RESEARCH

Open Access

# Probabilistic data linkage to study the epidemiology of unintentional fatal drowning in a large metropolitan region

Rohit P. Shenoi[1*], Ned Levine[2], Elizabeth A. Camp[1], Linh Nguyen[3], Sandra McKay[4] and Shaila Zaman[1]

## Abstract

**Background** It is difficult to study the epidemiology of drowning at the regional level because of multiple data sources, many of which have a high degree of unstandardized and missing data. We aimed to link multiple datasets to identify demographics and geographic locations of unintentional fatal drowning in a metropolitan region and compare linked data with vital statistic data.

**Methods** This cross-sectional study included unintentional drowning fatalities among persons of all ages in metropolitan Houston between 2016 and 2022. Probabilistic linking was used to link multiple datasets and geographical mapping to identify drowning locations. The effectiveness of data linkage was assessed by the recall, precision, and F1-score (harmonic mean of precision and recall). Geographic location of drownings by aquatic body was studied. Drowning burden by demographics, county and aquatic body were compared between linked data and vital statistic data using Chi-square tests.

**Results** Data from 8 datasets were linked. The linkage metrics were Recall (0.88-1.00), Precision (0.91-1.00), and F1-score (0.91-1.00) for datasets. There were 790 drowning fatalities. The median age was 40 years (IQR: 18.5,60); 71% were males. Children aged 0–17 years constituted 24% of drowning fatalities. Drownings occurred in swimming pools (27%), bathtubs (19%) natural water (27%), flood control structures (20%), and during flooding events (6%). Adults commonly drowned in natural water, flood control structures, bathtubs, and during flooding events whereas most toddlers drowned in swimming pools and hot tubs. No significant differences in counts by age group, sex, county of drowning and body of water were observed between linked and vital statistic data. Drowning locations were geocoded in 769 of 790 records (97%).

**Conclusion** Probabilistic data linkage can accurately determine the epidemiology of fatal drowning in a metropolitan region. Identification of high-risk drowning subpopulations and locations can inform drowning countermeasures at the regional level.

**Keywords** Probabilistic data linkage, Drowning burden, Geocoding, Drowning epidemiology

*Correspondence:
Rohit P. Shenoi
rshenoi@bcm.edu
[1]Division of Emergency Medicine, Department of Pediatrics, Texas Children's Hospital, Baylor College of Medicine, 6621 Fannin St., Suite A 2210, Houston, TX 77030, USA
[2]Ned Levine and Associates, Houston, TX 77025, USA
[3]Center for Population Health Management & Quality, Department of Healthcare Transformation Initiatives, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[4]Department of Pediatrics at McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

## Background

Drowning is the leading cause of death in US children aged 1–4 years [1, 2] and the third-leading cause of injury death worldwide [3]. Between 2018 and 2023, there were an average of 4,430 unintentional drowning deaths annually in the United States [4]. Texas (1.44 per 100,000) and the Houston metropolitan region (1.60 per 100,000) had higher unintentional fatal drowning rates than the US average (1.31 per 100,000) [4, 5]. While fatal drowning rates among persons ≤ 29 years in the United States have decreased in recent times, disparities in drowning fatality rates between individuals belonging to minority and non-minority groups persist [6]. This underscores the need to study the epidemiology of drowning and the social context for implementing effective countermeasures.

There is no single data source for fatal and non-fatal drowning in the Houston metropolitan region. Multiple data sources such as hospital, emergency medical service (EMS), medical examiner (ME) records, police records, media reports, and data from maritime, recreational, and administrative entities exist. Data are collected separately, stored in unique systems, and are not standardized. Data are often sparse, missing, or lacking identifiers making the linking of multiple datasets challenging. Data linkage for drowning surveillance is a drowning prevention research priority of the Centers for Disease Control and the US National Water Safety Action Plan [7, 8].

Linking methodology has been used to study the circumstances, outcomes and economic costs of motor vehicle crashes [9–11]. This approach has also investigated firearm injuries [12]. Data linkage between ambulance and hospital medical records have studied the demographics and outcome of drowned patients [13, 14]. In this study, probabilistic linkage was utilized to study the epidemiology of fatal unintentional drowning in a large metropolitan area.

## Methods

### Setting

This was a cross-sectional study with retrospective data collection of unintentional fatal drownings that occurred in metropolitan Houston, Texas between January 1, 2016 to December 31, 2022. The metropolitan region comprises the eight counties of Brazoria, Chambers, Galveston, Fort Bend, Harris, Liberty, Montgomery, and Waller The region falls along the Texas Gulf coast and includes the very large Galveston Bay, three river systems, four large lakes and numerous estuaries, bays, and inlets (Supplemental Fig. 1). The population of the region in 2022 was 7,092,073. Children under age 18 constituted 26% of the population. The demographic composition consisted of non-Hispanic White (34%), African-American (19%), Asian (8%), American Indian/Alaskan Native (1%), and Hispanic (39%) [15].

### Data sources and quality

Data were obtained from multiple sources through freedom of information requests (US Coast Guard (USCG), Texas Parks and Wildlife (TPWL), police reports), publicly available sources (Consumer Product Safety Commission, National Oceanic and Atmospheric Administration (NOAA), county tax records, media reports, hurricane data, location of pools from schools, municipalities, and non-profit organizations, and online searches) and data-use agreements (hospital, EMS and ME records, syndromic surveillance data) for the Houston metropolitan region (Supplemental Table 1). Police reports, hurricane (flooding events) and beach rescue

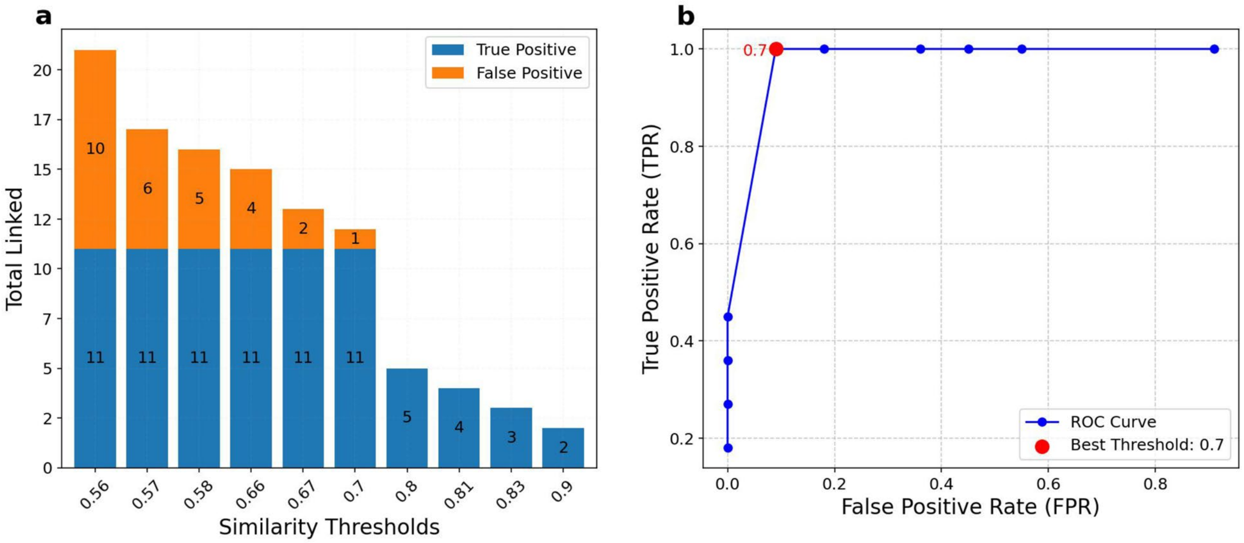

**Fig. 1** (**a**) True positive and false positive cases by similarity threshold; (**b**) Selection and validation of similarity threshold for linked medical examiner and EMS datasets using ROC curve

data were merged with media reports. Persons of all ages who unintentionally drowned in constructed bodies of water (swimming pools, hot tubs, bathtubs, buckets), natural water (gulf/bay, lake, river, other-open water), flood control structures (bayou, canal, creek, ditch, concrete waterway, reservoir, pond, retention pond), or during flooding events were included. This study received institutional review board approval from the Baylor College of Medicine (H-51536), University of Texas at Houston (HSC-MS-22-0427) and Department of State Health Services Texas (IRB# 24 – 020).

Study variables were selected based on existing literature [16, 17]. Variables included age, sex, race/ethnicity (White, Black, Asian, Hispanic, other), date of drowning event/date of death, body of water (swimming pool, hot tub, bathtub, natural water (gulf/bay, lake, river, other-open water), flood control structures (bayou, canal, creek, ditch, concrete waterway, reservoir, pond, retention pond), flooding event-related, other), address/geographic coordinates and county of drowning. Data on drowning risk and protective factors (supervision, swimming pool fencing), swimming ability, life jacket use, drug/alcohol as a contributing factor, comorbidities, CPR, adverse weather and water conditions), activity of deceased, and drowning circumstances were collected but not analyzed in this paper.

EMS records with a clinical assessment of drowning and hospital medical records of patients with a primary or secondary International Classification of Diseases, Clinical Modification (ICD-10-CM) discharge diagnosis code (ICD-10: W65-W74, V90, V92, W22.041, W16 (with 6th character = 1 except 16.4 and 16.9 where 5th character = 1, W22.041) and T75.1XXA) for fatal or non-fatal drowning were obtained and reviewed. Drownings due to intentional (suicide, homicide) and indeterminate causes (ICD-10: X71, X92, Y21) were excluded. One investigator and research associate, who were trained in data entry at each of two participating sites, entered data into a standardized electronic clinical research form (eCRF) in the Research Electronic Data Capture tool (Vanderbilt University, Nashville, TN), consistent with recommended standards [18, 19]. The eCRF was pilot tested for clarity and ease of usage. A data dictionary with definitions for study variables and a manual of operations were created and made available. Data discrepancies were resolved by contacting the principal investigator. Only the initial visit of the drowned patient was selected. Inter-hospital transfer cases were consolidated into one case record. Unavailable data were labeled missing.

For coding information from medical records, quality assurance included three iterations of random double entry to resolve errors and ensure consistency in coding practices. Both sites performed inter-rater reliability analysis of key variables on 10% of charts using Krippendorff's alpha and 1,000 bootstrapping samples. The inter-rater reliability analysis (Krippendorff's alpha) was near perfect for demographics (0.82-1.00), substantial for body of water, supervision, resuscitation (0.6–0.87), and moderate to substantial for protective device, swimming ability (0.43–0.68) between both sets of reviewers. After probabilistic linkage of cases, two researchers coded and compared drowning risk and protective factors in the linked dataset. The accuracy rate was very high (99%–100%). Risk and protective factors were also obtained from narratives using Large Language Model (LLM). The outputs of Open AI's LLM were validated by manually analyzing the narratives for ground truth in approximately 20% of media report charts. OpenAI API, 2024, GPT-4o, date of use: May, 2025 was used.

## Data linkage

Data linkage consisted of four steps: data preprocessing, probabilistic linking, post-linkage verification, and spatial validation.

### *Data preprocessing*

Data quality varied among multiple datasets. Raw data were unstructured, with inconsistencies, missing values, and varying formats that could lead to inconsistent record matching. Data preprocessing and cleaning were essential to ensure the accuracy and consistency of data linkage and analysis.

Selected relevant data columns that were vital for linkage and analysis were identified. Column names were standardized and merged with similar variables into a unified format. Inconsistencies in categorical data, such as sex labels, county names, bodies of water, and date formats were identified and standardized.

Free-text narratives and synopses were mapped to predefined categories, with missing values filled using keywords from each incident. Text parsing, and validation checks were applied to extract additional key information, resolve errors, address standardization issues, and to improve overall data quality. Duplicate entries were resolved by using a key variable for verification and by cross-referencing the narratives, retaining only one entry per incident. The final cleaned dataset was organized for integration with other data, ensuring its reliability for statistical analysis and predictive modeling.

### *Probabilistic linkage*

Since data were unstandardized, inconsistent, and often lacked personal identifiers, probabilistic linkage was used to link data across datasets. The method assigns weights to various matching criteria, such as date of drowning, sex, and age, and calculates a similarity score. The similarity scores estimate the likelihood that two records refer to the same event even if some details do not exactly match. Records with high similarity scores are considered matched.

The most common variables across all datasets that were chosen for linking included date of drowning/date of death, county of occurrence, sex, age, and body of water. This was supplemented with information from addresses and event narratives. Weights were assigned to each variable based on their importance for matching records. Each of the 5 key variables was assigned a weight of 0.2. A similarity threshold was set to optimize linkage and minimize the false negative cases. When gender information was missing, a probability of 0.5 was assigned to both male and female values to improve match accuracy.

Linkage between datasets began sequentially with datasets with most complete data followed by datasets with less complete data. Probabilistic linking was performed using Python 3 and open-source libraries for ease of comparison of string variables and to match records.

Parallel processing accelerated the matching by comparing multiple records simultaneously, making the linkage process efficient even on large datasets. For each pair of records, a weighted similarity score was calculated by applying predefined functions to selected variable pairs, assessing similarity based on criteria like string matching, numerical proximity, and categorical alignment. If the similarity score met the threshold, the records were considered linked. The results were saved and sorted, allowing for further review. The process reduced computation time considerably. To identify and manage duplicate entries and multiple matches, a combination of automated and manual validation was used to ensure accuracy.

To further optimize data linkage, early filtering was applied during processing. Once the similarity scores were computed, record pairs with a score $<0.3$ were removed to ensure that resources focused on more promising matches. Row access was streamlined, and results aggregated before conversion to a data frame, minimizing redundant updates and enhancing performance for large datasets.

Exponential decay and logarithmic functions were incorporated to manage discrepancies in continuous data involving date of drowning and age, respectively. This enabled reliable matching between datasets by penalizing data discrepancies in a flexible manner even when exact matching was not possible. Exponential decay data transformation preserves high similarity for small date differences (in days) but penalizes larger differences between dates. For age, logarithmic transformation of data allows for similarity between datasets when there are small age differences (e.g., 1–2 years), but similarity gradually decreases with larger age differences. Sex and county similarity functions incorporated domain-specific rules, such as assigning partial similarity scores to missing values or partial matches, thus allowing the linkage process to accommodate data irregularities.

In summary, beginning with the most complete data, using key variables and similarity functions, along with utilizing optimization methods (parallel processing, early filtering, and robust similarity measures), the datasets were combined in a structured and efficient manner. This approach yielded relevant connections between records while preserving critical information.

### *Post-Linkage verification*

The verification process involved two steps for each linked dataset pair: (1) visualizing true positive and false positive cases across all similarity thresholds, and (2) selecting and validating the optimal threshold. Figure 1a, b illustrates this process using the linked medical examiner and EMS datasets.

First, linked cases were sorted and evaluated across a range of similarity thresholds, and the distribution of true positive cases (TP) and false positive cases (FP) at each threshold was visualized using stacked bar plots (Fig. 1a). These plots illustrate how the proportion of false positive cases increase as the similarity threshold is reduced.

In the second step, the optimal similarity threshold was selected and validated by Receiver Operating Characteristic (ROC) analysis. The Youden's J statistic provided a measure of diagnostic accuracy at the optimal threshold (Fig. 1b). For each candidate threshold, the number of true positive cases were divided by the total number of actual positive cases to compute the True Positive Rate (TPR), and the number of false positive cases was divided by the total actual positive cases to calculate the False Positive Rate (FPR). Youden's J statistic, defined as the difference between the TPR and FPR, was computed for each similarity threshold. The threshold that corresponded to the maximum J statistic was selected as the optimal cut point. This threshold represented the cut point at which the linkage procedure achieved the best balance between correctly matched cases and incorrectly matched cases.

The procedure was performed separately for each pair of linked datasets, each of which had its unique optimum similarity threshold. The final selected similarity threshold varied based on differences in data quality, completeness, and record structure of linked datasets.

Some non-matched cases were not present in the initial dataset, even after manually verifying for possible mismatch using expert guidelines. These cases were retained as unique cases and added to the linked database. By integrating these non-matched cases, dataset alignment was improved both horizontally, through enhanced data linkage across datasets, and vertically, by incorporating unique cases that were not present in the initial or base dataset (Fig. 2). Linkage metrics were assessed by sensitivity (Recall), positive predictive value (Precision) and F1 score (Harmonic mean of precision and recall) (Supplemental Table 2).

### *Spatial validation*

Additional validation was conducted by geocoding each location of drowning. For those records with an address (i.e., drownings in pools, bathtubs, and hot tubs), geocoding was done in two steps. First, the address was located using standard address matching [20] for a rough approximation of the location. Second, for pool drownings, a database of swimming pools that the project had compiled was used to identify the parcel in which the drowning occurred. The pool location was very accurate since it was geographically mapped to the center of the land parcel. For non-pool drownings with addresses (e.g., bathtubs), a parcel GIS layer from each county was used to identify the exact parcel in which the drowning occurred. For these records, geographic accuracy is very high.

For those records that did not have an address, especially those in natural water, the best approximation to the location was selected based on the narrative. Some of the records (e.g., USCG) provide geographical coordinates for the drowning. Other records had the nearest street location for beach drownings; an estimate of where on the water's edge the drowning occurred was made. Clearly, estimation error for these cases is greater than for drowning locations with addresses.

The spatial geocoding was then used to identify drownings that had occurred outside our region and to provide a more accurate description of the body of water in which the drowning occurred.

### *Special considerations*

Linkage with syndromic surveillance and NEMSIS databases was not possible due to data being deidentified and unavailable drowning locations. Regulations prohibit linkage of hospital discharge data with other datasets [21, 22].

### *Data security*

After linkage, cases were assigned a unique number. Data files were transferred electronically using file encryption between research staff. Once linked, personal identifiers were stripped and a de-identified version of the dataset was prepared and used for analysis. To avoid re-identification of cases, results are reported by sex, age group, and race/ethnicity. Cell counts under ten are suppressed. Artificial intelligence testing was performed in a closed environment. Studying the epidemiology of fatal drowning from media reports using artificial intelligence was not considered human subjects research by the Baylor College of Medicine IRB (H-56709).

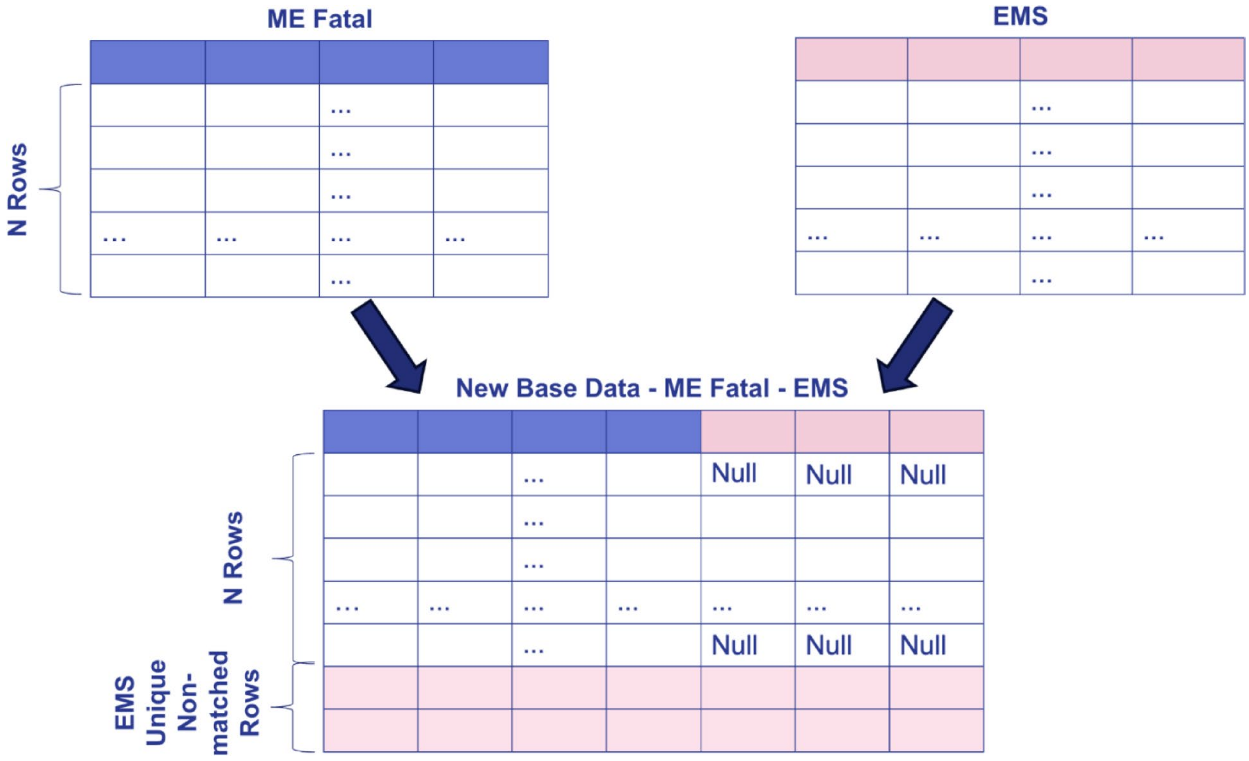

**Fig. 2** Probabilistic linkage of datasets and incorporation of new data into final database

### *Patient and public involvement*

Patient and public input were not sought since data were retrospective. Research questions were developed based on existing gaps in drowning surveillance [7, 8]. Study results will be disseminated in separate publications.

### *Statistical analysis and data verification*

No collinearity existed between the five key linking variables since the Variance Inflation Factor (VIF) was ≤ 5, the tolerance was > 0.8, and the condition index < 15. Supplemental Table 3 compares key study variables between linked data, National Vital Statistics System (NVSS) and Department of State Health Services, Texas (DSHS) data. The deceased's county of residence (2016–2017) [23] (as county of drowning was unavailable) and county of drowning (2018–2022) [24] were aggregated for NVSS data. Linked and DSHS data were based on county of drowning. Injury burden, age group, sex, county of drowning and body of water in the linked dataset were compared with NVSS and DSHS data. Race/ethnicity could not be compared between datasets because of different categories. For comparing linked and NVSS/DSHS datasets, hot tub and bathtub drownings were grouped together in the linked dataset since no category for hot tubs exists in NVSS or DSHS datasets; watercraft- and water-transport-related drownings were combined with natural water drownings in NVSS and DSHS datasets. The Pearson Chi-Square test was utilized when there were comparable values for all three datasets. A p-value < 0.05 was considered statistically significant. Statistical analyses were conducted using STATA version 15 (StataCorp, LP College Station, TX).

The Lincoln-Petersen estimator [25] was used to determine under-ascertainment of fatal drowning by individual dataset. The under-ascertainment rate varied from 0.02 to 0.46 among datasets.

## Results

From eight linked data sets, 790 drowning fatalities were identified in metropolitan Houston between 2016 and 2022. The linkage metrics for the datasets were recall (0.88-1.00), precision (0.91-1.00), and F1-score (0.91-1.00) (Table 1). Missing data among the 5 key linking variables was minimal, except for TPWL and USCG data which did not provide information on race and ethnicity (Supplemental Table 4).

There were no significant differences in injury burden by age group, sex, county of drowning and body of water between the linked data set and NVSS data (Table 2). Harris and Galveston counties accounted for the highest number of drowning fatalities (Table 2). Out of 790 records, 769 (97%) were geocoded (Supplemental Fig. 2).

The median age was 40 years (IQR: 18.5,60); 71% were males. Children aged 0–17 years constituted 24% of drowning fatalities. By race, White persons and Black persons constituted 44% and 19% of drowning patients, respectively. Persons with Hispanic ethnicity comprised 21% of drowning patients. Drownings occurred in swimming pools (27%), bathtubs (19%) natural water (27%), flood control structures (20%), and during flooding events (6%). Adults commonly drowned in natural water, flood control structures, bathtubs, and during flooding events whereas most toddlers drowned in swimming pools and hot tubs (Supplemental Table 5). The percentage of missing cases for drowning risk and protective factors was substantial and varied from 13.2% to 53.5% (Supplemental Table 6).

## Discussion

Overall, there was reasonable recall and precision in data linkage between datasets. The injury burden, demographics, and location of fatal drownings in the region closely matched NVSS and DSHS data. This study provides the methodology for data procurement, data quality, and probabilistic linkage of multiple datasets to accurately determine the epidemiology of fatal drowning in a large

**Table 1** Data linkage metrics

| Data Source | # Cases | True Positive | False Positive | False Negative | Recall | Precision | F1 |
|---|---|---|---|---|---|---|---|
| Medical Examiner (4 datasets) | 678 | - | - | - | - | - | - |
| EMS | 19 | 11 | 1 | 0 | 1.00 | 0.92 | 0.96 |
| Hospital | 62 | 40 | 4 | 1 | 0.98 | 0.91 | 0.94 |
| Media Reports | 142 | 74 | 5 | 10 | 0.88 | 0.94 | 0.91 |
| US Coast Guard | 35 | 29 | 1 | 0 | 1.00 | 0.97 | 0.98 |
| Texas Parks & Wildlife | 53 | 31 | 0 | 0 | 1.00 | 1.00 | 1.00 |
| National Oceanic & Atmospheric Data | 21 | 20 | 1 | 0 | 1.00 | 0.95 | 0.98 |
| Consumer Product Safety Commission | 874 | 207 | 9 | 0 | 1.00 | 0.96 | 0.98 |
| Hurricane data | 41 | 37 | 0 | 1 | 0.97 | 1.00 | 0.98 |

True negatives are excluded because they do not inform match probabilities or contribute to linkage decisions

F1: Harmonic mean of recall and precision

**Table 2** Comparison of drowning fatalities by dataset for the Houston metropolitan region (2016–2022)

| **Variable** | **Linked Data**<br>**(*n* = 790)** | **NVSS**[a]<br>**(*n* = 791)** | **DSHS**<br>**(*n* = 748)** | ***p*-value**[b] |
|---|---|---|---|---|
| Age | 189 (23.9%) | 200 (25.3%) | 180 (24.1%) | 0.83 |
| Children (0–17 yrs) | 591 (74.8%) | 591(74.7%) | 568 (75.9%) | |
| Adults (≥ 18 yrs) | 10 (1.3%) | 0 (0%) | 0 (0%) | |
| Unknown | | | | |
| Sex | 564 (71.4%) | 589 (74.4%) | 550 (73.5%) | 0.91 |
| Male | 201 (25.4%) | 202 (25.5%) | 198 (26.5%) | |
| Female | 25 (3.2%) | 0 (0%) | 0 (0%) | |
| Unknown | | | | |
| Race/Ethnicity[c] | 351 (44.4%) | 566 | 347 (46.4%) | - |
| White | 153 (19.4%) | 169 (21.4%) | 162 (21.7%) | |
| African-American | 169 (21.4%) | 192 (24.2%) | 188 (25.1%) | |
| Hispanic | 37 (4.7%) | 49 (6.2%) | 51 (6.8%) | |
| Asian/Other | 80 (10.1%) | - | - | |
| Unknown | | | | |
| County | 478 (60.5%) | 464 (58.6%) | 399 (53.3%) | 0.13 |
| Harris | 132 (16.7%) | 113 (14.2%) | 130 (17.4%) | |
| Galveston | 75 (9.5%) | 83 (10.4%) | 73 (9.8%) | |
| Montgomery | 37 (4.7%) | 53 (6.7%) | 55 (7.4%) | |
| Brazoria | 40 (5.0%) | 45 (5.7%) | 48 (6.4%) | |
| Fort Bend | 15 (1.9%) | *** | *** | |
| Chambers | *** | *** | 16 (2.1%) | |
| Liberty | *** | *** | *** | |
| Waller | | | | |
| Body of Water | 211 (26.7%) | 212 (26.8%) | 208 (27.8%) | 0.06 |
| Swimming Pool | 150 (19.0%) | 156 (19.7%) | 161 (21.5%) | |
| Bath Tub[d] | 215 (27.2%) | 310 (39.2%) | 281 (37.5%) | |
| Natural water | - | - | - | |
| [*Lake*, *river*, *gulf*, *bay*] | 157 (19.9%) | - | - | |
| Flood control structures | - | - | - | |
| [*Bayou*, *canal*, *creek*, | - | - | - | |
| *ditch*, *pond*, *concrete* | - | - | - | |
| *waterway*, *reservoir*] | 2 (0.3%) | - | - | |
| Other–natural water | - | 45 (5.7%) | 36 (4.8%) | |
| Watercraft, Water-Transport (V90,92)[e] | | | | |
| Other-specified | - | 19 (2.4%) | 20 (2.7%) | |
| Unspecified | - | 49 (6.2%) | 42 (5.6%) | |
| Flooding events | 49 (6.2%) | - | - | |
| Other-constructed water[f] | 2 (0.2%) | - | - | |
| Unreported | 4 (0.5%) | - | - | |

NVSS: National Vital Statistics System; DSHS: Department of State Health Services-Texas

[a] NVSS data is by deceased's county of residence for 2016–2017 and county of drowning for 2018–2022

Linked data and DSHS data are by county of drowning (2016–2022)

[b] Unknown or non-comparative categories were excluded from the Pearson Chi-square analysis

[c] The variable race/ethnicity was not compared due to different classification among datasets

[d] Includes 14 cases of hot tub drowning in linked dataset

[e] Transport-related drowning (V90,92) combined with Natural water for NVSS and DSHS data when comparing with linked data

[f] Includes basin and hole

*** Counts < 10 are suppressed

metropolitan area. This process is integral to conducting drowning surveillance, monitoring the effectiveness of drowning countermeasures and for disaster planning at the regional level.

Challenges in data linkage can arise from inadequate number of datasets, missing and unstandardized data within datasets and technical issues during probabilistic linkage. Metropolitan Houston has a large population with diverse aquatic systems and water recreational activities. In this study, successful linkage of multiple datasets which represented all bodies of water, different types of urban environments and encounters across the continuum of care helped mitigate the effect of missing data within each dataset.

Data quality was another challenge. Data were nonstandardized, incomplete, and missing across datasets. The five key variables (injury/death date, age, sex, body of water, county of drowning) were selected for linkage because they had the least amount of missing information Highest linkage rates are observed for demographic variables and temporal data [26]. Linkage was further augmented by extracting contextual information from case narratives using artificial intelligence. ME data were most comprehensive for Harris County, but less so for the other counties with MEs. USCG, TPWL and media report data lacked race and ethnicity information.

Finally, there were technical issues with probabilistic linkage. The success of data linkage is measured by reducing the number of false matches and missed matches [27]. False matches can occur when there are multiple drowning cases in a single incident. False matches lead to weaker associations between variables present in different datasets. Missed matches arise when data reports about the same individual are not linked. This happens when data lack personal identifiers. Excluding missed matches from analysis reduces the sample size and statistical power with under-ascertainment of exposures or outcomes. Using the five key linking variables minimized missed matches. The selection and validation of the optimum similarity thresholds for probabilistic linkage and manual verification of case narratives minimized the number of false positive matches and unmatched records. Furthermore, incorporating exponential decay and logarithmic functions to manage discrepancies in continuous data involving date of drowning and age respectively, enabled reliable matching even when exact matching was not possible.

Probabilistic linkage is superior to current surveillance methods that utilize death certificate data for fatal drowning surveillance [23, 24]. While providing comparable injury burden and demographics of patients of fatal drowning as conventional surveillance methods, linked data also provide vital information on drowning location, drowning circumstances and associated risk and protective factors. This information can identify high-risk subpopulations and drowning locations which, in turn, can inform countermeasures that address drowning.

There were small differences in the distribution of drowned patients by age group, sex, county of drowning and body of water between the linked data and NVSS and DSHS datasets. Comparison of race/ethnicity between linked and NVSS data was not possible. For NVSS data, bridged race [23] (2016–2017) was combined with single race [24] (2018–2022), thus precluding comparison with linked data which used a different combination of variables for race and ethnicity. For death certificate information, decedent race and ethnicity are reported by the funeral director based on next-of-kin or observation [23, 24]. Decedents with multiple race are imputed to a single race according to their combination of race, ethnicity, sex, and age as indicated on the death certificate [24]. Data on race and ethnicity were less complete in linked data compared to death certificate data.

With linked data, the date of injury and date of death may not coincide. This occurs in patients who die after a period of hospitalization. The date of injury is unavailable with NVSS data. NVSS data are also less robust than linked data in describing the drowning location and body of water. Geographic mapping of drowning events in linked data can accurately identify the approximate drowning location. Conversely, NVSS data which only uses three specific ICD-10 codes to describe the body of water (bathtub, swimming pool, and natural water) is less descriptive than linked data. Linked data provides information on drownings that occur in all bodies of water, including flood-related events. Some differences in drowning burden by county may occur because cases may be mistakenly assigned to counties that share natural water boundaries, such as the gulf coastline. An example is San Luis Pass, a coastal inlet between Brazoria and Galveston counties, which is a high-risk drowning location.

The study had limitations. It was performed in one metropolitan region and the results may not be generalizable to other metropolitan areas. However, the methodology of probabilistic linkage can be replicated elsewhere. Second, missing data were a problem. Race and ethnicity information were unavailable in TPWL and USCG datasets. Also, the exact geographic location of natural water drownings were difficult to obtain. Though data were reasonably complete for the five key linking variables, missing data were higher for drowning risk and protective factors. OpenAI's LLM was used to extract information from the case narratives of the probabilistically linked final dataset to mitigate missing data on risk and protective factors. Third, probabilistic linkage cannot identify every drowned patient. In-depth investigations were not possible in all instances (e.g., body not recovered prior to advanced decomposition) making it difficult to confirm injury intent [28]. Non-aquatic transport-related drownings may be missed in death certificate data [28]. Fourth, there was varying under-ascertainment of drowning fatalities among datasets. The causes were unique to each dataset. Not all hospitals and EMS systems participated in the study. Media reports were biased towards reporting more newsworthy items. TPWL and USCG only reported maritime and natural water drownings. CPSC data reported only on swimming pool and bathtub drowning and were based on a probability sample of cases. Conversely, hurricane and weather-related drowning were specific events and were reported by the medical examiners of the four largest counties. Probabilistic linkage of multiple datasets mitigated under-ascertainment of drowning fatalities. Finally, the relation of drowning risk and protective factors with at-risk populations were not studied being outside the scope of this paper.

The study has important public health implications. First, probabilistic linkage of multiple datasets can accurately describe the epidemiology of fatal drowning regionally and used for drowning surveillance. To scale linkage methodology, it will be necessary to employ data science strategies such as improving the availability of timely data, standardizing drowning variables, process documentation, improving analytic tools and expanding the capacity of state and regional entities in data science methodologies. The latter can be done by addressing funding and workforce constraints, fulfilling statutory requirements for obtaining and reporting data, and disseminating linked data for application in injury prevention [9, 29]. Second, high risk drowning locations and at-risk populations identified by spatial analysis can be a focus for drowning countermeasures and disaster planning. Third, the robustness of data on fatal drowning can be improved by using standardized data forms, such as those developed by the National Center for Fatality Review and Prevention during death scene investigations of fatal drowning [30]. Adoption of this form nationwide would improve the consistency of reporting.

## Conclusions

Probabilistic data linkage can accurately determine the epidemiology of unintentional fatal drowning at a metropolitan level. It is superior to surveillance methods that use death certificate data.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40621-025-00623-8.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Supplementary Material 7

Supplementary Material 8

**Acknowledgements**

The authors thank the research staff, City of Houston Health Department, Houston and Harris County EMS, Offices of the Sheriff, County Medical Examiner, Police departments and County Tax Offices of the Houston-Galveston metropolitan region, US Coast Guard and Texas Parks and Wildlife for providing data that was used in the study.

**Author contributions**

R.P.S, N.L and S.Z conceptualized the study and were responsible for the study methodology. S.Z implemented and optimized probabilistic linkage code, validated results, and improved methodology using metric-based evaluation. R.P.S obtained funding for the study and was responsible for project administration; R.P.S and N.L were responsible for study supervision. E.A.C was responsible for software and formal analysis. N.L and S.Z was responsible for software, formal analysis, and validation. All authors were responsible for data curation, writing the original draft and reviewing and editing the manuscript.

**Funding**

This study was funded by a grant from the CDC: U01 CE003504-01. The funding body did not have any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**Data availability**

Data will be available upon conclusion of the grant and study period on October 1, 2026 and upon request and data use agreement.

## Declarations

**Ethics approval**

This study received institutional review board approval from the Baylor College of Medicine (H-51536), University of Texas at Houston (HSC-MS-22-0427) and Department of State Health Services Texas (IRB# 24 – 020).

**Consent for publication**

Not applicable.

**Disclaimer**

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

**Competing interests**

The authors declare no competing interests.

Received: 25 June 2025 / Accepted: 19 September 2025
Published online: 14 October 2025

**References**

1. Centers for Disease Control and Prevention, National Center for Health Statistics. National Vital Statistics System, Provisional Mortality on CDC WONDER Online Database. Accessed on https://wonder.cdc.gov/ Accessed on July 24, 2024.
2. Centers for Disease Control and Prevention. Drowning Prevention. Available at: https://www.cdc.gov/drowning/prevention/index.html Accessed on July 24, 2024.
3. World Health Organization. Fact sheet Drowning 25. July, 2023 Available at: https://www.who.int/news-room/fact-sheets/detail/drowning Accessed on July 24, 2024.
4. Web-based Injury Statistics Query and Reporting System (WISQARS). available at: https://wisqars-viz.cdc.gov/create-visualizations/ Accessed on August 24, 2025.
5. Centers for Disease Control and Prevention, National Center for Health Statistics. National Vital Statistics System, Provisional Mortality on CDC WONDER Online Database. Data are from the final Multiple Cause of Death Files, 2018–2023, and from provisional data for years 2024 and later, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/mcd-icd10-provisional.html on Aug 24, 2025.
6. Clemens T, Moreland B, Lee R. Persistent racial/ethnic disparities in fatal unintentional drowning rates among persons aged ≤ 29 years - United States, 1999–2019. MMWR Morb Mortal Wkly Rep. 2021;70(24):869–74. https://doi.org/10.15585/mmwr.mm7024a1.
7. Drowning Prevention Research Priorities. NCIPC Board of Scientific Counselors Meeting August 23. 2022. Available at: https://www.cdc.gov/injury/pdfs/

bsc/drowning-prevention-research-priorities-bsc-presentation_cleared-a11y.pdf Accessed on May 17, 2024.

8. US National Water Safety Action Plan Available at. https://ndpa.org/safety-action-plan/ Accessed on August 22, 2025.
9. Milani J, Kindelberger J, Bergen G, Novicki E, Burch C, Ho S, West B. August). Assessment of characteristics of state data linkage systems. (Report No. DOTHS 812 180). Washington, DC: National Highway Traffic Safety Administration, and Atlanta: Centers for Disease Control and Prevention; 2015.
10. Cook L, Thomas A, Olson C, Funai T, Simmons T. July). Crash outcome data evaluation system (CODES): an examination of methodologies and multistate traffic safety applications. (Report No. DOT HS 812 179). Washington, DC: US Department of Transportation, National Highway Traffic Safety Administration; 2015.
11. US Department of Health and Human Services. Centers for Disease Control and Prevention (CDC). National Center for Injury Prevention and Control. Linking Information for Non-fatal Crash Surveillance. Available at LINCS. Linking Information for Non-fatal Crash Surveillance: A guide for integrating motor vehicle crash data to help keep Americans safe on the road https://www.cdc.gov/transportationsafety/linkage/Linking-Information-Nonfatal-CrashSurveillance.html. Accessed 5/20/2021.
12. Cain CM, Oluyomi AO, Levine N, Pompeii L, Rosales O, Naik-Mathuria B. Socioeconomic disparities based on shooting intent in pediatric firearm injury. J Trauma Acute Care Surg. 2024;97(3):440–4. Epub 2024 Jan 29.
13. Peden AE, Sarrami P, Dinh M, Lassen C, Hall B, Alkhouri H, et al. Description and prediction of outcome of drowning patients in New South Wales, Australia: protocol for a data linkage study. BMJ Open. 2021;11(1):e042489. https://doi.org/10.1136/bmjopen-2020-042489.
14. Shenoi RP, Koerner CE, Cruz AT, et al. Factors associated with poor outcome in childhood swimming pool submersions. Pediatr Emerg Care. 2016;32(10):669–74. https://doi.org/10.1097/PEC.0000000000000678.
15. US Census. Available at: https://www.census.gov/data.html Accessed on June 18, 2023.
16. Idris AH, Bierens JJLM, Perkins GD, Wenzel V, Nadkarni V, Morley P, Warner DS, Topjian A, Venema AM, Branche CM, Szpilman D, Morizot-Leite L, Nitta M, Løfgren B, Webber J, Gräsner JT, Beerman SB, Youn CS, Jost U, Quan L, Dezfulian C, Handley AJ, Hazinski MF. 2015 revised Utstein-style recommended guidelines for uniform reporting of data from drowning-related resuscitation: an ILCOR advisory statement. Resuscitation. 2017;118:147–58. https://doi.org/10.1016/j.resuscitation.2017.05.028. Epub 2017 Jul 17.
17. Denny SA, Quan L, Gilchrist J, McCallin T, Shenoi R, Yusuf S, Hoffman B, Weiss J, COUNCIL, ON INJURY, VIOLENCE, AND POISON PREVENTION. Prev Drowning Pediatr. 2019;143(5):e20190850. https://doi.org/10.1542/peds.2019-0850. Epub 2019 Mar 15.
18. Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J. Chart reviews in emergency medicine research: where are the methods? Ann Emerg Med. 1996;27:305–8.
19. Kaji AH, Schriger D, Green S. Looking through the retrospectoscope: reducing bias in emergency medicine chart review studies. Ann Emerg Med. 2014;64:292–8.
20. Levine N, Kim KE. The spatial location of motor vehicle accidents: A methodology for geocoding intersections, Computers, Environment, and Urban Systems. 1999, 22(6):557–576.
21. Texas Health Care Information Collection. Texas Health Care Information Collection Texas DSHS. (n.d.). https://www.dshs.texas.gov/texas-health-care-information-collection
22. Texas Inpatient Public Use Data File (PUDF). Available at: https://www.dshs.texas.gov/sites/default/files/thcic/hospitals/InpatientDataDictionary4Q2023.pdf
23. Centers for Disease Control and Prevention, National Center for Health Statistics. National Vital Statistics, System. Mortality 1999–2020 on CDC WONDER Online Database, released in 2021. Data are from the Multiple Cause of Death Files, 1999–2020, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/ucd-icd10.html on May 23, 2025.
24. Centers for Disease Control and Prevention, National Center for Health Statistics. National Vital Statistics System, Provisional Mortality on CDC WONDER Online Database. Data are from the final Multiple Cause of Death Files, 2018–2023, and from provisional data for years 2024 and later, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/mcd-icd10-provisional.html on May 23, 2025.
25. Chao A, Tsay PK, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. Stat Med. 2001;20(20):3123–57. https://doi.org/10.1002/sim.996.
26. Karimi S, Hosseinzadeh A, Kluger R, Wang T, Souleyrette R, Harding E. Accid Anal Prev. 2024;197:107461. https://doi.org/10.1016/j.aap.2024.107461. Epub 2024 Jan 9. A systematic review and meta-analysis of data linkage between motor vehicle crash and hospital-based datasets.
27. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 2017;46(5):1699–710. https://doi.org/10.1093/ije/dyx177.
28. Peden AE, Franklin RC, Mahony AJ, Scarr J, Barnsley PD. Using a retrospective cross-sectional study to analyse unintentional fatal drowning in Australia: ICD-10 coding-based methodologies verses actual deaths. BMJ Open. 2017;7(12):e019407. https://doi.org/10.1136/bmjopen-2017-019407.
29. Ballesteros MF, Sumner SA, Law R, Wolkin A, Jones C. Advancing injury and violence prevention through data science. J Saf Res. 2020;73:189–93. https://doi.org/10.1016/j.jsr.2020.02.018.
30. Water-Related Death Scene Investigation (DSI) Protocol) National Center for Fatality Review and Prevention. Available at: https://ncfrp.org/wp-content/uploads/Water-Related-Death-Scene-Investigation-Protocol.pdf Accessed on August 24, 2025.

## Publisher's note