An Efficient Sampling Approach to Surveillance of Non-Communicable Disease Risk Factors in Cienfuegos, Cuba

Luis Carlos Silva PhD DrSc, Mikhail Benet MD PhD, Alain Morejón MD, Pedro Ordúñez MD PhD

ABSTRACT

One of the most common shortcomings in non-communicable disease risk factor surveillance, especially in prevalence studies, is sampling procedure, which can and does compromise accuracy and reliability of derived estimates. Moreover, sampling consumes significant time and resources. Since the early 1990s, risk factor surveys in Cienfuegos province, Cuba have paid particular attention to careful sampling methods. The new survey conducted in 2011 was not only statistically rigorous but introduced an innovative, more efficient method. This article provides a detailed description of the sample design employed to optimize resource use without compromising selection rigor.

KEYWORDS Sampling, descriptive studies, risk factors, non-communicable diseases, hypertension, Cuba

INTRODUCTION

Basic science experimentation and clinical trials are the most robust sources of scientific evidence in medicine and health. Descriptive or observational studies, although lacking explanatory capacity per se, are also a legitimate type of research—moreover, one that is essential in epidemiology and public health for learning about health situations and shedding light on health system performance.[1]

Many efforts have been made recently to maximize quality of observational studies, significant among them the Strengthening the Reporting of Observational Studies in Epidemiology Statement (STROBE).[2] According to STROBE, prevalence studies—also known as cross-sectional studies—play a key role, especially when conducted rigorously enough to transcend merely quantitative aspects. Then, they permit evidence-based characterization of what is occurring in a population in a particular context in space and time, as well as development of useful judgments for optimizing health interventions by various actors.

Sampling procedures play a key role in observational studies; indeed, they are the bedrock of study quality. Sampling is one of the greatest Achilles' heels in non-communicable chronic disease (NCD) risk factor surveillance, essential for NCD prevention and control.[3] Yet, a 2001 review of surveillance studies in the Americas,[4] based on a tool specifically created for assessing prevalence studies,[5] yielded disturbing results regarding sample quality.

This assessment tool was fairly comprehensive, covering 19 points related to various technical areas: stated objectives, sampling design, data collection and processing methods, communication of results, etc. Four sections focused directly on sampling itself.

The 2001 surveillance study review involved a literature search in three databases (PubMed, Medline, BIREME) for articles on hypertension prevalence in Latin American and Caribbean countries published in the two preceding decades, yielding 58 articles, 48 (83%) of which appeared after 1990. Sampling design was not explained in 26% of articles; in 31%, sampling was not probabilistic; in 74%, formulas used for calculation of point estimates were inconsistent with design; and in 90%, errors and confidence intervals were not computed in accordance with sample design. In fact, only one of the 58 articles met all quality criteria, and ten met none.

This study was replicated in 2012[6] and, while it did not include a detailed analysis of sampling quality, its results suggest little progress in this regard. In fact, over one third of reports published in the past ten years lack information on sampling error associated with hypertension prevalence estimates.

Conventional techniques to meet the demands of sampling theory traditionally have required highly complex and costly fieldwork. Bearing this in mind, as well as the sampling problems outlined above, researchers in the city of Cienfuegos, Cuba, developed an innovative sampling method for the third Cienfuegos NCD risk factor study in 2011, aimed at achieving an approach both rigorous and efficient.

When the first two cross-sectional surveys were conducted in Cienfuegos (1991 and 2001), statistically rigorous but thoroughly conventional sampling designs were used. The sampling procedure in 2011 was considerably more efficient, especially because it minimized the number of household visits required, saving time and resources without sacrificing classical criteria for rigorous sampling design, such as probabilistic selection. This sampling innovation was created by one of the authors of this study (LCS) and was applied successfully in the Cardiovascular Risk Factor Multiple Evaluation in Latin America (CARMELA) study.[7] This paper reporting on its replication in Cienfuegos is the first published description of the technical details of the solution.

The purpose of this article is to outline the different alternatives considered for selecting the sampling procedure for such descriptive studies and, in particular, to describe in detail the procedure used and its results concerning sampling.

IMPLEMENTATION

Context: NCD risk factor surveillance in Cienfuegos, Cuba
Situated in the center of the island, with a population of nearly 150,000, Cienfuegos is one of Cuba's most important cities. Population-based NCD surveillance, particularly estimation of prevalence of main NCD risk factors, was first conducted there in 1991 under the general framework of the Cienfuegos Global Project.[8] Two more surveys followed, one in 2001 and another in 2011, the subject of this article. This latter study was conducted under the aegis of the CARMEN multi-country studies, a PAHO initiative for a multifactoral approach to NCD risk factors—Cienfuegos was designated the demonstration site for Cuba.[9] In fact, Cienfuegos is the only Cuban city that has systematically conducted

population-based surveillance of risk factor prevalence since the early 1990s, to inform programming for risk factor reduction. The studies were approved by the Medical University of Cienfuegos ethics committee.

A distinguishing features of NCD risk factor surveillance in Cienfuegos—permitting reliable, accurate estimates—is its careful sampling design process. This was documented in the first large-scale report of national significance, published in 1993;[10] and the essential results of the 1991 and 2001 surveys were published internationally.[11,12] These Cienfuegos studies, structured around an open-access instrument, also exceeded quality standards for NCD surveillance, as evidenced in the 2001 and 2012 reviews of surveillance studies.[4,6]

The general sampling problem The study's target population consisted of residents aged 15–74 in the urban zone of the Municipality of Cienfuegos. Studies of this type require analysis disaggregated by age group and sex, which in turn requires sufficiently large sample sizes for each of the 12 conventional groups: six ten-year age segments (15–24, 25–34, 35–44, 45–54, 55–64 and 65–74 years) and subgroups by sex for each age segment. The minimum desirable sample size calculated for each group was 180–200 subjects.[13]

Since, as is the case nearly everywhere, the Cienfuegos population pyramid is far from uniform, the general sample cannot be equiprobabilistic. Nevertheless, all subsamples in the respective groups should meet that criterion. In order to achieve sample sizes sufficient for more detailed analysis and assuming an appreciable number of nonresponses, we needed to select 240 subjects in each of the aforementioned sex and age groups.

Sampling design Cuba's National Statistics Office has a master sample that is regularly used for a wide range of purposes.[14] For this third Cienfuegos survey, a multistage sample was designed with its first three stages based on the master sample. The sampling units in the master sample are census districts (first-stage or primary sampling units), areas within the districts (second-stage units) and, finally, census sections (third-stage units) within the areas. Each census section consists of approximately five contiguous dwellings.

According to 2002 census data,[15] the smallest group was that of men aged 65–74 (4% of the population) and each dwelling housed an average of 2.5 adults in the target population, so it was calculated that to obtain 240 subjects in that group, 2400 dwellings would be required, finally including all subjects in that group in the sample. The other 11 groups would each need a minimum of 240 subjects. Thus, the set of 2400 dwellings would be sufficient to obtain the entire sample.

On this basis, it was calculated that some 500 census sections in the population would be needed. The selection procedure followed in this initial phase ultimately yielded an equiprobabilistic sample of 511 sections; the probability of selection in the master sample for each of these sections was 0.058.

Using this initial sample, a two-stage selection process was used for the purposes of our study, selecting dwellings first and then the subjects themselves. The sections selected contained 2540 dwellings, from which the necessary 2400 were selected at random. At this point we had an equiprobabilistic sample of dwellings. As for the subjects residing in these dwellings, the respective individuals (approximately 240) were selected for each of the groups using a probabilistic method.

The conventional method for handling this process would have been to first conduct a census of the 2400 dwellings to learn their composition, draw up separate lists for each of the 12 sex and age groups and, finally, to randomly or systematically select 240 subjects from each of the lists. However, this theoretically simple procedure is extremely complicated in practice; the idea was to avoid conducting such a costly initial census and save the time required to make two visits to each household—the first to determine the age and sex of their residents and compile the lists, and the second to interview those selected.

The approach used was devised by one of the authors of this article and, as far as we know, is original; broadly speaking, it involved first dividing the 2400 dwellings initially selected into 12 categories, randomly distributing the dwellings in these categories, whose sizes were determined by a system of equations. At this point, a sample of eligible people was selected in each dwelling. From each of these dwellings, subjects belonging to certain sex and age groups were selected, based on a rule established for each of the categories. For example, if the dwelling was in category 1, all eligible men were included (men aged 25–74); if the dwelling was in category 4, all women except those aged 35–44 were included, etc. This is only a general idea of how the mechanism operates. Fuller understanding of the procedure and its conceptual underpinnings can be obtained in the Appendix, describing a general solution to the problem, which was applied to the 12 sex and age groups corresponding to our case.

RESULTS

Weighting and corroboration of representativeness Having made the selection, the probability of inclusion was calculated for each person in the sample and, through its inverse, the weighting to use for overall estimates. Information was gathered in a two-step process. The first involved completing the general form; 2193 people gave informed consent and participated (933 men and 1260 women), distributed by age group and sex, as seen in Table 1, which shows that the desired sizes (180–200 subjects in each group) were obtained in most cells.

The second step involved physical measures (anthropometric, blood pressure and laboratory). A total of 1496 people (616 men and 880 women) were recruited and gave signed consent (Table 1). As can be seen, the second step had a high nonresponse rate: only 68.2% of subjects originally recruited appeared for physical measures (1496/2193). In such circumstances, possible differential impact of sample attrition must be analyzed.

The representativeness of the sample obtained was gauged through estimates of parameters for which census data were available. Results were highly satisfactory. For example, the 2002 national census, the last prior to the study, revealed that black and mestizo persons accounted for 27.9% of the population in the city of Cienfuegos. Dividing the sum of the weightings for all such persons by the sum of the weightings for all subjects in the sample yielded a ratio of 0.279, which coincides exactly with the census figure. Similar results were obtained for sex and educational level.

Table 1: Age and sex distribution of sample subjects

	Male			Female		
Age group	All subjects n (%)	Interview only n (%)	Interview and physical measures n (%)	All subjects n (%)	Interview Only n (%)	Interview and physical measures n (%)
15–24	163 (17.5)	67 (21.1)	96 (15.6)	192 (15.2)	74 (19.5)	118 (13.4)
25–34	131 (14.0)	63 (19.9)	68 (11.0)	172 (13.7)	70 (18.4)	102 (11.6)
35–44	176 (18.9)	65 (20.5)	111 (18.0)	250 (19.8)	76 (20.0)	172 (19.5)
45–54	195 (20.9)	58 (18.3)	137 (22.2)	234 (18.6)	72 (18.9)	162 (18.4)
55–64	146 (14.6)	41 (12.9)	105 (17.0)	206 (16.3)	46 (12.1)	160 (18.2)
65–74	122 (13.1)	33 (10.4)	99 (16.1)	206 (16.3)	40 (10.5)	166 (18.9)
Total	933 (100)	317 (100)	616 (100)	1260 (100)	380 (100)	880 (100)

Comparison of interviewed subjects who had physical measurements with those who did not found no significant differences between the two (at threshold of p = 0.05) by sex or skin color (Table 2). However, age structure did differ between the two subsamples, as can be seen in Table 1. In fact, particularly marked differences can be observed at age extremes. For example, the group aged 15–24 years represents 14.3% of people for whom measurements were taken (214 out of 1496) and 20.2% (141 out of 697) of those for whom they were not. Among people over 64, the opposite held true: this group accounted for 17.7% (265 out of 1496) and 10.5% (73 out of 697) respectively. The difference in age distribution between subgroups with and without physical measurements was statistically significant (p <0.001).

Table 2: Sex and skin color distribution (%) in subsamples with and without physical measures

	Physical measures		
	Yes	No	All
Sex			
Male	616 (41.2%)	317 (45.5%)	933 (42.5%)[a]
Female	880 (58.8%)	380 (54.5%)	1260 (57.5%)
Skin color			
White	1077 (72.0%)	507 (72.7%)	1584 (72.2%)[b]
Black or mestizo	419 (28.0%)	190 (27.3%)	609 (27.8%)
Total	1496 (100 %)	697 (100%)	2193 (100 %)

[a] $\chi^2(1) = 3.60$, p = 0.06
[b] $\chi^2(1) = 0.13$, p = 0.72

In principle, this result constitutes a study weakness, since it could bias results for hypertension. While it is impossible with available data to determine the degree of potential bias, a complementary analysis was done that consisted of estimating some prevalence rates related to variables measured in the first step (physical activity, smoking and educational level), independently using two samples: one consisting of subjects for whom measurements were taken in the second step and the other, of subjects for whom information was derived from self-report at interview in the first step.

The estimates were very similar [data not shown] for three parameters, suggesting that differences in age composition in these two subsamples might not affect overall results related to anthropometry and other variables not clearly related to age. However, blood pressure figures were likely overestimated to some degree, due to overrepresentation of older subjects in the group that appeared for measurement of this variable. In fact, the average age in this second group was approximately 46 years, while in the first group, it was only 41. Consequently, overall estimated hypertension prevalence for Cienfuegos in this study could be somewhat higher than true prevalence.

LESSONS LEARNED

The third survey of NCD risk factors in Cienfuegos (2011) was conducted using a probabilistic (not self-weighted) sample obtained through a five-stage selection process: districts, areas, sections, dwellings and subjects. The first three stages employed the procedures used for the master sample in the selection process, and the last two were specific to this study, using a novel approach that greatly facilitated field work and yielded substantial resource savings without compromising the probabilistic nature of the general sampling procedure.

The novelty of the sampling procedure used calls for careful study of its features. The detailed explanation of the method used in these stages is a resource that can be used in future to solve one of the most frequent problems in this type of research. The statistical rigor and efficiency of the procedure make it a useful tool for improving accuracy and reliability of NCD risk factor prevalence estimates in and beyond Cienfuegos.

REFERENCES

1. Grimes DA, Schulz KF. Descriptive studies: what they can and cannot do. Lancet. 2002 Jan 12;359(9301):145–9.
2. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. Ann Intern Med. 2007 Oct 16;147(8):573–7.
3. Action plan for global strategy for the prevention and control of non-communicable diseases. Geneva: World Health Organization; 2008.
4. Ordúñez P, Silva LC, Rodríguez P, Robles S. Prevalence estimates for hypertension in Latin America and the Caribbean: are they useful for surveillance? Rev Panam Salud Pública. 2001;10(4):226–31.
5. Silva LC, Ordúñez P, Rodríguez P, Robles S. A tool for assessing the usefulness of prevalence studies done for surveillance purposes: the example of hypertension. Rev Panam Salud Pública. 2001 Sep;10(3):152–60.
6. Burroughs Peña MS, Abdala CVM, Silva LC, Ordúñez P. Usefulness for surveillance of hypertension prevalence studies in Latin America and the Caribbean: the past 10 years. Rev Panam Salud Pública. 2012 Jul;32(1):15–21.
7. Schargrodsky H, Hernández R, Champagne BM, Silva H, Vinueza R, Silva LC, et al. CARMELA: Assessment of Cardiovascular Risk in Seven Latin American Cities. Am J Med. 2008 Jan;121(1):58–65.
8. Espinosa A, Ordúñez P. El Proyecto Global de Cienfuegos. Nuevas perspectivas para la salud de los cienfuegueros. Rev Finlay. 1991;5(4):379–84. Spanish.
9. CARMEN Network [Internet]. Washington D.C.: Pan American Health Organization; c1998-2012 [updated 2011 Oct 05; cited 2012 Jul 30]. Available from: http://new.paho.org/hq/index.php?option=com_content&task=blogcategory&id=1412&Itemid=1854
10. Ordúñez P, Espinosa A, Alvarez O, Apolinaire JJ, Silva LC. Marcadores múltiples de riesgo para enfermedades crónicas no transmisibles. Medición inicial del proyecto global de Cienfuegos 1991–1992. Havana: Editorial Ciencias Médicas; 1993. p 82. Spanish.
11. Ordúñez P, Espinosa A, Cooper R, Kaufman JS, Nieto FJ. Hypertension in Cuba: evidence of narrow black-white difference. J Hum Hypertens. 1998 Feb;12(2):111–6.
12. Ordúnez P, Bernal JL, Espinosa A, Silva LC, Cooper RS. Ethnicity, education and blood pressure in Cuba. Am J Epidemiol. 2005 Jul 1;162(1):49–56.
13. WHO STEPS surveillance manual: the WHO STEP wise approach to chronic disease risk factor surveillance / Noncommunicable Diseases and Mental Health. [Internet]. Geneva: World Health Organization; 2007 [cited 2012 Jul 30]. Available from: http://www.who.int/chp/steps/resources/updates/en/index.html

14. Center for Population and Development Studies (CU). Encuesta sobre indicadores de prevención de infección por el VIH/SIDA, 2009. Havana: National Statistics Bureau (CU); 2011. 171 p. Spanish.
15. Censo de población y viviendas. Cuba 2002 [Internet]. Havana: National Statistics Bureau (CU); 2002 [cited 2012 Jul 30]. Available from: http://www.cubagob.cu/otras_info/censo/index.htm. Spanish.

AUTHORS

Luis Carlos Silva Ayçaguer (Corresponding author: lcsilva@infomed.sld.cu), mathematician and biostatistician. Full professor and senior researcher, National Medical Sciences Information Center, Havana, Cuba.

Mikhail Benet Rodríguez, physiologist with a doctorate in health sciences, Medical University of Cienfuegos, Cuba.

Alain Morejón Giraldoni, internist, Medical University of Cienfuegos, Cuba.

Pedro O. Ordúñez García, internist specializing in public health with a doctorate in health sciences. Advisor, chronic disease prevention and control, PAHO, Washington, DC.

Submitted: August 13, 2012
Approved for publication: October 12, 2012
Disclosures: None

APPENDIX

General Theoretical Solution for Two-stage Selection of Subjects Based on a Cluster of Dwellings

Assume that the population is made up of *k* sex and age groups, which will be identified as $G_1, G_2, \cdots, G_k$. The equation $f_h = \frac{N_h}{N}$ is defined, where N_h is the size of group h (h:1, ···, k) and the total population is $N = \sum_{h=1}^{k} N_h$.

Let the total number of dwellings in which these N subjects reside be called *V,* and the average number of subjects per dwelling in the population η_p ($\eta_p = \frac{N}{V}$). Let the number of dwellings that would have to be selected in the first stage and would provide the framework for selecting the subjects to be studied in the second stage be called η. η would be determined by the necessary minimum number of dwellings containing at least *m* subjects in each sex and age group. The aim is to develop a sampling procedure that makes it possible to select m subjects from each of the groups and meets the following criteria:

Subject selection should be probabilistic, and the use of chance to meet this requirement should not be applied in the dwelling itself, but rather, before it is visited, so that once inside the dwelling, it is already known which subjects should be included in the sample, even though the specific composition of the dwelling's residents by age and sex is unknown.

Conducting a census in the η dwellings beforehand to learn the composition by groups is to be avoided; thus, it is unnecessary to have any data other than values for η_p, f_h and m.

In order to solve the problem, assume that the groups are organized such that $f_h \leq f_{h+1}$ for all h:1,2,···,k−1. Or to put it another way, call the group corresponding to the smallest of the f_h G_1, the one for which the fraction is smallest with the exception of f_1 G_2, and so forth, until G_k, which will be the group whose f_k fraction is the largest of all.

The proposed procedure involves the following 4 steps:

- Determine the number η of dwellings in the sample
- Select the dwellings using a probabilistic method
- Randomly divide the set of dwellings into *k* classes; if η_h is the number of dwellings in group h, we would have $\eta = \sum_{h=1}^{k} \eta_h$
- Next, select subjects from each dwelling according to the following rule: in the dwellings in the h^th^ class, all individuals in groups $G_1, G_2, \cdots, G_h$ would be selected (that is, the sample would include only subjects in that dwelling belonging to the first h groups).

The minimum number of necessary dwellings is determined by the frequency corresponding to group G_1. That is, a sufficient number of dwellings should be selected to ensure that the expected value for the total individuals residing in them that belong to the group with the lowest frequency in the population is equal to m. In formal terms, this is equivalent to meeting the condition $m = f_1 \eta_p \eta$, from which it is deduced that the minimum number is $\eta = \frac{m}{\eta_p f_1}$.

In order to determine the $\eta_1, \eta_2, \cdots, \eta_k$ values, point iv. above must be borne in mind; that is, it must be ensured that η_h dwellings in class h contain only individuals belonging to groups $G_1, G_2, \cdots, G_h$. Let n_{hj} be the number of subjects in group G_j furnished by the dwellings in class h. Let's then solve the system of k equations:

$m = \sum_{h=j}^{k} n_{hj}$ (for h =1, 2, ···, k) with the η_h as unknowns.

Bearing in mind that

$n_{hj} = \eta_h \eta_p f_j$, it can be seen that the solution is:

$$\eta_h = \frac{m}{\eta_p}\left(\frac{1}{f_h} - \frac{1}{f_{h+1}}\right) \text{ for h:1, 2, ···, k−1 and } \eta_k = \frac{m}{\eta_p f_k}$$

Knowing the values of η_h, we now proceed to randomly divide the set of dwellings into k classes, and to proceed in each dwelling as appropriate, depending on the class to which it belongs.