Linkage of NC Trauma Registry and NC Crash datasets

Sharon Schiro, PhD[1]
Mike Dolan Fliss, PhD, MSW, MPH[2]
Anna Waller, ScD[3]
Erika Redding, MSPH[3]
Katherine Harmon, PhD, MPH[4]

Statement of Funding: Funding for this project was provided by the NC Governors Highway Safety Program.

Acknowledgements: Our team would also like to thank our project's advisors and staff, Katherine Peticolas, Clifton Barnett, Dennis Falls, and Amy Ising from the Carolina Center for Health Informatics; Steve Marshall from the UNC Injury Prevention Research Center; Bevan Kirley, Nancy Lefler, Eric Rodgman, and Laura Sandt from the UNC Highway Safety Research Center; and Alan Dellapenna from the NC Division of Public Health.

[1] Department of Surgery, University of North Carolina School of Medicine
[2] University of North Carolina Injury Prevention Research Center
[3] Carolina Center for Health Informatics, Department of Emergency Medicine, University of North Carolina School of Medicine
[4] University of North Carolina Highway Safety Research Center

Table of Contents

Introduction

This project sought to link data from crash records and the NC Trauma Registry (NCTR) for people injured in NC crashes in 2018. Methods were developed, implemented, evaluated and refined. The resulting linkage was considered successful and produced a linked data set that we hope to use for further analysis. Overall, 55% of NCTR records for patients who were injured in a motor vehicle crash linked to a crash record and were deemed a good link based on our evaluation of the linked data. Table 1 presents an overall summary of the linkage results.

Table 1: Overall Linkage Summary

NCTR records			**NCTR records with crash mechanism of injury**		
# records	# linked	% linked	# records	# linked	% linked
35,386	5,735	16.2	10,234	5,630	55.0

Crash records			**Crash records with reported injury (A,B,C) or fatality (K)**		
# records	# linked	% linked	# records	# linked	% linked
832,058	5,735	0.7	227,927	5,653	2.5

Methodology

This linkage project involved developing and testing methods for linking the 2018 NC Crash and NC Trauma Registry (NCTR) datasets. All linkages were done with SAS (proc SQL). Variables from the crash and NCTR datasets used in the linkages are listed in Table 2 below. Linkage results in terms of counts of records matched are summarized in **Error! Reference s ource not found.**4.

Inclusion Criteria

All NCTR records for 2018 where the patient was not transferred to another trauma center were used in the linkage process. All records in the NCTR that were the result of a crash injury were flagged as “is crash”. Results for incidents resulting from a crash are independently reported, since these records should have a greater probability of linking correctly. Patients transferred to another trauma center were excluded from the dataset because we did not want patients represented twice.

All crash reports were included, structured to represent each person involved in a crash. People involved in crashes were eligible for linkage regardless of the level of injury recorded in the crash report.

Variables used for linkage

Variables used for linkage are presented in Table 2. These variables were harmonized across the two datasets to increase the odds of linkage. In this work, “harmonization” means making sure two variables meant to represent the same concept across two datasets (e.g. date of birth in crash and NCTR) are cleaned and represented in the same way so that they have the best chances of meaningfully matching. See section on data harmonization below.

Table 2: Variables used for linkage

Variable category	**Crash**	**NCTR**	**Definition/notes**
Age	L_age_num, Crash_Age	L_age_num_n	Non-numeric value of age
DOB	L_dob_date	L_Dob_Date_n	DOB
DOB	L_dobmd_fct	L_dobmd_fct_n	Month/day of DOB
Gender	L_gender_fct	L_gender_fct_n	Gender
Race and ethnicity	L_raceeth_fct	L_raceeth_fct_n	Race/ethnicity combined
Fatal crash	L_fatal_lgl	L_fatal_lgl_n (based on ED and hospital dispositions)	Did the patient die from injuries sustained in the crash? Y/N
Injury existence	L_isinj_lgl	L_isinj_lgl_n (based on existence of an injury diagnosis code, ie, an ICD-10 diagnosis code starting with letters S, T, U, V, W, X, or Y)	Does the patient have an injury? Y/N
Injury severity	L_issevere_lgl	L_issevere_lgl_n (ISS>15)	Is the patient's injury considered severe? Y/N
Is the injury due to a crash? Or is the event a crash?	L_iscrash_lgl	L_iscrash_lgl_n (flagged as a crash if ICD-10 injury mechanism in V02-V09, V12-V14, V20-79, X82, Y03)	Injury due to MVC event? Y/N
Position in vehicle	L_Crash_pos_fct	L_crash_pos_fct_n (based on PositionInVehicle field (primary) and InjuryMechanism1 field (secondary).)	Patient's position in or outside vehicle at time of crash
Residence state	L_rstate_fct	L_rstate_fct_n (PatientAddressState)	Patient state of residence
Crash/Residence county	L_rcounty_fct	L_rcounty_fct_n (PatientAddressFipsCounty – converted to text county)	County of crash/Patient county of residence
Residence zip	L_rzip5_fct	L_rzip5_fct_n (PatientAddressZip)	Patient Zip code of residence (5-digit)
Residence zip	L_rzip3_fct	L_rzip3_fct_n	First 3 digits of patient residence zip code
Residence city	L_rcity_fct	L_rcity_fct_n (PatientAddressCity)	Patient city of residence
Crash/injury state	L_cstate_fct	L_cstate_fct_n (InjuryAddressState)	State in which crash/injury occurred

Variable category	Crash	NCTR	Definition/notes
Crash/injury county	L_ccounty_fct	InjuryCounty (InjuryAddressFipsCounty)	County in which crash/injury occurred
Crash and injury dates	L_acc_date, AccDate	L_acc_date_n (InjuryDate)	Date of crash/injury
Crash and injury dates	L_accmd_fct	L_accmd_fct_n	Month/day of crash/injury date
Crash and hospital admission date	AccDate	Admission date (ED arrival or Facility arrival date)	Date of crash / date of arrival to hospital
Hospital	L_BestHospName	L_BestHospName_n or l_BestRfHospName_n (facilityid, RfFacilityName)	Harmonized facility names. Compared to both facility on record and referring facility names

Harmonization

The following steps were taken to harmonize the data between the two data sources.

- Age: In NCTR, age is calculated from the DOB and facility arrival date. In crash data, age was an integer number ranging from 1 to 999. If the age is > 120, then the age was set to blank for this project. If age was 0 to 0.9, then age was set to 1.
- Dates: In the crash data, dates are stored as character fields, so character versions of all NCTR date fields used for linkage were generated.
- DOB: In both data sets, birthdates > 12/31/2018 were set to blank. Since 1900 and January 1st are commonly used in place of unknown birthdates, DOBs containing these values also were set to blank.
- Gender: All values not equal to F or M were set to blank in the NCTR data, to match the Crash data. Note that the trauma data system will include “non-binary” for gender in the future, which will increase the mismatches unless the crash dataset adds a similar value.
- City: InjuryAddressCity (NCTR) and CityName (Crash) are free-text fields. These city names were standardized to a common list of city names.
- Race/Ethnicity: Race and ethnicity are stored as separate fields in NCTR, but as one field in Crash. NCTR also captures 2 race values for each person. The harmonized field l_raceeth_fct for NCTR data was generated using the two NCTR race fields and the ethnicity field using the following hierarchy: Native American, Hispanic, Asian, Black, Other, White. Pacific Islander was recoded to “Other”, since the “Pacific Islander” value is not captured in the Crash dataset.
- Hospital names: Both the name of the trauma center providing the NCTR record and the name of any referring facility were considered in the linkage. These facility names, as well as the hospital/facility names captured in the Crash data, were standardized to the same list. The Crash hospital name field is a free-text field; NCTR trauma center and referring facility names come from pick-lists. Some, but not all, of the crash record free-text names could be standardized to facility names on the NCTR pick-lists; some of the values in the crash record hospital name field were not hospital names or were not

complete enough to allow unique identification. Others were facility names that are also used by facilities in neighboring states and were therefore not harmonized since we could not be sure which facility it was.

Linkage methodology specifics

We used a cascading hierarchical deterministic (CHD) linkage approach. CHD consists of a series of linkage passes with variations on deterministic exact matches and filters (for fuzzy, window, or distance-based linking) for corresponding linkage variables. After each step, the linked records are removed from subsequent linkage passes, creating a cascading hierarchical linkage with decreasing specificity of linkage match requirements. A total of 34 different linkage passes were implemented and evaluated, with a final process including 20 passes, after evaluating the success of each linkage pass. Only linkage passes that resulted in >80% good matches based on our evaluation were included in our final linkage methodology.

Records were matched on variables representing a person (demographics, crash-related), place (residence, crash, and healthcare), and date (date of crash, date of arrival to hospital). These were flexed in various combinations, as shown in Tables 3 and 4.

Table 3: Cascading linkage pattern

Match pass # and criteria	Person (Demo)	Person (Crash)	Place (Resid)	Place (Crash)	Place (Healthcare)	Date
1. All match	Exact	Exact	Exact	Exact	Exact: Trauma center	Exact
2. Referring Facility	Exact	Exact	Exact	Exact	Exact: Trauma center or referring facility	Exact
3. Drop RCity	Exact	Exact	Drop RCity	Exact	Exact: Trauma center or referring facility	Exact
4. Drop R Zip5, keep R Zip3	Exact	Exact	Drop Zip5	Exact	Exact: Trauma center or referring facility	Exact
5. Drop R Zip5 and R Zip3	Exact	Exact	Drop R Zip5 and R Zip3	Exact	Exact: Trauma center or referring facility	Exact
6. Drop R County	Exact	Exact	Drop county	Exact	Exact: Trauma or referring facility	Exact

Match pass # and criteria	Person (Demo)	Person (Crash)	Place (Resid)	Place (Crash)	Place (Healthcare)	Date
7. Drop C County	Exact	Exact	Exact	Drop county	Exact: Trauma or referring facility	Exact
8. Drop R City, R Zip5, and R Zip3	Exact	Exact	Drop city, zip5, zip3	Exact	Exact: Trauma or referring facility	Exact
9. Drop R City and R County	Exact	Exact	Drop city, county	Exact	Exact: Trauma or referring facility	Exact
10. Drop R City and C County	Exact	Exact	Drop city	Drop county	Exact: Trauma or referring facility	Exact
11. Drop R City, R County, and C County	Exact	Exact	Drop city, county	Drop county	Exact: Trauma or referring facility	Exact
12. Drop Race_Ethnicity	Drop race_ethnicity	Exact	Exact	Exact	Exact: Trauma or referring facility	Exact
13. Drop Gender	Drop gender	Exact	Exact	Exact	Exact: Trauma or referring facility	Exact
14. Drop DoB and DoB_MD	Drop DOB, DOB_MD	Exact	Exact	Exact	Exact: Trauma or referring facility	Exact
15. Drop Age Number	Drop age	Exact	Exact	Exact	Exact: Trauma or referring facility	Exact
16. Drop Crash Pos	Exact	Exact	Exact	Drop crash position	Exact: Trauma or referring facility	Exact
17. Drop Crash and Crash Pos	Exact	Exact	Exact	Drop "iscrash" and crash position	Exact: Trauma center or referring facility	Exact
18. Drop Fatal and Severe	Exact	Exact	Exact	Drop fatal and severe	Exact: Trauma center or	Exact

Match pass # and criteria	Person (Demo)	Person (Crash)	Place (Resid)	Place (Crash)	Place (Healthcare)	Date
					referring facility	
19. Allow Accident Date to Flex +1 From Injury Date	Exact	Exact	Exact	Exact	Exact: Trauma center or referring facility	Flex +1 day from injury date
20. Acc date=Injury Date, DOB, and Trauma Facility Matching	DOB only	Drop all	Drop all	Drop all	Exact: Trauma center or referring facility	Exact

Match assessment

Linkage results are described in **Error! Reference source not found.**4. Some of the m atches reviewed were indeterminate in match quality; these were not flagged as matches or non-matches. Hence, the number of matches reviewed does not always equal the sum of the "match" and "non-match" review columns.

Matches were assessed as follows:

- A spreadsheet of all matches was created from the SAS table. Columns were organized so that similar variables from the two datasets were side-by-side.
- A subset of records was chosen for review: 3% of the total number of records with a minimum of 5 records. Three percent was used because it provided a reasonable number of records to be hand-reviewed.
- Records were hand-reviewed, looking for the level of discrepancy between non-matching columns. For example, a non-match on crash cities of Mebane and Graham was considered to be less significant than a non-match of Mebane and Asheville.
- Each record was flagged as appearing to be a good match or not – or as indeterminate. Because we do not have fully identifying information (e.g., name), we cannot determine match quality with absolute certainty.

Two matches in the "All match" group in Table 4 actually matched on all fields that were used for the match, but were duplicates; i.e. the crash record matched to two different NCTR records. Since we were unable to determine which match was correct, both were dropped from further analysis.

Final match process and results

The counts in Table 4 were generated by including only those linkage passes that resulted in >80% good matches based on our match assessment.

Table 4: Linkage process and results-final match process and assessment results

Match pass # and criteria	Match count	Non-dup trauma matches	Non-dup matches (trauma crashes)	# reviewed	Review result = match	Review result=non-match	% reviewed that matched	Final count: trauma matches	Final count: trauma-crash matches
1. All match	557	555	555	25	23	0	92	555	555
2. Referring Facility	143	143	143	66	66	0	100	143	143
3. Drop RCity	61	61	61	7	7	0	100	61	61
4. Drop R Zip5, keep R Zip3	54	54	54	6	6	0	100	54	54
5. Drop R Zip5 and R Zip3	1	1	1	1	1	0	100	1	1
6. Drop R County	214	213	213	210	209	0	99.5	213	213
7. Drop C County	66	66	66	6	6	0	100	66	66
8. Drop R City, R Zip5, and R Zip3	120	120	120	5	5	0	100	120	120
9. Drop R City and R County	16	16	16	16	16	0	100	16	16
10. Drop R City and C County	4	4	4	4	4	0	100	4	4
11. Drop R City, R County, and C County	59	59	59	5	5	0	100	59	59
12. Drop Race_Ethnicity	89	89	89	12	12	0	100	89	89
13. Drop Gender	4	4	4	4	4	0	100	4	4
14. Drop DoB and DoB_MD	26	26	26	17	16	0	94.1	26	26
15. Drop Age Number	2	2	2	2	2	0	100	2	2
16. Drop Crash Pos	264	264	264	8	8	0	100	264	264
17. Drop Crash and Crash Pos	11	11	0	5	5	0	100	11	0
18. Drop Fatal and Severe	921	921	921	29	28	0	96.6	921	921
19. Allow Accident Date to Flex +1 From Injury Date	1	1	1	1	1	0	100	1	1

20. Acc date=Injury Date, DOB, and Trauma Facility Matching	3133	3122	3031	122	99	12	81.1	3122	3031

Review of non-matches

- 89.4% of the NCTR records that did not match a Crash record were non-crashes in NCTR.
    - 98.2% of total linkages were marked in the trauma registry data as crashes.
- A small subset of the non-matches were pulled from NCTR and Crash for review. This subset was chosen so that the records matched on crash/injury date, gender, and DOB, and they were labeled as crashes in NCTR.
    - Of these records, seven out of seven were non-transports in the Crash data. AccSev in Crash data were: Fatal (5), Class C (1), and Unknown (1).
    - One of the seven non-matches reviewed matched on everything except residence county; the crash record, however, matched to two NCTR records, thus being a duplicate match and dropped from the matched group. One of the two records had a large discrepancy in residence county (Orange vs Caldwell) in terms of distance between counties.
- Future efforts should include evaluation of at least one linkage pass that does not require matching a hospital/facility name in the crash report and NCTR data to allow patients who do not require EMS transport from the scene of the crash but who sought treatment themselves to link. While there may be few of these cases, we know from other linkage efforts that many people seek emergency department care on their own following crashes, after receiving no EMS care at the scene of the crash. Some of these may end up in the trauma registry data so allowing the possibility of these linkages is important.

Table 5: Metrics on non-matching records

Metric	**Crash**	**NCTR**
Crash / Injury state	100% NC	NC=71%, missing=19.2%, SC=7.1%, VA=1.9%. Other states (GA, IN, PA, TN, WV): < 0.5% each.
Crash / Injury county	Mecklenburg (14%), Wake (12.6%), Guilford and Forsyth (5%). Remaining counties < 5% each.	Missing=40.9%, of which 17.2% were in NC
Severity – not severe	96.6%	71.1%
Injury status = injury	100%	97.2%
Fatal	0.1% (based on is_fatal_lgl field)	3.1% (based on hospital and ED disposition)
Race/Ethnicity	White=56.4%, Black=29.4%, Hispanic=8.3%, Other=2.4%, missing=1%	White=62.3%, Black=25%, Other=9%, missing=1.5%

Comparing linkages and non-linkages

Initial descriptive analyses of the linked crash-NCTR data are underway. Select results are presented in Tables 6-8, below. Not surprisingly, a higher proportion of those with A and B level injuries noted in the crash report were represented in the linked crash and NCTR records than those with either C level or fatal injury noted in the crash report (Table 6). Although pedestrians represent less than half of 1% of persons reported in crash records, they accounted for over 7% of those who linked to a NCTR record (Table 7). NCTR patients whose records linked to a crash report were more likely to have died in the ED or been admitted to the ICU than those who did not link to a crash record (Table 8). This is just a sampling of the sorts of analyses we can undertake with these linked crash report and NCTR record data.

Table 6: Crash injury severity by linkage status of crash data to NCTR data

Injury severity (Crash data)	**Linked to NCTR data**		**Did not link to NCTR data**		**Full Crash dataset***	
	N	%	N	%	N	%
A	1966	34.3	6973	0.8	8943	1.1
B	2159	37.6	53638	6.5	55801	6.7
C	922	16.1	162258	19.6	163183	19.6
Fatal	606	10.6	2827	0.3	3433	0.4
PDO	74	1.3	595010	72.0	595085	71.5
Unknown	8	0.1	5605	0.7	5613	0.7
Total	5735		826311		832058	

* There were 12 records that were not considered true links because they matched to more than one trauma or crash record. These records are not counted in the “Linked” and “Did not link” columns in Table 6.

Table 7: Role of injured person by linkage status of crash data to NCTR data

Role of injured person (Crash data)	**Linked to NCTR data**		**Did not link to NCTR data**		**Full crash dataset (including 12 records mentioned above)**	
	N	**%**	**N**	**%**	**N**	**%**
Driver	3371	58.8	589470	71.3	592846	71.2
Motorcyclist	802	14.0	4295	0.5	5099	0.6
Other	13	0.2	384	0	397	0
Passenger	1044	18.2	227750	27.6	228797	27.5
Pedal cyclist	89	1.6	849	0.1	938	0.1
Pedestrian	411	7.2	2968	0.4	3381	0.4
Unknown	5	0.1	595	0.1	600	0.1
Total	5735		826311		832058	

Table 8: NCTR ED disposition by linkage status of crash data to NCTR data

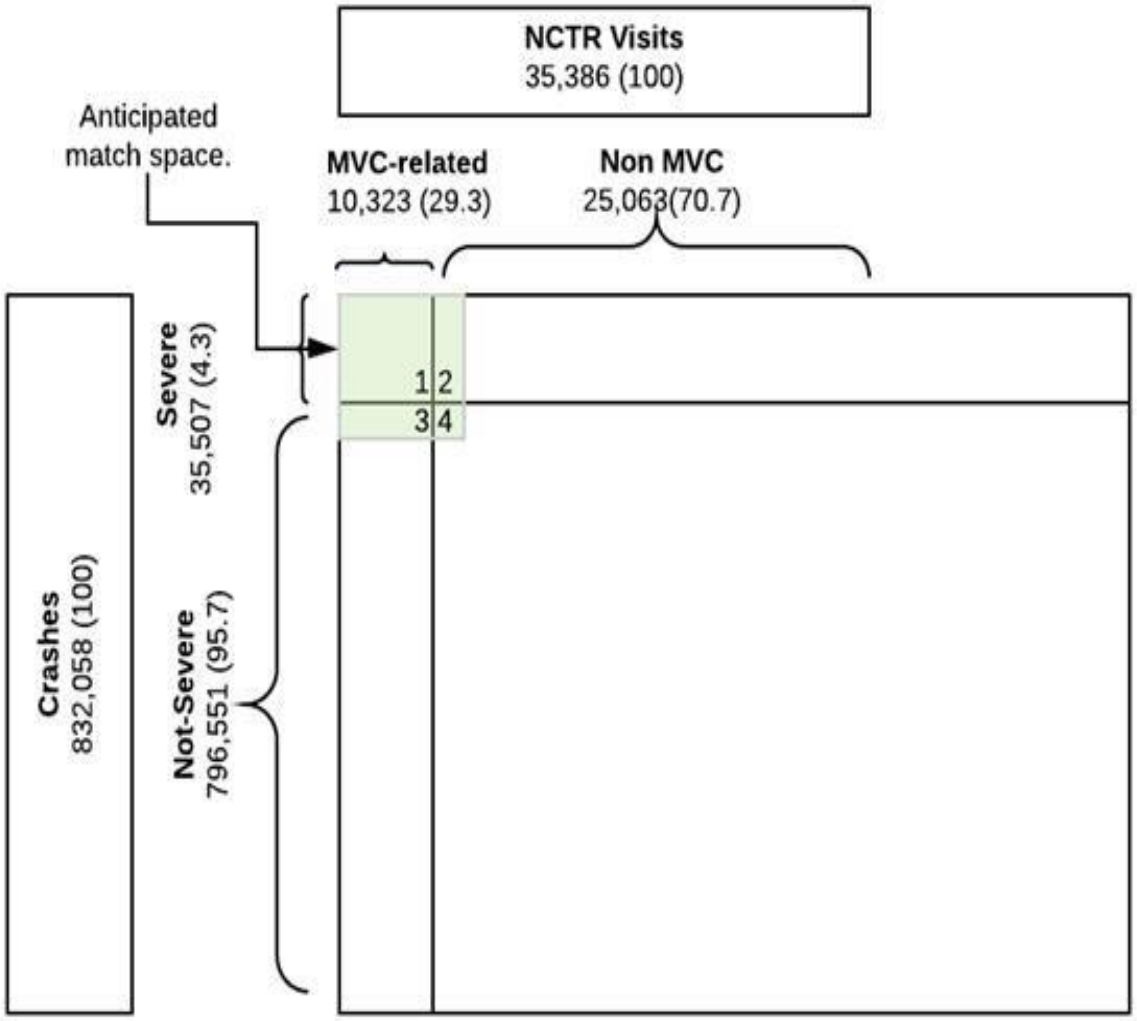
ED disposition (NCTR data)	Linked to crash data		Did not link to crash data	
	N	%	N	%
Against medical advice	20	0.4	55	0.2
Correctional facility	12	0.2	48	0.2
Hospital Floor	2167	37.8	13720	46.7
Home	751	13.1	2740	9.3
Home with services	1	0	8	0
Hospice	0	0	5	0
ICU	1185	20.7	3757	12.8
Labor & Delivery	38	0.7	71	0.2
Mental health facility	3	0	32	0.1
Morgue	121	2.1	211	0.7
Nursing home	0	0	2	0
Observation unit	176	3.1	1041	3.6
OR	816	14.2	3638	12.4
Pediatric ICU	68	1.2	349	1.2
Skilled Nursing Facility	0	0	7	0
Special procedure room	18	0.3	29	0.1
Step-down unit	117	2.0	449	1.5
Telemetry unit	203	3.5	1229	4.2
Transferred	17	0.3	174	0.6
Missing	22	0.4	1793	6.1
Total	5735		29358	

Recommendations

- Pick-lists in the crash dataset for hospital name would greatly improve linkages. Some of the free-text hospital names were able to be harmonized, but some could not because of vagueness of the name.
- Separate race and ethnicity data points in the crash report data would be an important improvement. Capturing race and ethnicity as separate variables is considered best practice for demographic variables.
- Including at least one linkage pass that does not require matching a hospital/facility name in the crash report and NCTR data would allow patients who did not require EMS transport from the scene of the crash but who sought treatment themselves to link. While there may be few of these cases, we know from other linkage efforts that many people seek emergency department care on their own following crashes, after receiving no EMS care at the scene of the crash. Some of these may end up in the trauma registry data so allowing the possibility of these linkages is important.

Appendix

Figure 1. Match space[5]



[5] Frequency and % presented, unless otherwise noted.

Figure 2. NCTR Inclusion criteria