SSM - Population Health 21 (2023) 101317

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: www.elsevier.com/locate/ssmph

Intra-cluster correlations in socio-demographic variables and their implications: An analysis based on large-scale surveys in India

Laxmi Kant Dwivedi [a,*], Bidhubhusan Mahaptra [b], Anjali Bansal [c], Jitendra Gupta [c], Abhishek Singh [d], T.K. Roy [c]

[a] Department of Survey Research & Data Analytics, International Institute for Population Sciences, Mumbai, India
[b] Population Council, New Delhi, India
[c] International Institute for Population Sciences, Mumbai, India
[d] Department of Public Health & Mortality Studies, International Institute for Population Sciences, Mumbai, India

ARTICLE INFO

Keywords:
Clustering
PSU
NFHS
Intra-class correlation coefficients
ICC
Multilevel analysis
Religion
Caste
Use of family planning
Anaemia
Full immunization
Precision loss
Cluster size
Sample size
Effective sample size
India

ABSTRACT

Individuals who share similar socio-economic and cultural characteristics also share similar health outcomes. Consequently, they have a propensity to cluster together, which results in positive intra-class correlation coefficients (ICCs) in their socio-demographic and behavioural characteristics. In this study, using data from four rounds of the National Family Health Survey (NFHS), we estimated the ICC for selected socio-demographic and behavioural characteristics in rural and urban areas of six states namely Assam, Gujarat, Kerala, Punjab, Uttar Pradesh, and West Bengal. The socio-demographic and behavioural characteristics included religion & caste of the household head, use of contraception & prevalence of anaemia among currently married women and coverage of full immunization services among children aged 12-23 months. ICC was computed at the level ofPrimary Sampling Units (PSUs), that is, villages in rural areas and census enumeration blocks in urban areas. Our research highlights high clustering in terms of religion and caste within PSUs in India. In NFHS-4, the ICCs for religion ranged from the lowest of 0.19 in rural areas of Kerala to the highest of 0.67 in urban areas of West Bengal. For the caste of the household head, the ICCs ranged from the lowest of 0.12 in the urban areas of Punjab to the highest of 0.46 in the rural areas of Assam. In most of the states selected for the study, the values of ICC were higher for the use of family planning methods than for full immunization. The value of ICC for use of contraception was highest for rural areas of Assam (0.15) followed by rural areas of Gujarat (0.13). A higher value of ICC has considerable implications for determining an effective sample size for large-scale surveys. Our findings agree with the fact that for a given cluster size, the higher the value of ICC, the higher is the loss in precision of the estimate. Knowing and taking into account ICCs can be extremely helpful in determining an effective sample size when designing a large-scale demographic and health survey to arrive at estimates of parameters with the desired precision.

1. Introduction

A cluster design (both a conventional cluster sampling design and a multistage design) poses the problem that units within clusters are generally dependent. The settlement pattern in a population-based survey is usually not random. Often, individuals sharing similar socio-economic and cultural characteristics prefer to stay close to each other and depict similar health outcomes (Donner, 1986; Gulliford et al., 1999; Johnson et al., 2015; Killip, Mahfoud, & Pearce, 2004; Pagel et al., 2011; Simpson et al., 1995). This propensity to cluster together results in a positive intra-cluster correlation in their socio-economic characteristics like religion, caste, education, and occupation. (Killip et al., 2004). When it comes to behavioural and attitudinal variables, a neighbour's influence is likely to be dented, although it will depend on the specific issue at hand (Benefo, 2006; Kravdal, 2002, 2007). Sometimes, the presence of certain external factors can enhance the clustering effect in a

Abbreviations: ICC:, Intra-class Correlation Coefficient; NFHS:, National Family Health Survey; PSU:, Primary Sampling Unit; CEB:, Census Enumeration Block; EPSEM:, Equal Probability of Selection Method.

* Corresponding author.
E-mail addresses: laxmikant@iipsindia.ac.in (L.K. Dwivedi), bbmahapatra@popcouncil.org (B. Mahaptra), anjalibansal35@gmail.com (A. Bansal), jiitend@gmail.com (J. Gupta), abhishek@iipsindia.ac.in (A. Singh), tarunkroy@yahoo.com (T.K. Roy).

https://doi.org/10.1016/j.ssmph.2022.101317
Received 16 December 2021; Received in revised form 15 October 2022; Accepted 12 December 2022
Available online 15 December 2022

variable. For example, the availability of a bank in some PSUs (villages) might lead to clustering in whether a person has a bank account or not. Similarly, if health facilities are scattered over rural areas, it can also lead to clustering in seeking maternal and child health care services (Roy, Acharya, & Roy, 2016). Intervention studies in which the cluster is the unit of allocation are increasingly being used to evaluate health interventions implemented at the level of a geographic area or a health service organization unit (Donner & Klar, 1994). Therefore, intra-class correlation (ICC) holds great importance for the present study.

The knowledge of ICC can be beneficial in many ways. First, it can assist in determining the sample size for the estimation of a parameter in a survey (Aliaga & Ren, 2006; Kerry & Bland, 1998; Liljequist et al., 2019; Pagel et al., 2011). Since clustering tends to increase the design effect, one needs to consider a larger sample size to estimate a given parameter with the required precision (Alegana et al., 2017). Second, because the value of ICC of a parameter is considered portable, its availability in a survey can facilitate fixing the sample size in another similarly designed survey (Aliaga & Ren, 2006).

Third, knowing the ICC can reveal how a parameter is distributed in a population, especially whether it is random or clustered (Eldridge et al. 2009; Eldridge, Ashby, & Kerry, 2006; Rutterford et al., 2015). A high ICC indicates that the parameter is not randomly distributed but concentrated in certain localities (Gulliford et al., 1999; Roy et al., 2016). Consider anaemia among women, which is a massive problem in India. A finding that its prevalence is highly clustered in rural areas would have considerable implications. It can assist in carrying out further research to gauge its causal linkages and formulate a suitable policy to control its prevalence. It can also help in organizing an effective intervention program to curb the menace. A high intra-cluster correlation in the prevalence of malaria can similarly be crucial in promoting sleeping under an insecticide-treated bed net.

Intervention studies in which the cluster is the unit of allocation are increasingly being used to evaluate health interventions implemented at the level of a geographic area or a health service organization unit (Donner & Klar, 1994). Given the importance of ICC in understanding the effect of clustering in a population, we endeavour to estimate the ICC for some selected socio-economic and demographic parameters for different states using data from four rounds of the National Family Health Survey (NFHS). The ICC values are estimated separately for the rural and urban areas for the selected states. For the sake of brevity, the present study shows results only for the second (NFHS-2) and the fourth (NFHS-4) rounds of NFHS to understand the changes in ICC over time.

2. Materials and methods

This study used data from the National Family Health Survey (NFHS), which is conducted at different points in time. NFHS is the Indian version of the Demographic and Health Surveys (DHS). The fourth round of the survey (NFHS-4) was conducted during 2015–16, and the previous three rounds were conducted in 1991–92 (NFHS-1), 1998–99 (NFHS-2), and 2005–06 (NFHS-3). NFHS is a nationally-representative cross-sectional survey that includes representative samples of households from all over India. Unlike the three previous rounds of the survey that provided national- and state-level estimates only, NFHS-4 also provided district-level estimates of demographic and health parameters and data on various socio-economic and program dimensions that are critical for implementing the desired demographic and health parameter changes. A stratified, two-stage sampling method is primarily used in NFHS to obtain a representative sample of households. In NFHS-4, probability proportional to size (PPS) sampling was used to select villages from the rural areas and census enumeration blocks (CEB) from urban areas. The number of households in a village was considered a size measure.

2.1. Statistical analysis

ICC measures the homogeneity of the variable under study within the PSUs. In practice, units within a PSU tend to be somewhat similar to each other for nearly all variables, although the degree of similarity varies from one variable to another (Roy & Pandey, 2008). In its broadest sense, ICC is also known as variance partition coefficient (VPC) and is used to analyse contextual phenomena. The percentage of an outcome's overall variance that can be attributed to the PSU level is disclosed by the ICC.

ICC = [(Between PSU level variance)/ (Within PSU variance + Between PSU level variance)]

We computed the intra-class correlation coefficient (ICC) to obtain the share of total variance in the outcome variable at the PSU level. For example, the ICC at the PSU level was defined as the PSU-level variance relative to the sum of the variance at all levels.

$$\text{ICC at the PSU level} = \frac{PSU\ level\ variance}{(PSU\ level\ variance + individual\ level\ variance)}$$

$$\text{ICC at the PSU level} = \left[\frac{\sigma^2_{0j}}{\sigma^2_{0j} + 3.29} \right]$$

To examine the ICC, it is important to understand the design effect. The design effect of the complex survey design was first studied by (Kish, 1965, p. 161) and is used as a measure of efficiency in most complex survey designs (Park & Lee, 2001; Särndal, Swensson, & Wretman, 2003). Kish, 1995 gave the following formula to understand the relation between the design effect (deff) of a variable and its ICC:

$$deff = \frac{The\ sampling\ variance\ in\ case\ of\ a\ cluster\ design\ of\ sample\ size\ 'n'}{The\ sampling\ variance\ in\ case\ simple\ random\ sampling\ of\ an\ equivalent\ size} = 1 + (b-1)\text{ICC} \quad (1)$$

Where 'b' is the average number of units selected in a PSU (referred to as the cluster size). The precision of a cluster sample design, in relation to a simple random sample of the same size, depends on ICC and the cluster size. Since ICC is generally positive, a higher value of either ICC and/or cluster size is detrimental for a design. This means that its number of iid (independently identically distributed) sample size will reduce. Kish (1995) gave a formula to obtain the reduced (effective) sample size in a cluster design (n_{eff}), which is expressed as:

$$n_{eff} = \frac{n}{deff} \quad (2)$$

This means, with a design effect of 2, the sample size in a cluster sampling design would have to be doubled so as to produce the same precision as a simple random sample of size n.

The ICC of a variable can be estimated in a number of ways. One approach is to utilize equation (1) for the estimation. Accordingly, the ICC can be estimated as:

$$\text{ICC} = \frac{(deff - 1)}{b - 1} \quad (3)$$

Here, the value of b, the cluster size, is the ratio of the total number of respondents in a survey design to the total number of PSUs selected. The design, however, should have a mechanism to control the variation in b (Roy et al., 2016). A survey generally provides information on both the sampling variance of the design and the deff of a variable. The sampling variance, and hence, the value of deff of an estimate, apart from the clustering, can be affected by factors such as:

1. Use of weights to adjust for unequal probability of selection
2. Adjustment for non-response and non-coverage

3. Stratification used in a survey

It is necessary to eliminate the effect, if present, of all such factors from the given value of deff to reflect the increase in sampling variance solely due to clustering. Kalton, Brick, and Lê (2005) has outlined the procedure to adjust the value of deff.

Positive clustering in a variable, say religion, means that people of a religion are more likely to be concentrated in a few PSUs rather than be distributed randomly over the PSUs. In statistical terminology, this means that between PSU variation, in terms of the religious composition of people, will be higher than within PSU variation. In other words, whether the selection of individuals is made randomly from a PSU (in cluster sampling) or the entire population (in simple random sampling) will make a difference. The ICC of a variable measures the extent to which the between-cluster variation differs from the within-cluster variation. The value of ICC can be computed from the following formula (Killip et al., 2004; Liljequist et al., 2019; Snedecor, 1989):

$$\mathrm{ICC} = \frac{\left(s_b^2 - s_w^2\right)}{\left(s_b^2 - (b-1)s_w^2\right)} \tag{4}$$

Where s_b^2 and s_w^2 are respectively the between and within sum of squares usually computed in an analysis of variance.

A multilevel analysis can also be used to get an idea about the extent of clustering in a variable at a PSU level. The ICC may be computed by applying multilevel (two level) regression analysis considering PSU at second level using NFHS data set and it can be interpreted as the correlation of the variable under research between two distinct women from the same PSU.

In multilevel regression analysis of two level data where women are nested within PSU takes the form

$$\log\left(\frac{\pi_{\mathrm{ij}}}{1-\pi_{\mathrm{ij}}}\right) = \beta_0 + u_{0\mathrm{j}}$$

Here, π_{ij} represents the probability that the ith women of jth PSU takes the value of 1if she uses any family planning method, for example, otherwise 0. $u_{0\mathrm{j}}$ represents the error terms at the PSU level.

$u_{0\mathrm{j}} \sim N(0, \sigma_{0\mathrm{j}}^2)$; $\sigma_{0\mathrm{j}}^2$ is variance at PSU level

To estimate the multi-level random intercept model, the Markov Chain Monte Carlo (MCMC) estimation process was adopted, and the model was executed using the runmlwin program in STATA 16.0.

It may be mentioned that the NFHS design is a self-weighing design for the estimation of parameters for both the rural and urban areas in a state. Any departure from proportionality, like increasing the sample size in some urban areas or metropolitan cities, is compensated using weights. The weights are also used to compensate for the non-response, which is generally less than 10%. The design has a built-in procedure to control the variation in the cluster size. In the first three rounds of the survey, a restriction was put to maintain the variation between b and 2b. Doing so was not necessary in NFHS-4 as its design was slightly different. As a result, the non-coverage error was non-existent for the six states considered in this paper.

In the present paper, the value of deff, representing equation (1), was estimated for a given domain (rural or urban area of a state) in a slightly different manner. Let n be the total sample size obtained from 'a' selected PSUs, from each of which an average of b ($b = \frac{n}{a}$) individuals are selected in a domain. Let y_{ij} denote the value of a parameter for the j^{th} individual in the i^{th} PSU and let x_i be the number of observations on y in the i^{th} PSU. Considering the PSU-level information, y_i ($= \sum_{j=1}^{x_i} y_{ij}$) and x_i, the sampling variance of the parameter, r, for the multistage design is obtained as:

$$\mathrm{V}(\mathrm{r}) = \frac{1}{x^2}\left[\mathrm{v(y)} + r^2\,\mathrm{v(x)} - 2\,\mathrm{r\,cov(yx)}\right] \tag{5}$$

Where, r, the ratio estimate is equal to:

$$\mathrm{r} = \frac{y}{x} = \frac{\sum_{i=1,j=1}^{a,x_i} y_{i,j}}{\sum_{i=1}^{a} x_i} \tag{6}$$

The variances, v(y) and v(x), and the covariance, cov(yx), can be computed in the usual manner (Kish, 1995; Roy, Acharya, & Roy, 2016). Thus, the sampling variance obtained and the estimated sampling variance in a simple random sampling provide the estimated value of deff for a variable. It may be mentioned that there was hardly any difference in the estimated value of ICC by the three methods, that is, by using equations (3) and (4) and by applying the multilevel analysis.

Many states in India exhibit considerable variation in their levels of development. Therefore, for brevity, only six states were selected for the analysis. The states were selected considering both their geographic location and level of development. This was done with the expectation to capture the variation in the ICC and examine its association, if any, with the level of development of an area. The six states included in the study were: Kerala from the South, Gujarat from the West, Punjab from the North, Uttar Pradesh from the Central, West Bengal from the East, and Assam from the Northeast. Compared to the national average, Kerala and Punjab rank much higher on Human Development Index (HDI). Gujarat ranks marginally above and West Bengal marginally below the national average. Assam and Uttar Pradesh fall well below the average.

While the first three rounds of NFHS provide estimates at the state level, the fourth round (NFHS-4) also gives estimates for each district in a state. Generally, NFHS follows a two-stage sampling design for selecting households from the rural areas in a state and a three-stage design for selecting households from the urban areas. In rural areas, villages form the PSUs, and households are selected from each selected PSU after listing the households in them. In urban areas, first wards are selected and, then, one census enumeration block (CEBs) is selected from each ward. Next, households are selected from a selected CEB after its listing. Since only one CEB is selected from a selected ward, for all practical purposes, CEBs can be treated as the PSUs in urban areas. In NFHS-4, the selection of households in urban areas was done in two stages instead of three by selecting the CEBs in the first stage itself. A uniform number of 20 households were selected from each selected PSU in NFHS-4. The number of households to be selected in each PSU, except in NFHS-4, is done to conform to the self-weighing design.

NFHS enables the estimation of various socio-economic and demographic variables in both the rural and urban areas of a state. We have illustrated the ICC values for five variables, namely religion, caste, use of contraception, use of immunization services, and prevalence of anaemia. These variables are expected to cover the range of variation in ICC. This effort was undertaken to understand how the precision of an estimate in a large-scale survey may get influenced by the presence of ICC. The importance of a parameter was also considered in the selection. The first two variables, as mentioned, are likely to provide the highest degree of ICC. Knowledge about the extent to which the use of contraception and the immunization of a child depends on the behaviour and practices of neighbours could help fine tune the government programs corresponding to these issues.

The first two variables are based on the household schedule and refer to the characteristics of the head of a household. The other three are based on the interviews of currently married women aged 15–49 years in a household. All the five variables are multinomial in the sense that there are more than two possible outcomes for each of them. However, while computing ICC, they were expressed as dichotomous variables. For example, in NFHS, the caste of the head of the household is shown across four categories as Scheduled castes (SC), Scheduled tribes (ST), Other backward classes (OBC), and 'others who do not belong to any of the previous three groups.' If caste is represented as a dichotomous variable by SCs (that is, y = 1 for SCs and 0 for groups other than SCs), the ICC can only explain the clustering among SCs but not the clustering in the remaining three categories. For that, it would be better to consider

two additional variables of those three categories and check for their ICCs.

In the present analysis, for the computation of ICC, various socio-demographic and behavioural variables were coded as dichotomous variables. Religion of household head was coded as 1 for the Hindu and 0 for other than the Hindu religion. Caste of household head was coded as 1 if the caste was SC, ST, or OBC and as 0 for caste other than SC/ST/OBC. Contraception use among currently married women aged 15–49 years was coded as 1 if a woman was using any modern method of contraception and as 0 if she was not using any modern method. Anaemia was categorized as 1 if a currently married woman was anaemic (<12.0 g/dl haemoglobin) and as 0 otherwise. Children aged 12–23 months, if fully vaccinated with BCG, measles-containing vaccine (MCV)/MR/MMR/Measles, and 3 doses each of polio (excluding polio vaccine given at birth) and DPT or penta vaccine, were coded as 1 and as 0 otherwise.

3. Results

At the outset, it is prudent to have an idea about the status of the variables under study across the six states. Table 1 gives an overview of the variables taken into account for evaluating ICC and the number of rural and urban PSUs in NFHS-4.

Most households in all the states belonged to the Hindu religion, except in Punjab, where Sikhs formed the majority and Hindus formed 37% of households. Muslim households formed about 1/3rd of households in Assam and a quarter in West Bengal. Uttar Pradesh also had a sizeable proportion of Muslim households (18%). Hindus and Muslims formed more than 95% of households in Gujarat, Uttar Pradesh, Assam, and West Bengal. Kerala had Hindus, Muslims, and Christians sharing a similar population. According to NFHS-4, the 'other' caste group constituted the highest percentage of households in West Bengal, Punjab, Kerala, and Assam. In Uttar Pradesh and Gujarat, OBCs were the dominant caste.

Table 1
Percentage of household heads who are Hindu and have 'other' as their caste; percentage of currently married women aged 15–49 years who use any contraception; percentage of women who are anaemic; and percentage of children aged 12–23 months who are fully vaccinated in the six selected states, NFHS-4 (2015–16).

	Assam	Gujarat	Kerala	Punjab	Uttar Pradesh	West Bengal
Percent of Hindu households	63.83	91.88	58.80	36.78	82.03	72.98
Percent of households with caste of head as 'other'[a]	36.24	30.16	35.24	42.80	22.18	41.74
Percent of currently married women using any contraception	52.40	46.89	53.11	75.80	45.47	70.90
Percent of women aged 15–49 years having anemia	46.01	54.93	34.31	53.52	52.43	62.46
Percent of children 12–23 months fully vaccinated	47.08	50.43	82.05	89.05	51.07	84.44
Number of rural PSUs	991	619	333	471	2659	510
Number of urban PSUs	170	369	198	289	979	212

[a] Persons in the 'other' group consist of those who do not belong to the scheduled castes, scheduled tribes, or other backward classes.

The use of any contraception ranged from a low of 46% in Uttar Pradesh to a high of 71% in West Bengal. Contraceptive use was generally widespread in all the states, and female sterilization happened to be the preferred method except in Assam, where traditional methods were more popular. The use of traditional methods was also prevalent in West Bengal. The states of Uttar Pradesh and Assam lagged behind in the immunization program. In almost all the states, at least half of the children had received all the vaccinations. The prevalence of anaemia was relatively low in Kerala; however, even there, it was substantially high, with around one-third of women having any anaemia.

3.1. ICC in socio-economic variables

Table 2 depicts the ICC in the distribution of households by religion for rural and urban areas for the six selected states based on two rounds of NFHS (NFHS-2 and NFHS-4) conducted two decades apart. There was a very high tendency among people to co-reside with those belonging to the same religion as theirs. This religious affinity was visible, with some variations, in both rural and urban areas of all the states. The ICC varied from a low of 0.19 in rural Kerala to a high of 0.71 in urban West Bengal in NFHS-2 and remained more or less the same in NFHS-4. There are two possible ways to hypothesize the state-level variation in ICC for religion. First is the existence of a negative association between ICC and the level of development in the area. To some extent, the ICC values tend to confirm it. With a higher level of development, Kerala showed a relatively lower correlation, and with a lower level of development, Assam depicted a high level of correlation. However, in West Bengal, the correlation was substantially higher than in the less developed state of Uttar Pradesh.

As mentioned earlier, two groups, Hindus and Muslims, constituted more than 90% of population in four of the states, except Punjab, where the two dominant groups were Hindus and Sikhs, and Kerala, where three religious groups (Hindus, Muslims, and Christians) had an almost equal share in the population. The second possible explanation for the variation in ICC relates to the religious composition of the population. The magnitude of ICC is expected to be a function of the proportion of a religious group in a population. An increase in the proportion of the population of a religious group (other than the main religious group) will tend to push up its concentration. This is likely because the migration of a religious group, which is expected to be the reason for its increasing size, tends to occur selectively in places where people belonging to similar socio-economic characteristics reside. The migration from the neighbouring country of Bangladesh to the states of Assam and West Bengal, which have a substantial proportion of Muslim population (Table 1), has been well documented (Gillan, 2002). It may be mentioned that the concentration of the Muslim population is quite high in Kerala as well [its ICC values (not shown in the table) were 0.41 and 0.34 in NFHS-2 and NFHS-4 respectively].

It is believed that the degree of the concentration is less in urban than in rural areas of a state because of the higher level of socio-economic development in the former. This was, however, not true in the case of the religious distribution in Assam and Punjab. One possible reason for the lower concentration in urban areas is the size of PSUs. PSUs are

Table 2
ICC in distribution of households by religion of household head by residence in six selected states, NFHS-2 & NFHS-4.

States	NFHS-2		NFHS-4	
	Rural	Urban	Rural	Urban
Kerala	0.19	0.27	0.19	0.28
Gujarat	0.21	0.36	0.34	0.53
Punjab	0.33	0.29	0.46	0.28
Uttar Pradesh	0.37	0.54	0.38	0.62
Assam	0.67	0.33	0.69	0.46
West Bengal	0.55	0.71	0.61	0.67

generally more compact and smaller in size in urban (considering CEB, as mentioned earlier, to be a PSU) than in rural areas and hence not exactly comparable. In Kerala, Gujarat, Uttar Pradesh, and West Bengal, the ICC values were higher in urban compared to rural areas and in West Bengal and Uttar Pradesh, the degree of ICC obtained for the urban areas was very high. Contrary to expectation, the extent of the correlation increased over the years. This was particularly true for Gujarat, where the ICC increased substantially in both rural and urban areas. It is worth mentioning that the ICC for rural Gujarat jumped from 0.21 in NFHS-2 to 0.60 in NFHS-3 (not shown in the table) and then declined to 0.34 in NFHS-4. In urban Gujarat, the ICC declined slightly to 0.35 in NFHS-3 (not shown in the table) from NFHS-2 and increased sharply in NFHS-4 (0.53).

Table 3 provides the ICC in the distribution of households by the caste of the head of the household for rural and urban areas for the six selected states based on NFHS-2 and NFHS-4. The value of ICC was substantially high in some cases. For example, in Assam, religious distribution had a high clustering than the caste-wise residential pattern in the rural areas. This concentration has increased considerably over the years, even in its urban areas. Gujarat also shown a significant increase in the value of ICC. Rural areas in every state showed an increase in ICC, although their extent was smaller in the remaining four states. Also, apart from Punjab and Kerala, the ICC in urban areas has also increased.

3.2. ICC among behavioural variables

Table 4 presents the ICC in the distribution of currently married women in terms of the use of any contraceptive method. The clustering effect was substantially high in rural areas, with all the states having a value between 0.10 and 0.20 in the earlier NFHS (table not shown). Its level declined moderately over the years and remained higher than 0.05, which is generally accounted for while fixing the sample size. The level of ICC was much lower in urban than in rural areas of every selected states in NFHS-2. However, the clustering increased over the years. For example, in Kerala, the ICC in urban areas increased substantially and was higher than in rural areas in NFHS-4.

In rural areas, the ICC in the distribution of currently married women by whether any anaemia was low and did not demonstrate any specific trend over time (Table 5). It increased in Punjab, where there was a substantially high clustering in NFHS-4. Even the urban areas of Punjab showed a high level of clustering of anaemia. Gujarat, where more than half of the women were found to be anaemic in NFHS-4, showed the lowest level of clustering. This suggests that anaemia was not concentrated in a group of villages but was widely spread. It can be mentioned, in this context, that people are generally vegetarian in this state (Agrawal, Millett, Dhillon, Subramanian, & Ebrahim, 2014), which contributes to their anaemic status.

Table 6 presents ICC in the distribution of children 12–23 months old by whether they were fully vaccinated. The percentage of children who had received all the basic vaccinations increased to 62% in NFHS-4 from a low of 44% in NFHS-3. The fourth round of the survey shows that the ICC values were high for Uttar Pradesh, Assam, and Gujarat both in rural and urban areas.

Table 3
ICC in distribution of households by caste of household head by residence in six selected states, NFHS-2 & NFHS-4.

States	NFHS-2		NFHS-4	
	Rural	Urban	Rural	Urban
Kerala	0.13	0.22	0.15	0.20
Gujarat	0.18	0.10	0.38	0.24
Punjab	0.14	0.18	0.16	0.12
Uttar Pradesh	0.17	0.18	0.20	0.21
Assam	0.34	0.09	0.46	0.19
West Bengal	0.11	0.07	0.19	0.16

Table 4
ICC in the distribution of currently married women aged 15–49 years in terms of use of any contraception by residence in six selected states, NFHS-2 & NFHS-4.

States	NFHS-2		NFHS-4	
	Rural	Urban	Rural	Urban
Kerala	0.11	0.03	0.05	0.12
Gujarat	0.15	0.02	0.13	0.12
Punjab	0.15	0.01	0.11	0.04
Uttar Pradesh	0.11	0.06	0.09	0.10
Assam	0.18	0.09	0.15	0.06
West Bengal	0.12	0.01	0.11	0.07

Table 5
ICC in the distribution of currently married women in terms of (any) anaemia by residence in six selected states, NFHS-2 & NFHS-4.

States	NFHS-2		NFHS-4	
	Rural	Urban	Rural	Urban
Kerala	0.03	0.04	0.07	0.07
Gujarat	0.02	0.03	0.04	0.06
Punjab	0.03	0.05	0.12	0.12
Uttar Pradesh	0.08	0.08	0.07	0.06
Assam	0.10	0.04	0.06	0.04
West Bengal	0.05	0.02	0.05	0.05

Table 6
ICC in distribution of children aged 12–23 months by whether fully vaccinated by residence in six selected states, NFHS-2 & NFHS-4.

States	NFHS-2		NFHS-4	
	Rural	Urban	Rural	Urban
Kerala	0.14	0.07	0.05	0.03
Gujarat	0.06	0.08	0.15	0.10
Punjab	0.07	0.00	0.03	0.04
Uttar Pradesh	0.07	0.07	0.15	0.12
Assam	0.09	0.0001	0.14	0.09
West Bengal	0.09	0.05	0.08	0.08

3.3. Adjustment for loss of precision due to ICC

As discussed above, the value of ICC for a majority of the variables varied from 0.05 to around 0.60. High values of ICC of a parameter have considerable implications for designing a sample survey to estimate the parameter in a population. For large-scale surveys, the general practice has been to assume a design effect of 2 for the parameter under consideration. The sample size initially estimated under the assumption of an Equal Probability of Selection Method (EPSEM) design with a given level of precision then provides the ultimate sample size (n). The inherent assumption in this is that the ICC for the variable would be around 0.05. Under such an assumption, the design effect with a cluster size of around 20 would be about 2 (equation (1)), and hence doubling the ultimate sample size would recover the loss in precision.

However, if the ICC is higher than 0.05, as is the case with many programmatic variables, the assumption of a deff of 2 would be untenable for a similar design. For example, if the ICC is 0.15, the deff would be 3.85 for a cluster size of 20 (equation (1)). The initial sample size would require almost quadrupling (n = 3.85n') to prevent a loss in precision, which will be equal to:

$$\text{Loss in precision} = \left(1 - 1 \Big/ \sqrt{deff}\right) x\ 100 = 49\%$$

There is, of course, another way to adjust for the loss of precision, which is to reduce the cluster size. For example, if the cluster size is reduced to 10 from 20, the deff would be 2.35. Considering n to be 2.35, it would reduce the loss in precision to 35%. The connection between ICC and cluster size with loss in precision of an estimate is further

illustrated in Table 7.

It can be seen that for a given cluster size, the higher the value of ICC, the higher is the loss in precision of the estimate. For example, with an ICC of 0.60 and a cluster size of 30 in a cluster design, the standard error increases more than four times and the loss of precision is to the extent of 77% compared to a design that uses SRS. In other words, the 95% confidence interval of 0.72 to 0.68 in the SRS will become 0.79 to 0.61 in a cluster design. Of course, decreasing the cluster size helps in restoring the precision to some extent. For example, if the cluster size is reduced to 10, the confidence interval shortens and varies between 0.75 and 0.65, and the loss in precision drops to 60%. If the level of ICC is 0.05 or less, as is the case in some of the demographic variables, a cluster size of 30 increases the standard error of the design only marginally.

4. Discussion

Due to the non-availability of a suitable frame for the direct selection of study units, a large-scale population-based survey usually adopts a macro-level unit, known as a cluster, to employ a probability sampling design (ICF International. (2012)). A conventional cluster sampling stops at one stage after the selection of the required number of clusters. Then, all the study units in a selected cluster are automatically included in the sample. In multistage sampling, which is generally preferred in large scale surveys, the selection of study units continues at least in two stages. Clusters are selected in the first stage also known as the PSUs. The PSUs or clusters can be any designated administrative units or units specifically prepared, like enumeration blocks (EBs) to facilitate census/surveys and used as proxies for 'neighbourhoods' or 'communities' (Roux, 2004; Montgomery & Hewett, 2005).

There is a significant advantage of studying ICC when planning a sample survey as there exists considerable clustering within a PSU in the distribution of the population by their socio-economic characteristics as people show a strong tendency to reside in the vicinity of those having similar characteristics as them. Even good communication among people within a cluster can influence their response to other behavioural and attitudinal characteristics.

This paper provides ICC estimates for a few selected socio-economic and demographic parameters for six states of India using data from NFHS-2 and NFHS-4. The analysis revealed that the ICC varied from a low of 0.07 to a high of 0.71 for variables such as religion and caste. Such a high value of ICC points towards a strong inclination of individuals to co-reside with similar socio-economic groups, particularly religious groups. The ICC values for religion are substantially high in every state. Contrary to expectation, the ICC values are higher in the urban areas with the exception of Assam and Punjab, where the ICC is higher in the rural areas. Except Kerala, there has been an increase in the extent of clustering in most of the other states over time. This is particularly evident in Gujarat and in the rural areas of Punjab. It is well known that there was a Hindu-Muslim riot in Gujarat in 2002, which may have played a role in the further concentration of the residential pattern in the state. Also, there has been a rural-urban migration among the Muslims there during NFHS-3 and NFHS-4, which too may have contributed positively to the concentration (Keshri & Bhagat, 2012).

The clustering was also found to have increased in West Bengal, especially in the rural parts. Inter-country migration in West Bengal may be one of the explanations for the high clustering in terms of religion as the state has witnessed vast amount of migration from Bangladesh (Gillan, 2002). The study also shows that the caste-wise affinity in the residential pattern was lower compared to the religious affinity. The highest caste-wise clustering was observed in Assam. While analysing the behavioural factors, we found clustering of contraception use among currently married women aged 15–49 years within a PSU. The highest clustering was observed in rural Gujarat and rural Assam at around 15% and 13% respectively. In urban Uttar Pradesh, 10% clustering was observed. It is observed that women living in the same environment typically have a similar cultural and socio-economic background, get information from similar sources, and possibly utilize health services from the same facilities, which explains the high level of clustering. In a Matlab study in Bangladesh, Munshi (2006) mentioned that women a village learnt about contraceptives from other women in the village. This study resonates with the previous literature that states that women from the same caste and religion tend to discuss family planning methods, resulting in clustering of the methods among them (Bhargava et al., 2005; Pinter et al., 2016; Sk, Jahangir, Mondal, & Biswas, 2018). A study based on the NFHS-4 data shows that contraception use can be clustered within the household as well (Ranjan et al., 2020).

While high clustering was observed for contraceptive use, clustering of anaemia within PSUs was found to be low in all the states. Clustering of immunization among children aged 12–23 months was also very high, especially in the states of Uttar Pradesh, Assam, and Gujarat. A study done in rural India suggested that the clustering effect in seeking maternal and health care services is present if the health services are scattered over a given area (Roy et al., 2016).

There is mounting evidence from various observational and intervention studies that people living within a PSU influence each other's behaviour. Several researchers (Murray & Hannan, 1990; Siddiqui et al., 1996) have studied the smoking behaviour of adolescents taking school classes as the unit of clustering (Murray & Hannan, 1990; Siddiqui et al., 1996). A study presented ICC for the vitamin A intake among children for several districts of India (Agarwal, Awasthi, & Walter, 2005). In the United States, mortality due to cardiovascular disease was presented by studying the variability in design effects (Mickey & Goodwin, 1993). A nutrition-based study analysed data from low-income countries and found clustering of nutritional status (Katz, 1995), diarrhoea (Katz, Carey, Zeger, & Sommer, 1993), cough, fever, and ocular diseases within households and villages. A study based on NFHS-4 studied the clustering of nutritional status among siblings in India (Banerjee & Dwivedi, 2020). Donner analysed data on hypertension, smoking, alcohol drinking, and body fatness and reported higher ICCs for spouse pairs and lower ICCs for counties in comparison with general practices (Donner, 1986). This reflects the need for a greater understanding of clustering in terms of socio-demographic variables within a PSU, which needs to be adjusted before providing unbiased estimates of demographic and health variables.

In summary, our study shows that the assumption of a design effect of 2 or less, which is often considered while determining the sample size in a survey, may not always be tenable, specifically if the objective of a study is to estimate a socio-economic characteristic like the religious or caste-wise distribution of a population in the country. Even for the estimation of contraceptive prevalence, the design effect will be larger than 2 unless the cluster size is minimized suitably.

In our study, we found that the values of ICC lie between 0.07 and 0.71. In case of high values of ICC, the assumption of a deff of 2 would be

Table 7
An illustration of variation in the loss of precision according to ICC and cluster size [a].

ICC	b	*deff* (from equation (1))	% loss in precision= $(1 - 1/\sqrt{deff})x\,100$	Standard error of the design	Standard error if SRS is adopted
0.6	30	18.40	77	0.04400	0.01025
	20	12.40	72	0.03609	"
	10	6.40	60	0.02593	"
0.4	30	12.60	72	0.03638	"
	20	8.60	66	0.03006	"
	10	4.60	53	0.02198	"
0.2	30	6.80	62	0.02673	"
	20	4.80	54	0.02246	"
	10	2.80	40	0.01715	"
0.05	30	2.45	36	0.01604	"
	20	1.95	28	0.01431	"
	10	1.45	17	0.01234	"

[a] Value of the parameter p = 0.7 & sample size n = 2000.

untenable for an EPSEM design. To adjust for the loss of precision in such a case, reducing the cluster size is the most effective method (Agarwal, Awasthi, & Walter, 2005; Gulliford et al., 1999; Katz, 1995; Katz & Zeger, 1994). It should also be noted that for a higher value of ICC, the loss of precision of the estimate will be more. It is clear that for a given cluster size, as the ICC increases, an increasing deff causes a more significant loss in precision. Simply increasing 'n' to regain the loss may not be an optimum solution as doing so will increase the survey cost also. An optimum solution will lie in decreasing the cluster size suitably before obtaining n. But diminishing the cluster size also leads to an increase in the number of clusters to be selected for a given sample size, which entails an increase in the fieldwork cost (Ram & Roy, 2004). This adds to the cost of household listing in the additional PSUs and the transportation cost of the field team moving to the additional PSUs.

Therefore, the decision regarding the appropriate cluster size should depend on:

a) The quality of data collection. This requires the team to visit a PSU for at least two days. This can help in minimizing the non-response bias apart from reducing the error (Ram & Roy, 2004). However, increasing the days of the visit has cost implications. First, it increases the transportation cost (a team needs to be provided with a vehicle for to-and-fro movement between their residence and the PSU).
b) The second related aspect is the formation of a field team to optimize their work. Apart from a health investigator when a biomarker is required, a field team generally consists of a few field investigators and a supervisor. It is better to have at least two investigators per team [this decision depends on the cost of vehicles per team and also on the question of safety and efficiency of a team; see (Roy et al., 2016)]. Since an investigator generally interviews around three respondents per day in a large-scale survey like NFHS, a cluster size lower than ten would tend to make the team less efficient in terms of the utilization of field workers.

Authorship contributions

Category 1.
Conception and design of study: LKD, TKR and AS.
Acquisition of data: LKD, TKR, JG and AB.
Analysis and/or interpretation of data: LKD, JG, AB, TKR.
Category 2.
Drafting the manuscript: LKD, TKR and AB.
Revising the manuscript critically for important intellectual content: LKD, AB and TKR.
Category 3.
Approval of the version of the manuscript to be published.
Laxmi Kant Dwivedi (LKD), Bidhubhusan Mahapatra (BM), Anjali Bansal (AB), Jitendra Gupta (JG), Abhishek Singh (AS) and T.K. Roy (TKR)

Financial disclosure statement

This paper was written as part of DataQi project of the Population Council funded by Bill & Melinda Gates Foundation (grant # OPP1194597).

Ethics statement

The authors used publicly available deidentified data for this analysis.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

Data will be made available on request.

Acknowledgements

This paper was written as part of DataQi project of the Population Council funded by Bill & Melinda Gates Foundation (grant # OPP1194597). The views expressed herein are those of the authors and do not necessarily reflect the official policy or position of the Bill & Melinda Gates Foundation and/or Population Council. The authors would like to acknowledge the support from National Data Quality Forum (NDQF) in providing feedback to initial draft of the paper and Ms. Roshni Subramanian for her editorial support.

The authors wish to acknowledge the valuable contribution of anonymous reviewers for their insightful comments and suggestions which have helped improve this research work tremendously.

References

Agarwal, G. G., Awasthi, S., & Walter, S. D. (2005). Intra-class correlation estimates for assessment of vitamin A intake in children. *Journal of Health, Population and Nutrition*, 66–73. https://www.ncbi.nlm.nih.gov/pubmed/15884754.

Agrawal, S., Millett, C. J., Dhillon, P. K., Subramanian, S. V., & Ebrahim, S. (2014). Type of vegetarian diet, obesity and diabetes in adult Indian population. *Nutrition Journal, 13*(1), 1–18. https://doi.org/10.1186/1475-2891-13-89

Alegana, V. A., Wright, J., Bosco, C., Okiro, E. A., Atkinson, P. M., Snow, R. W., … Noor, A. M. (2017). Malaria prevalence metrics in low-and middle-income countries: an assessment of precision in nationally-representative surveys. *Malaria Journal, 16*(1), 1–11. https://doi.org/10.1186/s12936-017-2127-y

Banerjee, K., & Dwivedi, L. K. (2020). Linkage in stunting status of siblings: a new perspective on childhood undernutrition in India. *Journal of Biosocial Science, 52*(5), 681–695. https://doi.org/10.1017/S0021932019000725

Benefo, K. D. (2006). The community-level effects of women's education on reproductive behavior in rural Ghana. *Demographic Research, 14*, 485–508.

Bhargava, A., Chowdhury, S., & Singh, K. K. (2005). Healthcare infrastructure, contraceptive use and infant mortality in Uttar Pradesh, India. *Economics and Human Biology, 3*(3), 388–404.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review/Revue Internationale de Statistique*, 67–82.

Donner, A., & Klar, N. (1994). Cluster randomization trials in epidemiology: Theory and application. *Journal of Statistical Planning and Inference, 42*(1–2), 37–56.

Eldridge, S. M., Ashby, D., & Kerry, S. (2006). Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International journal of epidemiology, 35*(5), 1292–1300. https://doi.org/10.1093/ije/dyl129

Eldridge, S. M., Ukoumunne, O. C., & Carlin, J. B. (2009). The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *International Statistical Review, 77*(3), 378–394.

Gillan, M. (2002). Refugees or infiltrators? The bharatiya janata party and "illegal" migration from Bangladesh. *Asian Studies Review, 26*, 73–95.

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: Data from the health survey for england 1994. *American Journal of Epidemiology, 149*(9), 876–883. https://doi.org/10.1093/oxfordjournals.aje.a009904

Johnson, J. L., Kreidler, S. M., Catellier, D. J., Murray, D. M., Muller, K. E., & Glueck, D. H. (2015). Recommendations for choosing an analysis method that controls Type I error for unbalanced cluster sample designs with Gaussian outcomes. *Statistics in Medicine, 34*(27), 3531–3545.

Kalton, G., Brick, J. M., & Lê, T. (2005). Estimating components of design effects for use in sample design. *Household Sample Surveys in Developing and Transition Countries*.

Katz, J. (1995). Sample-size implications for population-based cluster surveys of nutritional status. *American Journal of Clinical Nutrition, 61*(1), 155–160. https://doi.org/10.1093/ajcn/61.1.155

Katz, J., Carey, V. J., Zeger, S. L., & Sommer, A. (1993). Estimation of design effects and diarrhea clustering within households and villages. *American Journal of Epidemiology, 138*(11), 994–1006.

Katz, J., & Zeger, S. L. (1994). Estimation of design effects in cluster surveys. *Annals of Epidemiology, 4*(4), 295–301. https://doi.org/10.1016/1047-2797(94)90085-x

Kerry, S. M., & Bland, J. M. (1998). The intracluster correlation coefficient in cluster randomisation. *BMJ, 316*(7142), 1455. https://doi.org/10.1136/bmj.316.7142.1455

Keshri, K., & Bhagat, R. B. (2012). Temporary and seasonal migration: Regional pattern, characteristics and associated factors. *Economic and Political Weekly*, 81–88.

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *The Annals of Family Medicine, 2*(3), 204–208.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons, Inc.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics, 11*(1), pp 55–57.

Kravdal, Ø. (2002). Education and fertility in sub-saharan Africa: Individual and community effects. *Demography, 39*(2), 233–250. https://doi.org/10.1353/dem.2002.0017

Kravdal, Ø. (2007). A fixed-effects multilevel analysis of how community family structure affects individual mortality in Norway. *Demography, 44*(3), 519–537.

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation - a discussion and demonstration of basic features. *PLoS One, 14*(7), Article e0219854. https://doi.org/10.1371/journal.pone.0219854

Mickey, R. M., & Goodwin, G. D. (1993). The magnitude and variability of design effects for community intervention studies. *American Journal of Epidemiology, 137*(1), 9–18. https://doi.org/10.1093/oxfordjournals.aje.a116606

Montgomery, M. R., & Hewett, P. C. (2005). Urban poverty and health in developing countries: Household and neighborhood effects. *Demography, 42*(3), 397–425. https://doi.org/10.1353/dem.2005.0020

Munshi, K., & Myaux, J. (2006). Social norms and the fertility transition. *Journal of Development Economics, 80*(1), 1–38.

Murray, D. M., & Hannan, P. J. (1990). Planning for the appropriate analysis in school-based drug-use prevention studies. *Journal of Consulting and Clinical Psychology, 58*(4), 458–468. https://doi.org/10.1037//0022-006x.58.4.458

Pagel, C., Prost, A., Lewycka, S., Das, S., Colbourn, T., Mahapatra, R., Azad, K., Costello, A., & Osrin, D. (2011). Intracluster correlation coefficients and coefficients of variation for perinatal outcomes from five cluster-randomised controlled trials in low and middle-income countries: Results and methodological implications. *Trials, 12*, 151. https://doi.org/10.1186/1745-6215-12-151

Pinter, B., Hakim, M., Seidman, D. S., Kubba, A., Kishen, M., & Di Carlo, C. (2016). Religion and family planning. *The European Journal of Contraception and Reproductive Health Care, 21*(6), 486–495. https://doi.org/10.1080/13625187.2016.1237631

Ram, F., & Roy, T. K. (2004). *Comparability issues in large sample surveys-some observations. In T. K. Roy , M. Guruswamy & P. Arokiasamy (Eds.), Population, Health and Development in India: Changing Perspectives (pp. 40-56)*. Rawat Publications.

Ranjan, M., Mozumdar, A., Acharya, R., Mondal, S. K., & Saggurti, N. (2020). *Intrahousehold influence on contraceptive use among married Indian women: Evidence from the National Family Health Survey 2015–16, 11*. SSM-Population Health.

Roux, A. V. D. (2004). Estimating neighborhood health effects: the challenges of causal inference in a complex world. *Social science & medicine, 58*(10), 1953–1960. https://doi.org/10.1016/S0277-9536(03)00414-3

Roy, T. K., Acharya, R., & Roy, A. K. (2016). *Statistical survey design and evaluating impact*. Cambridge:Cambridge University Press.

Roy, T. K., & Pandey, A. (2008). National family health survey in India: Survey design and related issues. *Demography India, 37*(Supplement), 13–41.

Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology, 44*(3), 1051–1067. https://doi.org/10.1093/ije/dyv113

Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Siddiqui, O., Hedeker, D., Flay, B. R., & Hu, F. B. (1996). Intraclass correlation estimates in a school-based smoking prevention study. Outcome and mediating variables, by sex and ethnicity. *American Journal of Epidemiology, 144*(4), 425–433. https://doi.org/10.1093/oxfordjournals.aje.a008945

Simpson, J. M., Klar, N., & Donnor, A. (1995). Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *American Journal of Public Health, 85*(10), 1378–1383. https://doi.org/10.2105/ajph.85.10.1378

Sk, M. I. K., Jahangir, S., Mondal, N. A., & Biswas, A. B. (2018). Disparities in the contraceptive use among currently married women in Muslim densely populated States of India: An evidence from the nationally representative survey. *Epidemiology, Biostatistics and Public Health, 15*(3).

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed. ed.). Iowa State University Press.

ICF International. (2012). *Demographic and Health Survey Sampling and Household Listing Manual. MEASURE DHS*. Calverton, Maryland, U.S.A: ICF International.

Aliaga, A., & Ren, R. (2006). Optimal sample sizes for two-stage cluster sampling in demographic and health surveys. DHS Working Papers No. 30. Calverton, Maryland, USA: ORC Macro. Available at http://dhsprogram.com/pubs/pdf/WP30/WP30.pdf.

Park, I., & Lee, H. (2001). The design effect: Do we know all about it? Proceedings of the annual meeting of the American *Statistical Association*, August 5-9,2001.