Sampling design for the *Study of Cardiovascular Risks in Adolescents* (ERICA)

Desenho da amostra do *Estudo de Riscos Cardiovasculares em Adolescentes* (ERICA)

Diseño de la muestra del *Estudio de Riesgos Cardiovasculares en Adolescentes* (ERICA)

Mauricio Teixeira Leite de Vasconcellos [1]
Pedro Luis do Nascimento Silva [1]
Moyses Szklo [2]
Maria Cristina Caetano Kuschnir [3]
Carlos Henrique Klein [4]
Gabriela de Azevedo Abreu [2]
Laura Augusta Barufaldi [2]
Katia Vergetti Bloch [2]

[1] *Escola Nacional de Ciências Estatísticas, Fundação Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, Brasil.*
[2] *Instituto de Estudos em Saúde Coletiva, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.*
[3] *Núcleo de Estudos da Saúde do Adolescente, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil.*
[4] *Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.*

Correspondence
M. T. L. Vasconcellos
Escola Nacional de Ciências Estatísticas, Fundação Instituto Brasileiro de Geografia e Estatística.
Rua André Cavalcanti 106, Rio de Janeiro, RJ
20231-050, Brasil.
mautlv@gmail.com

Abstract

The Study of Cardiovascular Risk in Adolescents *(ERICA) aims to estimate the prevalence of cardiovascular risk factors and metabolic syndrome in adolescents (12-17 years) enrolled in public and private schools of the 273 municipalities with over 100,000 inhabitants in Brazil. The study population was stratified into 32 geographical strata (27 capitals and five sets with other municipalities in each macro-region of the country) and a sample of 1,251 schools was selected with probability proportional to size. In each school three combinations of shift (morning and afternoon) and grade were selected, and within each of these combinations, one class was selected. All eligible students in the selected classes were included in the study. The design sampling weights were calculated by the product of the reciprocals of the inclusion probabilities in each sampling stage, and were later calibrated considering the projections of the numbers of adolescents enrolled in schools located in the geographical strata by sex and age.*

Cardiovascular Diseases; Statistical Models; Sampling Studies; Adolescent

Resumo

O Estudo de Riscos Cardiovasculares em Adolescentes *(ERICA) objetiva estimar prevalência de fatores de risco cardiovascular e da síndrome metabólica em adolescentes (12 a 17 anos) matriculados em escolas públicas e privadas dos 273 municípios com mais de 100 mil habitantes no Brasil. A população de pesquisa foi estratificada em 32 estratos geográficos (27 capitais e cinco conjuntos com os demais municípios de cada macrorregião do país) e uma amostra de 1.251 escolas foi selecionada com probabilidade proporcional ao tamanho. Em cada escola foram selecionadas três combinações de turno (manhã e da tarde) e ano (série), e em cada uma destas combinações foi selecionada uma turma. Todos os alunos elegíveis das turmas selecionadas foram objeto de pesquisa. Os pesos amostrais do desenho foram calculados pelo produto dos inversos das probabilidades de inclusão em cada estágio da amostra e foram depois calibrados considerando as projeções do número de adolescentes matriculados em escolas localizadas nos estratos geográficos considerados por sexo e idade.*

Doenças Cardiovasculares; Modelos Estatísticos; Amostragem; Adolescente

http://dx.doi.org/10.1590/0102-311X00043214

Introduction

The *Study of Cardiovascular Risks in Adolescents* (ERICA) would in principle aim to provide national estimates about the prevalence of cardiovascular risk factors and metabolic syndrome by sex and age in adolescents aged 12 to 17 years.

To that end, it would be necessary to base the study on a national household sample, so that information could be collected in the subset of households with adolescents. However, considering the prohibitive cost of such an operation, it was decided to carry out the survey via schools and to limit the target population to adolescents enrolled in the last three grades of Basic Education and in all three grades of Secondary (High-School) Education of public or private schools in municipalities with population of over 100,000 inhabitants, grouped in 32 geographic strata described in this article.

A study based on a geographically stratified sample of students, clustered by school, shift and grade, and class, in addition to allowing inference to the set of most populated cities of Brazil, and to geographic strata that allow visualization of regional differences within the country was the solution to match ERICA's estimated costs to the available budget. It was thus decided to select the sample in three stages: schools, combinations of shift and grade, and classes. In the selected classes, all eligible students were invited to take part in the survey.

This article describes the study population and its geographic stratification; the calculation of the sample size and its allocation to the strata; the methods for the selection of schools, combinations of shift and grade, and classes; the treatment of non-response among the adolescents; calculation and calibration of sampling weights; and the estimation techniques recommended for analyzing the survey data.

Study population

The study population [1] includes the set of 12 to 17-year-old adolescents, with no temporary or permanent disability, enrolled in one of the last three grades of Basic Education or of the three grades of Secondary Education (High-School), in the morning or afternoon shifts, in public or private schools located in one of the 273 municipalities with population of over 100,000 inhabitants on July 1st, 2009 (the most recent population data available at the time ERICA's basic definitions were made).

The characterization of the study population is based on a file provided by the Anísio Teixeira National Institute of Educational Studies and Research (INEP/MEC), which was developed from data of the *2011 Educational Census*, (whose microdata are available at ftp://ftp.inep.gov.br/microdados/micro_censo_escolar_2011.zip, accessed on 16/Feb/2012), since this was used to select the combinations of shift and grade in each sampled school.

Considering students who are not lagging behind, adolescents in the 12 to 17 year-old range are expected to be enrolled in classes of one of the last three grades of Basic Education (grades 7th to 9th) or one of the three grades of Secondary Education, which were therefore defined as the eligible grades. Thus, from a total of 237,438 schools covered by the *2011 Educational Census*, there were 389,315 combinations of eligible shifts (morning or afternoon) and grades.

However, 63,912 of these combinations were related to remedial classes (8,114), non-serialized or mixed grade classes in Basic Education (54,973), and non-serialized classes in Secondary Education (825), which were discarded for not having a clear association with the eligible grades, and therefore with the target adolescent age range. In addition, 51,953 combinations were related to the evening shift (10,559 in Basic Education; 41,394 in Secondary Education), and were ruled out for operational reasons. Thus, in the whole country, there were 273,450 eligible combinations of shift and grade in 61,325 schools.

When considering only the schools located in the 273 municipalities with a population of over 100,000 inhabitants on July 1st, 2009, the number of combinations dropped to 117,726, and the number of schools to 24,441.

Table 1 shows the totals of schools, classes and students of the eligible shifts and grades, according to the *2011 Educational Census*, in the country and in the 273 municipalities considered (those with population over 100,000 inhabitants). It shows that limiting the geographic coverage of the study to these 273 municipalities accounts for coverage of a little over 54% (8.2 million) of the Brazilian population of adolescents enrolled in schools (15.3 million) in the eligible shifts and grades.

Stratification of the study population

The study population was stratified in 32 geographic strata, one for each of the 27 state capitals, and five strata with the schools in the set of municipalities with population over 100,000 in each of Brazil's five macro-regions.

The distribution of the study population over the geographic strata is presented, abridged, in Table 2. It shows that 44% of the students of the

Table 1

Number of schools, classes and students per urban or rural area according to the geographic level and the administrative dependency of the school.

Geographic level and		Schools			Classes			Students	
administrative dependency	Total	Urban	Rural	Total	Urban	Rural	Total	Urban	Rural
Total of the country	61,325	46,675	14,650	516,701	463,128	53,573	15,233,991	14,119,065	1,114,926
Public	48,440	33,955	14,485	431,969	379,153	52,816	12,960,385	11,862,698	1,097,687
Private	12,885	12,720	165	84,732	83,975	757	2,273,606	2,256,367	17,239
273 municipalities	24,441	22,698	1,743	261,086	253,051	8,035	8,273,275	8,076,757	196,518
Public	15,942	14,244	1,698	198,266	190,477	7,789	6,492,228	6,301,417	190,811
Private	8,499	8,454	45	62,820	62,574	246	1,781,047	1,775,340	5,707

Table 2

Distribution of the number of schools, classes and students per urban or rural location according to the group of geographic strata and administrative dependency of the school.

Group of strata and		Schools			Classes			Students	
administrative dependency	Total	Urban	Rural	Total	Urban	Rural	Total	Urban	Rural
Total (273 municipalities)	24,441	22,698	1,743	261,086	253,051	8,035	8,273,275	8,076,757	196,518
Public	15,942	14,244	1,698	198,266	190,477	7,789	6,492,228	6,301,417	190,811
Private	8,499	8,454	45	62,820	62,574	246	1,781,047	1,775,340	5,707
Capitals (27 strata)	9,771	9,520	251	111,591	110,292	1,299	3,616,486	3,583,046	33,440
Public	5,656	5,416	240	78,814	77,566	1,248	2,668,859	2,636,762	32,097
Private	4,115	4,104	11	32,777	32,726	51	947,627	946,284	1,343
Other municipalities (5 strata)	14,670	13,178	1,492	149,495	142,759	6,736	4,656,789	4,493,711	163,078
Public	10,286	8,828	1,458	119,452	112,911	6,541	3,823,369	3,664,655	158,714
Private	4,384	4,350	34	30,043	29,848	195	833,420	829,056	4,364

study population attended schools located in the capitals, whereas the remainder 56% were enrolled in schools of the other strata.

Sample size

Assuming that the prevalence of metabolic syndrome in adolescents is 4%, and specifying a maximum estimation error of 0.9% and a 95% confidence level, the required size for a simple random sample would be 1,821 students. Considering that the sample is clustered by school, shift and grade, and class, a design effect of 2.97 was calculated for the average of body mass. This design effect was obtained by processing the data from the 2007 survey of the surveillance system for risk factors to the health of adolescents developed in the city of Rio de Janeiro, Brazil (details in Castro et al. [2]). As the design effect changes according to the variable considered, and it is, therefore, a precaution for calculating the size of clustered samples, it was decided to use this figure, which leads to a sample size of 5,408 (≈ 1,821 x 2.97) students. In addition, to make up for expected non-response and other losses of up to 15%, the sample size was increased accordingly, reaching 6,219 adolescents. As the survey should produce estimates with the specified precision for each of the 12 domains (= 6 ages x 2 sexes), this led to a total sample size of 74,628 adolescents, which was rounded to 75,060 adolescents after allocation, as it was necessary to have multiples of 60 students sampled in each stratum, as indicated in Table 3.

Sample allocation

The total size of the sample of adolescents was calculated to allow the estimation with controlled precision for 12 domains defined according to

Table 3

Final size of the samples of schools, classes and adolescents by geographic stratum.

Geographic stratum	Final sample size		
	Schools	Classes	Students
Total	1,251	3,753	75,060
Porto Velho	24	72	1,440
Rio Branco	24	72	1,440
Manaus	42	126	2,520
Boa Vista	22	66	1,320
Belém	36	108	2,160
Macapá	25	75	1,500
Palmas	20	60	1,200
Other of the North Region	45	135	2,700
São Luís	34	102	2,040
Teresina	33	99	1,980
Fortaleza	44	132	2,640
Natal	30	90	1,800
João Pessoa	29	87	1,740
Recife	39	117	2,340
Maceió	32	96	1,920
Aracaju	26	78	1,560
Salvador	44	132	2,640
Other of the Northeast Region	68	204	4,080
Belo Horizonte	43	129	2,580
Vitória	22	66	1,320
Rio de Janeiro	56	168	3,360
São Paulo	71	213	4,260
Other of the Southeast Region	103	309	6,180
Curitiba	39	117	2,340
Florianópolis	23	69	1,380
Porto Alegre	33	99	1,980
Other of the South Region	66	198	3,960
Campo Grande	29	87	1,740
Cuiabá	27	81	1,620
Goiânia	36	108	2,160
Brasília	43	129	2,580
Other of the Central Region	43	129	2,580

the sex and age of the adolescents. Now the sample of adolescents is clustered by school, shift and grade, and class, because there were no frames of adolescents available for direct, unclustered selection of adolescents. Thus, the allocation of the total sample into the 32 geographic strata had to be made taking into account the characteristics of the *2009 Educational Census* (ftp://ftp.inep.gov.br/microdados/micro_censo_escolar2009.zip, accessed on 25/Feb/2010), the last one available at the time.

Four different alternatives were tested for the allocation of the total sample size to the strata: equal allocation (same sample size for each stratum); proportional allocation (the stratum sample size is proportional to the stratum population size); power allocation (the stratum sample size is proportional to a power of the stratum population size – two powers 1/3 and 1/2 were tried).

Of the four sample size allocations tested, power allocation with power 1/3 (cubic root) was the one that presented the best balance between precision and sample size available per estimation domain in each stratum.

In fact, equal allocation would ensure the same level of precision for all strata, whereas proportional allocation would ensure distribution proportional to the size of the stratum, but would imply in different precision levels per stratum, and an undesirable sample concentration in the most populated areas of Brazil. Finally, power allocation (with powers of 1/2 and 1/3) reduces the difference between the sample sizes of the various strata, while leading to a compatible size, which allows higher data disaggregation in the larger strata.

Considering that most schools have three or more classes of the grades and shifts being investigated, it was decided to limit the selection to three classes per school. To achieve this goal while spreading the sample over the eligible shifts and grades, three combinations of shift and grade were selected in each school. The average number of students per class, considering a 15% loss, was approximately twenty students. With these parameters, the sizes of the samples of schools and classes for each stratum were determined, by dividing the sample size of students by sixty and by twenty, respectively, as presented in Table 3.

Selection of school sample

The selection of the sample of schools was made using data from the *2009 Educational Census*, in order to determine the parameters for the cost of the project, so that the budget for the study could be prepared. Schools were sampled using probability proportional to size (PPS), with the measure of the size equal to the ratio between the number of the students of the school, in 2009, in the shifts and grades considered, and the distance, in kilometers, between the municipality where the school is located and the capital of the state. This composite size measure aimed to decrease travel costs between the capital and the municipalities with sample schools, by reducing the probability of selecting schools in municipalities that are farther away from the capital. To prevent excessive variability in the size measurements, which would lead to an undesirable variability in the sampling weights, distance ranges in kilometers were converted to numeric scores: (1) up to 10km, score 1; (2) between 11 and 50km, score 10; (3) between 51 and 200km, score 50; (4) between 201 and 400km, score 100; (5) between 401 and 600km, score 150; (6) between 601 and 800km, score 200; (7) between 801 and 1,000km, score 250; and (8) over 1,000km, score 300. The inclusion probabilities of schools are defined in equation (1) of Figure 1.

In order to ascertain the distribution of schools per location (urban or rural area) and administrative dependency (public or private), within each geographic stratum, the systematic PPS selection method was used, with schools in the frame sorted by geographic stratum, location (urban x rural) and administrative dependency (public or private). This corresponds to an implicit stratification by location and administrative dependency within each geographic stratum.

In the school selection process, 23 schools were large enough to be included in the sample with certainty. These schools were included in the sample, and the PPS sampling process was applied for sampling the remaining schools and with the remainder sample size. Note that these 23 schools included with certainty are no longer primary sampling units (PSUs) but become selection strata (Figure 1). For these schools, the PSUs are the combinations of shift and grade.

Once the schools were selected according to the above-mentioned criteria, schools from 124 (45.1%) municipalities were selected out of the 273 municipalities in the study population. This implied that the criteria allowed the sample of schools to be concentrated in less than half of the municipalities considered without loss of regional representation.

Selection of the combinations of shift and grade and classes

In the second stage of sampling, three combinations of shift and grade were sampled amongst those existing in the sampled school. This stage was necessary for two reasons: (1) to enable the blood testing of the students, considering that the 12-hour fasting made it infeasible for students enrolled in the afternoon shift; (2) to represent, in the sample, the different ages of the eligible adolescents, using the grades as approximations to the different ages. The design attempted to allocate about 2/3 of the sample in morning-shift classes, and about 1/3 in the afternoon-shift classes, with an equal allocation for the grades under consideration. The use of these fractions was defined in accordance with the resources available for the blood testing, which would not cover more than 2/3 of the size of the student sample.

Next, a selection algorithm was designed to ensure the selection of precisely three combinations of shift and grade per school, so that the size of the sample per shift and, whenever possible, per grade was respected. Using this selection algorithm, three combinations of shift and grade were selected from each sampled school, in accordance with the information about the existence of classes of the selected shifts and grades, as indicated in the formula (2) of Figure 1. The first ratio in expression (2) indicates the inclusion probability of the combination of shift and grade, while the second indicates the inclusion probability of the class in the combination of shift and grade. The selection of classes among the existing ones in each combination of shift and grade was made in the field, using Microsoft Excel spreadsheets (Microsoft Corp., United States) prepared for each sampled school (example in Figure 2).

These spreadsheets provided the full identification of the selected schools, and two tables for the selection of classes and the sub-sample of two students who should repeat the 24-hour food recall form. In the first Table, by typing, in the third column, the number of classes of the selected shift and grade indicated in the first two columns, the serial number of the selected class was automatically displayed, according to pre-programmed formulas and random numbers assigned for that school. The serial numbers were always associated to the list of eligible classes sorted by name in the selected shift and grade. The second Table was used to enter: the number of students attending classes at the time of data collection; the numbers of delivered, signed and accepted Parental Informed Consent Forms (ICF) and Adolescent Consent Forms; and the numbers of students who took part in each subsample of the investigation (filling out the questionnaire in the personal digital assistant (PDA), anthropometry, blood pressure, blood test, and 24-hour food recall). From the selected random numbers, pre-programmed formulas, and the total number of students who filled out the food recall form, the serial numbers of the two selected students to repeat the 24-hour food recall were displayed in the last two columns on the right. The serial numbers were always associated with the alphabetic order of the names of the students who did the first 24-hour food recall.

Figure 1

Probabilistic scheme of the *Study of Cardiovascular Risks in Adolescents* (ERICA) sample.

Considering h the geographic stratum index, *i* the school index, *e* the shift and grade index, *j* the class index, and *k* the adolescent index, the inclusion probability of any given adolescent in the sample, represented by $P(A_{hiejk})$ is equal to the product of the inclusion probabilities of school *i*, combination of shift and grade *e*, class *j*, and adolescent *k* respectively presented in the expressions (1) to (3):

$$P(E_{hi}) = \frac{n_h X M_{hi}}{M_h} \text{ or } P(E_{hi}) = 1, \text{ if } \frac{n_h X M_{hi}}{M_h} \geq 1; \tag{1}$$

$$P(T_{hiej} | E_{hi}) = \frac{n_{he}}{N_{he}} X \frac{n_{hie}}{N_{hie}} = \frac{n_{he} X n_{hie}}{N_{he} X N_{hie}}; \text{ and} \tag{2}$$

$$P(A_{hiejk} | T_{hiej} \cap E_{hi}) = \frac{n_{hiej}}{N_{hiej}}, \text{ where} \tag{3}$$

n_h is the size of the sample of non-certainty schools of geographic stratum *h* – may be obtained from Table 3 by subtracting the number of certainty schools in each stratum;

M_{hi} is the size measure of school *i* of geographic stratum *h*, defined as the number of enrollments in the eligible grades (grades 7 to 9 of Basic Education; grades 1 to 3 of Secondary Education) divided by the distance score, defined in the section "Selection of school samples";

M_h is the sum of the size measures of all non-certainty schools in geographic stratum *h*, this means, if N_h is the number of non-certainty schools in the population of the geographic stratum *h*, then $M_h = \sum_{i=1}^{N_h} M_{hi}$;

n_{he} *and* N_{he} are the number of eligible classes of the shift and grade *e*, of geographic stratum *h* in the sample and in the population, respectively;

n_{hie} *and* N_{hie} are the effective size of the sample of classes and the number of eligible classes in the shift and grade *e*, and of school *i* of geographic stratum *h* respectively;

n_{hiej} is the effective size of the sample of students of stratum and class *j* of shift and grade *e* of school *i* of stratum *h*;

N_{hiej} is the total number of students of class *j* of shift and grade *e* of school *i* of stratum *h*.

In principle, the expression (3) should be equal to 1, considering that all eligible students of the sampled classes are included in the sample. However, due to absences in the day of the survey or lack of parental consent to participate, the effective sample size may be smaller than the number of eligible students in the class. Expression (3) handles the sample of adolescents actually interviewed as an equiprobable subsample of the students of the class, assuming there is no major difference between those who did take part and those who did not take part.

With these hypotheses, the inclusion probability in the sample of any adolescent, represented by $P(A_{hiejk})$ is given by expressions (4a) and (4b):

$$P(A_{hiejk}) = P(E_{hi}) \times P(T_{hiej} | E_{hi}) \times P(A_{hiejk} | T_{hiej} \cap E_{hi}) = \frac{n_h \times M_{hi}}{M_h} \times \frac{n_{he} \times n_{hie}}{N_{he} \times N_{hie}} \times \frac{n_{hiej}}{N_{hiej}}; \text{ or} \tag{4a}$$

$$P(A_{hiejk}) = \frac{n_{he} \times n_{hie}}{N_{he} \times N_{hie}} \times \frac{n_{hiej}}{N_{hiej}}, \text{ if } \frac{n_h \times M_{hi}}{M_h} \geq 1. \tag{4b}$$

Thus, the design weight to be applied to adolescent *k*, of class *j*, of shift and grade *e*, of school *i*, of the geographic stratum *h*, represented by W_{hitjk}, is given by:

$$W_{hiejk} = 1 / P(A_{hiejk}) = \frac{M_h}{n_h \times M_{hi}} \times \frac{N_{he} \times N_{hie}}{n_{he} \times n_{hie}} \times \frac{N_{hiej}}{n_{hiej}}, \text{ or} \tag{5a}$$

$$W_{hiejk} = \frac{N_{he} \times N_{hie}}{n_{he} \times n_{hie}} \times \frac{N_{hiej}}{n_{hiej}} \text{ if } \frac{n_h \times M_{hi}}{M_h} \geq 1 \tag{5b}$$

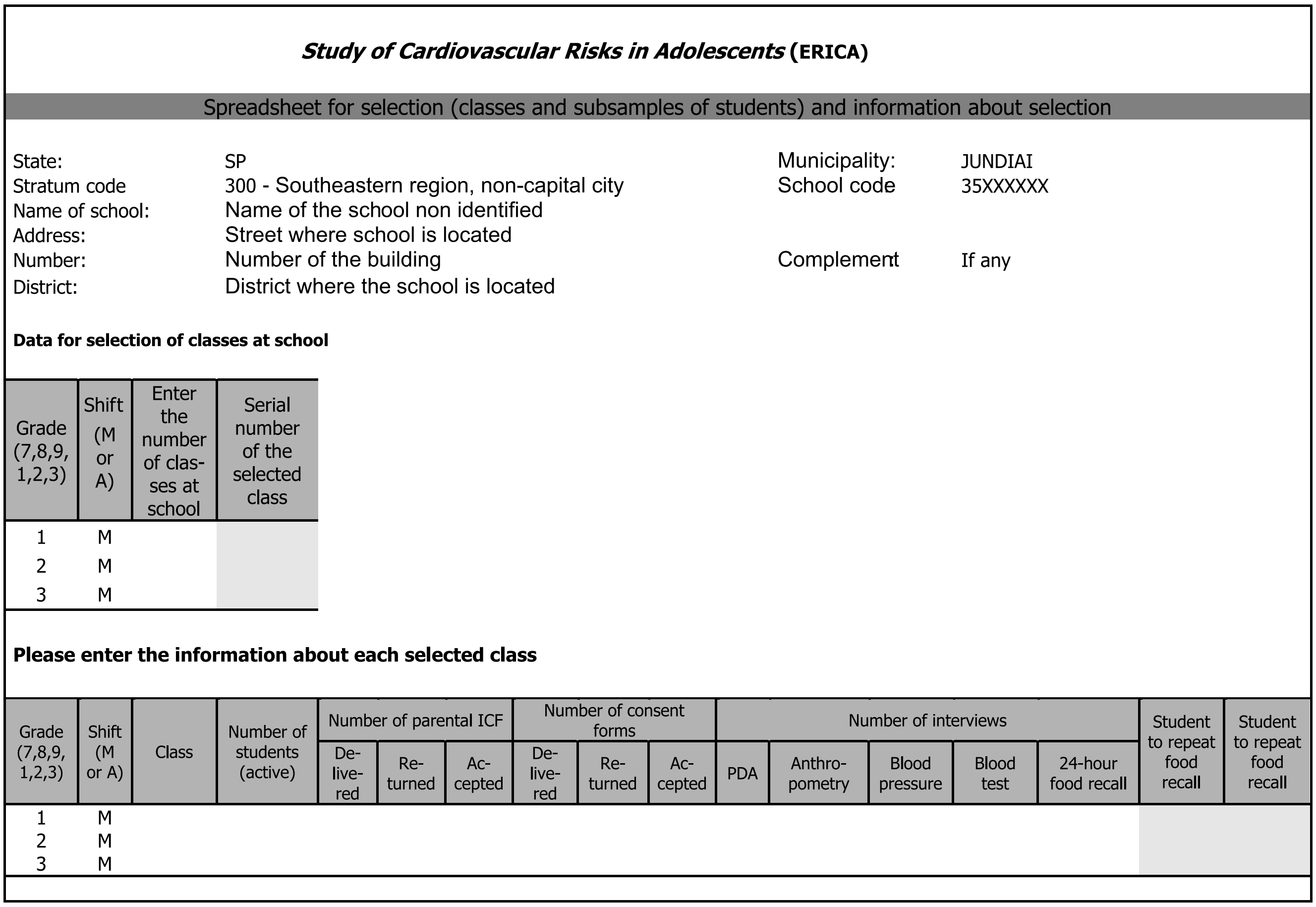
The calibrated weight to be applied to adolescent **k**, of class *j*, of shift and grade *e* of school *i*, of geographic stratum *h*, represented by W^d_{hiejk}, is given by:

$$W^d_{hiejk} = W_{hiejk} \times \frac{P_d}{\hat{N}_d}, \text{ where:} \tag{6}$$

P_d *and* $\hat{N}_d$ are respectively the population total and its estimate for domain *d* (d = 1, ... , 12), being the estimation obtained from the natural weight of the design given by expressions (5a) and (5b).

Figure 2

Example of spreadsheet for the selection of classes and subsamples of students.



***Study of Cardiovascular Risks in Adolescents* (ERICA)**

Spreadsheet for selection (classes and subsamples of students) and information about selection

State:	SP	Municipality:	JUNDIAI
Stratum code	300 - Southeastern region, non-capital city	School code	35XXXXXX
Name of school:	Name of the school non identified		
Address:	Street where school is located		
Number:	Number of the building	Complement	If any
District:	District where the school is located		

Data for selection of classes at school

Grade (7,8,9, 1,2,3)	Shift (M or A)	Enter the number of classes at school	Serial number of the selected class
1	M		
2	M		
3	M		

Please enter the information about each selected class

Grade (7,8,9, 1,2,3)	Shift (M or A)	Class	Number of students (active)	Number of parental ICF			Number of consent forms			Number of interviews					Student to repeat food recall	Student to repeat food recall
				Delivered	Returned	Accepted	Delivered	Returned	Accepted	PDA	Anthropometry	Blood pressure	Blood test	24-hour food recall		
1	M															
2	M															
3	M															

PDA: *personal digital assistant*; ICF: Informd Consent Forms.

Student selection

In principle, all eligible students of the sampled classes were included in the sample. However, due to absences in the day of the survey, or lack of parental consent for blood testing, the observed sample size may be smaller than the number of eligible students per class. Therefore, the inclusion probability of students in the class, presented in formula (3) of Figure 1, implies treating the sample of adolescents actually interviewed as an equi-probable subsample of the eligible students of the class, assuming there are no major differences between those who participated and those who did not participate in each class.

Treatment for non-response and for sub-sampling

The treatment used to compensate for non-response in each part of the study is based on the assumption that the surveyed adolescents are not significantly different from the non-surveyed adolescents in the same class. Thus, the set of surveyed adolescents may be treated as a random subsample of adolescents in each part of the study: filling out PDA questionnaires; anthropometry; blood pressure; blood testing (cases in which blood was collected and results were obtained), and 24-hour food recall.

Although Figure 1 shows a single conditional inclusion probability for an adolescent in his/her class, the formula (3) was calculated independently for each subsample. Therefore, formulas (4a) to (6) may also vary according to the subsample considered.

Item non-response and inconsistent values detected during data editing that occurred in each part of the study were treated by means of automatic correction or probabilistic imputation.

Calculation of sampling weights and their calibration

As indicated in the equations (5a) and (5b) of Figure 1, the design weight [1] is equal to the product of the reciprocals of the inclusion probabilities in each of the sampling stages, considering the treatment of the subsamples as selection stage.

However, like all clustered samples, these design weights do not reflect the distribution of the population of student adolescents (12 to 17 years in full) by sex and age. This occurs also in household surveys, because the inclusion probabilities reflect the total numbers of units in the different selection stages, and provide a good representation of the total (of residents or students), but distort the distribution by sex and age. According to Silva [3], this is the main reason that leads official statistics agencies to calibrate the sampling weights of their household surveys: to ensure that the estimates reflect population data of the elementary sampling units, known from exogenous sources to the investigation.

In the ERICA study, the population data of adolescents enrolled in public or private schools may be estimated considering the information from the last two population censuses held in the Brazil [4,5], and the linear trend model that the Brazilian Institute of Geography and Statistics (IBGE) uses to estimate the population of Brazilian municipalities [6], described by Madeira & Simões [7]. The estimates were made for the date of December 31st, 2013, considering that part of the sample was investigated in 2013, and part in 2014. In this case, the same type of strategy used by IBGE in its household budget surveys was adopted for ERICA: the chosen date is close to the mid-point of the data-collection period of the investigation.

However, the above method does not enable estimating the number of students in the morning shift. For this purpose, two files from the *2013 Educational Census* (ftp://ftp.inep.gov.br/microdados/micro_censo_escolar_2013.zip, accessed on 04/Mar/2014), namely the files with records for classes and the files with the enrolled students, were processed to produce some required estimates. In the classes file, the same criteria applied to define the eligible classes were applied, and in the second the age at December 31st, 2013, was calculated for all enrolled students. After merging the two files, the proportion of eligible students enrolled in the morning shift was calculated by geographic stratum, sex and age. This vector or proportions was applied to the projected population totals for December 31st, 2013, providing the vector of population totals of students in the morning shift by geographic stratum, sex and age.

A post-stratification estimator was used, which is a particular case of regression estimator described in Särndal et al. [8], which multiplies the design weight by a calibration factor that corresponds to the ratio between the known population total and its estimate obtained using by the design weights in each post-stratum. Twelve post-strata were defined, corresponding to the cross-classification of adolescents by six ages and two sexes considered.

Estimation from the ERICA data

The sample of ERICA is considered a complex sample [9], since it uses stratification, clustering and unequal probabilities in its selection stages. Unbiased (or at least approximately unbiased) point estimates of target population parameters may be calculated with the use of calibrated sampling weights by any statistical system that accepts weighting.

However, the usual estimates of variance and other statistics that depend on these (such as standard-deviations, standard errors, confidence intervals of point estimates, regression model parameter significance tests, among others) require special estimation procedures. This is due to the fact that two variability sources interfere in the estimates: (1) one is the complex sample design; and (2) the other concerns the residuals of the calibration equations.

The procedure suggested in the literature [8,9,10] to estimate variances from complex samples is the Ultimate Cluster method, which consists in the estimation of variance from the means calculated per PSU in each (geographic) stratum. Most statistics systems already incorporate routines for the estimation of variances using the Ultimate Cluster method. This method, however, does not consider the source of variability from the calibration, only from the complex sample design.

To consider all the variability sources of the ERICA's sample, the use *survey* library of the R language (The R Foundation for Statistical Computing, Vienna, Austria; http://www.r-project.org) is recommended. This library, developed by Thomas Lumley, allows replication of the sample-weight calibration, saving in an object of sample design both the complex sample design structural variables and the residuals of calibration [10].

For this purpose, the structural design variables of each sub-sample of data will be saved in the files: selection stratum, PSU code, and sampling weight (both the design and the calibrated ones), in addition to the poststratum codes and population totals used in the calibration of sampling weights. The selection strata are the 23 schools included with certainty in the sample and, for the other schools, the selection stratum matches the geographic stratum. For the 23 certainty schools, the PSUs are the eligible combinations of shift and grade. For all the other strata, the PSUs are the schools themselves. The recommended sampling weight is the calibrated weight, but the design weight is included in the files to allow replication of the calibration process in case the *survey* library is used.

Resumen

El Estudio de Riesgo Cardiovascular en Adolescentes *(ERICA) tiene como objetivo estimar la prevalencia de factores de riesgo cardiovascular y síndrome metabólico en adolescentes (12-17 años) matriculados en las escuelas públicas y privadas de 273 municipios con más de 100 mil habitantes en Brasil. La población de estudio fue estratificada en 32 estratos geográficos (27 capitales y cinco conjuntos con otros municipios de cada macrorregión del país); además se seleccionó una muestra de 1.251 escuelas con probabilidad proporcional a su tamaño. En cada escuela se seleccionaran tres combinaciones de horario (matutino y vespertino) con año de la clase, y en cada combinación se seleccionó una clase. Todos los estudiantes elegibles en las clases seleccionadas fueron objeto de la investigación. Los pesos de diseño de la muestra se calcularon por el producto de los inversos de las probabilidades de selección en cada etapa de la muestra y después se calibraron teniendo en cuenta las proyecciones del número de adolescentes inscritos en las escuelas ubicadas en los estratos geográficos por sexo y edad.*

Enfermedades Cardiovasculares; Modelos Estadísticos; Muestreo; Adolescente

Contributors

M. T. L. Vasconcellos participated in designing and selecting the sample, and wrote the manuscript. P. L. N. Silva, M. Szklo, K. V. Bloch, M. C. C. Kuschnir, C. H. Klein, G. A. Abreu and L. A. Barufaldi collaborated in designing the sample and in writing the manuscript.

Acknowledgments

We thank the field team, principals and teachers of the selected schools and classes, and, particularly, to the students, boys and girls, who responded the investigation, without whom ERICA would not exist. To the Brazilian Ministry of Health for their financial support.

References

1. Cochran WG. Sampling techniques. 3rd Ed. New York: John Wiley & Sons; 1977.
2. Castro IRR, Cardoso LO, Engstrom EM, Levy RB, Monteiro CA. Vigilância de fatores de risco para doenças não transmissíveis entre adolescentes: a experiência da cidade do Rio de Janeiro, Brasil. Cad Saúde Pública 2008; 24:2279-88.
3. Silva PLN. Calibration estimation: when and why, how much and how. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2004.
4. Instituto Brasileiro de Geografia e Estatística. Censo demográfico 2000: educação. Resultados da amostra. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2003.
5. Instituto Brasileiro de Geografia e Estatística. Censo demográfico 2010: educação e deslocamento. Resultados da amostra. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2012.
6. Instituto Brasileiro de Geografia e Estatística. Estimativas da população residente nos municípios brasileiros com data de referência em 1º de julho de 2013. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2013.
7. Madeira JL, Simões CCS. Estimativas preliminares da população urbana e rural segundo as Unidades da Federação, de 1960/1980 por uma nova metodologia. Revista Brasileira de Estatística 1972; 33: 3-11.
8. Särndal CE, Swensson B, Wretman JH. Model assisted survey sampling. New York: Springer Verlag; 1992.
9. Skinner CJ, Holt D, Smith TMF. Analysis of complex surveys. Chichester: John Wiley & Sons; 1989.
10. Lumley T. Complex surveys: a guide to analysis using R. Hoboken: John Wiley & Sons; 2010. (Wiley Series in Survey Methodology).

Submitted on 19/Mar/2014
Final version resubmitted on 12/Jan/2015
Approved on 14/Jan/2015