Mortality according to gender identity and sexual orientation: Data relationship strategies for Rio de Janeiro, Brazil

Ricardo de Mattos Russo Rafael¹, Kleison Pereira da Silva¹, Helena Gonçalves de Souza Santos¹, Davi Gomes Depret¹, Jaime Alonso Caravaca-Morera², Karen Marie Lucas Breda³

1. State University of Rio de Janeiro, School of Nursing, Public Health Nursing Department. Rio de Janeiro, Brazil.

Corresponding author
Ricardo de Mattos Russo Rafael
E-mail: prof.ricardomattos@gmail.com

Financial support
This study received financial support from the Brazilian National Council for Scientific and Technological Development (CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico) for the research productivity fellowship (Grant number: 312056/2022-2); Carlos Chagas Filho Research Support Foundation of the State of Rio de Janeiro (Grant number: 211.970/2021); and the Program for Incentives to Scientific, Technical, and Artistic Production of the Universidade do Estado do Rio de Janeiro (PROCIENCIA/UERJ) to Ricardo de Mattos Russo Rafael.
SUMMARY

Objective: To evaluate the accuracy, potential, and limits of probabilistic data relationships to yield information on deaths according to sex identity and sexual orientation in the state of Rio de Janeiro.

Methods: This study evaluated the accuracy of the probabilistic relationship of data to obtain information on deaths according to gender and sexual orientation. Data from two information systems were used from June 15, 2015 to December 31, 2020. We constructed nine probabilistic data relationship strategies and identified the performance and cutoff points of the best strategy.

Results: The best data blocking strategy was established through logical blocks with the first and last names, birthdate, and mother's name in the pairing strategy. With a population base of 80,178 records, 1556 deaths were retrieved. With an area under the curve of 0.979, this strategy presented 93.26% accuracy, 98.46% sensitivity, and 90.04% specificity for the cutoff point \( \geq 17.9 \) of the data relationship score. The adoption of the cutoff point optimized the manual review phase, identifying 2259 (90.04%) of the 2509 false pairs and identifying 1532 (98.46%) of the 1556 true pairs. A crude mortality rate of 19.11 deaths per 1,000 people was observed, where women who had sex with women, transvestites, and transgender women had higher mortality rates than heterosexual cisgender women. Neither men nor men with gender markers had higher mortality rates than heterosexual cisgender women.

Conclusion: With the identification of possible strategies for determining probabilistic data relationships, the retrieval of information on mortality according to sexual and sex markers has become feasible. Based on information from the daily routine of health services, the formulation of public policies that consider the LGBT+ population more closely reflects the reality experienced by these population groups.

Keywords: Gender identity, sexual orientation, probability, mortality records
INTRODUCTION

Mortality has been considered an important indicator in the context of public health, especially in the area of health surveillance and for public policymakers in Brazil and worldwide. It is not by chance that numerous studies address this topic (1–5). This is because monitoring of death rates not only helps to identify the patterns and causes of death of population groups but also is useful in defining priorities for the allocation of resources and evaluating the quality of life and well-being of populations, in addition to allowing the measurement of the efficiency of the programs implemented (6,7). Given the importance of monitoring these data, since the 1970s Brazil has implemented one of the most robust information systems for monitoring deaths, the mortality information system (SIM). Under the administration of the Brazilian Ministry of Health and fully integrated into the Unified Health System, SIM can publish detailed data on the cause of death using the International Classification of Diseases-10 and several individual and clinical characteristics (8–10).

On the other hand, it is still not possible to say that information on deaths truly reaches all population groups. Paradoxically, although the literature is vast that points out that gays, lesbians, bisexuals, travestis, and transgender people (LGBT+; using the “plus” sign to represent the variety of sexual orientations and gender identities that exist) have worse health indicators than others (11–16), SIM still does not monitor specific information on sexual orientation and gender identity. Thus, travestis and trans women who were unable (or did not have the desire to) to rectify their names had their death records classified as belonging to the male population. The same occurs for trans men, who are classified as part of the female population. This is because only the sex assigned at birth and recorded in Brazilian administrative systems is identified in SIM (17,18). This is the paradox of the country that, according to nongovernmental agencies, murders the most LGBT+ people (19).

In the absence of specific information and due to the urgency of providing data that support specific public policies for these groups, the National Association of Travestis and Transgenders (ANTRA – Assoiação Nacional de Travestis e Transexuais), as well as gay activist groups, monitors deaths based on newspaper reports and information captured from hospital nets (19). These cases, exclusively focused on murders, represent only a small part of the concrete reality experienced by these people. In other words, the country is very far from familiar with the mortality profile of the LGBT+ population, harming everything from the design of public policies to the direct
care provided by health professionals (still under the fog of the unknown due to the absence of data).

In all Brazilian health information systems, information on sexual orientation and gender identity has been systematically collected only from the interpersonal and self-inflicted violence module of the Notifiable Diseases Information System (SINAN) since June 15, 2015 (20). All suspected or confirmed violent events are recorded, providing a way to expand the dataset on the LGBT+ population. With possible inaccuracies due to the absence of a unique key to identify the population in information systems, the possibility of a probabilistic relationship of data may suggest a tool to estimate deaths in the Brazilian LGBT+ population. Data relationships are a statistical tool for combining different sets of information: deterministically, when there is a unique key (identifying variable) for linking, or probabilistically when keys are used to estimate the probability that records with different names refer to the same person (21–26). As a way of contributing to the systematic integration of the technique in health services, this study aims to evaluate the accuracy, potential, and limits of the probabilistic relationship of data to obtain information on deaths according to sex and sexual orientation in the state of Rio de Janeiro.

Methods

Study type

This study quantifies the accuracy with which the probabilistic relationships of data yield information on deaths according to gender identity and sexual orientation, which integrates the analyses of the first five years of the study "Effects of (trans)gender identity and sexual orientation on notification and mortality due to violence: a cohort study".

Sources, place, and period of study

This study used data from people registered in the SINAN for interpersonal and self-inflicted violence and from the SIM data of 92 municipalities in the state of Rio de Janeiro, Brazil, from June 15, 2015 to December 31, 2020.

Population and selection criteria

To determine the probabilistic relationships between the databases, the SINAN records and occurrence of interpersonal/self-inflicted violence between June 15, 2015
and December 31, 2020 and the times of introduction of the variables “sexual orientation” and “gender identity” in the notification form were included. The records of adults aged 19 to 59 years were included. Records without information or with invalid data on birthdate, name, mother’s name, and date of occurrence of violent acts were extracted. For SIM qualification, records without identification or with invalid names or mothers’ names (e.g., indigent, male, black male, white female, etc.) were excluded. To reduce the number of records and accelerate the data relationship process, records of people younger than 19 years of age were excluded, as this was the criterion for selection in SINAN.

**Study variables**

*Variables of the probabilistic relationship of the data*

The variables used as link keys between the databases were “name”, “mother’s name”, and “birthdate”. In addition, the variables “notification number”, “date of notification”, “date of occurrence”, “death number”, and “date of death” were used in SINAN to construct an identifier code and to verify duplicate records.

*Outcome and exhibits of interest*

The outcome variable was death (all causes), which was identified and retrieved from SIM through the database link. Variables such as sex, sexual orientation, and gender identity, which were exposures of interest in this investigation, were synthesized to build analytical groups approximating the realities of exposure of people to the concrete reality. Originally present in SINAN, the possible entries included female (S0) and male (S1) for the variable “sex”; heterosexual (O0), bisexual or homosexual (O1), and “unknown” or “omitted” (O2) for “sexual orientation”; and, for the gender variable, it is important to consider that the registration was made only for *travestis* and trans women (G1) and trans men (G2), with the interpretation that data “unknown” or “omitted” (G0) referred to the cisgender population, despite the validity problems that this may produce (18).

Thus, eight groups were created. The participants were cisgender and heterosexual women (S0+O0+G0), women who had sex with women (S0+O1+G0), cisgender women of unknown sexual orientation (S0+O2+G0), *travestis* and trans women (G1), cisgender and heterosexual men (S1+O0+G0), men who had sex with men (S1+O1+G0), cisgender men with unknown sexual orientation (S1+O2+G0), and trans
men (G2). Due to the limited number of cases retrieved from the databases, the sexual orientations of transgender people were not combined in the groups formed, limiting the definition of interest groups.

Other covariates

The other covariates tested in the study, all derived from SINAN, were “age group” based on the transformation of the continuous numerical variable “age” into four strata of adults (19-29; 30-39; 40-49; ≥50). The exclusion criteria for individuals were as follows: marital status (married, single, and other), time of schooling (up to 8 years, more than 8 years), person with disability (no, yes when at least one disability was recorded), people with mental disorders and behavioral disorders (no, yes when at least one disorder was recorded), and referral to the care network (no yes when at least one referral to social, legal, or health care services was documented). The variable “race/color” was classified as black (black and mixed-race) or nonblack, following an ethical-political and conceptual approach adopted in investigations with this focus in the Brazilian context (28). The variable “chronicity of violent episodes” was created from the repeated insertion of records into SINAN, as detailed in the section on probabilistic data relationships.

Data processing and analysis

Routines of the probabilistic relationship of data

The first stage, titled preprocessing, was intended for the preparation and standardization of the databases for the probabilistic relationship of the data (22). In this stage, computational scripts were built in Python 3.11.2 software in the Visual Studio Code 1.84.2 environment. Initially, the databases were reduced, leaving only the variables of interest for linking the databases (notification number/death number, date of notification, date of occurrence/date of death, name, mother’s name, and birthdate) to facilitate the process. The study selection criteria were also applied, and unique keys were constructed for each record in the database. This construction was necessary because, at least in SINAN, the notification number, which should be unique for each record, shares the same number with other people. In other words, different people had the same notification number without being homonymous. Thus, the unique key was generated based on the combination of the notification (or death) number, birthdate, first and last name, and mother’s name of each person.
With the expectation of reducing potential spelling errors in name records, a set of standardizations was implemented. All letters were converted to uppercase letters. Accents, special characters, excessive spaces, and other punctuation were eliminated (“Ç” became “C”, “Á” became “A”, etc.), as were prepositions (“DE”, “DE”, “ DO”, “DA”, “DOS”, “DAS”, “E”). Duplicate letters were kept only once (e.g., “TT” and “SS” became “T” and “S”, respectively). The letters were standardized according to Portuguese phonetics. In this context, the syllables “WA”, “KA”, “KO”, “KU”, “CE”, “CI”, “GE”, and “GI” were replaced by “VA”, “CA”, “CO”, “CU”, “SE”, “SI”, “JE”, and “JI”, respectively. The names starting with “H” had this letter deleted (e.g., “HUGO” became “UGO”), and the letter “Y” was replaced by “I”. To handle last names in the preprocessing phase, the names were separated into fragments (first name, middle name, and last name) to optimize the next steps of building data relationships. Dates were standardized to write the day, month, and year as two, two, and four numerals, without separation by slashes.

The following steps, which included the completion of database preparation and the steps related to the probabilistic relationship of the data itself, were conducted in Link Plus version 3.0. This software, developed by the USA’s Centers for Disease Control and Prevention, was initially designed for the probabilistic relationship of cancer registries in the USA (29). However, in the field of public health, its use has spread in other contexts and countries, including Brazil (23,26,30–32).

Thus, to finalize the preparation of the databases, the deduplication technique was applied due to the possibility of multiple records in SINAN since the recurrence of different violent events throughout life and multiple notifications of the same events are quite likely. Events may be recorded by more than one professional or health unit. The deduplication technique consists of verifying these repeated records in a database (22,24). Cases of duplicity were treated according to the following rules: 1) When the occurrence number was the same, with the same notification and occurrence dates, victim’s name, mother’s name, and birthdate, even allowing for spelling variations indicative of error, the record was considered the same, so only one such record was selected. 2) When the date of occurrence was different but the name, mother’s name, and birthdate were the same, even with the presence of spelling changes, it was considered a recurrence of the event, which we treated as a recurrent/chronic case. In this second rule, only the oldest record was selected, and we counted the number of repeated events in a new variable named “event chronicity”.

After standardization and preprocessing were completed, the blocking and pairing stage began. Blocking is a method of creating logical subsets based on specific criteria of variables to reduce the number of comparisons during the database relationship. This reduces the processing time. Pairing is intended to compare the corresponding records in the databases using preestablished algorithms in the relationship software (22,24).

In the specific case of this study, nine blocking and pairing strategies were compared. The blocking strategies combined the first name (FN), last name (LN), mother’s first name (FM), mother’s last name (LM), and birthdate (BD) using the soundex code to reduce potential spelling errors and nominal variations. For the pairing strategies, combinations of the full name (N), mother’s name (M), and birthdate (BD) were used, defining the minimum probabilities of agreement (M-probability) as 0.95 and 0.65 for the name and birthdate, respectively. In addition, the direct method of pairing and an initial cutoff point of 10 were used for the scores calculated from the data.

A manual peer review was performed for each strategy. The defining criteria for true pairs were as follows: 1) same name, birthdate, and mother's name; 2) same name, mother's name, and day and month of birth but with a variation of ±2 years for the last digit of the year of birth; 3) same name, mother's name, and year of birth, with inversion of the day and month digits; 4) same name, mother's name, and day and year of birth; 5) same name, mother's name, and month and year of birth; 6) similar name and mother's name with variation in the orthographic scan; and 7) unusual name and mother's name (foreign names, for example) that agree on the date of birth.

In the postpairing phase, a new deduplication was applied to the sets of pairs formed in each strategy. The selection of the best matching and blocking strategy was based on the shortest processing time and the most true pairs after deduplication. Once the best strategy was found, the files were combined based on the unique key constructed during the preparation phase, thus joining the true pairs (deaths) with the SINAN database and all its variables. The same occurred with the SIM variables, specifically concerning true pairs. The bank combinations were performed using Stata SE 15 software.

**Statistical analysis**
The processing times, minimum and maximum scores, absolute numbers, and proportions of true pairs formed by each blocking and pairing strategy of the probabilistic data relationship were analyzed. The criteria that made up the best strategy had receiver operating characteristic (ROC) curves drawn and the areas under the curves calculated to define the best cutoff point for the matching scores. In addition, the sensitivity (%), specificity (%), accuracy (%), positive and negative likelihood ratios, number of true pairs, and number of false pairs were calculated for each cutoff point of the scores of the chosen strategy.

After the SINAN database was integrated with the true pair data defined by the chosen strategy, the proportions corresponding to each covariate of interest in the study were calculated. This included the number of identified deaths and the crude mortality rates per thousand people, comparing the initial cutoff point with that selected in the most effective matching strategy. In addition, with the objective of observing the existence of differences between the cutoff points, the proportion of agreement, the kappa statistic, and the respective p values were calculated for each covariate of the study. The interpretation of the kappa statistic was classified as full agreement (0.81-1.00), substantial agreement (0.61-0.80), moderate agreement (0.41-0.60), fair agreement (0.21-0.40), slight agreement (0-0.20), or no agreement (<0) (33,34). All analyses were performed using Stata SE 15 software.

Ethical aspects
This study followed all the ethical principles of research involving human beings. Because this study was based on information from two databases that were systematically collected by health professionals, there was no need to sign a free and informed consent form. As personal and confidential information was needed, such as names and addresses, the nonanonymized data were manipulated only by the project coordinator and, later, coded and separated into separate files, aiming to maintain the privacy of the information. The research project from which the results of this study emerged was approved by the research ethics committee of the State University of Rio de Janeiro (Rio de Janeiro, Brazil) under protocol number 5,009,244.

Results

Figure 1 shows the sample composition that formed the baseline of the SINAN probabilistic relationship of interpersonal and self-inflicted violence, as well as the SIM rating. Note that 60.35% (n = 128,620) of the records were excluded from SINAN based
on the selection criteria. In addition, 3639 cases of notification recurrence were identified—i.e., when the case was the same but with different events and dates. In this case, only one record was kept, for which purpose we created a specific variable, which represented the chronicity of violent episodes. Thus, the baseline consisted of 80,178 records. In turn, SIM had 90,279 (10.12%) exclusions of records based on the qualification criteria of the database, totaling 737,493 death records (Figure 1).

Table 1 presents the results of the probabilistic relationship according to the 9 blocking and pairing strategies. strategy 1, although not the one with the highest percentage of true positives after deduplication, had the highest absolute number of true pairs (n = 1156) and the shortest data processing time (11 minutes and 50 seconds). Therefore, according to the preestablished criteria, the best strategy for further analysis was strategy 1.

Figure 2 shows the ROC curve of strategy 1. With an area under the curve of 0.979 (95% confidence interval: 0.976-0.983), the probabilistic relationship has an excellent ability to discriminate between true and false pairs. Table 2 lists the properties of the strategy. When a cutoff point of 17.9 was used for the score, the sensitivity was 98.46%, the specificity was 90.04%, and the positive and negative likelihood ratios were 9.84 and 0.02, respectively. This means this cutoff point allowed the correct identification of 98.46% of the true pairs and 93.26% of the false pairs. In addition, the likelihood ratio values show that a positive rating in the relationship significantly increased the chances of the pair being truly positive, while a negative classification, due to the very low negative likelihood ratio, provided strong evidence that the pair was not a true pair. Finally, when analyzing the accuracy, we observed that this cutoff point could correctly classify 93.26% of the observations, i.e., 1532 (98.46%) of the 1556 true pairs, as well as correctly classify 2259 (90.04%) of the 2509 false pairs, greatly reducing the need for manual review after the probabilistic relationship of the data.

Table 3 presents the characteristics of the study population; the crude mortality rates for the two cutoff points of the scores generated in the probabilistic relationship of the data; and the kappa statistic for each population subgroup. The study population was mostly composed of women (76.72%), whether cisgender or transgender, aged up to 39 years (73.74%), black (65.5%), single (54.4%), and having more than 8 years of school (66.2%). A total of 2.31% of people with disabilities were identified, 4.77% of the population reported some mental or behavioral disorders, and 3.77% had chronic
episodes of violence. Most patients (78.12%) had been referred to the assistance network.

Compared with those in the female population, mortality rates were ~3 times greater in the male population. Notably, after excluding groups of women with an unknown sexual orientation, the women who had sex with women and transvestites and transwomen had higher mortality rates than did cisgender women. On the other hand, this behavior was not observed in the male group. Higher mortality rates are also observed with advancing age; in black, married and nonsingle people; with less schooling; in people with disabilities or mental and behavioral disorders; and in cases of chronic violence. In contrast, people who received referrals to the health care network had lower mortality rates. With the kappa statistic indicating full agreement for all covariates (p<0.001), the adoption of the cutoff point of ≥17.9 did not produce substantial differences in the interpretation of the results.

Discussion

The main results of this study are focused on demonstrating the feasibility and good accuracy of determining the probabilistic relationship of data between the SINAN database of interpersonal and self-inflicted violence and the SIM database to obtain information on sexual orientation and gender identity. The study demonstrated the analytical limits of mortality as a function of these variables. Keeping a certain pioneering spirit, as the data linkage technique is used more frequently in health conditions without taking into account the markers of gender and sexuality, these results significantly contribute to overcoming important gaps on the subject in Brazil and in the world. The determination of a linkage strategy, with parsimonious parameters for the identification of true and false pairs, tends to contribute to the use of the technique in the daily routine of health services and in academic production on this topic, which is still considered a scarce area of knowledge.

By adopting preprocessing routines and considering the relationship time as a function of the number of true pairs detected, the recognition of the best blocking and processing strategies constitutes an advance for the systematic adoption of this method in the daily life of the services, especially when considering sexual orientation and gender identity as exposure variables in future analyses. These procedures resulted in a true-pair detection rate with proportions compatible with those found in other investigations whose outcome variable was the registration of death in SIM (21,35).
On the other hand, it is important to note that we are faced with slightly different connecting keys between these investigations. In this article, the linking keys were restricted to the variables “name”, “birthdate”, and “mother’s name”. Although the municipality of residence is traditionally used in probabilistic relationships, in the specific case of our study, this variable could bring more elements of uncertainty than solutions. This is because when reaching groups such as transvestites and transwomen, as well as people known to be victims of violence, depending on the population base used, it is imperative to consider the possibility of constant changes in addresses over time due to stigma and the structure of violence itself that surrounds them (36). There has been a consensus in Brazilian production that housing instability is a factor present in the lives of a significant portion of transvestites and transwomen (11,37). These distinctions in the probabilistic data matching strategies necessitate caution when comparing the data linkage results of different studies.

Complementing this analysis, the definition of a cutoff point in the scores of the probabilistic relationship of data constitutes another important aspect for the advancement of the adoption of this technique in analyses both by health surveillance services and by the academic sector. Recognizing the cutoff point, the manual peer review phase tends to be more efficient (5,25), significantly reducing processing time and improving accuracy in future data relationships. The adoption of scores ≥17.9 reduced 56.16% of the pairs to be reviewed, with 90.04% of the pairs being truly negative. In other words, with high specificity (>90.00%) and with the loss of only 25 true pairs erroneously classified as false (98.46% sensitivity), our strategy demonstrated excellent properties and elements compatible with other linkage strategies with Brazilian datasets (21,35). In addition, when incorporating the deaths into the SINAN database, the degrees of agreement for the kappa indicator are high and close to 1 for all the variables of interest in this study, revealing that the adoption of a higher cutoff point (≥17.9) does not imply losses for future analyses.

The mortality rates observed at the two cutoff points established by the linkage strategy revealed a scenario consistent with the investigations on the subject. That is, advanced age, black race/color, a shorter length of schooling, and disabilities and mental and behavioral disorders (38–40) were associated with higher mortality rates. Thus, these results corroborate the effectiveness of the probabilistic data relationship strategy in the production of reliable information on this topic.
The same occurs in relation to gender markers and sexual orientation. It was observed that transwomen and those who have sex with other women have higher mortality rates than heterosexual cisgender women. Men, whether they are cisgender or transgender, as well as men who have sex with men, also have a higher risk of dying than heterosexual cisgender women (41–44). Thus, having established the cutoff point and the best strategies for capturing information in large datasets, it is essential to scrutinize in more detail how sexual and gender markers affect the risk of dying in the population.

Furthermore, the number of deaths among *travestis* and transgender people is likely higher than that captured in the linkage process, resulting in classification bias. While cisgender men and women have their names established passively, that is, during birth (45), transgender and nonbinary people are actively involved in the choice of this name and still face important bureaucratic barriers to be able to make it compatible with their identity (17). Thus, when considering death as a time-dependent variable, it is possible that the name registered in the notification form, the name of the civil registry (name attributed to birth) (20), can be modified until the moment of death. Thus, in these cases, the death will not be attributed to the person, resulting in a false-negative result and, once again, approaching the risks between the cisgender and transgender people.

Additionally, although the use of the method is promising for the analysis of mortality according to sexual orientation and gender identity, other precautions are necessary. Although the quality of filling out the SINAN has improved over the years (46), notably, the present study found that sexual orientation was not reported (i.e., it was considered “unknown” or “omitted”) in 44.45% of the patients. A similar situation can be observed regarding “race/color”, “marital status”, and, even more markedly, regarding the variable “educational level”.

The quality and completeness of information in Brazilian health information systems have been constant and long-standing challenges for researchers and health managers since the omission of such data greatly compromises the analytical capacity and formulation of public policies (47–49). Because they are constant variables and are known to be associated with asymmetries and inequities in health (50–54), their omission in analytical models, either as a confounding variable or, in some cases, as an exposure variable, can produce strong biases in the production of knowledge on the subject. In this vein, reflecting on professional awareness strategies for the
incorporation of these practices of sensitive and attentive questioning to users, as well as awareness that these factors are preponderant for the construction of health actions, are currently urgent elements.

Special care should also be taken concerning the classification of the variable “gender identity”. Although it includes all the responses, the method for classifying the cisgender population in SINAN needs to be closely related since the form used in the system only takes into account the records of travestis, trans women and trans men (20). In other words, people who were not registered in the form, or rather, who had an “unknown” record (represented in the form by the options “unknown” or “omitted”), were deemed cisgender. On the other hand, people without a cisgender personality who had not reported their true identity or were not even asked about it, as well as people without a binary personality whose identity could not be recorded, also had an “unknown” record. This scenario especially lowers the sensitivity of the instrument, so in mortality analyses, it is possible that the cisgender data are skewed by miscategorized deaths of others, resulting in a potentially erroneous approximation of the risks among cisgender people and transvestites, transwomen, and transmen (18).

In addition to the limits already exposed and that must be taken into account when interpreting probabilistic relationships with these datasets, an important selection bias must be considered in future analyses. When studying the only database that contains information on sexual orientation and gender identity, the SINAN database of interpersonal and self-inflicted violence, we partially apprehended the actual information experienced by the population. This is because not all people suffer violence, and not all people who suffer violence are notified in health services.

Studies have shown substantial differences in the prevalence of violence between primary surveys and the results obtained by SINAN (55–57). A notable example of these differences is the fact that the surveys indicate psychological violence as the most prevalent (58), while the SINAN data indicate physical violence as the one with the greatest magnitude of violent events (46). This difference may be associated with health professionals’ own understanding of what the phenomenon of violence is, as well as the interpretation of which violence is worthy of reporting and which is not, since notions about this phenomenon are influenced by the sociohistorical conformations of society (59). In other words, culture, values, and social structure directly affect people’s perceptions of what is and is not considered violence. They also produce a kind of stratification of which types of violence are more violent than others.
and, consequently, those that are worthy of immediate reporting and those that do not need to be recorded urgently (60). Thus, physical and sexual violence, as well as cases that imminently threaten life, might be more common than others. This information bias results in the generation of SINAN results that are different from the reality estimated by surveys. In this context, analyses derived from linkages whose population base is based on SINAN should be interpreted with caution, as the records refer to a population subjected to specific contexts that may be different from those observed in the general population.

This bias can be circumvented by introducing a unique identification key (17) in information systems, such as civil registry numbers, the natural persons registry, or the national registry of the health system, as well as the introduction of the variables “sexual orientation” and “gender identity” into other data collection systems and not only in the SINAN database of interpersonal and self-inflicted violence (17,61). Until then, even with recognized inaccuracies and analytical limits, the performance of procedures involving probabilistic data relationships should be encouraged to better understand the needs of certain subpopulations, improve the technique for capturing information about the LGBT+ population, and thereby advance the formulation of public policies compatible with the needs of this group.

CONCLUSION

Despite the inaccuracies produced by the data collection format of the variables “gender identity” and “sexual orientation”, the technique of probabilistic data linkage is an important technique with high applicability in the daily routine of health services. With its identification of possible blocking and pairing strategies, as well as the detection of the best cutoff point for linkage scores, this technique becomes more useful in daily research and monitoring by health surveillance services. Although it was not the focus of this investigation, it is notable that women who have sex with women, transvestites, and transgender women have higher mortality rates than do cisgender and heterosexual women. Similarly, people with a male identity have higher mortality rates than women, with no substantial difference according to sexual orientation or gender identity.

REFERENCES


40. Ferdows NB, Aranda MP, Baldwin JA, Baghban Ferdows S, Ahluwalia JS, Kumar A. Assessment of Racial Disparities in Mortality Rates Among Older Adults Living in US


Figure 1. Flowchart of the study population composition for probabilistic record linkage. State of Rio de Janeiro, Brazil, 2015-2020.
Table 1. The results of probabilistic record linkage according to blocking and matching strategies. State of Rio de Janeiro, Brazil, 2015-2020.

<table>
<thead>
<tr>
<th>ID</th>
<th>Blocking strategy</th>
<th>Matching strategy</th>
<th>Score</th>
<th>Processing Time</th>
<th>P1</th>
<th>TP1</th>
<th>%TP1</th>
<th>PD</th>
<th>P2</th>
<th>TP2</th>
<th>%TP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FN + LM</td>
<td>N + BD + M</td>
<td>10.6-29.7</td>
<td>00:11:50</td>
<td>4087</td>
<td>1578</td>
<td>38.61</td>
<td>22</td>
<td>4065</td>
<td>1556</td>
<td>38.28</td>
</tr>
<tr>
<td>2</td>
<td>FN + LN + BD + FM + LM</td>
<td>N + BD + M</td>
<td>10.6-29.7</td>
<td>02:28:54</td>
<td>7296</td>
<td>1570</td>
<td>21.52</td>
<td>19</td>
<td>7277</td>
<td>1551</td>
<td>21.31</td>
</tr>
<tr>
<td>3</td>
<td>FN + LN + FM + LM</td>
<td>N + BD + M</td>
<td>10.6-29.7</td>
<td>00:25:46</td>
<td>7262</td>
<td>1570</td>
<td>21.62</td>
<td>20</td>
<td>7242</td>
<td>1550</td>
<td>21.4</td>
</tr>
<tr>
<td>4</td>
<td>FN + LN + FM + LM</td>
<td>N + BD</td>
<td>12.1-20.0</td>
<td>00:20:24</td>
<td>2388</td>
<td>1111</td>
<td>46.52</td>
<td>10</td>
<td>2378</td>
<td>1101</td>
<td>46.3</td>
</tr>
<tr>
<td>5</td>
<td>FN + LN + FM + LM</td>
<td>BD + M</td>
<td>10.6-18.5</td>
<td>00:19:11</td>
<td>5561</td>
<td>996</td>
<td>17.91</td>
<td>9</td>
<td>5552</td>
<td>987</td>
<td>17.78</td>
</tr>
<tr>
<td>6</td>
<td>FN + LN + BD</td>
<td>N + BD + M</td>
<td>10.6-29.7</td>
<td>02:19:07</td>
<td>7296</td>
<td>1570</td>
<td>21.52</td>
<td>20</td>
<td>7276</td>
<td>1550</td>
<td>21.3</td>
</tr>
<tr>
<td>7</td>
<td>FN + LN + BD</td>
<td>N + BD</td>
<td>12.1-20.0</td>
<td>01:32:52</td>
<td>2397</td>
<td>1103</td>
<td>46.02</td>
<td>10</td>
<td>2387</td>
<td>1093</td>
<td>45.79</td>
</tr>
<tr>
<td>8</td>
<td>FN + BD</td>
<td>N + BD + M</td>
<td>10.6-29.7</td>
<td>02:10:48</td>
<td>7296</td>
<td>1570</td>
<td>21.52</td>
<td>20</td>
<td>7276</td>
<td>1550</td>
<td>21.3</td>
</tr>
<tr>
<td>9</td>
<td>BD</td>
<td>N + BD + M</td>
<td>10.6-29.7</td>
<td>02:07:21</td>
<td>7296</td>
<td>1570</td>
<td>21.52</td>
<td>20</td>
<td>7276</td>
<td>1550</td>
<td>21.3</td>
</tr>
</tbody>
</table>

Legend:

**ID**: identification of the data linkage strategy; **P1**: pairs formed after data linkage; **TP1**: proportion of true pairs formed after data linkage; **PD**: number of duplicate pairs; **P2**: pairs formed after the deduplication strategy; **TP2**: true pairs formed after deduplication; **%TP2**: proportion of true pairs formed after deduplication; **DB**: birthdate; **N**: name; **FN**: soundex of the first name; **LN**: soundex of the last surname; **M**: mother's name; **FM**: soundex of the mother's first name; **LM**: soundex of the mother's last surname.
**Figure 2.** ROC curve of the best blocking and matching strategy (strategy number 1). State of Rio de Janeiro, Brazil, 2015-2020 (n=4065).
<table>
<thead>
<tr>
<th>Cutoff</th>
<th>S%</th>
<th>E%</th>
<th>Ac%</th>
<th>LR+</th>
<th>LR-</th>
<th>PT</th>
<th>PF</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 10.6</td>
<td>100.00</td>
<td>-</td>
<td>38.28</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>906</td>
</tr>
<tr>
<td>≥ 12.1</td>
<td>100.00</td>
<td>36.11</td>
<td>60.57</td>
<td>1.00</td>
<td>-</td>
<td>3</td>
<td>1.088</td>
</tr>
<tr>
<td>≥ 12.3</td>
<td>99.81</td>
<td>79.47</td>
<td>87.26</td>
<td>1.56</td>
<td>-</td>
<td>8</td>
<td>133</td>
</tr>
<tr>
<td>≥ 13.8</td>
<td>99.29</td>
<td>84.77</td>
<td>90.33</td>
<td>4.86</td>
<td>0.002</td>
<td>13</td>
<td>131</td>
</tr>
<tr>
<td>≥ 16.5</td>
<td>98.46</td>
<td>90.00</td>
<td>93.23</td>
<td>6.52</td>
<td>0.008</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>≥ 17.9</td>
<td>98.46</td>
<td>90.04</td>
<td>93.26</td>
<td>9.84</td>
<td>0.02</td>
<td>367</td>
<td>199</td>
</tr>
<tr>
<td>≥ 18.0</td>
<td>74.87</td>
<td>97.97</td>
<td>89.13</td>
<td>9.88</td>
<td>0.02</td>
<td>110</td>
<td>9</td>
</tr>
<tr>
<td>≥ 20.0</td>
<td>67.8</td>
<td>98.33</td>
<td>86.64</td>
<td>36.83</td>
<td>0.26</td>
<td>175</td>
<td>14</td>
</tr>
<tr>
<td>≥ 21.7</td>
<td>56.56</td>
<td>98.88</td>
<td>82.68</td>
<td>40.50</td>
<td>0.33</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>≥ 23.5</td>
<td>55.78</td>
<td>99.32</td>
<td>82.66</td>
<td>50.68</td>
<td>0.44</td>
<td>73</td>
<td>13</td>
</tr>
<tr>
<td>≥ 29.7</td>
<td>51.09</td>
<td>99.84</td>
<td>81.18</td>
<td>82.33</td>
<td>0.44</td>
<td>795</td>
<td>4</td>
</tr>
<tr>
<td>&gt; 29.7</td>
<td>0</td>
<td>100.00</td>
<td>61.72</td>
<td>320.48</td>
<td>0.49</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Legend:** S%: sensitivity value expressed as a proportion; E%: specificity value expressed as a proportion; Ac%: accuracy value expressed as a proportion; LR+: positive likelihood ratio; LR-: negative likelihood ratio; TP: true pairs; FP: false pairs.
Table 3. Characteristics of the study population, crude mortality rates (per 1000 persons), and agreement among the cutoff points of the best strategy (#1) for each population subgroup. State of Rio de Janeiro, Brazil, 2015-2020 (n=80.178)

<table>
<thead>
<tr>
<th>Variables</th>
<th>n</th>
<th>%</th>
<th>Cutoff ≥ 10.60</th>
<th>Cutoff ≥ 17.90</th>
<th>% Agreement</th>
<th>Kappa</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>80178</td>
<td></td>
<td>1556</td>
<td>19.41</td>
<td>1532</td>
<td>19.11</td>
<td>99.97</td>
</tr>
<tr>
<td>Groups</td>
<td>80169</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HCW</td>
<td>28987</td>
<td>36.2</td>
<td>285</td>
<td>9.83</td>
<td>281</td>
<td>9.69</td>
<td>99.99</td>
</tr>
<tr>
<td>WSW</td>
<td>1159</td>
<td>1.45</td>
<td>13</td>
<td>11.22</td>
<td>13</td>
<td>11.22</td>
<td>100</td>
</tr>
<tr>
<td>CWUnknow</td>
<td>30884</td>
<td>38.5</td>
<td>474</td>
<td>15.35</td>
<td>467</td>
<td>15.12</td>
<td>99.98</td>
</tr>
<tr>
<td>TGW</td>
<td>472</td>
<td>0.59</td>
<td>7</td>
<td>14.83</td>
<td>7</td>
<td>14.83</td>
<td>100</td>
</tr>
<tr>
<td>MHC</td>
<td>4496</td>
<td>5.61</td>
<td>172</td>
<td>38.26</td>
<td>171</td>
<td>38.03</td>
<td>99.98</td>
</tr>
<tr>
<td>MSM</td>
<td>534</td>
<td>0.67</td>
<td>11</td>
<td>20.60</td>
<td>10</td>
<td>18.73</td>
<td>99.81</td>
</tr>
<tr>
<td>CMUnknow</td>
<td>13550</td>
<td>16.9</td>
<td>591</td>
<td>43.62</td>
<td>580</td>
<td>42.80</td>
<td>99.92</td>
</tr>
<tr>
<td>TGM</td>
<td>87</td>
<td>0.11</td>
<td>3</td>
<td>34.48</td>
<td>3</td>
<td>34.48</td>
<td>100</td>
</tr>
<tr>
<td>Age group</td>
<td>80178</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19-29 years</td>
<td>35327</td>
<td>44.1</td>
<td>494</td>
<td>13.98</td>
<td>483</td>
<td>13.67</td>
<td>99.97</td>
</tr>
<tr>
<td>30-39 years</td>
<td>23800</td>
<td>29.7</td>
<td>360</td>
<td>15.13</td>
<td>353</td>
<td>14.83</td>
<td>99.97</td>
</tr>
<tr>
<td>40-49 years</td>
<td>13848</td>
<td>17.3</td>
<td>342</td>
<td>24.70</td>
<td>339</td>
<td>24.48</td>
<td>99.98</td>
</tr>
<tr>
<td>≥ 50 years</td>
<td>7203</td>
<td>8.98</td>
<td>360</td>
<td>49.98</td>
<td>357</td>
<td>49.56</td>
<td>99.96</td>
</tr>
<tr>
<td>Race/color</td>
<td>62620</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Nonblack</td>
<td>21629</td>
<td>34.5</td>
<td>384</td>
<td>17.75</td>
<td>377</td>
<td>17.43</td>
<td>99.97</td>
</tr>
<tr>
<td>Black</td>
<td>40991</td>
<td>65.5</td>
<td>818</td>
<td>19.96</td>
<td>808</td>
<td>19.71</td>
<td>99.98</td>
</tr>
<tr>
<td>Marital status</td>
<td>47110</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Married and others status</td>
<td>21461</td>
<td>45.6</td>
<td>347</td>
<td>16.17</td>
<td>340</td>
<td>15.84</td>
<td>99.97</td>
</tr>
<tr>
<td>Single</td>
<td>25649</td>
<td>54.4</td>
<td>404</td>
<td>15.75</td>
<td>399</td>
<td>15.56</td>
<td>99.98</td>
</tr>
<tr>
<td>Schooling</td>
<td>29326</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 8 years of study</td>
<td>19423</td>
<td>66.2</td>
<td>195</td>
<td>10.04</td>
<td>192</td>
<td>9.89</td>
<td>99.98</td>
</tr>
<tr>
<td>Up to 8 years of study</td>
<td>9903</td>
<td>33.8</td>
<td>235</td>
<td>23.73</td>
<td>231</td>
<td>23.33</td>
<td>99.96</td>
</tr>
<tr>
<td>Persons with disabilities</td>
<td>80178</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>No</td>
<td>78325</td>
<td>97.7</td>
<td>1463</td>
<td>18.68</td>
<td>1440</td>
<td>18.38</td>
<td>99.97</td>
</tr>
<tr>
<td>Yes</td>
<td>1853</td>
<td>2.31</td>
<td>93</td>
<td>50.19</td>
<td>92</td>
<td>49.65</td>
<td>99.95</td>
</tr>
<tr>
<td>Persons with a mental or behavioral disorder</td>
<td>80178</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>No</td>
<td>76350</td>
<td>95.2</td>
<td>1368</td>
<td>17.92</td>
<td>1346</td>
<td>17.63</td>
<td>99.97</td>
</tr>
<tr>
<td>Yes</td>
<td>3828</td>
<td>4.77</td>
<td>188</td>
<td>49.11</td>
<td>186</td>
<td>48.59</td>
<td>99.95</td>
</tr>
<tr>
<td>Chronicity of violent episodes</td>
<td>80178</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>No</td>
<td>77159</td>
<td>96.2</td>
<td>1493</td>
<td>19.35</td>
<td>1469</td>
<td>19.04</td>
<td>99.97</td>
</tr>
<tr>
<td>Yes</td>
<td>3019</td>
<td>3.77</td>
<td>63</td>
<td>20.87</td>
<td>63</td>
<td>20.87</td>
<td>100</td>
</tr>
<tr>
<td>Referral to the assistance network</td>
<td>80178</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>No</td>
<td>17545</td>
<td>21.9</td>
<td>438</td>
<td>24.96</td>
<td>428</td>
<td>24.39</td>
<td>99.94</td>
</tr>
<tr>
<td>Yes</td>
<td>62633</td>
<td>78.1</td>
<td>1118</td>
<td>17.85</td>
<td>1104</td>
<td>17.63</td>
<td>99.98</td>
</tr>
</tbody>
</table>

Legend: MR: Mortality Ratio; HCW - Heterosexual cisgender women; WSW - Woman who had sex with women; CWUnknow - Cisgender women of unknown sexual orientation; TGW - Transvestites and Transgender women; HCM - Heterosexual cisgender men; MSM - Men who had sex with men; CMUnknow - Cisgender men of unknown sexual orientation; TGM - Transgender men