

Bianca Schmid<sup>I</sup>

Nilza Nunes da Silva<sup>II</sup>

# Estimação de sub-registro de nascidos vivos pelo método de captura e recaptura, Sergipe

## Estimation of live birth underreporting with a capture-recapture method, Sergipe, Northeastern Brazil

---

### RESUMO

**OBJETIVO:** Estimar o número de ocorrências de nascidos vivos e, por consequência, o sub-registro civil de nascidos vivos.

**MÉTODOS:** As bases de dados do Sistema Nacional de Informação sobre Nascidos Vivos e do Registro Civil do Instituto Brasileiro de Geografia e Estatística, nos segundo e terceiro trimestres de 2006 do estado de Sergipe, foram pareadas por relacionamento determinístico a partir do número da Declaração de Nascido Vivo. A desagregação geográfica adotada foi a de microrregião de residência da mãe. Os modelos de Huggins para populações fechadas foram aplicados para estimar as probabilidades de captura em cada base e o total de nascidos vivos ocorrido no período, dentro de cada desagregação geográfica. O aplicativo utilizado para as estimações foi o Software MARK®.

**RESULTADOS:** O sub-registro civil no período analisado foi de 19,3%. A aplicação do método de captura-recaptura para estimar sub-registro de nascidos vivos é factível, inclusive para desagregações geográficas menores do que unidade da federação. O relacionamento determinístico foi prejudicado em quatro microrregiões, devido à falta de preenchimento do número da Declaração de Nascido Vivo na base do Instituto Brasileiro de Geografia e Estatística. Identificou-se que a idade da mãe afeta a probabilidade de captura pelo Registro Civil, característica de heterogeneidade na população de nascidos vivos.

**CONCLUSÕES:** O método de captura-recaptura mostrou-se viável para a estimação de sub-registro de nascidos vivos.

**DESCRIPTORIOS:** Nascimento Vivo. Declaração de Nascimento. Sistema de Registros. Registros como Assunto. Estatísticas Vitais. Captura-recaptura.

<sup>I</sup> Instituto Brasileiro de Geografia e Estatística. Unidade Estadual São Paulo. São Paulo, SP, Brasil

<sup>II</sup> Faculdade de Saúde Pública. Universidade de São Paulo. São Paulo, SP, Brasil

#### Correspondência | Correspondence:

Bianca Schmid  
R. Frei Caneca, 443 – Apto. 102  
Consolação  
01307-001 São Paulo, SP, Brasil  
E-mail: bika.schmid@gmail.com

Recebido: 13/12/2010

Aprovado: 27/7/2011

---

## ABSTRACT

**OBJECTIVE:** Estimate the number of live births and, therefore, underreporting of live births.

**METHODS:** The databases of the Live Birth Information System and the Civil Registry of the Brazilian Institute of Geography and Statistics, from the second and third trimesters of 2006 in Sergipe state (Northeastern Brazil) were paired by deterministic linkage based on the number of the Live Birth Declaration. The geographic disaggregation utilized was mother's microregion of residence. Huggins closed population models were used to estimate the capture probabilities for each database and the total live births during the period, within each geographic subdivision. MARK® software was used for the estimates.

**RESULTS:** Underregistration during the period studied was 19.3%. Application of the capture-recapture method to estimate underregistration of live births is possible, including for geographic disaggregations smaller than a state. The deterministic linkage was impaired in four microregions, due to non-inclusion of the Live Birth Declaration number in the database of the Brazilian Institute of Geography and Statistics. Maternal age, a heterogeneity characteristic in the population of live births, affected the probability of capture by the civil registry.

**CONCLUSIONS:** Capture-recapture was a viable method to estimate the underregistration of live births.

**DESCRIPTORS:** Live Birth. Birth Certificates. Underregistration. Registries. Records as Topic. Vital Statistics.

---

## INTRODUÇÃO

O sub-registro de eventos vitais ainda é uma realidade no País.<sup>a,b</sup> Segundo Simões,<sup>c</sup> a precariedade da cobertura dos sistemas de estatísticas vitais revela-se um entrave para o cálculo direto do nível e padrão de fecundidade e mortalidade no Brasil.

O cálculo de fecundidade e mortalidade infantil por meio de métodos diretos, ou seja, sem correção para o sub-registro de nascimentos e óbitos, respectivamente, pode mascarar a realidade demográfica de uma população.<sup>a</sup> Para o cálculo desses indicadores, lança-se mão de técnicas indiretas para a estimação, tomando-se como fonte de informação censos demográficos e pesquisas amostrais. Muitas vezes, a quebra dos pressupostos implícitos exigidos por tais técnicas causa distorções nas estimativas. Ainda, quando a demanda por estimativas se dá para os níveis de desagregação geográfica menores do que unidade da federação, o problema torna-se mais complexo, dado o pequeno

tamanho populacional de boa parte dos municípios brasileiros.<sup>c</sup>

Vários indicadores podem ser calculados a partir das estatísticas do Registro Civil, como taxas de fecundidade, coeficientes de mortalidade e esperança de vida ao nascer, por exemplo. A busca do diagnóstico dessa realidade de negligência do registro civil visa justamente subsidiar ações que permitam reverter essa situação, de maneira a colaborar no cumprimento dos Princípios de Padrões Profissionais e de Eficiência, dentre os Fundamentais das Estatísticas Oficiais da Organização das Nações Unidas.<sup>d</sup>

A estimação pelo método de captura-recaptura consiste em se aproveitar da superposição dos registros presentes em fontes sabidamente incompletas, mas disponíveis, para formalmente medir a subestimação dessas fontes. Isso permite a correção das estatísticas, de maneira

---

<sup>a</sup> Romero DEM. Vantagem e limitações do método demográfico indireto e dos dados da PNAD'98 para a estimativa da mortalidade infantil. In: Anais do XIII Encontro Nacional de Estudos Populacionais, Ouro Preto, BR. São Paulo: Associação Brasileira de Estudos Populacionais; 2002[citado 2007 jan 03]. Disponível em: [http://www.abep.nepo.unicamp.br/docs/anais/pdf/2002/gt\\_sau\\_st3\\_romero\\_texto.pdf](http://www.abep.nepo.unicamp.br/docs/anais/pdf/2002/gt_sau_st3_romero_texto.pdf)

<sup>b</sup> Instituto Brasileiro de Geografia e Estatística. Estatísticas do registro civil. Rio de Janeiro; 2005. v. 32.

<sup>c</sup> Simões CCS. Brasil: Estimativas da Mortalidade Infantil por microrregiões e municípios. Brasília: Ministério da Saúde; 1999.

<sup>d</sup> United Nations. Statistical Division. Official statistics: principles and practices, organization and management. New York; 2006[citado 2009 out 26]. Disponível em: <http://unstats.un.org/unsd/methods/statorg/default.htm>

a produzir indicadores mais próximos da realidade. Essas fontes (listas) disponíveis podem ser registros de doenças de notificação compulsória, estatísticas hospitalares e de outros serviços de saúde e registros de óbitos, entre outros.<sup>2,10,11,12</sup>

Resumidamente, método de captura-recaptura foi empregado na estimação da população da França, em 1793. A partir do século XIX começou a ser utilizado amplamente na estimação do tamanho de populações de animais selvagens<sup>3,8</sup> e em várias outras aplicações nas áreas da medicina, demografia e epidemiologia.

Petersen, em 1894, desenvolveu o modelo mais simples de estimação pelo método de captura-recaptura, o de duas amostras.<sup>8</sup> Na década de 1940, Sekar & Deming<sup>13</sup> estimaram o sub-registro de nascidos vivos e óbitos na Índia. A partir de dados censitários, Shapiro<sup>14</sup> aplicou a técnica para o cálculo do sub-registro de nascidos vivos nos EUA. Em 1968, Wittes & Sidel<sup>18</sup> introduziram uma generalização do método de captura-recaptura para as aplicações em epidemiologia, utilizando-se duas ou mais listas. O interesse no método foi crescente, e a partir da década de 1990 houve incremento considerável nas aplicações em pesquisas epidemiológicas.<sup>8</sup>

O objetivo do presente trabalho foi aplicar a metodologia de captura-recaptura para estimação das ocorrências de nascidos vivos

## MÉTODOS

Em Ecologia, o método mais simples consiste em realizar uma amostra da população, marcar os indivíduos, permitir que estes se misturem ao restante da população e, em seguida, realizar nova amostra. Os indivíduos marcados e recapturados são contados e, a partir do número de indivíduos presentes somente na primeira amostra ( $n_A$ ), somente na segunda amostra ( $n_B$ ) e daqueles presentes nas duas ( $n_{A \cap B}$ ), estima-se o tamanho total da população. Para o emprego dessa técnica, os seguintes pressupostos devem ser assumidos:<sup>4,8</sup>

1. a população é fechada, ou seja, não há nascimentos, óbitos e migrações no período compreendido entre as amostras;
2. a marcação é unívoca e inequívoca, isto é, cada indivíduo é identificado pela marcação e não há possibilidade de “perda” dessa marca;
3. em cada amostra, cada indivíduo tem a mesma probabilidade de pertencer à amostra (equi-probabilidade);
4. as duas amostras são independentes, ou seja, os eventos “um indivíduo é capturado em uma amostra” e “um indivíduo é capturado em outra amostra” são estocasticamente independentes; e

5. em cada amostra, qualquer indivíduo é capturado (recapturado) independentemente dos demais.

A idéia é que, se em determinado território a população é pequena, ao se realizar a segunda amostra haverá um grande número de indivíduos marcados na primeira amostra. Por outro lado, se a população é grande, então, na segunda amostra haverá um número pequeno de indivíduos marcados na primeira amostra.

Em epidemiologia, cada lista disponível é considerada uma amostra da população, e “estar registrado em uma lista” é equivalente a “ser capturado” na amostra. Para mais detalhes sobre o desenvolvimento do método em epidemiologia, citam-se os trabalhos publicados: Coeli et al,<sup>2</sup> Hook & Hegal,<sup>4,5</sup> *International Working Group for Disease Monitoring and Forecasting (IWGDMF)*,<sup>8,9</sup> Wittes et al<sup>17</sup> e Wittes & Sidel.<sup>18</sup>

Huggins<sup>6,7</sup> introduziu um procedimento para a estimação do tamanho de uma população fechada quando as probabilidades de captura são heterogêneas, modelando-as em termos de variáveis observáveis, como sexo, peso, história de captura, entre outros. A modelagem é realizada construindo a verossimilhança condicionada aos indivíduos capturados para estimar os parâmetros.

Seja  $p_{ij}$  a probabilidade de o indivíduo  $i$  ser capturado na amostra  $j$ , em que  $i = 1, 2, 3, \dots, N$  são os indivíduos da população ( $N$  é o tamanho da população) e  $j = 1, 2, \dots, t$  são as amostras. A verossimilhança condicionada aos indivíduos capturados pode ser escrita em termos de:

$$\gamma_{ij} = \frac{p_{ij}}{1 - (1 - z_{ij}) \prod_{i=j}^t (1 - p_{ij}^*)}$$

em que  $p_{ij}^*$  é  $p_{ij}$  quando  $z_{ij} = 0$ , com  $z_{ij}$  sendo o indicador da história anterior de captura do indivíduo  $i$ . Ou seja:

$$z_{ij} = \begin{cases} 1, & \text{se o indivíduo } i \text{ foi} \\ & \text{capturado antes da amostra } j \\ 0, & \text{se o indivíduo } i \text{ não foi} \\ & \text{capturado antes da amostra } j \end{cases}$$

Então,  $\gamma_{ij}$  é a probabilidade de o indivíduo  $i$  ser capturado na amostra  $j$  dado sua história anterior de captura e dado que foi capturado pelo menos uma vez durante o estudo.

Admitindo-se agora que  $x_{ij} = 1$  se o indivíduo  $i$  é capturado na amostra  $j$  e  $x_{ij} = 0$  se não o for, e renomeando os indivíduos capturados como  $1, 2, 3, \dots, n$  e os não capturados como  $n + 1, n + 2, n + 3, \dots, N$ , a verossimilhança condicional é proporcional a:

$$L = \prod_{i=1}^n \prod_{j=1}^t \gamma_{ij}^{x_{ij}} (1 - \gamma_{ij})^{(1-x_{ij})}$$

Que depende somente dos indivíduos presentes nas amostras. A função de ligação para o ajuste linear segundo características dos indivíduos e/ou ambiente é a logística  $\{\ln[p_{ij}/(1-p_{ij})]\}$ .<sup>7</sup> De acordo com o autor, os estimadores são assintoticamente normais e suas variâncias podem ser estimadas a partir da matriz das segundas derivadas. Vários modelos podem ser ajustados a partir das variáveis observáveis e da história de captura.

Para estimar o tamanho da população, a probabilidade de o indivíduo  $i$  ser capturado pelo menos uma vez durante o estudo será:<sup>7</sup>

$$P_i(\beta) = 1 - \prod_{j=1}^t (1 - p_{ij}^*)$$

em que  $\beta$  é o vetor dos parâmetros associados ao modelo ajustado. Um estimador não-viciado do tamanho da população será:

$$\hat{N}(\beta) = \sum_{i=1}^n p_i(\beta)^{-1}$$

E a variância:

$$\text{var}[\hat{N}(\beta)] = \sum_{i=1}^n p_i(\beta)^{-2} [1 - P_i(\beta)]$$

O erro-padrão de  $\hat{N}(\beta)$  é a raiz quadrada de sua variância. O intervalo com 95% de confiança é:

$$IC95\% = \hat{N}(\beta) \pm 1,96 * \text{ep}[\hat{N}(\beta)]$$

Os dados foram obtidos do Ministério da Saúde (dados preliminares de 2006 da base de dados do Sistema de Informações sobre Nascidos Vivos – Sinasc) e do Instituto Brasileiro de Geografia e Estatística (IBGE – Registro Civil de Nascidos Vivos do ano de 2006). Em 2006, o IBGE incluiu o número da Declaração de Nascido Vivo (DN) em seu formulário de coleta, a partir do qual é possível realizar o pareamento entre as duas bases e assim identificar os nascimentos vivos que não foram capturados por uma delas. Os detalhes sobre a limpeza, padronização e relacionamento entre os bancos foram descritos anteriormente.<sup>e</sup>

Estimou-se o número total de nascidos vivos com os modelos de Huggins, ajustados para o segundo e terceiro trimestres de 2006, em Sergipe. O sub-registro civil foi calculado a partir das estimativas de totais de nascidos vivos. Cada fonte de dados foi considerada uma amostra, ou ocasião, sendo o Sinasc a primeira ocasião (primeira captura) e o Registro Civil a segunda ocasião (recaptação). Para a desagregação geográfica, considerou-se a microrregião de residência da mãe. Cada microrregião foi considerada como grupo de indivíduos.

Vários fatores podem ser determinantes para o sub-registro de nascidos vivos, e algumas características

poderiam ser utilizadas nos modelos, como a escolaridade da mãe, raça, número de filhos tidos anteriormente pela mãe e a existência ou não de água encanada e esgotamento sanitário no domicílio de residência, por exemplo. Entretanto, para o emprego dessa técnica, é necessário que as co-variáveis individuais incluídas no modelo linear estejam disponíveis em todas as ocasiões de captura,<sup>7</sup> no caso, as duas fontes de dados. Somente o sexo do nascido vivo e a idade da mãe estavam disponíveis nas duas bases e foram considerados para os modelos de estimação. A inexistência de cartório de registro de pessoas naturais no município é um fator institucional que pode dificultar o registro civil do nascido vivo; em Sergipe, somente dois municípios não possuíam esse tipo de cartório, em 2006.<sup>f</sup> Dessa maneira, esse fator não foi considerado, uma vez que o nível de desagregação geográfica aplicado foi de microrregiões e, nesse nível, os cartórios se fazem presentes em todas as subdivisões.

Entre 1/4/2006 e 30/9/2006, o Sinasc capturou 19.502 nascidos vivos e o Registro Civil, 17.254. O pareamento entre as bases por meio do número da DN gerou 15.532 pares. A partir desse pareamento, as duas bases somam 21.224 registros de nascidos vivos de mães residentes em Sergipe, no segundo e terceiro trimestres de 2006. O Registro Civil, no período analisado, apresentou 808 registros com o número da DN ignorada, representando 4,7% dos registros nessa base. Quando se comparam os registros com DN preenchida e os que têm a DN ignorada, não há diferença estatística na média de idade da mãe ( $p = 0,992$ ) nem na proporção do sexo do nascido vivo ( $p = 0,510$ ).

A distribuição de registros com número de DN ignorada, no Registro Civil, revelou que algumas microrregiões tinham percentual de registros com o número da DN em branco acima de 5%: Agreste de Lagarto (31,9%), Tobias Barreto (9,2%), Boquim (8,2%) e Japaratuba (7,9%). Nas demais microrregiões variou de 0,4% em Nossa Senhora das Dores a 4,0% em Carira.

Para o sexo do nascido vivo, testou-se a hipótese nula de que a proporção de meninas é a mesma na microrregião/regional de saúde avaliada, em relação à proporção do estado como um todo. Já para a idade da mãe, a hipótese nula é de que a média de idade na microrregião/regional de saúde é a mesma de Sergipe. A proporção de crianças do sexo feminino não se mostrou estatisticamente diferente da média da unidade da federação. Já a média da idade da mãe diferiu estatisticamente. Portanto, apenas a idade da mãe foi considerada para inclusão nos modelos de estimação.

Considerando  $i = 1, 2, 3, \dots, N$  os nascidos vivos e  $b = \text{Sinasc, Registro Civil}$ , as bases de dados, o modelo

<sup>e</sup> Schmid B. Aplicação do método de captura-recaptação para estimar sub-registro de eventos vitais [tese de doutorado]. São Paulo: Universidade de São Paulo; 2010.

<sup>f</sup> Instituto Brasileiro de Geografia e Estatística. Estatísticas do registro civil. Rio de Janeiro; 2006. v. 33.

linear inicial (completo) para as probabilidades de captura pelo Sinasc e pelo Registro Civil foram expressos como:

$$\ln\left(\frac{p_{ib}}{1-p_{ib}}\right) = \beta_0 + \beta_1 g_1 + \beta_2 g_2 + \dots + \beta_{12} g_{12} + \beta_{13} \text{idmae}_i + \beta_{13}(g_1 * \text{idmae}_i) + \beta_{15}(g_2 * \text{idmae}_i) + \dots + \beta_{25}(g_{12} * \text{idmae}_i) \quad (1)$$

em que

$p_{ib}$  é a probabilidade de o indivíduo  $i$  estar na base  $b$  (Sinasc ou Registro Civil);

$\beta_0$  é o parâmetro do intercepto;

$\beta_k$  é o parâmetro estimado para o grupo  $k$  ( $k$  é a microrregião);

$g_k$  são os indivíduos que pertencem ao grupo  $k$ ;

$\beta_{13}$  é o parâmetro estimado para a idade da mãe;

$\text{idmae}_i$  é a idade da mãe do indivíduo  $i$ , em anos; e

$\beta_{k+13}$  é o parâmetro estimado para a interação entre o grupo  $k$  e a idade da mãe.

Nesse modelo a probabilidade de captura varia em função das características individuais. A notação adotada para o modelo completo foi  $[p(g+\text{idmae}+g*\text{idmae})c(g+\text{idmae}+g*\text{idmae})]$ .

Cada submodelo, de acordo com sua especificação, estimará um número determinado de parâmetros. Por exemplo, um modelo cuja probabilidade de captura pelo Sinasc depende do grupo (desagregação geográfica) e da idade da mãe e a probabilidade de captura pelo Registro Civil depende somente da idade da mãe  $[p(g + \text{idmae})c(\text{idmae})]$ , terá um número diferente de parâmetros estimados de outro modelo no qual as probabilidades de captura pelo Sinasc e pelo Registro não variam ao longo das microrregiões e são independentes da idade da mãe,  $[p(\cdot)c(\cdot)]$ . Quanto maior o número de parâmetros de um modelo, melhor este se ajustará aos dados, porém a precisão das estimativas dos parâmetros diminui.

Um método para se realizar um balanço entre ajuste e precisão é avaliar os modelos segundo critérios de informação. Um desses métodos é o *Akaike Information Criteria* (AIC), que consiste em se relacionar a verossimilhança do modelo (representando o seu ajuste) com o número de parâmetros que ele estima:

$$\text{AIC} = -2\ln(L) + 2k$$

em que  $L$  é a verossimilhança do modelo e  $k$  o número de parâmetros. Quanto maior o ajuste do modelo, maior será sua verossimilhança, diminuindo o valor  $[-2\ln(L)]$ . O termo aditivo  $[+2k]$  penaliza o valor do AIC. Quanto

maior o número de parâmetros, menor será o AIC, pois a verossimilhança aumentará. Porém, a soma do termo  $[2k]$  balanceia o valor do AIC. Assim, o modelo com menor AIC será aquele mais parcimonioso, em relação à sua verossimilhança e número de parâmetros.<sup>§</sup>

Apesar de simples a interpretação e escolha do modelo que melhor se ajusta aos dados, por vezes pode haver modelos com valores AIC muito próximos, o que dificulta a escolha. Os modelos podem, então, ser calibrados de maneira a fornecer um índice de relativa plausibilidade, empregando-se os pesos normalizados de Akaike. Esses pesos,  $w_i$ , são calculados para cada modelo no conjunto  $I$  de modelos candidatos de acordo com a fórmula:

$$w_i = \frac{\exp\left(\frac{-2\Delta\text{AIC}_i}{2}\right)}{\sum_{i=1}^I \left\{\exp\left(\frac{-2\Delta\text{AIC}_i}{2}\right)\right\}}$$

em que  $\Delta\text{AIC}_i$  é a diferença entre o valor de AIC entre o modelo  $i$  e o valor do modelo com menor AIC. O peso  $w_i$  é considerado como uma evidência de que o modelo  $i$  é o melhor modelo no conjunto dos candidatos. Quanto maior o peso do modelo, pode-se interpretar que ele melhor suporta os dados.<sup>f</sup>

Os modelos foram ajustados no programa MARK,<sup>®</sup> que estima as probabilidades de captura e recaptura, de acordo com o modelo linear que se deseja ajustar para cada probabilidade. Dessa maneira, vários modelos foram ajustados, em que a probabilidade de captura pelo Sinasc foi ajustada desde como uma constante até o modelo completo, descrito na equação (1). O mesmo procedimento foi adotado para a probabilidade de captura pelo Registro Civil, totalizando 30 submodelos para as microrregiões. Ainda, ajustaram-se quatro modelos para todo o estado de Sergipe, nos quais nenhum grupo foi considerado, somente foi investigada a influência da idade da mãe na captura pelas bases.

O Programa MARK<sup>®</sup> calculou o número de parâmetros de cada modelo, assim como a verossimilhança, o valor de AIC,  $\Delta\text{AIC}$ ,  $w_i$ , as estimativas das probabilidades de captura (Sinasc) e recaptura (Registro Civil), e também a estimativa derivada do total de nascidos vivos ( $\hat{N}$ ).

Uma vez obtidas as estimativas do total de nascidos vivos, partiu-se para o cálculo do sub-registro civil. O estimador é dado na forma de percentual:

$$\text{SR}_s = \frac{(\hat{N}_s - n_{\text{RCS}})}{\hat{N}_s} \times 100 \quad (2)$$

em que

$\hat{\text{SR}}_s$  = Percentual de sub-registro na subdivisão  $s$  (microrregião).

<sup>§</sup> Cooch E, White G, organizers. Program MARK: a gentle introduction. Ithaca: Cornell University; 2008 [citado 2009 mar 02]. Disponível em: <http://www.phidot.org/software/mark/docs/book>

$\hat{N}_s$  = Total de nascidos vivos estimados para a subdivisão s.

$n_{RCs}$  = Número de nascidos vivos capturados pelo Registro Civil, na subdivisão s.

O cartograma foi confeccionado no aplicativo Tabwin®.

## RESULTADOS

Dentre os quatro modelos ajustados para o estado de Sergipe como um todo, o modelo com maior peso  $\{p(idmae) c(\cdot)\}$  foi aquele em que a idade da mãe interfere na captura de nascidos vivos pelo Registro Civil, com cerca de 66% do peso total entre os modelos.

Esse modelo estimou 21.391 (IC95% 21.363;21.423) nascidos vivos no segundo e terceiro trimestres de 2006 em Sergipe, sendo a probabilidade de captura pelo Sinasc ( $\hat{p}$ ) estimada em 0,912 e a do Registro Civil, 0,804. Derivando as estimativas, a partir do número de

nascidos vivos, calculou-se que o sub-registro civil foi 19,3%. Ao aplicar o estimador dado em (2), nos limites do IC95% do total de nascidos vivos, estimou-se uma variação no sub-registro entre 19,2% e 19,5%.

Ao se incluírem as microrregiões de residência da mãe como agrupamentos de nascidos vivos, dos 30 modelos ajustados aos dados, apenas cinco mostraram algum peso relativo quando se avaliam os critérios AIC e a verossimilhança. O modelo com maior peso  $\{p(g) c(g + idmae + g*idmae)\}$ , 67%, foi selecionado (Tabela 1).

A probabilidade de captura pelo Sinasc foi alta em todas as microrregiões, variando entre 0,69 no Agreste de Lagarto até 0,95 em Estância e Nossa Senhora das Dores. Destaca-se que em Agreste de Lagarto há um grande número de registros no Registro Civil sem o número da DN preenchido (mais de 31%), o que prejudicou o pareamento das bases. Também em Tobias Barreto, Japaratuba e Boquim, o percentual de registros com a DN em branco foi maior do que 5%. Além dessas microrregiões, nas quais o pareamento foi afetado pela falta do número da DN, somente em Propriá a probabilidade de captura pelo Sinasc foi menor do que 0,90 (Tabela 2).

As probabilidades de captura pelo Registro Civil foram sensivelmente menores do que as do Sinasc. No Registro Civil, a maior probabilidade de captura estimada foi na microrregião de Aracaju (0,85) e a menor foi na Sergipana do Sertão do São Francisco (0,71), excetuando-se as microrregiões com pareamento problemático (Tabela 2).

O total de nascidos vivos estimados ( $\hat{N}$ ) foi bem próximo ao total medido ( $r$ ) em todas as microrregiões,

**Tabela 1.** Resultados dos modelos de Huggins para populações fechadas de nascidos vivos, que apresentaram algum peso, segundo as microrregiões de residência da mãe. Sergipe, segundo e terceiro trimestres de 2006.

Modelo	$w_i$	Nº parâmetros
$\{p(g) c(g + idmae + g*idmae)\}$	0,66837	39
$\{p(g + idmae) c(g + idmae + g*idmae)\}$	0,27672	40
$\{p(g + idmae + g*idmae) c(g + idmae + g*idmae)\}$	0,05485	52
$\{p(g) c(g + idmae)\}$	0,00004	27
$\{p(g + idmae) c(g + idmae)\}$	0,00002	28

**Tabela 2.** Estimativas da probabilidade de captura pelo SINASC ( $\hat{p}$ ) e probabilidade de captura pelo Registro Civil ( $\hat{c}$ ) do modelo  $\{p(g) c(g + idmae + g*idmae)\}$ , segundo a microrregião de residência da mãe. Sergipe, segundo e terceiro trimestres de 2006.

Microrregião	$\hat{p}$	IC95%	$\hat{c}$	IC95%
Sergipana do Sertão do São Francisco	0,912	0,896;0,926	0,714	0,692;0,735
Carira	0,937	0,912;0,956	0,822	0,787;0,852
Nossa Senhora das Dores	0,949	0,926;0,965	0,823	0,789;0,852
Agreste de Itabaiana	0,932	0,917;0,944	0,861	0,843;0,878
Tobias Barreto <sup>a</sup>	0,874	0,849;0,896	0,719	0,690;0,746
Agreste de Lagarto <sup>a</sup>	0,694	0,655;0,731	0,516	0,486;0,546
Propriá	0,876	0,849;0,898	0,807	0,779;0,832
Cotinguiba	0,914	0,883;0,938	0,829	0,793;0,860
Japaratuba <sup>a</sup>	0,874	0,839;0,903	0,775	0,738;0,809
Baixo Cotinguiba	0,921	0,900;0,938	0,870	0,847;0,890
Aracaju	0,944	0,938;0,949	0,848	0,840;0,856
Boquim <sup>a</sup>	0,886	0,868;0,902	0,749	0,728;0,770
Estância	0,949	0,935;0,960	0,799	0,777;0,820

<sup>a</sup> Estas microrregiões tiveram mais de 5% dos registros no Registro Civil com o número da DN em branco, o que prejudicou o pareamento.

devido à alta sobreposição das listas. Novamente no Agreste de Lagarto, o alto percentual de registros com DN em branco, no Registro Civil, levou a uma sobreposição menor ( $n_{S \cap RC}$  relativamente baixo) e conseqüentemente inflou a estimativa do total de nascidos vivos. A diferença absoluta entre os nascidos vivos estimados e capturados foi menor do que 20 em quase todas as microrregiões nas quais o percentual de DNs em branco no Registro Civil foi menor do que 5%, chegando a apenas 2 nascidos vivos em Carira e Nossa Senhora das Dores (Tabela 3).

O sub-registro ao longo das microrregiões variou entre pouco mais de 12% no Baixo Contiguiba até quase 27% na Sergipana do Sertão do São Francisco. O sub-registro estimado no Agreste de Lagarto superou os 40%, porém esse resultado deve ser avaliado com cautela. O sub-registro civil foi menor nas microrregiões localizadas na parte central do estado, Aracaju, Baixo Contiguiba e Agreste de Itabaiana (< 15% dos nascidos vivos). À medida que as microrregiões de residência da mãe se afastam dessas partes centrais, o sub-registro civil de nascidos vivos aumenta (Figura 1).

De acordo com a idade média das mães no Registro Civil, houve sutil tendência de diminuição do sub-registro à medida que a idade da mãe aumenta (Figura 2). A microrregião do Agreste de Lagarto não foi incluída na análise devido ao grande número de omissões da DN na base do Registro Civil.

## DISCUSSÃO

Os resultados obtidos sugerem que, quanto menor a idade da mãe, maior a probabilidade de sub-registro

civil (Figura 2), a qual também é influenciada pela proximidade com a área central do estado. A microrregião do Baixo Contiguiba, apesar de ter a média de idade das mães menor, não revelou sub-registro entre os mais elevados; o contrário ocorre em Tobias Barreto, microrregião mais afastada de Aracaju, que tem a idade média maior, mas também alto sub-registro. Entretanto, o pareamento em Tobias Barreto foi prejudicado por haver quase 10% dos registros na base do IBGE sem o número da DN.

É preciso discutir os pressupostos assumidos para a estimação pela captura-recaptura. Toma-se como uma população fechada aquela em que não há migração nem nascimentos ou mortes durante o período de estudo. Neste trabalho, a população que se quis estimar foi o número total de nascidos vivos. O evento nascer é único e número de nascidos vivos é fixo no período e na delimitação geográfica adotada. Claramente, óbitos neonatais e/ou infantis ocorreram no período analisado, assim como as famílias dos bebês podem ter mudado a residência para outro município ou unidade da federação, após o parto. Porém, esses fatos não alteram o tamanho da “população de nascidos vivos de mães residentes em Sergipe”, uma vez que óbito e mudança de endereço não podem modificar o fato de que, naquele momento, o bebê nasceu vivo e a mãe residia naquela localidade.

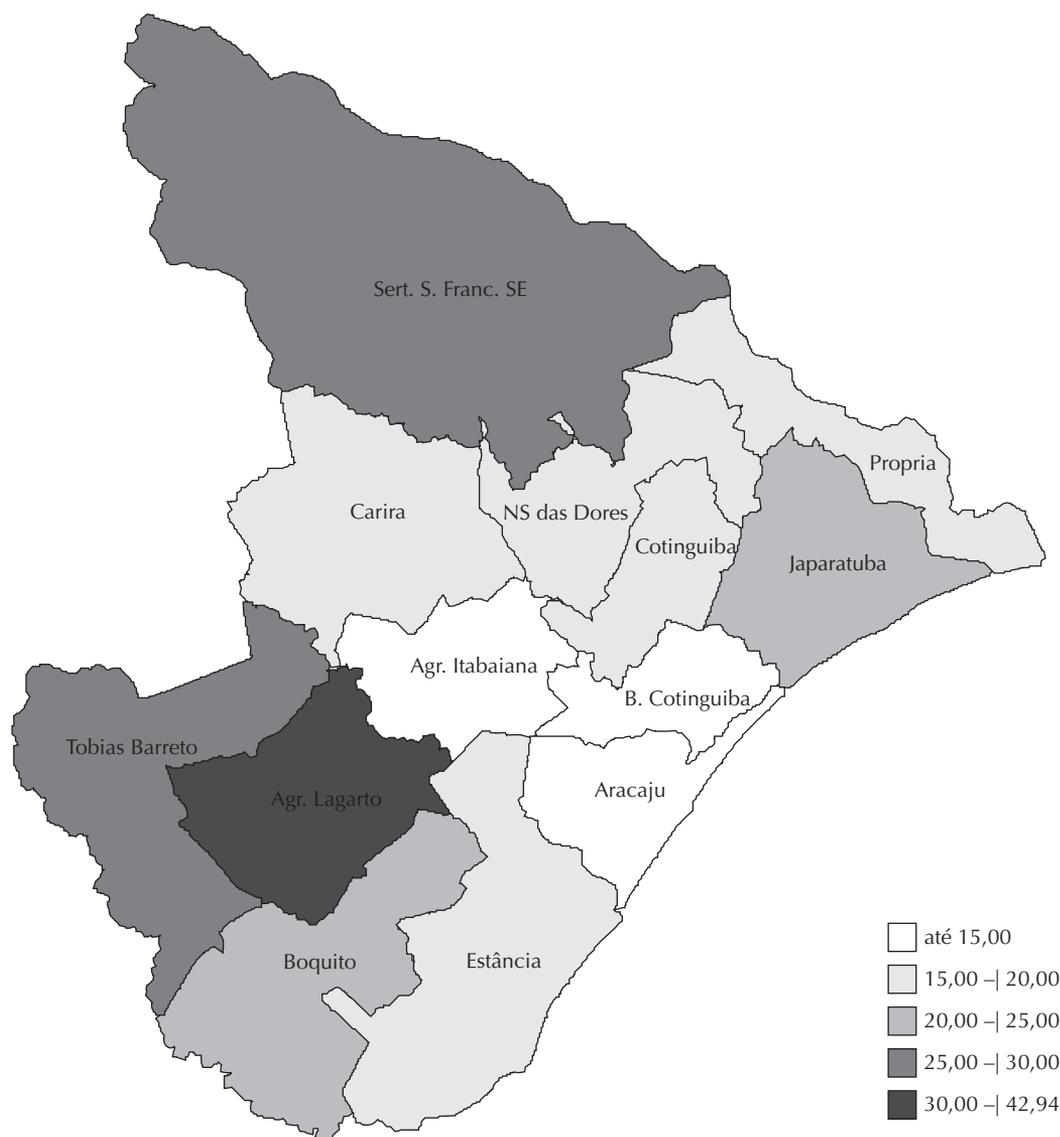
Em relação à marcação unívoca e inequívoca, as falhas no preenchimento do número da DN prejudicaram o relacionamento determinístico adotado. Dos 17.254 nascidos vivos presentes no Registro Civil, 808 (4,7%) tinham o campo DN em branco, o que leva a questionar quantos desses 808 registros no RC sem o número da DN são pareáveis com os registros do Sinasc.

**Tabela 3.** Distribuição da captura pelas duas bases [ $n_s$ ,  $n_{RC}$ ,  $n_{(S \cap RC)}$  e  $r$ ] e estimativas derivadas do modelo  $\{p(g) c(g + idmae + g * idmae)\}$ , segundo a microrregião de residência da mãe. Sergipe, segundo e terceiro trimestres de 2006.

Microrregião	$n_s$	$n_{RC}$	$n_{(S \cap RC)}$	$r$	$\hat{N}$	IC95%
Sergipana do Sertão do São Francisco	1.692	1.359	1.210	1.841	1.855	1.849;1.867
Carira	541	479	445	575	577	576;583
Nossa Senhora das Dores	565	493	464	594	596	594;601
Agreste de Itabaiana	1.505	1.399	1.296	1.608	1.616	1.611;1.625
Tobias Barreto <sup>a</sup>	986	833	709	1.110	1.128	1.120;1.142
Agreste de Lagarto <sup>a</sup>	1.056	868	545	1.379	1.521	1.482;1.575
Propriá	885	821	711	995	1.011	1.003;1.024
Cotinguiba	491	448	406	533	537	534;545
Japarutuba <sup>a</sup>	549	493	424	618	628	623;640
Baixo Cotinguiba	949	901	826	1.024	1.030	1.027;1.039
Aracaju	7.286	6.590	6.183	7.693	7.717	7.708;7.730
Boquim <sup>a</sup>	1.661	1.434	1.245	1.850	1.874	1.865;1.890
Estância	1.336	1.136	1.068	1.404	1.408	1.405;1.414

$n_s$  = total de nascidos vivos capturados pelo SINASC.  $n_{RC}$  = total de nascidos vivos capturados pelo Registro Civil.  $r$  = total de nascidos vivos distintos identificados ( $r = n_s + n_{RC} - n_{(S \cap RC)}$ )

<sup>a</sup> Estas microrregiões que tiveram mais de 5% dos registros no Registro Civil com o número da DN em branco, o que prejudicou o pareamento.

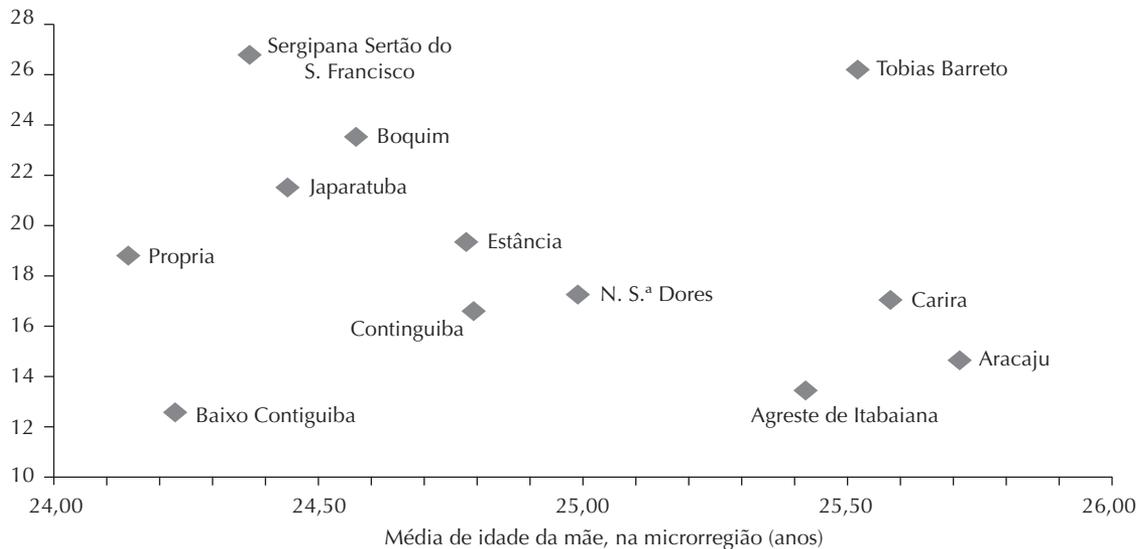


**Figura 1.** Sub-registro civil, segundo a microrregião de residência da mãe. Sergipe, segundo e terceiro trimestres de 2006.

Tomando como ponto de partida os resultados do modelo para todo o estado de Sergipe, a probabilidade estimada de captura pelo Sinasc é de 0,912 e a do Registro Civil é de 0,804. A probabilidade de um nascido vivo figurar nas duas bases, então, será  $0,912 \times 0,804 = 0,733$ . Logo, dos 808 registros sem número da DN, no Registro Civil,  $0,733 \times 808 = 592$  também teriam sido capturados pelo Sinasc. A partir dessa conta, aventa-se que o número correto de pares formados pelo relacionamento entre as duas bases seria  $15.532 + 592 = 16.124$  nascidos vivos. Apesar de ser possível estimar o provável número de pares, realizar tal relacionamento pós-estimação torna-se tarefa imprecisa, pois as variáveis disponíveis têm pouco poder discriminatório, o que implica encontrar mais de um registro com as mesmas características. O ano de 2006 foi o primeiro

de coleta do número da DN nos questionários do IBGE e espera-se que nos anos subsequentes haja melhora na qualidade dos dados, com menores discrepâncias entre a base do Sinasc e do Registro Civil.

Há duas questões relacionadas à equi-probabilidade, isto é, cada indivíduo tem a mesma probabilidade de ser capturado em uma dada amostra. A primeira questão refere-se à influência da idade da mãe na captura pelo Registro Civil. Entretanto, essa “fonte” de heterogeneidade pode ser incluída nos modelos e, de fato, mostrou-se significativa. A segunda remete ao intervalo de tempo tomado para o pareamento das bases. O banco de dados do Registro Civil no IBGE é organizado por ano de vigência da coleta dos dados. Os registros e suas respectivas datas de nascimento não são perdidos, porém constam no ano em que



Nota: Excluída a microrregião do Agreste de Lagarto, devido ao pareamento prejudicado.

**Figura 2.** Sub-registro civil nas microrregiões de residência da mãe. Sergipe, segundo e terceiro trimestres de 2006.

foram coletados. Dessa maneira, o número do total de nascidos vivos em 2006 irá variar à medida que os anos de coleta dos dados avançam. É evidente que essa variação não será muito grande, pois os registros tardios serão residuais com o passar dos anos. Como salientam Oliveira & Simões,<sup>h</sup> a cobertura do Registro Civil é sensivelmente incrementada quando se analisam os registros ocorridos até um ano após o nascimento vivo. Para aqueles que nasceram no segundo trimestre de 2006, cerca de 94% foram registrados até o terceiro trimestre. Como o acesso aos dados ficou restrito ao ano de coleta de 2006, os nascidos no terceiro trimestre de 2006 e registrados a partir de 2007 em diante não foram detectados neste trabalho. Acredita-se, então, que não houve equi-probabilidade de captura pelo Registro Civil entre os indivíduos que nasceram no terceiro trimestre de 2006 em relação aos que nasceram no segundo trimestre daquele ano. Uma vez que, como mencionado, o acesso ao banco do IBGE foi restrito ao ano de coleta de 2006, ressalta-se que o pareamento entre o Sinasc e o RC para os nascidos vivos no terceiro trimestre apresenta mais um revés, além da falta do número da DN. Deduz-se, então, que o número de pares, representado na tabela 2 x 2 de captura por  $n_{(A \cap B)}$ , será maior do que o medido, evidenciando ainda mais a grande sobreposição das duas bases.

Em relação ao Sinasc, os dados são consolidados a partir do envio das informações das secretarias estaduais de saúde para o Ministério da Saúde e eventualmente também são atualizados.<sup>i</sup> Assim, os números

de nascidos vivos ora disponíveis no sítio da internet também diferem dos dados preliminares disponibilizados para este trabalho, em 2007. Deve-se levar em consideração o dinamismo dos dois bancos de dados, sendo sempre abordada essa contínua mudança nos totais de registros.

O pressuposto de que as probabilidades de captura-recaptura de um indivíduo não afetam as probabilidades dos demais não foi violado, porque o nascimento de um bebê não deve interferir nem na identificação de outro no Sinasc, nem tampouco no registro civil. Em relação aos partos gemelares – o fato de os nascidos vivos serem registrados, ou não, em conjunto – não atenderia a esse pressuposto. Entretanto, a ocorrência de gestações gemelares é muito rara, e o número de gêmeos nascidos vivos não chega a prejudicar tal assunção.

Quanto à independência entre as amostras, aqui bases de dados, a grande sobreposição entre elas ( $n_{A \cap B}$  grande) sugere que há dependência positiva. Isso significa que o número de nascidos vivos estimados não será muito maior do que aqueles indivíduos distintos identificados nas duas bases. Com o objetivo de quantificar o vício gerado pela dependência entre duas fontes (listas) epidemiológicas, Brenner<sup>1</sup> incluiu um fator de correção, para a probabilidade de o indivíduo ser capturado nas duas listas. O autor simula situações nas quais variam as probabilidades de captura em cada fonte ( $n_A$  e  $n_B$ ) e o valor dos fatores de correção que modificam as probabilidades de inclusão nas duas fontes ( $n_{A \cap B}$ ), gerando dependência positiva e negativa, de maneira a observar

<sup>h</sup> Oliveira ATR, Simões CCS. Perfil dos municípios com informações precárias sobre eventos vitais. Brasília: Rede Interagencial de Informações para a Saúde, Organização Pan-Americana da Saúde; 2005.

<sup>i</sup> Ministério da Saúde. Departamento de Informática do SUS – DATASUS. Informações em Saúde. Brasília; 2009 [citado 2009 maio 20]. Disponível em: <http://www.datasus.gov.br/>

o comportamento do fator de sub e superestimação, respectivamente, das estimativas obtidas pelo método de captura-recaptura. Para o caso da dependência negativa, o investigador conclui que o vício de superestimação do tamanho total da população será mais grave quando as listas têm baixa cobertura de indivíduos, ou seja, têm pequena probabilidade de inclusão destes. O autor afirma que, no caso da dependência positiva, listas com probabilidade alta de inclusão têm fatores de subestimação menores e que, mesmo conscientes desse tipo de dependência, as estimativas do tamanho da população ainda serão mais próximas da realidade do que a simples agregação das fontes.<sup>1</sup>

No tocante ao modelo escolhido, Tilling & Sterne<sup>15</sup> e Tilling et al<sup>16</sup> aplicaram o modelo de Huggins para estimação em dados epidemiológicos, mostrando a viabilidade do modelo para esses tipos de dados. Ainda, a abordagem da verossimilhança condicionada aos indivíduos observados dá a flexibilidade da inclusão de co-variáveis para modelar as probabilidades de captura, por meio de ajustes de modelos lineares.

Acredita-se que o modelo de Huggins possa continuar a ser empregado para a estimação do total de nascidos vivos em associações futuras do Sinasc e do Registro Civil, desde que seja resolvido o problema da equi-probabilidade de captura pela base do IBGE no que tange ao período empregado para a identificação dos registros, enfrentado neste trabalho.

Em conclusão, os resultados do presente estudo sugerem valores mínimos de sub-registro civil e de cobertura do Sinasc, sendo possível a aplicação da metodologia de captura-recaptura para estimar sub-registro de nascidos vivos. No caso de grande sobreposição entre duas bases, o *International Working Group for Disease Monitoring and Forecasting*<sup>9</sup> recomenda que sejam agregadas e transformadas em uma só fonte. A saída seria o relacionamento determinístico entre o Sinasc e o Registro Civil e o relacionamento probabilístico entre essa nova base de dados e outras que podem ser empregadas para a aplicação da captura-recaptura, como os cadastros dos Programas de Saúde da Família e o Sistema de Informações Hospitalares, por exemplo.

## REFERÊNCIAS

1. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology*. 1995;6(1):42-8. DOI:10.1097/00001648-199501000-00009
2. Coeli CM, Veras RP, Coutinho ESF. Capture-recapture methodology: an option for surveillance of non-communicable diseases in the elderly. *Cad Saude Publica*. 2000;16(4):1071-82. DOI:0.1590/S0102-311X2000000400025
3. Dunn J, Andreoli SB. Método de captura e recaptura: nova metodologia para pesquisas epidemiológicas. *Rev Saude Publica*. 1994;28(6):449-53. DOI:10.1590/S0034-89101994000600009
4. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev*. 1995;17(2):243-64. Erratum in: *Am J Epidemiol*. 1998;148(12):1219.
5. Hook EB, Regal RR. Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *J Clin Epidemiol*. 1999;52(10):917-26. discussion 929-33.
6. Huggins RM. On the statistical analysis of capture experiments. *Biometrika*. 1989;76(1):133-40. DOI:10.1093/biomet/76.1.133
7. Huggins RM. Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*. 1991;47(2):725-32. DOI:10.2307/2532158
8. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation I: history and theoretical development. *Am J Epidemiol*. 1995;142(10):1047-58.
9. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation II: applications in human diseases. *Am J Epidemiol*. 1995;142(10):1059-68.
10. Laporte RE. Assessing the human condition: capture-recapture techniques. *BMJ*. 1994;308(6920):5-6.
11. Mccarty DJ, Tull ES, Moy CS, Kwoh CK, Laporte RE. Ascertainment corrected rates: applications of the capture-recapture methods. *Int J Epidemiol*. 1993;22(3):559-65. DOI:10.1093/ije/22.3.559
12. Pradel R. Utilization of capture-recapture for the study of recruitment and population growth rate. *Biometrics*. 1996;52(2):703-9. DOI:10.2307/2532908
13. Sekar CC, Deming WE. On a method of estimating birth and death rates and the extent of registration. *J Am Stat Assoc*. 1949;44(245):101-15.
14. Shapiro S. Estimating birth registration completeness. *J Am Stat Assoc*. 1949;45:261-4. DOI:10.2307/2280684
15. Tilling K, Sterne JAC. Capture-recapture models including covariate effects. *Am J Epidemiol*. 1999;149(4):392-400.
16. Tilling K, Sterne JAC, Wolfe CDA. Estimation of the incidence of stroke using a capture-recapture model including covariates. *Int J Epidemiol*. 2001;30(6):1351-9. DOI:10.1093/ije/30.6.1351
17. Wittes JT, Colton T, Sidel VW. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *J Chronic Dis*. 1974;27(1):25-36. DOI:10.1016/0021-9681(74)90005-8
18. Wittes JT, Sidel VW. A generalization of the simple capture-recapture model with applications to epidemiological research. *J Chronic Dis*. 1968;21(5):287-301. DOI:10.1016/0021-9681(68)90038-6

---

Artigo baseado na tese de doutorado de Schmid B, apresentada à Faculdade de Saúde Pública da Universidade de São Paulo em 2010.

Os autores declaram não haver conflitos de interesses.