SUPPLEMENTARY APPENDIX

ASSEMBLY OF THE LONGSHOT COHORT: PUBLIC RECORD LINKAGE ON A GRAND SCALE

Yifan Zhang, PhD
Erin E. Holsinger, MD
Lea Prince, PhD
Jonathan A. Rodden, PhD
Sonja A. Swanson, ScD
Matthew J. Miller, MD, ScD
Garen J. Wintemute, MD, MPH
David M. Studdert, LLB, ScD

I. Candidate Sources of Longitudinal Information

Before deciding to use the Statewide Voter Registration Database (SVRD) as the spine of the LongSHOT cohort, we investigated the suitability and accessibility of several other sources of longitudinal data on large numbers of adult residents of California. We considered driver license data from the California Department of Motor Vehicles but ultimately rejected it for two main reasons: snapshots of the dataset are not routinely archived and, despite legal requirements to do so within 10 days,[1] licensees frequently do not update their addresses when they move. (Recent reforms linking California's driver license database to the SVRD could ameliorate the latter problem.[2]) We also considered tax return data, but requests to the Internal Revenue Service and the California Franchise Tax Board were denied on the grounds that state and federal law precluded access to identifiable tax data for a study of this kind.

II. Extracts of the California Statewide Voter Registration Database

We obtained the voter file extracts from the Statewide Database, an organization that collects and analyzes official voter statistics in California.[3] The Statewide Database archives "15-day' reports of registration—so called because they are extracts of the SVRD taken 15 days prior to statewide elections in California. The extract dates in our series correspond to 6 general elections (11/2/04, 11/7/06, 11/4/08, 11/2/2010, 11/6/2012, 11/4/2014), 6 primary elections (2/5/2008, 5/3/2008, 5/8/2010, 5/5/2012, 5/3/2014, 5/7/2016), and 1 special election (11/8/2005).[4] The only statewide primary election during the study period not covered by the extracts in our series was the election on June 6 2006; we obtained it but did not use it in the study because of corruption issues with voter file numbers.

Table S1. Dates of extraction for 13 extracts of the the California Statewide Voter Registration Database

Extract no.	Date of extraction
1	October 18 2004
2	October 24 2005
3	October 23 2006
4	January 22 2008
5	May 19 2008
6	October 20 2008
7	May 24 2010
8	October 18 2010
9	May 21 2012
10	October 22 2012
11	May 19 2014
12	October 20 2014
13	May 23 2016

III. Interval-based Approach to Linkage

As noted in the manuscript, our linkage proceeded in discrete interval-based links that were sensitive to the time-varying nature of our principal datasets. In the first interval link, for example, persons who acquired handguns or died between 10/18/2004, the date of the earliest SVRD extract in our collection (and the first day of our study period), and 10/23/2005, the

Zhang Y, *et al. Inj Prev* 2020; 26:153–158. doi: 10.1136/injuryprev-2019-043385

day before the date of the second SVRD extract, were eligible to match to registrants named in the 10/18/2004 extract. This process repeated through successive inter-extract intervals. In the thirteenth and final interval link, persons who acquired handguns or died between 5/23/2016, the date of the latest SVRD extract, and 12/31/2016, the study end date, were eligible to match to registrants named in the 5/23/2016 extract.

In addition, to account for lags or errors in SVRD updating, before we commenced linkage, we searched for matches between persons who died in the two years leading up to the start of our study period and registrants named in the first SVRD extract, and removed any that were so identified.

IV. Development of the algorithms

We developed the algorithms iteratively, using training and validation datasets. The training datasets consisted of full SVRD extracts drawn from the middle of the study period. We applied sequences of draft algorithms, and continued to modify and refine them until manual review confirmed that the groups they produced met the following criteria: auto rule-in pairs had no or very few detectable non-matches; auto rule-out pairs had no or very few detectable matches; and manual check pairs consisted of a liberal mix of matches and non-matches, whose projected total across all interval links was not so large as to make them infeasible to manually review. We generally considered a manual check bin with more than 5,000 pairs to be infeasibly large, because across 13 interval links the algorithm would be expected to produce approximately 65,000 pairs for manual review.

Once the blocking keys and sorting algorithms were finalized in the training dataset we ran them on the validation dataset, checked that the resultant pairs had the desired mix of matches, non-matches, and bin sizes, and made minor refinements as necessary. The validation process sought to reduce the risk of overfitting the algorithms to the data, which may occur if, for example, there are substantial changes over time in the structure or completeness of the component datasets.

V. Linkage Steps and Algorithms

The algorithms were deployed in four consecutive steps (A-D). In this section, we describe the algorithms in a series of charts, organized by step. We begin with a guide to the terminology used in the charts.

Zhang Y, *et al. Inj Prev* 2020; 26:153–158. doi: 10.1136/injuryprev-2019-043385

Table S2. Glossary of terms and abbreviations used in descriptions of linkage algorithms

Abbreviation	Definition
/	or
+	and
address	Principal residential address in string text form (address1 refers to the address in record #1 of a candidate pair and address2 refers to the address in record #2 of the pair.)
DOB	Date of birth (DOB1 refers to date of birth in record #1 of a candidate pair and DOB2 refers to date of birth in record #2 of the pair.)
dp	Decimal places used to specify the precision of geocode matching. For e.g., geocode1=geocode2 @4dp means the geocode in record #1 must match the geocode in record #2 at the 4th decimal place of both geocodes.
FN	First name. (FN1 refers to first name in record #1 of a candidate pair and FN2 refers to first name in record #2 of the pair.)
geocode	Geocode of subject's residential address in the 1st record of a candidate pair
geodistance	Straight-line distance between two geocodes, expressed in miles. So for e.g., geodistance<0.02 means the geocoded residential address in record #1 is less than 0.02 miles from the geocoded residential address in record #2.
high MN threshold	Matching criteria for middle name, as defined in table following Step B
jw	Jaro-Winkler distance between values of 2 candidate pairs. For e.g., FN1=FN2 @jw≥0.8 means the Jaro-Winkler distance between the first name in record #1 and the first name in reord #2 is 0.8 or greater.
LN	Last name, or surname. (LN1 refers to last name in record #1 of a candidate pair and LN2 refers to last name in record #2 of the pair.)
low MN threshold	Matching criteria for middle name, as defined in table following Step B
MC	manual check (aka, manual review)
MI	Middle initial, which may be the only value in the middle-name field or the first letter of a more complete middle-name value. (MI1 refers to middle initial in record #1 of a candidate pair and MI2 refers to middle initial in record #2 of the pair.)
mid MN threshold	Matching criteria for middle name, as defined in table following Step B
MN	Middle name. (MN1 refers to middle name in record #1 of a candidate pair and MN2 refers to middle name in record #2 of the pair.)
NN	Nickname. A match between the first name of record #1 and a recognized nickname associated with that first name in record #2 is specified as FN1=NN2 (see section VII below).
RN screen	Rare name screen (see Section IX below)
zip	Zip code of the residential address

Zhang Y, *et al. Inj Prev* 2020; 26:153–158. doi: 10.1136/injuryprev-2019-043385

STEP A

Either record has no MN / MI

Blocking Run: geocode1=geocode2 @2dp + DOB1=DOB2

Both records have MN / MI

Primary auto rule-ins

same FN + same LN → auto rule-in

- same FN + same MN/MI + same LN
- same FN + diff MN/MI + same LN

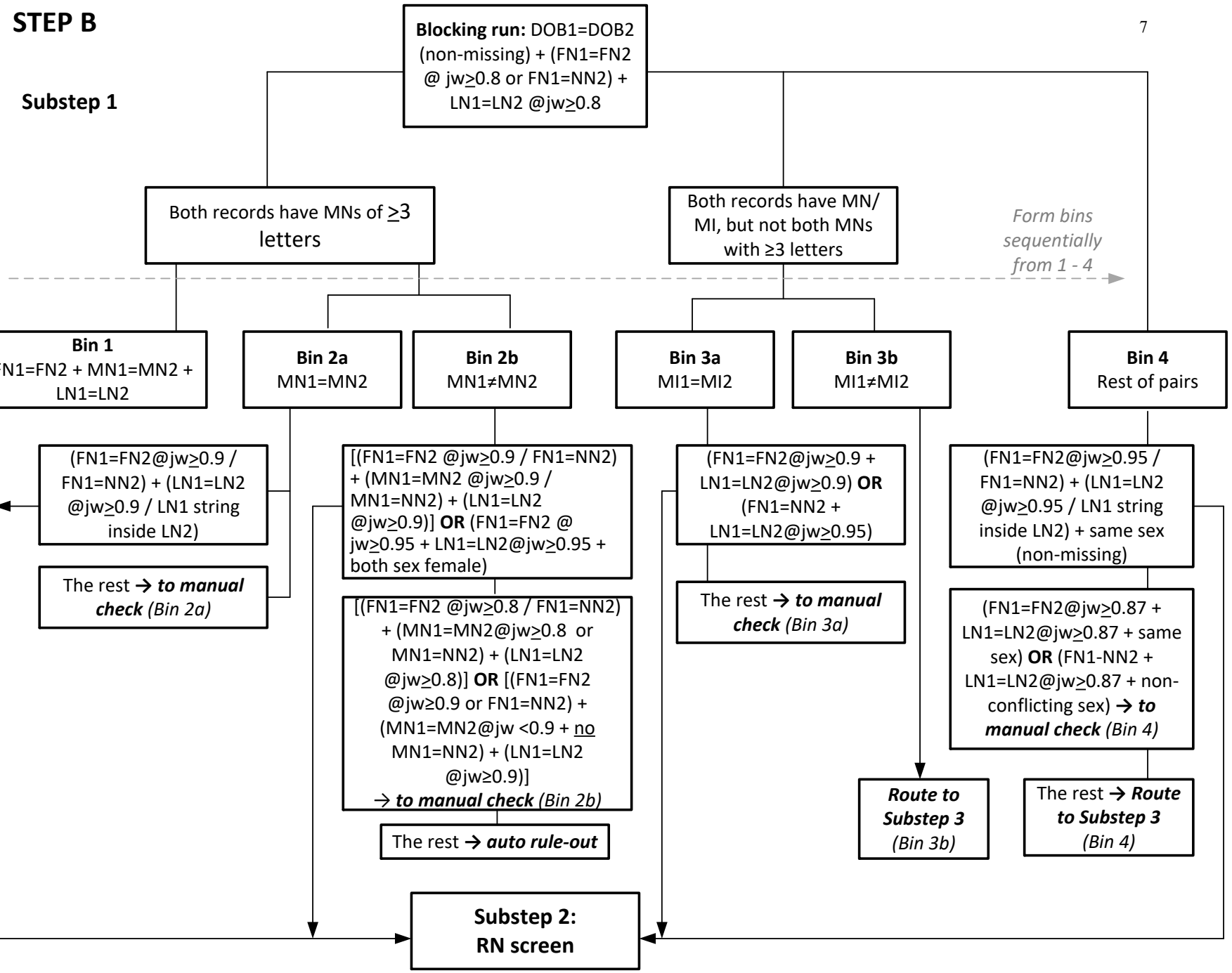
→ auto rule-in

Name bins

Bin	FN	MN/MI	LN		
1	X	--	X	Rule-in:	If any of following conditions met: • FN1=FN2 @jw≥0.8 + LN1=LN2 @jw≥0.8 • FN1=NN2 + LN1=LN2 @jw≥0.8 • FN1=LN2 @jw≥0.8 + FN2=LN1 @jw≥0.8
				Rule-out:	If geodistance≠0 + FN1=FN2 @jw<0.7 + FN1≠NN2 + LN1=LN2 @jw<0.7
				Manual check:	The rest *(notes: sort by geodistance, then by descending jw FN, then by descending jw LN)*
2	X	--	✓	Rule-in:	If any of following conditions met: • FN1=FN2 @jw≥0.8 • FN1=NN2 • FN1=MN2 @jw≥0.8 + FN2=MN1 @jw≥0.8
				Manual check:	The rest (notes: sort by geodistance, then by descending jw FN)
3	✓	--	X	Rule-in:	If sex=female for both records If sex=male/missing for either record + any of following conditions met: • LN1 string inside LN2 / LN2 string inside LN1 • LN1=LN2 @jw≥0.8 • MN1=LN2 @jw≥0.8 + MN2=LN1 @jw≥0.8
				Manual check:	The rest *(notes: sort by geodistance, then by descending jw LN)*
4	✓	✓	X	Rule-in:	If sex=female for both records If sex=male/missing for either record + any of following conditions met: • LN1=LN2 @jw≥0.8 • LN1 string inside LN2 / LN2 string inside LN1
				Manual check:	The rest *(notes: sort by geodistance, then by descending jw LN)*
5	✓	X	X	Rule-in:	If any of following conditions met: • LN1=LN2 @jw≥0.8 • LN1=MN2 @jw≥0.8 + LN2=MN1 @jw≥0.8 If geodistance=0 + any of following conditions met: • LN1=MN2 / LN2=MN1 • LN1 string inside LN2 / LN2 string inside LN1
				Manual check:	The rest *(notes: sort by geodistance, then by descending jw LN)*
6	X	X	✓	Rule-in:	If geodistance=0 + FN1=MN2 / FN2=MN1 If geodistance <0.02 + sex concordant + any of following conditions met: • FN1=NN2 • FN1=FN2 @jw≥0.8 • FN1=MN2 @jw≥0.8 + FN2=MN1 @jw≥0.8
				Manual check:	The rest *(notes: sort by geodistance, then by descending jw FN)*

STEP A - *continued*

Bin	FN	MN/MI	LN		
7	X	✓	✓	Rule-in:	If geodistance <0.02 + any of following conditions met: • FN1=NN2 • FN1=FN2 @jw≥0.8 • FN1 is 1 letter & = 1st letter of FN2 / FN2 is 1 letter & = 1st letter of FN1
				Manual check:	The rest *(notes: sort by geodistance, then by descending jw FN)*
8	X	X	X	Rule-out:	geodistance≠0 + FN1=FN2 @jw<0.7 + FN1≠NN2 + LN1=LN2 @jw<0.7
				Manual check:	The rest *(notes: sort by geodistance, then by descending jw FN, then by descending jw LN)*
9	X	✓	X	Rule-in:	If geodistance <0.02 + any of following conditions met: • FN1=FN2 @jw≥0.8 / FN1=NN2 + LN1=LN2 @jw≥0.8 • FN1=LN2 + FN2=LN1
				Rule-out:	geodistance≠0 + FN1=FN2 @jw<0.7 + FN1≠NN2
				Manual check:	The rest *(notes: sort by geodistance, then by descending jw FN, then by descending jw LN)*


STEP B
Substep 1
Blocking run: DOB1=DOB2 (non-missing) + (FN1=FN2 @ jw≥0.8 or FN1=NN2) + LN1=LN2 @jw≥0.8
Both records have MNs of ≥3 letters
Both records have MN/ MI, but not both MNs with ≥3 letters
Form bins sequentially from 1 - 4
Bin 1 FN1=FN2 + MN1=MN2 + LN1=LN2
Bin 2a MN1=MN2
Bin 2b MN1≠MN2
Bin 3a MI1=MI2
Bin 3b MI1≠MI2
Bin 4 Rest of pairs
(FN1=FN2@jw≥0.9 / FN1=NN2) + (LN1=LN2 @jw≥0.9 / LN1 string inside LN2)
The rest → to manual check (Bin 2a)
[(FN1=FN2 @jw≥0.9 / FN1=NN2) + (MN1=MN2 @jw≥0.9 / MN1=NN2) + (LN1=LN2 @jw≥0.9)] OR (FN1=FN2 @ jw≥0.95 + LN1=LN2@jw≥0.95 + both sex female)
[(FN1=FN2 @jw≥0.8 / FN1=NN2) + (MN1=MN2@jw≥0.8 or MN1=NN2) + (LN1=LN2 @jw≥0.8)] OR [(FN1=FN2 @jw≥0.9 or FN1=NN2) + (MN1=MN2@jw <0.9 + no MN1=NN2) + (LN1=LN2 @jw≥0.9)] → to manual check (Bin 2b)
The rest → auto rule-out
(FN1=FN2@jw≥0.9 + LN1=LN2@jw≥0.9) OR (FN1=NN2 + LN1=LN2@jw≥0.95)
The rest → to manual check (Bin 3a)
Route to Substep 3 (Bin 3b)
(FN1=FN2@jw≥0.95 / FN1=NN2) + (LN1=LN2 @jw≥0.95 / LN1 string inside LN2) + same sex (non-missing)
(FN1=FN2@jw≥0.87 + LN1=LN2@jw≥0.87 + same sex) OR (FN1-NN2 + LN1=LN2@jw≥0.87 + non-conflicting sex) → to manual check (Bin 4)
The rest → Route to Substep 3 (Bin 4)
Substep 2: RN screen

STEP B - *continued*

Substep 3

Applied to all pairs generated in blocking run that have not ruled-in to this point due to one of following:

- Failed RN screen (i.e. Substep 2)
- Did not rule-in through manual check of bins 2a, 2b, 3a, or 4
- Automatically ruled-out through bins 2b, 3b, or 4

Substep 3(1)

a.	Send to substep 3(2)(d):	If address1=address2 @jw<0.85 + discordant zip If address1=address2 @jw<0.75
b.	Rule-in:	If address1=address2 @jw≥0.85 + low MN threshold met
c.	Manual check:	If address1=address2 @0.75≤jw<0.85 + concordant zip If address1=address2 @jw≥0.85 + did not meet low MN threshold

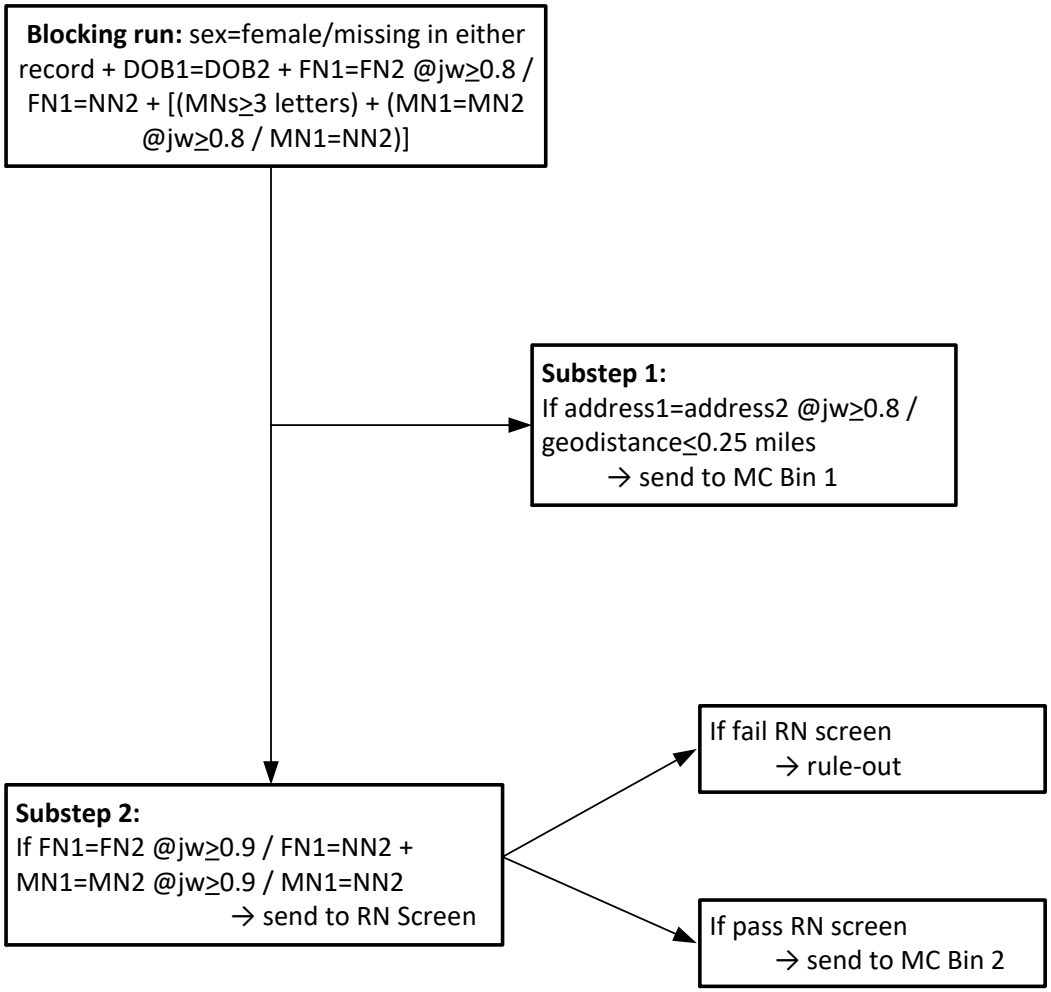
Substep 3(2)

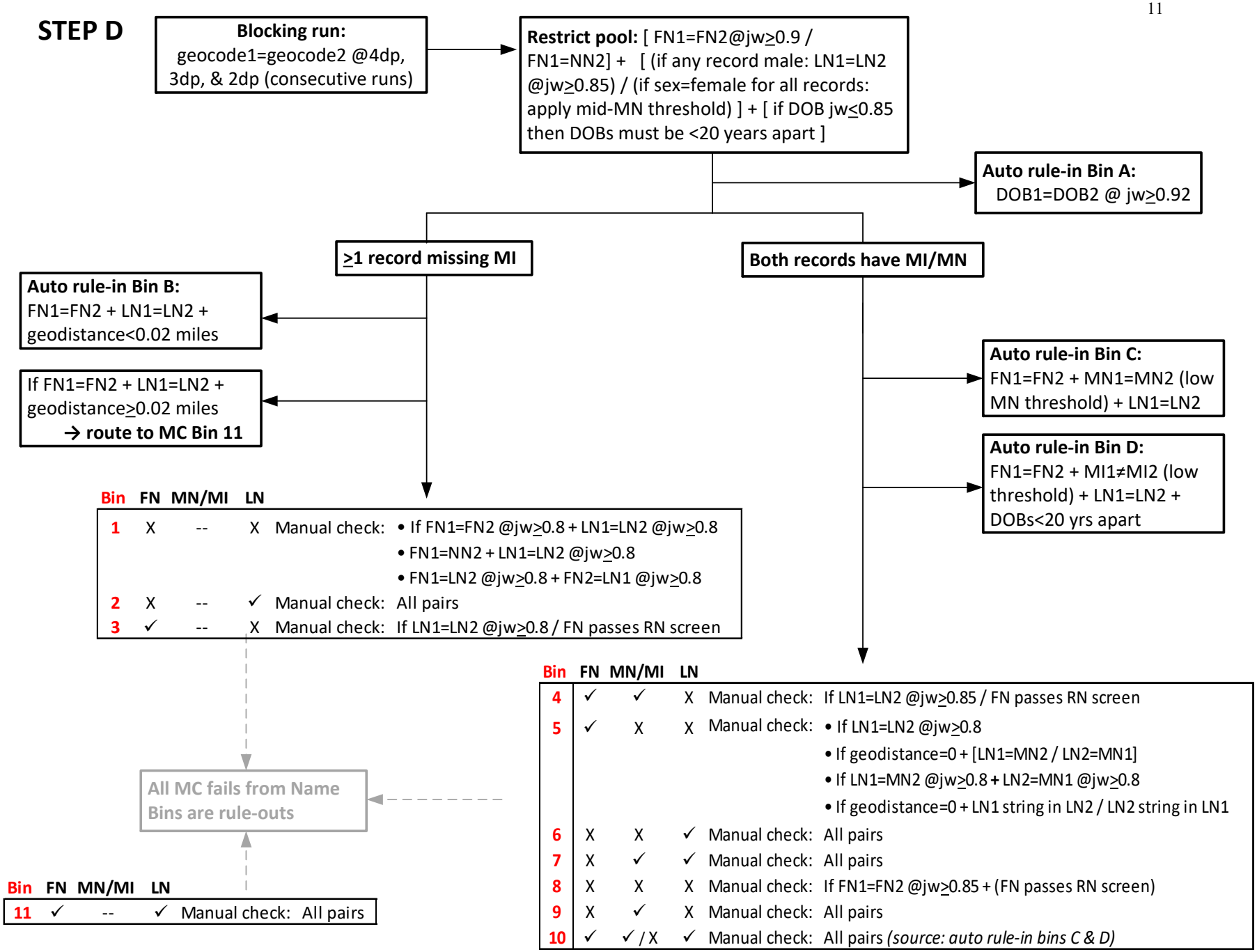
a.	Rule-in	If geodistance ≤2 miles + [FN1=FN2 @jw≥0.95 / FN1=NN2] + high MN threshold met + LN1=LN2 @jw≥0.95
b.	Rule-in	If geodistance≤0.25 miles + [FN1=FN2 @jw≥0.90 / FN1=NN2] + mid MN threshold met + LN1=LN2 @jw≥0.90
c.	Manual check	If did not rule-in above but jw address≥0.75
d.	For the rest:	
i.	Rule-in	If geodistance<50 miles + FN1=FN2 (no nn match allowed) + LN1=LN2 + mid MN threshold met
ii.	Rule-in	If geodistance≥50 miles + FN1=FN2 (no nn match allowed) + LN1=LN2 + MN1=MN2 (no NN match allowed)
iii.	Manual check all remaining pairs after RN screening and sorting into 8 bins:	*RN positive:* a. FN1=FN2 @jw≥0.90 + LN1=LN2 @jw≥0.90 + high MN threshold met b. FN1=FN2 @jw≥0.90 + LN1=LN2 @jw≥0.90 + mid MN threshold c. FN1=FN2 @jw≥0.90 + LN1=LN2 @jw≥0.90 + low MN threshold met d. The rest *RN negative:* a. FN1=FN2 @jw≥0.90 + LN1=LN2 @jw≥0.90 + high MN threshold met b. FN1=FN2 @jw≥0.90 + LN1=LN2 @jw≥0.90 + mid MN threshold c. FN1=FN2 @jw≥0.90 + LN1=LN2 @jw≥0.90 + low MN threshold met d. The rest

Definition of Low, Mid and High Middle Name Thresholds

Record A	Record B	Low MN threshold	Mid-MN threshold	High MN threshold
Missing	Missing	OK	OUT	OUT
1, 2, or ≥3 letters	Missing	OK	OUT	OUT
1 or 2 letters	≥1 letters	OK if first letter = match OUT if first letter ≠ match	Same as low	OUT
≥3 letters	1 or 2 letters	OK if first letter = match OUT if first letter ≠ match	Same as low	OUT
≥3 letters	≥3 letters	OK if MN≥0.90 OUT otherwise	Same as low	OK if MN1=MN2 @jw≥0.95

STEP C


Blocking run: sex=female/missing in either record + DOB1=DOB2 + FN1=FN2 @jw≥0.8 / FN1=NN2 + [(MNs≥3 letters) + (MN1=MN2 @jw≥0.8 / MN1=NN2)]
Substep 1:
If address1=address2 @jw≥0.8 / geodistance≤0.25 miles
→ send to MC Bin 1
Substep 2:
If FN1=FN2 @jw≥0.9 / FN1=NN2 + MN1=MN2 @jw≥0.9 / MN1=NN2
→ send to RN Screen
If fail RN screen
→ rule-out
If pass RN screen
→ send to MC Bin 2


STEP D
Blocking run: geocode1=geocode2 @4dp, 3dp, & 2dp (consecutive runs)
Restrict pool: [FN1=FN2@jw≥0.9 / FN1=NN2] + [(if any record male: LN1=LN2 @jw≥0.85) / (if sex=female for all records: apply mid-MN threshold)] + [if DOB jw≤0.85 then DOBs must be <20 years apart]
Auto rule-in Bin A: DOB1=DOB2 @ jw≥0.92
≥1 record missing MI
Both records have MI/MN
Auto rule-in Bin B: FN1=FN2 + LN1=LN2 + geodistance<0.02 miles
If FN1=FN2 + LN1=LN2 + geodistance≥0.02 miles → route to MC Bin 11
Auto rule-in Bin C: FN1=FN2 + MN1=MN2 (low MN threshold) + LN1=LN2
Auto rule-in Bin D: FN1=FN2 + MI1≠MI2 (low threshold) + LN1=LN2 + DOBs<20 yrs apart
Bin FN MN/MI LN
1 X -- X Manual check: • If FN1=FN2 @jw≥0.8 + LN1=LN2 @jw≥0.8
• FN1=NN2 + LN1=LN2 @jw≥0.8
• FN1=LN2 @jw≥0.8 + FN2=LN1 @jw≥0.8
2 X -- ✓ Manual check: All pairs
3 ✓ -- X Manual check: If LN1=LN2 @jw≥0.8 / FN passes RN screen
Bin FN MN/MI LN
4 ✓ ✓ X Manual check: If LN1=LN2 @jw≥0.85 / FN passes RN screen
5 ✓ X X Manual check: • If LN1=LN2 @jw≥0.8
• If geodistance=0 + [LN1=MN2 / LN2=MN1]
• If LN1=MN2 @jw≥0.8 + LN2=MN1 @jw≥0.8
• If geodistance=0 + LN1 string in LN2 / LN2 string in LN1
6 X X ✓ Manual check: All pairs
7 X ✓ ✓ Manual check: All pairs
8 X X X Manual check: If FN1=FN2 @jw≥0.85 + (FN passes RN screen)
9 X ✓ X Manual check: All pairs
10 ✓ ✓ / X ✓ Manual check: All pairs (source: auto rule-in bins C & D)
All MC fails from Name Bins are rule-outs
Bin FN MN/MI LN
11 ✓ -- ✓ Manual check: All pairs

VI. Selective Manual Review of Auto Rule-in and Auto Rule-out Pairs

We reviewed small random samples of candidate pairs ($n \approx 300$) from each of the auto rule-in bins and auto rule-out bins to verify that the assigning algorithms had performed as intended. In addition, we used any available match probability values to conduct targeted manual reviews; these reviews focused on candidate pairs whose values indicated that they were the least likely to be true matches (within auto rule-in bins) or the most likely to be matches (within auto rule-out bins). For the type of match probability values used to sort the pairs in this way, see manuscript tables 3 and 4.

VII. Nickname Matching

We obtained a database of "name-alias" pairs from American English Nickname Collection (Intelius, Inc.).[5] Researchers developed the database by identifying names that individuals used interchangeably in samples of government records, public web profiles, and financial and property reports. (Details of the data sources and the linkage methodology used to identify the pairs are available elsewhere.[6,7])

The list we obtained contained 331,236 name-alias pairs. We used the following steps to reduce this list to a smaller set of the most common and plausible combinations.

1. We selected pairs with at least one name that appeared in the list published by the Social Security Administration (SSA)[8] of the 500 most common boy and girl names, respectively, for babies born in 1974.*
2. The American English Nickname Collection database includes information on how frequently the name-alias pairs appear. We linked this information with the SSA name frequency data to restrict the list further to the most common pairs, defined according to the following criteria:
   - Conditional alias probability ($\text{Prob}(\text{alias}_j \mid \text{name}_i)$) in the American English Nickname Collection of $\geq 1\%$, &
   - At least one of the two names in the name-alias pair accounted for >0.5% of all male names or >0.5% all female names in 1974 (according to the SSA database).
3. Some name-alias pairs are duplicate combinations in the sense that they involve reversals of names and aliases (e.g. Kathleen-Kathy, Kathy-Kathleen). We identified these reversals and retained only one because our implementation of the rare name screen allowed for bidirectional matching.
4. Steps 1-3 reduced the initial list to 2,373 name-alias pairs. A review of this reduced list indicated that some pairs were unintuitive; they appeared to be chance matches of common names with little or no recognized connections (e.g. Mark-James, Yvette–Mary). Because name-nickname matches in our linkage algorithms are generally treated as equivalent to exact first name matches, unintuitive pairs create risks of over-matching (i.e. false positives). To ameliorate this risk, we manually reviewed the list of 2,373 pairs and removed 102 unintuitive ones, leaving a final list of 2,271 pairs.

* The highest-ranked names vary slightly from year to year. We chose 1974 because handgun buyers born in this year would have been aged 30-42 years during our study period (2004-16), an age group that has relatively high rates of handgun buying in California.

Zhang Y, *et al. Inj Prev* 2020; 26:153–158. doi: 10.1136/injuryprev-2019-043385

A review of the final list showed pairs with four main forms:

- Easily recognizable and widely used nicknames (e.g. William-Bill, Charles-Chad) and contractions (Joanne-Jo, Maxwell-Max)
- Common misspellings (e.g., Brian-Brain, Darcy-Darcey)
- Common phonetic mistakes, such as might arise when a name is being relayed orally (e.g., Shon-John, Mark-Mack)
- Invocations of middle or supplementary names in first name field (e.g., Mary - Mary-Ann)

Implementation: When the "nickname match" option was applied in the linkage algorithms, it allowed first name matches (and, in a few instances, middle name matches) within candidate pairs whenever the same name combination appeared in the final list of 2,271 name-alias pairs.

VIII. Fuzzy Date-of-birth Matches

Among records matched in Step D, common forms of birth date discrepancies included reversal of day and month fields, transcription errors (e.g., 3 instead of 8), and missing data (e.g., 5/7/178) or nonsensical data in one of the records (e.g., birth year of 1850). Males with the same name and address, and the same or very similar days and months of birth but different birth years, were particularly challenging because they may have been cohabitating fathers and sons. Hence, we generally did not accept such pairs as matches unless the birth years were less than 20 years apart; this restriction decreased the likelihood of falsely matching father-son cohabitants.

IX. Rare Name Screen

To facilitate decision-making about candidate pairs with the same or similar names but discrepancies on other link variables, we developed indicators of rarity for first and last names.

Rare first names

We obtained data on the frequency of baby names in California. The data came from first names provided in applications to the SSA for social security cards[9] and included all girl and boy names in each year over the period 1910-2016 that were given to 5 or more babies.

For each year in this 106 year period we classified as "rare" first names with counts that fell at or below the 15th percentile of the name frequency distribution for the year. So, for example, among 286,937 babies born in California in 1974, the most frequent names among those classified as rare were given to 61 babies or 0.02% of all births (e.g. Dan, Jean, Aurora) and the least frequent names among those classified as rare were given to 5 babies or 0.001% of all births (e.g. Sage, Glenna, Weldon).

Rare last names

We obtained a list of the 2,000 most frequently occurring surnames in the 2010 Census returns.[10] A total of 49.31% of the US population had one of these 2,000 surnames. The most common surname was Smith, which 2.44 million persons had (0.79% of the population). The 2000th most common surname was Kincaid, which 18,075 persons had (0.01% of the population).

Zhang Y, *et al. Inj Prev* 2020; 26:153–158. doi: 10.1136/injuryprev-2019-043385

Implementation

In the standard application of the rare name screen to first names, a candidate pair passed the screen if the name was classified by the above method as rare in the person's year of birth. In the standard application of the rare name screen to last names, any last name not on the list of the 2000 most common surnames passed the screen. A few of the algorithms screened for rarity of middle names by applying the same method and name lists as were used to screen for rarity of first names.

Zhang Y, *et al. Inj Prev* 2020; 26:153–158. doi: 10.1136/injuryprev-2019-043385

REFERENCES

1. California Department of Motor Vehicles, How to notify DMV when I change my address. https://www.dmv.ca.gov/portal/dmv/detail/faq/genfaq (accessed May 15 2019).
2. Assembly Bill No. 1461. California New Motor Voter Program (2015-16). https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201520160AB1461 (accessed May 15 2019).
3. Statewide Database (https://statewidedatabase.org/).
4. For a list of historical statewide elections and their dates, see https://www.sos.ca.gov/elections/voter-registration/voter-registration-statistics/
5. American English Nickname Collection. Philadelphia, PA: Linguistic Data Consortium 2017. https://catalog.ldc.upenn.edu/LDC2012T11
6. Carvalho V, Kiran Y, Borthwick A. American English Nickname Collection LDC2012T11. Philadelphia: Linguistic Data Consortium, 2012. https://catalog.ldc.upenn.edu/LDC2012T11 (accessed July 7 2019)
7. Carvalho VR., Kiran Y, Borthwick A. The Intelius Nickname Collection: Quantitative analyses from billions of public records. 2012, pp. 607–10, http://www.aclweb.org/anthology/N12-1075 (accessed July 7 2019).
8. Social Security Administration. Popular names by birth year. https://www.ssa.gov/cgi-bin/popularnames.cgi (accessed May 31 2019).
9. Social Security Administration. Baby names from Social Security Card Applications – State and District of Columbia Data. https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-data-by-state-and-district-of- (accessed May 31 2019).
10. United States Census Bureau. Frequently occurring surnames from the 2010 Census. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html (accessed May 31 2019).