

International Journal of Population Data Science

Journal Website: www.ijpds.org



Swansea University
Prifysgol Abertawe

Lessons in Linkage: combining administrative data using deterministic linkage for surveillance of sports and recreation injuries in Florida, United States

Charlotte Baker^{1,*}, Quinton Nottingham², and Jonathan Holloway¹

Submission History

Submitted:	03/03/2022
Accepted:	08/06/2022
Published:	26/07/2022

¹Virginia Polytechnic Institute and State University, Virginia-Maryland College of Veterinary Medicine, Department of Population Health Sciences, 205 Duck Pond Drive (0442), Blacksburg, Virginia, USA 24061-0442

²Virginia Polytechnic Institute and State University, Pamplin College of Business, Department of Business Information Technology, 880 W Campus Drive (0235), Blacksburg, Virginia, USA 24061-0235

Abstract

Objectives

Previous and ongoing epidemiological surveillance of sports and recreation injuries (SRI) has been cross-sectional in nature, utilised a subset of injuries based on athletic trainer availability, or focused on elite and professional athletes. In the United States, surveillance is often prohibitively expensive and not well funded by national organisations or agencies, except for the case of some professional and elite sports. This paper details the methodology, barriers, and successes of using deterministic linkage to combine emergency department and hospitalisation data with a single identifier for use in surveilling sports injuries for persons aged 5 to 18 years.

Design

Data linkage of a population cohort.

Methods

We performed deterministic linkage of administrative emergency department and hospitalisation data from the state of Florida in the US. Data was acquired from the Florida Agency for Health Care Administration. With limited identifiers available due to privacy, we combined data across multiple years using a near universal identifier. We identified sport and recreation injuries using a modified External Cause of Injury Morbidity Matrix and ICD codes across all possible diagnoses. Finally, we obtained descriptive statistics of records that were successfully linked and those that were not to assess similarities between the groups.

Results

We found 384,731 visits for SRI over a seven-year period. We were able to link approximately 70% of the records using a single identifier. There were statistically significant differences by age, sex, payer, and race/ethnicity for the records that were linked compared to the records that were not linked.

Conclusions

This study is significant because while similar methods have been used to examine other conditions (e.g. asthma), few have linked multiple types of administrative data especially with nearly no identifiers to examine sports and recreation injuries. This method was found useful to identify injuries over time for the same individuals seeking care in emergency departments, or in hospital inpatient settings, though future work will need to address the limitations of this method. If we expect to move health surveillance forward as budgets for it become even more limited, we must develop and improve methods to do it with fewer resources, including using data that has great limitations.

Keywords

deterministic linkage; epidemiological methods; epidemiology; surveillance; public health; data cleaning; data linkage; reformatting; SAS; Python; injury surveillance; disease surveillance; health surveillance; biostatistics

*Corresponding Author:

Email Address: bakerc@vt.edu (Charlotte Baker)

Introduction

De-identified hospital and emergency department data are often used when assessing the health of a population. This is advantageous when examining the prevalence of health problems because each visit is treated as a discrete occurrence, but disadvantageous when examining the issues such as the effect of repeat injuries on individuals (person-level). In certain situations, we have access to identifiable information and this permits a different type of epidemiological assessment that is often limited to primary data collection. It becomes possible to identify how many encounters an individual has with the medical system over time and a better understanding of the variance in health care in the population based on patient characteristics such as prior health history. This could be useful in determining, for example, how primary care physicians can better intercede with patients to reduce the number of preventable sports injuries.

Involvement in sport and recreation promotes health and overall well-being [1, 2]. However, these activities expose individuals to a variety of elements that heighten risk for injury. Studies indicate millions of sport and recreation related injuries (SRI) are treated in emergency departments (ED) each year [3–8] and the incidence and severity of injuries within sport continue to garner increased attention from prevention professionals, the media, and participants [9–11]. Data gathered from EDs often represent the best available resources for exploring SRIs across a broad geographical region. Administrative records from EDs have been utilized to establish the prevalence of SRIs and common mechanisms of injury among different populations [8, 12–16]. In the United States, all fifty states and Washington, DC capture administrative data related to clinical visits, though the information included from each state varies in depth and public availability. Despite continuing work to explore both the risk factors for SRIs as well as the nature and severity of these injuries, limited information exist regarding the burden of these injuries on the healthcare system, particularly the burden associated with recurring injuries to the same individual.

Public health surveillance is the ongoing systematic collection of data for quantifying the burden of a problem (and thus developing priorities for addressing it); for monitoring trends of an issue; for identifying new problems; and for the evaluation of prevention efforts [17]. Surveillance can take place at various levels, e.g. national, state, and local community, and can consider both geographic, demographic, and policy differences between groups. Surveillance provides the information necessary to change or support policy at an appropriate intervention level. It also identifies where more resources are necessary to make a difference. Some issues, such as sports and recreation injuries, require various interventions at multiple policy levels for prevention and evaluation. Previous and ongoing epidemiological surveillance of SRI tends to be cross-sectional in nature [18, 19], be collected for a limited geographic area [20], utilise a subset of injuries based on athletic trainer availability [21], or focus on elite and professional athletes [22–26]. This surveillance is necessary to better identify what injuries and other health concerns are happening to sports and recreation participants, identify factors for intervention, and test prevention programs [27–29]. But surveillance is also often prohibitively expensive

[30–33] and, in the United States, not well funded by national organizations or agencies except for the case of some professional and elite sport.

This problem is not without potential solutions. Multiple countries have identified that we should work more often to use existing, or secondary, data for injury surveillance and prevention [34]. It has also been recommended that we improve or establish injury surveillance systems given the high burden of morbidity and mortality due to unintentional injuries [34]. A sustainable injury surveillance system that captures the same or most of the same information as traditional active, passive, or syndromic surveillance could be a feasible way to reduce costs without reducing quality. Optimally, this alternative would also allow data to be combined with other data sources such as community level, census, education, or geographic data to provide richer context for health findings. Health administrative data, such as records about emergency department discharges, is an existing data source that is regularly updated. This type of data is used in the United States for required reporting, for research, and for policy development. Discharge data has been used to surveil various health issues [35] including diabetes [36], asthma [37], and arthritis [38]. The purpose of this paper is to describe the data linkage methodology we used to establish a surveillance data set for SRI using discharge data in a single state of the United States. We expect that this strategy can be applied to a variety of data in and outside of the US to study issues in SRI.

Methods

This paper follows the guidelines of the RECORD statement (REporting of studies Conducted using Observational Routinely-collected health Data), an extension of the STROBE statement (Strengthening the Reporting of Observational Studies in Epidemiology) [39] (<https://www.equator-network.org/reporting-guidelines/record/>).

Institutional review board approval

This study was approved by the Florida Agricultural and Mechanical University Institutional Review Board (#014-47). A HIPAA Waiver of Authorization was granted to permit the research without needing individual patient participation approval.

Data sources

In Florida, data from hospitals, ambulatory care centers, and EDs is captured on a quarterly basis by the Agency for Health Care Administration (AHCA) [40]. Inpatient data has been available since 1988, and ED data since 2005. The inpatient data includes short- and long-term acute care hospitals and psychiatric hospitals. The ED data includes hospitals that have EDs but does not include freestanding EDs. Freestanding EDs are licensed EDs that are physically separated from a hospital or other clinical facility and can be independently operated or as an offsite or satellite facility of a hospital [41]. Only facilities affiliated with hospitals are recognized by the Centers for Medicare and Medicaid Services (CMS), meaning

independently operated (non-hospital affiliated facilities) can be more selective about who they serve and cannot bill Medicaid (a primarily income-based insurance program) or Medicare (a primarily age-based insurance program). A listing of all independently owned and operated facilities is available from AHCA. By default, the AHCA data includes all discharges from non-independent inpatient, and ED facilities no matter the state or national residency of the patient. Diagnosis information in all AHCA data is codified using the International Classification of Diseases (ICD), specifically the United States Clinical Modification (CM). Procedure details are coded using Current Procedural Terminology (CPT) codes. A listing of variables captured can be found for each data set on the Florida Health Finder website for researchers using AHCA data [40] and in the Data Catalog [42]. Details on more secure variables such as identifiers are available after contacting the Agency. All variables included in this project can be found in Table 1.

Identifying a target population

Every data linkage project should clearly define the population under study. For surveillance purposes, the population source needs to be consistent over time and inclusive of all events in that time and space. This is in line with three of the criteria for a successful surveillance system – quality data, stable, and representative [43]. Optimally, a census (a complete enumeration of a population) rather than a sample (finite part of a statistical population) should be used to identify the prevalence (burden) of the health problem. Not only is it necessary to capture the universe of potential visits, it is imperative to ensure that the population under surveillance is consistent across databases used for linkage. The goal of this project was to create a data set that was representative of Florida residents aged 5 to 18 years of age (childhood/adolescence) seen in an ED or hospital with an unintentional SRI (Figure 1). Age in this study is captured as the patient age on the date of ED attendance or hospital admission. The age range was selected in this project to best align with childhood and adolescence in the US as aligned with typical primary and secondary school attendance.

A major step to assess data consistency is to determine the availability and format of data across time. As expected, the variables collected and the format of how those variables were collected/encoded changed in the duration of each AHCA database's (i.e. inpatient, ED) existence. The type of errors that this can cause are explained in detail here [44].

There are more codes for SRI in ICD-10-CM than exist in ICD-9-CM. The increased level of coding specificity allows for better capture of injuries, yet affects the comparability of rates across time periods. Future changes to the ICD (such as the use of ICD-11-CM) can create a similar problem and needs to be accounted for. As we started this work, the United States was using ICD-9-CM and migrating toward the use of ICD-10-CM. Now, that switchover is complete and if used for surveillance or for simply examining issues within this population, whatever system is created needs to be functional and easy to update. Due to the vast differences in the two ICD systems available (9 vs 10), this paper focuses specifically on our general linkage methodology using ICD-9-CM data. Future work will illustrate the methodology using ICD-10-CM and how

we approached linking data across ICD versions. Few, if any, data sources to date have converted older data in ICD-9-CM to ICD-10-CM or ICD-10 to ICD-11 though some crosswalks exist for these specific purposes [45–48].

Strategy to establish the study population

When linking data, an initial action should be distinguishing what the final data set should look like – from what variables and observations should be included to how they should be included and formatted. The RECORD Statement recognizes “three levels of population hierarchy that are relevant in studies using routinely collected data” – the source population [the population that the data comes from], the database population [people with data in the data source], and the study population [those identified within the database population that fit the study criteria] [39] (Figure 1). As previously stated, our goal was to create a data set that was representative of Florida residents aged 5 to 18 years of age injured during sport or recreational activities (*source population*). There currently is no surveillance system that completely tracks this population and measures every health issue or outcome. However, we had access to administrative data that could be used to address the topic. This data – hospitalisation and ED data from AHCA in Florida – is known as the *database population*. We further narrowed the database population down to the *study population* using a variety of ICD codes and tools discussed below. We were most concerned with keeping the information for each visit a person had to an ED or hospital during the period January 1, 2006 to December 31, 2012.

Data linkage often includes two steps – merging and concatenation. Merging refers to putting data together side by side that might be from the same or different sources; as an example, one data set contains demographic information and the other data set contains visit information for those same individuals. There are several types of merges. If one data set contains multiple rows for the same individual and other data has a single row, this is referred to as a ‘many to one’ or ‘one to many’ merge depending on which set is the ‘main’ data set. If both sets only have one row per individual, this is a ‘one to one’ merge. When merging data, if columns have the same name, same named columns will often be deleted.

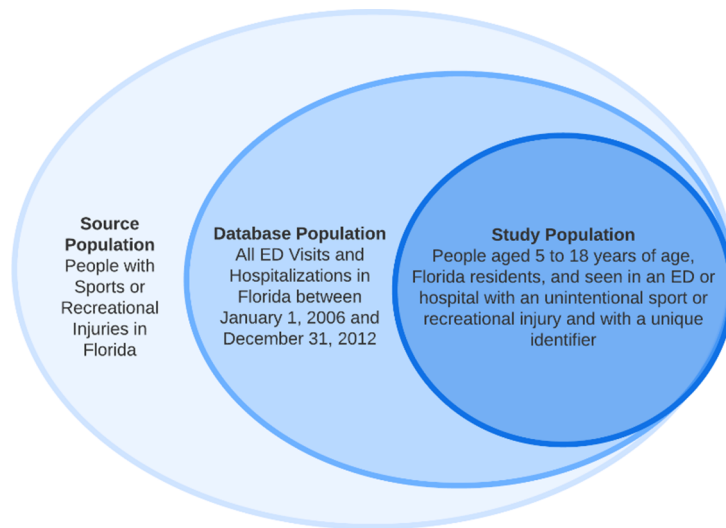
When information is available for different groups of people over time, such as health visits over multiple years, data should be stacked instead of being put side by side. This is called concatenation. If you are unsure whether observations in a given data set belong to an individual, they can be treated as discrete pieces of information and thus you will likely use concatenation more than merging. A major consideration when concatenating data is that the variables in each data set being stacked have the same names and formatting. Failure to abide by this can result in loss of data and/or a data set resulting in duplicate columns. For brevity, many existing papers refer to either of these processes simply as merging or data linkage. We will distinguish between merging and concatenation throughout this paper for clarity.

We obtained access to an identifier variable (the masked social security number is discussed below) and wanted our resulting data for an individual to be on a single row, meaning even if an individual had multiple visits their data would still be a single row. When using data from any clinical system,

Table 1: Variables included in the longitudinal sports and recreation injury study

Variable name	ED	Hospital	Variable name	ED	Hospital
Age	1	1	Race and Ethnicity	1	1
E-Code	3	3	ED Hour of Arrival	1	1
Length of Stay	1	1	Year of Service	1	1
Masked Identifier	1	1	Admission Source	1	1
Principal Diagnosis	1	1	Type of Admission/Priority of Admission	0	1
Other Diagnosis	9	30	Time of Admission	0	1
Principal Procedure	1	1	Days to Principal or Other Procedure	0	1
Procedure	4	30	Present on Admission Indicator for E-Codes	0	1
Payer	1	1	Present on Admission Indicator for Other Diagnoses	0	1
Patient Residence	1	1	Present on Admission Indicator for Principal Diagnosis	0	1
Patient Status at Discharge	1	1	ER Charges	1	1
Quarter of the Year	1	1	Trauma Charges	1	1
Admitting Diagnosis	1	1	Physical or Occupational Therapy Charges	0	1
Day of the Week	1	1	Total Gross Charges	1	1
Sex	1	1	Numbers indicate the number of variables for a particular item (e.g. there is 1 sex variable).		

Figure 1: Population hierarchy of the longitudinal sports and recreation injury study



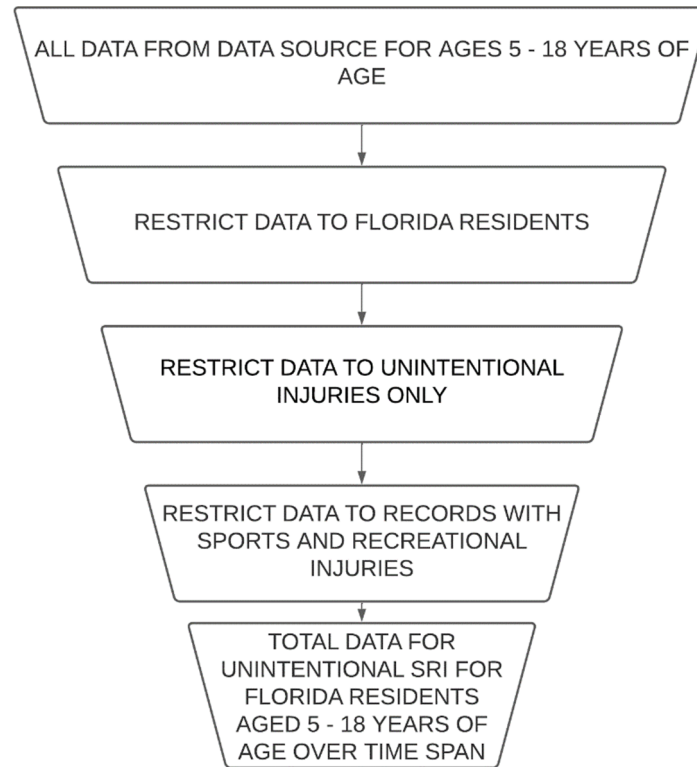
patients often do not have the same number of diagnoses or the same number of visits in a single year let alone across time. This meant the solution needed to be flexible enough to fit any possible situation, another criteria of a successful surveillance system [43]. As you can see in Figure 2, our process took all data from a particular source (e.g. hospitalisations), restricted the population based on our inclusion criteria (Florida residents between the ages of 5 and 18 years with unintentional sports and recreational injuries of an unintentional nature), and ended up with a set containing the total injuries of concern.

Identifying the identifiers

When linking administrative data, particularly discharge data for surveillance of a health issue, there are several important considerations including the necessity to have a consistently used universal patient level identifier, or combination of identifiers, across years of data collection such as a date of birth or a national identification number. In the United

States, there is no national identification number so the Social Security Number (SSN), a unique number allocated to each individual person, is often used a proxy. The SSN is issued by the US Social Security Administration, an agency established to oversee an anti-poverty program – Social Security – in 1935 via the Social Security Act [49]. Social Security includes benefits for retiring workers after age 65, disability insurance, health insurance (Medicare), survivors' insurance, and supplemental income for those with disabilities including children. SSNs are issued by the US Social Security Administration when applications are filed, often at birth, but registration is voluntary. An SSN is required for people that file US income taxes, open bank accounts, buy savings bonds, apply for government services, or get a job in the United States. It allows the agency to "identify and accurately record your covered wages or self-employment earnings" and "to monitor your record once you start getting benefits" [49]. Non-citizens that are authorized to work in the United States can get an SSN but, to be clear, as it is not required, there

Figure 2: Obtaining a study population from the database population in the longitudinal sports and recreation injury study



are people seeking health care in the US that do not have an SSN.

Something like a driver's license number or residential (or mailing) address is not sufficient as a universal identifier as 1) these values can change with varying frequency and 2) everyone does not have them. When possible, these identifiers can be used in combination with other identifiers such as the SSN to successfully identify a single individual. As societies change and values or constructs of those societies change, the utilisation of socio-demographic variables such as race, ethnicity, sex, and gender may be inconsistently captured. Race and ethnicity are defined and utilized in many ways across time, vary in terms of categorization, and vary by source; some data may be patient identified and other data may be nurse identified, among other ways of capture. Many scientists have previously misused the word gender to identify biological differences in people so many sets designed and collected in the past may interchangeably use the term sex and gender to mean the same thing – biology. However, as we know better, newer data sets are better defining and utilising the term gender and doing the same for sex. When needed, we also define and utilise sexual identity and sexual attraction. The identifiers and demographic variables are not just data points but markers of society. The longer the time period being linked, the more difficult it may be to identify a consistent set of identifiers no matter which ones are used.

In the present study, we had limited identifiers available in the data that is available to researchers. We chose to use a masked version of the SSN as the lone identifier. This was a value that was randomly and securely created using alphanumeric characters by AHCA in lieu of the actual SSN. This is done to anonymise the actual value and safeguard the patient's SSN. Every time data is drawn from the original

source for research purposes, the mask changes. For example, an actual SSN is a nine-digit number (e.g. 000000000) while the masked version might be Az024M12NGq81a. Note that the example value of 0 is not always recreated as the same alphanumeric digit but every visit for the individual with the SSN of 000000000 would be Az024M12NGq81a. Also note that the masked number is not the same length of the original SSN. The rationale behind using the masked SSN in this study was – 1) changes in the data across time to the race and ethnicity variables to make it difficult to know that using the variables would consistently match patients; 2) the benefit of being able to identify most individual patients using the masked SSN; 3) no other address information was available; and 4) no other consistent identifier was available in data researchers could access to prevent re-identification of patients.

Cleaning the study population data

Data obtained from AHCA included a variable to indicate residence of the patient which we used to exclude any visits that were not for Florida residents. We then identified the unintentional injury discharges using the ICD-9-CM external cause-of-injury codes (E-codes). E-codes are used to classify injury by mechanism (e.g. struck by/against) and intent (e.g. unintentional). These codes are in addition to the "main" ICD codes. In ICD-10-CM and beyond, external cause of injury codes are included as activity codes (V00-Y99) and provide a greater level of detail about the event. Because they are better integrated into ICD in versions 10 and beyond, we anticipate greater adoption and utilization which will improve linkage potential. Tools and recommendations to examine injury data can be found on the website of the Centers for Disease Control

Figure 3: Example study data set

ID	DX1	DX2	DX3	DX4	DX5	E1	E2	E3	...	X

Table 2: Altered external cause of injury morbidity matrix SAS syntax

Original^a:

```
IF (CAUSEDET GE* 800 AND CAUSEDET LE† 869) OR (CAUSEDET GE* 880
AND CAUSEDET LE† 929) THEN INJ = 1;
```

Updated^b:

```
IF (CAUSEDET GE* 800 AND CAUSEDET LE† 869) OR (CAUSEDET GE* 880
AND CAUSEDET LE† 929) THEN INJ = 1;
IF (CAUSEDET2 GE* 800 AND CAUSEDET2 LE† 869) OR (CAUSEDET2 GE* 880
AND CAUSEDET2 LE† 929) THEN INJ2 = 1;
IF (CAUSEDET2 GE* 800 AND CAUSEDET2 LE† 869) OR (CAUSEDET2 GE* 880
AND CAUSEDET2 LE† 929) THEN INJ3 = 1;
UNINTENTIONAL = 0;
IF INJ=1 OR INJ2 = 1 OR INJ3 = 1 THEN UNINTENTIONAL = 1;
```

*GE means Greater than or equal to.

†LE means Less than or equal to.

^a Original Syntax provided by the Centers for Disease Control and Prevention National Center for Injury Prevention and Control https://ftp.cdc.gov/pub/Health_Statistics/NCHS/injury/sascodes/icd9cm_external.sas https://www.cdc.gov/nchs/injury/injury_tools.htm.

^b Altered syntax copyrighted under the GNU GPL 3.0 license – <https://figshare.com/s/db7d50ee073360bd68bf> SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. indicates USA registration.

and Prevention (CDC) National Center for Injury Prevention and Control (NCIPC) [50, 51]. One of these tools, the External Cause of Injury Morbidity Matrix, was designed to use the first, or primary, of available E-codes in recorded data. Our data contained three E-code variables (Figure 3). In Figure 3, you can see an example data frame with an identifying variable (ID), five diagnosis variables (DX1-DX5), and three E-codes (E1-E3). As we found, the user may have an inconsistent number of diagnosis variables or E-codes across time. The method as described is adaptable to any number that the user has available [52]. It is also adaptable if the user has more than one identifying variable.

Due to the potential arbitrary placement of E-codes in the available fields and a want to include all possible instances of unintentional sport and recreational injury, we altered the SAS [53] syntax provided by the CDC [54] to include all E-code variables. The original syntax [54] created and used three variables – CAUSEDET (first 3 digits of the ICD code after the E), CAUSEDET4 (first 4 digits of the ICD code after the E for morbidity data), and PERSINJ (4th digit of the ICD code after the E) – to reorganize data. Our alteration was to make a CAUSEDET variable for all three E-codes we had available (Table 2). The original first E-code variable (E1) in Figure 3 was used to create the new variable CAUSEDET, the second (E2) was used to create CAUSEDET2, and the third (E3) was used to create CAUSEDET3. For example, if the data for the first visit for patient 'Az024M12NGq81a' had a valid response

for the first E-code (E1). This response is the value 'E849.4' and so the new variable CAUSEDET would contain the value 849 because they were the first three digits after the E. We could then use CAUSEDET, CAUSEDET2, and CAUSEDET3 to create variables (INJ, INJ2, and INJ3) indicating whether any of the three E-code variables contained data on an unintentional injury as opposed to some other external cause of injury [Table 2]. Since our intention was to only retain observations where an unintentional injury had occurred, we then created an indicator variable (UNINTENTIONAL) to identify and restrict the data set. Complete example syntax with instructions can be found online [52].

Finding and including sports and recreational injuries

The next step of the cleaning process was to identify which records contained SRI. A number of ICD-9-CM E-codes were available for this purpose (Table 3) [55]. While not the most inclusive, these variables narrowed down the number of observations to injuries that occurred in places for sport and recreation (E849.4), using recreational equipment (e.g. E885.2 [fall from a skateboard]), or while participating in sport or recreational activities without equipment (e.g. E910.2 [accidental drowning and submersion while engaged in other sport or recreational activity without diving equipment]). If the value of any E-code variable matched these codes singularly

Table 3: ICD-9-CM codes used to identify sports & recreational injuries

E849.4	E910.0
E885.0	E910.1
E885.1	E910.2
E885.2	E917.0
E885.3	E917.5
E884.4	E001-E010
E886.0	

Figure 4: Example of data changes – race and ethnicity 2007–2009 to 2010–2012

2006 – 2009	2010 – 2012	Final Version
Race and Ethnicity – Single Variable	Race – Variable 1	Race and Ethnicity – Single Variable
1. American Indian/Alaska Native	1. American Indian/Alaska Native	1. American Indian/Alaska Native
2. Asian/Pacific Islander	2. Asian	2. Asian/Pacific Islander
3. Black	3. Native Hawaiian/Pacific Islander	3. Black
4. White	4. Black	4. White
5. White Hispanic	5. White	5. White Hispanic
6. Black Hispanic	6. Other	6. Black Hispanic
7. Other		7. Other
	Ethnicity – Variable 2	
	E1. Hispanic	
	E2. Non-Hispanic	
	E3. Unknown	

or in combination (Table 3), the observations were retained. We identified these codes from prior works of SRI as well as a systematic examination of available codes in the ICD-9-CM system [2, 55].

Final cleaning steps

Normal considerations for combining data include standardizing inconsistent coding or formatting data sets and regular updates to data categories such as sociodemographic factors or causes of injury that improve the usefulness of the data. To combat the latter, we used a two-step process involving available codebooks and examination of the data sets. We identified every possible version of a variable and determined how we could make them consistent. For example, in Figure 4 you can see that from 2006 to 2009, race and ethnicity were in all of the data sets as a single combined variable (RACE). In 2010, however, they were separated into two separate variables (RACE and ETHNICITY). These two variables were more inclusive of possible categorizations. To reformat the 2006 to 2009 data into separate race and ethnicity variables likely would have resulted in a larger degree of misclassification of observations than combining the data from the race and the ethnicity variables from 2010–2012. Thus, we elected to use the older categories from the 2006–2009 RACE variable for the linkage. We updated all data prior to

concatenating years. While this method was often used to prevent a massive loss of data, it also meant we were unable to take advantage of newer and improved variable formats and categories.

Merging and concatenating data

Figure 5 shows each data set as it is restricted by patient residence, unintentional injury status, and then SRI status. Next the data is concatenated by source and finally the sources are merged into our final study population data set. The next section and Figure 6 provide more detail for preparing the data from each source for merging. In order to complete this process, we needed at least one identifying variable for individual people. We used the masked SSN.

Merging all the data together for surveillance

The data sets for each type (ED and Hospitalisation) were obtained in annual extractions (e.g. 2006, 2007, 2008) and needed to be combined (concatenated). After concatenating the data by type (Figure 6, Step 2) we then sorted each of the two new data sets (ED and Hospitalisation) by masked SSN (unique identifier). This allowed us to see how many total visits existed in each set and how many visits existed for each unique individual. After creating a visit specific variable (Figure 6,

Figure 5: The flow of the merges and concatenation steps to create the study population from the database population

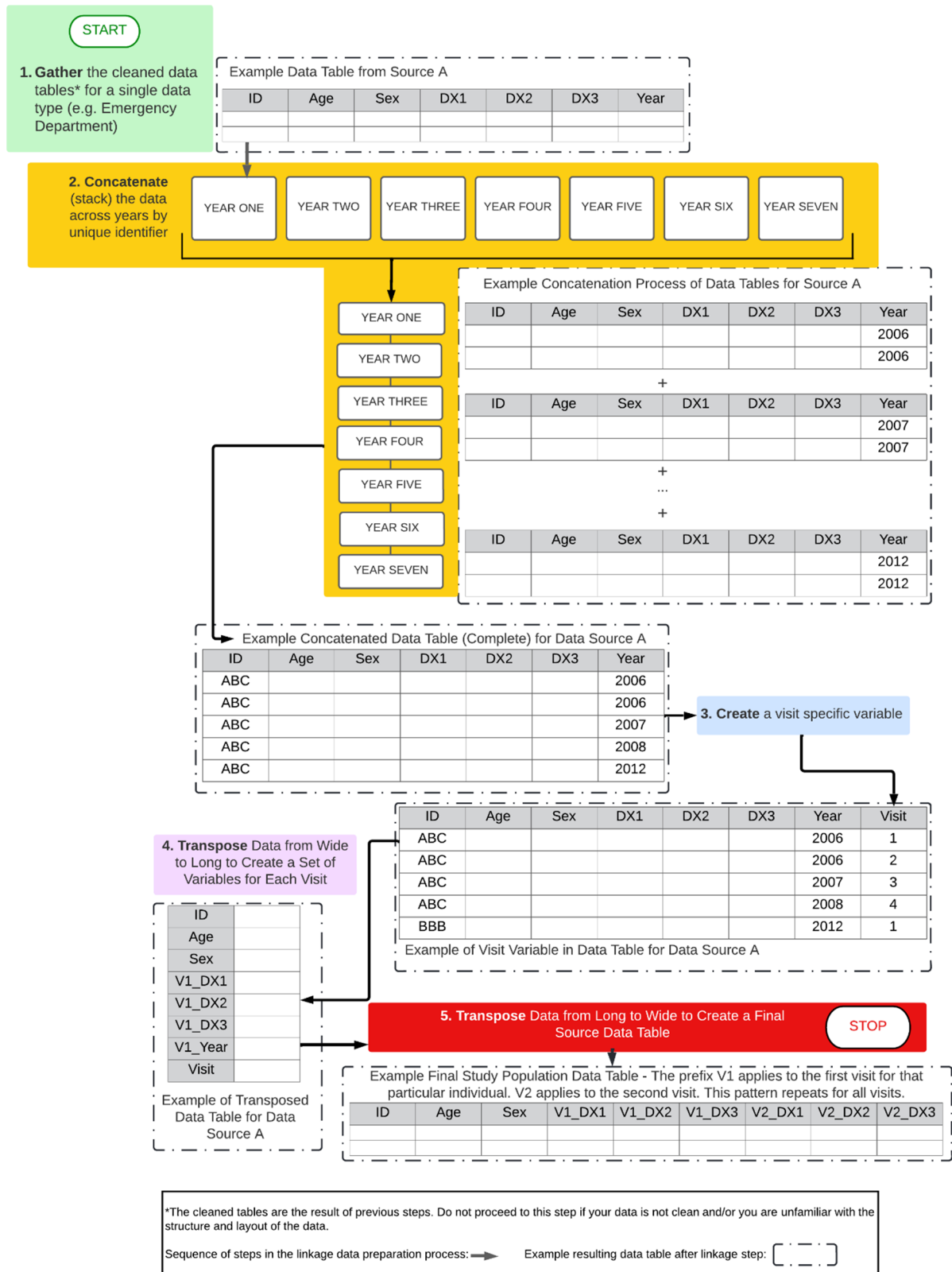


Step 3), we then used this new variable to transform (or rearrange) the data frame. The type of transformation used is also known as transposing. To transpose data generally means one of two things. The first is often called changing the data from wide and long (one row per visit) to long and narrow (one row per variable). This can be seen in Figure 6, Step 4. We change the data frame from a point where the columns are the individual variables (wide) and the rows represent individual visits (long) to a point where each row represents individual variables related to specific visits (long) and the columns are for each individual person (not visit) in the data set (narrow).

We used this transformation to add a prefix ('V1_' in Figure 6, Step 4) to each visit specific variable. For example, in Figure 6, Step 4 you can see that the first diagnosis variable (DX1) for the first visit for subject 'ABC' is now row V1_DX1 in the data frame.

The second type of transpose is known as changing the data from long and narrow to short and wide. Think of it as almost reversing the first type of transpose. We can see how this works as we move from Step 4 to Step 5 in Figure 6. Because we are using this transformation on our new data frame (the one with the prefixes), when we transform the table

Figure 6: Steps to transform the single data sets into concatenated data sources in preparation for merging

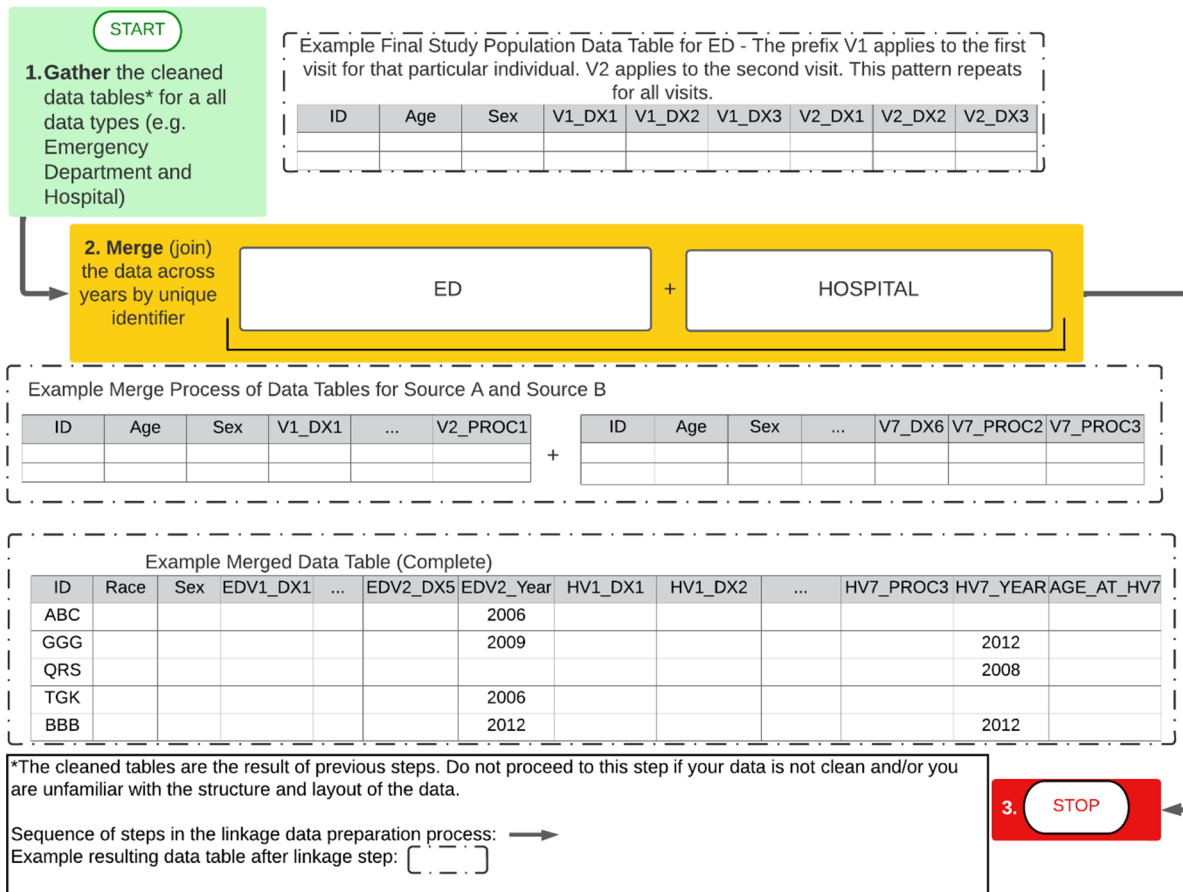


each individual person's data will be on a single row instead of across multiple rows. This makes the data set (top to bottom) shorter and (side to side) wider than the original data frame, hence the phrase 'short and wide'. The original data frame (the result of Figure 6, Step 2) is noticeably different than the resulting data frame (the result of Figure 6, Step 5). This

transformation put each individual person's discharges on a single row and created a requisite number of variables to retain all discharge information.

At this point either dataset (ED or hospital inpatient) could be analysed separately or merged to create one single and complete data set with ED and inpatient data

Figure 7: Steps to transform the concatenated data sources into a final merged data set



(Figure 7). The latter allows analysis across data sources while the former allows analysis by the source. Both are useful depending on the surveillance questions being asked.

Identifying the injuries

For injury surveillance purposes, it was important to align the now combined data to existing tools such as the Barell Matrix [56] to classify the injuries by body region (e.g. shoulder) and nature of the injury (e.g. fracture). One observed barrier to using the Barell Matrix (and the subsequent Injury Mortality Diagnosis (IMD) matrix for ICD-10) is that it was developed to only use the primary diagnosis on any record no matter how many diagnoses were available. However, our experience has indicated that the order of diagnoses, is arbitrary or based on the cost of the illness or injury to facilities and insurers rather than in the order of seriousness as intended. Because all aspects of injury regardless of seriousness impact the population, our goal was to capture and count all injuries, not just primary diagnoses or the total number of individuals injured. As such, we modified the Centers for Disease Control and Prevention's SAS syntax [57] for the Barell Matrix to include all available diagnoses in the data sets in a fashion similar to our alteration of the E-code variables. Because of the importance of how this was done and to not detract from the current work, specific methods for this process and population level statistics using this data will be published separately.

Results

Evaluating linkage quality

We calculated the frequency of the masked SSN in order to identify any outliers. We found that 2 identifiers had an improbable number of discharges (e.g. greater than 100 observations in a single year). We verified with the data owner, the Agency for Health Care Administration, that these particular identifiers were catch-all identifiers for people that did not have an SSN recorded on their discharge. In lieu of enough available data to perform probabilistic linkage, we decided to exempt these records (30.7% of ED data; 36.8% of hospitalization data) from the data set for individual level (person-level) analyses (Table 4). This data was still useful when looking at population wide measures such as counting the total number of SRI especially by sex or race/ethnicity.

As eloquently explained and demonstrated by both Gilbert [58] (GUILD guidance) and Harron [59], there is a significant need to not only report the link or the methods to achieve the link, but also certain factors such as bias that can indicate whether a link is low, medium, or high quality linkage. Lacking a gold standard for our project, we opted for sensitivity analyses that compared linked (those with a valid masked SSN) and unlinked data (those without a valid masked SSN) then identified potential biases that could occur and how they might affect the results. These analyses indicated that there were narrow but statistically significant differences by sex, age, and race/ethnicity (Tables 5, 6).

Table 4: Deterministic linkage statistics in the longitudinal sports and recreation injury study – individual persons represented

Source	Total visits	Excluded visits	Visits after exclusion	Individual people represented
Emergency Department	376,179	115,467	260,712	215,718
Hospital (Inpatient)	8552	3145	5407	5329

Table 5: Deterministic linkage statistics in the longitudinal sports and recreation injury study – differences in emergency department population demographics

Emergency department			
Linkage status	Excluded	Included	Comparison test
Age* in years	Mean 12.2	Mean 13.0	T-Test –75.4 $p < 0.0001$
*Normality tested using the KS test			
Sex (Male)	71.3%	74.1%	$X^2 315.7$ $p < 0.0001$
Race/Ethnicity			
American Indian/Alaska Native	0.2%	0.1%	$X^2 6062.8$ $p < 0.0001$
Asian/Pacific Islander	0.8%	0.5%	
Black	14.9%	26.5%	
White	60.7%	52.5%	
White Hispanic	16.2%	14.0%	
Black Hispanic	0.4%	0.4%	
Other	6.9%	6.2%	
Payer (Commercial health insurance)	32.9%	18.8%	$X^2 23535.91$ $p < 0.0001$

Table 6: Deterministic linkage statistics in the longitudinal sports and recreation injury study – differences in hospital population demographics

Hospital			
Linkage status	Excluded	Included	Comparison test
Age* in years	Mean 12.15	Mean 13.04	T-Test –11.2 $p < 0.0001$
*Normality tested using the KS test			
Sex (Male)	77.3%	81.0%	$X^2 16.6$ $p < 0.0001$
Race/Ethnicity			
American Indian/Alaska Native	0.2%	0.2%	$X^2 84.8$ $p < 0.0001$
Asian/Pacific Islander	0.9%	0.6%	
Black	16.3%	24.1%	
White	58.2%	53.6%	
White Hispanic	14.8%	14.4%	
Black Hispanic	0.5%	0.4%	
Other	9.2%	6.6%	
Payer (Commercial health insurance)	35.0%	22.8%	$X^2 294.3$ $p < 0.0001$

These differences persisted across years. This means that we must acknowledge that any results gained about the impact of sports and recreational injuries from this linked data is biased and may be less applicable to certain subgroups of the population. Further examination and potential linkages must be explored. With such a degree of missingness, imputation alone may not be a reasonable option, however may be one of the tools necessary to minimize loss of records and

improve both accuracy and representativeness of the study population.

Discussion

In this manuscript, we discuss in detail the methodology, barriers, and successes of using deterministic linkage to

combine emergency department and hospitalisation data with a single identifier for use in surveilling sports injuries across a seven-year period. This is significant because while similar methods have been used to examine other conditions (e.g. asthma), few have linked multiple types of administrative data especially with nearly no identifiers, nor did they specifically use this methodology to examine SRI. The steps described in this manuscript will not only allow the authors to complete future work looking at how well we can track SRI as the ICD continues to change, but others can use these methods with country specific data and expand the method to add other sources that use similar identifiers. If we expect to move data surveillance forward as budgets for data surveillance become even more limited, we must develop and improve methods to do it with fewer resources including data that has great limitations.

The use of deterministic linkage is not novel, but the particular combination of strategies employed here – using diagnoses and external cause of injury codes (E-codes) beyond the primary to make better use of all data; creating a static data set with all data over a period of time in one row per patient; and specifically applying these methods to the topic of sports and recreation injury – are novel in their use. We believe improved training of researchers and practitioners in these methods and their underlying theories will widen their use. For example, analysts need to understand that acquiring data can be much more complicated than just downloading from a website. Health data that is protected (e.g. contains information that could be used to directly identify an individual) in the US is subject to the Health Insurance Portability and Accountability Act (HIPAA) which dictates a set of rules for accessing and engaging this information. The particular data used in this study required multiple telephone and email conversations with AHCA to discuss the project and the feasibility using the data; a HIPAA waiver of authorisation from our institutional IRB; a separate application to the government of the State of Florida through AHCA that needed to be approved by multiple state officials; receipt of the data; and separate receipt of secure keys to access the data. This process took about three months. After data was received, any additions to who accessed the data needed to be confirmed with AHCA. At points where more data became available (e.g. a new year of data), the entire process started again and all analyses needed to be re-run because the masks for the SSNs change with every pull of the data. Both our institutional IRB and the State of Florida required annual renewals of approval. These were tedious though not impossible steps. Sometimes emails to generic agency email addresses (e.g. data@agency.gov) never get a response. We were lucky to have a known contact at the agency to work through our issues with and have fostered a long-standing relationship that we hope will be fruitful to make changes to how data is made accessible for all research and practice users and how we can make a greater number of identifiers available without compromising the integrity of the data. Future collaboration with AHCA and similar agencies is necessary to make previous linkages of this data available for all end users of the data without more barriers in place. We find that this is not impossible but the result of longstanding (though understandable) communication and trust issues between the data stewards and researchers. Secure,

source hosted versions of tools such as GitHub and volunteer moderators to work with state agencies could be the next step to make this a reality.

As data becomes more and more costly to obtain and funding for new surveillance sources is at a minimum, we find it important to have completed this work to provide data for future injury policy and prevention. We also feel these methods, as explicitly laid out here, can be applied to variety of data sources and across a wide variety of health topics. Most notably, these methods may be quite applicable to areas with limited resources or low prevalence such as rare diseases and disorders. This study was completed using data from the US, but as with many methods, we know it to be applicable across countries though with caveats of how data is captured and what identifiers are available to the public.

A goal of this work was to create steps that could be followed by people with some knowledge and expertise in data and statistical programming, yet also accessible to those without large financial resources. If the point of surveillance is to help identify health issues and places for improvement, we felt it was necessary to contribute to these achievements. Relational database languages such as SQL (structured query language) were considered for this project, but would have had significant limitations for our purposes. The data as obtained from AHCA was hierarchical in nature, meaning that there might be many pieces of data (known as 'children') but the data belonged to an individual (known as the 'parent'). The main limitation we considered was the difficulty in SQL to use hierarchical data without specified depth (e.g. various numbers of diagnosis variables per person and per data source) without expending a tremendous amount of resources (e.g. server or computer space, time, money, very specific software) [60]. The complexity of the data structure and the amount of data that SQL would need to handle it would require exorbitant funds to scale the system appropriately to support the size and the performance necessary for even simple queries because SQL's use of memory. In other words, every query would potentially add the need for more computing resources (think: hard drives and RAM). Performance is lost when data needs to be split up on too many servers or other computing machines. Newer additions to SQL since this work started have made dealing with hierarchical data easier but these additions are still dependent on the vendor for the SQL software [61]. Clinical organizations might be able to minimize these limitations, but more independent researchers, researchers at institutions with limited computing resources, and various government institutions such as health departments would not be able to consistently provide these types of funds or maintain these size systems long term. It would also be difficult to share data and methods as the data becomes even more complex. In the end, we decided to create methods that would benefit the many as opposed to the few. Users with the resources to use SQL or other relational database programs for data at scale may consider an alternative method more acceptable.

Successes and future potential

Some past studies of data linkage in sport and recreational injury, nearly all of them using data with many available identifiers, are linking data from an investigator led primary

data collection study with secondary data, or link data limited in scope. For example, Ackerman et al linked individual level data regarding knee injuries in Australia but achieved this by having the government agency link the data then provide it to the investigators [62]. In the present study, we obtained the data and performed the linkage. Both methods achieve an end goal, yet if researchers were able to obtain more unique information without overburdening data providers, it would be easier to achieve this type of work. The present study linked multiple data sets from the same agency together and linked data across sources (ED and hospitalisation). A specific goal was to only use secondary data which was achieved, but this manuscript did not demonstrate how to then take the linked data and combine it with data from other secondary data sources or with primary data as these very important tasks are outside the scope of this work. However, these are important aspects others have addressed and that need to be addressed as this project progresses.

Beauchamp et al used a stepwise approach in their study of de-identified records for hospital admissions in Victoria, Australia and compared the effectiveness of each of their methods [63]. Beauchamp had access to names, hospital names, full birth date, and zip codes (post codes) and were able to successfully link 87% of their cohort study population to hospitalisation data [63]. Clark et al matched ambulance data to emergency department records with no identifiers in England [64]. The process was time intensive (65 weeks), and required both a large team and external input [64], but resulting in an average of 81% data matched (range of 50–94% depending on the hospital trust) [64]. Though we matched approximately 70% of the data on a single identifier, the limitations of using state specific patient level data from Florida was that obtaining an identifier for linkage led to a vast amount of other identifiable data becoming unavailable for access for research purposes such as patient zip code, date of birth for more accurate age assessment, and zip code of the facility. The single identifier we had available, masked SSN, was not required to be recorded for patients. While residents of many other countries have national identification numbers, US citizens and residents do not. Had other variables been available along with the masked identifier, we could have potentially used probabilistic linkage or at least matched records with more certainty. It occurred to us that we potentially could have matched records using a combination of the masked identifier, sex, and race. However, discrepancies in how race and ethnicity were captured across time combined with how they are not self-identified in this data would have greatly reduced the match. Additional discrepancies for sex across time may not have occurred with the data we used in this project, but could be a factor in the future as people are able to more accurately identify sex and gender in medical records.

This project was achieved using SAS 9.4 software [53]. We have also begun to post equivalent examples using Python, specifically Python 3.10.2 [65]. One limitation to prior work by others in this space was to solely use a single software, often expensive software inaccessible to others around the world. We count using publicly available data for surveillance as a great success but to do it in a way that increases accessibility and applicability of these methods with multiple types of software as an even greater victory. This process certainly was not

without a number of challenges that must be addressed as this project moves forward.

Using data only from more recent years with consistent categories or improved categories will alleviate the problem of needing to recode or recategorise information, but it causes the loss of data that provides a better longitudinal outlook. This is not an uncommon problem – as we know better in science, especially when it comes to the capture of sociodemographic data, we do better. Progress should not be held up solely for need of long-term comparison. We address these limitations and potential methods for minimising their effect in future work. Furthermore, while it is outside the scope of this work, we do encourage continued efforts such as those by Lucyk [66] to qualitatively examine the processes that come together to make administrative data such as the discharge data we used in an effort to improve them. The complexity of linking data together across time cannot simply be explained by the methods as we describe them here, but in all the hands and applications it takes to get the data from point A to point Z.

There is quite a bit of future potential in this work. We did not address the application of these methods to ICD-10 data, combining ICD-9 and ICD-10 data for surveillance purposes, demonstration of the applicability of the data resulting from this data, and the application of existing tools (e.g. Barell Matrix) and new tools (Baker-10 Matrix) here. We will address each of these points in soon to come manuscripts. Syntax for this manuscript and future manuscripts will all be made available for use, for critique, and for modification with a GNU GPLv3 license in both SAS and open software (i.e. Python and R). Upon request, we will translate this work to other programs such as Stata and SPSS to increase usability.

Conclusion

We hope that readers are left with an understanding of and use of discharge data for surveillance of sport and recreation injuries. The quality of the data used for linkage and surveillance is always dependent on the source that the discharge data originates from, usually a medical provider. Because of this, these sources may not be as in depth or explanatory as data collected specifically for the purpose of surveillance or research, but has a great benefit in being affordable and quickly accessed. Reformatting the data into a usable format for surveillance requires several considerations in advance of acquiring and combining the data. Whether the data comes from a government, an organization, or directly from clinical records, it is important to be clear of your purpose up front and imperative to work with data managers to clarify any abnormal findings in the data.

Ethics statement

Ethics approval was granted by Florida Agricultural and Mechanical University (IRB # 630088).

Conflicts of interest

The authors have no conflicts to disclose.

Acknowledgements

The authors would like to thank several people and agencies for their contributions to this vision. The authors would like to thank Dr. Alex F. Howard and Dr. Jennifer Howard for their help and advice on how to make this overall project useful for clinical partners such as athletic trainers and sports medicine practitioners. The authors would also like to thank Dr. Robert Young for his expertise and contributions in translating our syntax into Python so the information could be available to more people around the world. The authors would like to thank Dr. C. Perry Brown and Dr. Jontae Sanders for sharing valuable lessons about this particular data source and for being there in the wee hours of the morning to troubleshoot issues. Finally, the authors would like to thank the State of Florida and the Florida Agency for Health Care Administration for the ability to access the identifiable and de-identified data that has made this entire project possible.

References

- Costa E Silva L, Fragoso MI, Teles J. Physical Activity-Related Injury Profile in Children and Adolescents According to Their Age, Maturation, and Level of Sports Participation. *Sports Health: A Multidisciplinary Approach*. 2017;9(2):118–25. <https://doi.org/10.1177/1941738116686964>
- Howard AF, Costich JF, Mattacola CG, Slavova S, Bush HM, Scutchfield FD. A statewide assessment of youth sports- and recreation-related injuries using emergency department administrative records. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*. 2014;55(5):627–32. <https://doi.org/10.1016/j.jadohealth.2014.05.013>. <https://pubmed.ncbi.nlm.nih.gov/25060289/>.
- Burt CW, Overpeck MD. Emergency visits for sports-related injuries. *Annals of emergency medicine*. 2001;37(3):301–8. <https://doi.org/10.1067/mem.2001.111707>. <https://pubmed.ncbi.nlm.nih.gov/11223767>.
- Fletcher EN, McKenzie LB, Comstock RD. Epidemiologic Comparison of Injured High School Basketball Athletes Reporting to Emergency Departments and the Athletic Training Setting. *Journal of athletic training*. 2014;49(3):381–8. <https://doi.org/10.4085/1062-6050-49.3.09>
- Fong DT-P, Man C-Y, Yung PS-H, Cheung S-Y, Chan K-M. Sport-related ankle injuries attending an accident and emergency department. *Injury*. 2008;39(10):1222–7. <https://doi.org/10.1016/j.injury.2008.02.032>
- Tirabassi J, Brou L, Khodaei M, Lefort R, Fields SK, Comstock RD. Epidemiology of High School Sports-Related Injuries Resulting in Medical Disqualification: 2005–2006 Through 2013–2014 Academic Years. *The American journal of sports medicine*. 2016;44(11):2925–32. <https://doi.org/10.1177/0363546516644604>
- Sabesan V, Steffes Z, Lombardo DJ, Petersen-Fitts GR, Jildeh TR. Epidemiology and location of rugby injuries treated in US emergency departments from 2004 to 2013. *Open Access J Sports Med*. 2016;7:135–42. <https://doi.org/10.2147/oajsm.S114019>. <https://www.ncbi.nlm.nih.gov/pubmed/27822128>.
- Kuczinski A, Newman JM, Piuze NS, Sodhi N, Doran JP, Khlopas A, et al. Trends and Epidemiologic Factors Contributing to Soccer-Related Fractures That Presented to Emergency Departments in the United States. *Sports health*. 2019;11(1):27–31. <https://doi.org/10.1177/1941738118798629>. <https://www.ncbi.nlm.nih.gov/pubmed/30247999>.
- Pagnotta KD, Mazerolle SM, Pitney WA, Burton LJ, Casa DJ. Implementing Health and Safety Policy Changes at the High School Level From a Leadership Perspective. *Journal of athletic training*. 2016;51(4):291–302. <https://doi.org/10.4085/1062-6050-51.2.09>. <https://www.ncbi.nlm.nih.gov/pubmed/27002250>.
- Bohne C, George SZ, Zeppieri G, Jr. Knowledge of injury prevention and prevalence of risk factors for throwing injuries in a sample of youth baseball players. *International journal of sports physical therapy*. 2015;10(4):464–75. <https://www.ncbi.nlm.nih.gov/pubmed/26345986>.
- LaBella CR, Myer GD. Youth sports injury prevention: keep calm and play on. *British journal of sports medicine*. 2017;51(3):145–6. <https://doi.org/10.1136/bjsports-2016-096648>. <https://www.ncbi.nlm.nih.gov/pubmed/27919920>.
- Sarmiento K, Thomas KE, Daugherty J, Waltzman D, Haarbauer-Krupa JK, Peterson AB, et al. Emergency Department Visits for Sports- and Recreation-Related Traumatic Brain Injuries Among Children - United States, 2010–2016. *MMWR Morbidity and mortality weekly report*. 2019;68(10):237–42. <https://doi.org/10.15585/mmwr.mm6810a2>. <https://www.ncbi.nlm.nih.gov/pubmed/30870404>.
- Kay AB, Wilson EL, White TW, Morris DS, Majercik S. Age is just a number: A look at "elderly" sport-related traumatic injuries at a level I trauma center. *Am J Surg*. 2019;217(6):1121–5. <https://doi.org/10.1016/j.amjsurg.2018.11.030>. <https://www.ncbi.nlm.nih.gov/pubmed/30528788>.
- Stephenson C, Rossheim ME. Brazilian Jiu Jitsu, Judo, and Mixed Martial Arts Injuries Presenting to United States Emergency Departments, 2008–2015. *J Prim Prev*. 2018;39(5):421–35. <https://doi.org/10.1007/s10935-018-0518-7>. <https://www.ncbi.nlm.nih.gov/pubmed/30043324>.
- Rowe BH, Eliyahu L, Lowes J, Gaudet LA, Beach J, Mrazik M, et al. Concussion diagnoses among adults presenting to three Canadian emergency departments: Missed opportunities. *The American journal of emergency medicine*. 2018;36(12):2144–51.

- <https://doi.org/10.1016/j.ajem.2018.03.040>. <https://www.ncbi.nlm.nih.gov/pubmed/29636295>.
16. Harmon KJ, Proescholdbell SK, Register-Mihalik J, Richardson DB, Waller AE, Marshall SW. Characteristics of sports and recreation-related emergency department visits among school-age children and youth in North Carolina, 2010–2014. *Injury epidemiology*. 2018;5(1):23. <https://doi.org/10.1186/s40621-018-0152-0>. <https://www.ncbi.nlm.nih.gov/pubmed/29761235>.
17. Medicine Io. Reducing the Burden of Injury: Advancing Prevention and Treatment. Bonnie RJ, Fulco CE, Liverman CT, editors. Washington, DC: The National Academies Press; 1999. 336 p.
18. Schneuer FJ, Bell JC, Adams SE, Brown J, Finch C, Nassar N. The burden of hospitalized sports-related injuries in children: an Australian population-based study, 2005–2013. *Injury Epidemiology*. 2018;5(1):45. <https://doi.org/10.1186/s40621-018-0175-6>
19. Akoto R, Lambert C, Balke M, Bouillon B, Frosch K-H, Höher J. Epidemiology of injuries in judo: a cross-sectional survey of severe injuries based on time loss and reduction in sporting level. *British journal of sports medicine*. 2018;52(17):1109–15. <https://doi.org/10.1136/bjsports-2016-096849>. <https://bjsm.bmj.com/content/bjsports/52/17/1109.full.pdf>.
20. Liller KD, Morris B, Yang Y, Bubu OM, Perich B, Fillion J. Injuries and concussions among young children, ages 5–11, playing sports in recreational leagues in Florida. *PloS one*. 2019;14(5):e0216217. <https://doi.org/10.1371/journal.pone.0216217>. <https://www.ncbi.nlm.nih.gov/pubmed/31091293>.
21. Warner K, Savage J, Kuenze CM, Erkenbeck A, Comstock RD, Covassin T. A Comparison of High School Boys' and Girls' Lacrosse Injuries: Academic Years 2008–2009 Through 2015–2016. *Journal of athletic training*. 2018;53(11):1049–55. <https://doi.org/10.4085/1062-6050-312-17>. <https://www.ncbi.nlm.nih.gov/pubmed/30451536>.
22. National Collegiate Athletic Association Sports Science Institute. NCAA Injury Surveillance Program. 2020. Available from: <http://www.ncaa.org/sport-science-institute/ncaa-injury-surveillance-program>.
23. Fulstone D, Chandran A, Barron M, DiPietro L. Continued Sex-Differences in the Rate and Severity of Knee Injuries among Collegiate Soccer Players: The NCAA Injury Surveillance System, 2004–2009. *Int J Sports Med*. 2016;37(14):1150–3. <https://doi.org/10.1055/s-0042-112590>. <https://www.ncbi.nlm.nih.gov/pubmed/27706548>.
24. Furlong L-AM, Rolle U. Injury incidence in elite youth field hockey players at the 2016 European Championships. *PloS one*. 2018;13(8):e0201834-e. <https://doi.org/10.1371/journal.pone.0201834>. <https://www.ncbi.nlm.nih.gov/pubmed/30138463>.
25. Calloway SP, Hardin DM, Crawford MD, Hardin JM, Lemak LJ, Giza E, et al. Injury Surveillance in Major League Soccer: A 4-Year Comparison of Injury on Natural Grass Versus Artificial Turf Field. *The American journal of sports medicine*. 2019;47(10):2279–86. <https://doi.org/10.1177/0363546519860522>. <https://www.ncbi.nlm.nih.gov/pubmed/31306590>.
26. Ekstrand J, Spreco A, Bengtsson H, Bahr R. Injury rates decreased in men's professional football: an 18-year prospective cohort study of almost 12 000 injuries sustained during 1.8 million hours of play. *British journal of sports medicine*. 2021;55(19):bjsports-2020-103159. <https://doi.org/10.1136/bjsports-2020-103159>. <https://bjsm.bmj.com/content/bjsports/early/2021/02/05/bjsports-2020-103159.full.pdf>.
27. Bolling C, van Mechelen W, Pasman HR, Verhagen E. Context Matters: Revisiting the First Step of the 'Sequence of Prevention' of Sports Injuries. *Sports Medicine*. 2018;48(10):2227–34. <https://doi.org/10.1007/s40279-018-0953-x>
28. Vriend I, Gouttebarger V, Finch CF, van Mechelen W, Verhagen EALM. Intervention Strategies Used in Sport Injury Prevention Studies: A Systematic Review Identifying Studies Applying the Haddon Matrix. *Sports Medicine*. 2017;47(10):2027–43. <https://doi.org/10.1007/s40279-017-0718-y>. <https://pubmed.ncbi.nlm.nih.gov/28303544/>.
29. van Mechelen W, Hlobil H, Kemper HC. Incidence, severity, aetiology and prevention of sports injuries. A review of concepts. *Sports medicine (Auckland, NZ)*. 1992;14(2):82–99. <https://doi.org/10.2165/00007256-199214020-00002>. <https://link.springer.com/article/10.2165/00007256-199214020-00002>.
30. Horan JM, Mallonee S. Injury Surveillance. *Epidemiologic reviews*. 2003;25(1):24–42. <https://doi.org/10.1093/epirev/mxg010>
31. Atherly A, Whittington M, VanRaemdonck L, Lampe S. The Economic Cost of Communicable Disease Surveillance in Local Public Health Agencies. *Health services research*. 2017;52 Suppl 2(Suppl 2):2343–56. <https://doi.org/10.1111/1475-6773.12791>. <https://pubmed.ncbi.nlm.nih.gov/29130264>.
32. Kirkwood A, Guenther E, Fleischauer AT, Gunn J, Hutwagner L, Barry MA. Direct Cost Associated With the Development and Implementation of a Local Syndromic Surveillance System. *Journal of Public Health Management and Practice*. 2007;13(2):194–9. <https://doi.org/10.1097/00124784-200703000-00017>. https://journals.lww.com/jphmp/Fulltext/2007/03000/Direct_Cost_Associated_With_the_Development_and_17.aspx.
33. University of Minnesota Center for Infectious Disease Research and Policy. Costs to develop and maintain a state biosurveillance system:

- The New York example. 2021. Available from: <https://www.cidrap.umn.edu/practice/costs-develop-and-maintain-state-biosurveillance-system-new-york-example>.
34. EuroSafe, Dinesh Sethi. Policy briefing 7 Injury surveillance: a health policy priority. EuroSafe.
35. and BAV, McBean M. Administrative Data for Public Health Surveillance and Planning. Annual Review of Public Health. 2001;22(1):213–30. <https://doi.org/10.1146/annurev.publhealth.22.1.213>. <https://www.annualreviews.org/doi/abs/10.1146/annurev.publhealth.22.1.213>.
36. Saydah SH, Geiss LS, Tierney E, Benjamin SM, Engelgau M, Brancati F. Review of the performance of methods to identify diabetes cases among vital statistics, administrative, and survey data. Annals of epidemiology. 2004;14(7):507–16. <https://doi.org/10.1016/j.annepidem.2003.09.016>. <https://www.ncbi.nlm.nih.gov/pubmed/15301787>.
37. Dombkowski KJ, Wasilevich EA, Lyon-Callo S, Nguyen TQ, Medvesky MG, Lee MA. Asthma surveillance using Medicaid administrative data: a call for a national framework. Journal of public health management and practice : JPHMP. 2009;15(6):485–93. <https://doi.org/10.1097/PHH.0b013e3181a8c334>. <https://www.ncbi.nlm.nih.gov/pubmed/19823153>.
38. Powell KE, Disiker RA, 3rd, Presley RJ, Tolsma D, Harris S, Mertz KJ, et al. Administrative data as a tool for arthritis surveillance: estimating prevalence and utilization of services. Journal of public health management and practice : JPHMP. 2003;9(4):291–8. <https://doi.org/10.1097/00124784-200307000-00007>. <https://www.ncbi.nlm.nih.gov/pubmed/12836511>.
39. Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS medicine. 2015;12(10):e1001885. <https://doi.org/10.1371/journal.pmed.1001885>.
40. Florida Agency For Health Care Administration. Florida Health Finder: Order Data / Data Dictionary. 2022. updated February 4, 2022. Available from: <https://www.floridahealthfinder.gov/Researchers/OrderData/order-data.aspx>.
41. Physicians ACoE. Freestanding emergency departments. Policy statement. Annals of emergency medicine. 2014;64(5):562
42. Florida Agency for Health Care Administration. Florida Center for Health Information and Transparency Data Catalog. 2019. Available from: <http://fhfstore.blob.core.windows.net/documents/researchers/OrderData/documents/Data%20Catalog.pdf>.
43. German RR, Horan JM, Lee LM, Milstein B, Pertowski CA. Updated guidelines for evaluating public health surveillance systems; recommendations from the Guidelines Working Group. MMWR Morbidity and mortality weekly report. 2001;50(RR13):1–35. <https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm>.
44. Sanders J, Baker C, Brown CP. The Nuances of Combining Multiple Hospital Data. SouthEast SAS User Group Conference 2014; Myrtle Beach, SC.2014. <https://analytics.ncsu.edu/sesug/2014/BB-10.pdf>.
45. Scott E, Hirabayashi L, Graham J, Krupa N, Jenkins P. Using hospitalization data for injury surveillance in agriculture, forestry and fishing: a crosswalk between ICD10CM external cause of injury coding and The Occupational Injury and Illness Classification System. Injury Epidemiology. 2021;8(1):6. <https://doi.org/10.1186/s40621-021-00300-6>
46. Sebastião YV, Metzger GA, Chisolm DJ, Xiang H, Cooper JN. Impact of ICD-9-CM to ICD-10-CM coding transition on trauma hospitalization trends among young adults in 12 states. Injury Epidemiology. 2021;8(1):4. <https://doi.org/10.1186/s40621-021-00298-x>
47. Slavova S, Costich JF, Luu H, Fields J, Gabella BA, Tarima S, et al. Interrupted time series design to evaluate the effect of the ICD-9-CM to ICD-10-CM coding transition on injury hospitalization trends. Inj Epidemiol. 2018;5(1):36. <https://doi.org/10.1186/s40621-018-0165-8>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6165830/>.
48. Glerum KM, Zonfrillo MR. Validation of an ICD-9-CM and ICD-10-CM map to AIS 2005 Update 2008. Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention. 2019;25(2):90–2. <https://doi.org/10.1136/injuryprev-2017-042519>. <https://injuryprevention.bmj.com/content/25/2/90>.
49. US Social Security Administration. Social Security. 2021. Available from: <https://www.ssa.gov/>.
50. Centers for Disease Control and Prevention. Recommended framework for presenting injury mortality data. MMWR Recomm Rep. 1997;46(RR-14):1–30. <https://www.ncbi.nlm.nih.gov/pubmed/9301976>.
51. Centers for Disease Control and Prevention. Matrix of E-code Groupings. 2018. Available from: <https://perma.cc/RVC7-TA4R>.
52. Baker C, Young R. Data Cleaning for Deterministic Data Linkage for Sports and Recreational Injuries Using ICD-9 and ICD-9-CM. 2022. University Libraries, Virginia Tech. Figshare. <https://doi.org/10.7294/19288343>.
53. SAS Institute, Inc.. SAS software. 2020. Cary, NC.
54. Centers for Disease Control and Prevention. ICD-9-CM External Cause-of-Injury (E-code) Matrix SAS

- Input Statements. 2018. updated 2018. Available from: <https://perma.cc/N75Z-2ZYB>.
55. Weiss AJ, Elixhauser A. Sports-Related Emergency Department Visits and Hospital Inpatient Stays, 2013: Statistical Brief #207. Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville (MD): Agency for Healthcare Research and Quality (US); 2006.
56. Barell V, Aharonson-Daniel L, Fingerhut LA, Mackenzie EJ, Ziv A, Boyko V, et al. An introduction to the Barell body region by nature of injury diagnosis matrix. Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention. 2002;8(2):91–6. <https://doi.org/10.1136/ip.8.2.91>. <https://www.ncbi.nlm.nih.gov/pubmed/12120842>.
57. Centers for Disease Control and Prevention. ICD-9-CM (Barell Matrix) SAS Input Statements. Available from: <https://perma.cc/YTS2-H3PS>.
58. Gilbert R, Lafferty R, Hagger-Johnson G, Goldstein H, Harron K, Li-Chun Z, et al. GUILD: GUIDance for Information about Linking Data sets. Journal of Public Health. 2018;40(1):191–8. <https://doi.org/10.1093/pubmed/fdx037>.
59. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology. 2017;46(5):1699–710. <https://doi.org/10.1093/ije/dyx177>
60. Špetič A, Gennick J. Transact-SQL Cookbook: O'Reilly; 2002.
61. Molinaro A, de Graaf R. SQL Cookbook, 2nd Edition: O'Reilly Media, Incorporated; 2020.
62. Ackerman IN, Bohensky MA, Kemp JL, de Steiger R. Likelihood of knee replacement surgery up to 15 years after sports injury: A population-level data linkage study. Journal of science and medicine in sport / Sports Medicine Australia. 2019;22(6):629–34. <https://doi.org/10.1016/j.jsams.2018.12.010>. [https://www.jsams.org/article/S1440-2440\(18\)30278-0/pdf](https://www.jsams.org/article/S1440-2440(18)30278-0/pdf).
63. Beauchamp A, Tonkin AM, Kelsall H, Sundararajan V, English DR, Sundaresan L, et al. Validation of de-identified record linkage to ascertain hospital admissions in a cohort study. BMC medical research methodology. 2011;11(1):42–9. <https://doi.org/10.1186/1471-2288-11-42>.
64. Clark SJ, Halter M, Porter A, Smith HC, Brand M, Fothergill R, et al. Using deterministic record linkage to link ambulance and emergency department data: is it possible without patient identifiers? A case study from the UK. International Journal of Population Data Science. 2019;4(1). <https://doi.org/10.23889/ijpds.v4i1.1104>. <https://ijpds.org/article/view/1104>.
65. Van Rossum G, Drake F. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
66. Lucyk K, Tang K, Quan H. Barriers to data quality resulting from the process of coding health information to administrative data: a qualitative study. BMC health services research. 2017;17(1):766. <https://doi.org/10.1186/s12913-017-2697-y>. <https://www.ncbi.nlm.nih.gov/pubmed/29166905>.

