) The absolute difference between population prevalence and sample prevalence (Abs diff); ii) The width of the 95% confidence interval (95%CI width); iii) The percentage coefficient of variation (CV %), which is determined based on the 100 estimates as (SD/$\bar{x}$) *100, where $\bar{x}$ is the mean of 100 prop$_i$ values and SD is the standard deviation of 100 prop$_i$ values; iv) The significance of the Z test between sample prevalence and population prevalence.

The Z test for single sample proportion compared to population proportion when n≥30 can be calculated as follows:

$$z = \frac{sample\ prop - population\ prop}{SE(sample\ prop)} = \frac{sample\ prop - population\ prop}{\sqrt{\frac{\text{sample prop X (1 - sample prop)}}{n}}} \qquad (6)$$

where n is the sample size; prop is the binomial proportion or prevalence estimate from a single sample; 1 – prop is 1 – prevalence estimate from a single sample; SE(prop) the standard error of prevalence estimate from a single sample is $\sqrt{\frac{prop\ X\ (1 - prop)}{n}}$; 'population prop' the prevalence in the population; and p the level of significance (=0.05 for a two-sided test). Z follows a normal distribution with the reference table value as 1.96 and the corresponding significant level p-value.

In the SSF we used vector datasets, the HH locations as point data and the medical conditions sought as binaries. In our semi-urban study setting, the spatial autocorrelation using local joint statistics for the HT distribution (99.0% of the points fall as insignificant) and DM (99.3% of the points fall as insignificant). In the rural setting, the outcomes were similar, i.e. 99.6% and 99.5% for HT and DM, respectively. Hence there is not enough evidence to say that the medical condition in question followed any pattern but occurs at random and may be due to non-communicable disease conditions. As a result, the estimation did not take into account the spatial autocorrelation factor.

All statistical analyses were two-tailed, and p<0.05 was considered statistically significant. SRS sample size calculation was done using OpenEpi version 3.01 software. For STRS, a Microsoft Excel spreadsheet was used to calculate the sample size. The statistical analyses were done using STATA SE and ArcGIS Desktop, version 10 (ESRI, Redlands, CA, USA); Geoda 1.18.0 (https://geodacenter.github.io/) and R 4.0.2 (https://cran.r-project.org/bin/windows/base/) software were used for the spatial analyses. Figure 6 shows the workflow of the study.

Results

HT in semi-urban setting (Ayapakkam)

The prevalence of HT by the two sampling methods (SRS and STRS) based on CSF ranged from 7.32% to 8.59%, whereas it ranged from 7.24% to 8.35% when based on SSF. All the 95%CIs based on both CSF and SSF contain the population prevalence (7.72%) (Table 1). All sample estimates based on CSF and SSF were not statistically significantly different from the population prevalence (7.72%) (Table 2). The minimum difference (0.01) was found in the combined proportion of 100 STRS based on CSF, whereas the maximum difference (0.87) was found in one single STRS based on CSF (Table 2). Considering the width of the

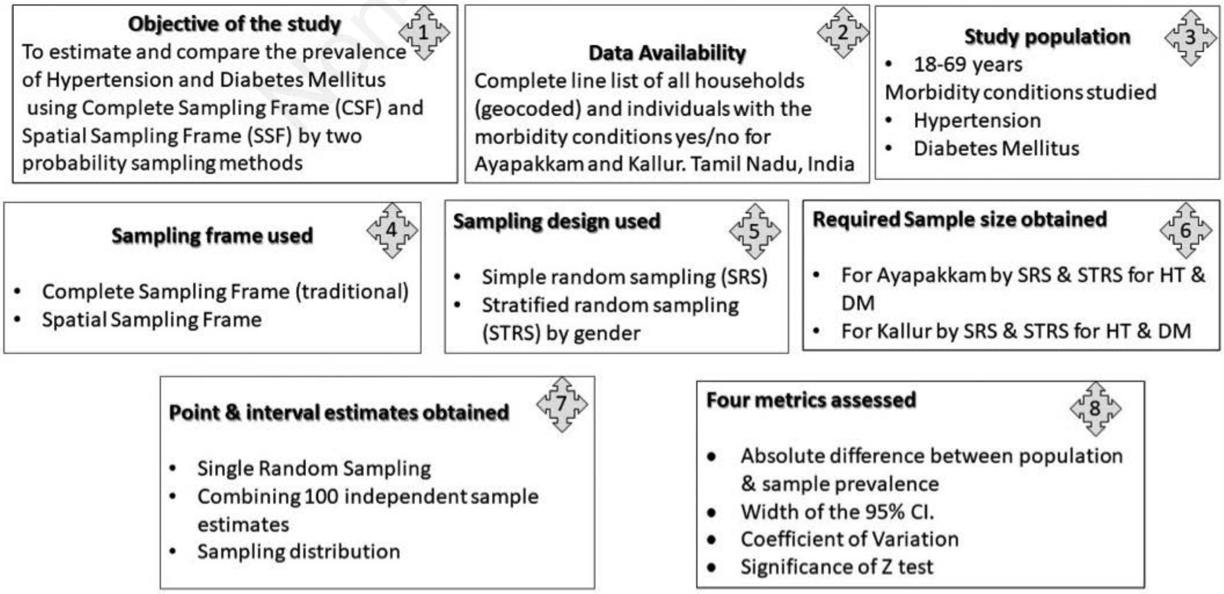

Objective of the study
To estimate and compare the prevalence of Hypertension and Diabetes Mellitus using Complete Sampling Frame (CSF) and Spatial Sampling Frame (SSF) by two probability sampling methods
1
Data Availability
Complete line list of all households (geocoded) and individuals with the morbidity conditions yes/no for Ayapakkam and Kallur. Tamil Nadu, India
2
Study population
• 18-69 years
Morbidity conditions studied
• Hypertension
• Diabetes Mellitus
3
Sampling frame used
• Complete Sampling Frame (traditional)
• Spatial Sampling Frame
4
Sampling design used
• Simple random sampling (SRS)
• Stratified random sampling (STRS) by gender
5
Required Sample size obtained
• For Ayapakkam by SRS & STRS for HT & DM
• For Kallur by SRS & STRS for HT & DM
6
Point & interval estimates obtained
• Single Random Sampling
• Combining 100 independent sample estimates
• Sampling distribution
7
Four metrics assessed
• Absolute difference between population & sample prevalence
• Width of the 95% CI.
• Coefficient of Variation
• Significance of Z test
8

Figure 6. Flow chart of the workflow of the study.

95%CI, the minimum width (2.67) occurred when using the spatial SRS with 100 replications based on SSF, while the maximum width (3.29) was from one single STRS based on CSF (Table 2). The CV ranged between 8.3% (spatial SRS sampling with 100 replications based on SSF) and 9.7% (STRS with 100 replications based on CSF) and all the CV values were <10.0%, which is clearly acceptable (Table 2).

DM in semi-urban setting (Ayapakkam)

The prevalence of DM by the two probability sampling methods based on CSF ranged from 9.24% to 10.86%, whereas it ranged from 9.94% to 10.10% when based on SSF. All the 95%CI contained the population prevalence (9.41%) in both sampling frames (Table 1). The sample estimates were not statistically significantly different from the population prevalence (9.41%) in any of the sampling frames (Table 2). The absolute minimum difference (0.03) was found in one single STRS based on CSF, whereas the absolute maximum difference (1.45) occurred in one single SRS on CSF (Table 2). Considering the width of the CI, the minimum width (3.41) was seen in STRS with100 replications based on CSF, and the maximum width (4.21) with single SRS based on CSF (Table 2). The CV ranged between 9.0% (STRS with 100 replications based on CSF) and 11.2% (SRS with 100 replications based on CSF), which was the only value higher than 10.0%; coefficient of variation all the other values were <10.0% (Table 2).

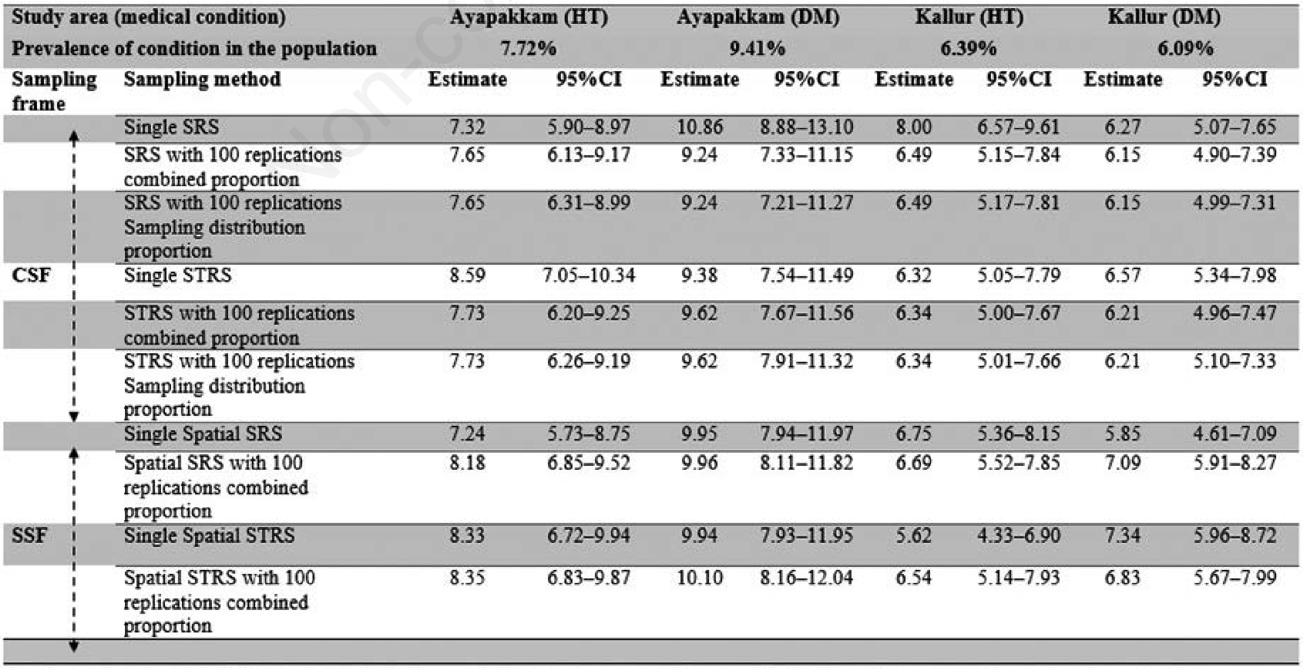
HT in the rural setting (Kallur Village)

The prevalence of HT by the two probability sampling methods using CSF ranged from 6.32% to 8.00%, whereas the use of SSF ranged from 5.62% to 6.75%. All the 95%CI contained the population prevalence (6.39%) except one single SRS based on CSF (Table 1). All sample estimates were not statistically significantly different from the population prevalence (6.39%) except for one single SRS based on CSF (8.00%; Kallur HT) (Table 1). The absolute minimum difference (0.05) was found in the combined proportion of 100 STRS based on CSF, whereas the absolute maximum difference (1.61) was found in a single SRS based on CSF (Table 2). Considering the width of the 95%CI, the minimum width (2.33) was from spatial SRS with 100 replications based on SSF, and the maximum width (3.04) was from one single SRS based on CSF (Table 2). The coefficient of variation ranges between 8.9% (spatial SRS with 100 replications based on SSF) and 10.9% (spatial STRS with 100 replications based on SSF). The majority of the CV values of probability sampling using CSF were >10%, whereas, in spatial sampling using SSF, all the CV values were <10% except one (spatial STRS with 100 replications; Kallur HT) (Table 2).

DM in a rural setting (Kallur Village)

The prevalence of DM by the two probability sampling methods using CSF ranged from 6.15% to 6.57%, whereas SSF ranged from 5.85% to 7.34%. All the 95%CI contained the population prevalence (6.09%) in both sampling frames (Table 1). All sample estimates were not statistically significantly different from the population prevalence (6.09%) in both sampling frames (Table 2). The absolute minimum difference (0.06) was found in the combined proportion of the 100 SRS replications based on CSF, whereas the absolute maximum difference (1.25) was found in one single

Table 1. Estimates of simple and stratified random sampling methods using complete and spatial sampling frames for two medical conditions.



Study area (medical condition)		Ayapakkam (HT)		Ayapakkam (DM)		Kallur (HT)		Kallur (DM)	
Prevalence of condition in the population		7.72%		9.41%		6.39%		6.09%	
Sampling frame	Sampling method	Estimate	95%CI	Estimate	95%CI	Estimate	95%CI	Estimate	95%CI
CSF	Single SRS	7.32	5.90–8.97	10.86	8.88–13.10	8.00	6.57–9.61	6.27	5.07–7.65
	SRS with 100 replications combined proportion	7.65	6.13–9.17	9.24	7.33–11.15	6.49	5.15–7.84	6.15	4.90–7.39
	SRS with 100 replications Sampling distribution proportion	7.65	6.31–8.99	9.24	7.21–11.27	6.49	5.17–7.81	6.15	4.99–7.31
	Single STRS	8.59	7.05–10.34	9.38	7.54–11.49	6.32	5.05–7.79	6.57	5.34–7.98
	STRS with 100 replications combined proportion	7.73	6.20–9.25	9.62	7.67–11.56	6.34	5.00–7.67	6.21	4.96–7.47
	STRS with 100 replications Sampling distribution proportion	7.73	6.26–9.19	9.62	7.91–11.32	6.34	5.01–7.66	6.21	5.10–7.33
SSF	Single Spatial SRS	7.24	5.73–8.75	9.95	7.94–11.97	6.75	5.36–8.15	5.85	4.61–7.09
	Spatial SRS with 100 replications combined proportion	8.18	6.85–9.52	9.96	8.11–11.82	6.69	5.52–7.85	7.09	5.91–8.27
	Single Spatial STRS	8.33	6.72–9.94	9.94	7.93–11.95	5.62	4.33–6.90	7.34	5.96–8.72
	Spatial STRS with 100 replications combined proportion	8.35	6.83–9.87	10.10	8.16–12.04	6.54	5.14–7.93	6.83	5.67–7.99

HT=hypertension; DM=diabetes mellitus; CSF=complete sampling frame; SSF=spatial sampling frame; SRS=simple random sampling; STRS=stratified random sampling. Table 2: Results based on metrics of simple and stratified random sampling methods using complete and spatial sampling frames.

spatial STRS based on SSF (Table 2). Considering the width of the 95%CI, the minimum width (2.24) was from STRS with 100 replications on CSF, and the maximum width (2.76) was from one single spatial STRS based on SSF (Table 2). The coefficient of variation ranged between 8.5% (spatial SRS sampling with 100 replications based on SSF) and 9.6% (SRS with 100 replications based on CSF), and all the values were <10.0%, which is clearly acceptable (Table 2).

Overall, 95%CI of all CSF and SSF contained the population prevalence rates with only one exception that occurred when using the CSF. It may be due to a 5% chance (Table 1). The Z test of significance between the sample estimates and population parameter was significantly different from the population prevalence on only one occasion for CSF (Table 2). In the entire exercise of results presented here, in two instances the minimum CIs noted concerned SSF in the case of HT and CSF in the case of DM. In contrast, CSF showed the maximum width for HT and DM (Table 2). The minimum CV was observed on three occasions in the case of SSF and on one when using CSF in the entire study presented here. Maximum CV was observed on three occasions using CSF and only once using SSF. Apart from that, on six occasions when using CSF, the CV exceeded 10.0%, whereas only once when SSF was used. Hence, SSF is as efficient as CSF. The findings are based on the metrics mentioned above and shown in Table 3.

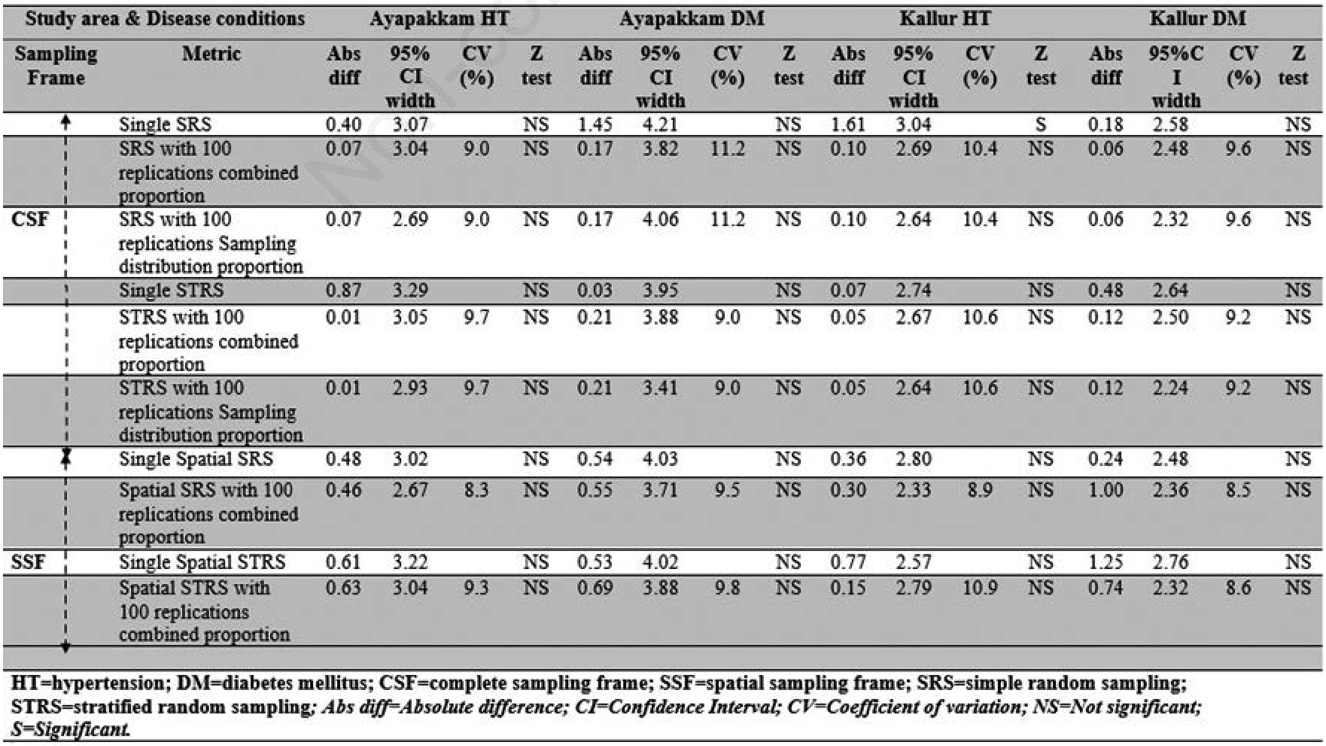
Discussion

Sampling methods lead to quick estimates at reduced costs but suffer from accuracy compared to complete enumeration. SRS is simple to apply, generalizable and still less commonly applied in practice. The construction of a complete sampling frame is a challenge and potentially expensive as it generates lower precision leading to low efficiency due to less representative data (Malhotra and Birks, 2006). In stratified sampling, heterogeneous data are divided into homogenous non-overlapping strata followed by SRS applied to each stratum (EPA, 2002). The heterogeneity in each stratum can be eliminated by splitting the data into homogeneous strata (Cochran, 1977), which makes SRS highly representative and easier to exercise. However, it can sometimes be difficult to identify relevant stratification variables or impossible to stratify all variables, which can turn out to be expensive.

The current study attempted an innovative approach by using a large spatial data repository to compare the prevalence of self-reported HT and DM in a semi-urban and a rural setting using two probability sampling methods. Except for one instance when using CSF, all 95%CI contained the population prevalence and were not significantly different when comparing the two sampling frames. CSF yielded a minimum 95%CI width for DM, whereas SSF yielded a minimum for HT. On six occasions when using the CSF and once when using the SSF, the CV exceeded 10%. Taking together, the results strongly support the notion that SSF is as efficient as CSF.

The error variance of the estimator of the population mean is reduced by stratification, according to a study employing raster data (or satellite imaging data) samples from a spatially autocorrelated homogenous surface (Wang *et al.*, 2010). Establishing a heterogeneous layer in a physical and logical framework in the spatial dimension is typically difficult. Furthermore, stratified sampling consistently exhibits lower sample variances than random sampling in geographical sampling (EPA, 2002). The spatial performance of grid-based models has been evaluated in an earlier study using the Spind software, a statistic for comparing spatial and clas-

Table 2. Results based on metrics of simple and stratified random sampling methods using complete and spatial sampling frames.



Study area & Disease conditions		Ayapakkam HT				Ayapakkam DM				Kallur HT				Kallur DM			
Sampling Frame	Metric	Abs diff	95% CI width	CV (%)	Z test	Abs diff	95% CI width	CV (%)	Z test	Abs diff	95% CI width	CV (%)	Z test	Abs diff	95%CI width	CV (%)	Z test
	Single SRS	0.40	3.07		NS	1.45	4.21		NS	1.61	3.04		S	0.18	2.58		NS
	SRS with 100 replications combined proportion	0.07	3.04	9.0	NS	0.17	3.82	11.2	NS	0.10	2.69	10.4	NS	0.06	2.48	9.6	NS
CSF	SRS with 100 replications Sampling distribution proportion	0.07	2.69	9.0	NS	0.17	4.06	11.2	NS	0.10	2.64	10.4	NS	0.06	2.32	9.6	NS
	Single STRS	0.87	3.29		NS	0.03	3.95		NS	0.07	2.74		NS	0.48	2.64		NS
	STRS with 100 replications combined proportion	0.01	3.05	9.7	NS	0.21	3.88	9.0	NS	0.05	2.67	10.6	NS	0.12	2.50	9.2	NS
	STRS with 100 replications Sampling distribution proportion	0.01	2.93	9.7	NS	0.21	3.41	9.0	NS	0.05	2.64	10.6	NS	0.12	2.24	9.2	NS
	Single Spatial SRS	0.48	3.02		NS	0.54	4.03		NS	0.36	2.80		NS	0.24	2.48		NS
	Spatial SRS with 100 replications combined proportion	0.46	2.67	8.3	NS	0.55	3.71	9.5	NS	0.30	2.33	8.9	NS	1.00	2.36	8.5	NS
SSF	Single Spatial STRS	0.61	3.22		NS	0.53	4.02		NS	0.77	2.57		NS	1.25	2.76		NS
	Spatial STRS with 100 replications combined proportion	0.63	3.04	9.3	NS	0.69	3.88	9.8	NS	0.15	2.79	10.9	NS	0.74	2.32	8.6	NS

HT=hypertension; DM=diabetes mellitus; CSF=complete sampling frame; SSF=spatial sampling frame; SRS=simple random sampling; STRS=stratified random sampling***; Abs diff=Absolute difference; CI=Confidence Interval; CV=Coefficient of variation; NS=Not significant; S=Significant.***

sical (non-spatial) indices. These differences were statistically significant at medium and high autocorrelation (Carl and Kuhn, 2017). In another study, the authors had non-autocorrelated data, and their investigation suggested no difference between the spatial and the classical (probability) measures of accuracy (Swacha, 2017). Other earlier studies located all residences within 20 yards or meters of the survey point (Kolbe *et al.,* 2006; Wang *et al.,* 2010;), whereas others (Grais *et al*., 2007) randomly selected one direction by the 'spinning pen' method and then picked the first household found following that route.

In another study (Grzegorz *et al.*, 2017), the samples were obtained using a raster dataset of vegetation from (5m x 5m) plots, comparing the performance of the methodology in assessing species composition patterns and environment vegetation. They employed non-probability preferred sampling and probabilistic sampling based on simple random sampling (SRS) and systematic sampling (SYS). Non-probability preferential sampling, according to their research, narrows the environmental gradient. An experiment was conducted in Minnesota to determine the effect of nitrogen fertilizer application rate on corn yield. The study showed that precision in the analysis of variance was significantly improved by compensating for these differences in spatial structure, with coefficients of variations of 11.4% versus 8.9% for classical probability sampling on the one hand with SRS and SYS on the other, respectively (Hernandex and Mulla, 2002).Under budget constraints and without an HH sampling frame, multilevel spatial sampling was used to sample more uniformly distributed women farmers and to collect demographic information. The spatial sampling method produced a nationally representative value that was comparable (Maduekwe and Vries, 2019).

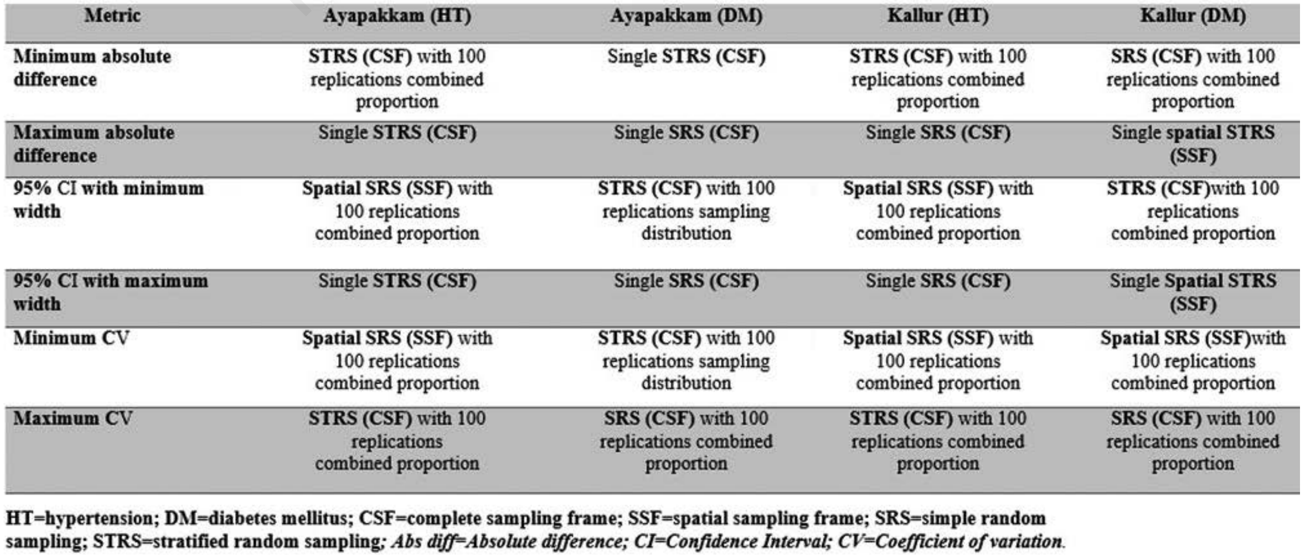
In 2014, National statistical agencies were motivated to replace the traditional census frame with a spatial sampling frame, namely a list of residential postal addresses (Australian Bureau of Statistics, 2014; Kalton *et al.*, 2014; Valliant *et al.,* 2014). Any community-level study requires a list of every individual household, and respondents can be randomly selected in the study community. For the past two decades, sampling households for survey-based research has gained much attention (Lee *et al.*, 2006). In developing a spatial sampling frame, the use of Google Earth satellite imagery and the geographical information system appears to be an efficient alternative in a demographic surveillance system (Escamilla *et al.*, 2014); and insecure environments and when census data are unavailable (Yihan and David, 2016). Earlier studies have employed different spatial sampling methodologies to choose a finite number of sample units by overlaying a regular geometric extent of the study area and assuming homogeneity within each unit (Desard and Bar-Hen, 2005). Kumar (2007) claims that regular grid cells do not match the very irregular geometries of residential neighbourhoods and suggests using discrete geographic space to generate a sampling frame of residential regions using GPS, GIS and remote sensing (RS). He showed that the method was robust and comparable.

Also, an investigation of *Plasmodium falciparum* transmission intensity was investigated in Lilongwe, Malawi with all HH structures digitized and assigned coordinates based on satellite imagery and GIS (Escamilla, 2014). Both Kumar (2007) and Escamilla *et al.* (2014) used discrete geographic grids of residential units (HH) as the spatial sampling frame for their research, accounting for spatial heterogeneity by giving weights when selecting a set of randomly selected geographic locations.

Advancements in 3S technologies, namely GIS, GPS and RS, offer a tremendous opportunity to create an efficient spatial sampling approach for demographic and health surveys. We also endorse that the use of GIS, GPS, and RS technologies, which should be applied to pinpoint the whereabouts of residential HHs as an indirect measure of a listing of these units.

We proposed a method for developing an HH spatial sampling frame (geographically constrained single shapefile of the residential population) in semi-urban and rural contexts. The spatial data consists of a single shape file rather than a geographic extent. Also, we compared the prevalence of self-reported HT and DM by two probability sampling methods using CSF with SSF. We show that the estimates are reliable, efficient, and comparable to those of traditional methods or CSF estimates. SSF would be helpful in situations with budgetary or other constraints, insecure environments, and when census data are unavailable. A CSF necessitates more

Table 3. Inference based on metrics of two sampling methods using two sampling frames.



Metric	Ayapakkam (HT)	Ayapakkam (DM)	Kallur (HT)	Kallur (DM)
Minimum absolute difference	STRS (CSF) with 100 replications combined proportion	Single STRS (CSF)	STRS (CSF) with 100 replications combined proportion	SRS (CSF) with 100 replications combined proportion
Maximum absolute difference	Single STRS (CSF)	Single SRS (CSF)	Single SRS (CSF)	Single spatial STRS (SSF)
95% CI with minimum width	Spatial SRS (SSF) with 100 replications combined proportion	STRS (CSF) with 100 replications sampling distribution	Spatial SRS (SSF) with 100 replications combined proportion	STRS (CSF)with 100 replications combined proportion
95% CI with maximum width	Single STRS (CSF)	Single SRS (CSF)	Single SRS (CSF)	Single Spatial STRS (SSF)
Minimum CV	Spatial SRS (SSF) with 100 replications combined proportion	STRS (CSF) with 100 replications sampling distribution	Spatial SRS (SSF) with 100 replications combined proportion	Spatial SRS (SSF)with 100 replications combined proportion
Maximum CV	STRS (CSF) with 100 replications combined proportion	SRS (CSF) with 100 replications combined proportion	STRS (CSF) with 100 replications combined proportion	SRS (CSF) with 100 replications combined proportion

HT=hypertension; DM=diabetes mellitus; CSF=complete sampling frame; SSF=spatial sampling frame; SRS=simple random sampling; STRS=stratified random sampling*; Abs diff=Absolute difference; CI=Confidence Interval; CV=Coefficient of variation.*

human resources, while the cost in the SSF is the requirement for personnel skilled in GIS use. In addition, there is increased need for desk work, primarily in preparing or mapping residential regions. Any desired variable sample size or sampling design would be achievable with the spatial sampling frame.

In the present study, we have used individuals as the unit of analysis using CSF. In contrast, in using SSF, we randomly selected the HH, and then an individual was selected and used as the unit of analysis. Malaria indicator surveys generally aim to estimate disease prevalence at the community level by sampling from a list of all HHs, primarily heads of households (Escamilla *et al.*, 2014).

We assumed that HHs (indirectly an individual) are selected directly from the SSF of solely HH regions of the population, as done in telephone surveys, which choose from a subscriber directory or utilize random digit dialling (indirectly an HH) (Lepowski, 1988) rather than going through multiple phases of selection. In large-scale national surveys, the HHs are selected first, followed by individuals. When HH is used as a unit of study, the variation is likely to be a little higher. Weights are typically assigned to account for unequal probability of selection and survey non-response. When the data exhibits spatial autocorrelation, they adjust the weighted sample distribution for key study variables. Since our study was based on the entirely surveyed list of HHs in the study area, non-response did not occur. The disease conditions were randomly dispersed in space. The spatial design sought random sampling of a single eligible individual per sampled HH, based on the assumption that a HH of individuals is relatively homogeneous compared to the broader population of individuals. As a result, no weights were assigned. We considered separate datasets for males and females. From a spatial perspective, HHs with males and those with females were investigated as separate layers. They were independent and can be overlaid to comprise the whole study population. Because individuals can only exist in one stratum, they were considered to be independent.

GIS, GPS and linking GPS to PDAs simplified the collection and framing of a complete enumeration list (SSF) and analysis of the population data. Although Google Earth Pro version 7.3.2 provided high-resolution satellite images, we would have missed a few new residential houses had we used only Google Earth images, however, this did not occur as we had access to the entire locations of all HHs in 2019 for both study settings. We used Garmin Handheld Trimble Juno SC GPS units to survey and observed that Google Earth streets could sometimes shift but by less than 10 m. The survey data could therefore be directly taken to the GIS environment with the GPS unit. Getting signals from satellites was much faster in urban settings than in rural ones. With a limited number of field staff, weather conditions and repeated visits of the demographic health survey, it took a long time to digitize all HHs, including the entire study area, and complete the survey. This method provides an opportunity for an efficient spatial approach comparable to the classical approach. However, it involves more desk work, primarily in preparing the digitized map of residential regions using Google Earth; GIS analyst or trained personnel is essential and requires GIS software. The urban setting was digitized with only eight polygons of residential areas to make a single polygon shapefile which was much easier. In contrast, combining 156 polygons into one polygon in the rural setting was cumbersome. Refusal or non-response error does not occur in this study; otherwise, non-response must be addressed at the analysis time. The vast majority of Kallur families live in single-family houses, with only a small percentage of the population living in multi-story or multi-family housing in each location or site. However, roughly 15% of the population in our semi-urban neighbourhood of Ayapakkam lives in multiple living units or HHs. If the study selected a location or residence with more than one HH, the first one with the required age group of 18-69 years was selected randomly. In a relatively small number of instances (<5%) in both study areas, randomly produced points do not correspond with HH locations and selected empty houses or petty shops. If the study area consists primarily of high-rise structures, it may oversample the sampling methods. Changing random points to coincide with the nearest households would also eliminate the sample's randomness.

Conclusions

Using SSF is as efficient as using CSF by two probability sampling methods in two study settings as shown by our study investigating two non-communicable diseases. However, this type of sampling frame should be further validated and compared with more probability sampling methods to determine its utility in communicable or infectious disease conditions and urban settings. This exercise would pave the way to understanding the pros and cons of using both frames in future research.

References

Abhijit B, Rohini P, Michael W, 2012. Delhi's slum-dwellers: deprivation, preferences and political engagement among the urban poor. Available from: https://www.theigc.org/wpcontent/uploads/2014/10/Banerjee-Et-Al-2012-Working-Paper.pdf Accessed: June 15, 2021.

Australian Bureau of Statistics, 2014. Sample and frame maintenance procedures for census and household surveys. Available from: www.abs.gov.au. Accessed: June 15, 2021.

Carl G, Kühn I, 2017. Spind: a package for computing spatially corrected accuracy measures. – Ecography 40:675-82.

Christakos G, 2005. Random field models in earth sciences. Dover Publications, New York.

Cochran WG, 1977. Sampling techniques. 3rd Edition, John Wiley and Sons, New York.

Cressie N, 1991. Statistics for Spatial Data. New York: Wiley. 900 pp.

Dessard H, Bar-Hen A, 2005. Experimental design for spatial sampling applied to the study of tropical forest regeneration. Can J For Res 35:1149-55.

EPA, 2002. Available fromhttps://www.epa.gov/sites/production/files/2015-06/documents/g5s-final.pdf Accessed: June 15, 2021.

Eric D, 2012. Spatial Sampling. Available from https://pages.uncc.edu/ eric-delmelle/wp-content/uploads/sites/150/2012/12/spatial-sampling-delmelle.pdf Accessed: June 15, 2021.

Escamilla V, Emch M, Dandalo L, Miller WC, Hoffman I, 2014. Sampling at community level by using satellite imagery and geographical analysis. Bull World Health Organ 92:690-4. Available from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4208578/Accessed: June 15, 2021.

Grais RF, Rose AM, Guthmann JP, 2007. Don't spin the pen: two alternative methods for second-stage sampling in urban cluster surveys. Emerg Themes Epidemiol 1;4:8.

Grzegorz S, Zoltán B, Zygmunt, Daniel P, Ludwik Ż, 2017. A performance comparison of sampling methods in the assessment of species composition patterns and environment–vegetation relationships in species-rich grasslands. Acta Soc Bot Pol 86:3561

Haining RP, 2001. Spatial sampling. In: International Encyclopedia of the Social & Behavioral Sciences. Available from https://www.sciencedirect.com/topics/computer-science/spatial-sampling Accessed: June 15, 2021.

Haining RP, 2003. Spatial data analysis: Theory and practice. Cambridge University Press. 452 pp.

Hernandez JA, Mulla DJ, 2002. Comparing statistical analysis for landscape scale experimental designs. 6th International Congress on Precision Agriculture and Other Precision Resource Management, July 15, Minneapolis, MN.

Improving Health in Slums Collaborative, 2019. A protocol for a multi-site, spatially-referenced household survey in slum settings: methods for access, sampling frame construction, sampling, and field data collection. BMC Med Res Methodol 19:109.

Joshua V, Sunitha K, Muthu G, Sinduja V, Venkatesh P, Nandini P, Shantaraman K, ManickamP, Murhekar MV, Yuvaraj J, 2020. Self-reported morbidity profile among geriatric population in ICMR-model rural health research unit, Kallur, Tirunelveli. J Indian Acad Geriatr 16:95-100.

Kalton G, Kali J, Sigman R, 2014. Handling frame problems when address-based sampling is used for in-person household surveys. J Surv Stat Methodol 2:283-304.

Kolbe AR, Hutson RA, 2006. Human rights abuse and other criminal violations in Port-au-Prince, Haiti: a random survey of households. Lancet 368:864-73.

Kumar N, 2007. Spatial Sampling Design for a Demographic and Health Survey. Popul Res Policy Rev 26:581-99

Lee C, Moudon AV, Courbois JYP, 2006. Built environment and behavior: Spatial sampling using parcel data. Ann Epidemiol 16:387-94.

Lepowski JM, 1988. Telephone sampling methods in the United States. Telephone Survey Methodology (Eds. Groves RM, Biemer PP, Lyberg LE, Massey JT, Nicholls II WL, Waksberg J). New York: Wiley 73-98 pp.

Maduekwe E, Vries WTD, 2019. Random spatial and systematic random sampling approach to development survey data: evidence from field application in Malawi. Sustainability 11:6899.

Malhotra NK, Birks, DF, 2006. Marketing research: an applied approach, 3rd Edition, Prentice Hall, Upper Saddle River.

Miller AC, Rohloff P, Blake A, Dhaenens E, Shaw L, Tuiz E, Grandesso F, Mendoza MC, Thomson DR, 2020. Feasibility of satellite image and GIS sampling for population representative surveys: a case study from rural Guatemala. Int J Health Geogr19:56.

Muller W, 1998. Collecting spatial data: optimal design of experiments for random Fields, revised edition. Contribution to statistics. Heidelberg: Physica-Verlag.

Ripley, B, 1981. Spatial Statistics. Wiley, New York, 252.

Rogerson PA, Delmelle E, Batta R, Akella M, Blatt A, Wilson G, 2004. Optimal sampling design for variables with varying spatial importance. Geogr Anal 36:177–94.

Speizer IS, Nanda P, Achyut P, Pillai G, Guilkey DK, 2012. Family planning use among urban poor women from six cities of Uttar Pradesh, India. J Urban Health 89:639-58.

Stehman SV, Overton SW, 1996. Spatial sampling. In: Arlinghaus SL (ed.) Practical Handbook of Spatial Statistics; Boca Raton, FL: CRC Press 31–64 pp.

Swacha G, Botta-Dukát Z, Kącki Z, Pruchniewicz D, Żołnierz L, 2017. A performance comparison of sampling methods in the assessment of species composition patterns and environment–vegetation relationships in species-rich grasslands. Acta Soc Bot Pol 86:3561.

Thomas B, Alexander K, Magdalena S, Giovanna B, Danuté K, 2019. Available from https://ec.europa.eu/eurostat/cros/system/files/qgfss-v1.51.pdf Accessed June 15, 2021.

Thomson DR, Shitole S, Shitole T, Sawant K, Subbaraman R, David EB, Anita PM, 2014. A system for household enumeration and re-identification in densely populated slums to facilitate community research, education, and advocacy. PLoS One 9:e93925.

Valliant R, Hubbard F, Lee S, Chang C, 2014. Efficient use of commercial lists in US Household Sampling. J Surv Stat Methodol2:182- 209.

Wang JF, Haining RP, Cao ZD, 2010. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. Int J Geogr Inf Sci 24:523-43.

WHO STEPS Surveillance Manual,2017. The WHO STEP wise Approach non-communicable disease risk factor surveillance. Geneva: World Health Organization. Available at https://www.who.int/teams/noncommunicable-diseases/surveillance/systems-tools/steps/manuals Accessed: June 15, 2021.

Yihan L, David PK, 2016. Using satellite imagery and GPS technology to create random sampling frames in high-risk environments Int J Surg 32:123-8.