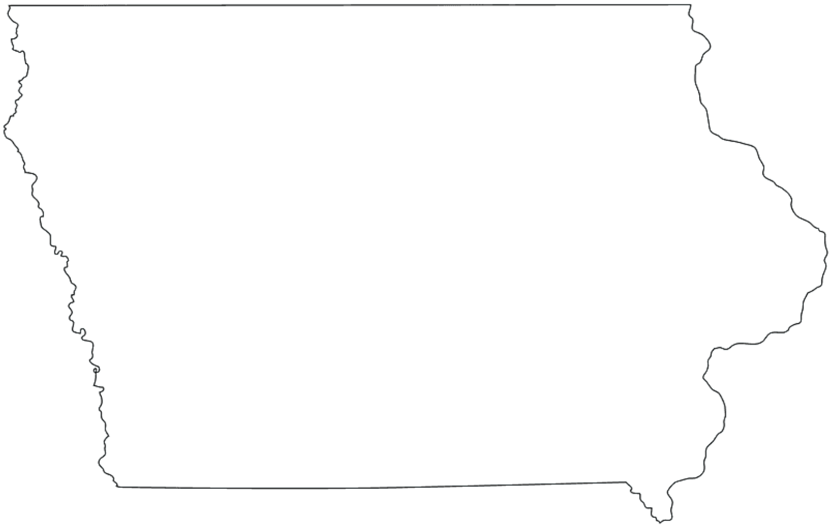
CODES

Iowa Crash Outcomes Data Evaluation System



2024
Year-End Report

Prepared by the
University of Iowa Injury Prevention Research Center
Under contract to Iowa Governor's Traffic Safety Bureau and Iowa Department of Transportation

IOWA

Injury Prevention
Research Center

Executive Summary

The objective of the Iowa CODES project is to examine outcomes related to motor vehicle crash-related injuries in the State of Iowa. These analyses are based on datasets formed through the probabilistic linkage of Iowa's crash data with the Iowa State Emergency Department Data and Iowa State Inpatient Data. Specialized analyses can be facilitated through additional linkages with judicial data, Fatality Analysis Reporting System (FARS) data, U.S. Census Bureau data, or other datasets. By linking these various data sources and using them for analyses, the team is in a unique position to discover and report data quality issues related to accessibility, accuracy, completeness, and integration.

During the current performance period (FY24), the UI CODES team:

- Worked to obtain updated Z table (crash) data and new Accident Processing System (APS; personally identifiable data from crashes) data from the Iowa DOT
- Explored potential strategies for improving the quality of the probabilistic linkage of crash data and hospital data, and
- Assessed the level of agreement between a number of variables in the crash, hospital, and judicial data.

Work completed by the team in FY23 concluded that a small amount of personally identifiable information (PII), specifically the month and year of birth and the zip code, are essential to successfully linking the crash and the hospital data. However, the date of birth and the zip code (along with most PII) have been removed from the Z table crash data provided to the UI by the Iowa Department of Transportation (DOT)'s Bureau of Traffic and Safety (TAS). This situation required the UI to obtain PII data from another Iowa DOT source (Motor Vehicle Division). The UI team regularly corresponded with IT personnel associated both these Iowa DOT units and provided feedback based on our data quality reviews, especially completeness, as the IT professionals worked to establish procedures that would automate the processing and transfer of these two datasets.

In FY23, probabilistic linkage was performed to link the crash records for injured persons with the HCUP data that were identified as transport accidents. The probabilistic record linkage software LinkSolv from Strategic Matching was used to perform the linkage. The primary linkage variables were the month and year of the crash, the month and year of the injured person’s birth, their sex, and their zip code. This was a good start, and the team was able to identify areas of improvement to get more matches and higher quality matches. Thus, FY24 saw major refinements to the linkage process while building on the foundation of the work from FY2023. Improvements to quality of the DOT crash data and an overhauled preprocessing of the HCUP data improved the reliability of the matches. The team explored ways to capture more matches, increasing the number of matches by 11.6%. To assess the quality of links made, several avenues were pursued involving the creation of new variables that could be compared between multiple data sources. The team also summarized how race and ethnicity data could be compared between CODES linked records and judicial records to assess match quality.

Project Description

The goal of the Iowa CODES project is to examine outcomes (injuries, long-term disability, hospital charges, discharge status) related to motor vehicle crash-related injuries in the State of Iowa. This is accomplished by probabilistically linking deidentified person (patient)-level crash and hospital data. These data are also sometimes linked with other datasets, e.g., judicial data, Fatality Analysis Reporting System (FARS) data, U.S. Census Bureau data, or other datasets, to conduct specialized analyses.

The hospital data sources are the Iowa State Inpatient Database (SID) and Iowa State Emergency Department Database (SEDD), which comes from the Agency for Healthcare Research and Quality (AHRQ)'s Healthcare Cost and Utilization Project (HCUP). Crash data have traditionally been obtained from the Iowa Department of Transportation in the form of "Z tables" generated by the Bureau of Traffic and Safety (TAS). Mortality data are obtained from the NHTSA's Fatality Analysis Reporting System (FARS).

By linking these various data sources and using them for analyses, the team is in a unique position to discover and report data quality issues related to accessibility, accuracy, completeness, and integration.

Accessibility to CODES Data Sources

The crash data and hospital data that are essential to this effort are regularly updated. Both sources of data have established procedures for requesting and accessing the data and both have established memorandum of understanding (MOU) or data use agreement (DUA). However, this performance period, accessibility proved to be extremely challenging.

Access to Iowa Crash Data

For more than a decade, researchers at the UI have received crash data in the form of Z tables provided by the Bureau of Traffic and Safety (TAS). Over the last few years, TAS has worked with the Institute for Transportation (InTrans) at Iowa State University to create new processes for generating the Z table data from the Iowa DOT's Accident Processing System (APS). During the FY24 performance period, the UI team collaborated heavily with two different teams at the Iowa DOT to initiate new processes and methods for obtaining crash data.

Access to Z table data

Throughout the FY24 performance period, TAS provided Z table data 5 different times. Each of these datasets had issues that prevented the data from being imported or used for analysis.

The first set of Z tables was provided to the UI on Oct 30, 2023. Within 2 two weeks, the UI reported that one of the tables could not be imported into SAS (a statistical analysis program). Within a few days, TAS provided a corrected file. About two weeks after that, the UI reported to TAS that the vehicle-level Z tables (i.e., the data tables containing data pertaining to units involved in crashes) had major issues with the unit identifiers. As the label suggests, the unit identifier should identify a specific unit consistently across all the vehicle-level Z tables. For example, unit 123456 in the driver Z table and unit 123456 in the vehicle Z table should refer to the same unit. However, our team discovered that within some multiple vehicle crashes, the unit identifiers were inconsistent across the various vehicle-level tables. This was especially problematic since these data had been used to filter specific types of crashes and drivers for a research project.

The second set of Z tables was received on May 3. The next day the UI reported issues with 4 of the Z tables, which were not formatted correctly and could not be imported into SAS. Two days later, TAS provided corrected files. However, none of these tables could be imported into SAS and this was immediately reported back to TAS.

The third set of Z tables was received on May 24. Again, there were issues that prevented 5 of the Z tables from being able to be imported into SAS. Within a few days, TAS provided corrected files. About a month later (June 27), after working with the data files, the UI discovered that the vehicle-level tables still had inconsistencies with the vehicle-level identifiers, and that many of the vehicle-level tables were missing the unit identifier altogether (see Completeness of Iowa Crash (Z Table) Data below for additional information).

The fourth set of Z tables was received on August 8. The issues with the vehicle-level identifiers were not resolved. In addition, the Narrative Z table had been reformatted and the new format could not be imported into SAS.

The fifth set of Z tables was received on September 16, 2024. Upon visual inspection, it appears the issues with the vehicle-level identifiers may now be resolved, but given the closeness to the end of the FY24 performance period, the UI team was unable to fully investigate. In addition, the Narrative table could again not be imported into SAS.

In summary, the first four Z table datasets that were provided by TAS could not be used for analyses and the final Z table dataset arrived very late in the performance period. For the most part, the TAS team was responsive to addressing the feedback from the UI. Unfortunately, TAS was not quite able to achieve their objective of getting the new processes for generating the Z tables stabilized so it could be run automatically each month.

Access to PII data

In the last quarter of the FY23 performance period, the UI team was informed that TAS could no longer provide personally identifiable information (PII) in the Z tables. This is problematic because date of birth and zip code, two of the data elements that are essential to linking the crash data to other data sets, were no longer provided in the Z tables. When the FY23 performance period ended, the UI team was investigating alternative ways to obtain DOB and zip code for individuals involved in crashes. A test file provided to the UI in the summer of 2024 suggested that it would be feasible to integrate data directly from the Iowa DOT's Accident Processing System (APS) with the Z table data. Therefore, in January of 2024 the UI team sent a data request to Motor Vehicle Division, the unit responsible for APS, for the data elements covered by the Memorandum of Understanding (MOU) between the UI and the Iowa DOT that were no longer being provided in the Z tables.

A little over a month after the UI made the data request, two members of the UI team met with a group of individuals from MVD and IT professionals supporting MVD. During the meeting the UI clarified the data request in the context of the standing MOU.

Three weeks later (3/19/2024) the UI and MVD met again. MVD described their progress and asked the UI to clarify some additional points.

Two months later (5/20/2024) MVD asked for another meeting. The main agenda item was discussing how to handle the transfer of crash diagrams, which are very large files. This was surprising to the UI team; while the ability for the UI to receive crash diagrams is described in the MOU, the UI team did not include crash diagrams in our data request. The IT personnel reported that they were trying to establish processes to facilitate all the data types covered in the MOU. While the UI appreciates and understands that rationale, the data request that we made was very specific and included only the PII data elements that we were no longer receiving from TAS.

On 5/30/2024 the MVD provided the UI with a test data file and informed us that we would have to access the actual data from an FTP site hosted by the Iowa DOT rather than the FTP location the UI had provided. Upon reviewing the test file, the UI discovered that the Accident Party data set did not include any non-motorists. This was reported to MVD on June 11. However, because the IT unit releases changes on a bi-monthly basis, the automated script could not be corrected until late August. In the meantime, MVD IT manually generated the Accident Party data set, which caused it to be mis-formatted and unable to be imported into SAS.

In early September, the UI was provided with a complete APS dataset for crashes from 2016-present. In all, it took just over 7 months from the initial meeting with MVD (8 months from when the UI made the data request) to obtain datasets containing date of birth and zip code that could be imported.

The UI has been reviewing the APS data and learning how the various tables in the data relate to each other. For example, the UI learned through observation that the data element labeled as "Accident Party ID" is not actually an identifier for a person involved in a crash but rather the identifier for a record related to that person. Each time a record is updated in the APS system, the Accident Party ID changes to a new number. These updates are usually but not always indicated with the previous record being assigned an "End Date." Since the end of the FY24 performance period, the UI met with an Iowa DOT data manager and then personnel from InTrans to better understand how to interpret and extract the PII data from APS and integrate it with the Z table data.

Renewal of MOU with Iowa DOT

The current MOU between the UI and the Iowa DOT detailing the use of crash, driving licensing history, and other data was set to expire on August 31, 2024. In March the UI inquired with the appropriate individual at Iowa DOT on the timeline of the process. In April, they replied they would prepare the form to extend the terms of the current MOU another two years and provide it within a month of the expiration date. The MOU renewal was received by the UI on July 12. The UI returned the signed agreement on July 19. The Iowa DOT signed the agreement on August 15. The UI received the finalized agreement on Sept. 15.

Access to HCUP in-patient and emergency room data

Due to obstacles with obtaining the Iowa Hospital Association data for use in this project, for several years the UI has been obtaining the state in-patient data (SID) and state emergency department data (SEDD) for Iowa from the Agency for Healthcare Research and Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP). In May, the UI requested the 2022 SID and SEDD datasets for Iowa using the same application language that we have used in the past. After initially receiving notification that the application was approved, the UI was informed that our intended use of the data, which had not changed from previous years, was not allowed.

Within two weeks the UI submitted a revised data request fully detailing how our procedures would adhere to the HCUP Statement of Intended Use (which was the same as previous years), including the deidentification of datasets. The request was again denied. In mid-July the UI team met with HCUP personnel to talk through the data use agreement, our intended uses of the data, and the restrictions on linking the data. During the meeting the HCUP staff agreed to speak with the appropriate person at Iowa Department of Health and Human Services (DHHS) to see if they would agree that our intended uses of the data were allowable. That meeting was held in August and the HCUP director reported back that the HCUP contact at Iowa DHHS would not allow the Iowa SID and SEDD data to be linked with crash data. The UI team modified our HCUP data request on August 27 and received the data in mid-September. The UI team has discussed several potential strategies for moving forward including novel linkage strategies, population-level analyses, and alternative data sources.

Crash Outcome Data Evaluation System (CODES)

Accuracy of Iowa Crash (Z Table) Data

Given the high level of effort put into obtaining usable versions of the Z table and APS data sets this performance period, there were few opportunities to assess the accuracy of the data. One accuracy issue was discovered when the UI team linked date of birth data from APS to Z table data from TAS and checked against driver age. This assessment revealed that driver age was incorrect for 18.6% of the drivers who were involved in a motor vehicle crash on their birthday. This issue was reported to TAS. However, given the access issues described above, the UI team has not been able to repeat this assessment and verify that the issue has been resolved.

Completeness of Iowa Crash (Z Table) Data

As described above, the UI team assessed the completeness of identifiers in the vehicle-level Z tables three different times as a quality control check on various iterations of the files. Table 1 shows the results of the first of these assessments, which was shared with the Iowa DOT TAS on June 27, 2024. All of the zdrv (driver) and most of the zvdm (vehicle damage) Z tables were missing the Accident Unit ID for at least some of the records, but none of the 2023 records included that identifier. In addition, there were numerous instances across multiple Z tables where the values entered for the Accident Unit ID did not match the value in the UNITKEY.

Table 1. Results of UI's evaluation of the unit identifiers in the vehicle-level Z-tables for data years 2020-2023.

Year	Vehicle-level Z table	Number of records	ACCIDENTUN = UNITKEY	ACCIDENTUN ne UNITKEY	ACCIDENTUN = .
2023	zveh	89634	89631	3	0
2023	**zdrv**	**89634**	**12909**	**1971**	**74754**
2023	zcit	89634	89631	3	0
2023	zctb	89634	89614	20	0
2023	zvdm	89634	14877	3	**74754**
2023	zrdb	89634	89631	3	0
2023	zcvo	2333	2321	21	0
2022	zveh	91503	91503	0	0
2022	**zdrv**	**91503**	**78920**	**12136**	**447**
2022	zcit	91503	91503	0	0
2022	zctb	91503	9149	64	0
2022	zvdm	91503	91056	0	**447**
2022	zrdb	91503	91503	0	0
2022	zcvo	2534	2457	77	0
2021	zveh	93158	93158	0	0
2021	**zdrv**	**93158**	**81017**	**11917**	**224**
2021	zcit	93158	93158	0	0
2021	zctb	93158	93119	39	0
2021	zvdm	93158	92934	0	**224**
2021	zrdb	93158	93158	0	0
2021	zcvo	2517	2517	0	0
2020	zveh	80167	0	0	0
2020	**zdrv**	**80167**	**70641**	**9516**	**10**
2020	zcit	80167	80167	0	0
2020	zctb	80167	80128	39	0
2020	zvdm	80167	80157	10	0
2020	zrdb	80167	0	0	0
2020	zcvo	2241	0	0	0

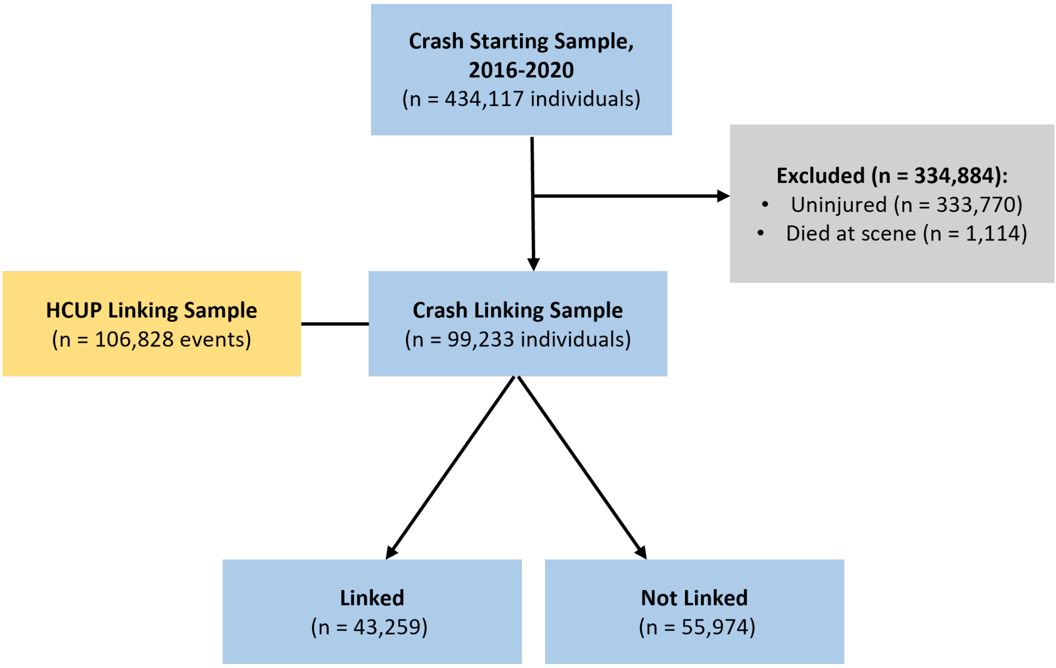
Data Linkage and Data Quality Checks

Probabilistic linkage was performed to link the crash records for injured persons with the HCUP data that were identified as transport accidents. The probabilistic record linkage software LinkSolv from Strategic Matching was used to perform the linkage. The primary linkage variables were the month and year of the crash, the month and year of the injured person's birth, their sex, and their zip code.

This performance period, major steps were taken to enhance the linkage. Corrections to DOT crash data and overhauled preprocessing of the HCUP data improved the reliability of the matches. Modification of the linkage process in LinkSolv through adjustment of the linkage parameters was explored, ultimately increasing the number of matches by 11.6% from 43,259 to 48,262. The team also identified several avenues to assess the quality of links made, involving the creation of new variables that could be compared between multiple data sources. Finally, traffic citations data was incorporated into the linked CODES data and a preliminary assessment was made to demonstrate how race and ethnicity data could be compared between sources as an additional test of match quality.

Updated linkage

In FY24, we completed new linkages with updated crash data and reprocessed HCUP data. Altogether, 48,262 (48.6%) of the injured persons in the crash records were linked to a record from HCUP. Figure 1 below shows the linkage process starting from the crash data.



Crash Starting Sample, 2016-2020
(n = 434,117 individuals)

Excluded (n = 334,884):
- Uninjured (n = 333,770)
- Died at scene (n = 1,114)

HCUP Linking Sample
(n = 106,828 events)

Crash Linking Sample
(n = 99,233 individuals)

Linked
(n = 43,259)

Not Linked
(n = 55,974)

Figure 1: Diagram of CODES data linkage, updated 2024

Improvements to data preprocessing

There are several major refinements we made to the data processing to improve linkage process and the quality of results in this performance period. First, we added further restrictions to our selection of HCUP records that are available to be linked. We have narrowed our scope to HCUP records that had ECODES matching the CDC's "ICD-10 113 Cause List: Motor Vehicle Accidents." In previous linkages, we restricted down to ECODES starting with "V," which included general motor vehicle injuries that may or may not have been crash-related. This new restriction ensures that we are not linking crashes to hospital records that did not result from crash-related injuries.

Previously, some matched records in the linkage would contain conflicting information because separate linkages were done between crash/SID and crash/SEDD, with the information combined at the end. This could result in a single crash record linking to both an SID and SEDD record where the HCUP records clearly did not correspond to the same person. These false links could be confirmed because SID and SEDD are deterministically related through VisitLink, an identifier that allows multiple SID and SEDD visits to be to the attributed to the same uniquely identified individual. We completely rectified issue of conflicting records by linking the HCUP inpatient (SID) and emergency (SEDD) data deterministically prior to performing the probabilistic between crash and hospital records.

Since there could be multiple visits for a single VisitLink, we grouped together visits that were within 30 days of each other, under a newly created "Event ID". Then, we linked records on the Event level to crash. For example, an individual who was in a crash may have an emergency visit on the day of the crash, be transferred to an inpatient visit on the same day, and have another inpatient visit a week later. In this case, these visits would be grouped together as being related to the single initial crash event. Although our 30-day grouping method may need refinement, it is currently still helpful in ensuring that some related hospital records are not being incorrectly attributed to unique crashes.

The flowchart in Figure 2 below displays the order of the new HCUP data preprocessing steps and it narrows down the records for linkage.

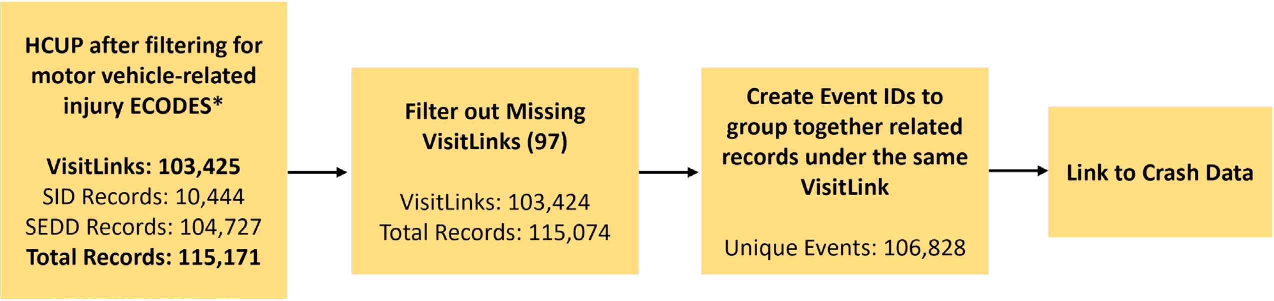

HCUP after filtering for motor vehicle-related injury ECODES*
VisitLinks: 103,425
SID Records: 10,444
SEDD Records: 104,727
Total Records: 115,171
Filter out Missing VisitLinks (97)
VisitLinks: 103,424
Total Records: 115,074
Create Event IDs to group together related records under the same VisitLink
Unique Events: 106,828
Link to Crash Data

Figure 2: Flowchart of HCUP data preprocessing steps

Increasing the Number of Linked Records with Zip Code

Without zip code, the combination of crash month, crash year, year and month of birth, and sex was unique for only 37.3% of the records. With zip code, the combination of the four linkage variables was unique for 92% of the crash records. For HCUP Emergency Department and Inpatient data, the combination of the four linkage variables identified a unique individual for 36% of the records. With zip code, uniqueness in HCUP jumps to 88.3%. This finding indicates that it should be possible to achieve a high-quality linkage using this combination of variables.

Therefore, we believe that there is a large number of true matches that were missed previously. Zip code is a highly discriminatory criteria in determining whether a pair of records match. In the DOT crash data, zip code often comes from the driver's license, while in the HCUP data, it usually reflects the patient's contemporaneous zip code of residence. We expect that a person's zip code from these two sources will often not match. For instance, they may have changed residences since they last updated their driver's license. In past linkages, we were overly strict with zip code, being about as stringent with it as we were the other linking attributes (year, month, birthday, gender). This year, we toggled the probability that we expected zip code to differ between true matches. A higher probability can help us to capture new links that match on all other criteria besides zip code. This "probability different" is available in the LinkSolv software as an input. We performed separate linkages with the probabilities 0.0 (the default), 0.1, 0.2, and 0.5.

Table 2, on the following page, compares results from testing linkages with varying values of "probability different" for zip codes. As the probability increased, so did the number of links between crash and hospital records. Although we gained links that did not match on zip code, the percentage of matching in the other linkage attributes remained consistently high.

Table 2: Linkage quality check under range of prespecified probabilities for mismatching zip code

	Linkage 1	Linkage 2	Linkage 3	Linkage 4
Probability of Different Zip Code	**0.0**	**0.1**	**0.2**	**0.5**
Number Linked Records (% of Crash records linked to HCUP)	43,259 (43.6%)	44,821 (45.2%)	45,923 (46.3%)	48,262 (48.6%)
Linked Attribute	**Records Matching on Attribute N (%)**			
Year	43,259 (100%)	44,821 (100%)	45,923 (100%)	48,262 (100%)
Month	43,259 (100%)	44,821 (100%)	45,923 (100%)	48,262 (100%)
Birth Year & Month	42,969 (99%)	44,517 (99%)	45,615 (99%)	47,953 (99%)
Gender	42,467 (98%)	43,983 (98%)	45,107 (98%)	47,420 (98%)
Zip code	40,384 (93%)	40,425 (90%)	40,396 (88%)	40,393 (84%)

Linkage 5 was ultimately chosen for further analysis in this report because the prespecified settings most accurately reflected our belief about the likelihood of zip code matching between records. Moving forward, further accuracy assessment should be performed on all linkage options.

Assessing Linkage Quality with New Variables: role in crash and vehicle type

One way to assess the quality of the linkage is to compare variables among linked records that correspond but were not used for matching. In this performance period, we created two new variables that could be compared between the HCUP and DOT crash records: "vehicle type" (car, truck, etc.) and "role in crash" (driver, passenger, etc.). The two data sources contain the same information on these attributes in some fashion, but the information needed to be extracted and standardized in categories that would correspond between HCUP and crash so that they could be directly compared. For example, the variable for vehicle type in the crash data has more than 20 categories while the vehicle type information contained by the ECODES in the HCUP data was more general with only six categories. Therefore, the crash data's vehicle type variable had to be condensed to match that in

HCUP. The process of making both new variables required a manual recoding so that they would be able to correspond between data sources. Table 3 below shows the origin of the information used to develop the new Vehicle Type and Role in Crash variables.

Table 3. Data origins for Vehicle Type and Role in Crash.

	Data Source	
	DOT Crash	**HCUP SID/SEDD**
Vehicle Type	Based on existing VEHTYPE and NM_TYPE variables	Based on vehicle information contained in ECODES
Role in Crash	Based on information from SEATING and NM_TYPE	Based on information contained in ECODES

Once these variables were created, they were descriptively assessed in the crash and HCUP data sets separately before the linkage was performed. The two tables below show aggregate statistics for the pre-linkage datasets the HCUP linking sample and the Crash linking sample, as labeled in Figure 1.

In general, the crash data is more descriptive on these variables, with very low proportions of role in crash and vehicle attributes being recorded as unknown (0.8% and <0.1%, respectively). In HCUP, there is both unknown (missing data) and attributes recorded as Unspecified (18% for role in crash, 13% for vehicle type).

Table 4: Role in Crash in linked crash and HCUP data (N = 48,262)

Role in Crash	Crash	HCUP
Driver	37,024 (77%)	30,105 (62%)
Other	206 (0.4%)	--
Passenger	9,313 (19%)	5,934 (12%)
Pedal Cycle	588 (1.2%)	246 (0.5%)
Pedestrian	713 (1.5%)	221 (0.5%)
Person Outside/Person Boarding	40 (<0.1%)	583 (1.2%)
Unknown	378 (0.8%)	2,392 (5.0%)
Unspecified Occupant	--	8,781 (18%)

Table 5: Vehicle Type in linked crash and HCUP data (N = 48,262)

Vehicle Type	Crash	HCUP
Bus	108 (0.2%)	182 (0.4%)
Car	20,414 (42%)	33,186 (69%)
Heavy Transport Vehicle	1,144 (2.4%)	781 (1.6%)
Motorcycle	2,931 (6.1%)	3,283 (6.8%)
Other	136 (0.3%)	--
Other Motor Vehicle	2,195 (4.5%)	--
Other Non-Motorist	75 (0.2%)	--
Pedal Cycle	588 (1.2%)	266 (0.6%)
Pedestrian	713 (1.5%)	556 (1.2%)
Three Wheel Motor Vehicle	22 (<0.1%)	24 (<0.1%)
Truck/Van/SUV	19,923 (41%)	3,947 (8.2%)
Unknown	13 (<0.1%)	--
Unspecified Vehicle Type	--	6,037 (13%)

For the Role in Crash and Vehicle Type variables, agreement was assessed for all linked records (Tables 6 and 7). Among the 48,262 linked records, 65.1% had agreement in "Role in Crash Variable" and 49.2% had agreement in Vehicle Type. Sometimes, the lack of agreement is due to the records having directly conflicting information ("Percent Not Matching" column), and other times, is due to lack of specification or missing data ("Percent Unknown" column). See tables 6 and 7 for a breakdown of this by category.

Table 6 shows that the subgroup of drivers has 72.2% of records matching, with only 4.2% explicitly not matching. Passengers have only 46.9% matching, and 31.0% not matching. Pedestrians and Pedal Cyclists have low matching rates and high not matching rates.

Table 7 shows information on matching in linked records based on Vehicle Type. Cars, Buses, and Motorcycles all have matching rates above 80%, while three-wheel motor vehicles and Trucks/Vans/SUV have very low matching rates.

These tables show that there is great potential in using these to new variables to assess linkage quality. Records that fall under "Not Matching" can be marked for further investigation. Manual review of some of these cases can help pinpoint issues relating or linkage performance or data quality in either source. Furthermore, additional assessment should be done especially on non-motorists (pedestrians, cyclists, etc.) understand why their records often do not line up between data sources.

Table 6. Comparison of matching on Role in Crash from crash to HCUP data.

	Percent Matching	Percent Not Matching	Percent Unknown
Role in Crash (as identified in crash data)			
Driver	72.2	4.2	23.6
Passenger	46.9	31.0	22.1
Pedal Cycle	0.0	79.4	20.6
Pedestrian	17.1	69.0	13.9
Person Outside	0.0	72.5	27.5
Other	0.0	79.6	20.4

Table 7. Comparison of matching on Vehicle Type from crash to HCUP data.

	Percent Matching	Percent Not Matching	Percent Unknown
Motor Vehicles (as identified in crash data)			
Car	82.7	3.9	13.4
Truck/Van/SUV	15.8	71.3	12.9
Heavy Transport Vehicle	51.0	36.8	12.2
Bus	80.6	17.6	1.9
Three Wheel Motor Vehicle	13.6	86.4	0.0
Motorcycle	89.0	6.4	4.5
Other Motor Vehicle	0.0	86.9	13.1
Non-Motorists (as identified in crash data)			
Pedal Cycle	31.1	60.5	8.3
Pedestrian	41.7	50.6	7.7
Other Non-Motorist	0.0	74.7	25.3
Other	0.0	66.9	33.1

Assessment of Race and Ethnicity data quality when compared to Iowa courts data:

Assessment of completeness of race and ethnicity data in hospital data from CODES compared to courts data. A subgroup of charged drivers who were treated in the hospital were examined to assess how hospital data match to courts data on race/ethnicity. This process will also allow for additional quality assessment of the CODES data linkage.

As a proof-of-concept, a set of crash records from 2016-2019, which were already linked by the Iowa Department of Human Rights to deidentified Iowa Courts System citations data from 2016-2019, were then merged with the HCUP records that were linked to crash data.

Tables 8 and 9 show aggregate descriptives on the race/ethnicity variables that are available from each data source.

The courts charge data observes over 533,948 charges from 185,431 people over 148,456 unique crashes. Because the total number of observed charges encompasses each driver's charge history from 2016-2019, many of the charges are not related directly to a crash (but each driver is directly related to a crash).

The race information in the courts data is unfortunately not very high quality. Firstly, there is only one variable containing race/ethnicity information (RACE_CD) and only one value can be recorded. Second, no distinction is made between race and ethnicity. An individual of multiple races and/or ethnicities is only represented by a single value that does not provide full information. A third issue that persists in this data is that an individual's race/ethnicity information may not be coded consistently between charge records collected over time, and sometimes there is conflicting information. For example, a single individual can have several citations over the years. In one instance, their race is recorded as "Caucasian" and for another charge occurring a year later, their race is recorded as "Hispanic." (In Table 8 below, this situation has been categorized as "2+ Races Recorded"). This could indicate that either one of the listed races is an error, that the individual's multiple races or ethnicities were not properly recorded due to the aforementioned limitations.

In contrast, the race and ethnicity data from HCUP sets a higher standard. There are separate variables for race and Hispanic ethnicity, and race has an option for multiracial individuals. There is also a lesser amount of missing data in comparison to the courts data.

Table 8: Descriptive Statistics on Courts Race Variable, on Courts data that is linked to drivers in crashes from 2016-2019

Courts Race	N = 185,431
White	134,174 (72%)
Black	13,591 (7.3%)
Asian	2,889 (1.6%)
Hispanic	1,636 (0.9%)
Native American	462 (0.2%)
Other	2,724 (1.5%)
2+ Races Recorded	3,762 (2.0%)
Unknown	26,193 (14%)

Table 9: Descriptive Statistics on HCUP Race and Hispanic Ethnicity Variables, on HCUP Linking Sample Subset from 2016-2019

HCUP Variable	N = 103,425
Race	
White	83,030 (80%)
African American/Black	12,017 (12%)
Asian	1,260 (1.2%)
American Indian/Alaska Native	695 (0.7%)
Native Hawaiian/Other Pacific Islander	232 (0.2%)
Multiracial/Two or More Races	1,517 (1.5%)
Declined to Answer	456 (0.4%)
Unavailable/Unknown	4,218 (4.1%)
Hispanic Ethnicity	
Hispanic or Latino	6,420 (6.2%)
Non-Hispanic or Latino	93,161 (90%)
Declined to Answer	586 (0.6%)
Unavailable/Unknown	3,258 (3.2%)

There are 17,396 records for crash-involved drivers from 2016-2019 who were linked to both HCUP and courts data. While this represents a small subset, it is useful because it shows a comparison of race and ethnicity made on an individual level. Table 10 shows the aggregate proportions of race and ethnicity attributes for the courts and HCUP independently, and Table 11 shows matching percentages from an individual-level comparison. The "Percent Unknown" column refers to records that have a race or ethnicity listed in the courts data, but were linked to an HCUP record where race/ethnicity information was unknown.

In terms of race, White has the highest percentage of matching (93%) and. Racial minorities, such as Asian and Native American, have lower matching percentages (53% and 34%, respectively) that may warrant further investigation.

This examination of race and ethnicity reveals another method to help assess the quality of the linkage. Records that fall under "Not Matching" can be marked for further investigation. Manual review of some of these cases can help pinpoint issues relating or linkage performance or data quality in either source. However, due to the courts data oversimplifying race and ethnicity, and the fact that only a small subset of crash records link to both HCUP and courts, the use of this as a linkage quality assessment is somewhat limited. This investigation also highlights how greater care should be given to the recording of racial and ethnic data, especially when concerning minority status.

Table 10: Matching Assessment of Race/Ethnicity data

Race/Ethnicity	Courts (N = 17,396)	HCUP (N = 17,396)
White	14,631 (84%)	12,778 (73%)
Asian	182 (1.0%)	168 (1.0%)
Black	1,566 (9.0%)	1,467 (8.4%)
Native American	77 (0.4%)	50 (0.3%)
Pacific Islander	63 (0.4%)	--
Hispanic	--	202 (1.2%)
Other	186 (1.1%)	308 (1.8%)
Refused	76 (0.4%)	1 (<0.1%)
Unknown	615 (3.5%)	2,422 (14%)

Table 11: Race/Ethnicity in Courts Matching to HCUP

Race/Ethnicity (as indicated in courts data)	Percent Matching	Percent Not Matching	Percent Unknown
White	93.0	3.7	3.2
Asian	53.0	39.3	7.7
Black	76.3	20.8	2.9
Native American	34.0	54.0	12.0
Hispanic	63.4	36.6	0.0
Other	6.8	80.2	13.0
Refused	0.0	100	0.0

Summary

During this performance period, the UI team established new collaborations with personnel at the Iowa to obtain updated Z table (crash) data and new Accident Processing System (APS; personally identifiable data from crashes) data from the Iowa DOT. Through this collaboration, the UI helped assess aspects of crash data quality. The team's efforts also supported the Iowa DOT's objectives of establishing automated processes, which should improve accessibility to crash Z tables and PII in the future. With the data processes becoming more stable, the UI team will be able to perform more assessments of data accuracy, completeness, and integration in the next performance period.

The crash outcome data linkage

- explored potential strategies for improving the quality of the probabilistic linkage of crash data and hospital data, and

- assessed the level of agreement between a number of variables in the crash, hospital, and judicial data.

Appendix A: 2024 CODES Project Personnel

Key Investigators

Cara J. Hamann, MPH, PhD; Principal Investigator
Associate Professor
The University of Iowa Injury Prevention Research Center

Michelle Reyes
Senior Research Associate
University of Iowa Driving Safety Research Institute

Jonathan Davis
Assistant Professor
The University of Iowa Injury Prevention Research Center

Data Analysts

Stephanie Jansson

Research Team

Ryan Dusil
Elizabeth O’Neal
Gilsu Pae

GTSB Contact

Mick Mulhern

Appendix B: Research Presented at ATSIP Traffic Records Forum

Pae G, Reyes M, Hamann C (August 2024). Exploring varied definitions of rurality and their impacts on key factors of road traffic crashes, ATSIP Traffic Records Forum, San Diego, CA.

In 2021, according to National Highway Traffic Safety Administration (NHTSA), rural areas exhibited a fatality rate per vehicle-traveled mile (VMT) 1.5 times higher than that in urban areas. Rurality has been a priority for crash injury prevention, given the higher rates of crash involvement and injury/fatality as well as distinct road infrastructure, driving conditions, and crash patterns compared with urban areas. However, the classification of rurality varies across crash data dictionaries and motor vehicle collision (MVC) research. This inconsistency is an obstacle to understanding the impact of rurality on crashes and makes it challenging to compare statistics based on rurality for comprehensive road safety policy development. This study aims to examine how the varying definitions of rurality can affect the findings from MVC research. Our approach included data review from several states and a comparison of different definitions of rurality for one state's crash data.

We gathered and examined crash data and associated data dictionaries from 12 states (CO, IL, IA, KS, MN, MO, NE, ND, OH, SD, UT, WY) spanning from 2010 to 2020. Additionally, we analyzed Iowa crash data to compare the different urban proportions of key factors for crash studies, based on 6 different definitions of rurality. We examined 1) incorporated urban boundaries which is a unique rural/urban classification used by Iowa, 2) Federal Highway Administration (FHWA) urban areas, 3) urban municipality boundaries with a population of more than 2,500, 4) Census urban areas, 5) Rural-Urban Commuting Area (RUCA) urban areas, and 6) Rural-Urban Continuum Code (RUCC) urban areas. The first three definitions are used in crash data, and the last three definitions are often employed in academic research. The last version of RUCA code was published in 2010, and RUCC was published in 2013 and 2023. Therefore, we selected the year 2013 as a sample for comparison.

Out of the 12 states, 8 states (CO, IL, IA, MN, MO, NE, ND, SD) provided rural/urban classification in the crash data with definitions of rurality, such as adjusted urban boundaries approved by FHWA, roadway functional classifications (e.g., urban arterial), incorporated boundaries, and urban municipality boundaries. In four states (IL, MN, NE, ND), either the definitions changed, or they stopped providing the rural/urban classification between 2010 and 2020. Recently, many states are likely to focus on providing coordinates of crash locations rather than offering limited definitions of rurality, allowing researchers to code the location according to definitions best fit for specific research purposes. In 2020, 10 out of the 12 states provided coordinates of crash locations with less than 10% missing, excluding CO and NE.

In 2013, there were 288 fatal crashes in Iowa (2 cases were removed due to missing coordinates). Of these, 61(21%) fatal crashes occurred in Census urban areas, 78 (27%) in FHWA urban areas, 84 (29%) in incorporated urban areas, 138 (48%) in RUCA urban areas, 266 (92%) in RUCC urban areas, and 269 (93%) within urban municipality boundaries. Additionally, there were 1,940 crashes involving motor vehicle drivers in motion who were reported not to be using a seat belt. Of these unprotected driver-related crashes, 994(51%) occurred in Census urban areas, 1,054(54%) in FHWA urban areas, 1,144 (59%) occurred in incorporated urban areas, 1,093 (56%) in RUCA urban areas, 1,832(94%) in RUCC urban areas, and 1,856(96%) within urban municipality boundaries. The proportion of fatal crashes designated as urban areas ranged from 21% to 93% depending on the definitions of rurality, while the proportion of unprotected driver-related crashes ranged from 51% to 96%.

Rurality is an essential factor in crash studies and can be defined differently for varying research purposes. Given the potential impact on analytical results, selecting appropriate definitions for specific purposes is critical, and declaring definitions used, especially in the crash data dictionary when a state has applied one or more definitions, is necessary for the audience to clearly understand the results. Moreover, additional studies are needed to explore which definitions are most appropriate for certain research topics.