Desenho da amostra do *Estudo do Risco Cardiovascular em Adolescentes* (ERICA)

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

Diseño de la muestra del *Estudio de Riesgo Cardiovascular 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]

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

[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.*

Correspondência
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

Introdução

O *Estudo de Riscos Cardiovasculares em Adolescentes* (ERICA) pretendia fornecer estimativas nacionais da prevalência de fatores de risco cardiovascular e da síndrome metabólica por sexo e idade em adolescentes (12 a 17 anos).

Para tal fim, seria necessário basear o estudo em uma amostra domiciliar de nível nacional, que coletasse informações no subconjunto de domicílios com adolescentes. Entretanto, considerando o custo proibitivo de uma pesquisa domiciliar em nível nacional, optou-se por uma pesquisa através de escolas e por limitar a população de pesquisa aos adolescentes matriculados nos três últimos anos do Ensino Fundamental e nos três anos do Ensino Médio de escolas (públicas ou privadas) localizadas em municípios com mais de 100 mil habitantes, agrupados em 32 estratos geográficos descritos neste artigo. Um estudo baseado em amostra de alunos estratificada geograficamente e conglomerada por escolas, turnos e anos, e turmas, além de permitir inferência para o conjunto dos municípios mais populosos do Brasil e para estratos geográficos que permitem visualizar as diferenças regionais do Brasil, foi a solução para adequar os custos estimados do ERICA ao orçamento disponível. Assim optou-se por uma amostra selecionada em três estágios: escolas; combinações de turno e ano; e turmas. Nas turmas selecionadas todos os alunos elegíveis foram convidados a participar da pesquisa.

Este artigo descreve a população do estudo e sua estratificação geográfica; o cálculo do tamanho de amostra e sua alocação entre os estratos; os métodos para seleção das escolas, combinações de turno e ano, e turmas; o tratamento da não resposta entre adolescentes; o cálculo e calibração dos pesos amostrais; e as técnicas recomendadas para análise dos dados coletados.

População de pesquisa

A população de pesquisa [1] corresponde ao conjunto de adolescentes de 12 a 17 anos, que não possuem qualquer deficiência provisória ou definitiva, e que cursam um dos três últimos anos do Ensino Fundamental ou dos três anos do Ensino Médio nos turnos da manhã ou da tarde, em escolas públicas ou privadas localizadas em um dos 273 municípios com mais de 100 mil habitantes em 1º de julho de 2009 (dado populacional mais recente à época das definições básicas do ERICA).

A caracterização da população de pesquisa é feita com base em arquivo fornecido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP/MEC), que foi produzido com dados do *Censo Escolar 2011*, cujos microdados são disponíveis em ftp://ftp.inep.gov.br/microdados/micro_censo_escolar_2011.zip (acessado em 16/Fev/2012), visto ter sido o usado na seleção das combinações de turno e ano em cada escola selecionada.

Considerando alunos sem defasagem escolar, espera-se que adolescentes de 12 a 17 anos estejam matriculados em turmas de um dos três últimos anos do Ensino Fundamental (7º ao 9º ano) ou de um dos três anos do Ensino Médio, que foram então definidos como os anos elegíveis. Assim, do total de 237.438 escolas investigadas pelo *Censo Escolar 2011*, existiam 389.315 combinações de turnos e anos elegíveis.

No entanto, 63.912 dessas combinações referiam-se a turmas de correção de fluxo (8.114), Fundamental não seriado ou multi-seriado (54.973) e Médio não seriado (825) que foram desconsideradas por não terem uma associação explícita com o ano ou série e, portanto, com a idade do aluno. Além disso, existiam 51.953 combinações relativas a turno noturno, que também foram desconsideradas por razões operacionais: 10.559 no Ensino Fundamental; e 41.394 no Ensino Médio. Assim, em todo o país ficaram 273.450 combinações elegíveis de turno e ano, em 61.325 escolas.

Ao considerar apenas as escolas localizadas nos 273 municípios com mais de 100 mil habitantes em 1º de julho de 2009, o número dessas combinações caiu para 117.726 e o de escolas para 24.441.

A Tabela 1 indica o total de escolas, turmas e alunos dos turnos e anos elegíveis, observados pelo *Censo Escolar 2011* no país e no conjunto dos 273 municípios considerados (os com população maior que 100 mil habitantes). Nela pode-se observar que o recorte geográfico feito no estudo significa que a população de pesquisa representa pouco mais de 54% (8,2 milhões) da população brasileira de adolescentes matriculados em escolas (15,3 milhões) nos turnos e anos elegíveis.

Estratificação da população de pesquisa

A população de pesquisa foi estratificada em 32 estratos geográficos assim constituídos: cada um dos 27 municípios de capital e cinco estratos com o conjunto de municípios de mais de 100 mil habitantes de cada uma das cinco macrorregiões do país.

A distribuição da população de pesquisa pelos estratos geográficos, sumarizada na Tabela 2, indica que 44% dos alunos da população de pes-

Tabela 1

Número de escolas, turmas e alunos por situação urbana e rural, segundo o nível geográfico e a dependência administrativa da escola.

Nível geográfico e dependência administrativa	Escolas			Turmas			Alunos		
	Total	Urbana	Rural	Total	Urbana	Rural	Total	Urbana	Rural
Total nacional	61.325	46.675	14.650	516.701	463.128	53.573	15.233.991	14.119.065	1.114.926
Públicas	48.440	33.955	14.485	431.969	379.153	52.816	12.960.385	11.862.698	1.097.687
Privadas	12.885	12.720	165	84.732	83.975	757	2.273.606	2.256.367	17.239
273 municípios	24.441	22.698	1.743	261.086	253.051	8.035	8.273.275	8.076.757	196.518
Públicas	15.942	14.244	1.698	198.266	190.477	7.789	6.492.228	6.301.417	190.811
Privadas	8.499	8.454	45	62.820	62.574	246	1.781.047	1.775.340	5.707

Tabela 2

Distribuição do número de escolas, turmas e alunos por situação urbana e rural, segundo o grupo de estratos geográficos e a dependência administrativa da escola.

Grupo de estratos e dependência administrativa	Escolas			Turmas			Alunos		
	Total	Urbana	Rural	Total	Urbana	Rural	Total	Urbana	Rural
Total (273 municípios)	24.441	22.698	1.743	261.086	253.051	8.035	8.273.275	8.076.757	196.518
Públicas	15.942	14.244	1.698	198.266	190.477	7.789	6.492.228	6.301.417	190.811
Privadas	8.499	8.454	45	62.820	62.574	246	1.781.047	1.775.340	5.707
Capitais (27 estratos)	9.771	9.520	251	111.591	110.292	1.299	3.616.486	3.583.046	33.440
Públicas	5.656	5.416	240	78.814	77.566	1.248	2.668.859	2.636.762	32.097
Privadas	4.115	4.104	11	32.777	32.726	51	947.627	946.284	1.343
Outros (5 estratos)	14.670	13.178	1.492	149.495	142.759	6.736	4.656.789	4.493.711	163.078
Públicas	10.286	8.828	1.458	119.452	112.911	6.541	3.823.369	3.664.655	158.714
Privadas	4.384	4.350	34	30.043	29.848	195	833.420	829.056	4.364

quisa estavam em escolas dos estratos de capital, enquanto os restantes 56% pertenciam a escolas dos demais estratos.

Tamanho da amostra

Considerando a prevalência da síndrome metabólica em adolescentes de 4%, a ser estimada com erro máximo de 0,9% e com nível de 95% de confiança, o tamanho requerido para uma amostra aleatória simples seria de 1.821 alunos. Considerando que a amostra é conglomerada por escola, turno e ano, e turma, foi calculado um efeito de desenho de 2,97 para a média de massa corporal, obtido a partir do processamento dos dados do inquérito de 2007 do sistema de vigilância para fatores de risco à saúde de adolescentes, implementado no Município do Rio de Janeiro, Brasil (detalhes em Castro et al. [2]). Como o efeito do desenho muda segundo a variável considerada e é, portanto, uma precaução no cálculo do tamanho de amostras conglomeradas, decidiu-se usar esse valor, o que conduziria a um tamanho de amostra de 5.408 (≈ 1.821 x 2,97) alunos, que, acrescido de 15% para compensar perdas esperadas de não resposta e outras, alcançou o valor de 6.219 adolescentes. Como a pesquisa deve produzir estimativas com a precisão especificada para cada um de 12 domínios (= 6 idades x 2 sexos), isto conduziu a um tamanho total de amostra de 74.628 adolescentes, que, após sua alocação, foi arredondado para 75.060 adolescentes, pois tamanhos múltiplos de 60 eram necessários em cada estrato, como indicado na Tabela 3.

Alocação da amostra

O tamanho total da amostra de adolescentes foi calculado para permitir estimar com precisão controlada para 12 domínios definidos em

Tabela 3

Tamanho final da amostra de escolas, turmas e adolescentes, segundo o estrato geográfico da amostra.

Estrato geográfico	Tamanho final da amostra		
	Escolas	Turmas	Alunos
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
Outros da Região Norte	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
Outros da Região Nordeste	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
Outros da Região Sudeste	103	309	6.180
Curitiba	39	117	2.340
Florianópolis	23	69	1.380
Porto Alegre	33	99	1.980
Outros da Região Sul	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
Outros da Região Centro-oeste	43	129	2.580

função de sexo e idade dos adolescentes. Entretanto, a amostra de adolescentes é conglomerada por escolas, turnos e anos, e turmas, já que não havia cadastro de adolescentes disponível para seleção direta de uma amostra não conglomerada de adolescentes. Sendo assim, a alocação da amostra total nos 32 estratos geográficos teve que ser feita considerando características do *Censo Escolar 2009* (ftp://ftp.inep.gov.br/microdados/micro_censo_escolar2009.zip, acessado em 25/Fev/2010), último disponível à época.

Para alocação do tamanho total da amostra foram testadas diferentes alternativas: alocação igual (mesmo tamanho de amostra para cada estrato); alocação proporcional (o tamanho da amostra do estrato é proporcional ao seu tamanho populacional); alocação potência (o tamanho da amostra é proporcional a uma potência do tamanho do estrato – duas potências 1/3 e 1/2 foram testadas).

Dentre as quatro formas de alocação testadas, a alocação com potência 1/3 (raiz cúbica) foi a que apresentou melhor equilíbrio entre a precisão e o tamanho de amostra disponível por domínio de estimação previsto para cada estrato.

De fato, a alocação igual garantiria o mesmo nível de precisão em todos os estratos, enquanto a proporcional asseguraria distribuição proporcional ao tamanho do estrato, mas implicaria em precisão diferenciada por estrato e uma indesejável concentração nos estratos mais populosos do país. Por fim, as alocações potência testadas (1/2 e 1/3) reduzem a diferença entre os tamanhos de amostra dos estratos, caminhando na direção da alocação igual e, portanto, na direção de uma amostra que mantém a mesma precisão por estrato geográfico, mas que conduz a um tamanho compatível que permite maior desagregação dos dados nos maiores estratos.

Considerando que a grande maioria das escolas tem três ou mais turmas dos turnos e anos considerados decidiu-se fixar a seleção de três turmas por escola. Para alcançar este objetivo e, ao mesmo tempo, manter um bom espalhamento da amostra ao longo dos turnos e anos, foram selecionadas inicialmente em cada escola três combinações de turno (manhã x tarde) e ano (EF-7, EF-8, EF-9, EM-1, EM-2, EM-3), onde EF corresponde ao Ensino Fundamental (7º ao 9º ano) e EM corresponde ao Ensino Médio (1º ao 3º ano). A média de alunos por turma, considerando a perda de 15%, ficou em torno de vinte alunos. Com esses parâmetros, foram obtidos os tamanhos das amostras de escolas e de turmas em cada estrato dividindo o tamanho da amostra de alunos por 60 e 20, respectivamente, como apresentados na Tabela 3.

Seleção da amostra de escolas

A seleção da amostra de escolas foi feita com base nos dados do *Censo Escolar 2009*, a fim de determinar os parâmetros de custo do projeto que permitiram elaborar o orçamento do estudo. Foi feita com probabilidade proporcional ao tamanho (PPT), sendo a medida de tamanho correspondente à razão entre o número de alunos que a escola possuía, em 2009, nos turnos e anos considerados e a distância em quilômetros entre

a sede do município onde se localiza a escola e a sede do município de capital. Essa medida de tamanho objetivou reduzir o custo do deslocamento entre a capital do estado e os municípios selecionados, por meio da redução da probabilidade de seleção das escolas em municípios mais afastados da capital. Para evitar uma variabilidade grande nas medidas de tamanho, o que acarretaria variabilidade indesejável nos pesos amostrais, as faixas de distância em quilômetros foram associadas a valores: (1) até 10km, valor 1; (2) de 11 a 50km, valor 10; (3) de 51 a 200km, valor 50; (4) de 201 a 400km, valor 100; (5) de 401 a 600km, valor 150; (6) de 601 a 800km, valor 200; (7) de 801 a 1.000km, valor 250; e (8) acima de 1.000km, valor 300. As probabilidades de inclusão das escolas constam da Figura 1.

Para tentar preservar a distribuição de escolas por situação (urbana ou rural) e dependência administrativa (pública ou privada) na amostra dentro de cada estrato geográfico, foi usado o método de seleção PPT sistemático, com ordenação prévia das escolas do cadastro de seleção por estrato geográfico, situação e dependência administrativa. Essa combinação de seleção sistemática com ordenação prévia do cadastro de seleção corresponde a uma estratificação implícita por situação e dependência administrativa dentro de cada estrato geográfico.

No processo de seleção da amostra, observou-se que 23 escolas tinham uma medida de tamanho grande o suficiente para serem incluídas na amostra com certeza. Essas escolas foram incluídas na amostra e o processo foi repetido para as demais escolas e com o tamanho restante da amostra. É importante registrar que essas 23 escolas não são mais unidades primárias de amostragem (UPA) e sim estratos de seleção (Figura 1). Nestas escolas, a UPA será a combinação de turno e ano.

Feita a seleção das escolas com os critérios indicados acima, observou-se que foram selecionadas escolas em 124 (45,1%) dos 273 municípios considerados na população de pesquisa, mostrando que os critérios concentraram a amostra nos municípios sem perder a representatividade regional.

Seleção da amostra de combinações de turno e ano e de turmas

No segundo estágio de seleção foram selecionadas, inicialmente em cada escola, três combinações de turno (manhã x tarde) e ano (EF-7, EF-8, EF-9, EM-1, EM-2, EM-3) entre as existentes na escola. Esse estágio foi necessário por dois motivos: (1) para viabilizar os exames de sangue dos alunos, visto que o jejum de 12 horas inviabilizava a seleção de alunos dos turnos da tarde; e (2) para representar na amostra as diferentes idades dos adolescentes elegíveis usando o ano da turma como uma aproximação da idade. Nas marginais desse cruzamento, buscou-se uma alocação de aproximadamente 2/3 das turmas no turno da manhã, com 1/3 no da tarde, e uma alocação igual dos anos considerados. O uso dessas frações foi definido em função dos recursos disponíveis para os exames de sangue, que não cobririam mais de 2/3 do tamanho da amostra de alunos.

Em seguida, foi elaborado um algoritmo de seleção para assegurar a seleção de exatamente três combinações de turno e ano por escola de forma a respeitar o tamanho da amostra por turno e, sempre que possível, por ano. Com o emprego desse algoritmo de seleção, para cada escola da amostra foram selecionadas três combinações de turno e ano, respeitando as informações sobre a existência de turmas nos turnos e anos selecionados, como indicado na expressão (2) da Figura 1. A primeira razão na expressão (2) indica a probabilidade de inclusão da combinação turno e ano na escola e a segunda razão indica a probabilidade de inclusão da turma na combinação de turno e ano. A seleção das turmas dentre as existentes em cada combinação de turno e ano foi feita em campo, com apoio de planilhas Microsoft Excel (Microsoft Corp., Estados Unidos) preparadas para cada escola amostrada (ver exemplo na Figura 2).

Essas planilhas continham a identificação completa das escolas selecionadas e duas tabelas para seleção de turmas e da subamostra de dois alunos que deveriam repetir o recordatório alimentar de 24 horas. Na primeira tabela, ao digitar, na terceira coluna, o número de turmas da combinação de turno e ano selecionada, indicada nas duas primeiras colunas, automaticamente aparecia o número de ordem da turma selecionada, em função de fórmulas pré-programadas e de números aleatórios selecionados para a escola. O número de ordem estava sempre associado à sequência das designações das turmas na escola. Na segunda tabela, foram digitados os números de alunos frequentando as aulas no momento da coleta de dados; os números de alunos relacionados a cada situação relativa ao Termo de Consentimento Livre e Esclarecido (TCLE) e de assentimento; bem como o número de alunos que participaram de cada subamostra da pesquisa (preenchimento do questionário no *personal digital assistant* – PDA), antropometria, pressão arterial, exame de sangue, e recordatório alimentar de 24 horas). Em função dos números aleatórios selecionados, das fórmulas pré-programadas e do total de alunos que preencheram

Figura 1

Esquema probabilístico da amostra do *Estudo de Riscos Cardiovasculares em Adolescentes* (ERICA).

Denotando por *h* o índice do estrato geográfico, por *i* o índice da escola, por *e* o índice da combinação de turno e ano, por *j* o índice da turma, e por *k* o índice do adolescente elegível, a probabilidade de inclusão na amostra de um adolescente elegível qualquer, representada por $P(A_{hiejk})$ é igual ao produto das probabilidades de inclusão da escola *i*, da combinação turno e ano *e*, da turma *j*, e do adolescente *k* apresentadas respectivamente nas expressões (1) a (3):

$$P(E_{hi}) = \frac{n_h \times M_{hi}}{M_h} \text{ ou } P(E_{hi}) = 1, \text{ caso } \frac{n_h \times M_{hi}}{M_h} \geq 1; \quad (1)$$

$$P(T_{hiej} \mid E_{hi}) = \frac{n_{he}}{N_{he}} \times \frac{n_{hie}}{N_{hie}} = \frac{n_{he} \times n_{hie}}{N_{he} \times N_{hie}}; \text{ e} \quad (2)$$

$$P(A_{hiejk} \mid T_{hiej} \cap E_{hi}) = \frac{n_{hiej}}{N_{hiej}}, \text{ onde} \quad (3)$$

n_h é o tamanho da amostra de escolas do estrato geográfico *h*, indicado na Tabela 3 subtraído do número de escolas incluídas com certeza na amostra;

M_{hi} é a medida de tamanho associada à escola *i* do estrato geográfico *h*, definida como o número de matrículas nos anos elegíveis (7 a 9 do Ensino Fundamental; 1 a 3 do Ensino Médio) dividido pelo valor associado à classe da distância, em quilômetros, entre a cidade da escola e a capital da UF;

M_h é a soma das medidas de tamanho de todas as escolas não incluídas com certeza na amostra do estrato geográfico *h*, ou seja, se N_h for o número de escolas não certas na população do estrato geográfico *h* então $M_h = \sum_{i=1}^{N_h} M_{hi}$;

n_{he} e N_{he} são os números de combinações elegíveis de turno e ano selecionadas e existentes na escola *i* do estrato geográfico *h*, respectivamente;

n_{hie} e N_{hie} são o tamanho da amostra efetiva de turmas e o número de turmas elegíveis na combinação de turno e ano e da escola *i* do estrato geográfico *h*, respectivamente;

N_{hiej} é o número total de alunos elegíveis da turma *j* na combinação de turno e ano e da escola *i* do estrato *h*.

Em princípio, a expressão (3) deveria ser igual a 1, visto que todos os alunos das turmas amostradas são incluídos com certeza na amostra. No entanto, em decorrência de faltas no dia da pesquisa, do não consentimento dos pais para participação na pesquisa, ou de o aluno não ter idade para ser elegível, o tamanho da amostra efetiva pode ser menor do que o número de alunos da turma. A expressão (3) equivale a tratar a amostra de adolescentes efetivamente entrevistados como uma subamostra equiprovável dos alunos elegíveis da turma, pressupondo que não há diferença importante entre os que participaram e os que não participaram.

Com essas hipóteses, a probabilidade de inclusão na amostra de um adolescente qualquer, representada por $P(A_{hiejk})$, é dada pelas expressões (4a) ou (4b):

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

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

Dessa forma, o peso natural do desenho da amostra a ser aplicado ao adolescente *k* da turma *j* na combinação de turno e ano e da escola i do estrato geográfico *h*, representado por W_{hitjk}, é dado por:

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

$$W_{hiejk} = \frac{N_{he} \times N_{hie}}{n_{he} \times n_{hie}} \times \frac{N_{hiej}}{n_{hiej}} \text{ caso } \frac{n_h \times M_{hi}}{M_h} \geq 1 \quad (5b)$$

O peso calibrado a ser aplicado ao adolescente *k* da turma *j* na combinação de turno e ano e da escola *i* do estrato geográfico *h*, representado por W^d_{hiejk}, é dado pela expressão abaixo:

$$W^d_{hiejk} = W_{hiejk} \times \frac{P_d}{\hat{N}_d}, \text{ onde} \quad (6)$$

P_d e $\hat{N}_d$ são, respectivamente, a população e a estimativa para o domínio de sexo e idade d (d = 1, ... , 12), sendo a estimativa obtida com o peso natural do desenho dado pelas expressões (5a) e (5b).

Figura 2

Exemplo de planilha de seleção de turmas e subamostras de alunos.

***Estudo de Riscos Cardiovasculares em Adolescentes* (ERICA)**

Planilha de seleção (turmas e subamostras de alunos) e de informações sobre as seleções

Sigla da UF:	SP	Município:	JUNDIAI
Código do Estrato	300 - Sudeste não-capital	Código da Escola:	35XXXXXX
Nome da escola:	Nome da escola não identificado		
Endereço da escola:	Rua onde está situada		
Número da escola:	Número do prédio	Complemento:	Se houver
Bairro da escola:	Bairro onde está localizada		

Dados para seleção das turmas na escola

Ano (7,8,9, 1,2,3)	Turno (M ou T)	Digite o nº de turmas na escola	Nº de ordem da turma sele-cionada
1	M		
2	M		
3	M		

Digite os dados das turmas selecionadas

Ano (7,8,9, 1,2,3)	Turno (M ou T)	Turma	Nº de alunos (Frequen-tando)	Nº de TCLE			Nº de Termos de Assentimento			Número de entrevistas					Aluno a fazer 2º recor-datório	Aluno a fazer 2º recor-datório
				Entre-gues	Devol-vidos	Com aceite	Entre-gues	Devol-vidos	Com aceite	PDA	Antro-pometria	Pressão arterial	Sangue	Recordatório 24 horas		
1	M															
2	M															
3	M															

PDA: *personal digital assistant*; TCLE: Termo de Consentimento Livre e Esclarecido.

o recordatório, apareciam nas duas colunas da direita o número de ordem dos dois alunos selecionados para repetir o recordatório. O número de ordem estava sempre associado à ordem alfabética dos nomes dos alunos que fizeram o primeiro recordatório.

Seleção dos alunos

Em princípio, todos os alunos das turmas amostradas foram incluídos com certeza na amostra. No entanto, em decorrência de falta nos dia da pesquisa, do não consentimento dos pais para coleta de sangue, ou de o aluno não ter idade elegível (12 a 17 anos), o tamanho da amostra efetiva pode ser menor do que o número de alunos da turma. Assim, probabilidade de inclusão dos alunos elegíveis nas turmas da amostra, apresentada na expressão (3) da Figura 1, equivale a tratar a amostra de adolescentes efetivamente entrevistados como uma subamostra equiprovável dos alunos elegíveis da turma, pressupondo que não há diferença importante entre os que participaram e os que não participaram na turma.

Tratamento da não resposta e das subamostras

O tratamento dado às não respostas em cada parte do estudo decorre do pressuposto de que os adolescentes pesquisados não diferem de forma importante dos não pesquisados da mesma classe. Assim, o conjunto de adolescentes pesquisados pode ser tratado como uma subamostra da amostra de adolescentes em cada parte do estudo: questionários em PDA; antropometria;

pressão arterial; exame de sangue (casos coletados e que tiveram os resultados obtidos) e recordatório 24 horas.

Apesar de a Figura 1 apresentar uma única probabilidade de inclusão do adolescente, a expressão (3) foi calculada independentemente para cada subamostra. Em consequência, as expressões (4a) a (6) também podem variar segundo a subamostra considerada.

As não respostas de itens pontuais que ocorreram dentro de cada parte do estudo foram tratadas durante a crítica dos dados, da mesma forma que valores inconsistentes encontrados, por meio de correções ou imputação probabilística.

Cálculo dos pesos amostrais e sua calibração

Como indicado nas expressões (5a) e (5b) da Figura 1, o peso natural do desenho 1 é igual ao recíproco do produto das probabilidades de inclusão em cada estágio da amostra, considerando o tratamento das subamostras como um estágio de seleção.

No entanto, como em toda amostra conglomerada, os pesos calculados não refletem a distribuição da população de adolescentes estudantes (12 a 17 anos completos) por sexo e idade. Isto ocorre, também, nas pesquisas domiciliares, visto que as probabilidades de inclusão refletem os quantitativos nos vários estágios de seleção e representam bem os totais (de moradores ou de alunos), mas distorcem as distribuições por sexo e idade. Segundo Silva [3], esta é a principal razão que conduz os órgãos oficiais de estatística a calibrarem os pesos amostrais de suas pesquisas demográficas: assegurar que as estimativas reflitam dados populacionais das unidades elementares da amostra, conhecidos por fontes exógenas à pesquisa.

No ERICA, os dados populacionais de adolescentes matriculados em escolas públicas ou privadas podem ser estimados considerando as informações dos dois últimos censos demográficos realizados no país [4,5], e o método da tendência linear, que a Fundação Instituto Brasileiro de Geografia e Estatística (IBGE) usa para projetar a população dos municípios brasileiros [6], descrito por Madeira & Simões [7]. As projeções foram feitas para a data de 31 de dezembro de 2013, visto que parte da amostra foi pesquisada em 2013 e parte em 2014. Adotou-se neste caso o mesmo tipo de estratégia que o IBGE utiliza em suas pesquisas de orçamentos familiares: uma data próxima ao centro do período de coleta da pesquisa.

No entanto, o método acima não permite obter totais populacionais de estudantes do turno da manhã. Para este caso, foram processados dois arquivos do *Censo Escolar 2013* (ftp://ftp.inep.gov.br/microdados/micro_censo_escolar_2013.zip, acessado em 04/Mar/2014): (1) turmas; (2) matrículas. No primeiro, manteve-se o mesmo conjunto de critérios para selecionar as turmas elegíveis, enquanto no segundo calculou-se a idade para 31 de dezembro de 2013. Com a junção dos dois arquivos, a proporção de alunos do turno da manhã foi calculada por estrato geográfico, sexo e idade. Esse vetor de proporções foi aplicado sobre os totais populacionais projetados para 31 de dezembro de 2013, fornecendo o vetor de totais populacionais de estudantes do turno da manhã.

Foi utilizado um estimador de pós-estratificação, que é um caso particular do estimador de regressão descrito em Särndal et al. [8], que modifica o peso natural do desenho por um fator de calibração que corresponde à razão entre o total populacional e o total estimado pelo peso natural do desenho para o pós-estrato ou domínio de estimação considerado. Foram definidos 12 domínios de estimação correspondentes às seis idades consideradas e os dois sexos.

Estimação a partir dos dados do ERICA

A amostra do ERICA é uma amostra considerada complexa [9], uma vez que emprega estratificação e conglomeração e probabilidades desiguais em seus estágios de seleção. Estimativas pontuais não enviesadas (ou pelo menos aproximadamente não enviesadas) de parâmetros populacionais, exceto variância, podem ser obtidas com o emprego dos pesos amostrais calibrados, usando qualquer sistema estatístico que aceite ponderação.

No entanto, as estimativas usuais de variância e demais estatísticas que delas dependem (como desvios-padrão, erros padrão, intervalos de confiança de estimativas pontuais, testes de significância de parâmetros de modelos de regressão, entre outras) demandam procedimento especial de estimação. Isto ocorre porque duas fontes de variabilidade interferem nas estimativas: (1) a decorrente do desenho complexo da amostra; e (2) a advinda dos resíduos das equações de calibração.

O procedimento sugerido na literatura [8,9,10] para estimação de variâncias a partir de amostras complexas é o método do conglomerado primário, que consiste na estimação da variância entre as médias obtidas por UPA dentro de cada estrato geográfico. A maior parte dos sistemas estatísticos atualmente já incorpora rotinas para estimação de variâncias pelo método do conglomerado primário. No entanto, este método não considera a fonte de variabilidade decorrente da

calibração, apenas a decorrente do desenho de amostra complexo.

Para considerar todas as fontes de variabilidade da amostra do Erica é recomendada, dentre outras, a biblioteca *survey* da linguagem R (The R Foundation for Statistical Computing, Viena, Áustria; http://www.r-project.org). Esta biblioteca, desenvolvida por Thomas Lumley permite replicar a calibração dos pesos amostrais, guardando em um objeto de desenho da amostra os resíduos de calibração [10].

Para tanto, as variáveis estruturais do desenho de cada subamostra de dados serão gravadas nos arquivos: estrato de seleção, código da UPA, e peso amostral (o natural e o calibrado), além dos totais populacionais usados na calibração dos pesos amostrais. Os estratos de seleção são as 23 escolas incluídas com certeza na amostra e, para as demais escolas, o estrato de seleção coincide com o estrato geográfico. Nas 23 escolas incluídas com certeza na amostra, as UPA são as combinações de turno e ano. Nos demais estratos de seleção, as UPA são as próprias escolas. O peso amostral recomendado é o peso calibrado, mas o peso natural do desenho foi mantido nos arquivos para permitir a replicação da calibração no caso de uso da biblioteca *survey*.

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

Colaboradores

M. T. L. Vasconcellos participou da concepção do desenho da amostra, selecionou a amostra e redigiu o manuscrito. P. L. N. Silva, M. Szklo, K. V. Bloch, M. C. C. Kuschnir, C. H. Klein, G. A. Abreu e L. A. Barufaldi colaboraram na concepção do desenho da amostra e na redação do manuscrito.

Agradecimentos

Agradecemos à equipe de campo, aos diretores e professores das escolas e turmas selecionadas e, sobretudo, aos alunos e alunas que responderam à pesquisa, sem os quais o ERICA não existiria. Ao Ministério da Saúde pelo apoio financeiro.

Referências

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).

Recebido em 19/Mar/2014
Versão final reapresentada em 12/Jan/2015
Aprovado em 14/Jan/2015