Received: 31 July 2022 | Revised: 15 February 2023 | Accepted: 26 February 2023

DOI: 10.1111/ppe.12968

SPECIAL ISSUE ARTICLE

Paediatric and Perinatal Epidemiology WILEY

National data linkage assessment of live births and deaths in Mexico: Estimating under-five mortality rate ratios for vulnerable newborns and trends from 2008 to 2019

Lorena Suárez-Idueta[1] | Robespierre Pita[2,3] | Hannah Blencowe[4] | Arturo Barranco[5] | Jesus F. Gonzalez[1] | Enny S. Paixao[2,4] | Mauricio L. Barreto[2] | Joy E. Lawn[4] | Eric O. Ohuma[4]

[1]Mexican Society of Public Health, Mexico City, Mexico

[2]Centre of Data and Knowledge Integration for Health (CIDACS), Salvador, Brazil

[3]Computing Institute, Federal University of Bahia, Salvador, Brazil

[4]Maternal, Adolescent, Reproductive & Child Health (MARCH) Centre, Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK

[5]Ministry of Health, Population and Health Information, Ministry of Health, Mexico City, Mexico

Correspondence
Lorena Suárez-Idueta, Mexican Society of Public Health, Mexico City, Mexico.
Email: lorena.idueta.19@alumni.ucl.ac.uk

Funding information
The Children's Investment Fund Foundation (United Kingdom), Grant/Award Number: 1803-02535; ESP ESP is funded by the Wellcome Trust, Grant/Award Number: 213589/Z/18/Z

A commentary based on this article appears on pages 287-291

Abstract

Background: Linked datasets that enable longitudinal assessments are scarce in low and middle-income countries.

Objectives: We aimed to assess the linkage of administrative databases of live births and under-five child deaths to explore mortality and trends for preterm, small (SGA) and large for gestational age (LGA) in Mexico.

Methods: We linked individual-level datasets collected by National statistics from 2008 to 2019. Linkage was performed based on agreement on birthday, sex, residential address. We used the Centre for Data and Knowledge Integration for Health software to identify the best candidate pairs based on similarity. Accuracy was assessed by calculating the area under the receiver operating characteristic curve. We evaluated completeness by comparing the number of linked records with reported deaths. We described the percentage of linked records by baseline characteristics to identify potential bias. Using the linked dataset, we calculated mortality rate ratios (RR) in neonatal, infants, and children under-five according to gestational age, birthweight, and size.

Results: For the period 2008–2019, a total of 24,955,172 live births and 321,165 under-five deaths were available for linkage. We excluded 1,539,046 records (6.2%) with missing or implausible values. We succesfully linked 231,765 deaths (72.2%: range 57.1% in 2009 and 84.3% in 2011). The rate of neonatal mortality was higher for preterm compared with term (RR 3.83, 95% confidence interval, [CI] 3.78, 3.88) and for SGA compared with appropriate for gestational age (AGA) (RR 1.22 95% CI, 1.19, 1.24). Births at <28 weeks had the highest mortality (RR 35.92, 95% CI, 34.97, 36.88). LGA had no additional risk vs AGA among children under five (RR 0.92, 95% CI, 0.90, 0.93).

Lorena Suárez-Idueta and Robespierre Pita Joint first author

Conclusions: We demonstrated the utility of linked data to understand neonatal vulnerability and child mortality. We created a linked dataset that would be a valuable resource for future population-based research.

KEYWORDS
infant, medical record linkage, mortality, perinatal

1 | BACKGROUND

Innovations in technology have offered the possibility to analyse large volumes of health records collected for administrative purposes to find meaningful information, conduct research, and support decision-making.[1] The availability of public individual-level datasets enables the linkage of routine records to describe disease patterns and track public health outcomes across time and space.[2] In maternal and child health, innovative linkage methods are useful to collate common variables across different data sources to form cohort datasets with information on pregnant women and their babies.[3,4] Linked datasets are widely used in high-income countries to describe perinatal outcomes, maternal risk factors, evaluate health services, and to assess critical goals such as those of the Sustainable Development Target 3.2 to end preventable deaths of newborns and children under five by 2030.[2,3,5,6] More recently, innovative linkage methods have been applied in low-and middle-income countries to evaluate social, financial, and public health interventions.[7–9]

Record linkage success depends on the availability of variables to accurately identify and link individuals (e.g., name of the child, mother's name, date of birth, address of residency, etc). However, the pseudo-anonymised and publicly available administrative data often lack some of these variables. In these datasets where unique identifiers are not available, score-based linkage methods are typically preferred to rank the most similar pairs of records based on existing variables. This ranking is used to define which cut-off point better separates true from false matches through a clerical review which poses a challenge for large-sized datasets. Beyond the linkage errors assessment, a proper validation scheme is needed to identify potential biases.[10]

We aimed to assess the linkage of two national, individual level and publicly available Mexican datasets of live births and deaths collected from 2008 to 2019 where unique identifiers and names of patients are unavailable. Using the linked dataset, we further aimed to examine the association of preterm, small for gestational age (SGA) and large for gestational age (LGA) with mortality at neonatal (0-27 days), infancy (0-365 days) and five years of age (0-1825 days).

Synopsis

Study question

To assess the linkage of two national datasets on live births and under-five deaths in Mexico to examine mortality rate ratios and trends in childhood mortality from 2008 to 2019.

What is already known

Record linkage techniques enable the creation of cohorts from individual-level datasets for tracking the progress of key health targets, identifying priority areas, and enabling planning. Such linked datasets are available and widely used in high-income countries but less so in low and middle-income settings.

What this study adds

This study shows the feasibility of nationwide record linkage in Mexico. As an illustration of the utility of this 25 million birth-linked dataset, we have calculated mortality rate ratios and trends for preterm, small, and large for gestational age.

2 | METHODS

2.1 | Data sources

Individual-level data on live births and deaths were obtained from open-access hubs administered by the Mexican Government.[11,12] We considered all live births and deaths registered from 2008 when the registry of birth certificates was implemented in Mexico to December 2019, the latest data release available when this analysis was performed. The two datasets were described according to the Dublin Core Metadata Initiative[13] to encourage best practices of interoperability (Table S1).

Live birth records were collected through the National Information Subsystem of Live births (Subsistema de Información Sobre Nacimientos, SINAC).[14] Live birth notification is mandatory for births occurring in clinics, hospitals or the community and must be recorded 24–48 hours after the event. Birth certificates include information regarding the mother's characteristics, information on the neonate, institution, and person who attended the delivery.[15] Deaths were identified by the National Institute of Statistics and Geography, Statistics of Deaths (Instituto Nacional de Estadística Geografía e Informática, Registros administrativos de defunción, INEGI).[11] Death certificates include information on demographic details regarding the person who died, the institution, and the cause of death according to the International Classification of Diseases.[16] Data from each state are collated at the national level. Every year,

pseudonymised and validated datasets of live births and deaths are published online separately.[11,12] Given that most live births and deaths occurred in healthcare facilities, national coverage has been estimated to reach more than 90.0% of the target population, but this percentage varies at the subnational level.[17,18]

2.2 | Record linkage

2.2.1 | Score-based record linkage software

Record linkage of national datasets is complex. In our linkage analysis, we applied indexing techniques such as rule-based blocking and Eelasticsearch heuristics to prevent unnecessary comparisons, and retain accurate linkages using the Centre for Data and Knowledge Integration for Health – Record Linkage (CIDACS-RL) software.[19] This software has been developed to integrate high-dimensional data sources and has been demonstrated to be effective in managing large administrative data in middle-income settings.[20] This approach can address the issue of scalability in big data linkages through massive processing and indexing.

2.2.2 | The linkage process

Data engineering tasks, often illustrated in a pipeline schematic, play a crucial role in enabling the use of administrative datasets for research through record linkage. These tasks harmonise the syntactic and semantics of variables from different sources and purposes resulting in higher-quality data. Figure 1 describes our data linkage pipeline with five steps that perform key transformations and filtering of live birth records from SINAC and death records from INEGI.

Attribute selection

First, we reviewed raw versions of SINAC and INEGI datasets including 24,955,172 and 7,636,151 initial unique records of live births and deaths, respectively. We identified common attributes such as birthday, sex, place of residence and place of occurrence based on their discriminative power, by calculating the variability and number of different possible values and missing values (Figure S1).

Pre-processing

According to National guidelines, we transformed addresses into numeric codes considering names and codes of a universe of 32 states, 2472 municipalities, and 300,338 localities.[11] For instance, the encoded residential address value "08–036-732" refers to an event that occurred in the locality of Guadalupe on Jimenez, Chihuahua represented by the individual codes 732, 36, and 8. Subsequently, we created five new variables to help with the linkage process. The variables *state_municipality_residence* and *state_municipality_occurrence* resulted from the concatenation of state and municipality codes, whereas the variables *day of birth*, *month of birth*, and *year of birth* resulted from breaking the birthday variable. Additionally, we reorganised the mortality data according to the year of birth instead of the year of notification. We excluded records for people older than 5 years of age as our analyses were focused on mortality under-5 years.

Deterministic linkage

We executed an exact search of *birthday*, *sex*, and the code generated from the *residential address* (e.g., 08-036-732) to match each

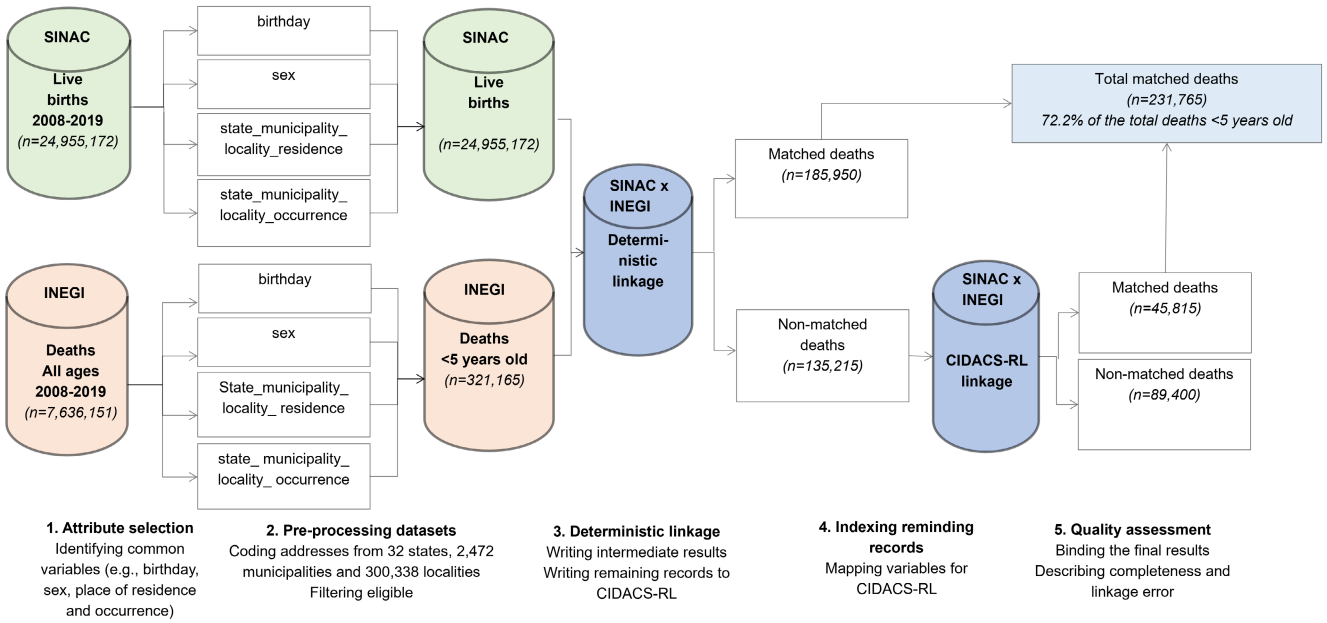

FIGURE 1 Mexico national data linkage pipeline for administrative datasets of live births and deaths (2008–2019).

record from INEGI to the indexed SINAC. We assumed that these three attributes might be capable of identifying records of the same person. We considered matched records those with exact agreement upon this set of attributes whereas non-matched records were those with any level of disagreement. Non-matched records were considered for further linkage.

Score-based linkage

Score-based linkage was applied to records that were not matched through deterministic linkage. We used the CIDACS-RL software to perform a semi-exact search using a broader set of attributes. We hypothesized that adding the variable address of occurrence, which we decoded into the state, municipality, and locality separately, and boosting some attributes, might enable a less restrictive searching process to return similar candidates, highlight differences, and identify which agreements are most probable to be records of the same individual.

We computed a similarity score for each possible pair of death and live birth candidates produced from the semi-exact search. This similarity score estimated the likelihood that a given pair of records represented the same individual, and it consisted of a weighted average from 0 to 1. Weights, parameters from CIDACS-RL, were used to determine the importance of each attribute, and penalties were used to prevent pairs with missing information from reaching elevated scores (Table 1). All produced pairs were sorted by their score to perform a quality assessment.

Quality assessment

We selected a sample of 3000 pairs to perform a clerical review to identify possible misclassifications. Each pair was stratified based on the similarity score as high: above 0.97, intermediate: between 0.92 to 0.97, and low: below 0.92.[21] Figure 2 shows the decision tree created to identify false and true matches. False matches were those pairs produced by CIDACS-RL that disagree on crucial attributes, such as date of birth, sex, state or municipality of residence.[22] This manual evaluation was used to build a receiver operator characteristic curve (ROC), calculate the area under de ROC curve (AUROC), and provide the threshold to better separate true and false matches.

As part of the quality assessment, we assessed linkage completeness and performance. To assess completeness we divided the number of linked deaths by the number of reported deaths per 100. To evaluate performance, we compared the percentage of linked records that were joined by deterministic versus CIDACS-RL methods and the percentage of non-linked records according to maternal clinical features such as age, parity, educational attainment, mode and place of delivery, number of babies, antenatal care, and sociodemographic characteristics using data on the social deprivation index. This multidimensional poverty index is calculated by the National Council of Evaluation of Public Policies considering the number of people without formal education, the number of people without access to health care, households with essential services, dirt floors, piped water, drainage, electricity and access to basic food supplies across the 32 states and reported as the following five categories: very low, low, medium, high and very high deprivation.[23]

2.3 | Exclusions

Once the data were linked, we excluded records with missing values on birthweight, gestational age and/or sex because these three variables are key to defining the size of the newborn. We also excluded improbable birthweight values defined as birthweight above or below five standard deviations from the mean birthweight at a given gestational age, or birthweight <250 g. We excluded values of gestational age < 22 + 0 weeks or >44 + 6 weeks.

2.4 | Exposure

Preterm birth was defined as birth at <37 completed weeks of gestation and size for gestational age was defined using birthweight, gestational age, and sex according to the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st) standards as follows: small for gestational age (SGA, <10th centile), appropriate for gestational age (AGA, 10th to 90th centiles), and large for gestational age (LGA, >90th centile).

A secondary analysis was conducted by fine strata of gestational age and birthweight. The following groups were used for the gestational age analysis: one baseline category (39–40 weeks) and seven comparison groups in completed weeks (41–44, 37–38, 34–36, 32–33, 30–31, 28–29 weeks, and below 28 weeks). The birthweight analysis included one baseline category (3000–3499 g) and fine strata of 500 g each for comparison (≥5000, 4500–4999, 4000–4499, 3500–3999, 2500–2999, 2000–2499, 1500–1999, 1000–1499, and <1000 g).

2.5 | Outcomes

A neonatal death was defined as a death that occurred during the first 28 days after birth (0–27 days). We also assessed deaths that occurred among infants (0–365 days after birth), and children under 5 years of age (≤1825 days after birth).

2.6 | Statistical analysis

To choose the best cut-off point balancing sensitivity and specificity, we built a ROC curve and calculated the area under the curve (AUROC).[20]

We explored mortality using a person-time approach. Given that this is a complete case analysis, we considered the time that each individual was at risk durign the study. We performed the person-time analysis with Stata version 16.1. We used the date of birth and date of death or the last day of the study (31st December 2019) to calculate the time at risk. The vital status of every child was considered as alive, dead, or censored at the end of each period of follow-up; the neonatal period (0–27 days), infancy (0–365 days), or 5 years of age (≤1825 days). Rates were calculated and reported as the number of deaths per 1000 person-years. Rates, rate ratios (RR), and 95%

TABLE 1 Set of common variables used to perform deterministic and score-based linkage.

	Query			Pairwise comparison		
Attributes	Exact match	Semi-exact match	Boost	Weight	Penalty	Similarity measure
Birthday (day-month-year)	yes	yes	4	4	0.1	Hamming
Residential address (state-municipality-locality)	yes	yes	4	3	0.05	Jaro Winkler
Sex	yes	yes	2	3	0.025	Overlap
Address of occurrence (state-municipality-locality)	no	yes	1	1	0.01	Jaro Winkler
State of residence	no	yes	1	1	0.01	Overlap
Municipality of residence	no	yes	1	0.5	0.05	Overlap
Locality of residence	no	yes	1	0.5	0.05	Overlap
State of occurrence	no	yes	0.5	0.5	0.025	Overlap
Municipality of occurrence	no	yes	0.5	0.05	0.05	Overlap
Locality of occurrence	no	yes	1	0.5	0.05	Overlap
State and municipality of residence	no	yes	3	2	0.01	Jaro Winkler
State and municipality of occurrence	no	yes	0.5	0.5	0.01	Jaro Winkler
Day of birth	no	yes	1	1	0.05	Overlap
Month of birth	no	yes	1	1.5	0.05	Overlap
Year of birth	no	yes	1	1.5	0.05	Overlap

confidence intervals (CI) were calculated according to birthweight, gestational age, size for gestational age, state of residence, and social deprivation areas. For comparability with the standard approach, we calculated the overall absolute risk (number of deaths per 1000 live births) among neonates, infants and children under five.

3 | RESULTS

3.1 | Linkage performance

A total of 24,955,172 live births and 7,636,151 deaths of all ages were recorded in Mexico from 2008 to 2019. Of the total deaths, 321,165 deaths were of children under five, 185,950 (57.9%) records were linked by deterministic linkage and the CIDACS-RL technique provided 45,815 (14.3%) additional matches with high accuracy as indicated by the AUROC. Table 2 shows how the model was consistently capable of distinguishing between true and false matches from 2008 to 2019 (AUROC, range 0.98–0.99). Overall, the proportion of linked records was 72.2% over the period, with the lowest proportion of 57.1% in 2009 and the highest of 84.3% in 2011. The distribution of linked and non-linked records for babies from SINAC was similar across baseline characteristics, but linked records were less likely to occur among records with missing values (Table S2). At the subnational level, we found a gradient in the proportion of linked records according to poverty using data on the social deprivation index.[23] The highest completeness was found among states with very low social deprivation (median 86.1, IQR 81.0, 90.0), followed by areas with low (median 84.3, IQR 79.1, 87.6), medium (median 74.4, IQR 71.6, 80.9), high (median 73.0, IQR 66.6, 81.7) and very high deprivation (median 72.0, IQR 58.4, 78.9) (Figure 3; Table S3). We were able to assess high-accuracy linkage performance across the different areas (Figure S2).

3.2 | Mortality rate ratios and trends

From the 24,955,172 live births overall, 1,539,046 (6.2%) were excluded due to missing or implausible values. Of the 23,416,126 included live births, 6.4% were preterm, 7.1% were SGA and 9.1% were LGA (Figure S3).

Preterm babies were almost at a 4-fold increased mortality rate compared to term (RR 3.83, 95% CI, 3.78, 3.88) and SGA babies had 1.22 times the rate compared to AGA (RR 1.22, 95% CI, 1.19, 1.24) (Table S4). There was a progressively declining mortality risk with increasing gestational age (Table S5).

Spatial analysis showed substantial disparities in preterm survival. Compared with babies born at term, the neonatal mortality rate was higher among preterm babies living in very highly deprived and highly deprived states and lower in states with low and very low social deprivation (Tables 3 and S6). A similar pattern was observed for SGA neonates compared with AGA neonates. Neonatal mortality rate ratios had a gradient from 1.31 among very high deprivation 1.08 in very low deprivation areas (Table 3).

LGA births were not associated with an increased mortality risk compared to AGA. However, the analysis stratified by gestational age showed that babies born from 41 to 44 weeks had a modestly higher risk of dying during the neonatal period (RR 1.08, 95% CI, 1.06, 1.11), infancy (RR 1.05, 95% CI, 1.03, 1.07) and under-five

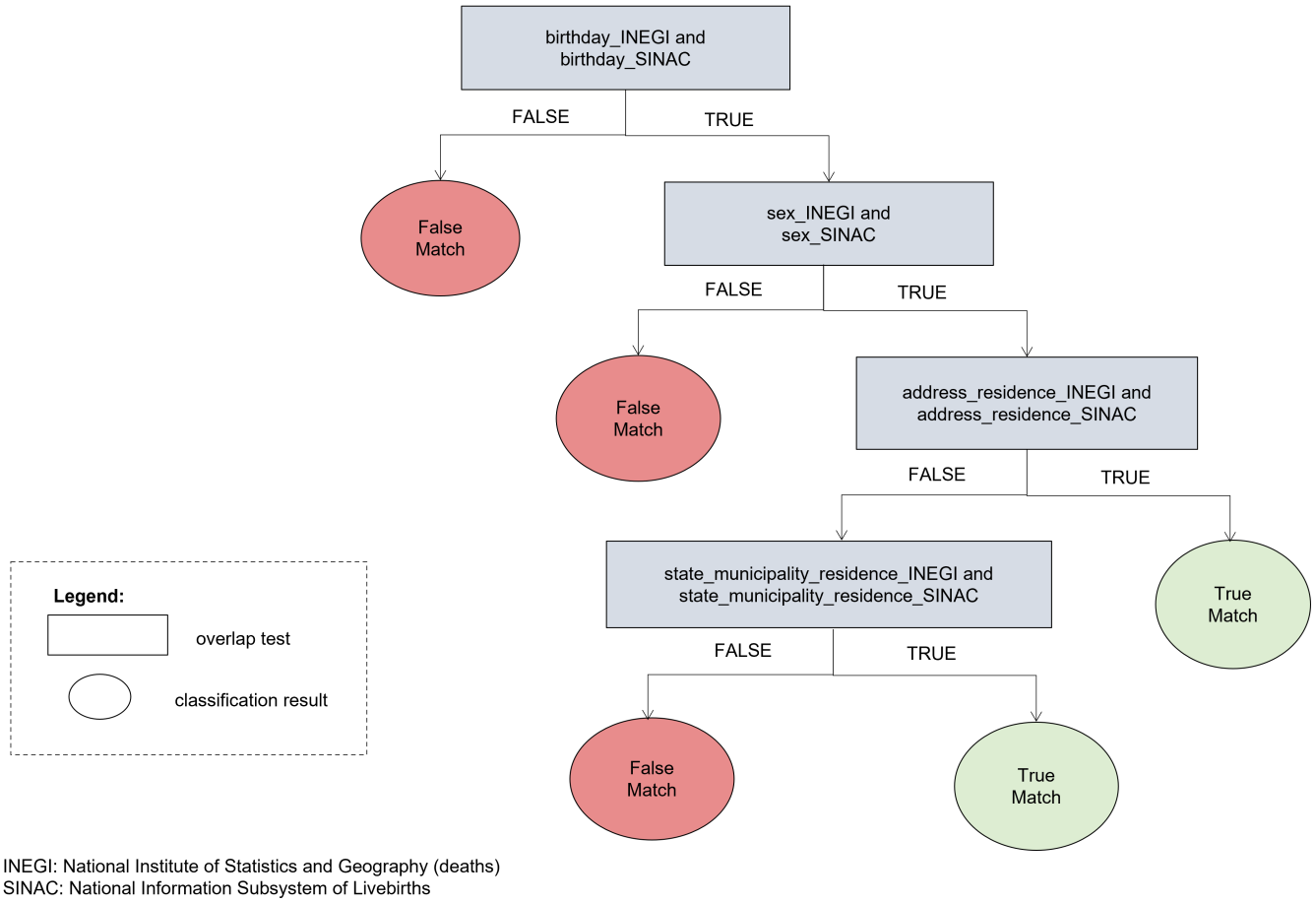

FIGURE 2 Decision tree learned from clerical review labelled data.

(1.04, 95% CI, 1.02, 1.06) compared to those born at 39–40 weeks (Tables S4 and S5).

3.3 | Comment

3.3.1 | Principal findings

This analysis found that is feasible to link large administrative datasets of live births and deaths in Mexico. We found that preterm birth was the strongest driver of mortality with the highest mortality rate among those born at less than 28 weeks gestation or weighing less than 1000 g at birth.

3.3.2 | Strengths of the study

Mexican registries collect information country-wide, reaching an estimated coverage above 90% of their total population.[17] This study brought together these vital statistics to describe how birth outcomes impact neonatal and child deaths. The indexing frameworks employed on CIDACS—RL software made this Big Data analysis scalable, feasible, and accurate[20] as it has been reported using national-wide Brazilian data.[26,39] These findings are, therefore, potentially useful to explore mortality rate ratio patterns and identify the most vulnerable children, which are the key target of the strategic plan towards 2030 and Sustainable Development Goals in Mexico.[24]

3.3.3 | Limitations of the data

We experienced limitations on linking public medical records due to the lack of unique identifiers, names, missing values, inaccurate registration, and demographic dynamics. We hypothesized that deterministic matching using sex (girl vs boy), birthday (day-month-year), and mother's residential address (state-municipality-locality) would result in few false-matches (records linked from different individuals),[3] but this approach resulted in 41% of non-matched records. Then, to increase the number of linked records, we incorporated approximate matching by boosting attributes such as the address of occurrence, splitting birthdays into the day, month and year of birth and adding similarity scores.[25] Even though we used clerical review to minimise missed matches,[26] we recognise that our process may offer a residual number of false-positive matches, that would potentially introduce some bias to our results.[21,27]

Despite the percentage of linked records, the failure to link 27.8% of deaths remains a limitation for calculating mortality rates and comparing these indicators with those repoted by the Ministry of Health

TABLE 2 Number, percentage and area under the receiver operating curve of linked records by year of birth. Mexico from 2008 to 2019.

				Linked deaths among children under five						
Year	SINAC Total live births	INEGI Deaths <5 years	Non-linked records	By deterministic join	%	By CIDACS-RL	%	Total Linked records	%	AUROC
2008	1,978,380	24,196	1,959,617	14,673	60.6	4090	16.9	18,763	77.5	0.99
2009	2,058,708	34,183	2,039,183	15,245	44.6	4280	12.5	19,525	57.1	0.98
2010	2,073,111	23,699	2,056,166	13,347	56.3	3598	15.2	16,945	71.5	0.98
2011	2,167,060	24,142	2,146,717	16,613	68.8	3730	15.5	20,343	84.3	0.99
2012	2,206,692	23,241	2,188,857	14,043	60.4	3792	16.3	17,835	76.7	0.98
2013	2,195,073	21,767	2,178,063	13,394	61.5	3616	16.6	17,010	78.1	0.98
2014	2,177,319	23,625	2,158,031	15,325	64.9	3963	16.8	19,288	81.6	0.98
2015	2,145,199	30,857	2,123,750	16,898	54.8	4551	14.7	21,449	69.5	0.98
2016	2,080,253	29,795	2,060,171	15,874	53.3	4208	14.1	20,082	67.4	0.98
2017	2,064,507	30,210	2,041,234	19,375	64.1	3898	12.9	23,273	77.0	0.99
2018	1,940,656	28,098	1,920,050	17,227	61.3	3379	12.0	20,606	73.3	0.99
2019	1,868,214	27,352	1,851,568	13,936	51.0	2710	9.9	16,646	60.9	0.99
TOTAL	24,955,172	321,165	24,723,407	185,950	57.9	45,815	14.3	231,765	72.2	0.98

Abbreviations: AUROC, area under the receiver operating curve; CIDACS-RL, Centre for Data and Knowledge Integration for Health; INEGI, National Institute of Statistics and Geography; SINAC, National Information Subsystem of Live births.

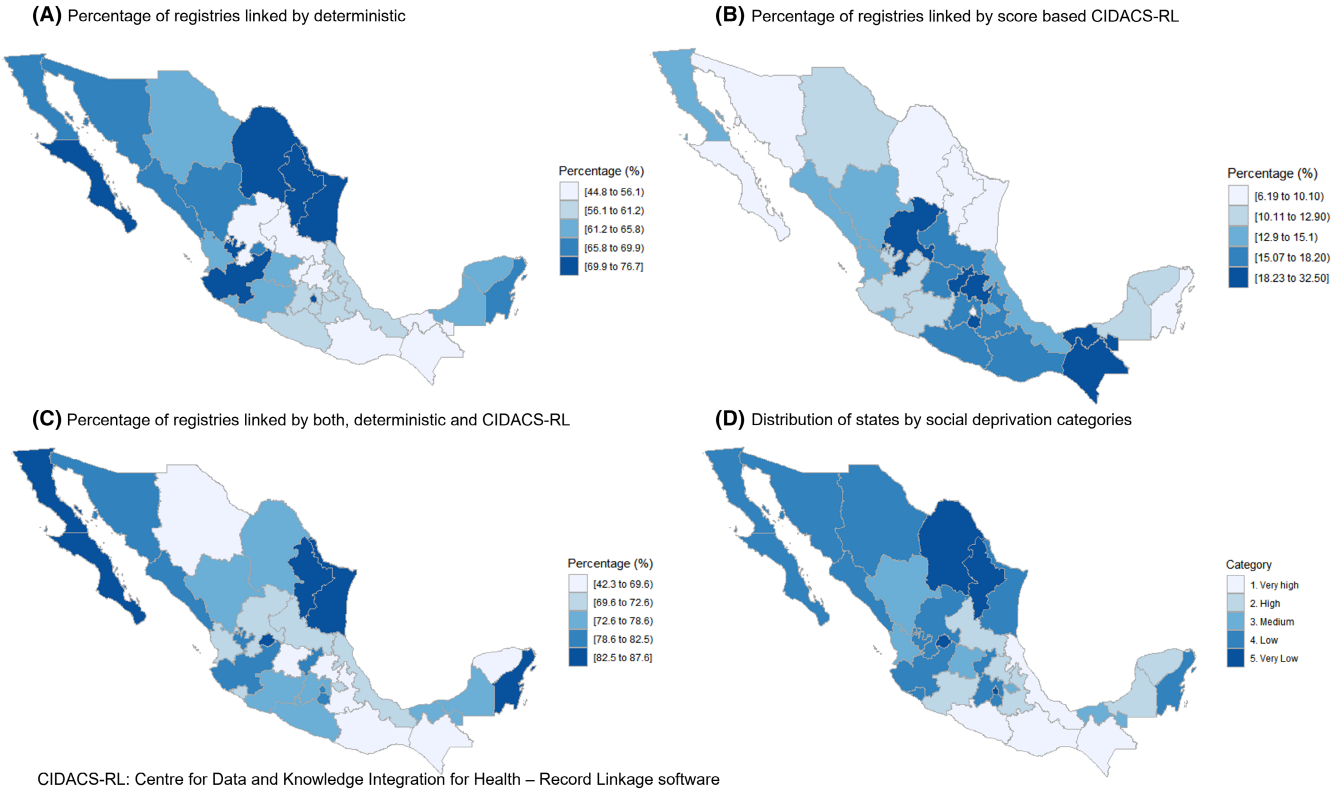

FIGURE 3 Geographic distribution of linked records by deterministic or score-based techniques and social deprivation areas.

TABLE 3 Number of deaths, mortality rate (deaths per 1000 person-years) and rate ratios of preterm vs term and SGA vs AGA according to social deprivation areas, Mexico from 2008 to 2019.

Social deprivation	Number of deaths	Deaths per 1000 person-years (95% CI)	Rate ratio (95% CI) Preterm vs term	Rate ratio (95% CI) SGA vs AGA
Neonatal mortality				
Very high	18,907	72.5 (71.46, 73.53)	6.80 (6.58, 7.03)	1.31 (1.25, 1.38)
High	18,651	71.4 (70.39, 72.44)	5.62 (5.44, 5.81)	1.26 (1.20, 1.33)
Medium	14,121	73.1 (71.94, 74.35)	4.81 (4.63, 5.00)	1.30 (1.22, 1.38)
Low	53,604	79.2 (78.49, 79.83)	3.07 (3.01, 3.14)	1.20 (1.16, 1.23)
Very low	18,714	84.0 (82.82, 85.23)	2.02 (1.94, 2.11)	1.08 (1.02, 1.14)
Infant mortality				
Very high	26,970	7.7 (7.57, 7.75)	5.50 (5.34,5.66)	1.31 (1.26, 1.36)
High	27,758	7.9 (7.79, 7.97)	4.52 (4.39,4.65)	1.34 (1.29, 1.39)
Medium	19,937	7.7 (7.55, 7.76)	4.08 (3.95,4.22)	1.36 (1.30, 1.43)
Low	78,339	8.6 (8.52, 8.64)	2.69 (2.64,2.74)	1.22 (1.19, 1.25)
Very low	26,919	8.9 (8.86, 9.07)	1.82 (1.76,1.89)	1.11 (1.06, 1.17)
Children under five mortality				
Very high	31,330	1.9 (1.76, 1.80)	4.83 (4.70, 4.96)	1.29 (1.24, 1.33)
High	31,911	1.8 (1.79, 1.83)	4.06 (3.95, 4.17)	1.33 (1.28, 1.38)
Medium	22,476	1.7 (1.71, 1.75)	3.75 (3.63, 3.87)	1.34 (1.28, 1.41)
Low	88,776	1.9 (1.93, 1.96)	2.50 (2.46, 2.55)	1.20 (1.17, 1.23)
Very low	30,070	2.0 (1.98, 2.03)	1.72 (1.67, 1.78)	1.12 (1.07, 1.17)

Note: Mortality is presented by neonatal, infant and under-five periods. Social deprivation included very high ($n = 4$ states), high ($n = 6$ states), medium ($n = 5$ states), low ($n = 13$ states) and very low ($n = 4$ states) deprivation areas.

Rate (deaths per 1000 person-years) calculated with the linked dataset using the person-time approach, for comparability with reported rates (deaths per 1000 livebirths) see supplementary material (Table S7).

(Table S7). Some part of this percentage of non-linkage may be explained by changes in the address of residence between the child's birth and death (Table S8 S7). Two years had a particularly low percentage of linked records: 2009 (57.1%) when the public health data systems had significant challenges due to the emergence of the novel influenza A(H1N1) pandemic in Mexico,[28] and 2019 (60.9%) the last year available by the time this study was performed which might be partially explained by the delay in death notification, for example, deaths that occurred in 2019 might be notified from 2020 onwards. At the sub-national level, areas with very high deprivation had lower levels of linkage completeness.

3.4 | Interpretation

In this population-based analysis, we described the linkage between births and deaths, including information about 24,955,172 live births and 231,765 deaths of children under five that were reported over more than a decade in Mexico. We have successfully performed a record linkage of two very large national pseudo-anonymised datasets, linking more than 70% of records with high accuracy (0.98 AUROC).

This study highlights the elevated risk of early and potentially preventable deaths among those who have been born with adverse conditions such as prematurity, low birthweight, and SGA.[29–31] Given the lack of gold standard rate ratios for Mexico, we found that our results are consistent with recent findings using a large administrative dataset in Brazil.[32] The risk of death was driven particularly by small babies and it persisted during the neonatal period, infancy, and five years of age (Table S4). Analysis by social deprivation also showed geographical disparities related to the place of the mother's residence, with higher mortality rate ratios among small babies living in the poorest areas compared with those living in areas with less deprivation. This may be explained by reduced infrastructure, food supplies, and access to health care in areas with high deprivation.[33,34]

For big babies, it is well described that LGA is associated with short-term complications including shoulder dystocia, brachial plexus injury, birth fractures, hypoglycaemia, and hospitalisation.[35–37] Even though the threshold of the size >90th centile used in this study does not seem to add value to delineate mortality, gestational age from 41 to 44 weeks was a better marker for this outcome. These plausible results, besides the quality assessment, support the overall performance of the linkage model.

4 | CONCLUSIONS

We have successfully performed a record linkage of two very large administrative datasets using common variables with high accuracy. We have demonstrated the utility of this linked dataset to investigate the association between adverse birth outcomes and the risk of dying prematurely during the first 5 years of age. Findings at the subnational level should be interpreted with caution as the completeness of the linkage varies between states, particularly among more deprived populations. This linked dataset can be used by other researchers for other related research and future collaborations. For example, further analyses considering children born preterm and SGA simultaneously would be valuable to provide a piece of more detailed information on childhood mortality.

AUTHOR CONTRIBUTIONS

LSI, RP, HB, EOO developed the study concept, building on the Vulnerable Newborn types from JEL and others. AB acquired the data. RP performed the linkage process with inputs from LSI and EOO. MB proved guidance of linkage techniques. JFG, ESP, JEL contributed to the interpretation of the results. EOO provided statistical guidance and oversight of the study. All authors decided to publish, revised the manuscript and approved the final version.

ACKNOWLEDGEMENTS

We want to acknowledge to Marcelino Esparza Aguilar, who provided valuable feedback to the manuscript, the Centre for Data and Knowledge Integration for Health (CIDACS) Fiocruz, Bahia, Brazil, and the Mexican Society of Public Health for their support, guidance, and time invested in this study.

FUNDING INFORMATION

This study has been funded by The Children's Investment Fund Foundation (United Kingdom) "Improving National Data and Measurement on LBW" prime grant EPIDZT83. The funder had no involvement in the planning, analysis, report or publication of results. ESP is funded by the Wellcome Trust: 213589/Z/18/Z.

CONFLICT OF INTEREST STATEMENT

We declare that we have no conflict of interest.

DATA AVAILABILITY STATEMENT

The final dataset will be available in a public repository with a technical note describing the linkage process and quality assessment. https://drive.google.com/drive/folders/1TCOUaaqcNsWeMgLISXEEWQUmREevQsUA?usp=sharing

ORCID

Lorena Suárez-Idueta https://orcid.org/0000-0003-0909-7737
Robespierre Pita https://orcid.org/0000-0002-0616-620X
Hannah Blencowe https://orcid.org/0000-0003-1556-3159
Jesus F. Gonzalez https://orcid.org/0000-0003-1932-1202
Enny S. Paixao https://orcid.org/0000-0002-4797-908X
Mauricio L. Barreto https://orcid.org/0000-0002-0215-4930
Joy E. Lawn https://orcid.org/0000-0002-4573-1443
Eric O. Ohuma https://orcid.org/0000-0002-3116-2593

REFERENCES

1. Zhang Y, Guo SL, Han LN, Li TL. Application and exploration of big data Mining in Clinical Medicine. *Chin Med J (Engl)*. 2016;129(6):731-738.
2. Bushnik T, Yang S, Kramer MS, Kaufman JS, Sheppard AJ, Wilkins R. The 2006 Canadian birth-census cohort. *Health Rep*. 2016;27(1):11-19.

3. Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking data for mothers and babies in De-identified electronic health data. *PLoS One*. 2016;11(10):e0164667.
4. Macfarlane A, Dattani N, Gibson R, et al. Health Services and Delivery Research. *Births and their Outcomes by Time, Day and Year: a Retrospective Birth Cohort Data Linkage Study*. NIHR; 2019.
5. Herman AA, McCarthy BJ, Bakewell JM, et al. Data linkage methods used in maternally-linked birth and infant death surveillance data sets from the United States (Georgia, Missouri, Utah and Washington state), Israel, Norway, Scotland and Western Australia. *Paediatr Perinat Epidemiol*. 1997;11(Suppl 1):5-22.
6. United Nations. Department of Economic and Social Affairs 2022. Accessed October 12, 2022. https://sdgs.un.org/goals
7. Ali MS, Ichihara MY, Lopes LC, et al. Administrative data linkage in Brazil: potentials for health technology assessment. *Front Pharmacol*. 2019;10:984.
8. Ramos D, da Silva NB, Ichihara MY, et al. Conditional cash transfer program and child mortality: A cross-sectional analysis nested within the 100 million Brazilian cohort. *PLoS Med*. 2021;18(9):e1003509.
9. Pescarini JM, Williamson E, Nery JS, et al. Effect of a conditional cash transfer programme on leprosy treatment adherence and cure in patients from the nationwide 100 million Brazilian cohort: a quasi-experimental study. *Lancet Infect Dis*. 2020;20(5):618-627.
10. Harron KL, Doidge JC, Knight HE, et al. A guide to evaluating linkage quality for the analysis of linked data. *Int J Epidemiol*. 2017;46(5):1699-1710.
11. INEGI. Administrative registries -statistical 2021. Accessed November 18, 2021. http://en.www.inegi.org.mx/programas/mortalidad/
12. Direccion General de Información en Salud. Sistema de Información en Salud 2020. Accessed November 18, 2021. http://www.dgis.salud.gob.mx/contenidos/sinais/subsistema1.html
13. Dublin Core Metadata Initiative. Dublin Core. 2021 Accessed October 12, 2022. https://dublincore.org/
14. Direccion General de Información en Salud. Manual de Llenado del Certificado de Nacimiento. Subsistema de Información sobre Nacimientos (SINAC). 2015 [25 September, 2021]. Accessed October 12, 2022. https://www.gob.mx/cms/uploads/attachment/file/16345/CN_ManualLlenado.pdf
15. Salud DGIS. Nacimientos Datos Abiertos. 2021 Accessed November 18, 2021. http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_nacimientos_gobmx.html
16. World Health Organization. ICD-10 Version 2019. 2012 Accessed October12, 2022. https://icd.who.int/browse10/2019/en#/XVII
17. Murguía-Peniche T, Illescas-Zárate D, Chico-Barba G, Bhutta ZA. An ecological study of stillbirths in Mexico from 2000 to 2013. *Bull World Health Organ*. 2016;94(5):322-30a.
18. Suarez-Idueta L, Bedford H, Ohuma EO, Cortina-Borja M. Maternal risk factors for small-for-gestational-age newborns in Mexico: analysis of a Nationwide representative cohort. *Front Public Health*. 2021;9:707078.
19. Paixao ES, Harron K, Andrade K, et al. Evaluation of record linkage of two large administrative databases in a middle income country: stillbirths and notifications of dengue during pregnancy in Brazil. *BMC Med Inform Decis Mak*. 2017;17(1):108.
20. Barbosa GCG, Ali MS, Araujo B, et al. CIDACS-RL: a novel indexing search and scoring-based record linkage system for huge datasets with high accuracy and scalability. *BMC Med Inform Decis Mak*. 2020;20(1):289.
21. Almeida D, Gorender D, Ichihara MY, et al. Examining the quality of record linkage process using nationwide Brazilian administrative databases to build a large birth cohort. *BMC Med Inform Decis Mak*. 2020;20(1):173.
22. Silva AA, Batista RF, Simões VM, et al. Changes in perinatal health in two birth cohorts (1997/1998 and 2010) in São Luís, Maranhão state. *Brazil Cad Saude Publica*. 2015;31(7):1437-1450.
23. Consejo Nacional de Evaluacion de la Política de Desarrollo Social (CONEVAL). Indice de Rezago Social 2020. Accessed November 18, 2021. https://www.coneval.org.mx/Medicion/IRS/Paginas/Indice_Rezago_Social_2020.aspx
24. United Nations. Sustainable Development Goals. 2021 Accessed October 12, 2022. https://sustainabledevelopment.un.org/memberstates/mexico
25. Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. *AHRQ methods for effective health care. Linking data for health services research: A framework and instructional guide*. Agency for Healthcare Research and Quality (US); 2014.
26. Bentley JP, Ford JB, Taylor LK, Irvine KA, Roberts CL. Investigating linkage rates among probabilistically linked birth and hospitalization records. *BMC Med Res Methodol*. 2012;12:149.
27. Adams MM, Kirby RS. Measuring the accuracy and completeness of linking certificates for deliveries to the same woman. *Paediatr Perinat Epidemiol*. 2007;21(Suppl 1):58-62.
28. Franco-Paredes C, del Río C, Carrasco P, Preciado JI. Response in Mexico to the current outbreak of AH1N1 influenza. *Salud Publica Mex*. 2009;51(3):183-186.
29. Blencowe H, Cousens S, Chou D, et al. Born too soon: the global epidemiology of 15 million preterm births. *Reproductive Health*. 2013;10(Suppl 1):1-14.
30. Vogel JP, Chawanpaiboon S, Moller A-B, Watananirun K, Bonet M, Lumbiganon P. The global epidemiology of preterm birth. *Best Pract Res Clin Obstet Gynaecol*. 2018;52:3-12.
31. Vilanova CS, Hirakata VN, de Souza Buriol VC, Nunes M, Goldani MZ, da Silva CH. The relationship between the different low birth weight strata of newborns with infant mortality and the influence of the main health determinants in the extreme south of Brazil. *Popul Health Metr*. 2019;17(1):15.
32. Paixao ES, Blencowe H, Falcao IR, et al. Risk of mortality for small newborns in Brazil, 2011-2018: A national birth cohort study of 17.6 million records from routine register-based linked data. *The Lancet Regional Health - Americas*. 2021;3:100045.
33. Consejo Nacional de Evaluacion de la Política de Desarrollo Social (CONEVAL). Índice de Rezago Social 2015. Presentacion de Resultados 2016. Accessed October 12, 2022. https://www.coneval.org.mx/Medicion/Documents/Indice_Rezago_Social_2015/Nota_Rezago_Social_2015_vf.pdf
34. Burstein R, Henry NJ, Collison ML, et al. Mapping 123 million neonatal, infant and child deaths between 2000 and 2017. *Nature*. 2019;574(7778):353-358.
35. Younes S, Samara M, Salama N, et al. Incidence, risk factors, and feto-maternal outcomes of inappropriate birth weight for gestational age among singleton live births in Qatar: A population-based study. *PLoS One*. 2021;16(10):e0258967.
36. Scifres CM. Short- and long-term outcomes associated with large for gestational age birth weight. *Obstet Gynecol Clin North Am*. 2021;48(2):325-337.
37. Poston L, Caleyachetty R, Cnattingius S, et al. Preconceptional and maternal obesity: epidemiology and health consequences. *Lancet Diabetes Endocrinol*. 2016;4(12):1025-1036.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Suárez-Idueta L, Pita R, Blencowe H, et al. National data linkage assessment of live births and deaths in Mexico: Estimating under-five mortality rate ratios for vulnerable newborns and trends from 2008 to 2019. *Paediatr Perinat Epidemiol*. 2023;37:266-275. doi:10.1111/ppe.12968