OPEN ACCESS

Assembly of the LongSHOT cohort: public record linkage on a grand scale

Yifan Zhang,[1] Erin E Holsinger,[1] Lea Prince,[1] Jonathan A Rodden,[2] Sonja A Swanson,[3] Matthew M Miller,[4] Garen J Wintemute,[5] David M Studdert [1,6]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/injuryprev-2019-043385).

[1]Medicine, Stanford University, Stanford, California, USA
[2]Political Science, Stanford University, Stanford, California, USA
[3]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands
[4]Health Sciences, Bouvé College of Health Sciences, Boston, Massachusetts, USA
[5]Violence Prevention Research Program, UC Davis, Sacramento, California, USA
[6]Stanford Law School, Stanford University, Stanford, California, USA

Correspondence to
Dr David M Studdert, Center for Health Policy, Stanford University, Stanford, CA 94305, USA; Studdert@stanford.edu

Received 9 July 2019
Revised 18 September 2019
Accepted 21 September 2019

To cite: Zhang Y, Holsinger EE, Prince L, *et al*. *Inj Prev* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/injuryprev-2019-043385

ABSTRACT

Background Virtually all existing evidence linking access to firearms to elevated risks of mortality and morbidity comes from ecological and case–control studies. To improve understanding of the health risks and benefits of firearm ownership, we launched a cohort study: the Longitudinal Study of Handgun Ownership and Transfer (LongSHOT).

Methods Using probabilistic matching techniques we linked three sources of individual-level, state-wide data in California: official voter registration records, an archive of lawful handgun transactions and all-cause mortality data. There were nearly 28.8 million unique voter registrants, 5.5 million handgun transfers and 3.1 million deaths during the study period (18 October 2004 to 31 December 2016). The linkage relied on several identifying variables (first, middle and last names; date of birth; sex; residential address) that were available in all three data sets, deploying them in a series of bespoke algorithms.

Results Assembly of the LongSHOT cohort commenced in January 2016 and was completed in March 2019. Approximately three-quarters of matches identified were exact matches on all link variables. The cohort consists of 28.8 million adult residents of California followed for up to 12.2 years. A total of 1.2 million cohort members purchased at least one handgun during the study period, and 1.6 million died.

Conclusions Three steps taken early may be particularly useful in enhancing the efficiency of large-scale data linkage: thorough data cleaning; assessment of the suitability of off-the-shelf data linkage packages relative to bespoke coding; and careful consideration of the minimum sample size and matching precision needed to support rigorous investigation of the study questions.

INTRODUCTION

Rates of civilian gun ownership are far higher in the USA than in any other country[1] and rates of firearm-related death and injury in the USA are among the world's highest.[2] Over the last 30 years, evidence linking access to firearms to elevated risks of death and injury has grown. Nearly all of this evidence comes from ecological[3–5] and case–control[6–13] studies. Only one cohort study[14] has been conducted; this should not be surprising given the substantial data demands of the cohort design, legal barriers to the collection of population-wide information on firearm purchasing and ownership (ie, exposure data)[15] and the dearth of funding in the USA for large-scale research on firearm violence.[16 17]

To help improve understanding of the health risks and benefits of firearm ownership, we launched the Longitudinal Study of Handgun Ownership and Transfer (LongSHOT) in 2016. The study's broad goal is to produce the most complete and robust estimates to date of the causal effects of firearm ownership on the health of owners and their family members. Our first task was to assemble a cohort by linking three sources of individual-level, state-wide data from California: official voter registration records, an archive of firearm transactions and mortality data. With nearly 29 million unique voter registrants, 5.5 million handgun transfers and 3.1 million deaths during our study period (18 October 2004 to 31 December 2016), cohort assembly was large in scale and complex.

In this article, we describe the linkage methodology we developed and implemented to create the cohort. We conclude with some lessons learnt, which may be useful to other researchers embarking on large-scale data linkage projects involving public records.

DATA SOURCES

Voter registration data

We sought to build the cohort around a source of longitudinal, individual-level information on California residents—one that captured as much of the adult population as possible, while also providing accurate, up-to-date information on individuals' residential location. We considered several possible data sources (see section I of the online supplementary appendix) before settling on California's Statewide Voter Registration Database (SVRD).[18] The SVRD has several features that made it attractive for our purposes. First, it captures a majority of adult residents of the state: in our study period, registrants accounted for approximately 74% of voting-eligible residents and 62% of all adult residents.[19 20] Second, the SVRD contains information on each registrant's name, sex, date of birth and principal residential address—all important variables to our study goals. Third, the California Secretary of State is required to keep the SVRD up to date with additions (eg, registrations by new state residents and residents who attain voting age) and removals (eg, deregistrations due to death, relocation or incarceration). The mandatory updates include weekly cross-checks against death and felony records[21] and monthly cross-checks against the US Postal Service Change of Address Database.[22] Registrants cannot receive a mail-in ballot or cast a valid vote if they are not registered at the correct address, so they face relatively strong incentives to update their

address information. Finally, snapshots of the SVRD are taken regularly and archived.

In sum, SVRD extracts present a large sample of adults known to be alive and residing in California on the extract date. We obtained 13 historical extracts of the SVRD that spanned the study period and were spaced approximately 1 year apart (see section II of the online supplementary appendix).

Dealer Record of Sale database

Nearly all transfers of firearms in California—including transfers between private parties, gun show sales, gifts and loans—must be transacted through a licensed dealer.[23] Dealers relay electronically details of transfers and transferees to the California Department of Justice (CalDOJ), where they are logged into the Dealer Record of Sale (DROS) database and stored permanently.[24] Handgun transfers in California have adhered to this process for decades, creating a state-wide archive of lawful handgun transfers. It was optional for licensed dealers to log information on long gun transfers into the DROS database until 1 January 2014, when it became mandatory. We obtained DROS records on over 10 million handgun and long gun transfers made over a 32-year period (1985–2016), although this report focuses on the 5.5 million transfers recorded during the study period.

Mortality data

The California Death Statistical Master Files are the state's official mortality records.[25] They contain detailed information on deaths among state residents, including deaths that occur out of state. The records include the decedent's name, sex, date of birth, race, and residential address, as well as the date, cause (International Classification of Diseases, Tenth Revision code) and location of death. We obtained data on all recorded adult deaths from 2000 through 2016.

OVERVIEW OF LINKAGE PROCESS

We used probabilistic data linkage methods to match the firearm transfer records and mortality records, respectively, to the SVRD extracts at the person level. The link variables, available in all three principal data sets, were: first name, middle name (or initial), last name (and former last name), date of birth (day/month/year), sex and geocoded residential address. Candidate pairs that matched on all link variables were automatically accepted as matches.

However, the bulk of the linkage effort involved developing and applying algorithms to detect matches with imperfect agreement on one or more of the link variables. Variation across public records in how an individual's information is recorded is common and occurs for a variety of reasons, including recording mistakes (eg, misspellings, entry errors), inconsistent use of certain identifying fields (eg, middle name, residential unit number) and normal temporal change among accurate identifiers (eg, new residential address, changes of last name).

The mortality–SVRD linkage was conducted between November 2017 and October 2018 and the DROS–SVRD linkage was conducted between October 2018 and March 2019. Study data were stored on secure servers at Stanford's Center for Health Policy and all linkage work was performed in a secure computing environment.

TEMPORAL STRUCTURE OF LINKAGE

We conceived LongSHOT as an open cohort in which cohort members would come under observation on the date of the first SVRD extract in which they appeared and remain under observation until the day before the date of the next voter file extract in which they did not appear, death, or the study end date, whichever came first. Our approach to data linkage mapped on to this design. Linkage was segmented according to the time intervals between consecutive SVRD extracts, with purchasers and deceased within each time interval eligible to match to voter registrants named in the SVRD extract that marked the beginning of the interval. This segmented approach meant that assembly of the cohort proceeded in 26 discrete 'interval links'—13 for the mortality–SVRD linkage and 13 for the DROS–SVRD linkage (see section III of the online supplementary appendix for further details).

LINKAGE STEPS AND ALGORITHMS

Within each interval link, we applied a suite of linkage algorithms. The algorithms were organised into four consecutive steps (table 1). The chief function of the algorithms was to sort candidate pairs into three groups: (1) those with very high probability of being matches (which we called 'auto rule-ins'); (2) those with very low probability of being matches ('auto rule-outs'); and (3) the rest ('manual checks'). The methodology used to develop the algorithms is described in section IV of the online supplementary appendix, and the algorithms themselves are described in section V.

MANUAL REVIEW

A member of our study team examined each candidate pair assigned to manual check bins in the DROS–SVRD linkage (n≈90 000) and the mortality–SVRD linkage (n≈276 000), comparing the available information to decide whether the records referred to the same person. We also subjected subsamples of pairs assigned to the auto rule-in and auto rule-out bins to manual review (details of those reviews are provided in section

Table 1 Summary of data linkage steps used to assemble the LongSHOT cohort

Step	Blocking key	Focus of linkage algorithms and manual reviews	Step rationale	Percentage of all matches identified	
				Purchasers-voter file (n≈1.2 m)	Deaths-voter file (n≈1.7 m)
A	Same date of birth+same residential address	Name fields	Identifies highest probability matches	80.12%	79.74%
B	Same date of birth+similar first and last names	Name fields, address fields	Removes address as a blocking criterion to capture relocators between SVRD extract date and purchase/death date.	17.92%	15.64%
C	Same date of birth+both persons female+similar first and middle names	Name fields, address fields	Same as step B, except also allows changes of last names among women.	0.73%	0.06%
D	Same address	Name fields, date of birth field	Removes date of birth from blocking key to allow for errors in this field.	1.24%	4.56%

LongSHOT, Longitudinal Study of Handgun Ownership and Transfer; SVRD, Statewide Voter Registration Database.

Table 2 Inter-rater and intrarater reliability of manual review

	DROS–SVRD linkage (n=500 candidate pairs)		Mortality–SVRD linkage (n=500 candidate pairs)	
	Inter-rater	Intrarater	Inter-rater	Intrarater
Expected agreement by chance alone (%)	50.51	50.60	51.79	52.10
Observed agreement (95% CI)	94.20% (91.78% to 96.08%)	93.40% (90.86% to 95.41%)	91.80% (89.04% to 94.05%)	92.60% (89.94% to 94.74%)
Kappa (95% CI)	0.88 (0.83 to 0.91)	0.87 (0.82 to 0.90)	0.83 (0.77 to 0.87)	0.85 (0.79 to 0.89)

DROS, Dealer Record of Sale database; SVRD, Statewide Voter Registration Database.

VI of the online supplementary appendix). During the DROS–SVRD linkage, manual reviewers were blinded to the results of the earlier mortality–SVRD linkage.

In the context of a study whose goal is to quantify the relationship between firearm ownership and injury risk, it is unclear whether overmatching or undermatching poses the greater threat of bias. (The answer depends on the exposure and outcome profiles of mismatched records, which is unknown.) A plausible consequence is a bias to the null in analyses estimating differences in mortality risk between handgun owners and non-owners. Given these considerations, we adopted a simple balance of probabilities standard: if the reviewer judged a candidate pair more likely than not to be a match, it was called one, otherwise it was called a non-match.

To assess inter-rater and intrarater reliability of the manual reviews, we randomly selected 1000 candidate pairs from across all manual check bins—500 from the DROS–SVRD linkage and 500 from the mortality–SVRD linkage. The two reviewers who conducted the original manual reviews reviewed all 1000 pairs. Inter-rater reliability measures were calculated by comparing reviewer A's determination in the original review to reviewer B's determination in the reliability review; intrareliability measures compared reviewer A or B's determination in the original review to the same reviewer's determination in the reliability review. Table 2 reports the results of the reliability testing.

PREMATCHING FIREARM PURCHASERS

In the DROS database, each firearm transferee has a unique identification number that allows CalDOJ to easily identify multiple acquisitions by the same person over time. Voter registrants in the SVRD also have a unique identification number. Together, these two identifiers provided an efficient method of linking to the SVRD purchasers who acquired multiple handguns during the study period. If purchaser X was matched to voter registrant Y in the first interval link, for example, our first move after generating the pool of candidate pairs in subsequent interval links was to 'pre-match' all X-Y candidate pairs in the pool. Since many firearm owners—both nationally[26] and within California[27]—acquire multiple weapons, this short cut helped reduce the manual review workload, especially in later intervals.

PROBABILISTIC MATCHING OF KEY VARIABLES

Approximately 72% of the matches identified in the DROS–SVRD linkage and the mortality–SVRD linkage, respectively, matched exactly on all link variables. While these proportions are high and bolster confidence in match accuracy, they also indicate that limiting the linkage to such 'perfect' matches would have missed matching a non-trivial number of purchases and deaths in the cohort. To avoid these false negatives, our linkage algorithms applied fuzzy matching techniques to each link variable.

Fuzzy matching of names

Table 3 summarises the techniques used to retrieve matches with name field discrepancies. The most widely used of these techniques was an edit distance measure of the degree of discrepancy between imperfectly matched names. After testing several options (eg, Levenshtein, Damerau-Levenshtein, Jaro, Jaro-Winkler), we chose the Jaro-Winkler distance algorithm because it performed best in our data. This algorithm scores similarity between two

Table 3 Techniques for identifying matched records with discrepant first, middle and/or last names

Extent of discrepancy	Source of name discrepancy	Retrieval technique	Place of application*
Slight or moderate	Misspelling, entry errors, and so on	Jaro-Winkler distance	Throughout all steps
	Phonetically similar but different spelling	Encoded all names using the Soundex function in R to allow matching of phonetically similar names[36]	Throughout all steps
	Shortened or expanded/hyphenated versions of same names	Allowed for substring matches between name fields	Step A: name bins 3, 4, 5 Step B: name bins 2a, 4 Step D: name bin 5
Extreme	Use of nicknames and contractions (eg, Elizabeth—Betty, Tommy Joe—TJ)	Allowed for matches to common nicknames (see section VII of online supplementary appendix)	Step A: name bins 1, 2, 6, 7, 9 Step B: blocking key; substep 1 bin 2a, 2b, 3a, 4; substeps 3(2)(a) and (b) Step C: blocking key; substep 2 Step D: blocking key; name bin 1
	Change or concatenation of last names among females	Relaxed last name matching criteria	Step C
		Allowed for matches between current last names (in purchaser and mortality records) and former last names (in voter records)	Throughout all steps
	Switches in name order	Allowed for reverse matching of first-middle and first-last	Step A: name bins 1, 2, 3, 5, 6, 9 Step D: name bins 1, 5

*Refers to locations in the charts of linkage algorithms provided in section V of the online supplementary appendix.

character strings on a scale from 0 (none) to 1 (exact match); the score is based on the number of characters the strings have in common and places extra weight on matches between early characters in the strings.[28] We incorporated scores from the Jaro-Winkler algorithm into several blocking keys and many algorithms. In our data, name fields with scores between 0.90 and 0.99 generally indicated a high likelihood that the names had minor discrepancies but were the same, scores between 0.77 and 0.89 indicated possible name matches, and name matches among pairs with scores below 0.77 were uncommon.

Edit distance metrics such as the Jaro-Winkler do not help identify matches between name fields that are very or completely different. Extreme discrepancies in name fields across public records pertaining to the same person occur for various reasons. For example, first and middle names are frequently used with variations (eg, nicknames, initials only) or interchangeably. Also, in our linkage, some people (mostly women) changed their recorded last name in the interval between the date of the SVRD extract and the date of their gun purchase or death. The lower section of table 3 describes the techniques we used to detect matches among records with extreme name mismatches. The most important of these techniques was nickname matching, details of which are provided in section VII of the online supplementary appendix.

Fuzzy matching of residential addresses

To facilitate address matching we geocoded residential addresses for all DROS, mortality and voter records using StreetMap Premium for ArcGIS software[29] and OpenCage Geocoder.[30] A total of 98% of the geocodes assigned to records were based on exact matches to a dwelling rooftop; 1% of geocodes were 'ties', indicating a location very near the address but uncertainty over the specific dwelling; geocodes could not be identified for the remaining 1% of records.

To avoid missing matches with slight geocode discrepancies—owing, for example, to minor discrepancies in the address strings or inconsistent use of unit/apartment numbers—the step A blocking key and several of the algorithms relaxed the number of decimal places to which the geocodes of candidate pairs had to match. (Given California's location on the globe, geocodes are precise to approximately 10 m at the fourth decimal place, 100 m at the third decimal place and 1 km at the second decimal place.)[31] To avoid overmatching, use of fuzzy geocode matching triggered stringent match requirements on other link variables. We also generated 'geodistances'—a measure of the distance between imperfectly matched geocodes in candidate pairs—and used these both to constrain fuzzy geocode matches and prioritise pairs in manual review.

Fuzzy matching of birth dates

Unlike name and address, record error is usually the only explanation for the same person having discrepant birth dates across public records. We insisted on exact date of birth matches in three of the four linkage steps. The blocking key for step D used less stringent criteria on this variable to create a pool of candidate pairs with probable errors in one of the birth dates and exact or high-probability matches on the other link variables. Examples of errors in birth dates within pairs that were judged to be matches are described in section VIII of the online supplementary appendix.

RESOLVING UNCERTAIN MATCHES

Do two records for Jane L Garcia with the same date of birth but residential addresses a mile apart refer to the same person? What about two records for Abdul Horatio Jones with birth dates 5 months apart and two different addresses that are located along the same small street? Neither computer algorithms nor manual review can confidently answer these questions.

We generated several additional variables to aid our decision-making in such 'hard' cases. The variables and the probabilistic intuition that motivated them are described in table 4. We made some use of these variables in the linkage algorithms, particularly name rarity (see section IX of the online supplementary appendix), but their primary use was to inform subjective decision-making during manual review.

MULTIMATCHES AND CONFLICTING MATCHES

We generally matched with replacement. Thus, within interval links we allowed a deceased or purchaser to match to more than one voter registrant and, conversely, for a registrant to match to multiple deceased or purchasers; for purchasers, both forms of multimatching were also allowed *across* interval links. After all linkage steps were complete, manual review of these anomalous clusters functioned as a form of quality control, allowing identification of the true matches and elimination of the false ones and, in a few instances, highlighting errors (eg, duplicate records) in a component data set.

LESSONS LEARNT

We began assembly of the LongSHOT cohort armed with a good working knowledge of data linkage methods. The literature in this area has blossomed in recent years,[32–34] and several members

Table 4 Additional variables used to inform match determinations in hard cases

Consideration	Intuition	Place of application*
Rarity of name in the population (see section IX of the online supplementary appendix)	Two records with same name but minor discrepancies on another link variable are more likely to be the same if first, middle or last name is uncommon.	Step B: substep 2, substep 3(2)(d) Step C: substep 2 Step D: name bins 3, 4, 8
Geodistance between discrepant addresses	Persons who move addresses are more likely to relocate near (eg, same city or county) than far (eg, distant city or county).	Step A: all name bins Step B: substep 3(2) Step C: substep 1 Step D: auto rule-in bin B; routing rule for name bin 11; name bins 1, 5
Geodistance+rurality	All else equal, two records that match on all variables except address are more likely to be true matches if both are in the same sparsely populated area than if both are in the same densely populated area.	Manual review only
Time interval between discrepant dates of birth	When errors in (or intended alternate uses of) birth dates occur, the conflicting dates are more likely to be proximate than distant.	Step D: blocking key; auto rule-in bin D

*Refers to locations in the charts of linkage algorithms provided in section V of the online supplementary appendix.

of our team had record linkage experience from previous projects (although on a much smaller scale). None of this fully prepared us for the scale and complexity of the LongSHOT linkage, nor spared us from a multitude of wrong turns and mistakes along the way. Lessons were hard won. Here are four we wished we had known or more fully appreciated at the outset.

First, we spent about 10 person-months cleaning the component data sets before commencing linkage. This was enough time to correct many errors and irregularities; we planned to deal with the rest once the analytical data set was formed. Deferral was a mistake. As the linkage progressed, problems discovered in match results exposed additional anomalies in the underlying data sets, forcing us to pause several times for supplementary cleaning. Most of these anomalies could have been found and addressed with more thorough prelinkage cleaning, and dealing with them at that stage would have been far more efficient. A related lesson is that presumptions about the cleanliness of vital administrative databases, including those for recording deaths and voter registration, should be set aside.

Second, we spent time in the first year experimenting with off-the-shelf matching packages (eg, Link Plus, G-Link, Record Linkage package in R) before eventually deciding to write our own code. Some of the packages had to be ruled out because they could not accommodate the volume demands of our linkage. However, our main reservation turned out to be an inability to clearly see, understand and, when necessary, modify the matching machinery in these products. We took too long to figure out that this was not a linkage for point-and-shoot mode and that we needed full manual control of the settings.

Third, there were several opportunities to pare back the scale of the linkage. In particular, it was always evident that we would have abundant non-owners, and in the final cohort 90.3% of members experienced neither the exposure nor the outcome of interest. A reduced form approach, such as a matched cohort design, would have alleviated substantial manual review burden, probably without materially compromising statistical power or precision. We chose not to downsize in this way because future phases of LongSHOT will consider additional research questions—including risks of household-level exposure to firearms—for which a less restricted design will have important advantages. Had such ancillary considerations not been on the horizon, however, a reduced form design would have been the smart choice.

Finally, the single most important determinant of workload in a linkage of this kind is the degree of matching precision sought. As noted above, nearly three-quarters of the purchaser and mortality matches came from perfect matches on our link variables. Stopping there would have dramatically reduced the workload, and doing so may be appropriate for studies in which the loss of statistical power is acceptable and risks of bias and generalisability from false negatives are relatively low.

We pressed on to retrieve imperfect matches as best we could for several reasons. The public health importance and political sensitivity of our topic summoned a high degree of precision. In addition, we predicted, correctly, that fuzzy matches recovered through steps B, C and D would differ systematically from those identified in the relatively pristine step A matches. Table 5 shows that purchasers matched in later steps tended to be younger, and both purchasers and deceased matched in later steps were more likely to be members of racial or ethnic minorities. These are important population subgroups for understanding patterns and causes of gun violence. Moreover, disproportionately excluding them from consideration would have compounded the fact that these same subgroups are already under-represented in a cohort anchored in voter registration.[35]

Table 5 Characteristics of sharp and fuzzy matches*

	DROS–SVRD linkage		Mortality–SVRD linkage	
	Matches identified in step A	Matches identified in step B, C or D	Matches identified in step A	Matches identified in step B, C or D
Male (%)	86.05	85.45	51.96	48.81
Age†				
Mean (years)	45	38	73	73
Median (years)	45	35	76	77
Race/ethnicity				
White (%)	74.52	69.33	70.73	66.33
Hispanic (%)	13.18	15.76	13.30	14.41
Black (%)	3.78	5.79	8.34	11.82
Asian (%)	5.73	5.80	7.00	6.55
Other (%)	2.79	3.32	0.63	0.90
Residential location‡				
Urban (%)	81.96	81.16	87.80	87.41
Suburban (%)	12.22	10.95	6.94	6.10
Large rural town (%)	3.11	3.46	3.11	2.64
Small rural town (%)	2.28	2.48	1.61	2.37

*In both linkages, all variables in the step A match are significantly different ($p<0.01$) from step B/C/D matches.
†Refers to cohort members' age at the midpoint of their observation period.
‡Categories are based on the US Census Bureau's urban-rural classification system and refer to cohort members' residential location at time they entered the cohort.
DROS, Dealer Record of Sale database; SVRD, Statewide Voter Registration Database.

CONCLUSION

Over 3 years of concerted effort we assembled the LongSHOT cohort, which consists of 28.8 million adults followed for up to 12.2 years. A total of 1.2 million cohort members purchased at least one handgun during the study period and 1.6 million died—nearly 14 500 of them from firearm-related injuries. Analyses of the cohort will help advance understanding of the effects of handgun ownership on cause-specific mortality risks; in the long run, it will serve as a platform for addressing other questions about the health risks and benefits of firearm ownership for owners and households.

Although the cohort is the largest assembled to date for addressing these questions, certain design choices we made and limitations of the data sets we used to form the cohort mean that future analyses of cohort data must grapple with various methodological challenges; table 6 foreshadows several key ones. We hope that this account of our methods and travails in creating

Table 6 Key challenges to address in future analyses of the LongSHOT cohort exploring the relationship between handgun ownership and mortality

Challenge	Description
Mismeasurement of exposure	Non-handgun owners may in fact be owners due to, for example, failure to match their purchases in probabilistic linkage, unlawful handgun acquisition and purchases made prior to 1985. Non-handgun owners may own unobserved long guns.
Unobserved confounding	Relevant differences may exist between handgun owners and non-owners that are not measured in the linked data (eg, incidence of mental illness, risk-taking propensity).
Restriction of cohort to voter file registrants in California	Generalisations to non-registrants in California and to residents of other states may be impaired by relevant unobserved heterogeneity.

LongSHOT, Longitudinal Study of Handgun Ownership and Transfer.

the LongSHOT cohort may help other public health researchers improve the quality and efficiency of their own data linkage efforts.

What is already known on this subject

- Existing research provides substantial evidence of a positive association between firearm availability and risk of firearm-related death and injury.
- Virtually no cohort studies of this relationship have been conducted—chiefly, because population-wide information on firearm availability is difficult to obtain.

What this study adds

- We demonstrate the feasibility of linking public records from multiple sources (voter registration files, archival information on firearm transfers, and mortality data) to produce a large cohort in which handgun ownership and death are observed
- Future analyses of the cohort will help advance understanding of the effects of handgun ownership on cause-specific mortality risks.

Acknowledgements The authors thank Hitsch Daines, Anunay Kulshrestha and Zach Templeton for research assistance; Stace Maples at Stanford Geospatial Center and Claudia Engel at the Stanford Libraries for assistance with geocoding; Michael Francis at the Office of the Secretary of State and Karin McDonald at the California Statewide Database for assistance with voter registration data; and staff at the Bureau of Firearms, California Department of Justice for assistance with Dealer Record of Sale data.

Contributors YZ, EEH, LP and DMS conducted all data cleaning. YZ, EEH and DMS developed and implemented the linkage algorithms. DMS obtained project funding, with assistance from YZ, GJW, JAR, MJM and SAS. DMS, YZ, GJW and LP obtained the study data. JR provided expert advice regarding voter registration data and geocoding. MJM, JAR, SAS and GJW advised on study design and helped troubleshoot issues arising in the data linkage. DS wrote the first draft of the manuscript. YZ, EEH, LP, JAR, SAS, MJM and GJW contributed revisions relating to important intellectual content. DMS is the guarantor of the study.

Funding The study was funded by the Fund for a Safer Future (Grant No GA004696) and the Joyce Foundation (Grant No 17-37241).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The LongSHOT project was approved by the Institutional Review Board at Stanford University.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD
David M Studdert http://orcid.org/0000-0003-0585-5537

REFERENCES

1 Karp A. *Estimating global civilian-held firearms numbers. Briefing paper*. Geneva: Small Arms Survey, 2018.
2 Naghavi M, Marczak LB, Kutz M, *et al*. Global mortality from firearms, 1990-2016. *JAMA* 2018;320:792–814.
3 Miller M, Lippmann SJ, Azrael D, *et al*. Household firearm ownership and rates of suicide across the 50 United States. *J Trauma* 2007;62:1029–35.
4 Miller M, Hemenway D, Azrael D. State-level homicide victimization rates in the US in relation to survey measures of household firearm ownership, 2001-2003. *Soc Sci Med* 2007;64:656–64.
5 Miller M, Azrael D, Hemenway D. Firearm availability and unintentional firearm deaths. *Accid Anal Prev* 2001;33:477–84.
6 Kellermann AL, Rivara FP, Somes G, *et al*. Suicide in the home in relation to gun ownership. *N Engl J Med* 1992;327:467–72.
7 Kellermann AL, Rivara FP, Rushforth NB, *et al*. Gun ownership as a risk factor for homicide in the home. *N Engl J Med* 1993;329:1084–91.
8 Wiebe DJ. Homicide and suicide risks associated with firearms in the home: a national case-control study. *Ann Emerg Med* 2003;41:771–82.
9 Dahlberg LL, Ikeda RM, Kresnow M-J. Guns in the home and risk of a violent death in the home: findings from a national study. *Am J Epidemiol* 2004;160:929–36.
10 Anglemyer A, Horvath T, Rutherford G. The accessibility of firearms and risk for suicide and homicide victimization among household members: a systematic review and meta-analysis. *Ann Intern Med* 2014;160:101–10.
11 Miller M, Swanson SA, Azrael D. Are we missing something pertinent? A bias analysis of unmeasured confounding in the Firearm-Suicide literature. *Epidemiol Rev* 2016;38:mxv011–19.
12 Miller M, Azrael D, Hemenway D. Firearms and violence death in the United States. In: Webster DW, Vernick JS, eds. *Reducing gun violence in America*. Baltimore MD: Johns Hopkins University Press, 2013.
13 Cummings P, Koepsell TD, Grossman DC, *et al*. The association between the purchase of a handgun and homicide or suicide. *Am J Public Health* 1997;87:974–8.
14 Wintemute GJ, Parham CA, Beaumont JJ, *et al*. Mortality among recent purchasers of handguns. *N Engl J Med* 1999;341:1583–9.
15 Giffords Law Center to Prevent Gun Violence. Registration. Available: https://lawcenter.giffords.org/gun-laws/policy-areas/gun-owner-responsibilities/registration/
16 Stark DE, Shah NH. Funding and publication of research on gun violence and other leading causes of death. *JAMA* 2017;317:84–5.
17 Kellermann AL, Rivara FP. Silencing the science on gun research. *JAMA* 2013;309:549–50.
18 California code of regulations §20108 et seq.
19 California Secretary of State. Voter registration statistics. Available: http://www.sos.ca.gov/elections/voter-registration/voter-registration-statistics/
20 United States Census Bureau. American community survey 1-year estimates, tables B01001. Available: https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t
21 California code of regulations §20108.55.
22 California code of regulations §20108.50.
23 Cal Penal code §§26500, 27545.
24 Bureau of Firearms, California Department of Justice. DROS entry system (des) – firearms and ammunition dealer user guide. rev. 4, 6/27/2019. Available: https://oag.ca.gov/sites/all/files/agweb/pdfs/firearms/pdf/dros-des-firearms-ammunition-dealer-user-guide.pdf [Accessed 5 Jul 2019].
25 California Department of Public Health. Vital records data and statistics. Available: https://www.cdph.ca.gov/Programs/CHSI/Pages/Data%20Types%20and%20Limitations.aspx [Accessed 19 May 2019].
26 Azrael D, Hepburn L, Hemenway D, *et al*. The stock and flow of U.S. firearms: results from the 2015 national firearms survey. *RSF: The Russell Sage Foundation Journal of the Social Sciences* 2017;3:38–57.
27 Kravitz-Wirtz N, Pallin R, Miller MJ, *et al*. Firearm ownership and acquisition in California: findings from the 2018 California safety and wellbeing survey (unpublished manuscript on file with authors) 2019.
28 Winkler WE. *String comparator metrics and enhanced decision rules in the Fellegi-Sunter model ofrecord linkage. Proceedings of the Section on Survey Research Methods*. American Statistical Association, 1990: 354–9.
29 ESRI, ArcGIS Enterprise. StreetMap premium, 2018. Available: http://enterprise.arcgis.com/en/streetmap-premium/
30 OpenCage Geocoder. API. Hertford, United Kingdom. Available: https://opencagedata.com/
31 Goodchild MF, Gopal S. *The accuracy of spatial databases*. London: Taylor and Francis, 1989.
32 Herzog TN, Scheuren FJ, Winkler WE. *Data quality and record linkage techniques*. New York: Springer, 2007.
33 Christen P. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Heidelberg: Springer, 2012.
34 Dusetzina SB, Tyree S, Meyer AM, *et al*. *Linking data for health services research: a framework and instructional guide. AHRQ publication No. 14-EHC033-EF*. Rockville, MD: Agency for Healthcare Research and Quality, 2014.
35 Citrin J, Highton B. *How race, immigration, and ethnicity shape the California electorate*. San Francisco: Public Policy Institute of California, 2002.
36 Howard JP. Phonetic spelling algorithm implementations for R (J STAT software, 2019 – forthcoming).