Comparing the usefulness of two mortality registries for data-linkage for prospective cohorts in Mexico

Susana Lozano-Esparza, MD, PhD,[1] Oscar E Zazueta, MD, MPH,[2]
Juan Eugenio Hernández-Ávila, MS, PhD,[3] Martín Lajous, MD, ScD.[4,5]

Abstract

Objective. To estimate the sensitivity and specificity of two Mexican death registries for the identification of vital status using a publicly available record-linkage tool. **Materials and methods.** We selected all reported deaths (n=581) and 575 alive participants in an epidemiologic cohort with active follow-up. Individual records were cross-linked to two mortality registries. **Results.** A sensitivity of 87.2% (95%CI: 84.7, 90.2) and specificity of 99.3% (95%CI: 98.2, 99.8) were jointly achieved with both registries. Major discrepancies in cause of death were observed in 10.8% of deaths. **Conclusion.** There is initial evidence that Mexican death registries are a valuable resource for mortality follow-up in epidemiologic studies.

Keywords: longitudinal studies; Mexico; mortality registries

Resumen

Objetivo. Estimar la sensibilidad y especificidad de dos registros mexicanos de mortalidad para la identificación de muertes. **Material y métodos.** Se seleccionaron al azar 575 participantes vivos y todas las muertes notificadas (n=581) de una cohorte con seguimiento activo. Se vinculó cada individuo utilizando una herramienta públicamente disponible. **Resultados.** Se obtuvo una sensibilidad de 87.2% (IC95%: 84.7, 90.2) y una especificidad de 99.3% (IC95%: 98.2, 99.8) con ambos registros. Hubo discrepancias en la causa de muerte en 10.8% de las defunciones. **Conclusión.** Existe evidencia inicial de que los registros mexicanos de mortalidad son un recurso valioso para el seguimiento en estudio epidemiológicos.

Palabras clave: estudios longitudinales; México; registros de mortalidad

In Mexico, the *Subsistema Epidemiológico y Estadístico de Defunciones* (SEED) and the *Instituto Nacional de Estadística y Geografía* (Inegi) independently collect and code death certificates. These death registries were created for official statistics and mortality surveillance.

Their usefulness for the linkage of individual records to external databases is unclear. As public health research capacity is strengthened in Mexico, understanding how these databases can be used for vital status assessment is essential. Following a previously established strategy,[1,2]

(1) Department of Epidemiology, University of Washington. Seattle, USA.
(2) Secretaría de Salud de Baja California. Mexicali, Mexico.
(3) Centro de Investigación en Evaluación y Encuestas, Instituto Nacional de Salud Pública. Cuernavaca, Morelos, Mexico.
(4) Centro de Investigación en Salud Poblacional, Instituto Nacional de Salud Pública. Mexico City, Mexico.
(5) Department of Global Health and Population, Harvard T.H. Chan School of Public Health. Boston, USA.

we sought to estimate the sensitivity and specificity of these databases for the identification of vital status comparing them to an epidemiologic cohort with active follow-up using a publicly available record-linkage tool.

Materials and methods

SEED

SEED is Mexico's mortality surveillance tool. Until 2014, standardized coders in health districts coded death certificates using International Classification of Diseases (ICD)-10 codes and manually attributed the underlying cause of death.[3] The database was updated continuously.[4] We accessed 2006-2014 databases after approval by the *Instituto Nacional de Salud Pública* (INSP) (CI-249-2016-1396).

Inegi

Inegi generates Mexico's official death statistics and processes death certificates independently of SEED. Information from death certificates is coded at regional offices using an adapted version of the Mortality Medical Data System for automated data entry and classification based on ICD-10 codes. Inegi's central office validates the databases periodically.[5] We obtained on-site access at Inegi to death records between 2006 and 2014.

Mexican Teachers' Cohort (MTC)

The MTC is a prospective cohort of 115 314 female teachers established in 2006-2008 with a follow-up response of 83% for the 2011-2014 questionnaire cycle.[6] Deaths were identified yearly through data linkage to human resource databases from education authorities, a pension fund database, and next-of-kin reports. As of December 31, 2014, we had identified 581 deaths and randomly selected a sample of 575 participants known to be alive (*i.e.*, answered follow-up questionnaire and were not reported dead). At the time, the national identifier (*Clave Única de Registro de Población*, CURP) were available for 70% of study participants.

Mortality linkage

We used Centers for Disease Control and Prevention's (CDC) probabilistic record linkage software Registry Plus™ Link Plus (Version 2.0) to search for deaths and women known to be alive.[7] A probability score for record pairs was generated based on the probability that the matched records belonged to the same person. We hoped to identify the ideal probability score cut-point by using

5, 7.5, and 10 (recommended range is 7-10). For SEED, we used female sex as a matching variable along with: CURP (on average 12.8% of records), names (name and two last names), and CURP plus names. For Inegi, we used female sex and names (CURP were unavailable). The linkage software used a phonetic algorithm known to accommodate Hispanic names.[8] Potential matches were manually confirmed using date and state of birth. Minor mismatches on a single field were allowed. Two independent reviewers conducted the manual assessment for 2010 and reached perfect reproducibility. See supplemental materials for more information on our linkage process.[9] We estimated the sensitivity and specificity (and 95% confidence intervals; 95%CI) of SEED (CURP, names, CURP + names), Inegi (names), and SEED (names) plus Inegi (names) to identify deaths and women known to be alive. We sought to identify a process to minimize the manual review burden. We assessed discrepancies in the underlying cause of death by comparing deaths found on both databases. Major discrepancies were a difference in the first digit of the ICD-10 or cancer site. We subclassified discrepancies in adjudication when there was no difference between the multiple causes and in coding when multiple causes´ codes differed.

Results

We found 509 out of 581 deaths: 305 appeared on both databases, while 72 were not identified by either. Fifty of these unidentified deaths (69%) occurred in the two most recent years (2013 and 2014). Three participants known to be alive were found in both databases.

Using the most inclusive probability matching score, the SEED achieved the highest sensitivity when using names (84.9% [95%CI: 81.7, 87.7]), although this method had the most records to review. When using only CURP, we achieved a low sensitivity (due to missingness), and when using CURP + names, the number of potential matches, relative to names, decreased without affecting the sensitivity substantially. In contrast, Inegi had a sensitivity of 51.1% (95%CI: 47.0, 55.3, 118 840 matched records). Using both databases, a sensitivity of 87.6% (95%CI: 84.7, 90.2) and specificity of 99.3% (95%CI: 98.2, 99.8) were achieved by reviewing 224 645 matched records. We found that using 7.5 and 10 probability scores dramatically reduced the number of records for review (38 944 and 8 595, respectively) while keeping the sensitivity above 80% (table I). Among the 305 records identified in both registries, the underlying cause of death was discordant in 18.3% (n=56), but only 10.8% (n=33) were major discrepancies (mainly occurring in cancer). Overall, we found that potential

Table I
Specificity, sensitivity and possible matched revised by search method. Mexico, 2006-2014*

	SEED			Inegi	SEED (N)+ Inegi (N)
	CURP	Name	Name + CURP	Name	
Cutoff score 5					
Sensitivity (95%CI)	6.5 (4.7, 8.9)	84.9 (81.7, 87.7)	84.7 (81.7, 87.7)	51.1 (47.0, 55.3)	87.6 (84.7, 90.2)
Specificity (95%CI)	99.3 (98.2, 99.8)	99.3 (98.2, 99.8)	99.3 (98.2, 99.8)	99.5 (98.5, 99.9)	99.3 (98.2, 99.8)
Potential matches	42	105 805	50 345	118 840	224 645
Cutoff score 7.5					
Sensitivity (95%CI)	0.0	81.8 (77.7, 84.2)	81.1 (77.7, 84.2)	49.6 (45.3, 53.7)	84.5 (81.3, 87.4)
Specificity (95%CI)	100.0	99.3 (98.2, 99.8)	99.3 (98.2, 99.8)	99.7 (98.8, 100)	99.3 (98.2, 99.8)
Potential matches	0	18 370	7 985	20 574	38 944
Cutoff score 10					
Sensitivity (95%CI)	0.0	78.8 (75.3, 82.1)	76.9 (73.3, 80.3)	47.9 (43.7, 52.0)	81.6 (78.2, 84.7)
Specificity (95%CI)	100.0	99.3 (98.2, 99.8)	99.3 (98.2, 99.8)	99.7 (98.8, 100)	99.3 (98.2, 99.8)
Potential matches	0	3 167	1 941	5 428	8 595

*All search methods included gender variable
SEED: *Subsistema Epidemiológico y Estadístico de Defunciones*
Inegi: *Instituto Nacional de Estadística y Geografía*
CURP: *Clave Única de Registro de Población*

errors in adjudication and coding were roughly similar in frequency (adjudication 30; coding 26).

We conducted a stepwise method aimed at reducing the manual record review burden where matched records were sequentially removed. We first used CURP as the matching variable in SEED and identified 38 deaths (table II). After removing these deaths, we repeated the matching process and matched-record removal using SEED (CURP + names), SEED (names only), and Inegi (names) sequentially. This method identified the same death as our previous method with an important reduction in manual record review (from 224 645 records to 23 412).

Discussion

Linkage of Mexican mortality registries using a publicly available probability record matching tool may be useful to determine vital status in epidemiologic cohorts. Strategies to increase the efficiency of manual record review can be implemented.

The sensitivity of SEED and Inegi for cohort mortality follow-up approached 90%, which is somewhat lower than the 97-98% observed in the U.S National Death

Table II
Stepwise matching strategy. Mexico, 2006-2014

Steps	Deceased participants searched	Deceased participants found	Potential matches found
1. SEED, CURP	581	38	42
2. SEED, CURP + names	543	441	7 868
3. SEED, names only	102	4	5 989
4. Inegi, names	98	26	9 513
Total		509	23 412

SEED: *Subsistema Epidemiológico y Estadístico de Defunciones*
Inegi: *Instituto Nacional de Estadística y Geografía*
CURP: *Clave Única de Registro de Población*

Index.[1,2] However, this is probably an underestimate. Most unidentified deaths were most likely due to a mortality reporting lag. Also, we expect the sensitivity to have increased after 2014 because usage of the national identifier has increased with each year. SEED was significantly better at identifying deaths relative to Inegi.

This was expected since the latter's collection of death certificate data probably does not emphasize including identifiers because these are unnecessary for national statistics. Major discrepancies between registries were among the lower range reported in the literature and are consistent with prior work.[9–11]

Our study has limitations.[10-12] We assumed as gold standard for vital status, employer and pension fund manager information, and next-of-kin reports. While 95% of participants considered alive answered a follow-up questionnaire between 2011 and 2014, three participants found in both databases were misclassified as alive by the gold standard. Our analysis assumes that the probabilistic record linkage tool is adequate for Mexico. While this tool accommodates Hispanic names, this and other database characteristics may have affected our capacity to identify some deaths. Finally, our study only included middle-aged women and results may not be fully transportable to children or men.

Conclusion

Our study provides initial evidence that national mortality databases can be used for mortality follow-up with reasonable use of human resources. SEED performs better than Inegi, but when possible, these registries should be used jointly. Our results require confirmation in other Mexican prospective studies that include different populations and age groups.

Funding

References

1. Stampfer MJ, Willett WC, Speizer FE, Dysert DC, Lipnick R, Rosner B, Hennekens C. Test of the national death index. Am J Epidemiol. 1984;119(5):837-9. https://doi.org/10.1093/oxfordjournals.aje.a113804
2. Rich-Edwards JW, Corsano KA, Stampfer MJ. Test of the National Death Index and Equifax Nationwide Death Search. Am J Epidemiol. 1994;140(11):1016-9. https://doi.org/10.1093/oxfordjournals.aje.a117191
3. World Health Organization. International Statistical Classification of Diseases and Related Health Problems (ICD). Geneva: WHO, 2021 [cited 2021 Aug 12]. Available from: https://www.who.int/standards/classifications/classification-of-diseases
4. Secretaría de Salud. Manual de Procedimientos Estandarizados para el Sistema Estadístico Epidemiológico de las Defunciones (SEED). Mexico City: SS, 2012 [cited 2021 Aug 12]. Available from: https://epidemiologiatlax.files.wordpress.com/2013/08/26_2012_manual_seed.pdf
5. Instituto Nacional de Estadística y Geografía. Estadística de defunciones generales. Síntesis metodológica. Mexico City: Inegi, 2014:55 [cited 2021 Aug 12]. Available from: http://internet.contenidos.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825063597.pdf
6. Lajous M, Ortiz-Panozo E, Monge A, Santoyo-Vistrain R, García-Anaya A, Yunes-Díaz E, *et al.* Cohort Profile: The Mexican Teachers' Cohort (MTC). Int J Epidemiol. 2017;46(2):e10. https://doi.org/10.1093/ije/dyv123
7. Centers for Disease Control and Prevention. Registry Plus TM Link Plus Technical Information and Installation. USA: CDC, 2007 [cited 2021 Aug 12]. Available from: https://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm
8. Patman F SL. Is soundex good enough for you? the hidden risks of soundex-based name searching. IBM Information Management Software, 2006.
9. Lozano-Esparza S, Zazueta OE, Hernández-Ávila JE, Lajous M. Supplemental materials: Comparing the usefulness of two mortality registries for data-linkage for prospective cohorts in Mexico [cited 2022 Feb 10]. Available from: https://figshare.com/articles/journal_contribution/Supplemental_materials_Comparing_the_usefulness_of_two_mortality_registries_for_data-linkage_for_prospective_cohorts_in_Mexico/19149566/1
10. Rutqvist LE. Validity of certified causes of death in breast carcinoma patients. Acta Radiol Oncol. 1985;24(5):385-90. https://doi.org/10.3109/02841868509134405
11. Phillips DE, Lozano R, Naghavi M, Atkinson C, Gonzales-Medina D, Mikkelsen L, *et al.* A composite metric for assessing data on mortality and causes of death: the vital statistics performance index. Popul Health Metr. 2014;12(1):14. https://doi.org/10.1186/1478-7954-12-14
12. Wise RA, Kowey PR, Austen G, Mueller A, Metzdorf N, Fowler A, *et al.* Discordance in investigator-reported and adjudicated sudden death in TIOSPIR. ERJ Open Res. 2017;3(1):00073-02016. https://doi.org/10.1183/23120541.00073-2016