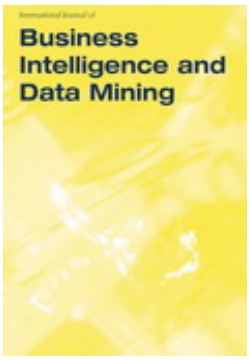



**International Journal of Business Intelligence and Data Mining**

ISSN online: 1743-8195 - ISSN print: 1743-8187
https://www.inderscience.com/ijbidm

# Application of a record linkage software to identify mortality of enrolees of large integrated healthcare organisations

Yichen Zhou, Zhi Liang, Sungching Glenn, Wansu Chen, Fagen Xie

**DOI:** 10.1504/IJBIDM.2022.10042864

**Article History:**
Received: 18 July 2021
Accepted: 14 September 2021
Published online: 30 November 2022

# Application of a record linkage software to identify mortality of enrolees of large integrated healthcare organisations

Yichen Zhou, Zhi Liang, Sungching Glenn, Wansu Chen and Fagen Xie*

Department of Research and Evaluation,
Kaiser Permanente Southern California,
100 S. Los Robles Ave., 2nd Floor,
Pasadena CA 91101, USA
Email: yichen.zhou@kp.org
Email: zhi.liang@kp.org
Email: sungching.c.glenn@kp.org
Email: wansu.chen@kp.org
Email: fagen.xie@kp.org
*Corresponding author

**Abstract:** Information on mortality is important for the improvement of public health and the conduct of medical research. Healthcare organisations typically lack complete and accurate information on mortality. This paper proposes a comprehensive process to link the records of the enrolees of a healthcare organisation with the death records of 2015 obtained from the California State via a commercial data linkage software. The developed linkage process has successfully identified 23,628 and 21,009 death records of health plan enrolees from the state file after the initial and second post-linkage, respectively. Validation of the linkage process against the deaths records documented in the internal systems of the organisation achieved a sensitivity of 97.5% and a positive predictive value of 88.7% at the time of initial linkage but increased to 99.4% in three years using more information available later. The linkage process demonstrated high accuracy and can be utilised to support various business needs.

**Keywords:** data cleaning; data standardisation; data matching; mortality linkage.

**Reference** to this paper should be made as follows: Zhou, Y., Liang, Z., Glenn, S., Chen, W. and Xie, F. (2023) 'Application of a record linkage software to identify mortality of enrolees of large integrated healthcare organisations', *Int. J. Business Intelligence and Data Mining*, Vol. 22, Nos. 1/2, pp.264–285.

**Biographical notes:** Yichen Zhou is an analytical programmer in the Department of Research and Evaluation at Kaiser Permanente Southern California. She has a Master of Science degree in Biostatistics from the University of Connecticut. She has been an Informatics Analyst for several federal funded grants. Her research work focuses on the areas of healthcare data linkage, information retrieval, prediction modelling and natural language processing.

Zhi Liang is a data specialist in the Department of Research and Evaluation at Kaiser Permanente Southern California. He received his Doctorate degree in Mathematical Sciences from New Jersey Institute of Technology. His research work focuses on the areas of scientific computing, specialising in large-scale simulation and healthcare informatics data linkage.

Sungching Glenn is a senior SAS programmer in the Department of Research and Evaluation at Kaiser Permanente Southern California. She is the site data manager for a large-scale federal funded research project. During her tenure, she has developed and maintained many comprehensive infrastructure data files.

Wansu Chen is a Research Scientist II in the Department of Research and Evaluation at Kaiser Permanente Southern California. She received her Doctorate degree in Biostatistics from University of Southern California. Her research interest includes pancreatic cancer, asthma, heart failure, atrial fibrillation, mortality trend analysis, predictive modelling and natural language processing.

Fagen Xie is a Senior Lead Architect Research Informatics in the Department of Research and Evaluation at Kaiser Permanente Southern California. He received his Doctorate degree in Physics from Beijing Normal University, China. His current research interest includes healthcare data management and integration, health informatics, information retrieval, predictive modelling and clinical natural language processing.

## 1 Introduction

Information on mortality is important for assessing community health status, developing health policy, improving practice guidelines, and conducting healthcare research (Sorlie et al., 1995). Examples of how mortality data are used in research include the determination of death as an outcome or a censoring event and studying causes of death and their associations with other factors (Benjamin et al., 2017; Go et al., 2004). However, healthcare organisations typically lack complete and accurate information on mortality. For example, the electronic systems of these organisations may not capture deaths that occur after enrolees are disenrolled from the health plan or deaths that occur to enrolees with dual health insurance coverage. On the other hand, death certificates legally issued by the State of California contain complete and official information including cause of death, which is often utilised to conduct research (California, 2020). Therefore, a process that links the records of deceased individuals retrieved from the healthcare organisation with the death certificate data obtained from the State has the potential to collate complete and accurate mortality information for deceased enrolees (Alonso-Sardón et al., 2015; Go et al., 2004; Krewski et al., 2005), which can mitigate the misclassification bias and facilitate medical research studies to examine the more accurate causes and factors associated with deaths.

Record linkage is typically a lengthy and challenging process because no unique and high-quality identifier can be simply used (Bohensky et al., 2010; Christen and Goiser, 2007; Dunn, 1946; Harron et al., 2017). Although social security number (SSN), is available in both data sources, it is incomplete and subject to errors. Thus, a common practice is to use multiple identifiers, namely SSN, names, gender, date of births, race/ethnicity, and/or addresses to conduct the linkage. Although one can write a 'join' or 'merge' statement using various programming languages to link a large volume of records with multiple identifiers from different data sources, even when all or some of these identifiers are incomplete or inaccurate (i.e., misspelled names), the process is inefficient and requires the extensive theoretical knowledge of record linkage. Therefore, a number of linkage algorithms and corresponding software packages have been developed for the data linkage process (Arellano et al., 1984; ChoiceMaker, 2021; Choi et al., 2017; Dusetzina et al., 2014; Fair, 2004; Fair et al., 2000; Fellegi and Sunter, 1969; Gidding et al., 2017; Kara, 1996; Karr et al., 2019; Mamun et al., 2016; Wunsch and Gourbin, 2018). Linkage systems were also developed within government agencies or academia (Alur et al., 2008; Arellano et al., 1984; CDC, 2021; Fair, 2004; Herzog et al., 2007; Kara, 1996; Washington, 2021). Some of these linkage systems are publicly available (CDC, 2021; ChoiceMaker, 2021; Washington, 2021). One of these automated record linkage systems, named AutoMatch, was developed by the national agricultural statistics service (Kara, 1996). AutoMatch was integrated into a commercial software called INTEGRITY by Vality Technology Inc. and subsequently into the IBM InfoSphere software (Alur et al., 2008; Herzog et al., 2007).

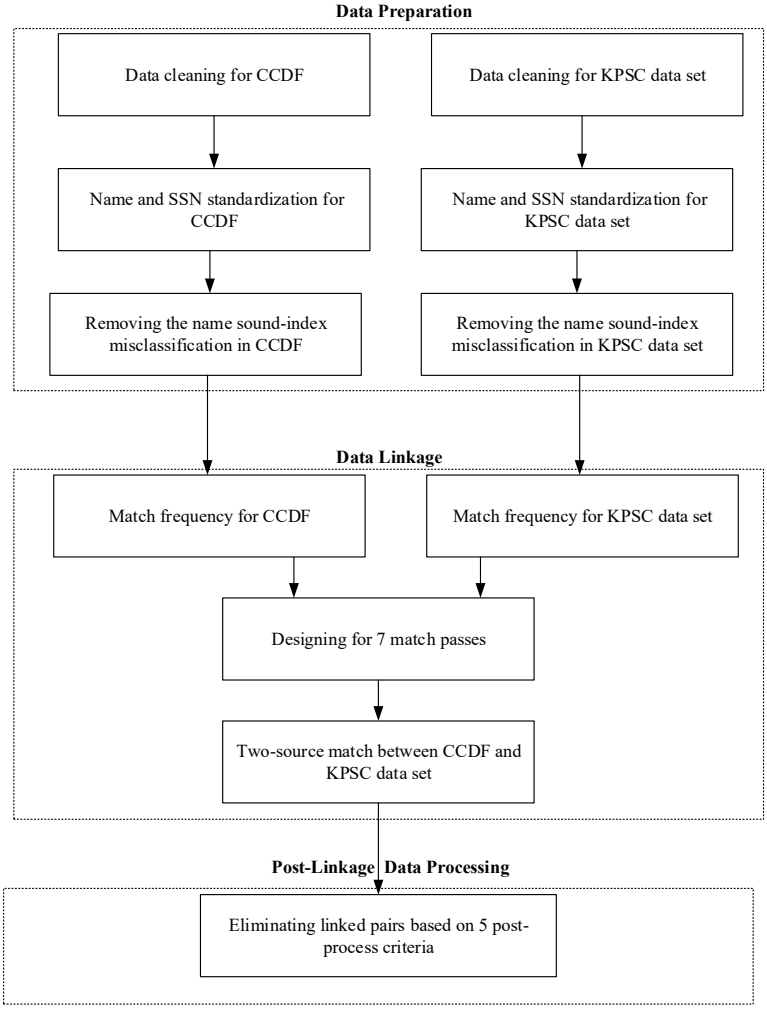
In this study, we sought to leverage the IBM InfoSphere software (IBM, 2021) to design a comprehensive process to link decedents in the California comprehensive death file (CCDF) with health plan enrolees within a large integrated healthcare organisation, Kaiser Permanente Southern California (KPSC). We evaluated the performance of the linkage process. Our purpose is to demonstrate the linkage process using the software and to shorten the users' learning curve by informing the potential pitfalls and some technical details that are not well documented/released by the software vendor.

## 2 Methods

### *2.1 Study setting*

KPSC is an integrated healthcare system that provides medical services to over 4.6 million members (Koebnick et al., 2012) through its 15 hospitals and over 220 satellite medical offices throughout Southern California. Information on patient demographics is collected through a comprehensive electronic medical record (EMR) system and an electronic system called foundation system. The foundation system manages information on insurance plans, enrolments, benefits, purchasers and contracts, and routinely exchanges information with the EMR system. The study was approved by KPSC's institutional review board (IRB) with waivers of informed consent and health insurance portability and accountability act authorisation.

**Figure 1** Diagram illustrating the entire data linkage process




## *2.2 Eligibility and data sources*

To conduct linkage, relevant data were gathered from the KPSC's EMR system and foundation system (referred to as the 'KPSC systems' below) and from the State of California. From the KPSC systems, we first identified all individuals who had at least one day of KPSC enrolment in 2015. We then extracted demographic data including SSN, name (last, first, and middle), date of birth, gender, race/ethnicity, and county of residence. From the State of California, four California death data sources released by the California department of public health were available from the centre for health statistics and informatics (California Department of Public Health, 2020). The CCDF was obtained for individuals who died in 2015 and used by this study. This file also contained out-of-state deaths submitted by other states or jurisdictions; however, the personal identities of the out-of-state deaths were redacted. The same demographic data mentioned above were

also extracted from the CCDF. Each data source also had an internal identifier uniquely associated with each individual. In addition, the file from the State of California also contained information on the underlying cause of death, location of death, and many other useful data elements.

### *2.3 Design of data linkage process*

The diagram illustrating the entire data linkage process included data preparation, data linkage, and post-linkage data processing is shown in Figure 1.

#### *2.3.1 Data preparation*

Data preparation was the essential first step for achieving good performance in data linkage (Playford et al., 2016; Randall et al., 2013) because the identifiers extracted from the two sources varied in structure, format, and quality. The quality stage (QS) component of IBM InfoSphere software contains several standardised rules for correctly parsing and identifying each element or token and placing them in the appropriate column in the output file. The standardised rule sets can assimilate the data and append additional information from the input data. For both datasets (one dataset from each data source), the first and last names were first cleaned by removing the suffix and any special characters and then standardised by applying the USNAME rules pre-defined in the QS component (top of Figure 1). The standardisation process outputted the reverse Soundex of the first name (SFUSNAM) and last name (SLUSNAM). The SSN field was also cleaned and standardised based on the pre-defined USTAXID rules (Alur et al., 2008).

When we initially applied the above process to the CCDF file, 10.9% of the standardised first name or reverse Soundex of the first name (SFUSNAM) was null. The errors were mainly introduced by multiple word tokens of the standardised first names or last names, either from multiple word tokens of original names or from the concatenated first names and last names according to the USNAME standardised rules. For example, if CONCEPCION LEE was an input name, where CONCEPCION was the first name and LEE was the last name, the rules did not consider CONCEPCION as a valid first name. Thus, the standardised first name was null and the standardised last name was CONCEPCION LEE. The possible scenarios that introduced errors of name reverse Soundex are described in Supplementary Table 1. As a result, the following additional steps were introduced to create the name reverse Soundex to avoid potential errors:

1 Removed suffixes like ‘SR’, ‘JR’, ‘II’, ‘III’, ‘IIII’, ‘IV’, ‘V’, ‘MD’, ‘PhD’, and special characters and compress all spaces in the first name and the last name.

2 If there was only one word in the standardised last name, the reverse Soundex of the standardised last name was used as the reverse Soundex of the last name (SLUSNAM).

3 If there were two words in the standardised last name and no standardised first name was returned by the software, the reverse soundex of the first word of the standardised last name was used as the reverse soundex of the first name (SFUSNAM) and the reverse soundex of the second word in the standardised last name was used as the reverse soundex of the last name (SLUSNAM).

### *2.3.2 Data linkage*

The IBM InfoSphere software allows users to define a set of blocking and matching variables. The set of blocking variables is required to be identical while the set of matching variables can vary between the two data sources. These variables are typically selected based on the uniqueness and the reliability of the variables involved in the matching process. To balance completeness and efficiency, we designed a seven-step process in which variables with a higher level of uniqueness and quality had a higher priority to serve as blocking variables. SSN was the blocking variable for pass 1 since it contained the most unique information. As we mentioned earlier, SSN was still incomplete in the KPSC system and was also subject to error. Therefore, we designed passes 2–7 to match records that possibly belonged to the same individuals but did not have the same SSNs from the two sources, such as SSN missed from one source. For example, in pass 2, the month and year of birth date, as well as the last name, was used as the set of blocking variables because they were considered to be more reliable compared to the rest of other variables (e.g., day of birth date, first name). Because the last name could be misspelled, we added pass 3 to replace the last name with the reverse soundex of the last name. In pass 4, the month and year of birth date, as well as the reverse soundex of the first name, formed the set of blocking variables. The reverse soundex was assigned the same value for formal name versus nickname by the USNAME rules. For example, Elizabeth and Betty received the same reverse soundex as 'H312'. In passes 5–7, we further relaxed the set of blocking variables based on the same concepts being utilised to design passes 1–4. The specified blocking and matching variables for each pass are described in Supplementary Tables 2 and 3.

A person could have registered in different KPSC health plans at different times and thus be represented by several KPSC membership records. By contrast, a person in the CCDF was uniquely collected. Thus, it is likely that multiple individuals in the KPSC membership can link to a single individual in the CCDF. However, it is unlikely that more than one individual in the CCDF can link with the same individual in the KPSC membership dataset. This requirement can be met *by the many-to-one two-source matching stage setting*. Once an individual in the KPSC dataset was matched with an individual in the CCDF, the corresponding KPSC record was removed from the input dataset for the next pass. By contrast, all records in the CCDF were retained for all passes.

The designed linkage process between KPSC and CCDF is represented below.

1 Determined block variables, matching variables, matching comparison, and cutoff weight for each pass.

2 For each pass, compared record x from the CCDF and record y from the KPSC dataset:

   a Compared whether all the block variables $\{a_1, a_2, \ldots, a_n\}$ for two records were identical. If yes, go to step b.

b Calculated the contributed link weight of each matching variable $\{b_1, b_2, \ldots, b_m\}$ for $x$ and $y$. For each matching variable $b_i$, $i \in \{1, 2, \ldots, m\}$, two statistical properties were considered: *m-probability* ($m$) measuring the reliability of the field and *u-probability* ($u$) measuring the probability of a random agreement of values (Brown et al., 2017). For example, the probability of variable 'month of birth' agreed purely by chance for a linkage pair not belonging to the same individual was 1/12 (or 0.083). Hence the u-probability of this variable was set to be 0.083. The *u-probability* and *m-probability* probabilities were used to determine the (*dis*)*agreement weights*. Based on the match comparison, if $x_{b_i} = y_{b_i}$ *agreement* weight $log_2 (m/u)$ was added to the overall link weight $w(x, y)$, otherwise, *disagreement weight* $log_2 (1\text{-}m/1\text{-}u)$ was subtracted from the overall link weight $w(x, y)$.

3 Summarised the link weight from matching variables for records $x$, $y$ as overall match weight $w(x, y)$ and compared with the predefined cutoff weight $w$. If $w(x, y) \geq w$, $(x, y)$ was outputted as a matched pair and y was removed from the KPSC dataset. Otherwise, $y$ was kept in the KPSC dataset as candidates for following matching passes. Meanwhile, $x$ from the CCDF was be kept for all matching passes.

The total linkage weight (referred to as linkage weight thereafter) can range from a negative number (as a result of many disagreed matching variables) to a positive number (as a result of many agreed matching variables). We found that matched pairs with a linkage weight below 2.0 had a very small chance (less than 0.1%) of being a true death linkage. Therefore, 2.0 was set as the cutoff value to output potential matched pairs in our study. We also defined linkage weights between 2.0 and 4.9, between 5.0 and 9.9, and $\geq$ 10.0 as low, medium, and high link weight match pairs, respectively.

### *2.3.3 Post-linkage data processing*

Because the match was based on probabilities, the matched pairs were not necessarily true matches. We applied additional information available electronically to further eliminate matched pairs that were not likely to be true matches. The following exclusion criteria were applied to the matched pairs:

1 The date of birth in the KPSC dataset occurred later than the date of death in the CCDF.

2 The enrolee joined KPSC for the first time at least 1 month after the date of presumed death, and the link weight was less than 10.0.

3 The enrolee renewed the health insurance coverage at least 1 month after the date of presumed death, and the linkage weight was less than 5.0.

4 The individual had a face-to-face visit at one of the KPSC facilities for medical services at least 1 week after the presumed date of death, and the linkage weight was less than 10.0.

5 When multiple KPSC enrolees were found to match the same individual from the CCDF, the matches not having the highest link weight were dropped if the link weight was less than 5.0. Multiple KPSC enrolees were kept for users who apply mortality linkage results in their studies and could be further excluded by their study requirements.

Because the additional information used in the post-linkage data processing to further eliminate false positive matches becomes more complete over time, we reported the results based on two sets of 'additional data' representing the following two-time points.

1 April of 2017 (referred to as the Y2017 dataset)

2 April of 2020 (referred to as the Y2020 dataset)

The corresponding death records after the post-linkage data processing were referred to as D2017 and D2020, respectively. The April of 2020 dataset contained information on medical utilisation and health plan enrolment/disenrollment between April of 2017 and April of 2020, and thus, D2020 was expected to provide a higher positive predictive value (PPV) (i.e., less false positive matches), compared to D2017.

## *2.4 Data analysis and validation*

The linkage weights of the matched pairs were finally analysed by match pass number and age group at death before and after the post-linkage processing. Age at death was calculated by subtracting the date of birth in the KPSC dataset from the date of presumed death in the CCDF.

The linkage results before and after the post-linkage data processing were compared with the known deaths in 2015 recorded in KSPC systems. These included deaths that occurred at KPSC-owned facilities and at outside facilities which submitted medical claims to KPSC, or deaths reported to the KPSC Health Plan. The overall and linkage weight percentage of deaths found before and after the post-linkage process was calculated by the number of deaths found divided by the number of matched pairs.

To assess positive predictive value (PPV), a total of 400 deaths identified by the linkage process were randomly sampled from D2017 for manual reviews, with 100 deaths from each of the following linkage weight groups: 2.0–4.9, 5.0–9.9, 10.0–14.9 and 15.0 +. The manual review results were served as the gold standard. In each linkage weight group, the raw PPV is calculated by the confirmed true deaths divided by 100. The weighted PPV was calculated by the raw PPV and the sampling weight in each linkage weight group.

## *2.5 Linkage software and server*

The linkage process demonstrated in this study was performed on a Linux server with 4 CPUs of 3.07 GHz (2 cores) based on the IBM InfoSphere Information Server version 11.5.0. Data manipulation and analysis were conducted in SAS 9.4. Please visit online website for more information on the IBM InfoSphere information server (IBM, 2021).

**Table 1** Distribution of link weight in groups by match pass number and age at death before any post-linkage data processing

| | *Link weight (n, %)* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *2.0 – 4.9* | *5.0 – 9.9* | *10.0 – 14.9* | *15.0 – 19.9* | *20.0 – 24.9* | *25.0 – 29.9* | *30.0 – 34.9* | *35.0 – 39.9* | *40.0+* | *Overall* |
| Overall | 14,954 (40.0) | 4,928 (11.8) | 849 (2.0) | 515 (1.2) | 655 (1.6) | 439 (1.1) | 882 (2.1) | 8,963 (21.5) | 9,431 (22.7) | 41,616 |
| | *Match pass* | | | | | | | | | |
| 1 | 3 (0.0) | 2 (0.0) | 16 (0.1) | 49 (0.3) | 272 (1.4) | 398 (2.0) | 768 (3.9) | 8,812 (44.6) | 9,431 (47.8) | 19,751 (47.5) |
| 2 | 1,757 (38.8) | 1,282 (28.3) | 407 (9.0) | 403 (8.9) | 376 (8.3) | 41 (0.9) | 111(2.5) | 150 (3.3) | 0 (0.0) | 4,527 (10.9) |
| 3 | 2,033 (61.0) | 1,165 (34.9) | 111 (3.3) | 24 (0.7) | 0 (0.0) | 0 (0.0) | 1 (0.0) | 0 (0.0) | 0 (0.0) | 3,334 (8.0) |
| 4 | 7,369 (49.3) | 1,601 (32.5) | 311 (3.3) | 39 (0.4) | 6 (0.1) | 0 (0.0) | 2 (0.0) | 1 (0.0) | 0 (0.0) | 9,329 (22.4) |
| 5 | 312 (82.8) | 64 (17.0) | 1 (0.3) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 377 (0.9) |
| 6 | 2,661 (85.0) | 467 (14.9) | 3 (0.1) | 0 (0.0) | 1 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 3,132 (7.5) |
| 7 | 819 (70.2) | 347 (29.8) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 1,166 (2.8) |
| | *Age at death** | | | | | | | | | |
| < 1 | 193 (67.7) | 51 (17.9) | 9 (3.2) | 8 (2.8) | 12 (4.2) | 0 (0.0) | 0 (0.0) | 9 (3.2) | 3 (1.1) | 285 (0.7) |
| 1–12 | 224 (65.9) | 50 (14.7) | 12 (3.5) | 6 (1.8) | 7 (2.1) | 0 (0.0) | 3 (0.9) | 19 (5.6) | 19 (5.6) | 340 (0.8) |
| 13–17 | 237 (64.2) | 72 (19.5) | 8 (2.2) | 4 (1.1) | 2 (0.5) | 0 (0.0) | 4 (1.1) | 21 (5.7) | 21 (5.7) | 369 (0.9) |
| 18–24 | 539 (56.0) | 184 (19.1) | 33 (3.4) | 13 (1.4) | 15 (1.6) | 3 (0.3) | 16 (1.7) | 100 (10.4) | 59 (6.1) | 962 (2.3) |
| 25–44 | 2,075 (58.2) | 700 (19.6) | 117 (3.3) | 30 (0.8) | 32 (0.9) | 22 (0.6) | 44 (1.2) | 311 (8.7) | 234 (6.6) | 3,565 (8.6) |
| 45–74 | 9,480 (45.3) | 3,224 (15.4) | 546 (2.6) | 198 (1.0) | 203 (1.0) | 189 (0.9) | 300 (1.4) | 3,165 (15.1) | 3,628 (17.3) | 20,933 (50.3) |
| 75+ | 2,206 (14.8) | 647 (4.3) | 124 (0.8) | 256 (1.7) | 384 (2.5) | 225 (1.5) | 515 (3.4) | 5,338 (35.2) | 5,467 (36.1) | 15,162 (36.4) |

Notes: * Age at death was calculated by subtracting the date of birth in the KPSC data set from the date of presumed death in the CCDF.

**Table 2** Distribution of link weight in groups by match pass number and age at death after post-linkage data processing based on Y2017 dataset

| | *Link weight (n, %)* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *2.0 – 4.9* | *5.0 – 9.9* | *10.0 – 14.9* | *15.0 – 19.9* | *20.0 – 24.9* | *25.0 – 29.9* | *30.0 – 34.9* | *35.0 – 39.9* | *40.0+* | *Overall* |
| Overall | 1,171 (5.0) | 726 (3.1) | 846 (3.6) | 515 (2.2) | 655 (2.8) | 439 (1.9) | 882 (3.7) | 8,963 (37.9) | 9,431 (39.9) | 23,628 |
| | | | | | *Match pass* | | | | | |
| 1 | 0 (0.0) | 2 (0.0) | 16 (0.1) | 49 (0.3) | 272 (1.4) | 398 (2.0) | 768 (3.9) | 8812 (44.6) | 9431 (47.8) | 19,748 (83.6) |
| 2 | 148 (8.1) | 191 (10.5) | 406 (22.2) | 403 (22.1) | 376 (20.6) | 41 (2.3) | 111 (6.1) | 150 (8.2) | 0 (0.0) | 1,826 (7.7) |
| 3 | 154 (35.3) | 146 (33.5) | 111 (25.5) | 24 (5.5) | 0 (0.0) | 0 (0.0) | 1 (0.2) | 0 (0.0) | 0 (0.0) | 436 (1.8) |
| 4 | 534 (47.2) | 241 (21.3) | 309 (27.3) | 39 (3.5) | 6 (0.5) | 0 (0.0) | 2 (0.2) | 1 (0.1) | 0 (0.0) | 1,132 (4.8) |
| 5 | 19 (63.3) | 10 (33.3) | 1(3.3) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 30 (0.1) |
| 6 | 235 (75.8) | 71 (22.9) | 3 (1.0) | 0 (0.0) | 1 (0.3) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 310 (1.3) |
| 7 | 81 (55.5) | 65 (44.5) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 146 (0.6) |
| | | | | | *Age at death** | | | | | |
| < 1 | 3 (7.3) | 0 (0.0) | 6 (14.6) | 8 (19.5) | 12 (29.3) | 0 (0.0) | 0 (0.0) | 9 (22.0) | 3 (7.3) | 41 (0.2) |
| 1–12 | 21 (22.8) | 5 (5.4) | 12 (13.0) | 6 (6.5) | 7 (7.6) | 0 (0.0) | 3 (3.3) | 19 (20.7) | 19 (20.7) | 92 (0.4) |
| 13–17 | 26 (26.3) | 13 (13.1) | 8 (8.1) | 4 (4.0) | 2 (2.0) | 0 (0.0) | 4 (4.0) | 21 (21.2) | 21 (21.2) | 99 (0.4) |
| 18–24 | 94 (23.5) | 67 (16.8) | 33 (8.3) | 13 (3.3) | 15 (3.8) | 3 (0.8) | 16 (4.0) | 100 (25.0) | 59 (14.8) | 400 (1.7) |
| 25–44 | 242 (19.9) | 185 (15.2) | 117 (9.6) | 30 (2.5) | 32 (2.6) | 22 (1.8) | 44 (3.6) | 311 (25.6) | 234 (19.2) | 1,217 (5.2) |
| 45–74 | 700 (7.5) | 422 (4.5) | 546 (5.8) | 198 (2.1) | 203 (2.2) | 189 (2.0) | 300 (3.2) | 3,165 (33.9) | 3,628 (38.8) | 9,351 (39.6) |
| 75+ | 85 (0.7) | 34 (0.3) | 124 (1.0) | 256 (2.1) | 384 (3.1) | 225 (1.8) | 515 (4.1) | 5,338 (43.0) | 5,467 (44.0) | 12,428 (52.6) |

Notes: * Age at death was calculated by subtracting the date of birth in the KPSC data set from the date of presumed death in the CCDF.

**Table 3** Distribution of link weight in groups by match pass number and age at death after another round of post-linkage data processing based on Y2020 dataset

| | *Link weight (n, %)* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *2.0 – 4.9* | *5.0 – 9.9* | *10.0 – 14.9* | *15.0 – 19.9* | *20.0 – 24.9* | *25.0 – 29.9* | *30.0 – 34.9* | *35.0 – 39.9* | *40.0+* | *Overall* |
| Overall | 76 (0.4) | 44 (0.2) | 109 (0.5) | 420 (2.0) | 647 (3.1) | 438 (2.1) | 882 (4.2) | 8962 (42.7) | 9,431 (44.9) | 21,009 |
| | *Match pass* | | | | | | | | | |
| 1 | 0 (0.0) | 2 (0.0) | 16 (0.1) | 47 (0.2) | 271 (1.4) | 397 (2.0) | 768 (3.9) | 8,811 (44.6) | 9,431 (47.8) | 19,43 (94.0) |
| 2 | 13 (1.2) | 15 (1.3) | 69 (6.1) | 355 (31.5) | 372 (33.0) | 41 (3.6) | 111 (9.9) | 150 (13.3) | 0 (0.0) | 1,126 (5.4) |
| 3 | 11 (28.2) | 6 (15.4) | 12 (30.8) | 9 (23.1) | 0 (0.0) | 0 (0.0) | 1 (2.6) | 0 (0.0) | 0 (0.0) | 39 (0.2) |
| 4 | 39 (50.0) | 12 (15.4) | 12 (15.4) | 9 (11.5) | 3 (3.9) | 0 (0.0) | 2 (2.6) | 1 (1.3) | 0 (0.0) | 78 (0.4) |
| 5 | 0 (0.0) | 1 (100) | 0(0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 1 (0.0) |
| 6 | 7 (53.9) | 5 (38.5) | 0 (0.0) | 0 (0.0) | 1 (7.7) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 13 (0.1) |
| 7 | 6 (66.7) | 3 (33.3) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 9 (0.0) |
| | *Age at death** | | | | | | | | | |
| < 1 | 1 (3.0) | 0 (0.0) | 0 (0.0) | 8 (24.2) | 12 (36.4) | 0 (0.0) | 0 (0.0) | 9 (27.3) | 3 (9.1) | 33 (0.2) |
| 1–12 | 0 (0.0) | 0 (0.0) | 1 (1.9) | 5 (9.4) | 6 (11.3) | 0 (0.0) | 3 (5.7) | 19 (35.9) | 19 (35.9) | 53 (0.3) |
| 13–17 | 0 (0.0) | 1 (2.0) | 0 (0.0) | 3 (5.9) | 1 (2.0) | 0 (0.0) | 4 (7.8) | 21 (41.2) | 21 (41.2) | 51 (0.2) |
| 18–24 | 1 (0.5) | 1 (0.5) | 2 (1.0) | 8 (3.9) | 14 (6.9) | 3 (1.5) | 16 (7.8) | 100 (49.0) | 59 (28.9) | 204 (1.0) |
| 25–44 | 2 (0.3) | 6 (0.9) | 3 (0.5) | 20 (3.0) | 29 (4.3) | 22 (3.3) | 44 (6.6) | 311 (46.4) | 234 (34.9) | 671 (3.2) |
| 45–74 | 27 (0.4) | 21 (0.3) | 49 (0.6) | 134 (1.7) | 201 (2.6) | 189 (2.5) | 300 (3.9) | 3,165 (41.0) | 3,628 (47.0) | 7,714 (36.7) |
| 75+ | 45 (0.4) | 15 (0.1) | 54 (0.4) | 242 (2.0) | 384 (3.1) | 224 (1.8) | 515 (4.2) | 5,337 (43.5) | 5,467 (44.5) | 12,283 (58.5) |

Notes: * Age at death was calculated by subtracting the date of birth in the KPSC data set from the date of presumed death in the CCDF.

**Table 4** Validating the linkage results against the known deaths in 2015 recorded in the KPSC systems (n = 19620)

| | *Link weight* | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Not found** | *2.0–4.9* | *5.0–9.9* | *10.0–14.9* | *15.0–19.9* | *20.0–24.9* | *25.0–29.9* | *30.0–34.9* | *35.0–39.9* | *40.0+* | *Overall** |
| *Post-linkage processing based on Y2017 dataset* | | | | | | | | | | | |
| Deaths in D2017 | | 1171 | 726 | 846 | 515 | 655 | 439 | 882 | 8,963 | 9,431 | 23,628 |
| Deaths recorded in the KPSC systems | 486 | 11 | 9 | 49 | 357 | 571 | 382 | 756 | 8,215 | 8,784 | 19,620 |
| Deaths in D2017 but not in deaths recorded in the KPSC systems | | 1160 | 717 | 797 | 158 | 84 | 57 | 126 | 748 | 647 | 4,494 |
| Percentage of KPSC recorded deaths among algorithm identified deaths (%) | NA | 0.9 | 1.2 | 5.8 | 69.3 | 87.2 | 87.0 | 85.7 | 91.7 | 93.1 | 81.0 |
| *Post-linkage processing based on Y2020 dataset* | | | | | | | | | | | |
| Deaths in D2020 | | 76 | 44 | 109 | 420 | 647 | 438 | 882 | 8,962 | 9,431 | 21,009 |
| Deaths recorded in the KPSC systems | 488 | 11 | 9 | 49 | 357 | 571 | 381 | 756 | 8,214 | 8,784 | 19,620 |
| Deaths in D2020 but not in deaths recorded in the KPSC systems | | 65 | 35 | 60 | 63 | 76 | 57 | 126 | 748 | 647 | 1,877 |
| Percentage of KPSC recorded deaths among algorithm identified deaths (%) | NA | 14.5 | 20.5 | 45.0 | 85.0 | 88.3 | 87.2 | 85.7 | 91.7 | 93.1 | 91.1 |

Notes: * Deaths missed (not found) by the linkage process.
* The overall percentage of deaths found after post-linkage data processing 80.0% and 91.1% was calculated by (19620–486)/23628, (19620–488)/21009 respectively.

## 3 Results

### 3.1 Matched pairs before post-linkage processing

A total of 4,470,873 KPSC enrolee records and 260,217 death records from the CCDF were input into the IBM InfoSphere software for matching. A total of 41,616 matched records resulted from the seven matching passes. Table 1 shows the distribution of the link weight in groups by match pass and age at death before post-linkage processing. Overall, 48.2% had a high link weight (≥ 10.0), 11.8% had a medium link weight (5.0–9.9), and 40.0% had a low link weight (2.0–4.9). Almost all matched pairs contributed by pass 1 had a high link weight (≥ 10.0). Most of the matched pairs from passes 2 to 7 had either medium link weight (5.0–9.9) or low link weight (2.0–4.9)

### 3.2 Matched pairs after post-linkage processing

Table 2 summarises the distribution of link weight in groups by match pass and age at the death after the post-linkage processing based on the Y2017 dataset. A total of 17,985 matched pairs with link weight less than 10.0 were eliminated after the post-linkage data processing. Most of these eliminated matches were from passes 2 to 7. Only three linked pairs from the first pass with low weight (2.0–4.9) were eliminated. After the post-linkage processing, 92.0% of matched pairs had a high link weight (≥10.0), 3.0% had a medium link weight (5.0–9.9), and 5.0% had a low link weight (2.0–4.9). When stratified by age at death, 52.6% of matched pairs had died at age 75 years or older, 39.6% had died between 45 and 74 years of age, 5.2% between 25 and 44 years of age, 1.7% between 18 and 24 years of age, 0.8% between 1 and 17 years of age, and 0.2% were infant deaths. Table 3 summarises the distribution of link weight in groups by match pass and age at the death after the post-linkage processing based on the Y2020 dataset. Compared with the results based on the Y2017 dataset, an additional 2,619 matched pairs were eliminated.

### 3.3 Cross-validation and performance of finalised matched pairs

The cross-validation of the linkage results against the 19,620 deaths recorded in the KPSC mortality database is shown in Table 4. After the post-linkage process based on the Y2017 dataset, the linkage process identified 19,134 deaths (97.5%) and missed 486 (2.5%) of total deaths recorded in the KPSC systems. Of these identified deaths (n = 19,134), only 11 had a low link weight (2.0–4.9), 9 had a medium link weight (5.0–9.9), and the rest had a high link weight (≥ 10.0). In the low linkage weight (2.0–4.9) group, only 0.9% of deaths in D2017 were in deaths recorded in the KPSC. In the medium link weight (5.0–9.9) group, 1.2% of deaths in D2017 were in deaths recorded in the KPSC. While in the high link weight ( ≥ 10.0) group, 87.9% of D2017 were in deaths records in the KPSC. In addition, the linkage process identified additional 4,494 potential deaths, and 58.2% of these had a link weight ≥10.0. After the post-linkage data processing based on the Y2020 dataset, the linkage process missed two additional true deaths (increased from 486 to 488). The potential additional deaths identified by the linkage process were reduced to 1,877, and 94.7% of these had a link weight ≥ 10.0.

Table 5 summarises the performance of the linkage process against the manual review results. The overall weighted PPV after the post-linkage process based on the Y2017

dataset was 88.7% and was increased to 99.4% after the post-linkage process based on the Y2020 dataset.

**Table 5** Estimation of positive prediction value (PPV)

| | *Link weight* | | | | |
|---|---|---|---|---|---|
| | *2.0–4.9* | *5.0–9.9* | *10.0–14.9* | *15.0+* | *Weighted PPV£* |
| Validation samples: | | | | | |
| Deaths sampled from D2017 | 100 | 100 | 100 | 100 | |
| Sampling weight* (%) | 5.0 | 3.1 | 3.6 | 88.4 | |
| Deaths remained in D2020 | 4 | 9 | 11 | 100 | |
| Sampling weight* (%) | 0.4 | 0.2 | 0.5 | 98.9 | |
| Deaths verified by manual review | 1 | 3 | 7 | 100 | |
| PPV: | | | | | |
| Based on D2017 (%) | 1.0 | 3.0 | 7.0 | 100.0 | 88.7 |
| Based on D2020 (%) | 25.0 | 33.3 | 63.6 | 100.0 | 99.4 |

Notes: * Sampling weight (%) in deaths sampled from D2017 is calculated by 100% × 1,171/23,628, 100% × 726/23,628, 100% × 846/23,628, 100% × 20,885/23,628 for the four link weight groups. Sampling weight (%) in deaths sampled from D2020 is calculated by 100% × 76/21,009, 100% × 44/21,009, 100% × 109/21,009, 100% × 20,780/21,009 for the four link weight groups. £ Weighted PPV is the summation of the product of PPV and the sampling weight from each linkage weight group. The weighted PPV 88.7% is calculated by (1.0% × 5.0%+3.0% × 3.1%+7.0% × 3.6%+100.0% × 88.4%). Add the same for 99.4.

## 4 Discussion

### *4.1 Study findings*

In this study, we demonstrated the use of the IBM InfoSphere software by designing a comprehensive algorithm and process to link decedents in the California comprehensive death file (CCDF) with health plan enrolees of KPSC. The linkage process has successfully identified more than 20k potential deaths of health plan enrolees in 2015. 97.5% of the true deaths recorded in the KPSC internal systems can be found by the linkage process.

Although there was only 5.5% of health plan enrolees who did not have information on SSN, the information on known SSN may not be accurate in both data sources. In the current study, after the post-linkage process based on Y2017, we identified a total of 21,731 matches with a high likelihood to be true matches (link weight > = 10), and 90.9% of them matched exactly on SSN. The linkage process identified an additional 9.1% high link weight match not based on SSN. Meanwhile, after the post-linkage process based on Y2020, we identified a total of 20,889 matches with a high likelihood to be true matches (link weight > = 10), and 94.5% of them matched exactly on SSN. The linkage process identified an additional 5.5% high link weight match not based on SSN.

The performance of the current study was consistent with those reported previously based on the predecessors of the software. For example, in the late 1990s, the National

agricultural statistics service (NASS) utilised AutoMatch and AutoStan to link PLMA and Ohio list files (Kara, 1996). The AutoMatch achieved a false match error rate of 1.1 % and a false nonmatch error rate of 4.9 % (Kara, 1996). In a study of the health effect of potentially less hazardous cigarettes involving 4,696 Kaiser Permanente members, the linkage utilising CAMLIS achieved a sensitivity of 89.0% and 0.1% of death were missed by CAMLIS (Arellano et al., 1984).

### *4.2 Linkage pair cutoff values and post-linkage reprocessing*

Selecting a reasonable threshold cutoff value of link weight is critical to maintaining the balance between the number of false matches and the number of missed matches (Krewski et al., 2005). But choosing an effective threshold is not straightforward and typically requires manual review of a subset of matched and unmatched pairs (Harron et al., 2017). In the current study, our validation of the data linkage results against the internal death records showed that the chance of missing any true deaths was very small (near or less than 0.1%) as long as the link weight was greater than 2.0. Therefore, we are confident that the cutoff value of 2.0 was acceptable for our study purpose. However, such a low cutoff value brought in a large number of false matches with link weights between 2.0 and 9.9. To remove the false-positive match pairs, we applied a post-linkage process based on available medical utilisation and health plan enrolment/disenrollment in April of 2017 that removed 92.2%, 85.3%, and 0.01% of the low (2.0–4.9), medium (5.0–10.0), and high (≥ 10.0) weight match pairs. We applied another post-linkage process based on available medical utilisation and health plan enrolment/disenrollment between April of 2017 and April of 2020 that further removed 93.5%, 93.9%, and 3.9% of the low (2.0–4.9), medium (5.0–10.0), and high (≥ 10.0) weight match pairs. We recommend inspecting matched records with linkage weights of less than 15.0 manually before making inferences about deaths based on validation results shown in Table 5. In addition, we also recommend that users select a reasonable cutoff value that addresses both sensitivity and specificity if information for conducting post-linkage processing is not available.

False positive matches could largely impact the quality of research. We demonstrated that using additional information available in three years after the initial linkage process could improve the PPV from 88.7% to 99.4%. Therefore, we strongly recommend users who conduct any data linkage to update results using the most recently available information on an ad hoc or regular basis. Given the number of known deaths in 2015 (n=19,620) documented within the KPSC internal systems, an addition of 1,877 deaths identified by the linkage process has the potential to increase the sensitivity by about 9.0% for any research studies when death was an outcome of interest or a censoring event.

### *4.3 String comparisons*

The IBM InfoSphere software provides several methods for string comparisons, such as MULT_EXACT, MULT_UNCERT, and NAME_UNCERT, etc. (IBM, 2021). Our study applied NAME_UNCERT comparison to match first names and last names to capture all possible matches included the partially matching. There are some other methods for string comparisons (Winkler, 2006). Fellegi and Sunter (1969) provided some intuition on how to get crude estimates of typographical error rates. Jaro–Winkler string

comparator (Winkler, 1990) enhancing the decision rules in the Fellegi-Sunter model improved matching efficacy in comparison to situations when the string comparators are not used. Cohen et al. (2003) investigated several different string distance metrics including edit-distance metrics, fast heuristic string comparators, token-based distance metrics, and hybrid methods for name matching. Overall, the best-performing method was a hybrid scheme combining a term frequency-inverse document frequency weighting scheme with the Jaro-Winkler string-distance scheme. Unfortunately, these methods were not available in the IBM InfoSphere software and thus beyond the scope of the current study.

Deterministic and probabilistic approaches are two main types of linkage process that have been successfully implemented in previous studies (Arellano et al., 1984; Choi et al., 2017; Dusetzina et al., 2014; Fair et al., 2000; Gidding et al., 2017; Kara, 1996). Deterministic linkage compares an identifier or a group of identifiers across databases and a link is made if they all agree. The deterministic approach ignores the fact that certain identifiers or certain values have more discriminatory power than others do. On the other hand, probabilistic strategies assess:

1 the discriminatory power of each identifier

2 the probability that two records are a true match based on whether they agree or disagree on various identifiers.

Public or commercial software based on these two approaches was summarised in the review paper by Dusetzina et al. (2014). However, testing and comparison of these applications are beyond the scope of the current study.

### *4.4 IBM InfoSphere software issues*

Our study found that the name standardisation procedure in the IBM InfoSphere software (version 11.5.0) may cause some name reverse soundex errors. in some cases, the usname rules within the software did not recognise the part of the first name and treated it as part of the last name. as a result, the first reverse soundex of the last name was actually the reverse soundex of the first name, as shown in the CONCEPCION LEE example in the ‘data preparation’ section. When the data linkage process used the reverse soundex of the last name as the blocking variable and the first name as a matching variable, the first name actually contributed to both the matching variable and blocking variable when the reverse soundex misclassification of the last name occurred. Therefore, the link weight for the corresponding match was highly inflated. The error could occur similarly for the reverse soundex of the first name. Because our linkage process used both the reverse soundex of the last name and the reverse soundex of the first name, our study designed an additional pre-processing step as described in the Methods section to avoid these errors without changing the default USNAME rules. We urge the developer of the software to resolve this issue in future software releases and recommend that, until a solution is provided by the software, users take an approach similar to ours if the reverse soundex of names or standardise names are used.

### *4.5 Study limitations*

We acknowledge several potential limitations of our study. First, our data linkage process used a limited set of variables to conduct the probabilistic linkage, and the performance of the linkage process relied mainly on the data quality of these matching variables. If the information coded by the matching variables is incorrect, incomplete, or missing, then the data records could also be linked incorrectly. Second, our study applied a post-linkage process to eliminate false-positive match pairs by using medical utilisation and health plan enrolment/disenrollment. The availability and quality of the information are important to ensure accurate results. Third, the linkage process missed about 2.5% of true deaths compared with the individual deaths collected from the internal systems. These cases were missed because of:

1 incorrect or incomplete individual information, for example, some neonatal deaths may have an incomplete first name

2 individuals who died outside the State of California and who therefore were not in the CCDF.

Last, our linkage process was applied to enrolees with at least one day of enrolment in 2015. Extension of this linkage process to individuals who were unenrolled or terminated from the health plan poses additional challenges because updated information may not be available for these people.

### *4.6 Implementation and automation*

It is worthwhile to point out that the linkage algorithm has no specific data collected time requirement for the matching process as long as the two data sources are available. But the performance can vary because the post-linkage steps of the algorithm apply additional available information after the death date to eliminate potential false matches. The performances can be improved as the time lag between the two data sources is large. Therefore, users who conduct any data linkage based on real-time records or closed time window of the two death sources should consider update results later using the most recently available information. In addition, although the linkage process was built based on mortality data, a similar approach can be applied or adapted for other types of data linkage. Once the linkage algorithm is finalised, it can be easily set up as an automated process to run sequentially on an ad hoc or regular basis.

## 5 Conclusions

We demonstrated the application of commercial software, the IBM InfoSphere, to link large volumes of data from multiple data sources. The linkage process demonstrated high accuracy and can be used to support various business needs. Users who conduct mortality data linkage should consider updating results using the most recently available information. Users who use the name standardisation function within the IBM InfoSphere software should be aware of the errors related to the name reverse soundex. In addition, users should consider the NAME_UNCERT function offered by the software to increase the accuracy of matching.

## References

Alonso-Sardón, M., Iglesias-de-Sena, H., Sáez-Lorenzo, M. et al. (2015) 'B-learning training in the certification of causes of death', *Journal of Forensic and Legal Medicine*, Vol. 29, pp.1–5.

Alur, N., Jha, A.K., Rosen, B. and Skov T. (2008) *IBM WebSphere QualityStage Methodologies, Standardization, and Matching, Poughkeepsie*, IBM Corp., NY.

Arellano, M.G., Petersen, G.R., Petitti, D.B. et al. (1984) ''The California automated mortality linkage system (CAMLIS)', *American Journal of Public Health*, Vol. 74, No. 12, pp.1324–1330.

Benjamin, E.J., Blaha, M.J., Chiuve, S.E. et al. (2017) 'Heart disease and stroke statistics–2017 update: a report from the American heart association', *Circulation*, Vol. 135, No. 10, pp.e146–603.

Bohensky, M.A., Jolley, D., Sundararajan, V. et al. (2010) 'Data linkage: a powerful research tool with potential problems', *BMC Health Service Research*, Vol. 10, No. 1, p.346.

Brown, A.P., Randall, S.M., Ferrante, A.M. et al. (2017) 'Estimating parameters for probabilistic linkage of privacy-preserved datasets', *BMC Medical Research Methodology*, Vol. 17, No. 1, p.95.

California Department of Public Health (2020) *Center for Health Statistics and Informatics. Comparison of California Death Data Sources* [online] https://data.chhs.ca.gov/dataset/statewide-death-profiles (accessed 12 July 2020).

CDC (Centers for Disease Control) (2021) *Link Plus* [online] https://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm (accessed 3 September 2021).

Choi, S.T., Lin, Y. and Mulrow, E. (2017) 'Comparison of public-domain software and services for probabilistic record linkage and address standardization', in Holzinger, A., Goebel, R., Ferri M., Palade, V. (Eds.): *Towards Integrative Machine Learning and Knowledge Extraction*, Vol. 10344, Lecture Notes in Computer Science, Springer, Cham.

ChoiceMaker Technologies (2021) *ChoiceMaker 2* [online] https:// www.sourceforge.net/projects/oscmt/ (accessed 3 September 2021).

Christen, P. and Goiser, K. (2007) 'Quality and complexity measures for data linkage and deduplication', in Guillet, F. and Hamilton, H. (Eds.): *Quality Measures in Data Mining*, Vol. 134, Studies in Computational Intelligence, Springer, Berlin.

Cohen, W., Ravikumar, P. and Fienberg, S. (2003) 'A comparison of string metrics for matching names and records', in *KDD Workshop on Data Cleaning and Object Consolidation*, Vol. 3, pp.73–78.

Dunn, H.L. (1946) 'Record linkage', *American Journal of Public Health*, Vol. 36, No. 12, pp.1412 –1416.

Dusetzina, S.B., Tyree, S., Meyer, A.M. et al. (2014) 'Linking data for health services research: a framework and instructional guide [Internet]', *An Overview of Record Linkage Methods*, Agency for Healthcare Research and Quality (US), Rockville (MD).

Fair, M. (2004) 'Generalized record linkage system – statistics canada's record linkage software', *Australian Journal Statistics*, Vol. 33, No. 1, pp.37–53.

Fair, M., Cyr, M., Allen, A.C. et al. (2000) 'Fetal and infant health study group, an assessment of the validity of a computer system for probabilistic record linkage of birth and infant death records in Canada', *Chronic Disease Canada*, Vol. 21, No. 1, pp.8–13.

Fellegi, I.P. and Sunter A.B. (1969) 'A theory for record linkage', *Journal of American Statistics Association*, Vol. 64, No. 328, pp.1183–1210.

Gidding, H.F., McCallum, L., Fathima, P. et al. (2017) 'Probabilistic linkage of national immunization and state-based health records for a cohort of 1.9 million births to evaluate Australia's childhood immunization program', *International Journal of Population Data Science*, Vol. 2, No. 1, pp.1–13.

Go, A.S., Chertow, G.M., Fan, D. et al. (2004) 'Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization', *New England Journal of Medicine*, Vol. 351, No. 13, pp.1296–1305.

Harron, K., Dibben, C., Boyd, J. et al. (2017) 'Challenges in administrative data linkage for research', *Big Data and Society*, Vol. 4, No. 2, DOI: 2053951717745678.

Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007) 'Review of record linkage software', in *Data Quality and Record Linkage Techniques*, pp.201–207, Springer, New York, NY.

IBM (2021) *IBM InfoSphere Information Server Version 11.5.0 documentation* [online] https://www.ibm.com/docs/en/iis/11.5; *Match Comparisons* [online] https://www.ibm.com/docs/en/iis/11.7?topic=stages-match-comparisons (accessed 12 July 2021).

Kara, B. (1996) *Record Linkage III: Experience Using AUTOMATCH for Record Linkage in a State Office Setting*, US Department of Agriculture, Washington, DC.

Karr, A.F., Taylor, M.T., West, S.L. et al. (2019) 'Comparing record linkage software programs and algorithms using real-world data', *PloS One*, Vol. 14, No. 9, DOI: e0221459.

Koebnick, C., Langer-Gould, A.M., Gould, M.K. et al. (2012) 'Sociodemographic characteristics of members of a large, integrated health care system: comparison with US census bureau data', *Permanente Journal*, Vol. 16, No. 3, pp.37–41.

Krewski, D., Dewanji, A., Wang, Y. and et al. (2005) 'The effect of record linkage errors on risk estimates in cohort. Mortality Studies', Survival Methodology, Vol. 31, pp.13 - 21.

Mamun, A., Aseltine, R. and Rajasekaran, S. (2016) 'Efficient record linkage process using complete linkage clustering', *PLoS One*, Vol. 11, No. 4, DOI: e0154446.

Playford, C.J., Gayle, V., Connelly, R. and Gray, A.J.G. (2016) 'Administrative social science data: the challenge of reproducible research', *Big Data Society*, Vol. 3, No. 2, pp.1–13.

Randall, S.M., Ferrante, A.M., Boyd, J.H. and Semmens, J.B. (2013) 'The effect of data cleaning on record linkage quality', *BMC Medical Informatics and Decision Making*, Vol. 13, No. 1, pp.64.

Sorlie, P.D., Backlund, B.J. and Keller, J.B. (1995) 'US mortality by economic, demographic, and social characteristics: the national longitudinal mortality study', *American Journal of Public Health*, Vol. 85, No. 7, pp.949–956.

Washington State, (2021) *Division of Alcohol and Substance Abuse*, *the Link King* [online] https://http://www.the-link-king.com (accessed 3 September 2021).

Winkler, W.E. (1990) 'String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage', *Proceedings of the Section on Survey Research*, Washington, DC, USA, pp.354–359.

Winkler, W.E. (2006) *Overview of Record Linkage and Current Research Directions*, Bureau of the Census.

Wunsch, G. and Gourbin, C. (2018) 'Mortality, morbidity and health in developed societies: a review of data sources', *Genus*, Vol. 74, No. 1, p.2.

## Appendix

**Supplementary Table 1** Errors introduced by standardisation and name sound index

| *Scenarios* | *First name* | *Last name* | *Standardised first name* | *Standardised middle name* | *Standardised last name* | *First Name sound index* | *Last name sound index 1* | *Last name sound index 2* |
|---|---|---|---|---|---|---|---|---|
| 1 | A | B | NULL | NULL | S(A)<br>S(B) | NULL | SI(A) | SI(B) |
| 2 | A | B | NULL | NULL | S(B) | NULL | SI(B) | NULL |
| 3 | A | B C | S(A) | S(B) | S(C) | SI(A) | SI(C) | NULL |
| 4 | A | B C | NULL | NULL | S(A)<br>S(B)<br>S(C) | NULL | S(A) | S(B) |
| 5 | A B | C | S(A) | S(B) | S(C) | SI(A) | SI(C) | NULL |
| 6 | A B | C | S(A) | NULL | S(B)<br>S(C) | SI(A) | SI(B) | SI(C) |
| 7 | A B | C | NULL | NULL | S(A)<br>S(B)<br>S(C) | NULL | SI(A) | SI(B) |

Notes: A, B, and C denote individual names (e.g., MICHEAL, JOHN). S(A), S(B), and S(C) represent the corresponding standardised names. SI(A), SI(B), and SI(C) symbolise the corresponding sound indexes. NULL means empty or no return value, indicating that the software system is unable to recognise the name as a valid one.

The first name and last name columns are the original first name and last name submitted for standardisation. The middle name is not shown in this table because including the middle name in the input for standardisation will result in more complex error scenarios in the name sound indexes. Only the initials of the middle name are kept for linkage. Words in the first and the last name are separated by a space. The next 3 columns are the standardised first name, standardised middle name, and standardised last name, respectively. For example, the standardised name of BESSIE or BETTY is ELIZABETH. If the USNAME rules are unable to standardise a first name or a last name, the corresponding standardised name field will be assigned NULL. The far right columns represent the sound index of the standardised first name (First name sound index), the sound index of the first word in the standardised last name (Last name sound index 1), and the sound index of the second word in the standardised last name (Last name sound Index 2), respectively. For example, the sound index of ROBERTA is A361. The sound index of BESSIE and BETTY are the same because they are first standardised into ELIZABETH.

10% of the name standardisation steps result in errors in name sound indexes, and around 95% of these errors come from scenarios 1 and 4.

1 In some cases, the IBM InfoSphere information server software is unable to standardise both first name and last name. In these cases, the returned standardised first and last name and all sound indexes are null. These cases are not listed in the table above.

2 In scenarios 1, 4, and 7, the first name is not recognised as a valid first name. Instead, the standardised last name combines both first name and last name. In these cases, two sound indexes are generated for the standardised last name. These cases were the main cause of name misclassification in our dataset.

3 In scenario 2, the software is unable to standardise the first name and standardises the last name only.

4. In scenario 3, the software incorrectly identifies the first word of the last name as the middle name and keeps only the second word of the last name in the standardised last name and the corresponding last name sound index.

5 In scenario 5, the software incorrectly identifies the second word of the first name as the middle name and keeps only the first word of the first name in the standardised first name.

6 In scenario 6, the software incorrectly identifies the second word of the first name as the first word of the standardised last name and the last name as the second word of the standardised last name. This also results in the sound index of the second word of the first name being treated as the first sound index of the standardised last name and the sound index of the last name being treated as the second sound index of the standardised last name.

**Supplementary Table 2** Block variables and matching variables for each pass

| *Pass number* | *Block variables* | *Matching variables* |
|---|---|---|
| 1 | SSN | DOBYYMM, DOBDD, LNAME, FNAME, MI, SEX, RACE CTY |
| 2 | DOBYYMM, LNAME | SSN, DOBDD, FNAME, MI, SEX, RACE, CTY |
| 3 | SLUSNAM, DOBYYMM | SSN, DOBDD, FNAME, MI, SEX, RACE, CTY |
| 4 | SFUSNAM, DOBYYMM | SSN, DOBDD, LNAME, MI, SEX, RACE, CTY |
| 5 | LNAME, FNAM3, DOBYY | SSN, DOBMM, DOBDD, MI, SEX, RACE, CTY |
| 6 | LNAME, FNAM3, DOBMM | SSN, DOBYY, DOBDD, MI, SEX, RACE, CTY |
| 7 | MI, FNAME, DOBYY | SSN, DOBMM, DOBDD, LNAME, SEX, RACE, CTY |

Notes: SSN = social security numberDOBDD = day of birth date
DOBMM = month of birth date DOBYY = year of birth date
DOBYYMM = concatenate year and month of birth date
FNAME = first name LNAME = last name
MI = initial of middle name FNAM3 = first three letters of first name
SLUSNAM = sound index of last name SFUSNAM = sound index of first name
SEX = gender CTY = residence county code RACE = race and ethnicity cod.

**Supplementary Table 3** Match comparisons for matching variables

| *Matching variable* | *Match comparison* |
|---|---|
| DOBYYMM | CHAR |
| DOBDD | CHAR |
| LNAME | NAME_UNCERT |
| FNAME | NAME_UNCERT |
| MI | CHAR |
| SEX | CHAR |
| RACE | CHAR |
| CTY | CHAR |
| DOBMM | CHAR |
| DOBYY | CHAR |
| SSN | UNCERT |

Notes: Description of match comparisons: CHAR: Compares data values on a character-by-character basis. This comparison is often used to catch spelling mistakes or inverted letters. UNCERT: Evaluates the similarity of two character strings based on the string length, the number of transpositions, and the number of unassigned insertions, deletions, or replacement of characters between the two strings. NAME_UNCERT: Compares two strings. First, it right-truncates the longer string so that it contains the same number of characters as the shorter string. If that comparison is not an exact match, it evaluates the similarity of the strings by doing an UNCERT comparison.