OPEN

# Record Linkage Approaches Using Prescription Drug Monitoring Program and Mortality Data for Public Health Analyses and Epidemiologic Studies

*Sarah Nechuta, Sutapa Mukhopadhyay, Shanthi Krishnaswami, Molly Golladay, and Melissa McPheeters*

**Background:** The use of Prescription Drug Monitoring Program (PDMP) data has greatly increased in recent years as these data have accumulated as part of the response to the opioid epidemic in the United States. We evaluated the accuracy of record linkage approaches using the Controlled Substance Monitoring Database (Tennessee's [TN] PDMP, 2012–2016) and mortality data on all drug overdose decedents in Tennessee (2013–2016).

**Methods:** We compared total, missed, and false positive (FP) matches (with manual verification of all FPs) across approaches that included a variety of data cleaning and matching methods (probabilistic/fuzzy vs. deterministic) for patient and death linkages, and prescription history. We evaluated the influence of linkage approaches on key prescription measures used in public health analyses. We evaluated characteristics (e.g., age, education, sex) of missed matches and incorrect matches to consider potential bias.

**Results:** The most accurate probabilistic/fuzzy matching approach identified 4,714 overdose deaths (vs. the deterministic approach, n = 4,572), with a low FP linkage error (<1%) and high correct match proportion (95% vs. 92% and ~90% for probabilistic approaches not using comprehensive data cleaning). Estimation of all prescription measures improved (vs. deterministic approach). For example, frequency (%) of decedents filling an oxycodone prescription in the last 60 days (n = 1,371 [32%] vs. n = 1,443 [33%]). Missed overdose decedents were more likely to be younger, male, nonwhite, and of higher education.

**Conclusion:** Implications of study findings include underreporting, prescribing and outcome misclassification, and reduced generalizability to population risk groups, information of importance to epidemiologists and researchers using PDMP data.

**Keywords:** Record linkage; Data linkage methods; Epidemiologic methods; Bias; Data accuracy; Linkage error; Prescription drug monitoring programs

(*Epidemiology* 2020;31: 22–31)

In the late 1990s and 2000s, use of prescription opioids for acute and chronic pain, and associated morbidity and mortality, increased substantially in the United States, resulting in the need for regulation and policy to reduce opioid misuse and abuse.[1,2] Prescription Drug Monitoring Programs (PDMPs), now operational in all 50 states and one United States territory,[3] support the monitoring of prescribing practices and potential misuse of controlled substances. PDMPs provide information to health care professionals about a patient's prescription history to help identify potential misuse, abuse, and inappropriate prescribing.[3] Ideally, this information can be used to guide appropriate care and implement risk mitigation strategies for substance use disorder and overdose.[4–7]

Use of PDMP data for public health surveillance and epidemiologic studies has increased in recent years with the implementation of PDMPs through the United States, including cohort studies of linked PDMP and health outcome data.[8–14] Methods for data/record linkage (including de-duplication [matching individuals within the same data source] and linkage of individuals between data sources) [see Dusetzina et al.[15] and Sayers et al.[16] for a review of terms and concepts in linkage methodology]) can influence complete patient identification and medical history.[17–20] Accuracy of data/record linkage (hereafter referred to as record linkage) is particularly important for cohort studies and can influence estimation of incidence and mortality rates, and effect estimates due to misclassification of covariates, exposures, and outcomes, as well as generalizability of results if certain groups are excluded (e.g., individuals of lower socioeconomic status).[21–25]

Submitted March 4, 2019; accepted September 25, 2019.

From the Tennessee Department of Health, Office of Informatics and Analytics, 710 James Robertson Parkway, Nashville, TN.

Supported by the Centers for Disease Control & Prevention [*Prescription Drug Overdose* Prevention for States Program (5 NU17CE002731-02-00) to the Tennessee Department of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

The authors report no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

The data are not publicly available as the Controlled Substance Monitoring Database is not available for external data use at this time based on the current laws governing use of this database. Code used for this study (as feasible) and supplemental methodology information are available in the Supplemental Digital Content.

Correspondence: Sarah Nechuta, Department of Public Health, Grand Valley State University, 500 Lafayette Ave Northeast, Grand Rapids, MI 49503. E-mail: nechutas@gvsu.edu.



ISSN: 1044-3983/20/3101-0022

DOI: 10.1097/EDE.0000000000001110

The accurate identification of patient entities, and thereby patient prescription history, is a key challenge with PDMP data, which include millions of patient records with non-standardized data entry for identifying fields, such as name and address. We developed comprehensive strategies for record linkage using Tennessee's (TN) PDMP data (the Controlled Substance Monitoring Database [CSMD]) and vital statistics mortality data for all drug overdose decedents from TN's death statistical files. These methodologies and lessons learned can be helpful to public health epidemiologists and researchers using PDMP data. Our first objective was to evaluate the accuracy of several record linkage approaches using varying methods in data cleaning, standardization, and linkage (e.g., deterministic vs. probabilistic/fuzzy) and determine the most accurate approach for use in epidemiologic studies using linked PDMP data. Our second objective was to quantify the influence of the linkage approaches on the estimation of frequently reported opioid and benzodiazepine prescribing measures. Our third objective was to describe characteristics of matched, unmatched, and incorrectly matched records to evaluate the potential for measurement error and selection bias in analyses using PDMP and mortality data.

# METHODS

## Study Design and Data Sources

Figure 1 provides an overview of study design, population, and data sources.

### Controlled Substance Monitoring Database (2012–2016)

In accordance with the Controlled Substance Monitoring Act of 2002, the Tennessee Department of Health established the CSMD to monitor the dispensing of Schedule II–V controlled substances on December 1, 2006, with over 18 million prescriptions reported each year since 2011.[26] Data (eTable 1; http://links.lww.com/EDE/B594, provides the raw primary fields used for the present study) are collected using the American Society for Automation in Pharmacy specifications (https://www.asapnet.org). Prescriptions are reported by dispensers, which are largely pharmacies (although some veterinarians are dispensers). In the CSMD, there is no unique patient identifier required to be collected consistently such as social security number (SSN). Potentially identifying fields collected consistently for use in record linkage in the patient file of >13 million records included first name, middle name, last name, date of birth (DOB), and address.

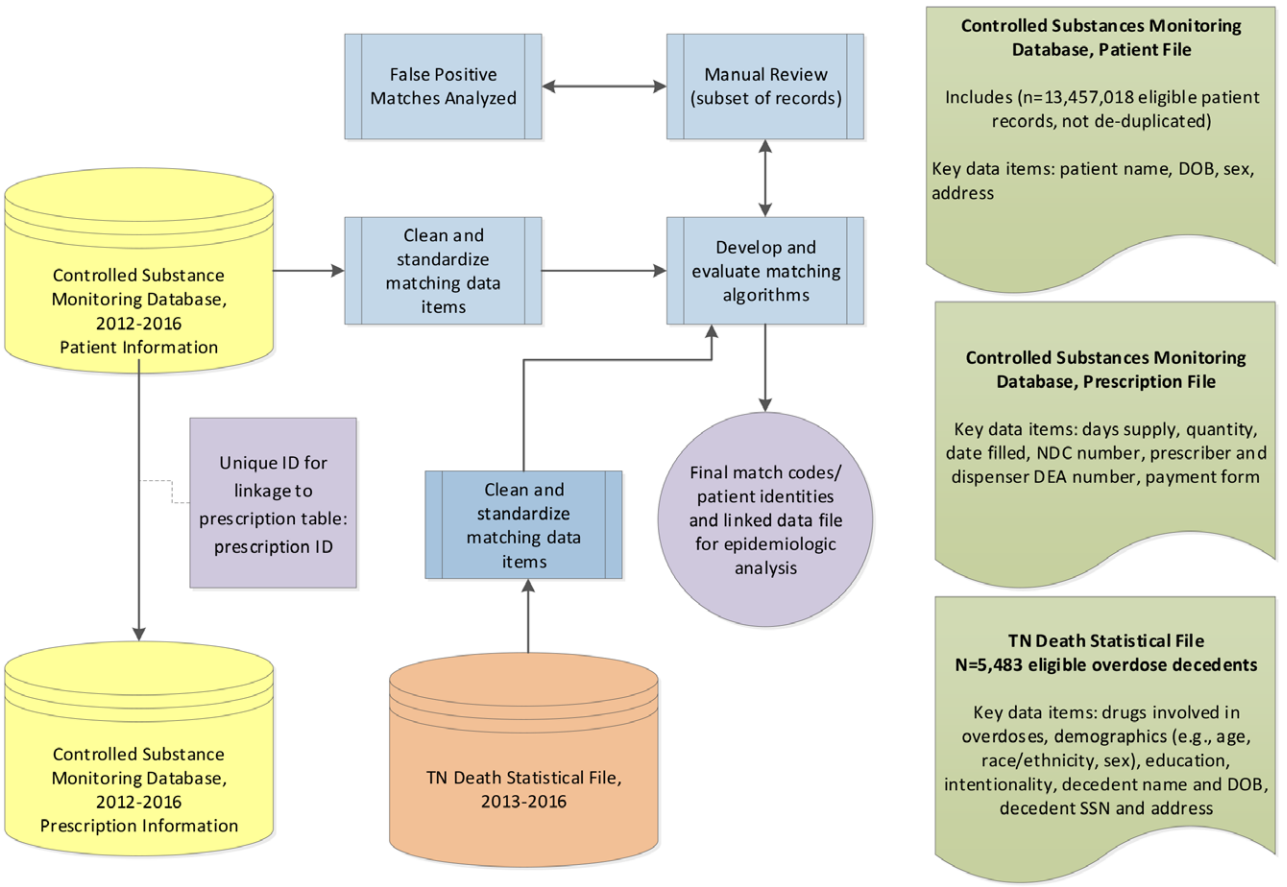


**FIGURE 1.** Study Data Sources, Study Population, and Entity Resolution/Record Linkage Methods

### Death Statistical Files (2013–2016)

TN's death certificates provide causes, place, and manner of death.[27] Each death certificate is classified with an underlying cause of death using the International Classification of Diseases, 10th Revision (ICD-10) codes and up to 20 additional multiple cause of death ICD-10 codes by the National Center for Health Statistics. Additional data collected include name, address, DOB, date of death, and sociodemographic information (e.g., sex, race, ethnicity, education, and marital status). Eligible decedents included TN residents aged ≥18 years with an underlying cause of death due to all drug overdose during 2013–2016 (n = 5,483). ICD-10 codes for all drug overdoses included X40–X44 (unintentional drug poisoning codes); X60–X64 (intentional drug poisoning codes); X85 (homicide/assault drug poisoning codes); and Y10–Y14 (undetermined intent drug poisoning codes).

## Record Linkage Methodology

### Overview

The first approach used comprehensive name cleaning with both standard[15] and data-specific programming techniques using Statistical Analysis Software (SAS) and a multi-step deterministic matching approach using Structured Query Language (SQL), used in our previous case-only study of overdose decedents.[28] The second through fifth approaches utilized probabilistic/fuzzy matching algorithms with varying sensitivity codes using SAS Data Management Studio Software.[29] We also tested our in-house data cleaning techniques developed in SAS vs. the SAS Data Management Studio standardization approach using raw data.

### Cleaning and Standardizing Matching Data Items

Name-cleaning strategies for the CSMD were described in detail previously[30] (with additional methodology and code available in the eAppendix 2; http://links.lww.com/EDE/B594). Briefly, extra spaces and non-alphanumeric characters, and non-name text were removed, and prefixes/suffixes were placed in separate fields. Similar approaches were used for cleaning and parsing name fields in the death files; however, name cleaning was more extensive in the CSMD largely due to lack of standardization at data entry and the large size of the data (>13 million patient records). Address was not used as a primary patient identification variable in the CSMD (due to potential changes in address across prescription records), but was used to identify potential false positive (FP) matches and to confirm match status via manual review, described below. Address fields were geocoded using ArcGIS, version 10.6 (ESRI, Redlands, CA), with a minimum match score of 85 and a spelling sensitivity of 80. Geocoding was successful for 94% of records in the death file and 91% of records in the CSMD patient file. For addresses that could not be geocoded, manual review was used to determine if the address was a match between records when assessing potential FP matches.

### Deterministic Data Matching (Approach 1)

The first approach used the comprehensive name cleaning protocol implemented in SAS and deterministic matching in SQL on exact DOB, any last name, and any first name. The matched set (regardless of approach) could have one death record matched to one patient record or multiple patient records, depending on the number of matched patient records and associated prescription histories in the CSMD for the decedent.

### Fuzzy/Probabilistic Data Matching (Approaches 2–5)

*Approach 2* used the same comprehensive name cleaning protocol, with probabilistic/fuzzy matching sensitivity codes of 85 for names and 95 for DOB in the SAS Data Management Studio software. It is worth noting that the SAS Data Management Studio-derived sensitivity codes are not the sensitivity (and corresponding specificity) metrics commonly used in epidemiology.[15] Sensitivity codes from SAS Data Management Studio are based on a combination of standardization/fuzzy matching techniques (e.g., regular expression processing, phonetics, transformations) using natural language processing to create a threshold value that ranges from 50 to 95 (in five-level increments).[29] *Approach 3* followed the same methods as approach 2, but varied on SAS Data Management Studio-derived sensitivity values (used 85 for both names and DOB). *We conducted Approaches 4 and 5* using SAS Data Management Studio software for standardization and matching, with unclean names and DOB entered directly. Standardization in SAS Data Management Studio uses the SAS Data Quality Knowledge Base, a repository of rules, and reference data.[29] The difference between approach 4 and approach 5 was in the sensitivity values. Specifically, approach 4 used sensitivity values of 85 for names and 95 for DOB; while approach 5 used the lower sensitivity value of 85 for both names and DOB.

## Evaluation of the Accuracy of Matching Algorithms

Accuracy measures included number of matches, FPs (i.e., incorrect matches) and false negatives (i.e., missed matches) for both overdose deaths and prescriptions.[15,31] These were evaluated for patient to death matches, unique overdose decedents, and prescriptions to eligible decedents. Many duplicate patient records exist, due to name and address variations entered by pharmacy staff, and each of these is linked to one or more prescriptions filled.[32] eFigure 3(http://links.lww.com/EDE/B594) provides example patient records from the CSMD within a de-duplicated matched set to illustrate how data issues in available patient information may affect identification of patient entities and complete prescription history.

### FP Method for Identification of Matches for Manual Review (Death and CSMD)

Determining if a match was a true FP was not straightforward, due to lack of a unique identifier in the CSMD and

data quality issues with names (e.g., nicknames, misspellings) and address (e.g., missing data, wrong street number, change in address). Therefore, we developed an approach to identify potential FPs, which were then reviewed manually to classify as follows:(1) not an FP (able to confirm a correct match), (2) possible FP (i.e., unable to confirm true match status based on available information), and (3) true FP (confident an incorrect match). One author manually reviewed all potential FP matches for each approach, and a second author confirmed all potential and true FP matches for approaches 2–5.

We used differences in either full name or DOB to identify potential FPs for patient and death record matches. To reduce burden of manual review for FPs, we used a systematic stepwise approach considering name, DOB, original and geocoded address, and SSN (when available in both death and the CSMD [SSN is missing for close to 70% of records as it is not a required variable]), to reduce the pool of potential FPs for manual review. For the first matching approach (deterministic only), we also evaluated the usefulness of middle names in identifying FPs and confirming match status. For approaches 2–4, which incorporated probabilistic/fuzzy matching, we evaluated the primary data issues resulting in potential FP identification and summarized the frequency of these by approach.

### Statistical Analysis

We calculated descriptive statistics for continuous variables (e.g., median, and interquartile ranges [IQR]) for total, opioid, and benzodiazepine prescriptions, and opioid and benzodiazepine days' supply. Total morphine milligram equivalents for opioid analgesics were calculated for the last 60 and 180 days before overdose. Chi-square analyses were conducted for categorical variables, including number of prescribers or dispensers in the year before overdose (1,2,3,4, ≥5), any prescription use (any, opioid analgesic, buprenorphine for medication-assisted treatment, oxycodone, hydrocodone, and benzodiazepine use) in the last 60 days before overdose, active opioid or benzodiazepine prescription at overdose (where prescription end date overlapped date of death by at least one day), and cash payment for an opioid analgesic. Drug classifications used the Centers for Disease Control Drug Classification table for controlled substances.[33] We calculated the frequency of matches, missed matches, FPs, and the FP linkage error rate by linkage approach. We conducted data management and/or statistical analyses using SAS version 9.4 (Cary, NC), Microsoft SQL Server Management Studio Version 17 (Redmond, WA), and SAS Data Management Studio version 2.7 (Cary, NC). The primary site Institutional Review Board for human subjects research approved this study.

## RESULTS

Figure 1 provides an overview of study design and population. Table 1 displays number of deaths and prescriptions, and potential FP matches identified for manual review for each linkage approach. Approach 3 (comprehensive name cleaning/standardization and probabilistic/fuzzy matching) resulted in the highest number of matched overdose deaths (n = 4,714) (versus the deterministic approach (n = 4,572)) and prescriptions (n = 250,082), with 154 more overdose deaths and 17,632 more prescriptions included than the deterministic approach 1. Approach 5 had the highest number of potential FP matches identified that required manual review.

We conducted a comprehensive evaluation and analysis to identify possible and true FP patient (CSMD) and death (TN death statistical file) matches. For approach 1 (deterministic record linkage), we identified the initial potential FPs matches as those with discrepancies in full names between the patient and death record (largely due to situations where an individual had 2 last names or differences in middle name fields) (n = 14,906). After removing records confirmed as true matches based on comparison of address and/or SSN, a total of 8,671 potential FPs matches remained (Table 1). We evaluated the use of middle name to help identify FPs and confirm true match status. However, this matching variable was found to provide limited additional utility to determine match status, and was not used subsequently. After manual review, we identified 344 matched records that we considered possible FPs (could not completely confirm match status without additional identifying information) and 0 true FP matches, representing 161 unique overdose decedents and 3,134 prescriptions.

Figure 2 and Table 2 display results regarding the evaluation, analysis, and final classification of FP matches (i.e., patient [CSMD] and death [TN death statistical file] matches) after manual review for the linkage approaches that utilized fuzzy/probabilistic matching. The total number of FPs between a death and patient record based on name and/or DOB ranged from 1,500 for approach 2 to 3,935 for approach 5. Approach 2 resulted in the fewest and approach 5 the most potential FP matches for manual review (i.e., match status could not be confirmed using available address or SSN [Figure 2]).

As shown in Table 2, the most common data quality issues resulting in a potential FP match included missing and alternative spellings, nicknames, multiple last names (and where applicable, different DOB). For approaches 3 and 5, which used raw data without cleaning, the primary data quality issue was unclean text data issues (addressed in our other approaches that implemented comprehensive name cleaning). The true FP error rate among all CSMD patient and death record matches was <1.0%, regardless of approach. The correct match proportion ranged from 91% in approach 5 to 97% in approach 2.

Table 3 displays prescription characteristics for approach 1 compared with approach 3 (identified as the most accurate approach based on number of matches and false negative and FP linkage errors). Approach 1 underestimated all prescription measures, including number of prescriptions, days' supply for prescriptions, total morphine milligram equivalents, and frequency of prescriptions across

**TABLE 1.** Results from Deterministic and Probabilistic Record Linkage Approaches for All Drug Overdose Decedents in the TN Death Statistical Files (2013–2016) and CSMD (2012–2016)

| | Influence on Fatal Overdose Deaths | | | Influence on Prescription History | |
|---|---|---|---|---|---|
| **Linkage Approach** | **Matched Overdose Deaths** | **New Overdose Deaths Matches in Subsequent Step** | **Potential FPs Matches (Death and Patient Record) for Manual Review[a]** | **Matched Total Prescriptions** | **New Prescriptions Matches/ Non-matches in Subsequent Step** |
| Deterministic record linkage approach | | | | | |
| Approach 1: Clean and standardized names in SAS, exact match on name, and DOB using SQL programming | 4,572 | NA | 8,671 | 233,486 | NA |
| Fuzzy/Probabilistic record linkage approaches | | | | | |
| Approach 2: Name cleaning and standardization in SAS[b], fuzzy matching in SAS DMS (name sensitivity[c] 85, DOB sensitivity 95) | 4,684 | 124 (compared with approach 1)[d] | 867 | 245,885 | 13,435/1,036 (compared with approach 1)[e] |
| Approach 3: Name cleaning and standardization in SAS[b], fuzzy matching in SAS DMS (name sensitivity[c] 85, DOB sensitivity 85) | 4,714 | 30 (compared with approach 2) | 1,235 | 250,082 | 4,197/0 (compared with approach 2) |
| Approach 4: Name cleaning, standardization[f], and fuzzy matching in SAS DMS (name sensitivity[c] 85, DOB sensitivity 95) | 4,680 | 0 (compared with approach 3) | 1,834 | 245,087 | 15/5,010 (compared with approach 3) |
| Approach 5: Name cleaning, standardization[f], and fuzzy matching in SAS DMS (name sensitivity[c] 85, DOB sensitivity 85) | 4,710 | 30 (compared with approach 4)[d] | 2,197 | 249,257 | 4,170/0 (compared with approach 4) |

[a] See Figure 2 for how this number was identified for approaches 2–4, and results text for approach 1. See Table 2 for FP match analysis and classification results.
[b] Name cleaning and standardization conducted using statistical programming techniques in SAS.
[c] Sensitivity values from SAS DMS using the data quality knowledge base are based on a combination of standardization and fuzzy/probabilistic matching techniques. A context-specific matching algorithm that combines parsing, standardization, regular expression processing, and phonetic (e.g., phonetics, transformations) using natural language processing.
[d] Twelve matches in approach 1 were not captured in any of the subsequent approaches. Four matches in approach 3 were not captured in approaches 4 or 5.
[e] A total of 1,036 prescriptions were matched in approach 1, but were not captured in any of the subsequent approaches.
[f] Name and DOB cleaning and standardization conducted using SAS DMS, and based on SAS data quality knowledge base, a repository of rules and reference data. SAS Data Management Studio (SAS DMS).

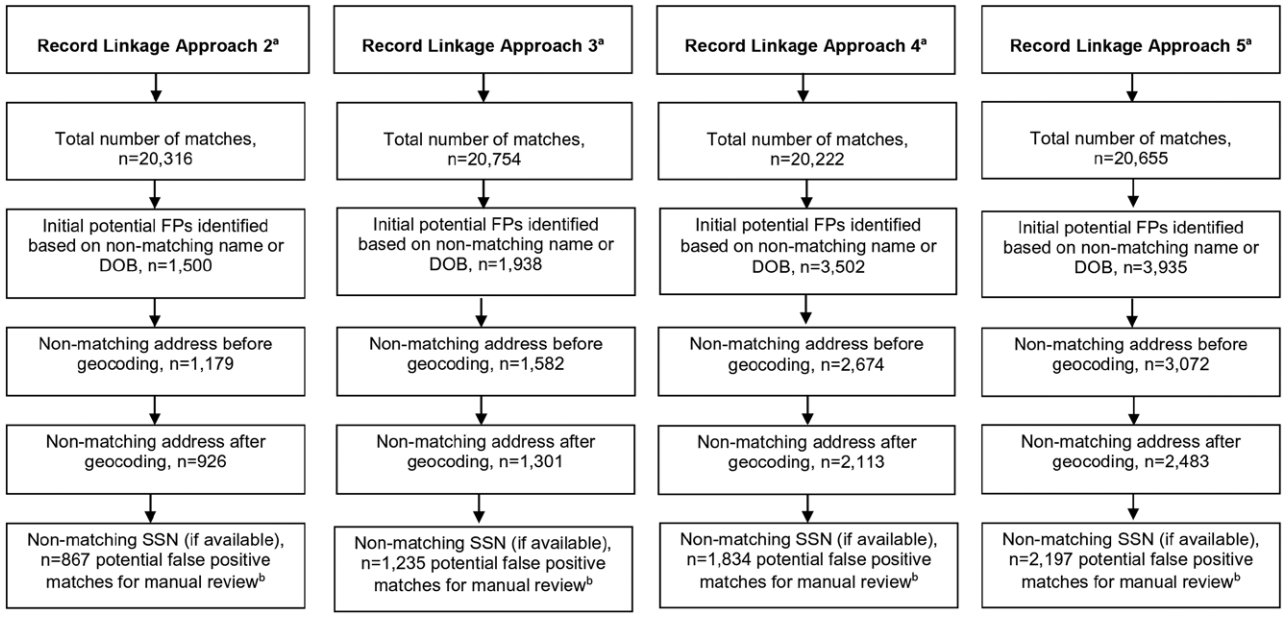


**FIGURE 2.** Identification of FP Matches for Manual Review for Four Fuzzy/Probabilistic Record Linkage Approaches [a]See Table 1 for approach descriptions. [b]Results from manual review are summarized in Table 2.

**TABLE 2.** Manual Review of Potential FP Matches from Fuzzy/Probabilistic Matching Algorithms and Final Classification

| | Approach 2 | | Approach 3 | | Approach 4 | | Approach 5 | |
|---|---|---|---|---|---|---|---|---|
| | n = 867[a] | | n = 1,235[a] | | n = 1,834[a] | | n = 2,197[a] | |
| **Data Quality Issue** | **n** | **%** | **n** | **%** | **n** | **%** | **n** | **%** |
| Addressed matched via manual review[b] | 53 | 6.1 | 62 | 5.0 | 95 | 5.2 | 103 | 4.7 |
| Misspellings/alternative spellings[b] | 388 | 45 | 407 | 33 | 416 | 23 | 482 | 22 |
| Nicknames[b] | 369 | 43 | 406 | 33 | 353 | 19 | 392 | 18 |
| Multiple last names[b] | 110 | 13 | 134 | 11 | 140 | 7.6 | 144 | 6.6 |
| Switched order[b] | 7 | 0.81 | 7 | 0.60 | 6 | 0.30 | 12 | 0.60 |
| Different first name[b] | 8 | 0.92 | 44 | 3.6 | 6 | 0.30 | 47 | 2.1 |
| Different last name[b] | 4 | 0.46 | 16 | 1.3 | 4 | 0.20 | 44 | 2.0 |
| Different DOB[b] | 0 | | 368 | 30 | 0 | | 363 | 17 |
| Cleaning/parsing middle names, prefix, suffix, notes[b] | NA | | NA | | 978 | 53.3 | 1,019 | 46.4 |
| Final FP classification for patient matches after manual review[c] | n = 20,316[a] | | n = 20,754[a] | | n = 20,222[a] | | n = 20,655[a] | |
| Correct match | 19,603 | 97 | 19,695 | 95 | 18,609 | 92 | 18,688 | 91 |
| Possible FP | 706 | 3.5 | 1,004 | 4.9 | 1,607 | 7.9 | 1,881 | 9.1 |
| True FP | 7 | 0.03 | 55 | 0.30 | 6 | 0.03 | 86 | 0.41 |
| Number of possible and true FP prescriptions[d] | n = 245,885[d] | | n = 250,082[d] | | n = 245,087[d] | | n = 249,257[d] | |
| Correct prescriptions | 240,337 | 98 | 241,359 | 97 | 233,456 | 95 | 234,394 | 94 |
| Possible FP prescriptions | 5,509 | 2.2 | 8,368 | 3.3 | 11,612 | 4.7 | 14,313 | 5.7 |
| True FP Prescriptions | 39 | 0.02 | 355 | 0.14 | 19 | 0.01 | 550 | 0.22 |

[a] Numbers are for patient and death record matches, one unique overdose death may have >1 match to a patient record in the CSMD and each patient record may include one or more prescription records.
[b] Counts are not mutually exclusive.
[c] Excludes matches that were found to be correct matches based on address manual review.
[d] Number of prescriptions.

different types (opioids, benzodiazepines) and timing (active at overdose and 60 days before overdose). The overall differences in distributions, when compared with frequencies, were small. For example, 1,371 (32%) and 1,443 (33%) of decedents with a prescription filled for oxycodone in the 60 days before overdose for approach 1 and approach 3, respectively. While the proportions are similar, 72 additional patients were identified in approach 3 (when compared with approach 1). We also provide the difference in proportions and 95% confidence intervals in eTable 4; http://links.lww.com/EDE/B594.

We compared decedents missed in approach 1, but included in the most accurate approach (i.e., approach 3), to understand the characteristics of decedents excluded from the population when using the deterministic approach (Table 4). Missed overdose decedents were more likely to be younger, male, nonwhite, and of higher education. A higher proportion of eligible fentanyl and heroin decedents (i.e., with a prescription history in the CSMD who should have been linked), and unintentional overdoses, were found among missed matches. We also compared characteristics of overdose deaths with 1 or more possible/true FP matches to decedents with no incorrect matches (Table 4). Decedents with FP matches were more likely to be middle-aged, male, unmarried at time of overdose, and of lower education.

## DISCUSSION

In our study using PDMP and mortality data, we evaluated the accuracy of record linkage approaches using varying methods for cleaning, standardization, and matching, and developed a systematic process to identify and exclude FPs prior to analyses. The number of matches (for overdose decedents to their patient record[s] in the CSMD) was improved when we implemented a comprehensive name-cleaning protocol using database-specific techniques via statistical programming and incorporating probabilistic/fuzzy matching (when compared with deterministic). The most accurate approach identified an additional 142 overdose deaths, and the trade-off with increased FP matches (i.e., patient and death record matches) was small (FP linkage error of <1%). Further, the correct match proportion after manual review was ~95% (compared with 92% and ~91% for the probabilistic approaches not including comprehensive data cleaning).

Assessment of the accuracy of linkage approaches, including missed and FP matches, and consideration of potential misclassification of study variables and selection bias is needed to improve validity and results interpretation in public health analyses and epidemiologic studies.[15,17,19,25,34] Deterministic linkage generally has high specificity, with a trade-off in sensitivity,[15] potentially resulting in underestimation of health statistics.[24] However, this depends on the completeness

**TABLE 3.** Influence of Entity Resolution/Record Linkage Approaches on PDMP Prescription Measures for All Drug Overdose Decedents in Tennessee

| | Statistical Programming Approach (n = 4,259)[a] | Most Accurate Probabilistic Approach (n = 4,400)[a] |
|---|---|---|
| All prescriptions[b], mean (range) | 32.2 (1–224) | 33.1 (1–224) |
| Opioid prescriptions[b], mean (range) | 19.9 (1–182) | 20.4 (1–182) |
| Benzodiazepine prescriptions, mean (range) | 14.4 (1–96) | 14.7 (1–96) |
| Opioid days' supply, median (IQR) | 269 (30–705) | 278 (30–720) |
| Benzodiazepine days' supply, median (IQR) | 360 (110–634) | 381.5 (120–652) |
| Total morphine milligram equivalents for all opioid analgesics filled in the last 60 days before overdose, median (IQR) | 2,700 (750–7,890) | 2,745 (750–8,100) |
| Total morphine milligram equivalents for all opioid analgesics filled in the last 180 days before overdose, median (IQR) | 5,130 (675–16,800) | 5,205 (675–17,100) |
| Number of prescribers in the year before overdose[b], n (%) | | |
| 1 | 871 (22) | 876 (22) |
| 2 | 738 (19) | 760 (19) |
| 3 | 573 (15) | 604 (15) |
| 4 | 430 (11) | 459 (11) |
| ≥5 | 1,313 (34) | 1,375 (34) |
| Number of dispensers in the year before overdose[b], n (%) | | |
| 1 | 1,404 (36) | 1,409 (35) |
| 2 | 978 (25) | 1,018 (25) |
| 3 | 645 (16) | 676 (17) |
| 4 | 365 (9) | 385 (10) |
| ≥5 | 533 (14) | 586 (14) |
| Active prescription at overdose death, n (%) | | |
| No | 1,871 (44) | 1,898 (43) |
| Yes | 2,388 (56) | 2,502 (57) |
| Any prescription in the last 60 days before overdose, n (%) | | |
| No | 1,286 (30) | 1,283 (29) |
| Yes | 2,973 (70) | 3,117 (71) |
| Opioid prescription in the last 60 days before overdose, n (%) | | |
| No | 1,781 (42) | 1,794 (41) |
| Yes | 2,478 (58) | 2,606 (59) |
| Oxycodone prescription in the last 60 days before overdose, n (%) | | |
| No | 2,888 (68) | 2,957 (67) |
| Yes | 1,371 (32) | 1,443 (33) |
| Hydrocodone prescription in the last 60 days before overdose, n (%) | | |
| No | 3,251 (76) | 3,335 (76) |
| Yes | 1,008 (24) | 1,065 (24) |
| Buprenorphine for MAT prescription in the last 60 days before overdose, n (%) | | |
| No | 4,120 (97) | 4,255 (97) |
| Yes | 139 (3.3) | 145 (3.3) |
| Benzodiazepines in the 60 days before overdose, n (%) | | |
| No | 2,372 (56) | 2,409 (55) |
| Yes | 1,887 (44) | 1,991 (45) |
| Active opioid prescription at overdose, n (%) | | |
| No | 2,485 (58) | 2,525 (57) |
| Yes | 1,774 (42) | 1,875 (43) |
| Active benzodiazepine prescription at overdose, n (%) | | |
| No | 2,678 (63) | 2,735 (62) |
| Yes | 1,581 (37) | 1,665 (38) |
| Cash payment for a prescription in the year before overdose[c], n (%) | | |
| No | 1,830 (47) | 1,862 (46) |
| Yes | 2,060 (53) | 2,178 (54) |

Table excludes true FP prescriptions (excludes only 1 death that had FP prescriptions in the year before overdose).

[a] Decedents in this analysis had to fill at least 1 prescription on or before the date of death within 2 years of death with an eligible prescription defined as at least 1 days' supply and NDC number linked to the 2017 CDC Oral morphine milligram equivalents Drug Classification Table.

[b] Excludes decedents with no prescriptions in the year before death.

[c] Excludes prescriptions with no payment information.

Medication Assisted Treatment (MAT).

**TABLE 4.** Characteristics of Missed Matches and FP Matches for Overdose Decedents[a]

| | Overdoses Linked in Approach 1 (n = 4,572) | | Missed Eligible Overdoses[b] (n = 142) | | Overdose Decedents from Most Accurate Record Linkage Approach (n = 4,714) | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Deaths Excluding FP Matches (n = 4,243) | | Deaths with ≥1 Possible/ True FP matches[b] (n = 471) | |
| | n | % | n | % | n | % | n | % |
| Age at death | | | | | | | | |
| <25 | 243 | 5.3 | 10 | 7.0 | 232 | 5.5 | 17 | 3.6 |
| 25–34 | 823 | 18 | 35 | 25 | 757 | 18 | 85 | 18 |
| 35–54 | 2,471 | 54 | 71 | 50 | 2,266 | 53 | 279 | 59 |
| 55–64 | 778 | 17 | 18 | 13 | 741 | 18 | 77 | 16 |
| ≥65 | 257 | 5.6 | 8 | 5.6 | 247 | 5.8 | 13 | 3.0 |
| Race/ethnicity[c] | | | | | | | | |
| NonHispanic White | 4,111 | 91 | 120 | 85 | 3,806 | 91 | 427 | 92 |
| NonHispanic Black | 329 | 7.3 | 16 | 11 | 310 | 7.4 | 35 | 7.6 |
| Other | 64 | 1.4 | 5 | 3.6 | 63 | 1.5 | 1 | 0.2 |
| Gender | | | | | | | | |
| Male | 2,495 | 55 | 90 | 63 | 2,309 | 54 | 285 | 61 |
| Female | 2,077 | 45 | 52 | 37 | 1,934 | 46 | 186 | 40 |
| Marital status[c] | | | | | | | | |
| Never Married | 1,297 | 29 | 32 | 29 | 1,204 | 29 | 139 | 31 |
| Not Married | 1,780 | 40 | 47 | 42 | 1,653 | 40 | 190 | 42 |
| Married | 1,373 | 31 | 33 | 30 | 1,279 | 31 | 120 | 27 |
| Education[c] | | | | | | | | |
| <High school | 1,121 | 25 | 28 | 21 | 1,026 | 25 | 133 | 29 |
| High school | 2,085 | 46 | 66 | 49 | 1,928 | 46 | 223 | 49 |
| >High School | 1,300 | 29 | 42 | 31 | 1,230 | 29 | 102 | 22 |
| Opioid overdose[d] | | | | | | | | |
| No | 1,313 | 29 | 47 | 33 | 1,247 | 29 | 117 | 25 |
| Yes | 3,259 | 71 | 95 | 67 | 2,996 | 71 | 354 | 75 |
| Fentanyl overdose[e] | | | | | | | | |
| No | 4,073 | 89 | 113 | 80 | 3,788 | 89 | 412 | 88 |
| Yes | 499 | 11 | 29 | 20 | 455 | 11 | 59 | 13 |
| Heroin overdose[f] | | | | | | | | |
| No | 4,045 | 89 | 120 | 85 | 3,757 | 89 | 413 | 88 |
| Yes | 527 | 12 | 22 | 16 | 486 | 12 | 58 | 12 |
| Intentionality | | | | | | | | |
| Unintentional | 4,002 | 88 | 128 | 90 | 3,709 | 87 | 410 | 87 |
| Homicide | 370 | 8.1 | 9 | 6 | 349 | 8.2 | 33 | 7.0 |
| Suicide | 5 | 0.1 | 0 | | 5 | 0.1 | | |
| Undetermined | 195 | 4.3 | 5 | 3.5 | 180 | 4.2 | 28 | 5.9 |

[a] Here a match refers to death record in death statistical file and patient record in CSMD.
[b] Additional matches identified in most accurate fuzzy matching approach (Approach 3) and incorrect matches identified in approach 3.
[c] Proportions exclude missing data.
[d] Defined using ICD-10 codes T40.0-T40.4, T40.6.
[e] Defined using literal text searches.
[f] Defined using ICD-10 code T40.1.

and accuracy of identifiers and deterministic linkage approaches (e.g., exact matching or multi-stage). Probabilistic linkage is generally more sensitive (fewer missed matches), with a potential increase in FPs.[15,16,31] In our study, the most accurate approach resulted in improved estimation of descriptive prescription measures of common interest in public health and epidemiologic analyses (compared with the deterministic approach). Other studies have shown that probabilistic

matching can improve sensitivity and linkage rates, including record linkage studies utilizing health claims, vital statistics, and hospital administrative data.[19,20,25,35] The reduction in missed matches can reduce bias in study estimates and improve study generalizability by including eligible records/study participants that may have been missed,[20,24,25,35] which we demonstrate in our study. Specifically, we found that decedents who would have been excluded using only a deterministic approach but included using a probabilistic approach tended to be younger, nonwhite, male, and of higher education. We also found that the distributions for characteristics of decedents with one or more true or possible FP matches were potentially different from the population identified for analysis using the most accurate approach, with a higher proportion in the age group 35–54 years (59% vs. 53%), of male gender (61% vs. 54%), and of lower education (<high school 29% vs. 25%). Our findings, which should be interpreted with the caveat that they are descriptive, highlight the importance of considering both false negatives and FPs on study variable measurement and potential selection bias.

We show that approaches used to identify prescription history in PDMP data, and for linking patients to health outcomes, have important implications for accuracy and bias, including underreporting, prescribing and outcome misclassification, and reduced generalizability to all populations at risk. We believe the methods can be used by public health epidemiologists and researchers to improve validity and interpretation for public health analyses and epidemiologic studies using PDMP data, regardless of setting. Our study is unique in that we evaluated multiple record linkage approaches, considering both cleaning and standardization of fields before use (often overlooked but critical components of best practices in record linkage methodology[16]), as well as comparison of matching methods (deterministic and probabilistic/fuzzy matching). We demonstrated that our own in-house database-specific cleaning methodology improved accuracy (see eAppendix 2; http://links.lww.com/EDE/B594), including increased number of outcome linkages, and reduced FP linkage errors. We also developed an approach to identify and exclude FPs, which can improve quality of findings enabling exclusion of incorrect matches prior to analysis. We found that using additional available potential identifiers, even if only available for a subset of records, can reduce burden in manual review of potential FP records for match status validation. Further, we provide an analysis of the primary reasons for FP matches, which could be used to select high priority records for manual review when time and/or resources are limited.

A limitation of our study is that we did not have a true "gold standard" identifier that would enable us to confirm all death and patient records matches (such as SSN or health medical record number). This is generally required for calculation of sensitivity and specificity,[35,36] as it is a way to confirm true match status without manual review. However, we did manually review all potential FP matches for each linkage approach, with systematic evaluation and analysis to identify true FPs. Therefore, our estimates of correct matches (excluding any overdose deaths with one or more true FP match), should be quite accurate, although the possibility of some human error remains. Another limitation of our work is that we focused on mortality data only. It is important to note that this record linkage framework can be applied beyond mortality to other health data, such as infectious disease (e.g., blood-borne infections associated with injection drug use such as Hepatitis C) and administrative data, to enable evaluations of new risk factors and identification of susceptible populations and high risk groups.[37–39] Finally, we did not evaluate the influence of matching approaches on effect estimates for health outcomes associations as this was beyond the scope of the present study. Future studies using PDMP data, such as those of prescribing patterns and fatal and nonfatal overdose or use of prescription opioids during pregnancy and infant outcomes can provide opportunities for future research that apply the methodologies of the current work in multiple settings and populations.

## Public Health Implications

The use of PDMP data for epidemiologic studies and public health surveillance has greatly increased in recent years as these data have accumulated as part of the response to the opioid epidemic. We comprehensively evaluated multiple record linkage approaches using PDMP data, including an analysis of FPs and assessment of potential for bias by comparing prescribing measures, characteristics of missed and incorrect matches, and excluding potentially incorrect matches identified by manual review. We show that approaches used to identify prescription history in PDMP data, and for linking patients to mortality outcomes, a common outcome of interest, have important implications for accuracy and bias. These implications, which apply to public health analyses used to inform prevention and intervention efforts as well as epidemiologic studies, include underreporting, prescribing and outcome misclassification, and potential for reduced generalizability to all population risk groups.

## REFERENCES


1. Meldrum ML. The ongoing opioid prescription epidemic: historical context. *Am J Public Health*. 2016;106:1365–1366.
2. Dasgupta N, Beletsky L, Ciccarone D. Opioid crisis: no easy fix to its social and economic determinants. *Am J Public Health*. 2018;108:182–186.
3. Prescription Drug Monitoring Program Training and Technical Assistance Center. http://www.pdmpassist.org/pdf/PDMP_Program_Status_20170824.pdf. Accessed 20 January 2018.
4. Lin DH, Lucas E, Murimi IB, et al. Physician attitudes and experiences with Maryland's prescription drug monitoring program (PDMP). *Addiction*. 2017;112:311–319.
5. McCauley JL, Leite RS, Gordan VV, et al.; National Dental Practice-Based Research Network Collaborative Group. Opioid prescribing and risk mitigation implementation in the management of acute pain: results from the national dental practice-based research network. *J Am Dent Assoc*. 2018;149:353–362.
6. Christianson H, Driscoll E, Hull A. Alaska nurse practitioners' barriers to use of prescription drug monitoring programs. *J Am Assoc Nurse Pract*. 2018;30:35–42.

7. Suffoletto B, Lynch M, Pacella CB, Yealy DM, Callaway CW. The effect of a statewide mandatory prescription drug monitoring program on opioid prescribing by emergency medicine providers across 15 hospitals in a single health system. *J Pain*. 2018;19:430–438.
8. Dasgupta N, Funk MJ, Proescholdbell S, Hirsch A, Ribisl KM, Marshall S. Cohort study of the impact of high-dose opioid analgesics on overdose mortality. *Pain Med*. 2016;17:85–98.
9. Deyo RA, Hallvik SE, Hildebran C, et al. Association between initial opioid prescribing patterns and subsequent long-term use among opioid-naïve patients: a Statewide Retrospective Cohort Study. *J Gen Intern Med*. 2017;32:21–27.
10. O'Kane N, Hallvik SE, Marino M, et al. Preparing a prescription drug monitoring program data set for research purposes. *Pharmacoepidemiol Drug Saf*. 2016;25:993–997.
11. Fink PB, Deyo RA, Hallvik SE, Hildebran C. Opioid prescribing patterns and patient outcomes by prescriber type in the oregon prescription drug monitoring program. *Pain Med*. 2018;19:2481–2486.
12. Hallvik SE, Geissert P, Wakeland W, et al. Opioid-prescribing continuity and risky opioid prescriptions. *Ann Fam Med*. 2018;16:440–442.
13. Geissert P, Hallvik S, Van Otterloo J, et al. High-risk prescribing and opioid overdose: prospects for prescription drug monitoring program-based proactive alerts. *Pain*. 2018;159:150–156.
14. Deyo RA, Hallvik SE, Hildebran C, et al. Association of prescription drug monitoring program use with opioid prescribing and health outcomes: a comparison of program users and nonusers. *J Pain*. 2018;19:166–177.
15. Dusetzina S, Tyree S, Meyer A, Meyer A, Green L, Carpenter W. Linking data for health services research: a framework and instructional guide (Prepared by the University of North Carolina at Chapel Hill under Contract No. 290-2010-000141.) AHRQ Publication No. 14-EHC033-EF. Rockville, MD: Agency for Healthcare Research and Quality; September 2014. www.effectivehealthcare.ahrq.gov/reports/final.cfm.
16. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. *Int J Epidemiol*. 2016;45:954–964.
17. Joffe E, Byrne MJ, Reeder P, et al. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. *J Am Med Inform Assoc*. 2014;21:97–104.
18. McCoy AB, Wright A, Kahn MG, Shapiro JS, Bernstam EV, Sittig DF. Matching identifiers in electronic health records: implications for duplicate records and patient safety. *BMJ Qual Saf*. 2013;22:219–224.
19. Aldridge RW, Shaji K, Hayward AC, Abubakar I. Accuracy of probabilistic linkage using the enhanced matching system for public health and epidemiological studies. *PLoS One*. 2015;10:e0136179.
20. Hagger-Johnson G, Harron K, Goldstein H, Aldridge R, Gilbert R. Probabilistic linkage to enhance deterministic algorithms and reduce data linkage errors in hospital administrative data. *J Innov Health Inform*. 2017;24:891.
21. Brenner H, Schmidtmann I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. *Stat Med*. 1997;16:2633–2643.
22. Schmidlin K, Clough-Gorr KM, Spoerri A, Egger M, Zwahlen M; Swiss National Cohort. Impact of unlinked deaths and coding changes on mortality trends in the Swiss National Cohort. *BMC Med Inform Decis Mak*. 2013;13:1.
23. Miller EA, McCarty FA, Parker JD. Racial and ethnic differences in a linkage with the national death index. *Ethn Dis*. 2017;27:77–84.
24. Moore CL, Gidding HF, Law MG, Amin J. Poor record linkage sensitivity biased outcomes in a linked cohort analysis. *J Clin Epidemiol*. 2016;75:70–77.
25. Harron KL, Doidge JC, Knight HE, et al. A guide to evaluating linkage quality for the analysis of linked data. *Int J Epidemiol*. 2017;46:1699–1710.
26. Tennessee Department of Health. Tennessee Chronic Pain Guildelines (2017). https://www.tn.gov/content/dam/tn/health/healthprofboards/ChronicPainGuidelines.pdf. Accessed 1 May 2018.
27. Tennesseee Department of Health. Bureau of Policy, Planning and Assessment. Division of Health Statistics. *Death Statistical File User Manual*. January 2014.
28. Nechuta SJ, Tyndall BD, Mukhopadhyay S, McPheeters ML. Sociodemographic factors, prescription history and opioid overdose deaths: a statewide analysis using linked PDMP and mortality data. *Drug Alcohol Depend*. 2018;190:62–71.
29. SAS Insitute Inc. *SAS 9.4 Data Management: Overview*. Cary, NC: SAS Institute Inc; 2016.
30. Golladay M, Nechuta S. Lessons Learned: Deep Cleaning Procedure Design for Name Variables in the Tennessee CSMD. Available at: https://www.tn.gov/content/dam/tn/health/documents/opioid_response/CSMDNameCleaningReport.pdf. Accessed 29 April 2019.
31. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *Int J Epidemiol*. 2002;31:1246–1252.
32. Prescription Drug Monitoring Program Training and Technical Assistance Center. Technical Assistance Guide. PDMP Suggested Practices to Ensure Pharmacy Compliance and Improve Data Integrity. April 13, 2015. Available here: http://www.pdmpassist.org/pdf/Resources/Pharmacy_compliance_data_quality_TAG__FINAL_20150615_A.pdf. Accessed 29 April 2019.
33. National Center for Injury Prevention and Control. CDC compilation of benzodiazepines, muscle relaxants, stimulants, zolpidem, and opioid analgesics with oral morphine milligram equivalent conversion factors, 2018 version. Atlanta, GA: Centers for Disease Control and Prevention; 2018. Available at https://www.cdc.gov/drugoverdose/resources/data.html. Accessed 29 April 2019.
34. Campbell KM. Impact of record-linkage methodology on performance indicators and multivariate relationships. *J Subst Abuse Treat*. 2009;36:110–117.
35. Baldwin E, Johnson K, Berthoud H, Dublin S. Linking mothers and infants within electronic health records: a comparison of deterministic and probabilistic algorithms. *Pharmacoepidemiol Drug Saf*. 2015;24:45–51.
36. Silveira DP, Artmann E. Accuracy of probabilistic record linkage applied to health databases: systematic review. *Rev Saude Publica*. 2009;43:875–882.
37. Degenhardt L, Peacock A, Colledge S, et al. Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of HIV, HBV, and HCV in people who inject drugs: a multistage systematic review. *Lancet Glob Health*. 2017;5:e1192–e1207.
38. Lo-Ciganic WH, Huang JL, Zhang HH, et al. Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. *JAMA Netw Open*. 2019;2:e190968.
39. Moyo P, Zhao X, Thorpe CT, et al. Dual receipt of prescription opioids from the fepartment of veterans affairs and medicare part D and prescription opioid overdose death among veterans: a Nested Case-Control Study. *Ann Intern Med*. 2019;170:433–442.