North Carolina Crash Injury Surveillance System (NC-CISS)

Year 1 Final Report

Report Prepared By:

Mike Dolan Fliss
University of North Carolina Injury Prevention Research Center
North Carolina Division of Public Health, Injury and Violence Prevention Branch

Katherine J. Harmon
University of North Carolina Highway Safety Research Center

Kathy Peticolas
North Carolina Division of Public Health, Injury and Violence Prevention Branch

Anna E. Waller
Carolina Center for Health Informatics
University of North Carolina School of Medicine, Department of Emergency Medicine

Submitted August 27, 2020
Revised October 27, 2020
Revised March 16, 2021

This report was supported by the National Center for Injury Prevention and Control of the Centers for Disease Prevention and Control (CDC) under award number CE16-1602. The content is solely the responsibility of the authors and does not necessarily represent the official views of the CDC.

This report is a deliverable for Contract Number 00039605 from the Injury and Violence Prevention Branch of the NC Division of Public Health.

Table of Contents

Table of Figures

Collaboration and Funding

The project was a collaboration between North Carolina Department of Health and Human Services Injury and Violence Prevention Branch (NCDHHS IVPB), University of North Carolina Injury Prevention Research Center (UNC IPRC), the UNC Carolina Center for Health Informatics (CCHI), and the UNC Highway Safety Research Center (HSRC).

The funding source for this project was based on the year four motor vehicle supplement of the five year CDC grant Core State Violence and Injury Prevention Program (Core SVIPP), which was awarded to the North Carolina Department of Health and Human Services Injury and Violence Prevention Branch (NCDHHS IVPB). The University of North Carolina at Chapel Hill Injury Prevention Research Center (UNC IPRC) was contracted (contract number 00039605) to complete the work.

Acknowledgments

We would like to acknowledge our Project Team members: Clifton Barnett, Ingrid Bou-Saada, Alan Dellapenna, Jr., Dennis Falls, Amy Ising, Nancy Lefler, Steve Marshall, Scott Proescholdbell, and Eric Rodgman. We would also like to acknowledge Lana Deyneka and Zach Faigen from the Communicable Disease Branch of the North Carolina Division of Public Health, who provided permission for the NC DETECT emergency department visit data used in the data linkage, and Matt Avery of the State Center for Health Statistics, who provided the death certificate data used in the data linkage.

Finally, we would like to thank Dr. Larry Cook for sharing his expertise and time for this data linkage project.

NC DPH Data Attribution & Disclaimer

NC DETECT is a statewide public health syndromic surveillance system, funded by the NC Division of Public Health (NC DPH) Federal Public Health Emergency Preparedness Grant and managed through collaboration between NC DPH and UNC-CH Department of Emergency Medicine's Carolina Center for Health Informatics. The NC DETECT Data Oversight Committee does not take responsibility for the scientific validity or accuracy of methodology, results, statistical analyses, or conclusions presented.

List of Abbreviations

CCHI	Carolina Center for Health Informatics
CDC	Centers for Disease Control and Prevention
CHD	Cascading hierarchical deterministic linkage
DOT	Department of Transportation
ED	Emergency Department
GHSP	Governor's Highway Safety Program
HSRC	Highway Safety Research Center
IPRC	UNC Injury Prevention Research Center
IVPB	NC DHHS Injury and Violence Prevention Branch
MVC	Motor vehicle crash
NC	North Carolina
NC DHHS	North Carolina Department of Health and Human Services
NHTSA	National Highway Traffic Safety Administration
SCHS	State Center for Health Statistics
SVIPP	State Violence and Injury Prevention Program
UNC	University of North Carolina

Executive Summary

We have concluded the first year of the North Carolina Crash Injury Surveillance System (NC-CISS) project. The goal of the project was to create a sustainable linkage methodology to link crash report data with health outcome data. To fulfill this goal, we evaluated different linkage methodologies before creating final linked datasets. The linked datasets will provide a more complete picture of the circumstances and outcomes associated with motor vehicle crash injuries in North Carolina.

This Year 1 final report addresses the project requirements and deliverables in the contract between UNC and NC DHHS and describes data linkage activities, deviations from our implementation plan, data linkage results, project barriers and facilitators and next steps.

Descriptions of Source Data

The system in Year 1 is comprised of two health datasets linked with the crash report data: death certificate data and emergency department (ED) visit data from the North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT). All data were from 2018.

Crash Report Data

There were 832,058 persons involved in 355,571 crashes in the 2018 crash report data.

- 0.4% (N=1,468) of persons were reported as fatally injured in crashes.
- 16% (N=129,774) of crash victims were reported as having non-fatal injuries ranging from possible injury to serious injury.
- 84% of all persons in crashes were reported as having no injury (N=689,423) or were missing an injury designation (N=11,392).

Death Certificate Data

There were 94,867 deaths recorded in North Carolina in 2018. 8% (N=7,680) had an injury cause of death code and 1.8% (N=1,677) had a transportation-related cause of death code.

NC DETECT Emergency Department Visit Data

There were 5,084,987 emergency department (ED) visits in the 2018 dataset, among which 803,649 (16%) had at least one injury diagnosis code. 126,734 (2%) contained at least one transportation-related external cause of injury code. In North Carolina, the reporting of external cause of injury codes is not mandatory; therefore, some transportation crash injury-related ED visits may be missing these codes.

Evaluating Linkage Methodologies

We tested three different data linkage methodologies for the linkage of crash and death certificate data:

a. Recursive partitioning trees (RPT) in R
b. Probabilistic linkage using LinkSolv
c. Cascading hierarchical deterministic linkage with block filtering in R (CHD)

For crash and ED visit data linkage, we tested only the last two methods (b and c above). Because RPT and LinkSolv required more computing resources and may be more challenging

for multi-disciplinary audiences, we focused on refining the CHD linkage methodology to achieve our desired results.

We compared the results of the linkage using the different methodologies and analyzed the results for missingness and data mismatch patterns. For the crash and death certificate linkage, we also created a hand-reviewed set of 1,483 matched records to use as a comparison set. All analyses using the 2018 death data will use the hand-reviewed dataset. Due to the amount of resources required for creating the hand-reviewed dataset, future years of analyses will use the CHD method.

We applied a similar process to the crash and ED data linkage. We also sampled 1% of the results from LinkSolv and CHD linkage methods and used the verification of those results to inform the process. We also reviewed LinkSolv matches with ED visits four or more days after the crash and subsequently revised the algorithms for both methodologies based on the results.

Linkage Results

Linked Crash and Death Certificate Data

The CHD-generated dataset matched 1,375 death certificates to persons from 1,274 crash events.

- 92% (N=1,355) of crash-reported fatalities were matched with death certificates.
- Less than 1% (N=13) of persons reported as having non-fatal injuries, ranging from possible injury to serious injury, were matched with death certificates.
- Less than 1% (N=7) of persons reported as having no injury or were missing an injury designation were matched with death certificates.

Linked Crash and NC DETECT Emergency Department Data

The CHD-generated dataset matched 83,998 emergency department visits to persons from 63,681 crash events.

- 10% (N=158) of crash-reported fatalities matched with ED visits.
- 40% (N=52,502) of persons reported as having non-fatal injuries, ranging from possible injury to serious injury, were matched with ED visits.
- 4% (N=31,338) of persons reported as having no injury or were missing an injury designation were matched to ED visits related to the crash.

Note: Linkage totals reflect the linked datasets at the time of this report. Future analysis using later iterations of the linked datasets may produce different totals due to additional data validation and linkage evaluation.

Next Steps

We plan to analyze the data linkage and linked datasets and share the results with local, state, and federal partners in transportation, planning, and public health. We will also leverage state funding to create a data dashboard and assemble a research advisory board to help identify important research and policy questions to be addressed with the linked data system. If funding is available, we plan on adding additional years of linked data and, eventually, additional data sources.

Project Requirements

Activity 1: Develop a description of datasets used for data linkage by October 15, 2019

Tasks

a. Obtain written approvals from the Communicable Diseases (CD) Branch and the State Center for Health Statistics (SCHS) in order for (1) NC DETECT emergency department (ED) visit data and (2) SCHS death certificate/vital records data to be linked to NC motor vehicle crash data.
b. Complete data documentation that assesses data source 1: NC DETECT ED visit data.
c. Complete data documentation that assesses data source 2: SCHS death certificate/vital records data.
d. Submit descriptive data documentation to IVPB.

Narrative

We capitalized on established relationships and previous work to develop descriptions of datasets for data linkage. Approval letters from the Communicable Diseases Branch and the State Center for Health Statistics were included as part of our proposal to the CDC for this project. We had already created data documentation for the two health data sources through a previous project funded by the NC Governor's Highway Safety Program. These documents were then used to create a list of probable data elements that could be used to link the health data to crash report data.

Deliverables

Deliverable	**Due date**	**Date completed**	**Location**
Written approval from Communicable Diseases Branch to use NC DETECT Emergency Department (ED) visit data and from the State Center for Health Statistics (SCHS) to use death certificate data	9/17/2019	10/10/2019	Appendix A: Written approval to use NC DETECT and death certificate data
Completed and submitted data documentation for NC DETECT ED visit data and SCHS death certificate data	10/15/2019	10/10/2019	Appendix B: Data Documentation for NC DETECT Appendix C: Data Documentation for Death Certificate Data Appendix D: Data Linkage Data Elements

Activity 2: Develop data linkage methodology by February 27, 2020

Tasks

a. Obtain data use approval for the overall linkage project from the UNC Institutional Review Board.
b. Receive at least one full year of motor vehicle crash report data from the UNC Highway Safety Research Center (HSRC).
c. Obtain data use approval from the Communicable Disease (CD) Branch for at least one year of NC DETECT ED visit data.
d. Obtain data use approval from NC SCHS for at least one year of death certificate/vital records data.
e. Complete implementation plan for data linkage analysis.

Narrative

The first half of the project year was dominated by our efforts to gain access to the data and to plan out how to approach the problem of developing linkage methodology. The process of obtaining access to the three datasets varied.

An IRB request for the data linkage project was submitted to UNC on October 7, 2019 and approved October 29, 2019 (IRB 19-2675)

UNC HSRC provided crash data on November 1, 2019. A formal data request was not needed due to HSRC's participation in this project.

North Carolina death certificate data are considered part of public health surveillance and do not require formal approval. A signed SCHS Data Request for Identified Data for Research Purposes form F-14 was submitted to SCHS on October 17, 2019 and the data were obtained the same day. A revised request to include names (for linkage verification) was submitted and the data received on December 6, 2019.

NC DETECT ED visit data are considered to be sensitive health data. The data request was submitted via the online application on October 22, 2019 to the NC DETECT Data Oversight Committee, Communicable Disease Branch, NC DPH. The NC DETECT Data Oversight Committee approved the request on November 18, 2019. The DUA was signed and executed on December 9, 2019.

A key component of implementing the project was inviting Dr. Lawrence (Larry) Cook, the Principal Investigator of Utah CODES and a Professor in the University of Utah School of Medicine, to consult on different aspects of the project. Dr. Cook provided his insight and experience in an extended onsite visit February 10-14, 2020.

Part of Dr. Cook's visit included consultation on our draft implementation plan (Appendix H), which we wrote using the public template on the CDC's website. We also solicited feedback from the entire project team. The implementation plan was finalized on March 3, 2020.

Deliverables

Deliverable	Due date	Date completed	Location
Received at least one year of motor vehicle crash report data from the NC Highway Safety Research Center	11/13/2019	11/1/2019	Appendix E: Crash data delivery
Completed data use approvals for NC DETECT and SCHS data	12/31/2019	12/9/2019	Appendix F: Data request for NC DETECT Appendix G: Data request for death certificate data
Completed implementation plan for data linkage analysis	2/27/2019	3/3/2020	Appendix H: Implementation Plan

Activity 3: Identify and document barriers and facilitating factors to data linkage analysis

Narrative

By the middle of the project year we were able to identify barriers and facilitating factors to our project. These were submitted both to the CDC via their Partner Portal and in a separate report created for this project (Appendix I). We were also able to document in the same report the four different methods of data linkage we had considered.

This report was submitted to NCDHHS on April 29, 2020.

Deliverables

Deliverable	Due date	Date completed	Location
Submitted mid-year brief report outlining progress to date on all performance requirements	3/17/2020	3/3/2020	Appendix H: Mid-Year Report
Submitted report documenting initial data linkage methods for two health outcome data sources (NC DETECT ED visit data and SCHS death certificate data), including any barriers and facilitating factors to data linkage analysis	4/30/2020	4/29/2020	Appendix I: Data Linkage and Barriers and Facilitators report

Activity 4: Complete data linkage analysis

Tasks

a. Receive health outcome data sources on NC DETECT ED visit data and SCHS death certificate data.
b. Complete analysis and linkage of two health outcome data sources: NC DETECT ED visit data and SCHS death certificate data.
c. Summarize data linkage activities and results.

Narrative

Death certificate data was received from SCHS on October 9, 2019. An updated data file was requested and received on December 9, 2019 which included the names of the deceased. The names were requested in order to do linkage verification. The NC DETECT file was received from CCHI on December 16, 2019.

This final report was compiled from May until August 2020. We requested and received a one month extension. The majority of the data linkage work occurred in 2020. Dr. Larry Cook's visit in February 2020 was instrumental in our approach over the remaining months of the project. The following two sections describe the data linkage activities and results.

Deliverables

Deliverable	**Due date**	**Date completed**	**Location**
Received NC DETECT ED visit and SCHS death certificate data	1/30/2020	12/16/2019	Appendix K: Receipt notices of NC DETECT and SCHS death certificate data
Submitted final report summarizing data linkage activities and results in Year 1 (addresses all Performance requirements) and a summary handout of data linkage activities and results reported in Year 1	7/31/2020	8/27/2020	Appendix L: Summary handout

Process for Linking Data

Although we had a general implementation plan for linking the data, the final process flow and coding of the linkage were developed throughout Year 1. We chose to test different data linkage methodologies to find the optimum one for our needs. The methodologies we tested were:

a. Recursive partitioning trees (RPT) in R
b. Probabilistic linkage using LinkSolv
c. Cascading hierarchical deterministic linkage with block filtering in R (CHD)

We also reviewed probabilistic linkage packages in R, but did not include them in the testing.

See Appendix J for a description of these methods. Because the recursive partitioning trees and LinkSolv had complicated processes that required more computing resources and might be unclear to a multidisciplinary audience, we chose to focus on creating a CHD methodology that would have comparable results. See figure 1 below for a general flow chart of the overall process.

In practice, we added a fourth linked dataset for the crash and death certificate linkage: a hand-reviewed set. We did not include an RPT-linked dataset for comparison with the crash and ED visit linkage, though RPT principles learned through the death work were applied to ED-crash linkage as well. All coding of the RPT and CHD processes in R was done by Mike Fliss.

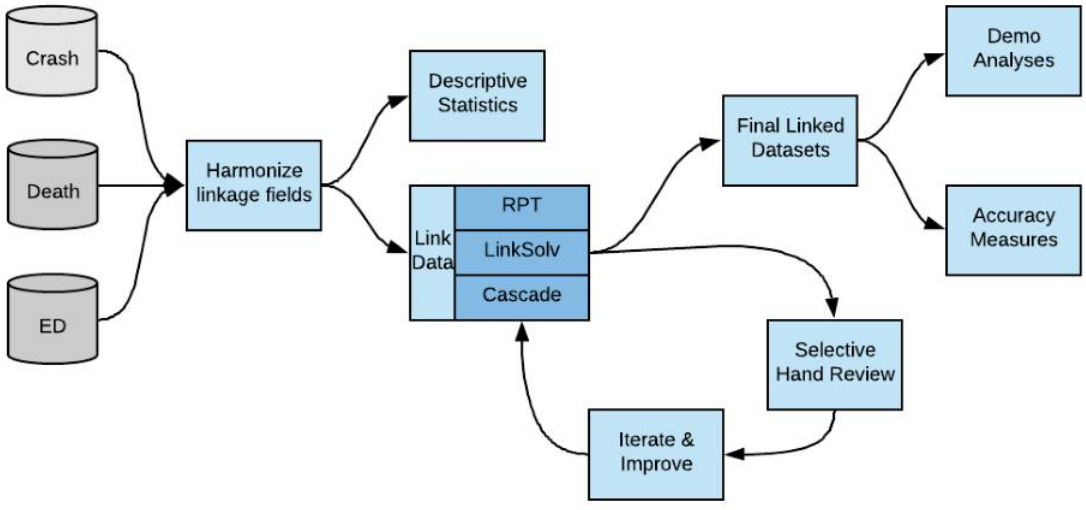


Figure 1: General linkage study process

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 steps, creating a cascading hierarchical linkage with decreasing specificity of linkage match requirements. See figure 2 for a description of the process.

The process iterates over four inputs: (1) dataset A, e.g. crash report data; (2) dataset B, e.g. health data; (3) the linkage command table (described below); and (4) a known links table, which is initially empty or 'null' on the first linkage pass, but grows as links are added with each pass. The process repeats through each of the linkage passes in the linkage command table to create the final linked dataset. See figure 1 below for a general flow chart of the process.

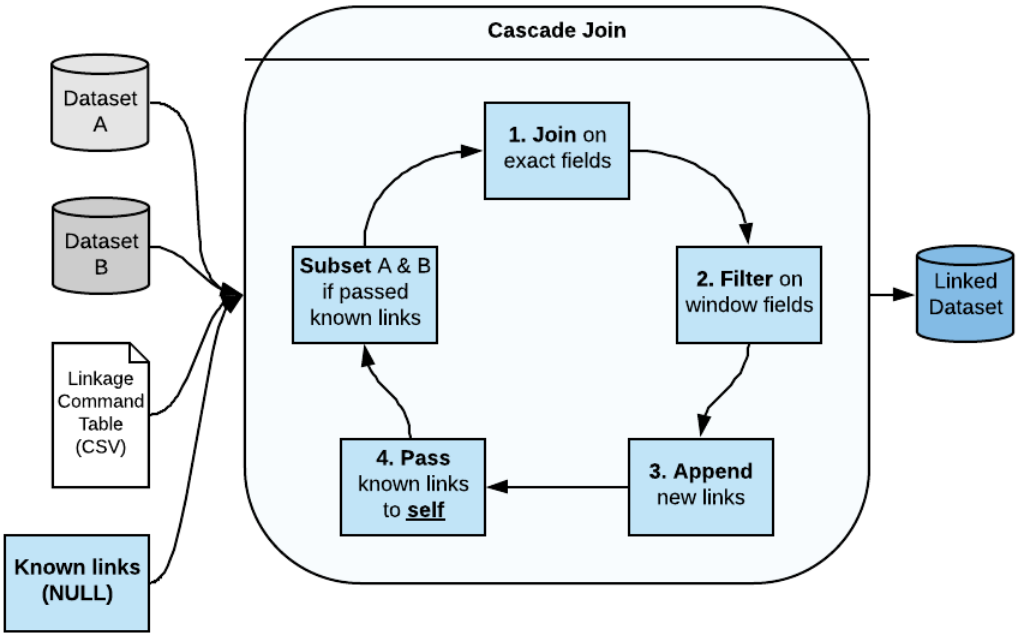


Figure 2: Cascade join function

To build in flexibility, this process introduces the novel use of a "linkage command table" in comma separated value (CSV) format that lists the linkage passes (i.e. "cascades"). The linkage command table is utilized directly by the R-coded linkage process as the instructions for how the data are to be linked on each linkage pass. The linkage command table allows for changes to be made to the linkage solely by editing the CSV file.

See figure 3 for an example of a linkage command table. Each row in the example represents a pattern of linking shared variables by different criteria or "distances". The linkage is run sequentially, starting at linkage ID row 1. The type of required match between corresponding linkage variables are represented in each cell. Exact matches have a distance of 0 and are shown in green, fuzzy/window/distance-match in yellow (e.g. number of years, number of days, etc.), and variables to drop in that pattern in red (not applicable = NA).

link_id	link_name	l_age_num	l_dob_date	l_dobmd_fct	l_gender_fct	l_raceeth_fct	l_fatal_lgl	l_isinj_lgl	l_issevere_lgl	l_iscrash_lgl	l_crashpos_fct	l_rstate_fct	l_rcounty_fct	l_rzip5_fct	l_rzip3_fct	l_rcity_fct	l_cstate_fct	l_ccounty_fct	l_acc_date	l_accmd_fct
1	Exact: completely matching	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Place-2: Drop R city, R zip	0	0	0	0	0	0	0	0	0	0	0	0	NA	NA	NA	0	0	0	0
3	Place-2: Drop R city, C county, R county	0	0	0	0	0	0	0	0	0	0	0	NA	0	0	0	0	NA	0	0
4	Demo-1: drop race	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Crash-1: crash pos	0	0	0	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0
6	Crash-2: crash & crash pos	0	0	0	0	0	0	0	0	NA	NA	0	0	0	0	0	0	0	0	0
7	Date-1: Day 30 away	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	NA
8	Place-Max: Just states, crash pos, date+1	0	0	0	0	0	0	0	0	0	NA	0	NA	NA	NA	NA	0	NA	1	NA
9	Demo-Max: Age only	1	NA	NA	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Crash drop fatal/severe, date 30 days	0	0	0	0	0	NA	0	NA	0	0	0	0	0	0	0	0	0	30	NA
11	Demo-Max: Age only, crash flex	1	NA	0	0	0	NA	NA	NA	0	NA	0	0	0	NA	0	0	NA	1	NA
12	Crash drop inj/crash	0	0	0	0	0	0	NA	0	NA	NA	0	0	0	0	0	0	0	0	0
13	Date-Max: No date, MD match, near ccount	0	0	0	0	NA	0	0	0	0	NA	NA	NA	NA	NA	NA	NA	NA	30	NA
14	Date-Max: No date, MD match, near ccount	NA	0	0	0	NA	0	0	0	0	NA	NA	NA	NA	NA	NA	NA	NA	30	NA

Figure 3: Example linkage command table

An additional process was developed in R to determine the linkage patterns necessary to get as close as possible to a comparison linked dataset, such as a hand-reviewed set or one created via other methodologies, such as LinkSolv. Attempting to match the comparison dataset exactly can result in an extensive list of passes and increases the risk of false positives. To reduce the number of passes, a combination of individual record review and data analysis determines which linkage variables are the most critical and which should be dropped. Linkage passes are typically dropped when they are not generating many true positives and/or creating too many false positives relative to the hand-reviewed datasets and/or LinkSolv. This process results in a manageable number of passes by identifying the most effective linkage pass patterns.

Some benefits from the CHD linkage process include:

- **The process is transparent.** Each match includes clear documentation as to what linkage criteria were used to link the records.
- **The links are high quality.** We are showing comparable results to other established linkage methods, such as a hand-reviewed set or a LinkSolv-generated one.
- **The linkage process is fast.** Once the data are cleaned and ready to be linked, the linkage process runs in less than a minute for the crash and death certificate linkage and less than two minutes for the crash and ED visit linkage.
- **The process is replicable for future linkage work.** The code can be used to link other years of the same data sources or revised to link other data sources.

The activities for Year 1 and for linking crash report data with death certificate and ED visit data are described in more detail below.

Data Linkage Activities

Provide overall planning and coordination for the implementation
The core team members met bi-weekly or as needed. Project meetings with all project members were held monthly. Project updates were also emailed to all the project team members between meetings.

Obtain and Prepare the Data
See Activity 2 for a description of the process to obtain the three datasets. All datasets were obtained by December 16, 2019. We prepared the data by converting all the files to comma-separated value (CSV) text files.

Unique identifiers were assigned to each record, using the person as the base unit. The unique identifier for the crash data was a combination of the crash ID, the vehicle position number, and the person number. This information was captured in the new variable "uid_crash_veh_per". The unique identifier for NC DETECT data was the visit ID. The unique identifier for death certificate data was created by combining "2018-" with a counter for each record in the original files. The unique identifiers for death certificate data and NC DETECT data were called "uid_death_yearrow" and "uid_visit_id", respectively.

Identify and Clean Linkage Data Elements
We began with a preliminary list of probable linkage elements based on the available linkage fields in the three datasets. Based on our data linkage results throughout the linkage process and with input from data linkage expert Dr. Larry Cook, the following data elements were used to link the data.

Variables & Cleaning / Recoding notes, if any

		Harmonized Variable	Dataset basis variable		
			Crash	Death	ED
	Unique ID	NA	crsh_id, vehpos, pernum	age, agetype	age
		Unique ID for each dataset. Create if necessary	uid_crash_veh_per. Note: still seem to be 5 duplicate UIDs	non-year (month, day, hour, minute) are calculated in years and rounded to integer.	
Person Demographics	**Age in years**	l_age_num	age	age, agetype	age
	age in integer years	Drop ages < 0 or greater than 120. Promote '0' to 1 year old. Function: get_clean_age_num		non-year (month, day, hour, minute) are calculated in years and rounded to integer.	
	Date of Birth	l_dob_date	perdob	dob_yr, dob_mo, dob_dy	dob
	Date.	Cannot be > today. Dropping (as missing) any Jan 1			

			or year 1900 DOBs.			
		Date of Birth	l_dobmd_fct	perdob	dob_yr, dob_mo, dob_dy	dob
		Month-Day factor	(same drops as above)			
		Gender	l_gender_fct	sex	sex	sex
		M/F factor	Male or female only			
		Race-Ethnicity	l_raceeth_fct	race	racer, hisp	racedesc., ethnicitydesc.
		Combined race and ethnicity	Harmonizing to: Nat Amer, Hispanic, Black (non-Hispanic), White (non-Hispanic), Asian, Other. Create in descending order of population (e.g. in NC start with Native American / Indiginous, end with White non-Hispanic) to maximize group size.	Race already combined	deaths$l_raceeth_fct = case_when(….etc.) Follow small numbers / IVPB guidelines.	Create c_anyhisp variable, then follow same code pattern
Person	Person-crash characteristics	**Is Fatality**	l_fatal_lgl	inj	(none)	dispositiontext
			Note some fatalities may come after the crash or ED visit. Target construct here is a death within 2 weeks of a crash.	inj is "Fatal Injury"	All deaths are fatalities.	"died" --> fatal
		Is Injury	l_isinj_lgl	(none)	ICD variables (cod*)	ICD diag/injurycode
		*Is injury *mechanism*. Not a measure of severity (e.g. no injury crashes = injury)*	Note this is more about the mechanism (injury related) than severity (was there an injury severe enough to "count").	All crashes flagged as (true), given what is an "injury" in other data sets, e.g. "no injury" -> injury..	Regex: [S][0-9][0-9]|T[0-8][0-8]|V[0-8]|V9[8-9]	Regex: ^[S|T]. May subset to only injuries given time to run. Note that cascading deterministic+filter linkage can handle not subsetting.
		Is Severe	l_issevere_lgl	inj	(none)	dispositiontext
			T/F	inj is Fatal, A, or B	All deaths are severe	Admitted|Died
		Is Crash	l_iscrash_lgl	(none)	ICD variables (cod*)	ICD diag/injurycode
			T/F	All are true (crashes)	Regex: U011|V[0-9][1-9]|X82|Y03|Y361	Regex: U011|V[0-9][1-9]|X82|Y03|Y361
		Crash Position	l_crashpos_fct	type, vehtype	ICD variables (cod*)	

			Harmonized to: Driver, Passenger, Motorcyclist, Pedestrian, Pedal cyclist, Unknown, Other, and Not Crash	Bicyclist->Pedal Cyclist, if vehtype = "MC|Moped|Scotter" -> "Motorcyclist", otherwise = type	See separate table. Note that unknown position is promoted to driver.	See separate table. Note that unknown position is promoted to driver.
Place	Resident of	**Residential Zipcode**	l_rzip5_fct	rzip	zipcode	zip
		5 digit factor	Moderate cleaning, pull first five. No longer validity checking (US ZCTAs are insufficient). Georef those we can.			
		Residential Zipcode	l_rzip3_fct	l_rzip5_fct	l_rzip5_fct	l_rzip5_fct
		3 digit factor	substring of cleaned, valid 5 digit zips			
		Residential Zipcode	l_rzip5_x, y	l_rzip5_fct	l_rzip5_fct	l_rzip5_fct
		lat & long	33,000 US ZCTAs are used, centering projection on NC (max accuracy), subsetting to NC zips			
		Residential County	l_rcounty_fct	county	countyc	county
		string name	Counties matched against 466 neighbor state counties (whitespace, punctuation, captailized). Typos are not fuzzy matched.	promote crash county (county) to a stand-in for the unavailable residential county.		
		Residential County	l_rcounty_x, y	l_rcounty_fct	l_rcounty_fct	l_rcounty_fct
		lat & long	Pulled from census county shapefile			
		Residential City	l_rcity_fct	rcity	state_fip+cityc+cityrestext	edfacilityid
		string name	Cleaned & matched to valid NC city		If city code is not in fips table, use city res text	**Facility Table C**: zip and county of facility (Table A/B)
		Residential City	l_rcity_x, y	l_rcity_fct	l_rcity_fct	l_rcity_fct
		lat & long	Unmatched cities are moderately cleaned (whitespace, punctuation, captailized) and passed on.	**City Table B:** census city file. Lat-long centroids joined in as x, y.		**Facility Table C**: zip and county of facility (Table A/B)
		Resident State	l_rstate_fct	rzip	state_fip	state

	string name	Code neighbor states, everything else is other.	If zip is in NC zips -> "NC"; if not missing, "out of state"; otherwise missing.		Already clean, simple rename
Place / Crash	**Crash County**	l_ccounty_fct	county	cod	edfacilityid
	string name	Moderate cleaning, remove non alpha-numerics, cleaned against national county name file			assume facility location is county of crash. Not perfect but better than nothing (works for spatial prox)
	Crash County	l_ccounty_fct	l_ccounty_fct	l_ccounty_fct	l_ccounty_fct
	lat & long				**Table B**: (1) zip and county of facility + (2) census shape files
	Crash State	l_cstate_fct	(none)	dstate	(none)
	Crash state as string	Largely uninformative, but there's nuance. Injury county could be out of state, for example, and ED/death in NC.	100% of dataset for study are NC crashes	100% of dataset for study are NC deaths. In rare cases with this dataset we could get crashes elsewhere.	facility IDs all in NC.
Time	**Crash Date**	l_acc_date	accdate	dod_yr, dod_mo, dod_dy	arrivaldate
	Day date	crash date			
	Crash Date	l_accmd_fct	accdate	dod_mo, dod_dy	arrivaldate
	Month-Day factor	month-day string			

Helper Tables (not included)

Table A: Census city data.

Table B: Census county shapefile

Table C. Facility table used as proxy for ED crash location

Crash and Death Certificate Linkage Activities

Systematically Test Linkage Methodologies

We started with the death certificate and crash linkage, because of the lower number of linked records and the simpler expected result of a 1:1 linkage of crashes and death certificates. It was anticipated that the linkage would yield between 1300 and 1600 linked records, based on the 1,442 reported motor vehicle crash fatalities in the North Carolina 2018 Crash Facts published by the North Carolina Division of Motor Vehicles and the 1,468 records listed as fatalities in the 2018 crash data file.

We began with the plan to test three different data linkage methodologies:

d. Recursive partitioning trees (RPT) in R
e. Probabilistic linkage using LinkSolv

f. Cascading hierarchical deterministic linkage with block filtering in R (CHD)

We identified the records we anticipated linking in each dataset, illustrated below. We expected to link the majority of fatalities in the crash data with the majority of persons with motor vehicle crash (MVC)-related causes of death (box 1 in figure 4 below). We also anticipated that we would link some number of non-fatally injured crash victims and persons with non-MVC-related causes of death (boxes 2, 3, and 4).

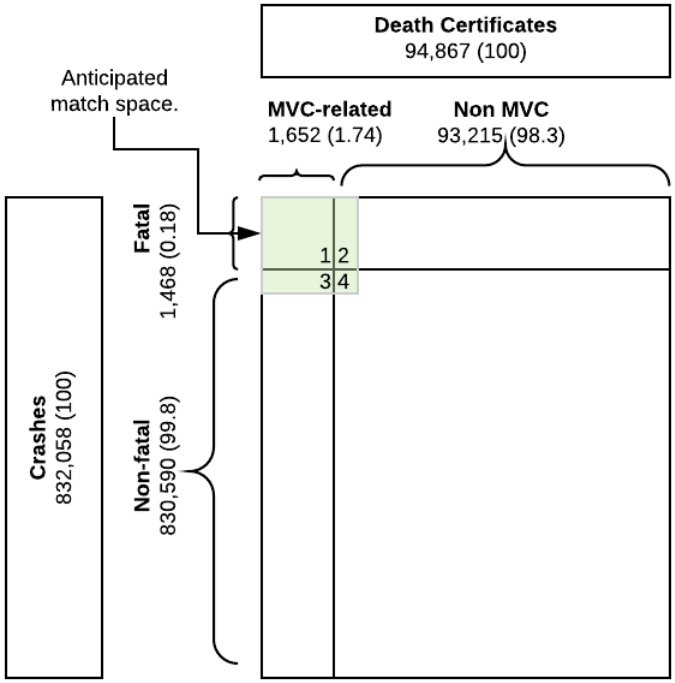


Figure 4: Anticipated match space, crash-to-death certificate linkage. Not to scale.

Based on its use in other linkage projects and its endorsement by the CDC and Dr. Larry Cook, we initially considered the LinkSolv results to be a good source for comparison with other linkage methods.

Verify linkage

We linked crash and death certificate data using all three methods and systematically reviewed linkages found by each method. By comparing the linkage found via each method, we were able to identify the common linkages between methods. See figure 5 for a Venn diagram of the comparison.

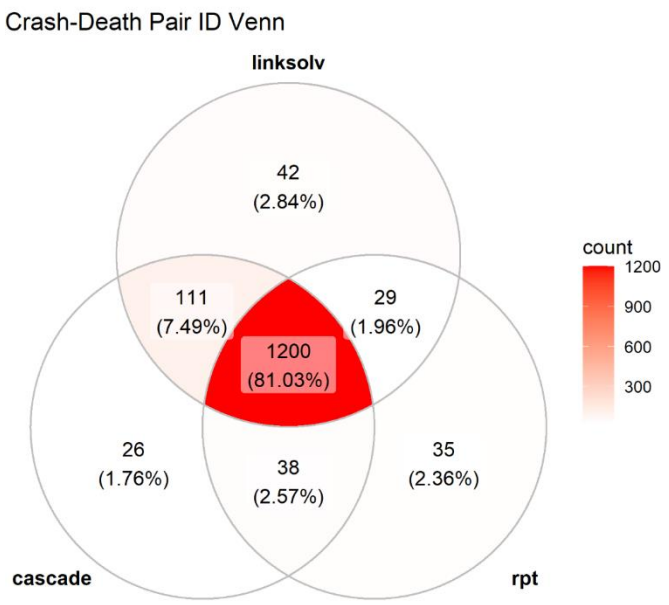


Figure 5: Venn diagram of linkage matches via three methodologies. Not to scale.

Each record was reviewed individually and assigned a rating of "likely match", "unlikely match", or "undetermined". The availability of the deceased's name and residential street address in the death certificate data was helpful in classifying matches and allowed us to further verify matches by viewing full crash reports for fatal events investigated by the NC State Highway Patrol (https://www2.ncdps.gov/Index2.cfm?a=000003,000014,002745).

By reviewing each match and matching up unlinked records when possible, we were able to create a fourth comparison linked dataset, a "hand-reviewed" set of 1,490 linked records. This was reduced to 1,483 after applying the following criteria.

	Count
Identified through individual record review	1,490
Not reported as fatal by law enforcement officers and the death certificate did not list a transportation-related cause of death	-5
Records likely matched but the time of death was from day prior to crash	-2
Final hand-reviewed dataset	**1,483**

This hand-reviewed dataset of 1,483 records was then used as the "gold standard" comparison dataset. We were then able to compare the three methodologies to this gold standard set. See figure 6 for this results of that comparison (not to scale).

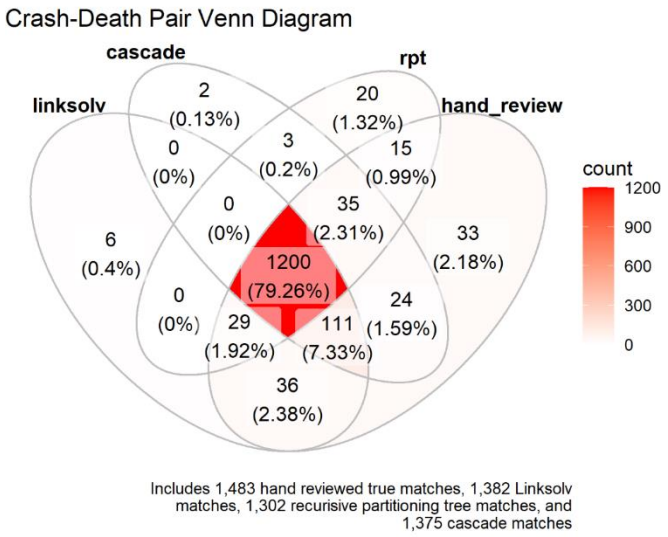


Figure 6: Venn diagram comparing three methodologies to the hand-reviewed dataset. Not to scale.

We also reviewed all unlinked fatal crash records and unlinked death certificate records which indicated the presence of a motor vehicle crash-related cause of death. See figure 7 for the results of that review.

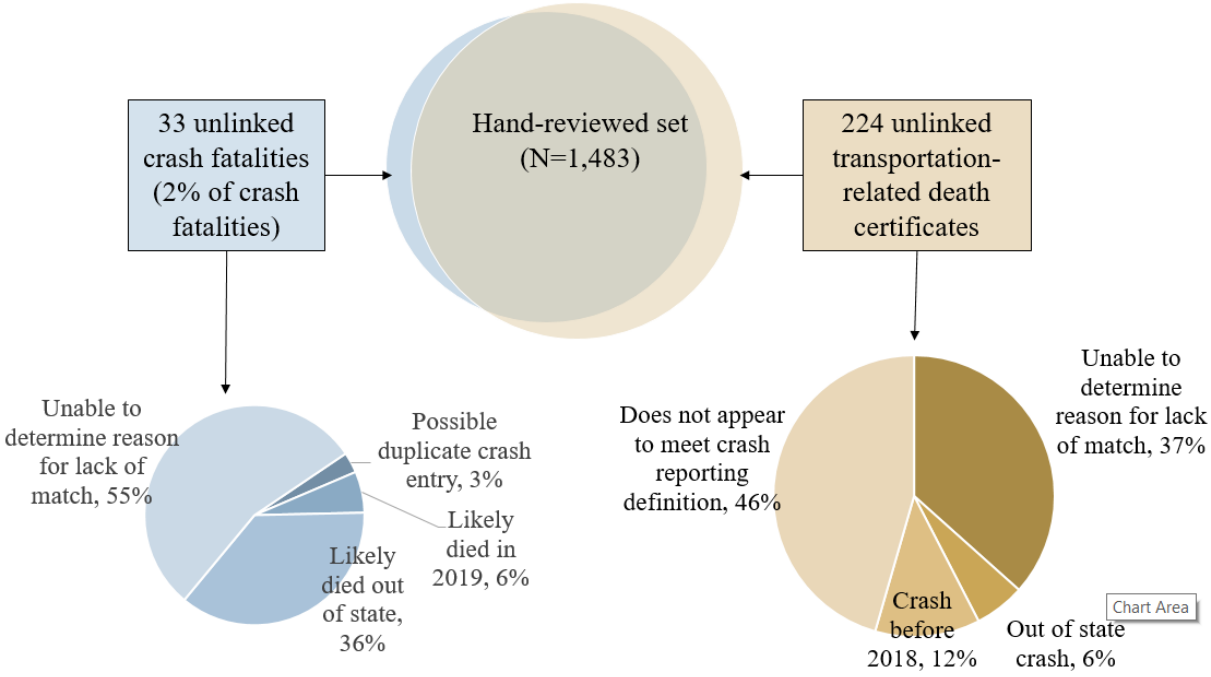


Figure 7: Review of unlinked crash and death certificate records

Determine the Final Methodology

Using the 1,483 hand-reviewed dataset as a goal, we tested different deterministic linkage passes that would have comparable results without a high number of passes. By trying many different combinations of joins and comparing the results to the hand-reviewed dataset, we were able to identify the linkage passes that would come close to the desired results with the least number of passes. See figure 8 for the final sequence of fourteen linkage passes.

Crash - Death Cascading Linkage Pattern

Linkage step, starting with step 1 and going down ... ->

	Person (Demographics)	Person (Crash)	Place (Residence)	Place (Crash)	Time
1	Exact (5 variables)	Exact (5 vars)	Exact (5 vars)	Exact (2 vars)	Exact (2 vars)
2	Exact	Exact	Same state, county	Exact	Exact
3	Exact	Exact	Drop county	Drop county	Exact
4	Drop race-eth.	Exact	Exact	Exact	Exact
5	Exact	Drop crash pos.	Exact	Exact	Exact
6	Exact	Drop is crash, pos.	Exact	Exact	Exact
7	Exact	Exact	Exact	Exact	w/in 30 days
8	Exact	Drop crash pos.	Drop all but state	Drop county	w/in 1 day
9	Drop DOB/MD, race-eth, age+/-1	Exact	Exact	Exact	Exact
10	Exact	Drop fatal, severe	Exact	Exact	w/in 30 days
11	Drop DOB, age+/-1, keep MD, gender,RE	Drop all but crash	Exact	Drop county	w/in 1 day
12	Exact	Drop inj, crash, pos	Exact	Exact	Exact
13	Drop race-eth.	Drop crash pos.	Drop entirely	Drop entirely	w/in 30 days
14	Drop age, race-eth.	Drop crash pos.	Drop entirely	Drop entirely	w/in 30 days

Figure 8: Crash/Death certificate linkage pattern

There were no exclusionary criteria applied prior to data linkage; 100% of both crash records and death certificate records were included in the CHD linkage. There are also currently no post-linkage restrictions or processing, although we may consider them in future linkage work.

Because we had our hand-reviewed dataset, we were able to calculate the sensitivity and specificity of the CHD-generated match.

	n		n
True Positive (a)	1,370	False Positive (c)	5
False Negative (b)	113	True Negative (d)	830,688

The results were scored according to the following formulas:

Measure	Formula	Results
Sensitivity	$\frac{a}{a+b}$	92.38%
Specificity	$\frac{d}{c+d}$	>99.99%
Positive Predictive Value (PPV)	$\frac{a}{a+c}$	99.64%
Negative Predictive Value (NPV)	$\frac{d}{c+d}$	>99.99%
Accuracy	$\frac{(a+d)}{(a+b+c+d)}$	99.99%
Cohen's Kappa[1,2]	$\frac{(p_o-p_e)}{(1-p_e)}$	1.00

[1] p_o =Observed agreement (identical to accuracy).

[2] p_e = Probability of chance agreement.

Crash and Death Certificate Linkage Results

The hand-reviewed linked dataset represents 1,483 persons from 1,370 crashes. The CHD-generated linked dataset represents 1,375 persons from 1,274 crashes. Both datasets are saved and available for analysis on the UNC OneDrive.

Note: Linkage totals reflect the linked crash/death certificate visit datasets (Versions: 20200706_crashdeath_handreviewed.xlsx; 2020-08-16 cascade_linked_deaths.csv) at the time of this report. Future analysis using later iterations of the linked datasets may produce different totals due to additional data validation and linkage evaluation.

Linkage Metrics of Hand-Reviewed and CHD-Generated Linked Data

Frequency of persons in crash report data linking with death certificate (health) data, by law enforcement-reported injury severity

Crash injury severity	Hand-reviewed dataset				CHD-generated dataset			
	Linked to health data		Did not link to health data		Linked to health data		Did not link to health data	
	N	%	N	%	N	%	N	%
K-Fatal injury	1,433	96.6	35	0.0	1,355	98.5	113	0.0
A-Serious injury	15	1.0	4,685	0.6	2	0.1	4,698	0.6
B-Minor injury	14	0.9	29,325	3.5	8	0.6	29,331	3.5
C-Possible injury	10	0.7	95,725	11.5	3	0.2	95,732	11.5
O-No injury	7	0.5	689,416	83.0	6	0.4	689,417	83.0
Unknown	4	0.7	11,389	11.5	1	0.1	11,392	1.4
Total	**1,483**	**100**	**830,683**	**100**	**1,375**	**100**	**830,683**	**100**

Frequency of persons in crash report data linking with death certificate (health) data, by road user type

Road User / Position*	Hand-reviewed dataset				CHD-generated dataset			
	Linked to health data		Did not link to health data		Linked to health data		Did not link to health data	
	N	%	N	%	N	%	N	%
Motor vehicle driver	782	52.7	592,064	71.3	731	53.2	592,115	71.3
Motor vehicle passenger	247	16.7	228,550	27.5	219	15.9	228,578	27.5
Motorcyclist	193	13.0	4,906	0.6	189	13.7	4,910	0.6
Pedal cyclist	19	1.3	919	0.1	14	1.0	924	0.1
Pedestrian	237	16.0	3,144	0.4	217	15.8	3,164	0.4
Other	5	0.3	595	0.1	5	0.4	392	0.0
Unknown	0	0.0	397	0.0	0	0.0	600	0.1
Total	**1,483**	**100**	**830,575**	**100**	**1,375**	**100**	**830,683**	**100**

*Road user type is determined by the vehicle type and person type variables in the crash report data.

Frequency of persons in death certificate data linking with crash report data, by cause of death*

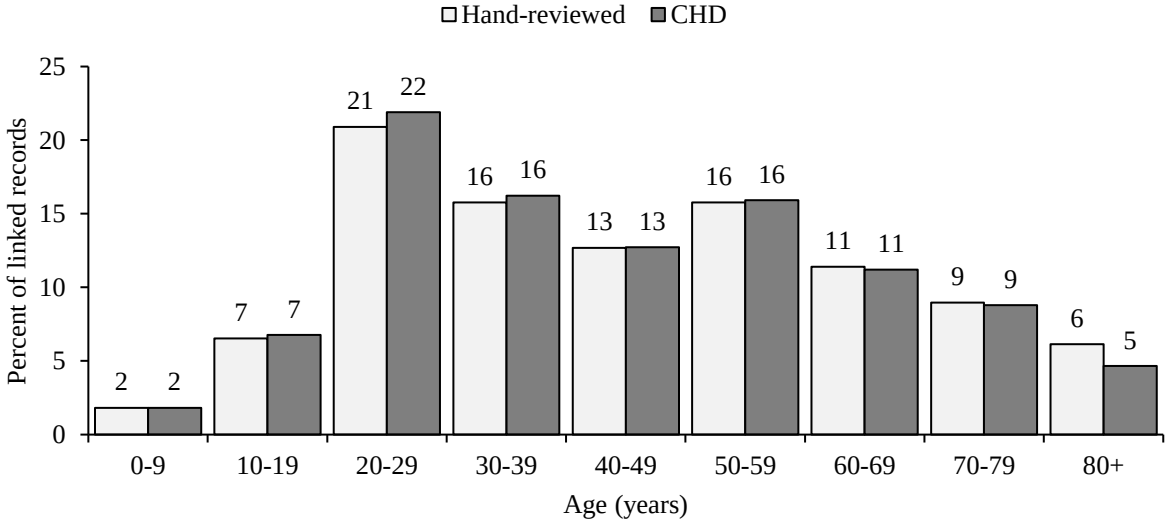
	Hand-reviewed linked dataset				CHD-generated linked dataset			
	Linked to crash data		Did not link to crash data		Linked to crash data		Did not link to crash data	
Causes of death	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
No injury or transportation-related cause of death	9	0.6	85,501	91.6	5	0.4	85,505	91.5
Injury cause of death without a transportation-related cause of death	23	1.6	7,657	8.2	9	0.7	7,671	8.2
Any transportation-related cause of death	1,451	97.8	226	0.2	1,361	99.0	316	0.3
Total	**1,483**	**100**	**93,384**	**100**	**1,375**	**100**	**93,492**	**100**

Injury causes of death include any of the following cause of death ICD-10 codes: S, T[0-8][0-8], V0*-V8*, V988 or V989.
Transportation-related causes of death include any of the following cause of death ICD-10 codes: U011, V[0-9][1-9], X82, Y03, or Y361.

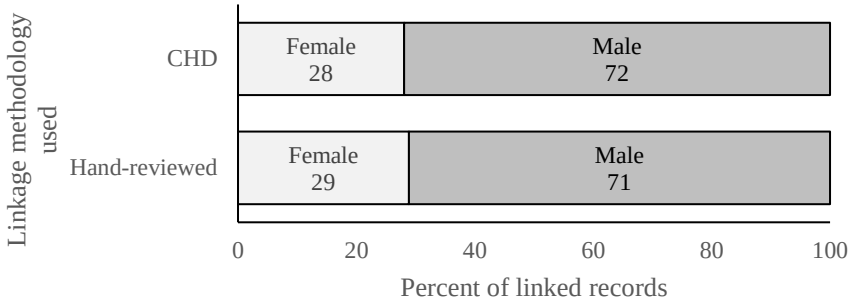
Demographics of Linked Data

Health data was used for all demographic data except when it was unknown. All percentages have been rounded to the nearest integer value, so percentage totals may not sum to 100%.

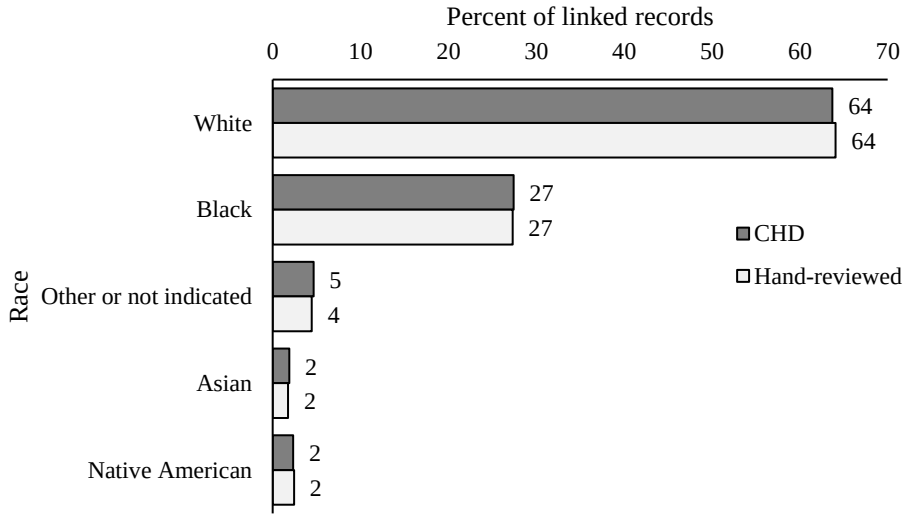
Percentage of persons in linked datasets by age (n=1,483 hand-reviewed set, n=1,375 CHD set)



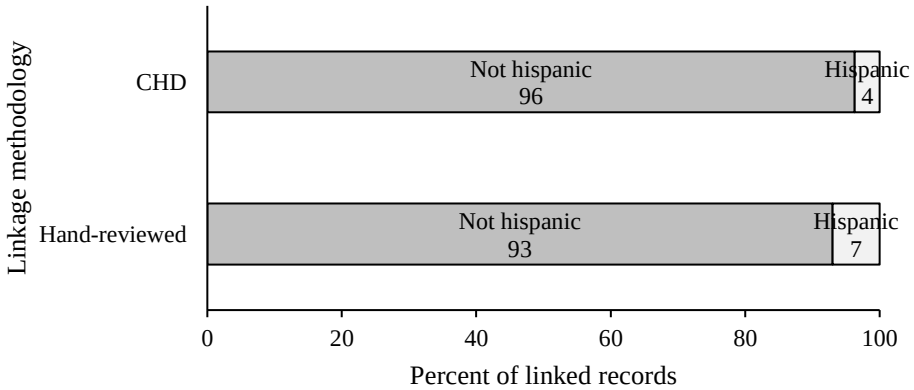
Percentage of persons in linked datasets, by sex (n=1,483 hand-reviewed set, n=1,375 CHD set)



Percentage of persons in linked datasets, by race (n=1,483 hand-reviewed set, n=1,375 CHD set)



Percentage of persons in linked datasets, by ethnicity (n=1,483 hand-reviewed set, n=1,375 CHD set)



Crash and Emergency Department Linkage Activities

Systematically Test Linkage Methodologies

We tested two different data linkage methodologies:

a. Probabilistic linkage using LinkSolv
b. Cascading hierarchical deterministic linkage with block filtering in R (CHD)

We linked crash report and ED visit data using both methods to compare the results. We chose to focus on creating a CHD methodology that would have comparable results with LinkSolv.

The final expected number of linked records was unknown, because persons at any level of law enforcement-reported injury could potentially go to the ED after a crash. Through past experience with ED visit data, the MVC-related reasons for the visit are not always well-documented. We anticipated a higher linkage rate for more serious injuries.

The records we anticipated linking in each dataset are illustrated in figure 9 below. Crash victims with serious injuries linking with MVC-related ED visits (box 1) were our expected group of core records. However, we also anticipated that we would link some number of crash victims with non-severe injuries and persons with non-MVC-related diagnosis codes (boxes 2, 3, and 4).

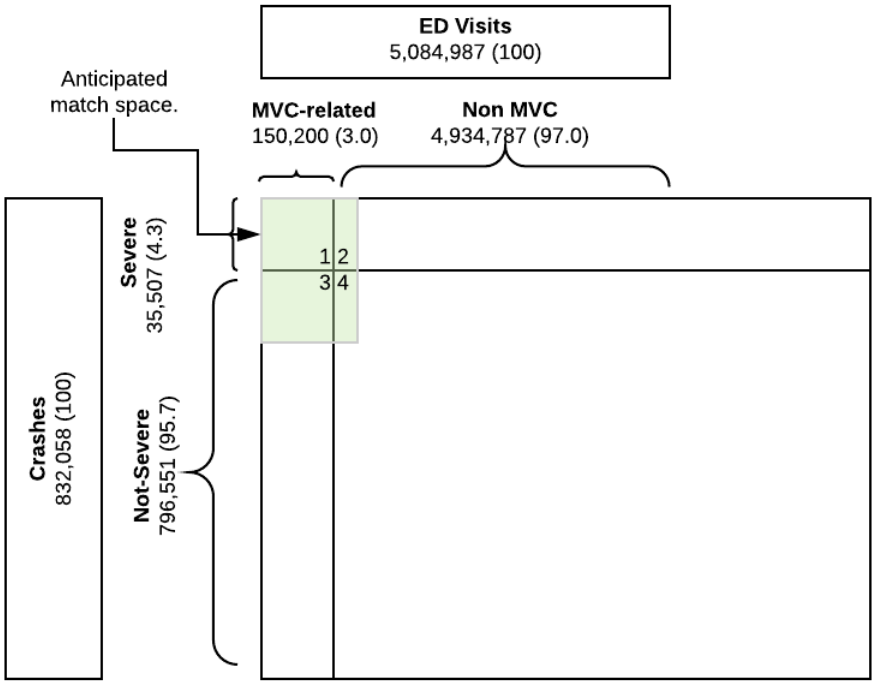


Figure 9: Anticipated match space, crash to ED visit linkage. Not to scale.

Verify linkage

We linked crash and ED visit data using LinkSolv and CHD and selected 1% of matches found in both methodologies and 1% found only in one methodology for individual review. Each of the

resulting 1,083 matches was reviewed individually and assigned a rating of “likely match”, “unlikely match”, or “undetermined”.

Because ED visit data have fewer identifiers than the death certificate data, the individual review is more challenging. Determination was made based on the overall level of matching on linkage variables and the congruency of the chief complaint and disposition diagnosis entries with the crash information. A few were checked using the record-level review available on the security-restricted NC DETECT online site but this was a time-consuming process and often added no relevant information.

The review found that 82.6% were likely matches, 10.7% were unlikely matches and 6.6% were undetermined. These results were used to inform the ongoing trials of linkage passes.

The first linkage comparison restricted the timeframe from the date of the crash to the date of the ED visit to three days. Because LinkSolv had also found more than 20,000 matches beyond three days, we were interested in understanding the nature of these matches. A selection of 100 of those matches were reviewed individually and assigned a rating of “likely match”, “unlikely match”, or “undetermined”.

The review of these 100 matches found that only 14 were likely matches, 51 were unlikely matches and 35 could not be determined. Many of the unlikely matches were found to be ED visits without any transportation-related diagnosis code and with a clearly non-crash-related chief complaint. This review prompted a reconfiguring of the LinkSolv linkage algorithm to require a transportation-related diagnosis code for ED visits past the initial 0-2 days. The results of the new LinkSolv match were again used to compare with the CHD methodology results.

Determine the Final Methodology

The final methodology is a cascading hierarchical deterministic linkage with the following linkage passes.

Crash - ED Cascading Linkage Pattern

Linkage step, starting with step 1 and going down ... ->

	Person (Demographics)	Person (Crash)	Place (Residence)	Place (Crash)	Time
1	Exact (5 variables)	Exact (5 vars)	Exact (7 vars)	Exact (2 vars)	Exact (1 var)
2	Drop race-eth.	Drop inj	Exact	Exact	Exact
3	Drop age	Exact	Exact	Exact	Exact
4	Exact	Exact	Drop R county	Exact	Exact
5	Exact	Exact	Drop city	Exact	Exact
6	Exact	Exact	Flex zip5 to 50mi	Exact	Exact
7	Exact	Exact	Drop county & city	Exact	Exact
8	Exact	Exact	Drop county & city, flex zip5 to 50mi	Exact	Exact
9	Exact	Exact	Drop zip	Exact	Exact
10	Exact	Exact	Drop county	Drop county	Exact
11	Drop race-eth.	Exact	Drop zip5	Exact	Exact
12	Drop age & race-eth.	Exact	Drop zip5	Exact	Exact
13	Drop race-eth.	Drop crash pos.	Exact	Exact	Exact
14	Drop race-eth.	Drop is_crash, pos.	Exact	Exact	Exact
15	Exact	Exact	Exact	Exact	w/in 1 day
16	Exact	Exact	Exact	Exact	w/in 2 days
17	Exact	Exact	Exact	Exact	w/in 7 days
18	Exact	Drop crash pos.	Flex zip5 to 50mi	Exact	Exact
19	Drop race-eth.	Drop crash pos.	Drop city, zip5	Exact	Exact
20	Drop race-eth.	Drop is_crash, pos.	Flex zip5 to 50mi	Exact	w/in 1 day
21	Exact	Drop crash pos.	Drop city, flex zip5 to 50mi	Drop county	Exact
22	Exact	Drop is_crash, pos.	Drop city, flex zip5 to 50mi	Drop county	Exact
23	Drop race-eth.	Drop is_fatal, is_severe	Exact	Exact	Exact
24	Exact	Drop is_crash, pos., is_severe	Drop city	Drop county	w/in 1 day
25	Drop race-eth.	Drop crash pos.	Flex zip5 to 50mi	Drop county	w/in 1 day
26	Exact	Drop is_inj, pos.	Exact	Exact	Exact
27	Drop race-eth.	Drop is_inj, is_crash	Drop city	Drop county	w/in 1 day
28	Drop race-eth.	Drop is_inj, pos.	Drop city, flex zip5 to 50mi	Exact	Exact
29	Drop DOB, race-eth.	Drop all but crash	Exact	Drop county	w/in 1 day
30	Drop DOB, race-eth.	Drop all but crash	Flex zip5 to 50mi	Drop county	w/in 2 days
31	Drop DOB, race-eth.	Drop all but crash	Flex zip5 to 50mi	Drop county	w/in 7 days

Figure 10: Crash/ED visit linkage pattern

No exclusionary criteria was applied to the crash data prior to data linkage; 100% of crash records were included in the match. ED data was first filtered to exclude any records with a disposition code of 20, representing a "Transferred/discharged to another short-term general hospital." This filter removed 1% of all ED visits (N= 58,547). There are currently no post-linkage restrictions or processing, although we may consider them in the future.

Crash and Emergency Department Linkage Results

The CHD-generated linked dataset represents 83,998 persons from 63,681 crashes. The dataset is saved and available for analysis on the UNC OneDrive.

Note: Linkage totals reflect the linked crash/ED visit datasets (Versions: 20200828_crashed_LinkSolv.xlsx; 2020-08-20 cascade_linked_ed.csv) at the time of this report. Future analysis using later iterations of the linked datasets may produce different totals due to additional data validation and linkage evaluation.

Linkage Metrics of LinkSolv and CHD-Generated Linked Data

Frequency of persons in crash report data linking with ED visit (health) data, by law enforcement-reported injury severity

Crash injury severity	LinkSolv-generated dataset				CHD-generated dataset			
	Linked to health data		Did not link to health data		Linked to health data		Did not link to health data	
	N	%	N	%	N	%	N	%
K-Fatal injury	268	0.3	1,200	0.2	158	0.2	1316	0.2
A-Serious injury	2,928	3.0	1,772	0.2	2,329	2.8	2,414	0.3
B-Minor injury	18,221	18.4	11,118	1.5	14,141	16.8	15,383	2.1
C-Possible injury	41,458	41.8	54,277	7.4	36,032	42.9	59,911	8.0
O-No injury	34,860	35.2	654,563	89.3	30,163	35.9	659,385	88.1
Unknown	1,372	1.4	10,021	7.4	1,175	1.4	10,220	1.4
Total	**99,107**	**100**	**732,951**	**100**	**83,998**	**100**	**748,629**	**100**

Frequency of persons in crash report data linking with ED visit (health) data, by road user type

Road User / Position	LinkSolv-generated dataset				CHD-generated dataset			
	Linked to health data		Did not link to health data		Linked to health data		Did not link to health data	
	N	%	N	%	N	%	N	%
Driver	70,263	70.9	522,927	71.3	61,570	73.3	531,620	71.0
Passenger	23,934	24.1	205,029	28.0	18,499	22.0	210,464	28.1
Motorcyclist	2,636	2.7	2,500	0.3	2,121	2.5	3,015	0.4
Pedal cyclist	419	0.4	525	0.1	353	0.4	591	0.1
Pedestrian	1,776	1.8	1,619	0.2	1,392	1.7	2,003	0.3
Other	49	163.3	349	0.0	43	215.0	355	61.1
Unknown	30	0.0	571	0.1	20	0.0	581	0.1
Total	**99,107**	**100**	**733,520**	**100**	**83,998**	**100**	**748,629**	**100**

*Road user type is determined by the vehicle type and person type variables in the crash report data.

Frequency of ED visits linking with crash report data, by presence of diagnosis codes*

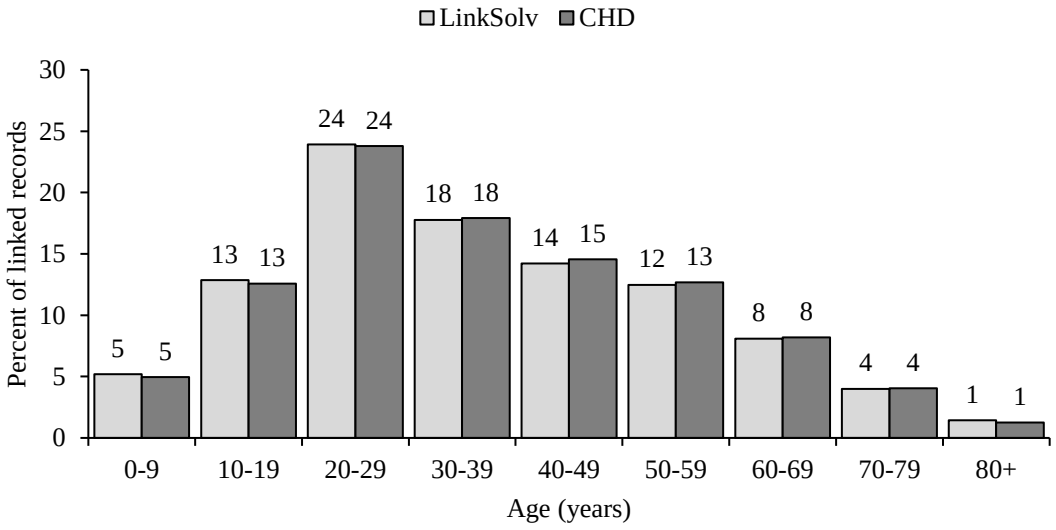
	LinkSolv-generated dataset				CHD-generated dataset			
	Linked to crash data		Did not link to crash data		Linked to crash data		Did not link to crash data	
Diagnosis code	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
No injury or transportation-related code	0	0.0	3,949,107	81.9	51	0.1	3,949,056	81.6
Injury code without a transportation-related code	10,331	10.4	814,135	16.9	7,133	8.5	817,333	16.9
Any transportation-related code	88,776	89.6	58,896	1.2	76,814	91.4	70,858	1.5
Total	**99,107**	**100**	**4,822,138**	**100**	**83,998**	**100**	**4,837,247**	**100**

*Injury diagnosis codes include any ICD-10-CM codes starting with S or T.
Transportation-related diagnosis codes include any of the following ICD-10-CM codes: U011, V[0-9][1-9], X82, Y03, or Y361.

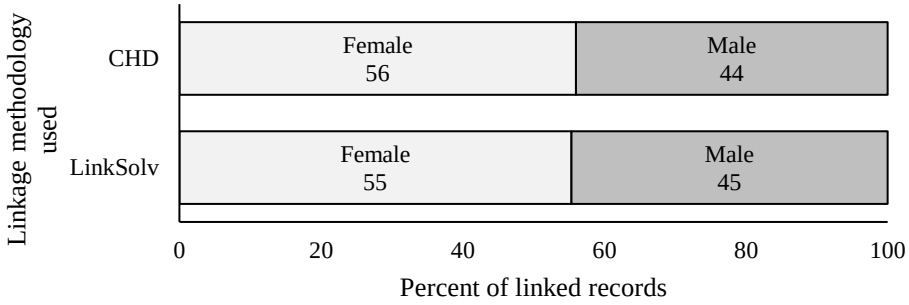
Demographics of Linked Data

Health data was used for all demographic data except when it was unknown. All percentages have been rounded to the nearest integer value, so percentage totals may not sum to 100%.

Percentage of persons in linked datasets by age (n=99,107 LinkSolv dataset, n=83,998 CHD dataset)

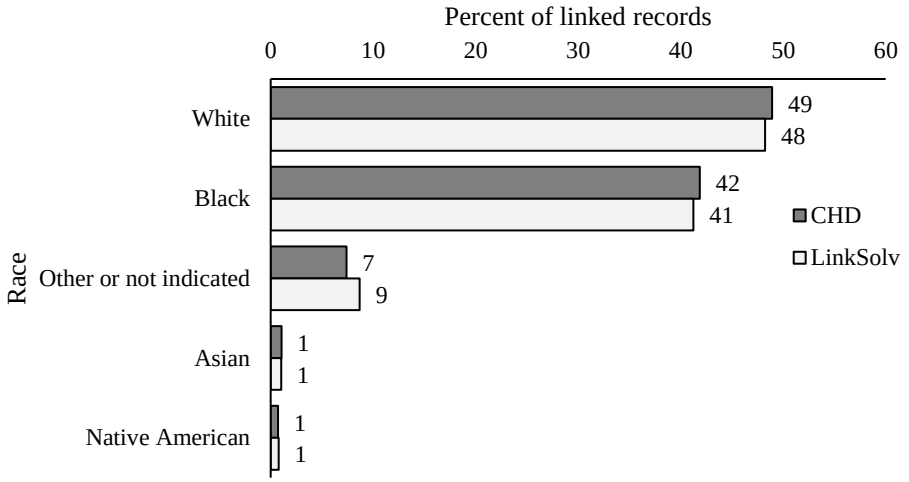


Percentage of persons in linked datasets, by sex (n=99,107 LinkSolv dataset, n=83,997* CHD dataset)

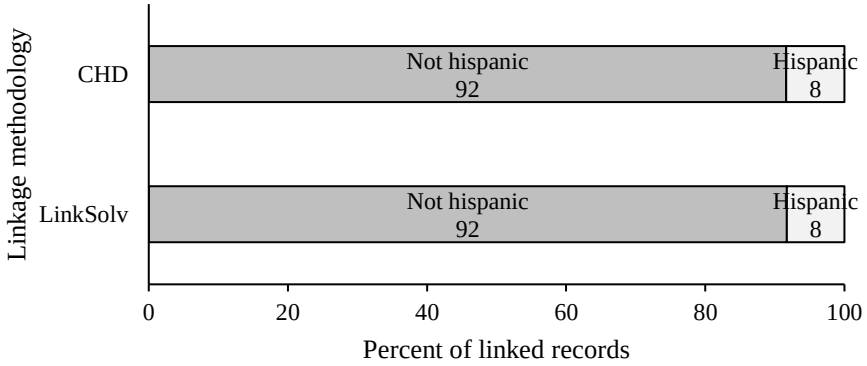


*One linked record lacked a sex designation in both data sources.

Percentage of persons in linked datasets, by race (n=99,107 LinkSolv dataset, n=83,998 CHD dataset)



Percentage of persons in linked datasets, by ethnicity (n=99,107 LinkSolv dataset, n=83,998 CHD dataset)



Deviations from the Implementation Plan

We deviated from our implementation plan in the following ways:

1. **The creation of a hand-reviewed linked dataset for crash report and death certificate data**
   We did not anticipate the creation of a hand-reviewed linked dataset for crash and death certificate linkage which went beyond sampling for sensitivity and specificity. However, this curated linked dataset will be extremely useful in year 2 of the project for exploring important research questions involving fatal motor vehicle crashes. It is unlikely that we will produce a similar dataset in the future due to the amount of resources required and the difficulty in replicating methodology.

2. **The crash/ED visit linkage inclusion of initial visits only, after any transfers**
   The complexity of the crash/ED visit linkage and limited time necessitated restricting the results of the linkage to initial visits only. The original plan was to include subsequent ED visits.

3. **The one-month delay of the final report**
   The final report was delayed due to multiple factors: competing projects and initiatives, the disruption and demands of the COVID-19 global pandemic, the increased need for resources towards the end of the project in terms of time, and the challenges of the complexity of the crash/ED visit linkage.

4. **All work, including meetings, became virtual**
   We had planned on using the office space at 100 Market Street in Chapel Hill as a meeting space, but all meetings became virtual due to the 2020 COVID-19 pandemic.

Facilitating factors, barriers and challenges

Facilitating factors

1. **A coalition of program leadership committed to data linkage**
   The support of program leadership from the following organizations has been critical to moving this project forward:
   - the NC DHHS Injury and Violence Prevention Branch (DHHS IVPB),
   - the UNC Carolina Center for Health Informatics (CCHI),
   - the UNC Highway Safety Research Center (HSRC),
   - the UNC Injury Prevention Research Center (IPRC),
   - the North Carolina Traffic Records Coordinating Committee (TRCC),
   - the NC DHHS Communicable Disease Branch, and
   - the State Center for Health Statistics (SCHS).

2. **Experience with and funding for previous data linkage projects**

Prior data linkage projects provided critical experience in data linkage methodology. These projects also established relationships with the data owners, as well as providing experience working with each dataset. Most of the prior projects were part of the Linking Motor Vehicle Crash Data to Health Outcome Data in North Carolina project, funded from 2016 to 2020 by the NC Governor's Highway Safety Program (GHSP).

3. **Staffing with personnel who have the appropriate skills**
   The core project team included individuals with high levels of expertise in project management, epidemiology, statistics, knowledge of transportation safety, the data sources being linked, and presentation and marketing skills.

4. **The input of data linkage practitioners**
   This project included funding for a one-week consultancy with data linkage expert Dr. Larry Cook from the University of Utah School of Medicine. Dr. Cook spent a week in Chapel Hill working closely with the project team and also meeting with others, including program leaders and staff of related data linkage projects. Dr. Cook provided insight into his use of probabilistic linkage methodology, feedback on our progress, and direct instruction on the use of LinkSolv.

Project Challenges and our Response

1. **Lack of common unique identifiers across data sources**
   The lack of common unique personal identifiers on the crash report and either of the health data sources necessitated a more complex data linkage methodology. It also made verification of the accuracy of the linkage results more challenging.

   Our approach to this issue was to require combinations of linkage variables to increase the likelihood of true matches. We also used other linkage variables in the datasets to do verification. This was particularly effective with the crash and death certificate linkage, because death certificate data includes the person's name and street address.

2. **Data quality and completeness**
   Missing data and data entry errors impacted our ability to link the datasets. These issues were present for both crash report and health data.

   We took these issues into account when we created our linkage passes. Fields with known issues with missingness and data entry errors were allowed to not match when a sufficient number of other variables matched.

3. **Lack of program ownership and stable funding**
   The lack of an identified long-term program owner with stable funding made planning more challenging. Planning for sustainability is made more difficult when the resources that will be available to implement the project into the future are unknown or unpredictable.

We discussed possible sources for long term funding at a meeting that included our guest Dr. Larry Cook in February 2020. This remains an open issue and we will continue to pursue stable, long term funding.

Recommendations

1. **Data Acquisition**
    a. Ensure that potential data linkage datasets have adequate candidate variables for linkage.

2. **Data Preparation**
    a. Prepare data to allow for easily adding additional years. For example, the unique IDs for each cleaned dataset should include the year. Currently, only the death certificate data unique ID includes the year.
    b. Include motor vehicle crash (MVC)-related keywords in the open text fields in the health data to meet the "is crash" linkage variable, rather than relying only on diagnosis codes. For death certificates, this includes the variable labeled "Describe how injury occurred" ("Injury_how" in the death certificate data) and the literal cause of death variables "COD1a"-"COD1d". For ED visit data, this includes the chief complaint and discharge diagnosis variables ("ChiefComplaint" and "DischargeDiagnosisDesc" in the ED visit data).
    c. Add a variable denoting whether a pain-related diagnosis code is present in the ED visit data; some ED visits for post-crash pain may not include a transportation-related diagnosis code.
    d. Change the default crash position in the health data to "passenger" instead of "driver" for child motor vehicle occupants.
    e. Evaluate whether the death certificate field labeled "If transportation injury, specify: Driver/Operator, Passenger, Pedestrian, Other" ("AccidentLiteral" in the data) could be used to determine the crash position in conjunction to the assignment based on cause of death codes.
    f. Create linkage flags for pedestrians, bicyclists and motorcyclists, so they can be assigned more lenient linkage requirements on linkage passes. These road users are less common, so inexact linkages are likely still good matches. Also, since pedestrians and bicyclists may not be carrying a driver's license or other form of identification at the time of crash, they are more likely to have missing or incorrect personal information, making them more difficult to link.
    g. Consider improving the geospatial linkage variables, including geocoding the locations lacking existing geocoding and performing additional verification on FIPS data with low match rates. Evaluate how to code non-North Carolina states of residence.

3. **Data Linkage**
    a. For ED visit data linkage, perform a self-linkage of ED visits to identify multiple visits by the same person before linking the data to crash data. This would allow tracking visits by the patients beyond the initial visit after the crash. We could

also consider using the "InternalTrackingID" variable as a way of following patients within the same system.

b. In combination with 2f above, create more lenient linkage requirements for pedestrians, bicyclists and motorcyclists.
c. Continue working to simplify the linkage passes where possible. For example geospatial data was dropped from the crash and death certificate linkage when it was found not to be needed.
d. Review a sample of the matches found by LinkSolv but not found by the CHD methodology to determine if any changes to the CHD methodology would improve the linkage.
e. Continue to review and analyze the linked and unlinked data to determine possible ways the linkage could be improved.

4. Post-Linkage Processing

a. Remove duplicates created in the same linkage pass.
b. Filter the results to meet minimum linkage requirements, such as the following drafted minimum linkage standards:
    i. Crash/death certificate linkage
        1. Must have 2 demographic matches, 1 must be DOB or age
        2. Death data must indicate MVC OR crash report data must indicate that the person died
        3. Crash county=death county OR match on residence zip, city or state if out of state
        4. Death must occur within 30 days of crash
    ii. Crash/ED visit linkage
        1. Must match on 3 demographic variables, 1 must be DOB or age
        2. ED visit must indicate MVC or crash must indicate injury
        3. Crash county=hospital county OR match on residence zip, city or state if out of state
        4. ED visit must occur within an agreed upon time period such as 3 or 7 days after the crash

5. Coding

a. Continue cleaning and documenting the R code. Consider more robust functions or a package for wider use.
b. Document the code in a shared repository and include a change log.
c. Consider converting the process from R to other formats such as SQL or SAS.

6. Project overall

a. Evaluate and document any ongoing infrastructure needs, including hardware (e.g. specific desktop-based or distributed server-based systems) and supporting, non-linkage software (e.g. operating systems, database software like SQL server).
b. Encourage and participate in efforts to improve data quality of the original data.
c. Pursue future funding based on results analysis now that we have established linkage methodology.

Next steps

Based on the approved CDC funding for a year 2 of the Motor Vehicle Supplement and extended contract between NC DHHS and UNC-Chapel Hill, we will:

1. Obtain additional year(s) of data used for data linkage in Year 1 (2019-2020) and submit all documented changes in the mid-period and final progress reports.
2. Analyze the linked 2018 motor vehicle crash and health outcomes data in order to assess health outcomes and/or evaluate the data linkage.
3. Produce at least two data linkage products (e.g. reports, fact sheets, presentations) based on the linked data. At least one product shall focus on a prioritized research question as determined in collaboration between IPRC and IVPB. One of these products shall be an online presentation (webinar or conference presentation) that can be archived.
4. Disseminate the data linkage products to stakeholders and include in the final progress report lists of products and stakeholders with whom the information is shared.

In addition, we will review our recommendations and implement them as time and resources allow. We will also continue to pursue stable, long term funding.

Appendices

Appendix A: Written approval to use NC DETECT and death certificate data

NC DEPARTMENT OF HEALTH AND HUMAN SERVICES

ROY COOPER • Governor
MANDY COHEN, MD, MPH • Secretary
BETH LOVETTE, MPH, BSN, RN• Acting Director,
Division of Public Health

April 10, 2019

Alan Dellapenna, RS, MPH
Branch Head, Injury and Violence Prevention Branch
Chronic Disease and Injury Section
NC Division of Public Health
1915 Mail Service Center
Raleigh, NC 27699-1915

RE: CDC-RFA-CE16-1602- Motor Vehicle Data Linkage

Dear Mr. Dellapenna:

As the state's Enhanced Surveillance Director, I am writing to offer our support for activities outlined in your CDC grant application, Motor Vehicle Data Linkage.

We have worked closely with NC Injury and Violence Prevention Branch (IVPB) for many years providing the North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT) data to prevention programs. We are happy to continue our successful collaboration with IVPB as it moves forward with motor vehicle and injury prevention activities. In this capacity, we will ensure that NC DETECT system continues to collect high quality emergency department data. We will work in close partnership with IVPB and their staff as we move this project forward. We will also be available to provide technical assistance, access and training for the NC DETECT system, as needed.

I look forward to our continued partnership.

Sincerely,

Lana Deyneka, MD, MPH
Director
Statewide Enhanced Surveillance and
Hospital Based Public Health Epidemiologist Programs
Division of Public Health, Communicable Disease Branch
North Carolina Department of Health and Human Services

NC DEPARTMENT OF HEALTH AND HUMAN SERVICES • DIVISION OF PUBLIC HEALTH
LOCATION: 225 N. McDowell St, Raleigh, NC 27613
MAILING ADDRESS: 1932 Mail Service Center
www.ncdhhs.gov • TEL: 919-707-5425 • FAX: 919-870-4803

NC DEPARTMENT OF HEALTH AND HUMAN SERVICES

ROY COOPER • Governor
MANDY COHEN, MD, MPH • Secretary
BETH LOVETTE, MPH, BSN, RN• Acting Director,
Division of Public Health

April 3, 2019

Alan Dellapenna, RS, MPH
Branch Head, Injury and Violence Prevention Branch
Chronic Disease and Injury Section
NC Division of Public Health
1915 Mail Service Center
Raleigh, NC 27699-1915

RE: CDC-RFA-CE16-1602- Motor Vehicle Data Linkage

Dear Mr. Dellapenna,

As director of the North Carolina State Center for Health Statistics (SCHS), I am writing to offer our support for activities outlined in your CDC grant application, Motor Vehicle Data Linkage.

We have worked closely with NC Injury and Violence Prevention Branch (IVPB) for many years around motor vehicle and injury prevention issues. I have participated in the North Carolina Motor Vehicle Crash Injury Data Linkage Project Stakeholder meetings for the past few years.

We are committed to provide death certificate data to be linked to the motor vehicle crash data. As NC implements its electronic death certificate system, we will work with the project to continue data sharing with even more timely and robust data.

We will work in close partnership with IVPB and their staff as we move this project forward. We will also be available to provide technical assistance as needed around health data.

I look forward to our continued partnership.

Sincerely,

Eleanor Howell

Eleanor Howell, MS
Director, State Center for Health Statistics
Division of Public Health
North Carolina Department of Health and Human Services

Appendix B: Data Documentation for NC DETECT

Emergency Department visits in NC DETECT

Overview: General description of data source

The North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT) is North Carolina's statewide syndromic surveillance system. The North Carolina Division of Public Health (NC DPH) created NC DETECT in 2004 in collaboration with the Carolina Center for Health Informatics (CCHI) in the UNC Department of Emergency Medicine. Its purpose was to address the need for early event detection and timely public health surveillance in North Carolina using a variety of secondary data sources. Authorized users are currently able to view data from emergency departments (EDs), the Carolinas Poison Center, and the Pre-hospital Medical Information System (PreMIS), as well as pilot data from select urgent care centers. Only ED visit data are available to request for research. NC DETECT is designed, developed and maintained by CCHI staff with funding and program oversight by NC DPH.

Data owner

NC Department of Health and Human Services; Division of Public Health

Data description and collection criteria

Extract from the North Carolina Healthcare Association's (NCHA) dataset consisting of select data elements from all ED visits to 24/7 civilian acute care hospital affiliated EDs in NC

Type of data: source or compiled/abstracted

Source

Are the data available to outside parties for analytical purposes?

Yes

Process to obtain the data for research

After obtaining IRB approval from home institution, a data request/proposal must be submitted for review and approval by the NC DETECT Data Oversight Committee (DOC). A Data Use Agreement (DUA) is generally required.

The DOC membership includes representatives from NC DPH, UNC Chapel Hill Department of Emergency Medicine, NCHA and others identified at the discretion of NC DPH. Full process described in detail at: https://www.ncdetect.com/ncd/public/dataRequestPresentation.action

Website

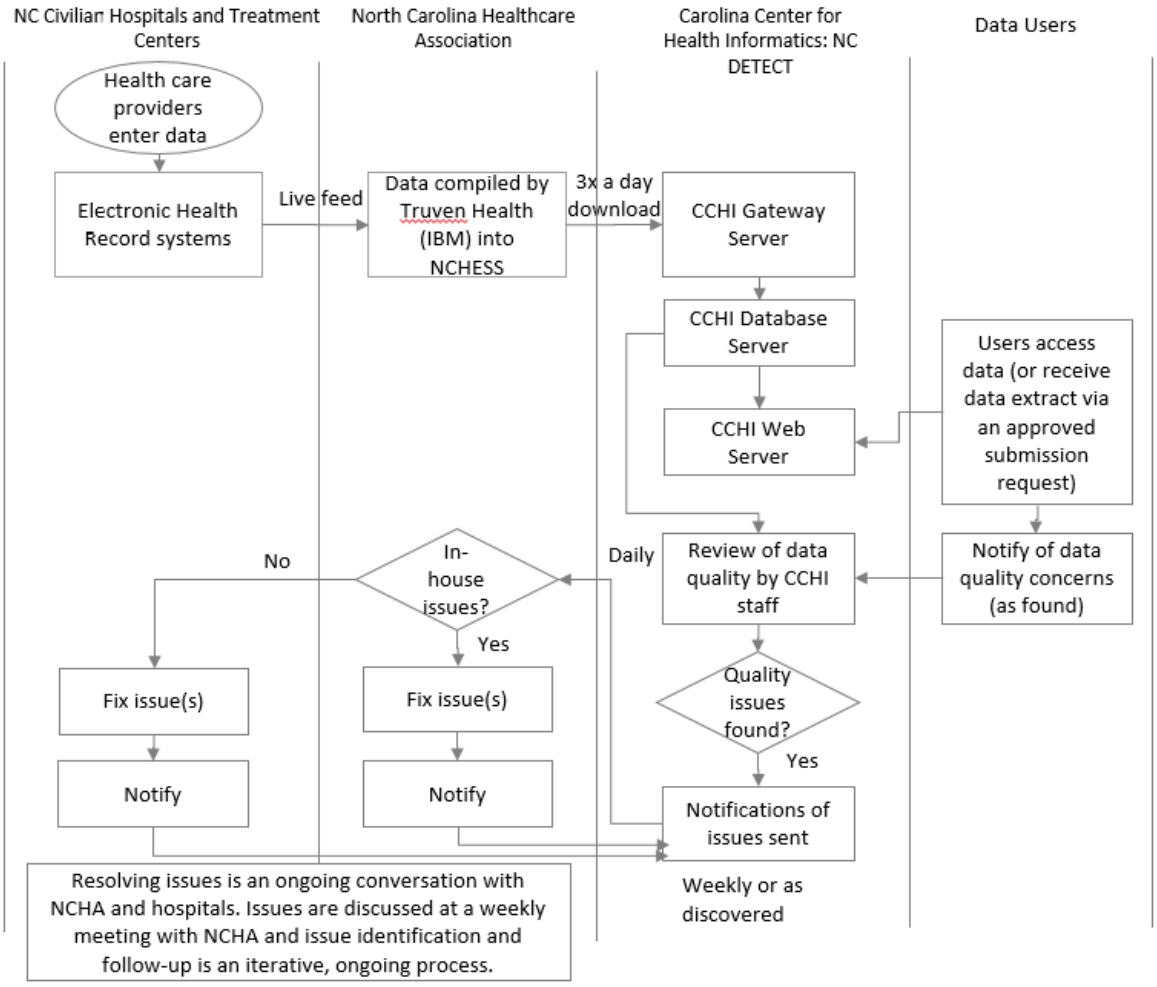
http://ncdetect.org/

Contact(s)

Clifton Barnett, MSIS
Data Quality Manager
Carolina Center for Health Informatics
919-843-0867
cbarnett@ad.unc.edu

Anna Waller, ScD
CCHI Executive Director
919-843-0389
anna_waller@med.unc.edu

Data Flowchart



Who enters the original data (Highway patrol officers, Healthcare providers, etc.)?
Emergency Department healthcare providers

Injury classification: Injury classification method (ICD-10-CM, etc.)
ICD-10-CM codes for diagnosis and external mechanism

Collection timeframe: when the data were entered after original event
Most data elements are collected close to real-time, but some data such as diagnosis codes may not be available for up to 3 months following event

Years available: Description of timespan for which data are available
Roughly 2008-present; there is variance when hospitals began participating. Full list of participating hospitals can be found here: http://ncdetect.org/participating-hospitals/ However, by 2008, NC DETECT estimated it captured over 99% of all ED visits in the state for that year. Most data are transmitted to NC DETECT within 48 hours of data collection, although there may be up to a three month lag in some diagnosis codes.

Data History: Key changes in the data that would affect research use
Hospitals sent ICD-9-CM codes through September 30, 2015; on October 1, 2015, they began sending ICD-10-CM codes.
Hospitals began sending PHIN codes for disposition instead of DEEDS codes in 2015-2016. However, for consistency, DEEDS codes are used for data requests.
Ethnicity and Race not reliably available until mid-2016.

Is a data dictionary available?
Yes

Dictionary
NCHESS Cookbook, CIHAtoNC_DETECT_fileformat_DRAFT_20170822

Field Mapping from Source Documentation

Source documentation field map

Report labels	Source labels (where available)
Table or category	
Field	Field Name (edited)
Field-Literal	Field Name
Description	Description
Source comments	Other Notes + Answer Options (if applicable)
Format	Data Type
Length	Length
Required (Y/N)	Required?
Sensitive (Y/N)	
Unique key (Y/N)	
Retired Field (Y/N)	
Retired Date	

Additional fields available in source documentation
None

Quality and Performance Measures

Known data quality issues
Up to 3 month delay in some diagnosis codes
Hospitals were added at different times
Initial ED Acuity is unreliable
Procedure codes are sparsely reported
Initial Temperature Read Route is unreliable
Ethnicity and Race not reliably available until mid-2016
Various documented gaps in provision of total data or significant individual fields (chief complaint, ICD codes)

Appendix C: Data Documentation for Death Certificate Data

Death registration data

Overview: General description of data source
N.C. Vital Records is part of the DHHS Division of Public Health and is located in Raleigh. In partnership with county registers of deeds offices, local health departments, and birthing facilities throughout the state, we are responsible for recording North Carolina vital events. This includes responsibility for legally registering all births, deaths, fetal deaths, marriages, and divorces which occur in North Carolina; coding these events for statistical purposes; maintaining these records; and providing certified or uncertified copies to individuals, researchers, and public health programs. http://vitalrecords.nc.gov/aboutus.htm

Data owner
NC Department of Health and Human Services; State Center for Health Statistics: https://schs.dph.ncdhhs.gov/

Data description and collection criteria
Data from death certificates filed with the North Carolina Vital Records office

Type of data: source or compiled/abstracted
Source data

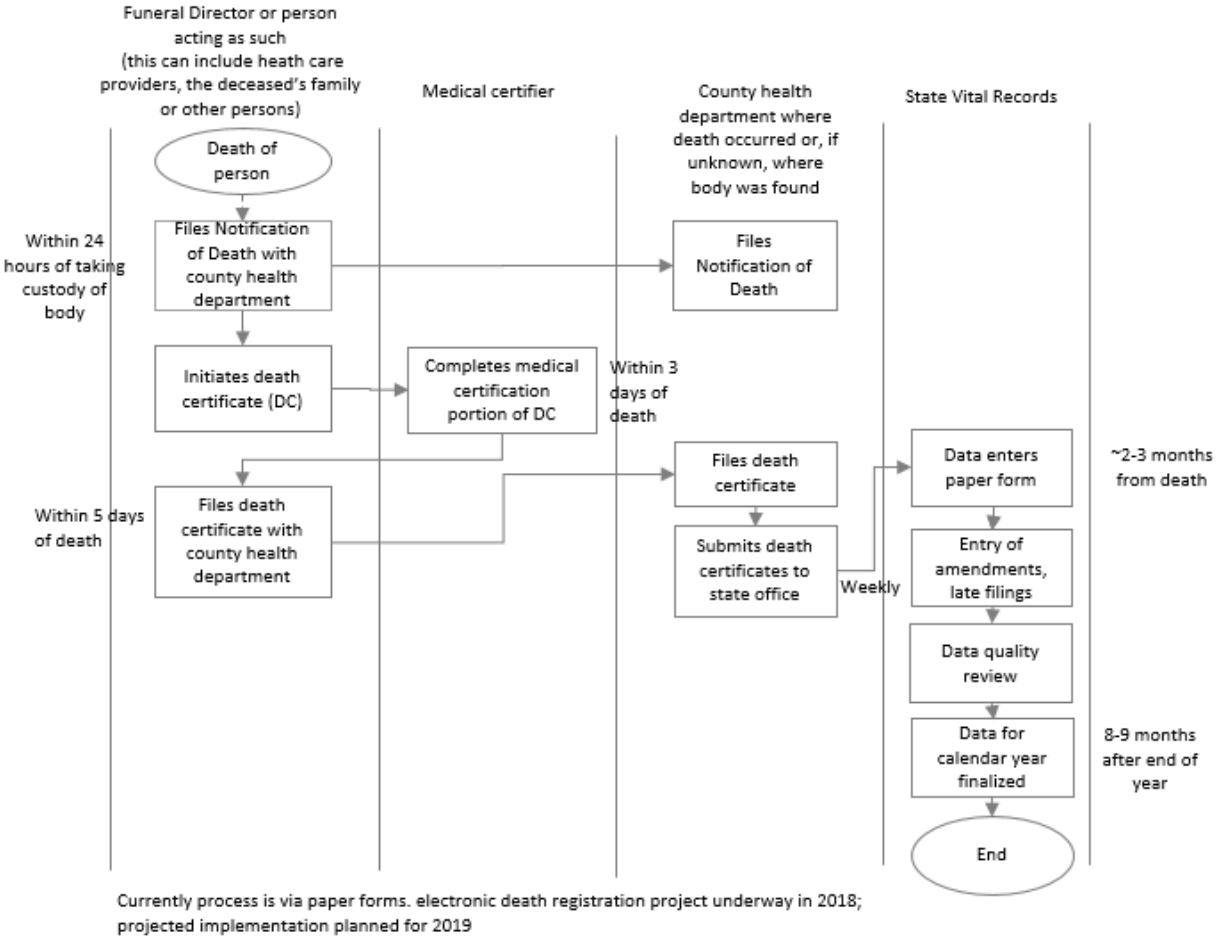
Are the data available to outside parties for analytical purposes?
Yes

Process to obtain the data for research
Death certificate data is available upon request. Per the definition of redaction in GS § 75-61, requestors may request the last four (4) digits of the deceased's social security number (SSN) to be included in the data. Requests for full SSN are reviewed and approved dependent upon the nature of the data requested. If you plan to conduct a study analyzing death certificate data, please contact the State Center for Health Statistics at SCHS.Info@dhhs.nc.gov Staff will review the study protocol and respond promptly.
Preliminary data may also be available monthly via a web server. Contact Matt Avery to obtain access to this data.

Website
http://www.schs.state.nc.us/units/stat/vital.htm

Contact(s)
Matt Avery, MA
Supervisor, Vital Statistics
Division of Public Health, State Center for Health Statistics
North Carolina Department of Health and Human Services
919-715-4572
Matt.Avery@dhhs.nc.gov

Data Flowchart



Who enters the original data (Highway patrol officers, Healthcare providers, etc.)?
Death registration form completed by funeral home directors or persons acting as such, medical certifiers and state and local registrars

Injury classification: Injury classification method (ICD-10-CM, etc.)
ICD-10-CM codes for diagnosis and external mechanism

Collection timeframe: when the data were entered after original event
The medical certification portion of the death certificate must be completed within three days of death. Death certificates must be filed with the local office within five days of death. The paper copies are typically filed at the state office within 2-3 months of the death.

Years available: Description of timespan for which data are available
Registration data are available starting from 1968. Each year of data are finalized around August or September for the previous year (2017 data should be finalized around September 2018).

Data History: Key changes in the data that would affect research use
The death certificate form was updated in 2014.

Is a data dictionary available?
Yes

Dictionary
2014-present Mortality File Layouts_Detail.xls

Documentation

Source documentation field map

Report labels	Source labels (where available)
Table or category	
Field	Variable
Field-Literal	SAS
Description	
Source comments	Code structure description
Format	
Length	Format
Required (Y/N)	
Sensitive (Y/N)	
Unique key (Y/N)	
Retired Field (Y/N)	
Retired Date	

Additional fields available in source documentation
None

Quality and Performance Measures

Known data quality issues
None

Appendix D: Data Linkage Data Elements

NC-CISS Linkage Data Elements: Crash, NC DETECT and Death Registration

Please note: This list should be considered preliminary only. The final linkage methodology is still being determined, but will be fully documented once complete.

Data Elements for Linkage

	Health Data Sources to be linked to Crash data	
Crash	**NC DETECT**	**Death Registration**
• Crash Date • Crash Time	Crash date on or prior to: Time of visit	Crash date on or prior to: • Date of Death--Month • Date of Death--Day • Time of Death
Zip code	Patient zip	Decedent's Residence—Zip code
Date of Birth	Patient Date of Birth	• Date of Birth--Year • Date of Birth--Month • Date of Birth--Day
Approximate Age	Patient age	• Decedent's Age--Type • Decedent's Age--Units
Gender	Patient sex	Sex
Possible Use in Linkage		
Ethnicity	• Race • Ethnicity	• Race of Decedent • Hispanic Origin of Decedent
• Address 1 • Address 2		Decedent's Residence— • Building/Street Number • Pre-Directional (North, West etc.) • Name of Street • Street Suffix (e.g. Street, Avenue, Blvd, etc.) • Post-Directional (SW, NE etc.) • Apartment/Lot Number
Person Address: City	Patient city	Decedent's Residence--City
Person Address: State	Patient state	Decedent's Residence--State
Crash county		County of Occurrence
Crash location (multiple data elements)		• Place of Death • Place of Injury

Potential Data Elements Used for inclusion criteria in Linked dataset

Data Source	Data Element	Comment
Crash	Injury Status	Possibly exclude O's
NC DETECT	Chief complaint text	Possible use of keywords
NC DETECT	Triage notes	Possible use of keywords
NC DETECT	ICD-10-CM codes 1 - 26	Injury diagnosis codes/MVC mechanism codes
Death Registration	ACME Underlying Cause of Death	Injury diagnosis codes/MVC mechanism codes
Death Registration	1st – 20th Mentioned Cause of Death	Injury diagnosis codes/MVC mechanism codes

Appendix E: Crash data delivery

From: Rodgman, Eric A <rodgman@hsrc.unc.edu>
Sent: Thursday, October 31, 2019 2:53 PM
To: Waller, Anna E <anna_waller@med.unc.edu>; Harmon, Katie Jean <harmon@hsrc.unc.edu>; Peticolas, Katherine Alice <kathy_peticolas@med.unc.edu>
Cc: Redding, Erika Megan <eredding@live.unc.edu>
Subject: RE: GHSP Project Manager - Erika

Katie and Anna::

Many thanks -- check to see if the 2018 file is there. Name = pers2018 .

This is a comma delimited file for all the records exported from SAS with the actual BAC results (variable name is bacres).

Let me know.

Eric

Appendix F: Data request for NC DETECT

Title: North Carolina Crash Injury Surveillance System (NC-CISS)

Data description: All ED visits in NC DETECT between January 1, 2018 and December 31, 2018. List of data elements (reasons for including in request are provided below): Chief complaint, ICD-10-CM diagnosis and injury mechanism codes (all), 5-digit ZIP code of residence, city of residence, county of residence, state of residence, ED disposition, insurance/expected source of payment, mode of transport, patient sex, patient DOB, patient age (years), internal tracking ID, visit ID, visit date, and visit time (1-hour blocks).

Variables:
Chief complaint – The recipients will use this additional text information to identify road user injury-related ED visits that are not captured using ICD-10-CM injury mechanism/diagnosis codes. The recipients will identify these visits using common crash keywords (e.g. MVC, MVA, CAR VS PED, etc.)
ICD-10-CM diagnosis and injury mechanism codes (ALL available codes) - The recipients will use these codes to identify crash-related injuries and to describe the nature and location of these injuries.
5-digit ZIP code – In order to integrate/link NC DETECT ED visit and crash report data, the recipients will require patient ZIP code. ZIP code of residence will be stripped from the analytical data set after data integration.
City of residence – In a previous data integration study using NC DETECT data, Dr. Harmon found that sometimes hospitals default to the ZIP code of the facility rather than the ZIP code of the patient residence; however, she found that city of residence is generally accurate. Therefore, for cases in which ZIP code of residence does not match, city of residence is a good alternative matching variable. City of residence will be stripped from the analysis data set after integration.
Patient county of residence – The recipients will use patient county of residence to identify geographic hot spots of crash injuries. The recipients will not display counts <10 by county.
State of residence -The recipients will need to know if the patient is a NC or out-of-state resident.
ED disposition – Disposition is an important indicator of injury severity.
Insurance/Expected source of payment – This will be another helpful descriptor of NC ED visits related to crash injuries. In addition, self-pay is an important indicator of potential health disparities.
Mode of transport – Mode of transport (e.g. arrive via ambulance) is another indicator of crash injury severity.
Patient sex – Important demographic descriptor of the ED visits among crash patients.
Patient date of birth (DOB) – In order to integrate/link NC DETECT ED visit and crash report data, the recipients will require patient DOB. Patient DOB will be stripped from the analysis data set after integration.
Patient age (in years) – Important demographic descriptor of ED visits. In addition, it is a key integration variable if patient DOB is missing.
Internal tracking ID and visit ID – These variables will be used to identify return visits to the same ED or hospital system by the same patient. Return/follow-up visits is also an indicator of injury severity.

Unique facility tracking ID – The recipients will require this ID to adjust for facility reporting issues, data outages, and known data quality issues.
Visit date and time (1-hour blocks) – In order to integrate/link NC DETECT ED visit and crash report data, we will require Visit Date and Time. In addition to data integration, visit date and time will help us to describe seasonal and time trends. Both factors have been associated with crash frequency and injury severity.

Project Title:
North Carolina Crash Injury Surveillance System (NC-CISS)

DESCRIPTION OF STUDY:
Motor vehicle crashes are a significant source of injury and fatality in North Carolina, but examination of the crash and socio-demographic factors involved with various health outcomes is limited by the lack of connection between crash and health outcome data. Emergency department (ED) visit data provide valuable insight into the demographics, treatment, and the nature and severity of injuries among persons treated for motor vehicle crash injuries, while police-reported motor vehicle crash data provide relevant details about the circumstances of the crash. This study will link NC DETECT ED visit data with crash data as part of a North Carolina Crash Injury Surveillance System (NC-CISS). NC-CISS will be used to obtain a more complete picture of the circumstances and outcomes associated with motor vehicle crash injuries in North Carolina.
The NC-CISS project is a collaboration between the NC Injury and Violence Prevention Branch (NC IVPB), Carolina Center for Health Informatics (CCHI), the UNC Highway Safety Research Center (UNC HSRC), the UNC Injury Prevention Research Center (UNC IPRC), the CDC, and others. It expands on a recent demonstration project, funded by the Governor's Highway Safety Program, which demonstrated the usefulness of NC DETECT ED visit data for integration with motor vehicle crash report data for a study of pedestrian and bicycle crash injuries.
The NC-CISS project will integrate NC DETECT ED visit data with motor vehicle crash report data for *all* road user types. Specifically, we would like to integrate NC DETECT ED visit and UNC HSRC crash data for all crash victims seeking treatment at 24/7 hospital affiliated civilian acute care EDs in NC. The data linkage process may include some case definition evaluation. The integrated data, as part of the NC-CISS, will be used to 1) describe the characteristics of medically attended injuries among patients involved in crashes and 2) to identify predictors of hospital admission and death among patients involved in crashes. Data linkage will be performed by Drs. Mike Fliss and Katherine Harmon.

PRINCIPAL RESEARCHER AND CO-INVESTIGATORS:
Kathy Peticolas, Project Coordinator/Data Analyst/CCHI Liaison
Injury and Violence Prevention Branch
Division of Public Health, Chronic Disease and Injury Section
NC Department of Health and Human Services
100 Market Street, 1st floor Chapel Hill, NC 27516
Email: kathy_peticolas@med.unc.edu Phone: (919) 448-5314

Co-Requesters:
Clifton Barnett cbarnett@ad.unc.edu (919) 843-2360 UNC CCHI

Will Curran-Groome wcggcw@live.unc.edu (919) 962-2202 UNC HSRC
Dennis Falls dennis_falls@med.unc.edu (919) 843-0815 UNC CCHI
Mike Fliss mike.dolan.fliss@unc.edu (919) 843-6755 UNC IPRC / NC IVPB
Katherine Harmon harmon@hsrc.unc.edu Phone: (919) 962-0745 UNC HSRC
Amy Ising amy_ising@med.unc.edu (919) 843-0814 UNC CCHI
Kendall Knuth Kendall.Knuth@dhhs.nc.gov NC IVPB
Anna Waller anna_waller@med.unc.edu (919) 843-0389 CCHI
DPH contact:
Alan Dellapenna alan.dellapenna@dhhs.nc.gov 919-707-5441 NC IVPB

PUBLIC HEALTH SIGNIFICANCE:
In 2016, motor vehicle crashes were the second leading cause of fatal injury (N=1,472 deaths) and the second and third leading mechanism of injury for hospitalizations (N= 7,963 hospitalizations) and emergency department visits (N= 67,784 visits) in North Carolina, respectively. Emergency department visit data provide valuable insight into the demographics, treatment, and the nature and severity of injuries among persons treated for motor vehicle crash injuries, while police-reported motor vehicle crash data provide relevant details about the circumstances of the crash. This study will link NC DETECT emergency department visit data with crash data as part of a North Carolina Crash Injury Surveillance System (NC-CISS). NC-CISS will be used to obtain a more complete picture of the circumstances and outcomes associated with motor vehicle crash injuries in North Carolina. Ultimately, it is hoped that this work will lead to ongoing annual data linkage and analysis of crash and emergency department visit data to inform prevention efforts in NC.

RESULTS OF PROJECT WILL BE USED FOR:
Contract requirements and presentation. Contract number: UNC/DHHS contract 00039605. Project aims include: 1. Integrate NC DETECT ED visits and crash data for all road users as part of a new North Carolina Crash Injury Surveillance System (NC-CISS). 2. Describe the characteristics of those persons injured in MVCs. 3. Identify predictors of hospital admission among persons injured in MVCs. Through integrating NC DETECT ED visit and crash data, the recipients will perform a comprehensive analysis of factors associated with crash morbidity and mortality. This information is needed to inform system-level interventions aimed at reducing crash injuries and fatalities. In addition to fulfilling our contract requirements, the recipients will provide regular updates to our partners at CCHI and NC DPH about our progress.

TIME PERIOD FOR DATA REQUESTED:
January 1, 2018 – December 31, 2018.

ESTIMATED STUDY COMPLETION DATE:
12/31/2020

Appendix G: Data request for death certificate data

Study background and description

Study Title: North Carolina Crash Injury Surveillance System (NC-CISS)

In 2016, motor vehicle crashes were the second leading cause of fatal injury (N=1,472 deaths) and the second and third leading mechanism of injury for hospitalizations (N= 7,963 hospitalizations) and emergency department visits (N= 67,784 visits) in North Carolina, respectively. While mortality and morbidity data provide valuable insight into the demographics, treatment, and the nature and severity of injuries among persons treated for motor vehicle crash injuries, these data sources provide little information about the actual crash event. On the other-hand, police-reported motor vehicle crash data provide many relevant details about the circumstances of the crash, but little information about the patient's outcome. In order to obtain a more complete picture of the circumstances and outcomes associated with motor vehicle crash injuries in North Carolina, this study will link death registration data with crash data as part of a statewide integrated motor vehicle crash injury surveillance system. We will conduct descriptive and inferential analyses of the linked and unlinked data to expand our understanding of the crash and socio-demographic factors associated with crash fatalities. We will disseminate findings, pending review and approval by State Center for Health Statistics staff, to research, policy, and advocacy communities, as well as the general public.

We are requesting all publicly available upon request death records within the period January 1, 2018 - December 31, 2018. We understand that the 2018 dataset is still being processed and we are willing to wait until its completion before receipt of the data. We are specifically interested in demographics (age, sex, race/Hispanic ethnicity, county of residence, occupation, and education), immediate and underlying causes of death (ICD-10 codes and literal text fields), date/time of injury (if available), location of injury (if available), and date/time of death. We are requesting the text fields, because the use of these fields was suggested by the NC Injury & Violence Prevention Branch for identifying alcohol-involved crash fatalities, as the ICD-10 codes indicating alcohol involvement (Y90-Y91) are frequently underreported.

The research team for this project includes Dr. Anna Waller, ScD, Executive Director and Research Professor at the UNC Carolina Center for Health Informatics and Department of Emergency Medicine; Dr. Katie Harmon, PhD, Postdoctoral Research Associate at the UNC Highway Safety Research Center (HSRC); Dr. Mike Fliss, PhD, MPS, MSW, Research Scientist at the UNC Injury Prevention Research Center and Epidemiologist & Public Health Informatician at the Injury and Violence Prevention Branch in the Division of Public Health, North Carolina Department of Health and Human Services; Amy Ising, MSIS, Associate Director of the UNC Carolina Center for Health Informatics; Dennis Falls, Database Administrator and Security and Privacy Officer for the UNC Carolina Center for Health Informatics; Clifton Barnett, MSIS, Data Quality Manager and Data Analyst for the UNC Carolina Center for Health Informatics; and Kathy Peticolas, MPS, PMP, Project Manager at the Injury and Violence Prevention Branch in the Division of Public Health, North Carolina Department of Health and Human Services. Dr. Waller will be responsible for guiding the study, developing research questions, and reviewing the analysis plan, findings, and final

publications. Drs. Fliss and Harmon will perform the data linkage, data analysis and will develop publications. Ms. Ising will assist with project guidance and preparation of reports. Mr. Falls will assist in the data linkage and insure data security and integrity throughout the process. Mr. Barnett will assist with data analysis, report preparation, and data quality assurance. Ms. Peticolas will manage the project and assist with data preparation, analysis, and writing of results.

Provide a brief description of the public health benefits of this study:
Resulting analyses, publications, and reports will be targeted toward other researchers, policy makers, advocates, and the lay public in order to provide a more data-driven conversation around opportunities to better protect the safety of North Carolina's residents and visitors. Further, findings from the proposed analyses may serve to inform ongoing research and policy relating to road safety in other states throughout the US. Ultimately, it is hoped that the results of this work will lead to ongoing annual data linkage and analysis of crash and death certificate data to inform prevention efforts in NC.

Provide a brief account of security measures, including the conditions under which the records or data will be used, stored, and disposed of and any other security precautions in place to ensure the confidentiality of the data:
The CCHI Security and Privacy Officer, Dennis Falls, is a member of this project team and will have responsibility for insuring the security of these data. We employ multiple levels of information security and redundancies, including: password-protecting the data so that only those individuals directly involved in the project have access; storing all files behind UNC at Chapel Hill's secure firewalls; full-disk encrypting all laptops used by project staff; and deleting all data at the conclusion of the project. All staff involved in the proposed project have up-to-date certifications in social and behavioral research from the CITI Program, the premier research ethics and compliance training organization; these certifications include coursework in privacy, confidentiality, and data protection practices. Lastly, we will only present aggregated results from analyses, and will present all draft publications or other public-facing materials to State Center for Health Statistics staff for review and approval prior to dissemination. We request that the death data be provided to UNC CCHI as a SAS or excel file. It can be burned to a CD-rom or sent via secure FTP site.

Appendix H: Implementation Plan

1 Introduction

1.1 Purpose

This plan describes the development and implementation of a statewide integrated motor vehicle crash injury surveillance system called the North Carolina Crash Injury Surveillance System (NC-CISS). The purpose of NC-CISS is to obtain a more complete picture of the circumstances and outcomes associated with motor vehicle crash injuries in North Carolina.

1.2 System Overview

1.2.1 System Description

The first year of NC-CISS will use three data sources: crash report data from the UNC Highway Safety Research Center (HSRC), NC DETECT emergency department (ED) visit data from NC Division of Public Health (NC DPH), and death certificate data from the State Center for Health Statistics (SCHS). Crash report data will be separately linked to NC DETECT ED visit data and to death certificate data to create data analysis files for motor vehicle crash injury research.

1.2.2 Assumptions and Constraints

The following are the assumptions regarding the development and execution of this plan:

1. The schedule allows for sufficient time to complete the implementation.
2. The budget is sufficient to complete the project and no additional sources of funding are needed.
3. The persons assigned to this project will be available as needed. In particular, key subject matter experts, Mike Fliss and Katie Harmon, will be available to devote time during the critical stages of the data linkage planning and implementation.
4. Secure data storage and support will be available for use in this project.

1.2.3 System Organization

1.3 Glossary

CCHI	Carolina Center for Health Informatics
CDC	Centers for Disease Control and Prevention
DMV	Division of Motor Vehicles
DOB	Date of birth
DOD	Date of death

DOT	Department of Transportation
ED	Emergency Department
EMS	Emergency Medical Services
GHSP	Governor's Highway Safety Program
HSRC	Highway Safety Research Center
IPRC	Injury Prevention Research Center
IVPB	Injury and Violence Prevention Branch of NCDHHS
MVC	Motor vehicle crash
NC	North Carolina
NC DETECT	North Carolina Disease Event Tracking and Epidemiologic Collection Tool
NCDHHS	North Carolina Department of Health and Human Services
NC DPH	North Carolina Division of Public Health
NHTSA	National Highway Traffic Safety Administration
PVA	Public Vehicular Area
SCHS	State Center for Health Statistics
UNC	University of North Carolina

2 Management Overview

2.1 Description of Implementation

After gaining access to the three data sources, we will systematically determine a sustainable process to link each of the two health outcome datasets to the crash dataset to create linked data that are as complete, accurate, and representative of the crash injury population as possible. After the linkage methodology for each is determined, linked and cleaned analysis datasets will be created. The process will be documented so that it is replicable and sustainable. The process may be updated if the data change, or if other data sources are added in the future.

2.2 Points-of-Contact

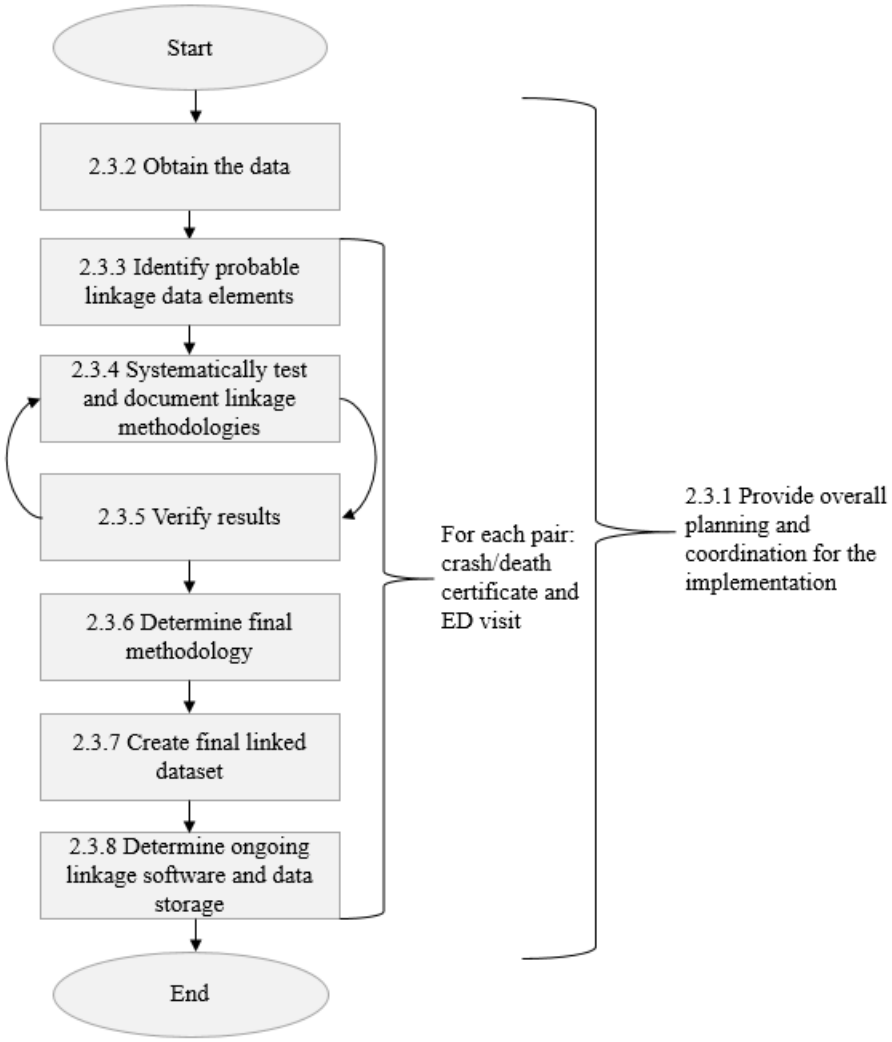
The project is a collaboration between the NC DHHS Injury and Violence Prevention Branch (DHHS IVPB), the UNC Carolina Center for Health Informatics (CCHI), the UNC Highway Safety Research Center (HSRC) and the UNC Injury Prevention Research Center (IPRC). See Table 1 for the list of project team members.

Table 1: Project Team Members

Role	Agency/Dept.	Name	Email
Business Sponsor	DHHS IVPB	Alan Dellapenna	alan.dellapenna@dhhs.nc.gov
Project Manager	DHHS IVPB	Kathy Peticolas	Kathy_peticolas@med.unc.edu
UNC/DHHS Contract Contact	DHHS IVPB	Ingrid Bou-Saada	Ingrid.Bou-Saada@dhhs.nc.gov
Agency Contact	DHHS IVPB	Scott Proescholdbell	Scott.Proescholdbell@dhhs.nc.gov
Agency Contact	DHHS IVPB	Kendall Knuth	Kendall.Knuth@dhhs.nc.gov
Principal Investigator	CCHI	Anna Waller	Anna_waller@med.unc.edu
Subject Matter Expert	CCHI	Amy Ising	amy_ising@med.unc.edu
Quality Assurance Manager	CCHI	Clifton Barnett	Clifton_barnett@med.unc.edu
Security Officer/DBA	CCHI	Dennis Falls	Dennis_falls@med.unc.edu
Subject Matter Expert/Data User	HSRC	Katie Harmon	harmon@hsrc.unc.edu
Agency Contact	HSRC	Nancy Lefler	lefler@hsrc.unc.edu
Subject Matter Expert	HSRC	Eric Rodgman	rodgman@hsrc.unc.edu
Agency Contact	IPRC	Steve Marshall	smarshall@unc.edu
Subject Matter Expert	DHHS IVPB /IPRC	Mike Fliss	mike.dolan.fliss@unc.edu

2.3 Major Tasks

The following lists the major implementation tasks to be completed for each pair of datasets (crash/NC DETECT and crash/death certificate data). Although the tasks are generally sequential, multiple tasks may be implemented simultaneously.



2.3.1 Provide overall planning and coordination for the implementation

What the task will accomplish	*Ensure the project meets deliverable timelines*
Resources required to accomplish the task	*Input from team members*
Key person(s) responsible for the task	*Core team members*
Criteria for success	*Project completed on time and meeting all deliverables*

The following core team members will meet bi-weekly or as needed. Project meetings with all project contacts will be held monthly.

Role	Agency/Dept.	Name
Project Manager	DHHS IVPB	Kathy Peticolas
Principal Investigator	CCHI	Anna Waller
Subject Matter Expert/Data User	HSRC	Katie Harmon
Subject Matter Expert	DHHS IVPB /IPRC	Mike Fliss

2.3.2 Obtain the data

What the task will accomplish	*Obtain access to the three datasets*
Resources required to accomplish the task	*Time to submit applications; input from team members*
Key person(s) responsible for the task	*Kathy Peticolas*
Criteria for success	*Access to all data by 12/31/2019*

Table 2 describes the three data sources used and the selection criteria for this project.

Table 2 Data Description

Data	***Data description***	***Data requested***
Crash report data	North Carolina crash records meeting at least one of the following criteria: the crash resulted in a fatality, the crash resulted in a non-fatal personal injury, the crash resulted in total property damage amounting to $1,000.00 or more, the crash resulted in property damage of any amount to a vehicle seized, or the vehicle has been seized and is subject to forfeiture under G. S. 20-28.3. In addition, a reportable motor vehicle traffic crash must occur on a trafficway (any land way open to the public as a matter of right or custom for moving persons or property from one place to another) or occur after the motor vehicle runs off the roadway but before events are stabilized.	All persons listed in 2018 crashes, including crashes on public vehicular areas (PVAs), such as parking lots, which are not on a government-maintained road. PVA crashes do not meet the definition of a reportable crash, so their entry is not consistent.
Death certificate data	Data from death certificates filed with the North Carolina Vital Records office.	All 2018 deaths in North Carolina
NC DETECT ED visit data	As mandated by North Carolina Statute § 130A-480 Emergency Department Data Reporting, NC DETECT receives daily extracts through the North Carolina Healthcare Association (NCHA), consisting	All 2018 ED visits; does not include visits to the Cherokee Indian Hospital

	of select data elements from all ED visits to 24/7 civilian acute care hospital affiliated EDs in NC.

Table 3 describes the process to obtain each dataset.

Table 3 Process to obtain data

Data	*Data owner*	*Process to obtain*
Crash report data	UNC Highway Safety Research Center (HSRC)	Crash report data are not considered sensitive and do not require IRB approval. A formal data request was not needed due to HSRC's participation in this project. Data were requested and received from HSRC employee and project team member Eric Rodgman.
Death certificate data	NC DHHS State Center for Health Statistics (SCHS)	North Carolina death certificate data are considered part of public health surveillance and do not require IRB approval. A signed SCHS Data Request for Identified Data for Research Purposes form F-14 was submitted to SCHS employee Matt Avery and the data were obtained on October 17, 2019. A revised request to include names (for linkage verification) was submitted and the data received on December 6, 2019. See APPENDIX A: Death Certificate Data Request for the submitted form. No approval document was received from the SCHS.
NC DETECT ED visit data	NC DHHS NC DPH Communicable Disease Branch & Injury and Violence Prevention Branch (IVPB)	NC DETECT ED visit data are considered to be sensitive health data. An IRB request for the project was submitted to UNC on October 7, 2019 and approved October 29, 2019 (IRB 19-2675). The data request was submitted via the online application on October 22, 2019 to the NC DETECT Data Oversight Committee, Communicable Disease Branch, NC DPH. The NC DETECT Data Oversight Committee approved the request on November 18, 2019. The DUA was signed and executed on December 9, 2019. See APPENDIX B: NC DETECT Data Request Application.

See below for a description of the three datasets.

Data	*Number of Records*	*Number of Variables*	*File type*
Crash report data	832,058	69	Tab-delimited text file
Death certificate data	94,867	127	SAS file
NC DETECT ED visit data	5,084,987	63	Pipe-delimited text file

2.3.3 Identify probable linkage data elements

What the task will accomplish	*Identify the data elements most likely to be used for linkage*
Resources required to accomplish the task	*Input from team members; documentation from prior linkage projects*
Key person(s) responsible for the task	*Core team members*
Criteria for success	*A list of data elements to be used for linkage per dataset*

Data elements to be used for linkage were selected based on the available data elements in the HSRC crash report data and on previous data linkage projects. Although person name is a common data element used for linking datasets, person name is not available in HSRC crash report data or NC DETECT ED visit data in 2019/2020. If person name becomes available in the future, it should be considered for data linkage where applicable and/or used for linkage verification. The list of linkage data elements may be revised throughout the project. Table 3 identifies the data elements identified as the most likely to be used for linkage.

Table 4 Linkage data elements

Data Element	***Crash Report Data***	***NC DETECT***	***Death Certificate Data***
Event date	*accdate*	*Crash date on or prior to:* *TIME_OF_VISIT*	*Crash date on or prior to date of death:* *DOD_MO* *DOD_DY* *DOD_YR*
Zip code of residence	*RZIP*	*PATIENT_ZIP*	*ZIPCODE*
Date of birth	*perdob*	*PATIENT_DOB*	*DOB_MO* *DOB_DY* *DOB_YR*
Age	*AGE*	*PATIENT_AGE*	*AGE* *AGETYPE (units, 1=Years)*
Sex	*SEX*	*PATIENT_SEX*	*SEX*
Race/ethnicity	*RACE*	*RACE* *ETHNICITY*	*RACER* *RACE1-RACE23* *DETHNIC1-DETHNIC5*
Street address of residence	*ADDR1* *ADDR2*		*ADDRNUM (street number)* *ADDRPRED (pre direction)* *ADDRNAME (street name)* *ADDRSUFF (e.g. Street, Blvd)* *ADDRPOST (post direction)* *ADDRAPT (apt/lot number)*
City of residence	*CITYNAME*	*PATIENT_CITY*	*CITYRESTEXT* *CITYC (coded)*
County of crash	*COUNTY*		*COD (county of death)*

County of residence	*COUNTY (of crash, could be used for verification)*	*PATIENT_COUNTY*	*COUNTYC*

2.3.4 Systematically test and document linkage methodologies

What the task will accomplish	*Provide the basis and justification for the final methodology*
Resources required to accomplish the task	*Input from team members; programming and processing time*
Key person(s) responsible for the task	*Mike Fliss (expertise and implementation); Katie Harmon (expertise)*
Criteria for success	*A documented process of determining how to link the two datasets*

This step describes an overview of the core process of determining the final linkage methodology. The final methodology will be documented in a separate report. The methodology may be determined prior to the selection of the software that will be used for ongoing data linkage.

This process will be informed by the following:

1. Literature related to data linkage and motor vehicle crash projects.
2. Project member experience with these data sources and data linkage in general.
3. Consultation and feedback from the CDC and representatives from other states doing similar data linkage projects.
4. Consultation with data linkage expert Dr. Larry Cook, a professor at the University of Utah School of Medicine. Dr. Cook will be visiting in order to consult on the project in February of 2020.
5. Comparison with the results of the probabilistic linkage software LinkSolv and with previous data linkage projects that used hierarchical deterministic linkage.

The following lists the steps towards determining the methodology.

1. Prepare the data.
    a. Identify or assign a unique identifier(s).

Data will be linked using the person as the base unit. The expectation is that crash/death certificate data will have one to one matches and crash/NC DETECT data will have one to many matches, because persons may seek medical treatment multiple times for the same crash. A

person may also have more than one crash that resulted in an ED visit and a person may have had crashes prior to the crash that resulted in their death.

Table 5 Identifiers

Level	***Crash report data***	***NC DETECT***	***Death Certificate Data***
Person	*CRSH_ID & VEHPOS & PERNUM*	*InternalTrackingID (within same hospital facility)*	*No unique identifier in source data; must be created*
Vehicle in Crash	*CRSH_ID & VEHPOS*		
Crash	*CRSH_ID*		
Health Care Facility		*EDFACILITYID*	
ED visit		*VISITID*	

b. Remove duplicate records.

c. Harmonize the core linkage datasets by cleaning and standardizing linkage variables. How the data are cleaned may be adjusted as different methodologies are tried. The final cleaning algorithm will be documented.

2. Determine the usefulness of the different linkage variables and how they will be linked.

Models will be generated in R to determine the best way to link the data. This may include deterministic matching, decision tree-based methodologies and probabilistic linkage.

3. Determine the selection criteria and linkage variables for an initial high confidence linkage. This may include human hand review of a well-selected (e.g. some random, some targeted) subset of the potential linkages pairs. The human hand reviewed set may be used as a test and training set for models.

4. Determine subsequent linkage algorithms to link further records, recording the number of linked records as a result of each step.

5. Adjust all aspects of the process as needed, including the selection criteria, how the data are cleaned, linkage variables used, approaches to missingness, how one-to-many matches are handled and the linkage algorithms until a viable methodology is attained.

6. The initial data linkage methodology and the barriers and facilitators of the process will be documented in a report scheduled to be completed by April 30, 2020.

2.3.5 Verify results.

What the task will accomplish	*Verify the accuracy of linkage*

Resources required to accomplish the task	*Review time, records pulled by HSRC*
Key person(s) responsible for the task	*Mike Fliss to identify when it is time to test; review done by other team members*
Criteria for success	*Documentation of verification results*

The results of data linkage will be verified to evaluate the linkage methodology. Once a viable methodology is determined, a selection of linked and unlinked records will be pulled for evaluation by team members who did not implement the linkage. For the crash/death certificate linkage, a selection of corresponding full crash reports will be requested from HSRC for individual record review. Because both datasets include the person name, this should increase the accuracy of the verification. Because no names are available with NC DETECT data, the data may be limited to a visual review of whether the match was likely correct, utilizing all available information for that record.

Each record will be reviewed individually and, when possible, assigned as a true positive, a false positive, a false negative and true negative. This process will also be followed if different methodologies for the same linkage are compared.

	n		n
True Positive (a)	**?**	**False Positive (c)**	**?**
False Negative (b)	**?**	**True Negative (d)**	**?**

The results will be scored according to the following formulas:

Measure	**Formula**
Sensitivity	$\frac{a}{a+b}$
Specificity	$\frac{d}{c+d}$
Positive Predictive Value (PPV)	$\frac{Sensitivity}{1-Specificity}$
	$\frac{1-Sensitivity}{Specificity}$

Negative Predictive Value (NPV)	
Accuracy	$\frac{(a+d)}{(a+b+c+d)}$
Cohen's Kappa[1,2]	$\frac{(p_o - p_e)}{(1-p_e)}$

[1] p_o=Observed agreement (identical to accuracy).

[2] p_e= Probability of chance agreement.

2.3.6 Determine final methodology.

What the task will accomplish	*Provide a plan for linkage for this year and future years*
Resources required to accomplish the task	*Input from all team members*
Key person(s) responsible for the task	*All core team members*
Criteria for success	*Documentation report on the final linkage methodology for both dataset pairs*

The final methodology will be determined based on an evaluation of the results of the model testing and the data verification. It will also include an evaluation of the sustainability of the methodology. The core team will document the final methodology in a final report scheduled to be completed by July 31, 2020. A handout summarizing the results will accompany the full report.

2.3.7 Create final linked datasets.

What the task will accomplish	*Create analysis datasets from linked data*
Resources required to accomplish the task	*Input from team members*
Key person(s) responsible for the task	*Mike Fliss*
Criteria for success	*Linked datasets ready for analysis*

The final linked datasets will be created using the final determined methodology and stored securely on the UNC OneDrive, where it will be available for data analysis.

2.3.8 Determine ongoing linkage software, data storage, and infrastructure

What the task will accomplish	*Establish a working framework for ongoing linkage*
Resources required to accomplish the task	*Input from all team members, consultation with CCHI team members who manage NC DETECT data*
Key person(s) responsible for the task	*All team members, Dennis Falls and Clifton Barnett*
Criteria for success	*A documented process for replicable linkage*

The software that will be used for ongoing data linkage may be the same as the software used to determine the methodology or it may differ. Project team members will determine the software to be used based on its cost, the need for training or specialized expertise, and ease of use and documentation. Infrastructure to be documented includes hardware (e.g. specific desktop-based or distributed server-based systems) and supporting, non-linkage software (e.g. operating systems, database software like SQL server).

2.4 Implementation Schedule

The following is the listing of the planned activities and tasks for this project.

Activity 1: Develop a description of datasets used for data linkage.		
Date to be Completed	Task	Description
Aug. 15, 2019	Task 1. Confirm two data sources to be linked to NC motor vehicle crash data (NC DETECT emergency department visit and SCHS death certificate/vital records data)	a. Receive written approval from the data owner (Communicable Disease Branch at NC DPH) to proceed with linking crash report data to NC DETECT emergency department visit data
		b. Receive written approval from the data owner (NC SCHS) to proceed with linking crash report data to death certificate/vital records data
		c. Inform state and federal partners of approval decisions
Aug. 30, 2019	Task 2. Complete documentation for data source # 1 (NC DETECT emergency department visit data)	a. Thoroughly document data source # 1 (NC DETECT emergency department) according to CDC guidelines

Sept. 15, 2019	Task 3. Complete documentation for data source # 2 (SCHS death certificate/vital records data)	a. Thoroughly document data source # 2 (SCHS death certificate/vital records data) per CDC guidelines
Sept. 30, 2019	Task 4. Submit data documentation to CDC	a. Review and edit drafts of data documentation
		b. Submit data documentation to CDC
Activity 2. Develop data linkage methodology.		
Date to be Completed	Task	Description
Aug. 31, 2019	Task 1. Obtain approval from UNC Institutional Review Board (IRB)	a. Submit IRB to UNC IRB
		b. Receive approval from UNC IRB
Sept. 30, 2019	Task 2. Obtain at least one full year of motor vehicle crash report data	a. Request and receive 1+ years of motor vehicle crash report data from HSRC
Oct. 31, 2019	Task 3. Obtain approval from the Communicable Disease Branch of NC DPH for 1+ year of NC DETECT emergency department visit data	a. Obtain approval from Communicable Disease Branch (may or may not involve formal DUA)
		b. Provide approval documentation in Implementation Plan
Nov. 30, 2019	Task 3. Obtain approval from NC SCHS for 1+ year of death certificate/vital records data	a. Obtain approval form NC SCHS (may or may not involve formal DUA)
		b. Provide approval documentation in Implementation Plan
March 31, 2020	Task 4. Complete data implementation plan	a. Review literature related to data linkage
		b. Establish communication with one or more data linkage experts; invite linkage expert to travel to NC to consult with Project Team members
		c. Develop methodology to link crash and health outcome data based on a) literature review, b) consultation with experts, and c) Project Team members' expertise
		d. Draft Implementation Plan, share with partners, and solicit feedback; revise, as needed

		e. Submit Implementation Plan to CDC
Activity 3: Identify barriers and facilitators.		
Date to be Completed	Task	Description
4/30/2020	Task 1. Document barriers and facilitators to data linkage	a. Document any barriers and facilitators related to planning, implementation, and linkage of crash report/ health outcome data sources for project duration
		b. Record all barriers/facilitators in a summary report
		c. Submit report to CDC
Activity 4: Summarize data linkage.		
Date to be Completed	Task	Description
Jan. 31, 2020	Task 1. Obtain health outcome data sources	a. Obtain NC DETECT emergency department visit data
		b. Obtain NC SCHS death certificate/vital records data
4/30/2020	Task 2. Link two health outcome data sources	a. Link NC DETECT emergency department visit data with crash data
		b. Link NC SCHS death certificate/vital records data
		c. Incorporate methodology developed as part of Implementation Plan (Activity 3)
		d. Document all deviations from Implementation Plan and explanation
7/31/2020	Task 3. Summarize data linkage activities	a. Prepare a report summarizing data linkage activities; include all required CDC indicators
		b. Prepare a handout succinctly summarizing data linkage activities
		c. Share report/handout with state and CDC partners

2.5 Security and Privacy

All listed participants on the project will abide by NC DPH's/NC DETECT's security requirements for the prevention of unauthorized disclosure of protected health information. The data will be stored on a limited access folder on the UNC OneDrive. Sensitive data elements will

be used for linkage only. These elements will be removed prior to analysis. All publications and presentations will be shared with NC DPH prior to submission.

3 Implementation Support

3.1 Hardware, Software, Facilities, and Materials

3.1.1 Hardware

No specific hardware is needed.

3.1.2 Software

The following software may be used:

Software	***Type***	***Licensing***
R	*Statistical/programming software*	*Free, open source*
SAS	*Statistical software*	*UNC*
LinkSolv	*Probabilistic matching software*	*License owned by UNC Department of Emergency Medicine*
SQL Server	*Database*	*License owned by UNC Carolina Center for Health Informatics*
Microsoft Office	*Various*	*UNC*
Tableau	*Data visualization*	*Public-use version only*

3.1.3 Facilities

The CCHI office at 100 Market Street in Chapel Hill will be used for meetings. The LinkSolv software may be on a desktop in this location.

3.2 Documentation

The first year of this project will produce the following list of documents.

	Document	**Planned completion date**	**Actual completion date**
1	Data documentation for NC DETECT ED visit data and SCHS death certificate data	9/30/2019	10/22/2019
2	Implementation plan for data linkage analysis	3/31/2020	

3	Mid-year brief report outlining progress to date on all performance requirements for NC DHHS	3/17/2020	
4	Report documenting initial data linkage methods for two health outcome data sources, including any barriers and facilitating factors to data linkage analysis	4/30/2020	
5	Final report summarizing data linkage activities and results in Year 1 (addresses all Performance requirements)	7/31/2020	
6	Summary handout of data linkage activities and results reported in Year 1	7/31/2020	

3.3 Personnel

3.3.1 Staffing Requirements

All relevant staff are included in 2.2 Points of Contact.

3.3.2 Training of Implementation Staff

No specific training needs have been identified for the project. Training needs will need to be reassessed once the final linkage methodology is determined.

3.4 Outstanding Issues

Funding after year one has not been procured. The project team will apply for additional years of funding, if offered.

3.5 Implementation Impact

The implementation is not expected to impact the network infrastructure, staff or any user communities. The implementation will require time from project members, but it is not anticipated that it will significantly impact other project requirements or Service Level Agreements in the first year.

3.6 Performance Monitoring

The success of the linkage itself will be monitored as described in section 2.3.5. Future years of data linkage will review a smaller number of individually linked records, if the linkage algorithm and the structure and content of the data sources has not substantially changed.

3.7 Configuration Management Interface

This is the first year of implementation for this system. This document will be updated with any changes to the system in future years, including adding new data sources.

4 Implementation Requirements

4.1 Carolina Center for Health Informatics

4.1.1 Site Requirements

Most of the project will be done virtually, with no site-specific requirements. The following lists the requirements of the project.

Type of Requirement	Description	Supplied by
Hardware	A desktop computer will be needed to use LinkSolv. This computer will be located in the CCHI office.	UNC CCHI
Processing	Distributed processing needs may be needed depending on the linkage methodology	UNC
Software	The initial linkage modeling will be done in R, which is open-source and does not require a license. Other software licenses that may be required (SAS, SQL Server, etc.) will be used under UNC's license.	UNC, unless open source
Database	The CCHI license of SQL Server may be used, depending on the final data storage determination	UNC CCHI
Facilities	CCHI offices at 100 Market Street in Chapel Hill will be used for meetings and to house the LinkSolv software.	UNC CCHI

4.1.2 Site Implementation Details

Because most of the project will be done virtually, there are no site-specific procedures. If the final linkage methodology and data storage require additional procedures, this will be revisited and revised.

4.1.3 Risks and Contingencies

The following includes the risks identified by the CDC in the 2015 "Assessment of Characteristics of State Data Linkage Systems." (1)

Risk	*Contingency Plan/Risk Mitigation*
Insufficient funding	Funding is sufficient for year one. We will apply for any additional years of CDC funding if it is offered and will explore other funding options.

Risk	*Contingency Plan/Risk Mitigation*
Staffing turnover	After determining the best data linkage methodology, we will have a documented process that can be followed by others. The process of determining the best linkage methodology, which requires the critical expertise of project team members, will take less than one year.
Lack of process documentation	We will thoroughly document the process, including what linkage algorithms were tried, but not chosen.
Long lag times in obtaining source data for linkage	The two selected health data sources are managed by NCDHHS, minimizing lag times. Gaining access to NC DETECT data can be accomplished within a few months. Because the death certificate data are not restricted, the data requests are typically handled within days. Crash data are obtained from project members from HSRC with minimal delay.
Statutory requirements for obtaining and reporting data	The project team has many years of experience obtaining NC DETECT data and getting IRB approval for its use. By documenting the process of obtaining the data, the process should be sustainable.
Complex linkage techniques such as probabilistic linkage	We will be testing a variety of linkage methodologies, including probabilistic linkage. If a more complex linkage technique is selected, we will have documentation showing the selection process and why it was the best methodology. Sustainability will be a major consideration when selecting the linkage methodology. The linkage technique chosen will be documented so that it is replicable by others.
Marketing linked data so that others understand how they can be used to increase traffic safety	We have experience creating targeted reports and factsheets based on the expressed areas of interest of our stakeholders on prior demonstration projects. Furthermore, our plan for year two, should funding be available, is to create a governance board to direct our research and ensure the results are shared with policy-makers who can make a difference in increasing traffic safety.

Risk	*Contingency Plan/Risk Mitigation*
Scaling issues: Piloting on smaller datasets and scaling to larger ones	Implementation started with relatively smaller datasets (e.g. deaths, crashes) may not scale without changes to larger datasets and have algorithms be able to run in a reasonable amount of time. Modern distributed architecture (which UNC has available) can assist with this challenge, as will simplifying algorithms for production after details, potentially less efficient algorithms are used for exploration and linkage model development.

4.1.4 Implementation Verification and Validation

The implementation of this project will include biweekly meetings of the core project team and monthly meetings with the entire project team. Project status emails will be sent monthly by the project manager to the entire team and will include the current project status and any roadblocks that are preventing project implementation.

The project status will also be communicated to the CDC during their scheduled monthly phone meetings.

4.2 Acceptance Criteria

The NC-CISS will be considered complete in its first year with the following benchmarks:

1. A documented methodology for linking death certificate data and emergency department visit data with 2018 crash data that can be replicated in future years.
2. The creation of linked datasets of crash and death certificate data and crash and emergency department visit data. The data should be cleaned and ready for analysis.

5 Sustainability

This Implementation Plan is intended to describe the plan for the first year of the NC-CISS, where the focus is on initiating a new process. Subsequent years should require fewer resources related to determining linkage methodology and to documentation, unless there are significant changes in the data, including adding new data sources, or with the process itself. The plan will be revised if such changes occur. All processes will be sufficiently documented so that the risk of staff turnover is mitigated.

In the absence of significant data and process changes, resources in subsequent years will shift to the use of the linked data to provide insight into motor vehicle crash injuries to increase traffic safety in North Carolina.

Two needs will need to be addressed for the NC-CISS to be sustainable: sustained funding and the input and support of stakeholders. The stakeholders should be able to direct research and have the ability to influence policymakers, so that the potential benefits of linking these data are realized. It is recommended that a new advisory group be established for the NC-CISS to

oversee the use of the NC-CISS data for research efforts, make recommendations about expansion of the NC-CISS to include additional data sources, and anticipate data needs to inform policy. It is also expected that this group, along with the existing state, university and community partners and stakeholders, will assist with the identification and attainment of ongoing funding.

APPENDIX A: Death Certificate Data Request

Study background and description

Study Title: North Carolina Crash Injury Surveillance System (NC-CISS)

In 2016, motor vehicle crashes were the second leading cause of fatal injury (N=1,472 deaths) and the second and third leading mechanism of injury for hospitalizations (N= 7,963 hospitalizations) and emergency department visits (N= 67,784 visits) in North Carolina, respectively. While mortality and morbidity data provide valuable insight into the demographics, treatment, and the nature and severity of injuries among persons treated for motor vehicle crash injuries, these data sources provide little information about the actual crash event. On the other-hand, police-reported motor vehicle crash data provide many relevant details about the circumstances of the crash, but little information about the patient's outcome. In order to obtain a more complete picture of the circumstances and outcomes associated with motor vehicle crash injuries in North Carolina, this study will link death registration data with crash data as part of a statewide integrated motor vehicle crash injury surveillance system. We will conduct descriptive and inferential analyses of the linked and unlinked data to expand our understanding of the crash and socio-demographic factors associated with crash fatalities. We will disseminate findings, pending review and approval by State Center for Health Statistics staff, to research, policy, and advocacy communities, as well as the general public.

We are requesting all publicly available upon request death records within the period January 1, 2018 - December 31, 2018. We understand that the 2018 dataset is still being processed and we are willing to wait until its completion before receipt of the data. We are specifically interested in demographics (age, sex, race/Hispanic ethnicity, county of residence, occupation, and education), immediate and underlying causes of death (ICD-10 codes and literal text fields), date/time of injury (if available), location of injury (if available), and date/time of death. We are requesting the text fields, because the use of these fields was suggested by the NC Injury & Violence Prevention Branch for identifying alcohol-involved crash fatalities, as the ICD-10 codes indicating alcohol involvement (Y90-Y91) are frequently underreported.

The research team for this project includes Dr. Anna Waller, ScD, Executive Director and Research Professor at the UNC Carolina Center for Health Informatics and Department of Emergency Medicine; Dr. Katie Harmon, PhD, Postdoctoral Research Associate at the UNC Highway Safety Research Center (HSRC); Dr. Mike Fliss, PhD, MPS, MSW, Research Scientist at the UNC Injury Prevention Research Center and Epidemiologist & Public Health Informatician at the Injury and Violence Prevention Branch in the Division of Public Health,

North Carolina Department of Health and Human Services; Amy Ising, MSIS, Associate Director of the UNC Carolina Center for Health Informatics; Dennis Falls, Database Administrator and Security and Privacy Officer for the UNC Carolina Center for Health Informatics; Clifton Barnett, MSIS, Data Quality Manager and Data Analyst for the UNC Carolina Center for Health Informatics; and Kathy Peticolas, MPS, PMP, Project Manager at the Injury and Violence Prevention Branch in the Division of Public Health, North Carolina Department of Health and Human Services. Dr. Waller will be responsible for guiding the study, developing research questions, and reviewing the analysis plan, findings, and final publications. Drs. Fliss and Harmon will perform the data linkage, data analysis and will develop publications. Ms. Ising will assist with project guidance and preparation of reports. Mr. Falls will assist in the data linkage and insure data security and integrity throughout the process. Mr. Barnett will assist with data analysis, report preparation, and data quality assurance. Ms. Peticolas will manage the project and assist with data preparation, analysis, and writing of results.

Provide a brief description of the public health benefits of this study:

Resulting analyses, publications, and reports will be targeted toward other researchers, policy makers, advocates, and the lay public in order to provide a more data-driven conversation around opportunities to better protect the safety of North Carolina's residents and visitors. Further, findings from the proposed analyses may serve to inform ongoing research and policy relating to road safety in other states throughout the US. Ultimately, it is hoped that the results of this work will lead to ongoing annual data linkage and analysis of crash and death certificate data to inform prevention efforts in NC.

Provide a brief account of security measures, including the conditions under which the records or data will be used, stored, and disposed of and any other security precautions in place to ensure the confidentiality of the data:

The CCHI Security and Privacy Officer, Dennis Falls, is a member of this project team and will have responsibility for insuring the security of these data. We employ multiple levels of information security and redundancies, including: password-protecting the data so that only those individuals directly involved in the project have access; storing all files behind UNC at Chapel Hill's secure firewalls; full-disk encrypting all laptops used by project staff; and deleting all data at the conclusion of the project. All staff involved in the proposed project have up-to-date certifications in social and behavioral research from the CITI Program, the premier research ethics and compliance training organization; these certifications include coursework in privacy, confidentiality, and data protection practices. Lastly, we will only present aggregated results from analyses, and will present all draft publications or other public-facing materials to State Center for Health Statistics staff for review and approval prior to dissemination. We request that the death data be provided to UNC CCHI as a SAS or excel file. It can be burned to a CD-rom or sent via secure FTP site.

APPENDIX B: NC DETECT Data Request Application

Title: North Carolina Crash Injury Surveillance System (NC-CISS)

Data description: All ED visits in NC DETECT between January 1, 2018 and December 31, 2018. List of data elements (reasons for including in request are provided below): Chief complaint, ICD-10-CM diagnosis and injury mechanism codes (all), 5-digit ZIP code of residence, city of residence, county of residence, state of residence, ED disposition, insurance/expected source of payment, mode of transport, patient sex, patient DOB, patient age (years), internal tracking ID, visit ID, visit date, and visit time (1-hour blocks).

Variables:

Chief complaint – The recipients will use this additional text information to identify road user injury-related ED visits that are not captured using ICD-10-CM injury mechanism/diagnosis codes. The recipients will identify these visits using common crash keywords (e.g. MVC, MVA, CAR VS PED, etc.)

ICD-10-CM diagnosis and injury mechanism codes (ALL available codes) - The recipients will use these codes to identify crash-related injuries and to describe the nature and location of these injuries.

5-digit ZIP code – In order to integrate/link NC DETECT ED visit and crash report data, the recipients will require patient ZIP code. ZIP code of residence will be stripped from the analytical data set after data integration.

City of residence – In a previous data integration study using NC DETECT data, Dr. Harmon found that sometimes hospitals default to the ZIP code of the facility rather than the ZIP code of the patient residence; however, she found that city of residence is generally accurate. Therefore, for cases in which ZIP code of residence does not match, city of residence is a good alternative matching variable. City of residence will be stripped from the analysis data set after integration.

Patient county of residence – The recipients will use patient county of residence to identify geographic hot spots of crash injuries. The recipients will not display counts <10 by county.

State of residence -The recipients will need to know if the patient is a NC or out-of-state resident.

ED disposition – Disposition is an important indicator of injury severity.

Insurance/Expected source of payment – This will be another helpful descriptor of NC ED visits related to crash injuries. In addition, self-pay is an important indicator of potential health disparities.

Mode of transport – Mode of transport (e.g. arrive via ambulance) is another indicator of crash injury severity.

Patient sex – Important demographic descriptor of the ED visits among crash patients.

Patient date of birth (DOB) – In order to integrate/link NC DETECT ED visit and crash report data, the recipients will require patient DOB. Patient DOB will be stripped from the analysis data set after integration.

Patient age (in years) – Important demographic descriptor of ED visits. In addition, it is a key integration variable if patient DOB is missing.

Internal tracking ID and visit ID – These variables will be used to identify return visits to the same ED or hospital system by the same patient. Return/follow-up visits is also an indicator of injury severity.

Unique facility tracking ID – The recipients will require this ID to adjust for facility reporting issues, data outages, and known data quality issues.

Visit date and time (1-hour blocks) – In order to integrate/link NC DETECT ED visit and crash report data, we will require Visit Date and Time. In addition to data integration, visit date and time will help us to describe seasonal and time trends. Both factors have been associated with crash frequency and injury severity.

Project Title:

North Carolina Crash Injury Surveillance System (NC-CISS)

DESCRIPTION OF STUDY:

Motor vehicle crashes are a significant source of injury and fatality in North Carolina, but examination of the crash and socio-demographic factors involved with various health outcomes is limited by the lack of connection between crash and health outcome data. Emergency department (ED) visit data provide valuable insight into the demographics, treatment, and the nature and severity of injuries among persons treated for motor vehicle crash injuries, while police-reported motor vehicle crash data provide relevant details about the circumstances of the crash. This study will link NC DETECT ED visit data with crash data as part of a North Carolina Crash Injury Surveillance System (NC-CISS). NC-CISS will be used to obtain a more complete picture of the circumstances and outcomes associated with motor vehicle crash injuries in North Carolina.

The NC-CISS project is a collaboration between the NC Injury and Violence Prevention Branch (NC IVPB), Carolina Center for Health Informatics (CCHI), the UNC Highway Safety Research Center (UNC HSRC), the UNC Injury Prevention Research Center (UNC IPRC), the CDC, and others. It expands on a recent demonstration project, funded by the Governor's Highway Safety Program, which demonstrated the usefulness of NC DETECT ED visit data for integration with motor vehicle crash report data for a study of pedestrian and bicycle crash injuries.

The NC-CISS project will integrate NC DETECT ED visit data with motor vehicle crash report data for *all* road user types. Specifically, we would like to integrate NC DETECT ED visit and UNC HSRC crash data for all crash victims seeking treatment at 24/7 hospital affiliated civilian acute care EDs in NC. The data linkage process may include some case definition evaluation. The integrated data, as part of the NC-CISS, will be used to 1) describe the characteristics of

medically attended injuries among patients involved in crashes and 2) to identify predictors of hospital admission and death among patients involved in crashes. Data linkage will be performed by Drs. Mike Fliss and Katherine Harmon.

PRINCIPAL RESEARCHER AND CO-INVESTIGATORS:

Kathy Peticolas, Project Coordinator/Data Analyst/CCHI Liaison
Injury and Violence Prevention Branch
Division of Public Health, Chronic Disease and Injury Section
NC Department of Health and Human Services
100 Market Street, 1st floor Chapel Hill, NC 27516
Email: kathy_peticolas@med.unc.edu Phone: (919) 448-5314

Co-Requesters:

Clifton Barnett cbarnett@ad.unc.edu (919) 843-2360 UNC CCHI
Will Curran-Groome wcggcw@live.unc.edu (919) 962-2202 UNC HSRC
Dennis Falls dennis_falls@med.unc.edu (919) 843-0815 UNC CCHI
Mike Fliss mike.dolan.fliss@unc.edu (919) 843-6755 UNC IPRC / NC IVPB
Katherine Harmon harmon@hsrc.unc.edu Phone: (919) 962-0745 UNC HSRC
Amy Ising amy_ising@med.unc.edu (919) 843-0814 UNC CCHI
Kendall Knuth Kendall.Knuth@dhhs.nc.gov NC IVPB
Anna Waller anna_waller@med.unc.edu (919) 843-0389 CCHI
DPH contact:
Alan Dellapenna alan.dellapenna@dhhs.nc.gov 919-707-5441 NC IVPB

PUBLIC HEALTH SIGNIFICANCE:

In 2016, motor vehicle crashes were the second leading cause of fatal injury (N=1,472 deaths) and the second and third leading mechanism of injury for hospitalizations (N= 7,963 hospitalizations) and emergency department visits (N= 67,784 visits) in North Carolina, respectively. Emergency department visit data provide valuable insight into the demographics, treatment, and the nature and severity of injuries among persons treated for motor vehicle crash injuries, while police-reported motor vehicle crash data provide relevant details about the circumstances of the crash. This study will link NC DETECT emergency department visit data with crash data as part of a North Carolina Crash Injury Surveillance System (NC-CISS). NC-CISS will be used to obtain a more complete picture of the circumstances and outcomes associated with motor vehicle crash injuries in North Carolina. Ultimately, it is hoped that this work will lead to ongoing annual data linkage and analysis of crash and emergency department visit data to inform prevention efforts in NC.

RESULTS OF PROJECT WILL BE USED FOR:

Contract requirements and presentation. Contract number: UNC/DHHS contract 00039605.

Project aims include: 1. Integrate NC DETECT ED visits and crash data for all road users as part of a new North Carolina Crash Injury Surveillance System (NC-CISS). 2. Describe the characteristics of those persons injured in MVCs. 3. Identify predictors of hospital admission among persons injured in MVCs. Through integrating NC DETECT ED visit and crash data, the recipients will perform a comprehensive analysis of factors associated with crash morbidity and mortality. This information is needed to inform system-level interventions aimed at reducing crash injuries and fatalities. In addition to fulfilling our contract requirements, the recipients will provide regular updates to our partners at CCHI and NC DPH about our progress.

TIME PERIOD FOR DATA REQUESTED:

January 1, 2018 – December 31, 2018.

ESTIMATED STUDY COMPLETION DATE:

12/31/2020

APPENDIX C: Bibliography

1. Milani J, Kindelberger J, Bergen G, Novicki EJ, Burch C, Ho SM, et al. Assessment of characteristics of state data linkage systems. Washington, DC and Atlanta, GA: National Highway Traffic Safety Administration and Centers for Disease Control and Prevention; 2015.

Appendix I: Mid-Year Report

Brief Project Description
The North Carolina Crash Injury Surveillance System (NC-CISS) project is a one-year CDC-funded project to create a sustainable methodology for linking crash report data with two health data sources: emergency department (ED) data from NC DETECT and death certificate data.

Project Status Summary
The project has proceeded according to the proposed timeline, other than some delay due to the time needed to assemble the core team and begin the process to obtain the data. We currently expect to complete the project on time.

Completed Project Milestones
The proposal for CDC funding included four high level activities. The following lists the activities and their status mid-year.

Activity	**Status**
1. Develop a description of datasets used for data linkage.	Complete
2. Develop data linkage methodology.	Complete
3. Identify barriers and facilitators.	In progress
4. Summarize data linkage.	In progress

The following lists the completed project tasks for each activity.

Activity 1: Develop a description of datasets used for data linkage.

Task	**Description**	**Planned**	**Completed**
Task 1. Confirm two data sources to be linked to NC motor vehicle crash data (NC DETECT emergency department visit and SCHS death certificate/vital records data)	a. Receive written approval from the data owner (Communicable Disease Branch at NC DPH) to proceed with linking crash report data to NC DETECT emergency department visit data b. Receive written approval from the data owner (NC SCHS) to proceed with linking crash report data to death certificate/vital records data c. Inform state and federal partners of approval decisions	8/15/2019	10/10/2019
Task 2. Complete documentation for data source # 1 (NC DETECT emergency department visit data)	a. Thoroughly document data source # 1 (NC DETECT emergency department) according to CDC guidelines	8/30/2019	9/25/2019

Task	Description	Planned	Completed
Task 3. Complete documentation for data source # 2 (SCHS death certificate/vital records data)	a. Thoroughly document data source # 2 (SCHS death certificate/vital records data) per CDC guidelines	9/15/2019	9/25/2019
Task 4. Submit data documentation to CDC	a. Review and edit drafts of data documentation b. Submit data documentation to CDC	9/30/2019	10/22/2019

Activity 2. Develop data linkage methodology.

Task	Description	Planned	Completed
Task 1. Obtain approval from UNC Institutional Review Board (IRB)	a. Submit IRB to UNC IRB b. Receive approval from UNC IRB	8/31/2019	10/29/2019
Task 2. Obtain at least one full year of motor vehicle crash report data	a. Request and receive 1+ years of motor vehicle crash report data from HSRC	9/30/2019	11/1/2019
Task 3. Obtain approval from the Communicable Disease Branch of NC DPH for 1+ year of NC DETECT emergency department visit data	a. Obtain approval from Communicable Disease Branch (may or may not involve formal DUA) b. Provide approval documentation in Implementation Plan	10/31/2019	12/9/2019
Task 3. Obtain approval from NC SCHS for 1+ year of death certificate/vital records data	a. Obtain approval form NC SCHS (may or may not involve formal DUA) b. Provide approval documentation in Implementation Plan	11/30/2019	10/17/2019
Task 4. Complete data implementation plan	a. Review literature related to data linkage b. Establish communication with one or more data linkage experts; invite linkage expert to travel to NC to consult with Project Team members c. Develop methodology to link crash and health outcome data based on a) literature review, b) consultation with experts, and c) Project Team members' expertise d. Draft Implementation Plan, share with partners, and solicit feedback; revise, as needed	3/31/2019	3/3/2020

	e. Submit Implementation Plan to CDC		

Activity 4. Summarize data linkage.

Task	Description	Planned	Completed
Task 1. Obtain health outcome data sources	a. Obtain NC DETECT emergency department visit data b. Obtain NC SCHS death certificate/vital records data	1/1/2020	12/16/2019

Other Project Activities or Related Data Linkage Work

Consultation with data linkage experts

- The core project team met on December 4th with Peter Leese, a Senior Programmer and Data Scientist at UNC Healthcare who has designed ongoing data linkage with their EHR and death certificate data and who shared his experience with sustainability.
- A key part of our proposal was having data linkage expert Dr. Larry Cook assist us in person. This took place February 10-14. Discussion and input from Dr. Cook was very influential on our planned methodology. While the visit focused on Dr. Cook's direct assistance with different aspects of this project, the visit included presentations from Larry on data linkage and meetings with other researchers doing data linkage projects.

Project meetings

Ten core team meetings were held on the following dates:

- 9/16/2019
- 9/24/2019
- 10/8/2019
- 10/25/2019
- 11/6/2019
- 12/4/2019
- 1/6/2020
- 1/29/2020
- 2/3/2020
- 2/26/2020

Four full project team meetings were held on the following dates:

- 10/28/2019
- 11/15/2019
- 12/17/2019
- 2/10/2019

Five CDC calls with representatives from the four funded states were attended on the following dates:

- 10/24/2019
- 11/18/2019
- 12/19/2019
- 1/23/2020
- 2/27/2020

Data linkage-related presentations

- Kathy Peticolas presented the poster "Beyond KABCO: Improving Our Understanding of Pedestrian and Bicyclist Injuries with Linked Crash and Hospital Data" the December 11-12, 2019 Conference on Health and Active Transportation in Washington, D.C.

- Katie Harmon presented on the "Benefits of Data Integration in Safety Decision Making" at the January 12-15, 2020 TRB Annual Conference in Washington, D.C.

Remaining Project Milestones

Activity 3: Identify barriers and facilitators.

Task	Description	Planned date
Task 1. Document barriers and facilitators to data linkage	a. Document any barriers and facilitators related to planning, implementation, and linkage of crash report/ health outcome data sources for project duration b. Record all barriers/facilitators in a summary report c. Submit report to CDC	4/30/2019

Activity 4: Summarize data linkage.

Task	Description	Planned date
Task 2. Link two health outcome data sources	a. Link NC DETECT emergency department visit data with crash data b. Link NC SCHS death certificate/vital records data c. Incorporate methodology developed as part of Implementation Plan (Activity 3) d. Document all deviations from Implementation Plan and explanation	4/30/2020
Task 3. Summarize data linkage activities	a. Prepare a report summarizing data linkage activities; include all required CDC indicators b. Prepare a handout succinctly summarizing data linkage activities c. Share report/handout with state and CDC partners	7/31/2020

Appendix J: Data Linkage and Barriers and Facilitators report

Overview

This report documents our initial data linkage efforts and reviews the barriers and facilitators we have encountered as part of the North Carolina Crash Injury Surveillance System (NC-CISS) project, which will link 2018 crash report data with two health data sources: death certificate data and emergency department (ED) data.

Data Linkage Methodology Evaluation

A key component of this project is testing different data linkage methodologies to evaluate their sustainability. Some key factors in evaluating the sustainability of a methodology include:

- Flexibility
- Ability to communicate process to audiences with diverse professional backgrounds
- Dependability and sustainability of linkage software
- Programming expertise required
- Data processing resources required

Four linkage methodologies have been evaluated, with the third deemed the most promising. We will initially focus on the crash report and death certificate linkage before moving on to the crash report and ED linkage.

The next phase of evaluation will compare the numbers and percentages of linked records, the accuracy of the linkage via review of a sample of linked pairs, and the representativeness of the linked records as compared to the crash records and linked datasets for key demographics and metrics. The four methodologies are described below.

Methodology 1: Recursive Partitioning Trees (RPT) in R

Description

A model was generated in R, which reviewed the match rates of different linkage variables. It calculated the degree of difference between matching variables. A score of zero indicated an exact match in that variable, which ranged unscaled from 0 to infinity. The higher the number, the higher the difference between the variables being compared. A total for each linkage indicated the total degree of difference across several variables. This total distance was treated alongside the variable-specific scores as another measure of linked pair quality.

Review

While this methodology provided a potentially powerful approach to data linkage, its sustainability was questionable. The process required considerable programming expertise and data processing needs. It was also highly complex and not necessarily understandable to a wider audience. See Appendix A: Notes on Preliminary RPT Linkage for a description of this method of linking crash report and death certificate data.

Methodology 2: Probabilistic Linkage in R

Description

We reviewed two probabilistic linkage packages built for R: RecordLinkage and fastLink. RecordLinkage is an older package associated with numerous peer reviewed articles (e.g. Sariyar and Borg, 2010), implementing a stochastic framework calculating weights through Expectation-Maximization algorithms and extreme value theory, including a number of machine learning methods. The fastLink package (described in Enamorado, Fifield, and Imai, 2019) also implements probabilistic linkage and supports parallel processing for larger datasets on one machine.

Review

Though there is good documentation for the packages, they require a high level of R skill. Moreover there seemed to be no ability to do fuzzy matching of dates or incorporate spatial distance. The fastLink

package, though newer (developed in 2016), had other challenges: it had not been significantly updated since 2018, it seemed particularly slow on larger datasets (e.g. ED records), and likewise seemed only to easily allow string distance matching. Though fastLink had some address-specific linking functions, it did not easily include fuzzy spatial proximity in its matching algorithm.

The R probabilistic linkage packages required strict blocking and needed more data processing resources than other methods. They also required more programming expertise and more expertise in probabilistic mathematical models. Probabilistic linkage is not as easily communicated to a wider audience. Because of these challenges, the R probabilistic linkage packages will not be not pursued further as a linkage methodology for this project.

Methodology 3: Hierarchical Deterministic Linkage with Block Filtering in R

Description

Hierarchical deterministic linkage matches data according to a list of linkage variables in a stepwise fashion, moving from most strict to least strict linkage. For a match to occur, the two data sources must have the same exact values or be within a specified distance for the linkage variables. However, if a match does not occur during the first linkage step, certain linkage criteria are relaxed (while other linkage criteria are tightened) and a second round of matching commences. If matching does not occur during the second round of matching, more rounds of matching can be performed. In each step, linkage variables are either (1) exactly matched, (2) required to be within a given distance (e.g. age distance or distance in miles), or (3) ignored entirely (allowing for missingness).

Because of the reliance on exact matching, the cleaning of the data is key. Some fuzzy cleaning will be used, particularly to account for misspellings in some variables (e.g. City of Residence). Including a filtering step after each linkage step allows more flexibility than deterministic linkage alone.

Missingness patterns, estimates of data quality, and results from other linkage techniques will inform the steps of the final linkage methodology.

Review

This methodology has many positives. We have had success using hierarchical deterministic linkage in prior projects. It is simpler and requires less programming time and expertise. It is more easily explained to others. The data processing needs are low. While this methodology is being developed in R, it can be translated to other software such as SAS. This methodology will continue to be pursued and evaluated for accuracy. See Appendix B: Notes on Preliminary Hierarchical Deterministic Linkage for a description of this process of linking crash report and death certificate data.

Methodology 4: Probabilistic Linkage Using LinkSolv

Description

LinkSolv is a probabilistic linkage software using Microsoft Access. It was created by Dr. Michael McGincy through his company Strategic Matching, Inc. The UNC Department of Emergency Medicine already owned a license, prior to the NC-CISS project. It has a yearly license and technical support fee of $1,680.

Review

LinkSolv has some positive qualities. With in-person instruction, the point and click user interface is mostly straightforward to use and customizable per project. The process is already developed, so no

programming or process development time is required other than preparing the data before and after the linkage. Also, the probabilistic linkage method has been promoted by the CDC as mathematically sound and has been used in published research.

However, the LinkSolv user interface is not very intuitive without direct instruction by an experienced user. The mechanism by which it works is not entirely transparent. Also, probabilistic linking is not as easily explained to audiences. Finally, Strategic Matching is a one-person run business, so future support and maintenance of the software is somewhat precarious.

LinkSolv will continue to be tested and compared with the hierarchical deterministic methodology, although the concerns with the software will be considered in the final decision on methodology. See Appendix C: Notes on Preliminary LinkSolv Linkage for a description of a preliminary linkage of crash report and death certification data using LinkSolv.

Project Facilitators

5. **A coalition of program leadership committed to data linkage**
   The support of program leadership from the following organizations has been critical to moving this project forward:
   - the NC DHHS Injury and Violence Prevention Branch (DHHS IVPB),
   - the UNC Carolina Center for Health Informatics (CCHI),
   - the UNC Highway Safety Research Center (HSRC),
   - the UNC Injury Prevention Research Center (IPRC),
   - the North Carolina Traffic Records Coordinating Committee (TRCC),
   - the NC DHHS Communicable Disease Branch, and
   - the State Center for Health Statistics (SCHS).

6. **Experience with and funding for previous data linkage projects**
   Prior data linkage projects provided critical experience in data linkage methodology. These projects also established relationships with the data owners, as well as providing experience working with each dataset. Most of the prior projects were part of the Linking Motor Vehicle Crash Data to Health Outcome Data in North Carolina project, funded from 2016 to 2020 by the NC Governor's Highway Safety Program (GHSP).

7. **Staffing with personnel who have the appropriate skills**
   The core project team included individuals with high levels of expertise in project management, epidemiology, statistics, knowledge of transportation safety, the data sources being linked, and presentation and marketing skills.

8. **The input of data linkage practitioners**
   This project included funding for a one week consultancy with data linkage expert Dr. Larry Cook from the University of Utah School of Medicine. Dr. Cook spent a week in Chapel Hill working closely with the project team and also meeting with others, including program leaders and staff of related data linkage projects. Dr. Cook provided insight into his use of probabilistic linkage methodology, feedback on our progress, and direct instruction on the use of LinkSolv.

Project Barriers

4. **Lack of common unique identifiers across data sources**
   The lack of common unique personal identifiers on the crash report and either of the health data sources necessitates a more complex data linkage methodology. It also makes verification of the accuracy of the linkage results more challenging.

5. **Data quality and completeness**
   Missing data and data entry errors impact our ability to link the datasets. These issues are present for both crash report and health data.

6. **Lack of program ownership and stable funding**
   The lack of an identified long-term program owner with stable funding has made planning more challenging. Planning for sustainability is made more difficult when the resources that will be available to implement the project into the future are unknown or unpredictable.

References

1. Sariyar M, Borg A. The RecordLinkage Package: Detecting Errors in Data. R journal. 2010;2(2). https://journal.r-project.org/archive/2010-2/RJournal_2010-2_Sariyar+Borg.pdf. Accessed July 9, 2017.

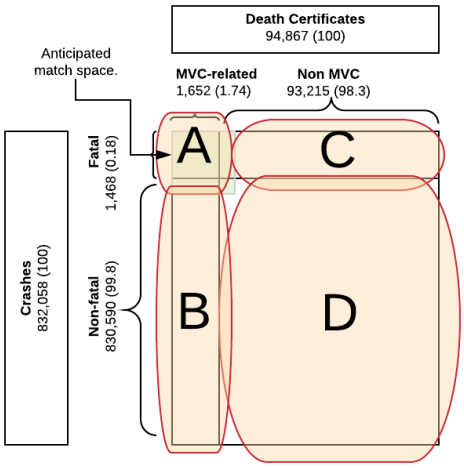
2. Enamorado T, Fifield B, Imai K. Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. Am Polit Sci Rev. 2019;113(2):353-371. doi:10.1017/S0003055418000783

3. Milani J, Kindelberger J, Bergen G, Novicki EJ, Burch C, Ho SM, et al. Assessment of characteristics of state data linkage systems. Washington, DC and Atlanta, GA: National Highway Traffic Safety Administration and Centers for Disease Control and Prevention; 2015.

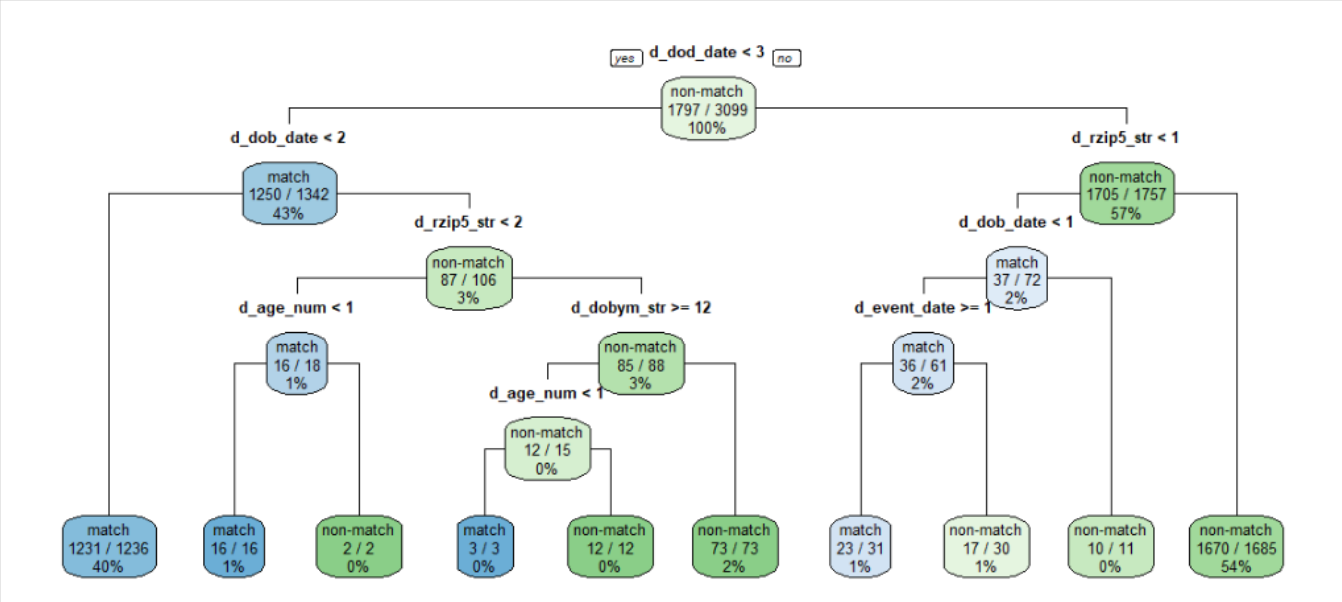
Appendix A: Notes on Preliminary RPT Linkage

We began the crash report and death certificate linkage by identifying the anticipated matches based on the numbers of fatally injured persons in the crash data and the numbers of deaths with motor vehicle crash (MVC) coding in the death certificate data. These matches, illustrated at right, informed our initial pair-wise four-part blocking scheme, listed below in descending order of expected matches and strictness of blocking:

1. MVC-related deaths and fatal crashes
2. Unmatched MVC-related deaths
3. Unmatched fatal crashes
4. All other unmatched



Below is an example of a simplified recursive partition tree that uses crash report and death certificate data. One model successfully match 99% of potential matches compared with a hand-reviewed set of 1,400 matches.



Future improvements will be made by reviewing match sets and re-running the RPT model.

Appendix B: Notes on Preliminary Hierarchical Deterministic Linkage

The hierarchical deterministic linkage in R uses a command table such as the one below, where each row represents one matching pass. Each column beginning with "l_" is a linkage variable shared between the datasets. A zero in the case definition's cell indicates it is required for the match. An "NA" means that variable is not used in that matching pass. A non-zero numeric entry allows for a type-specific distance (e.g. a person's age can be within 3 years). Records matched in a higher case definition row are considered linked and are not included in subsequent matching passes.

link_name	link_description	l_age_num	l_dob_date	l_dobmd_fct	l_gender_fct	l_raceeth_fct	l_fatal_lgl	l_isinj_lgl	l_issevere_lgl	l_isMVC_lgl	l_crashpos_fct	l_rzip5_fct	l_rzip3_fct	l_rzip_x	l_rzip_y	l_rcounty_fct	l_rcounty_x	l_rcounty_y	l_rcity_fct	l_rcity_x	l_rcity_y	l_acc_date	l_accmd_fct
all exact	exact	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no city	place: no city	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	NA	NA	0	0	
no zip	place: no zip	0	0	0	0	0	0	0	0	0	NA	NA	NA	NA	0	0	0	0	0	0	0	0	
no county	place: no county	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	NA	NA	0	0	0	0	0	
no geo	place: no county	0	0	0	0	0	0	0	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	
No DOB	person	0	NA	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
No DOB, city	person, place	0	NA	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	NA	NA	0	0	
missing_DOB	person	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
missing_DOB	person	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
missing_race	time	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
missing_DOB	time	0	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	
missing_race	person	0	0	0	0	0	0	0	NA	0	0	NA	NA	50	50	0	50	50	NA	50	50	0	0
missing_age	person	0	0	0	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0
fuzzy_date	no crash pos	0	0	0	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	14	0
fuzzier date	all three fuzzy	0	NA	NA	0	NA	0	0	0	0	NA	0	NA	NA	NA	0	0	0	0	50	50	14	0
fuzzier date	all three fuzzy	0	0	0	NA	NA	0	0	0	0	NA	0	NA	NA	NA	0	0	0	0	50	50	14	0

The linkage results include which pass each pair were linked on. The table below describes the number of linkages per pass for an initial crash report and death linkage.

link_name	n
all exact	660
no county	104
No DOB	8
no race	11
No DOB, city	4
no geo	271
fuzzier date	8

Because distance is only calculated for the filter-matched variables in the definitions, and the rest are exact joins, even a many step process can be completed in minutes. Because this model runs very quickly and is highly flexible, improvements can be made by exploring match types from other models and adding the necessary match definitions to the list.

Appendix C: Notes on Preliminary LinkSolv Linkage

As with the other methodologies, the first steps are to clean the data and determine the inclusion criteria for each dataset. After this is done, the data are imported into LinkSolv. LinkSolv analyzes the frequency of values in each variable. The user then indicates the expected error probability for each variable. The user creates blocking passes, which reduces the number of possible matches by requiring exact matches for each variable in the pass. It then compares all the linkage variables for each blocked dataset.

LinkSolv assigns or subtracts points per matching variable. Matching variables that are rarer are assigned higher points than more common matches (e.g. a small town match would have higher points than a large city match). LinkSolv combines the results of each pass into a table of matches with both the total points assigned to that match and the probability that the match is correct. Some data cleaning and removal of duplicates are still required after the match is complete. The use of LinkSolv to run an initial linkage of crash report and death certificate data is described below.

Data Cleaning

Linkage variables	*Data cleaning: crash data*	*Data cleaning: death certificate data*
Event date	None	None
Zip code of residence	First 5 characters only if length >5; drop values with fewer than 5 characters; drop if there are any non-numeric values	First 5 characters only if length >5; drop values with fewer than 5 characters; drop if there are any non-numeric values
Date of birth	Removed future DOBs & DOBs for ages > 110; Removed crash DOBS if 1/1/2000 or 1/1/2001	None
Age	Removed if>110 or <0	None
Sex	M, F or null	M, F or null
Race/ ethnicity	None	Combined some categories to match crash data
City of residence	None	Used CITYRESTEXT when coded rcity was missing
County of occurrence	None	Used coded county of death (cod) field
Person type	Used combination of vehicle and person type to indicate category (Bicyclist, Motor vehicle occupant, Motorcyclist, or Pedestrian)	Used CDC criteria to indicate category (Bicyclist, Motor vehicle occupant, Motorcyclist, or Pedestrian)

Inclusion Criteria

	Crash report data	*Health data*
Inclusion Criteria	Only persons designated as having some injury (K, A, B or C Injuries)	Meets CDC MVC cause of death in any cause of death variable (ACMECOD, COD1-COD20) Occupant V30-V79[.4-.9],V83-V86[.0-.3] Motorcyclist V20-V28[.3-.9],V29[.4-.9] Pedal cyclist V12-V14[.3-.9],V19[.4-.6] Pedestrian V02-V04[.1,.9],V09.2 Other V80[.3-.5],V81.1,V82.1

		Unspecified V87[.0-.8],V89.2 Suicide X82, Homicide Y03 Added V030, V040, V090, V099, V199, V890, V899 OR Any Cause of Death injury (first character in ('S','T','V','W','X','Y'))
Total records	**130,109**	**10,200**

Blocking Passes in LinkSolv

Number	*Blocking Pass*
1	1. Date of birth 2. 5-digit zip
2	1. Accident date/date of death 2. Age
3	1. Accident date/date of death 2. 5-digit zip

LinkSolv Linkage Variables (all passes)

Linkage variables	*Type of linkage*	*Error probability*
Event date	Up to 14 day after crash	0.01
Zip code of residence	Exact	0.10
3-digit zip code of residence	Exact	.10
Date of birth	Exact	0.01
Date of birth month and day	Exact	0.01
Age	-1, +1	0.02
Sex	Exact	0.01
Race/ ethnicity	Exact	0.15
City of residence	Exact	0.1
County of occurrence	Exact	0.05
Person type	Exact	0.10

Final criteria

Linked records	2,254
Date of death must be on or after the crash date	-55
Duplicate removal	-275
Removed age discrepancies (-2, +2)	-8
<90% probability of linkage	-548
Remaining	**1,368**

Appendix K: Receipt notices of NC DETECT and SCHS death certificate data

From: Barnett, Clifton A <cbarnett@ad.unc.edu>
Sent: Monday, December 16, 2019 12:14 PM
To: Peticolas, Katherine Alice <kathy_peticolas@med.unc.edu>
Subject: (Secure) NCDETECT Dataset

Hi Kathy,

Your file is available for download at:

[Removed]

The password is: [Removed]

Let me know if you have any questions.

Clifton

From: Avery, Matt <matt.avery@dhhs.nc.gov>
Sent: Friday, December 6, 2019 11:24 AM
To: Peticolas, Katherine Alice <kathy_peticolas@med.unc.edu>
Subject: RE: [External] New death data request - NC-CISS

Hi Kathy,

Sure, revised dataset is attached w/names. Let me know if you have any questions.

Have a good weekend!

Matt

Matt Avery, M.A.
Supervisor, Vital Statistics
Division of Public Health, State Center for Health Statistics
North Carolina Department of Health and Human Services

919 715 4572 office
919 733 8485 fax
Matt.Avery@dhhs.nc.gov

222 North Dawson Street
Raleigh, NC 27603

1908 Mail Service Center
Raleigh, NC 27699-1900

Appendix L: Summary Handout