Research note

Analysis of the application of a deterministic routine for identifying multiple pregnancies on Live Birth Information Systen, Brazil*

doi: 10.5123/S1679-49742020000200001

Fernanda Pinheiro Aguiar[1] - orcid.org/0000-0003-0197-1354
Patrícia Viana Guimarães Flores[2] - orcid.org/0000-0001-5074-5113
Luis Carlos Torres Guillen[1] - orcid.org/0000-0001-5246-733X
Helena Pereira da Silva Santos[1] - orcid.org/0000-0002-5712-590X
Luís Guilherme Santos Buteri Alves[1] - orcid.org/0000-0002-5810-8474
Kenneth Rochel de Camargo Jr.[3] - orcid.org/0000-0003-3606-5853
Rejane Sobrino Pinheiro[1] - orcid.org/0000-0002-3361-3626
Cláudia Medina Coeli[1] - orcid.org/0000-0003-1757-3940

[1]Universidade Federal do Rio de Janeiro, Programa de Pós-Graduação em Saúde Coletiva, Rio de Janeiro, RJ, Brazil
[2]Ministério da Saúde, Hospital Federal de Bonsucesso, , Rio de Janeiro, RJ, Brazil
[3]Universidade do Estado do Rio de Janeiro, Instituto de Medicina Social, Rio de Janeiro, RJ, Brazil

Abstract

Objective: to evaluate the application of a deterministic routine for identifying multiple pregnancies on the Brazilian Live Birth Information System (SINASC). **Methods:** SINASC data deduplication and linkage with the mortality database (fetal deaths) for Rio de Janeiro state for the period 2007-2008; we used a deterministic routine, using a key based on SINASC maternal and birth information, complemented by manual review. **Results:** of the 433,874 SINASC records, 9,036 (2.1%) were classified as multiple pregnancy newborns; after implementing the routine, we reclassified 385 records as twins, and 286 as singletons; accuracy of multiple pregnancy information on the SINASC database was high (sensitivity=95.8%; specificity=99.9%); applying the routine without the manual review process increased sensitivity by 4.2%, with no significant change of specificity. **Conclusion:** despite the accuracy of information regarding multiple pregnancy held on SINASC, we suggest the use of this routine as an option for improving classification of twins.

Keywords: Health Information Systems; Systems Integration; Health Evaluation; Pregnancy, Multiple.

*Article derived from the Ph.D. dissertation by Fernanda Pinheiro Aguiar entitled 'Maternal age and schooling at birth and infant mortality', defended at the Federal University of Rio de Janeiro Institute of Public Health Studies/Public Health Postgraduate Program (IESC/UFRJ), on May 22nd 2018.The authors receive research scholarships from the National Scientific and Technological Development Council/Ministry of Science, Technology, Innovation and Communications (CNPq/MCTIC) (Cláudia Medina Coeli, Process No. 447199/2014-5 and Process No. 305545/2015-9; Rejane Sobrino Pinheiro, Process No. 309728/2012-6; and Kenneth Rochel de Camargo Jr., Process No. 300686/2013-7) and from the Rio de Janeiro State Research Support Foundation (FAPERJ) (Cláudia Medina Coeli, Process No. E-26/203.195/2015; and Kenneth Rochel de Camargo Jr., Process No. E-26/102.900/2012), FAPERJ Scientific Initiation scholarship (Helena Pereira da Silva Santos, Process No. E-26/202.575/2016; and Luís Guilherme Santos Buteri Alves, Process No. E-26/219.275/2016) FAPERJ 'Nota DEZ' Ph.D. scholarship (Fernanda Pinheiro Aguiar, Process No. E-26/400.418/2016).

Correspondence:
Fernanda Pinheiro Aguiar – Rua Pajuçara, No. 600, apto. 302, Cocotá, Ilha do Governador, Rio de Janeiro, RJ, Brazil. Postcode: 21910-300
E-mail: aguiarfernandap@gmail.com

Introduction

The Live Birth Information System (SINASC) was implemented in order to bring together information on births in the entire country. With effect from 1990, SINASC has shown itself to be relevant regarding characterization and status of child deliveries and births, as well as for identifying at risk/vulnerable groups of mothers and children.[1,2]

Multiple pregnancy is a risk factor for negative outcomes at birth.3-6 The risk of multiple babies dying can be 12 times greater, when compared to the same risk for single pregnancy babies. The main explanation for this difference lies in the increased proportion of prematurity and restricted intrauterine growth in the case of twins.[7,8].

Multiple pregnancy children also have greater risk of developing undesirable conditions in the long term, such as cerebral palsy, cognitive dysfunction, language development disorders, learning difficulties, as well as psychiatric and socio-behavioral problems.[7,8] It is therefore important to identify multiple births in studies that use vital statistics databases.

> ***SINASC has a variable that indicates the number of children in the same pregnancy; however data completion errors lead to incorrect classification of information about twinning.***

Each twin is registered separately on the births database and has its own identification number. In addition, SINASC has a variable that indicates the number of children in the same pregnancy; however data completion errors lead to incorrect classification of information about twinning.[1,9,10] Database linkage is used to improve data information quality by retrieving it and confirming it in a single database (data duplication identification process) or comparing different databases.[11-14]

The objective of this study was to evaluate the application of a deterministic routine in order to identify multiple pregnancies on the SINASC database for Rio de Janeiro state for the years 2007 and 2008.

Methods

We conducted a descriptive study to evaluate improvement of information about multiple pregnancies held on the SINASC database by applying a deterministic routine (internal linkage).

We used SINASC data (N=433,882) for Rio de Janeiro state for the years 2007 and 2008. We also examined fetal death records (N=372) on the Mortality Information System (SIM), searching manually for twins, when multiple pregnancy was indicated on SINASC but with only one live birth recorded.

The deterministic routine was based on four processes: (i) record comparison (internal database linkage), using a deterministic key comprised of maternal information (soundex of mother's first name, soundex of mother's second name; soundex of mother's last name) and information about birth (full date of birth; code of the health establishment where birth occurred); (ii) automatic comparison of residential address, using a routine based on the Levenshtein edit distance; (iii) manual search for twins on the SIM system; and (iv) manual reviewing.

First of all the SINASC database was pre-processed with the aim of eliminating records with duplicated Live Birth Certificate numbers.

Records that had the same deterministic key were assessed according to information about pregnancy (single; multiple) held on SINASC. In the case of records having classification in agreement. i.e. with coinciding keys and indicating multiple pregnancy (key+/Sinasc+), and records having classification in disagreement, i.e. with coinciding keys but indicating single pregnancy (key+/Sinasc-), their addresses were compared automatically. When addresses coincided completely, the records were classed as being multiple pregnancy. When addresses were in disagreement, a manual review was performed in order to achieve final classification. During this manual stage, information regarding mother's name, maternal age, place of birth, type of delivery and type of pregnancy were used by the researcher to define classification as being or not being multiple pregnancy.

When the key did not identify twins and the information held on SINASC was for single pregnancy (key-/Sinasc-), records were classified as not being twins. In the case of records indicating multiple pregnancy but for which the key did not identify a record of twins (key-/Sinasc+), we performed a manual search on the SIM fetal deaths database to confirm for twins, given that babies from the same pregnancy might be found on different information systems. Those that were not found on the fetal deaths database underwent a manual review.

We assessed records that had their classification changed (multiple or single pregnancy) following application of the complete routine (deterministic key, address comparison, fetal deaths database search and manual review of pairs). Classification following application of the complete routine was considered to be gold standard for accuracy analyses, both for information on twins registered on SINASC and also for classified obtained by applying a reduced routine, based solely on information held on SINASC and deterministic key agreement, without performing the remaining procedures (address comparison, fetal deaths database search and manual review of pairs). In this case, SINASC records that indicated multiple pregnancy or had deterministic key record agreement were classified as multiple pregnancy. SINASC records with no information on multiple pregnancy and with no pairs indentified simultaneously by the deterministic key were considered not to be twins. We calculated sensitivity, specificity and positive predictive value and respective 95% confidence intervals (95%CI).

We used PostgreSQL 9.2 and Stata12 applications, respectively, to carry out the deterministic linkage routine and for analysis.

The study project, based on secondary data provided by the Rio de Janeiro State Health and Civil Defense Department and developed in accordance with research ethics principles, was submitted to the IESC/UFRJ Research Ethics Committee as an amendment to the project entitled 'Integrated Health Records: longitudinal evaluation of morbidity and mortality in a cohort of live born babies and their mothers - Phase 1' and was approved on October 3rd 2012 – Certification of Submission for Ethical Appraisal (CAAE) No. 07534512.9.0000.5286.

Results

Eight of the 433,882 records of live births in Rio de Janeiro state in 2007 and 2008 were excluded because of duplication and 9,036 (2.1%) were classified as multiple pregnancy on the SINASC system; 8,136 of these latter records had deterministic key agreement (key+/Sinasc+). Following application of the routine and following automatic address comparison, 6,508 records that had the same address were automatically classified as twins, and a further 1,628 that had different addresses were classified as twins following manual review (Figure 1).

All 385 records having key+/Sinasc- were classified as twins: 260 with the same address were classified automatically, and 125 following manual review (Figure 1).

We identified 816 records for which the routine did not indicate twins but for which the information held on SINASC referred to multiple pregnancy (key-/Sinasc+). Seventy-eight of these were found after searching on the SIM system. With regard to the other 738, manual review identified 452 twins and 286 non-twins.

There were 424,537 records in key-/Sinasc- category; 9,051 were classified as twins and 424,823 as non-twins, with change of initial status in 671.

Accuracy of multiple pregnancy information held on SINASC, when compared to the classification derived by applying the complete routine, was as follows: sensitivity=95.8% (95%CI 95.3;96.2%), specificity=99.9% (95%CI 99.9;99.9%) and positive predictive value=95.9% (95%CI 95.5;96.3%) (Table 1).

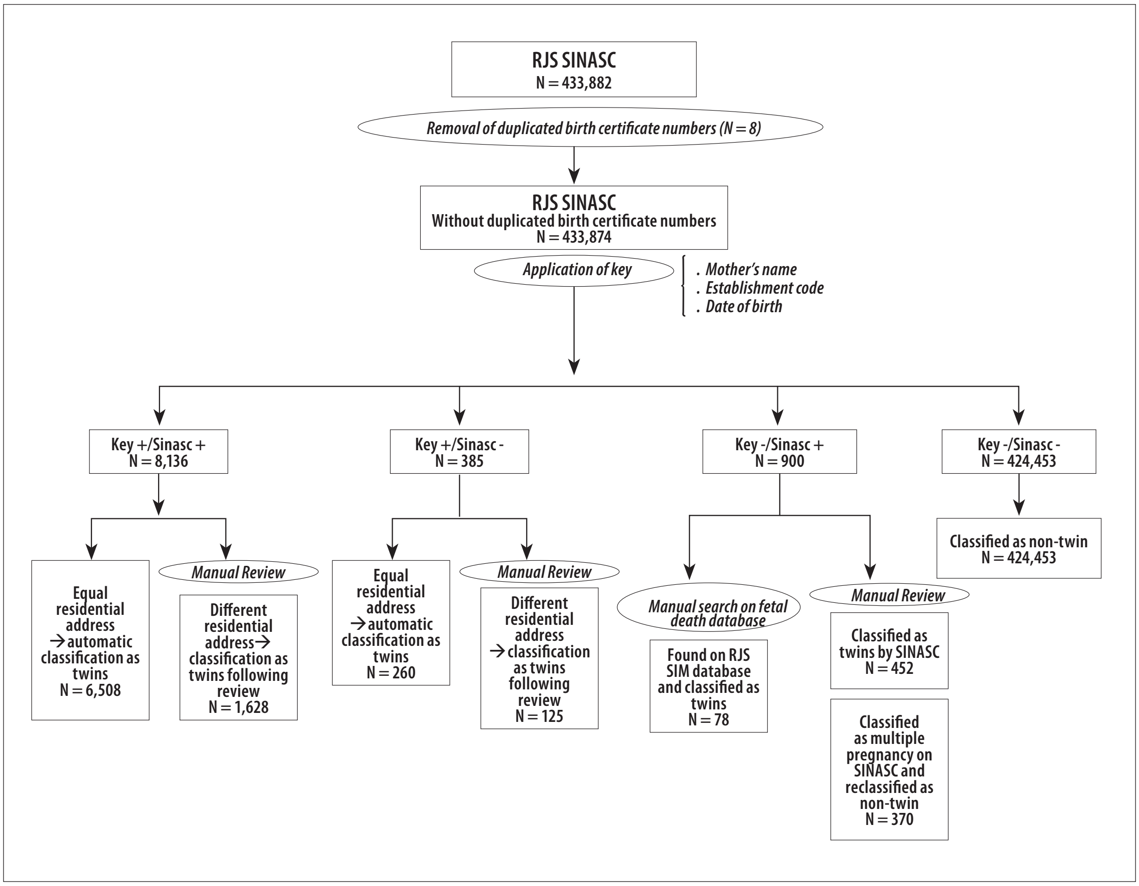
When applying the routine without manual review, accuracy was as follows: sensitivity=100.0%, specificity=99.9% (95%CI 99.9;99.9%) and positive predictive value=96.9% (95%CI 96.6;97.3%) (Table 2).

Discussion

The study used a deterministic routine which enabled better classification of multiple pregnancy information held on the SINASC system, avoiding both false-positive errors and also false-negative errors. Incorrect classification of a multiple pregnancy, such as duplications in data linkage processes, is a challenge for the development of algorithms for electronic health records.[15-17]

SINASC data coverage and quality are fundamental for its reliability as a source of substantial information for health evaluation and research.[18,19]

Although good quality information was found about pregnancy type on the Rio de Janeiro state SINASC system, a result in agreement with the literature,[1,20,21] applying the routine is nevertheless useful and easy to perform. However, given the peculiar characteristics of twins, in general studies of neonatal outcomes exclude records of this group, which should be analyzed separa-


RJS SINASC
N = 433,882
Removal of duplicated birth certificate numbers (N = 8)
RJS SINASC
Without duplicated birth certificate numbers
N = 433,874
Application of key
. Mother's name
. Establishment code
. Date of birth
Key +/Sinasc +
N = 8,136
Key +/Sinasc -
N = 385
Key -/Sinasc +
N = 900
Key -/Sinasc -
N = 424,453
Equal residential address → automatic classification as twins
N = 6,508
Manual Review
Different residential address → classification as twins following review
N = 1,628
Equal residential address → automatic classification as twins
N = 260
Manual Review
Different residential address → classification as twins following review
N = 125
Manual search on fetal death database
Found on RJS SIM database and classified as twins
N = 78
Manual Review
Classified as twins by SINASC
N = 452
Classified as multiple pregnancy on SINASC and reclassified as non-twin
N = 370
Classified as non-twin
N = 424,453

RJS: Rio de Janeiro state.

Figure 1 – Flowchart of application of deterministic routine to identify multiple pregnancies on the Rio de Janeiro state Live Birth Information System, 2007-2008

Table 1 – Accuracy of information about twins held on the Rio de Janeiro state Live Birth Information System database, 2007-2008

Sinasc[a]	Deterministic routine (gold standard)		Total
	Twin	Non-twin	
Twin	8,666	370	9,036
Non-twin	385	424,453	424,838
Total	**9,051**	**424,823**	**433,874**
Sensitivity = 95.8% (95%CI95.3;96.2%)			
Specificity = 99.9% (95%CI99.9 ;99.9%)			
Positive predictive value = 95.9% (95%CI95.5;96.3%)			

a) SINASC: Live Birth Information System.

Table 2 – Accuracy of information about twins held on the Rio de Janeiro state Live Birth Information System database, following application of an automatic deterministic routine without manual review, 2007-2008

Routine without manual review on SINASC[a]	Deterministic routine (gold standard)		Total
	Twin	Non-twin	
Twin	9,051	286	9,337
Non-twin	–	424,537	424,537
Total	**9,051**	**424,823**	**433,874**
Sensitivity = 100.0%			
Specificity = 99.9% (95%CI99.9;99.9%)			
Positive predictive value = 95.9% (95%CI95.7;96.4%)			

a) SINASC: Live Birth Information Systems.

tely.[22,23] Low frequency of twins in relation to total births results in changes in the number of cases being relatively important, even when small in absolute terms. Moreover, the Robson classification is now publicized on the sinasc system, since accurate information about twins is necessary for adequate categorization of women.[24]

Database linkage techniques, whether deterministic or probabilistic, are being used to improve information quality.[12,25] Deterministic routines have excellent performance when data quality is good:[26,27] their processing is rapid and they can be used without manual review of the links formed.

The routine developed in our study included a manual review stage which is only feasible for small or medium volume databases in situations of information disagreement. In situations involving databases with a larger volume of records, only applying the key without doing manual review increases sensitivity for identifying twins without significant alteration of specificity or positive predictive value. A midway alternative would be to manually process only records not identified by the key, although for these records multiple pregnancy information exists on the SINASC system.

A limitation of this study is that not all cases were manually reviewed. However, the likelihood of mistaken pregnancy type classification is very low when there is agreement between the key and the Live Birth Information System – SINASC.

Although the increase as a result of recovering twins appears small, the cost of doing this is low in view of the possibility of improving information. We suggest that the routine proposed be used habitually, especially in studies of neonatal outcomes among twins.

Authors' contributions

Aguiar FP and Coeli CM were responsible for the conception and structuring of the study and data analysis. Aguiar FP, Coeli CM, Flores PVG, Guillen LCT, Santos HPS, Alves LGSB, Camargo Jr KR and Pinheiro RS contributed to data analysis and interpretation, drafting the preliminary versions of the manuscript and critically reviewing it. All authors have approved the final version of the manuscript and declare that they are responsible for all aspects of the work, guaranteeing its accuracy and integrity.

References

1. Theme Filha MM, Gama SGN, Cunha CB, Leal MC. Confiabilidade do Sistema de Informações sobre Nascidos Vivos Hospitalares no Município do Rio de Janeiro, 1999-2001. Cad Saúde Pública. 2004;20(Supl. 1):83-91.
2. Costa JMBS, Frias PG. Avaliação da completitude das variáveis da Declaração de Nascido Vivo de residentes em Pernambuco, Brasil, 1996 a 2005. Cad Saúde Pública. 2009;25(3):613-24.
3. Morais Neto OL, Barros MBA. Risk factors for neonatal and post neonatal mortality in the Central-West region of Brazil: linked use of life-birth and infant death records. Cad Saúde Pública.2000;16(2):477-85.
4. Silva CF, Leite AJM, Almeida NMGS, Gondim RC. Fatores de risco para a mortalidade infantil em município do Nordeste do Brasil:linkageentre bancos de dados de nascidos vivos e óbitos infantis – 2000 a 2002. Rev Bras Epidemiol. 2006;9:69-80.
5. Ramos HÂDC, Cuman RKN. Risk factors for prematurity: document search. Escola Anna Nery. 2009;13(2):297-304.
6. Silva VFG. Complicações na gestação de gemelar. Fertilização in vitro versus espontânea. Instituto de Ciências Biomédicas Abel Salazar. Porto: Universidade do Porto; 2013.
7. Shinwell ES, Haklai T, Eventov-Friedman S. Outcomes of Multiplets. Neonatology. 2009;95:6-14.
8. Cooke RWI. Does neonatal and infant neurodevelopmental morbidity of multiples and singletons differ? Seminars in Fetal & Neonatal Medicine. 2010;15:362-6.
9. Barbuscia DM, Rodrigues-Júnior AL. Completeness of data on live birth certificates and death certificates for early neonatal and fetal deaths in the Ribeirão Preto Region, São Paulo State, Brazil, 2000-2007. Cad Saúde Pública. 2011; 27:1192-200.
10. Oliveira MM, Andrade SSCA, Dimech GS, Oliveira JCG, Malta DC, Rabelo Neto, DL, et al. Avaliação do Sistema de Informações sobre nascidos vivos. Brasil, 2006 a 2010. Epidemiol Serv Saúde. 2015;244(4):629-40.
11. Méray N, Reitsma JB, Ravelli AC, Bonsel GJ. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. Journal of Clinical Epidemiology. 2007;60:883-91.
12. Silva LP, Moreira CMM, Amorim MHC, Castro DS, Zandonade E. Avaliação da qualidade dos dados do Sistema de Informações sobre Nascidos Vivos e do Sistema de Informações sobre Mortalidade no período neonatal, Espírito Santo, Brasil, de 2007 a 2009. Ciênc Saúde Coletiva. 2014;19(7):2011-20.
13. Bartholomay P, Oliveira G.P, Pinheiro RS, Vasconcelos AMN. Melhoria da qualidade das informações sobre tuberculose a partir do relacionamento entre bases de dados. Cad Saúde Pública. 2014;30(11).
14. Souza Maia LT, Souza WV, Cruz AG.M. A contribuição do linkage entre SIM e SINASC para a melhoria das informações da mortalidade infantil em cinco cidades brasileiras. Revista Brasileira de Saúde Materno-Infantil. 2015;15(1):57-66.
15. Baldwin E, Johnson K, Berthoud H, Dublin S. Linking mothers and infants within electronic health records: a comparison of deterministic and probabilistic algorithms. Pharmaco epidemiology and Drug Safety. 2015;24:45-51.
16. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, Meulen JH. A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology. 2017, 46(5).
17. Harper G. Linkage of Maternity Hospital Episode Statistics data to birth registration and notification records for birth sin England 2005-2014: Quality assurance of linkage of routine data for singleton and multiple births. BMJ Open. 2018.
18. Silva RS, Oliveira CM, Ferreira DKS, Bonfim CV. Avaliação da completitude das variáveis do Sistema de Informações sobre Nascidos Vivos- SINASC- nos Estados da região Nordeste do Brasil, 2000 e 2009. Epidemiol Serv Saúde. 2013;22(2):347-52.
19. Mello Jorge MHP, Laurenti R, Gotlieb SLD. Análise da qualidade das estatísticas vitais brasileiras: a experiência de implantação do SIM e do SINASC. Ciênc Saúde Coletiva. 2007;12(3):643-54.
20. Gabriel GP, Chiquetto L, Morcillo AM, Ferreira MC, Bazan IG, Daolio LD, et al. Evaluation of data on live birth certificates from the Information System on Live Births (Sinasc) in Campinas, São Paulo, 2009. Revista Paulista de Pediatria. 2014;32(3):183-8.
21. Bonilha EA, Vico ESR, Freitas M, Barbuscia DM, Galleguillos TGB, Okamura MN, et al. Cobertura, completude e confiabilidade das informações do Sistema de Informações sobre Nascidos Vivos de maternidades da rede pública no município de São Paulo, 2011. Epidemiol Serv Saúde. 2018;27(1).

22. Gardner MO, Goldenberg RL, Cliver SP, Tucker JM, Nelson KG, Copper RL. The origin and outcome of preterm twin pregnancies. Obstetrics and Gynecology. 1995;85(4):553-7.

23. Garite TJ, Clark RH, Elliott JP, Thorp JA. Twins and triplets: the effect of plurality and growth on neonatal outcome compared with singleton infants. American Journal of Obstetrics and Gynecology. 2001;191(3):700-7.

24. Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância de Doenças e Agravos não Transmissíveis e Promoção da Saúde. Saúde Brasil 2017: uma análise da situação de saúde e os desafios para o alcance dos objetivos de desenvolvimento sustentável. Brasília, 2018.

25. Pedraza DF. Quality of the Information System on Live Births/Sinasc: a critical analysis of published studies. Ciênc Saúde Coletiva. 2012;17(10):2729-37.

26. Coeli CM, Pinheiro RS, Camargo Jr KR. Conquistas e desafios para o emprego das técnicas de record linkage na pesquisa e avaliação em saúde no Brasil. Epidemiol Serv Saúde. 2015;24:795-802.

27. Oliveira GP, Bierrenbach AL, Camargo Jr. KR, Coeli CM, Pinheiro RS. Acurácia das técnicas de relacionamento probabilístico e determinístico: o caso da tuberculose. Rev Saúde Pública. 2016;50(49)

Received on 20/12/2018
Approved on 24/09/2019

Associate Editor: Doroteia Aparecida Höfelmann - orcid.org/0000-0003-1046-3319