eAppendix 1

Extended introduction

Self-harm is a leading cause of morbidity and mortality in the United States (US), accounting for over 44,000 deaths and 500,000 injuries in 2015.¹ Rates of self-harm are also increasing,¹ for some groups rapidly,² but the reasons for these increases are not well-understood. The influence of social environments on self-harm has been recognized for over a century,³ but research to identify which features of the social environment are most influential is limited.

Community violence—defined as experiencing, witnessing, or hearing about violence in one’s community—is one potentially modifiable feature of the social environment that may influence self-harm. However, few studies have examined the association of community violence with self-harm,⁴–¹¹ and to our knowledge, no research has examined short-term, within-community variation in violence, as opposed to chronic or overall levels of violence.

Within-community variation in violence is directly relevant to common models of self-harm. The stress-diathesis model and its variants posit that incidents of self-harm are the confluence of long-term predisposition to self-harm (e.g., due to early life adversity) with stressful life events (e.g., loss of a loved one or psychosocial crisis) that trigger brief periods of elevated risk.¹² Increases in community violence—for example, having neighbors who were victims of a recent shooting—may trigger self-harm in a vulnerable individual. The effects of community violence on self-harm may also vary for different demographic subgroups, because self-harm is a heterogeneous condition with different drivers and manifestations.¹,¹³,¹⁴ Existing models of self-harm support the hypothesis that within-variation in violence may be associated with self-harm for some groups but not others, due to differences in vulnerability to stressors or in levels of predisposing risk factors.

From a methodological standpoint, chronic exposure to community violence is strongly associated with other self-harm risk factors such as economic opportunity, making the effects of these factors difficult or impossible to disentangle, a phenomenon known as structural confounding¹⁵ that has limited past research.⁴–¹¹ We address this limitation by investigating within-community variation in violence, allowing comparison of residents of the same community at times with relatively high and low levels of community violence, thereby controlling for community-level factors that are time-invariant over the study period.

Additionally, individual conditions such as mental and substance use disorders are risk factors for violence towards both self and others in the community, and thus are potential confounders of the relationship between community violence and self-harm. Because self-harm is rare, studies of self-harm are frequently retrospective and lack detailed data to adequately control individual-level confounding. Examining within-community changes in violence over relatively short time frames enables the use of designs such as the case-crossover to compare each individual’s exposures at different times while controlling for individual time-invariant risk factors over the study period.

Finally, community violence and self-harm both have long-term trends and seasonal patterning, peaking in summer and plunging in winter.¹⁶ Past research not accounting for this may thus have detected associations that are simply artefacts of this temporal patterning. Analyzing within-community variation in violence allows this temporal patterning to be explicitly modeled and removed to isolate the associations of interest.
We examined whether within-community variation in community violence is associated with fatal and nonfatal self-harm. To maximize control of individual and community confounders, we utilized a case-crossover approach with community-matched controls drawn in close time proximity to cases. We leveraged data from statewide population-based registries, surveys, and healthcare utilization data from California, a large and heterogeneous state with self-harm trends similar to those seen nationwide.

**Extended methods**

**Overall study designs and data sources**

We conducted a case-crossover study, comparing cases’ exposure at a time relevant to case occurrence to exposure at referent non-case times. We compiled data on self-harm and community violence for 2005-2013 from mortality, emergency department, and inpatient hospitalization discharge records from the California Office of Vital Records and the Office Statewide Health Planning and Development. Records included all deaths and hospital visits statewide, excluding active duty military hospitals, and captured medical information including external cause of death or injury, demographic characteristics, and location of residence. Cases were all deaths and hospital visits due to deliberate self-harm (ICD-10 death codes X6-X8; ICD-9-CM hospital visit code E95). Controls were the cases themselves at referent time periods (see Selection of control time periods). External cause of injury coding in California’s hospital discharge records is compulsory, entails ongoing quality assurance efforts, and is considered 100% complete.\(^{18}\) In mortality records, external cause of mortality codes for homicide and self-harm are also considered valid and complete.\(^{19}\) Emergency department records were available starting in 2005.

Cases were restricted to those occurring between March 1, 2005 and December 1, 2013, such that controls could be drawn from as early as January 31, 2005 and from as late as December 31, 2013, and exposure could be assessed for the 30 days before these dates. We restricted the study to California residents (those for whom we had community violence data) and to those aged 15 to 84 at the time of injury because there were few cases of self-harm outside of that range.

**Exposure characterization**

Community violence was assessed using deaths due to homicide (ICD-10 death codes X85-X99, Y00-Y09, Y35, U01, U02, Y871) and hospital visits due to assault (ICD-9 hospital visit codes E960-E969, E970-E977) in the Consistent Public Use Microdata Area (CPUMA) of residence. CPUMAs are geographic partitions designated by the US Census Bureau that include at least 100,000 residents. The 110 CPUMAs in California (Appendix Figure 1) are consistently defined over the study period, and correspond to known neighborhoods in urban areas (95% of the California population), and counties or aggregations of small counties in rural areas. We selected CPUMAs to define communities because they are recognized places of residence and large enough for stable estimation of community violence rates for short time periods. CPUMAs were also found to be meaningful geographic units in previous research on community violence and self-harm.\(^{11}\) CPUMAs are also the smallest geographic identifier available in the American Community Survey (ACS), an alternative source of controls considered in sensitivity analyses (see below). The CPUMA of residence was determined from the geocoded decedent address (mortality records) or the zip code of residence via geographic crosswalk (hospital records).\(^{20}\)
To our knowledge, there is no evidence on the critical exposure period (lag time and duration) for the association of within-community variation in violence with self-harm. Related literature on stressful life events and self-harm varies in the time frames assessed; self-harm has been associated with stressors occurring within a few hours and as long as several weeks prior. We hypothesized that any effects of within-community variation in violence would be immediate and of short duration. Thus, we selected a reasonable time frame of 30 days prior to injury/survey to balance capturing short-term, acute effects with pooling enough data to estimate stable rates of community violence. We then tested the sensitivity of our results to longer and shorter time windows.

We calculated community violence rates using ACS-based population estimates as denominators. To separate variation in community violence from predictable temporal patterning including seasonality, we de-trended these community violence rates by applying a Kalman smoother. To ensure temporal ordering, we applied the smoother to the unique time series of 30-day units spanning 2005 to 2013 and defined by the community and index day of the case. For example, for a case occurring on April 20 2007 in a given community, we constructed a time series of community violence rates in 30-day time units in the set {…, February 20 2007 – March 21 2007, March 22 2007 – April 20 2007, April 21 2007 – May 20 2007, …}, and applied the smoother to this series. We defined within-community variation in violence, or deviations from expected levels, as the difference between the observed rate and the modeled rate of community violence (i.e. the residuals of the Kalman smoother). Previous simulation work suggests that the Kalman smoother is superior to a range of other time series methods in the separation of unpredictable versus predictable patterning of violence in California populations. Appendix Figure 2 depicts an example community violence trend and residuals after applying the Kalman smoother. Violence residuals created using standard ARIMA models were highly correlated (Pearson’s correlation: 0.95).

**Figure 1: Geographic boundaries of Consistent Public Use Microdata Areas (CPUMAs) in California**
Confounder assessment

Potential confounders were identified a priori based on scientifically established risk factors for self-harm and factors that affect community violence or its determinants. The case-crossover design controls time-invariant community- and individual-level confounders because cases serve as their own controls at different time points within the same community. Remaining potential confounders are time-varying community and individual factors for both the case-crossover and case-control designs, and time invariant individual factors in the case-control design.

Time-varying individual factors (e.g. mental disorder symptoms) were not captured in death and discharge records. These may be mediators and/or confounders. Confounding by these factors was limited by drawing controls as close in time as possible to cases (see Selection of control time periods).

In the case control design, measured time-invariant individual-level confounders were controlled in the analysis. Variables controlled in the final case-control analysis depended on availability in death, discharge, and ACS records. Individual-level confounders included in case-control analyses of fatal self-harm were marital status (married/partnered, divorced/widowed, or single never-married), education (high school and 4-year-college completion), foreign born (yes/no), history of military service (yes/no), and recent immigration to the United States (years of residence in US is more/less than 5 years). Case-
control analyses of nonfatal self-harm controlled for individual-level primary language spoken (English/not English).

Time-varying community-level confounders were also controlled by time proximity of cases and controls, and additionally by controlling measured confounders. Community-level confounders in both case-control and case-crossover analyses included the following: percent male, percent Hispanic, percent non-Hispanic Black, percent non-Hispanic Asian or Pacific Islander, percent non-Hispanic American Indian/Alaskan Native, percent non-Hispanic multiracial, percent renters, percent single-parent households, percent foreign born, percent separated, divorced, or widowed, percent males aged 15 to 29, percent unaffiliated youth, percent moving residence in previous year, percent with a cognitive, ambulatory, independent living, self-care, vision, or hearing difficulty (source: ACS; time frame: annual estimates); population (US Census; annual); alcohol outlet density and social organization density (a proxy for social cohesion) (US Census Zip Code Business Patterns; annual); a validated proxy for firearm ownership constructed from percent firearm suicides and hunting licenses per capita (California Vital Records and California Department of Fish and Wildlife, and US Census; annual)\(^{23}\); mean self-reported mentally unhealthy days per month (California Health Interview Survey; bi-annual); primary care providers per capita (Health Resources and Services Administration Area Resource File; annual); unemployment (Bureau of Labor Statistics; monthly); and average temperature and average precipitation (WestMap Climate Analysis PRISM Climate Mapping Program; monthly).

We excluded covariates that were excessively correlated with other covariates in the control set. These were: median income, percent below poverty, racial/ethnic composition, percent English speaking, percent veterans, marital status composition, educational composition, percent employed, percent searching for work, percent living alone, population density, average number of physically unhealthy days in previous month, percent of suicides completed with firearms, alcohol outlet count, health food establishments count and density, social organizations count. We used the Missouri Census Data Center Geographic Correspondence Engine as needed to crosswalk covariate values from measured geographic units to CPUMAs.\(^{24}\) To ensure correct temporal ordering, monthly covariates were assigned for the month prior to injury/survey.

**Selection of control time periods**

We drew controls from two time periods: exactly 30 days before the case occurrence and exactly 30 days after (a “bidirectional” design;\(^ {25}\) see Appendix Figure 3 for depiction). By being as close in time as possible (in 30-day units) to the case, these control periods limit confounding by time-varying individual- and community-level factors and confounding by seasonal patterns and secular trends common to the exposure and outcome. In addition, although we hypothesized that any effects of within-community variation in violence on self-harm would be short-term, longer-lasting effects are possible. The symmetry of bidirectional controls may balance any carryover in exposure effects between cases and controls and thus is likely to generate conservative estimates of association. Relative to control time periods with longer lags or leads, the 30-day unit also minimizes the risk of exposure misclassification due to residential moves. In addition, previous simulation studies assessing the optimal source of controls in case-crossover studies with temporal patterning suggest that the bidirectional design with controls drawn as close in time as possible to the case provide superior control of confounding by trends and seasonality, compared to unidirectional controls, controls with longer lags or leads, or controls drawn from all or a random selection of control periods.\(^ {25-28}\)
Figure 3: Case and control exposure periods for case-crossover and case-control designs

Statistical analysis

We used conditional logistic regression to estimate the association between a community violence residual of 1 per 100,000 versus 0 (approximately the 80th percentile versus the median/expected level of community violence) and self-harm, accounting for the matched data structure. To allow for potential non-linearity, we considered linear, quadratic, and cubic terms for continuous variables (exposure, covariates) that improved model fit, a priori optimizing the Akaike Information Criterion (AIC). Exposures modeled with multiple terms were combined into a single summary measure of association. The intraclass correlation coefficients (ICC) for the occurrence of fatal and nonfatal self-harm across CPUMAs29 were negligible (ICC<0.001), so no modeling of within-community clustering was necessary.

We stratified analyses by self-harm type (fatal versus nonfatal) because determinants of self-harm differ by type.1,13,14 We also considered effect measure modification by age group, gender, race/ethnicity, and urbanicity, because the risk of self-harm differs substantially across these factors, and we hypothesized that demographic subgroups defined by these characteristics would respond differently to community violence.

We conducted analyses using R.30 Statistical code is provided below. This study was approved by the State of California and University of California at Berkeley Committees for the Protection of Human Subjects.

Sensitivity analyses

In sensitivity analyses, we first considered analyses using shorter (15-day) and longer (45-day) time units for exposure assessment, because the critical time period of exposure is unknown. Second, our approach assumes that exposure is uncorrelated across periods.27 Thus, we assessed the autocorrelation function (ACF) of each violence residual series and tested restriction to CPUMAs with an absolute value of the ACF<0.2 at one lag.31

Third, to confirm that null findings were not due to misspecification in the analytic approach, we applied the main design and analysis to an established risk factor for fatal self-harm–economic downturns.32 Specifically, we used the change in unemployed individuals (i.e., disemployments) per population, a metric based mainly on unemployment insurance claims, in the month prior to case occurrence, from
the California Employment Development Department, using a population-proportional geographic crosswalk from counties to CPUMAs.

**Case-control study**

As a final sensitivity analysis, we conducted a case-control study, drawing population-based controls from California resident participants of the ACS, matched to cases on community, age, gender, and race/ethnicity. In the case-crossover design, using controls drawn from after case occurrence is a violation of the Study Base Principle for fatal outcomes, because the case has died in the index time and thus is not a valid control eligible to become a case 30 days later. The case-control addresses these concerns; however, it does not control unmeasured time-invariant individual-level confounders.

The ACS is a continuous, national survey conducted by the US Census Bureau. It produces population-representative small-area estimates of demographic, economic, and social indicators, and serves as an efficient, existing, population-representative sampling frame from which to draw population-based controls. From 2005 to 2013, between 170,000 and 220,000 California residents participated in the ACS annually.

Consistent with previous research, we created a representative pseudo-population of California residents by duplicating each ACS record by the corresponding person weight and selected controls from this expanded ACS dataset. For statistical efficiency, ACS controls were matched to cases on confounders that are strongly associated with self-harm: gender, race/ethnicity, 5-year age group, and community. For this design, we assumed that controls were not also cases at the time they were selected as controls. This is reasonable, because self-harm was very rare (<0.5% in all strata).

Statistical analysis was identical to that in the main case-crossover study. See Confounder Assessment above for details on covariates controlled in case-control analyses. Cases with incomplete individual-level covariate data (2.8%) were excluded.

**Extended results**

There were 31,027 cases of fatal self-harm and 365,933 cases of nonfatal self-harm among adults aged 15 to 84 in California between March 1, 2005 and December 1, 2013, corresponding to crude annual rates of 12.2 per 100,000 and 144.2 per 100,000, respectively. Appendix Table 1 presents characteristics of study participants. Deviations from expected levels of community violence showed small positive associations with self-harm (median difference in within-community variation in violence of 0.016 for fatal self-harm and 0.018 for nonfatal). Across the 108 30-day time units in this study, communities experienced an average of 31 30-day time units with deviations from expected levels of community violence greater than 1 per 100,000.
Table 1: Sample sizes and distribution of within-community variation in violence across outcomes

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Fatal Cases</th>
<th>Fatal Controls</th>
<th>Nonfatal Cases</th>
<th>Nonfatal Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>31,027</td>
<td>62,054</td>
<td>365,933</td>
<td>731,866</td>
</tr>
<tr>
<td>Within-community violence (median [IQR])</td>
<td>-0.02</td>
<td>-0.03</td>
<td>-0.01</td>
<td>-0.03</td>
</tr>
<tr>
<td>Characteristics of cases</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (median [IQR])</td>
<td>48 (34, 59)</td>
<td></td>
<td>31 (21, 45)</td>
<td></td>
</tr>
<tr>
<td>Gender (% female)</td>
<td>23.3</td>
<td></td>
<td>58.5</td>
<td></td>
</tr>
<tr>
<td>Race/ethnicity (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White, NH</td>
<td>69.5</td>
<td>58.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black, NH</td>
<td>4.0</td>
<td>8.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian, NH</td>
<td>0.5</td>
<td>0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian or Pacific Islander, NH</td>
<td>8.3</td>
<td>4.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other/multi-race, NH</td>
<td>1.4</td>
<td>3.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>16.3</td>
<td>24.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IQR: interquartile range. NH: non-Hispanic. Within-community variation in violence was defined as the difference between the observed rate and the Kalman smoothed rate of community violence, across 30-day time units spanning 2005 to 2013.

In the main analysis, 30-day periods with higher-than-expected levels of community violence were associated with a fatal self-harm odds ratio (OR) of 1.004 (95% confidence interval [CI]: 0.997, 1.011) and a nonfatal self-harm OR of 1.005 (CI: 1.003, 1.007).

Sensitivity analyses using shorter (15-day) and longer (45-day) time units for exposure assessment were consistent with the main analysis for fatal self-harm, and showed slightly stronger associations for nonfatal self-harm (15-day OR: 1.007 [CI: 1.001, 1.012]; 45-day OR: 1.011 [CI: 1.003, 1.020]; Appendix Figure 4).

Figure 4: Adjusted odds ratios for sensitivity analyses of fatal and nonfatal self-harm associated with within-community variation in violence, California, 2005-2013
OR: Odds ratio. Bars indicate 95% confidence intervals. Point shapes indicate the width of the time units used to assess rates and deviations from expected levels of community violence across 2005 to 2013. The main analysis used 30-day time units. Within-community variation in violence was defined as the difference between the observed rate and the Kalman smoothed rate of community violence, across 30-day time units spanning 2005 to 2013.

Results from the case-control analysis with controls drawn from the American Community Survey were consistent with the main analysis. Specifically, 30-day periods with higher-than-expected levels of community violence were associated with a fatal self-harm OR of 1.007 (CI: 0.996, 1.017) and a nonfatal self-harm OR of 1.005 (CI: 1.003, 1.007).

Assessment of autocorrelation in the violence residuals after applying the Kalman smoother suggested that this approach successfully removed predictable temporal patterning, including autocorrelation, secular trends, and seasonality, from most series. There were 16 CPUMAs (15%) in which the absolute value of the autocorrelation at one lag was greater than 0.2 (Appendix Table 2). Exclusion of these CPUMAs did not alter the results (Appendix Table 3).

Table 2: Autocorrelation Functions (ACF) for Community Violence Exposure at Different Lags across Consistent Public Use Microdata Areas, California 2005-2013

<table>
<thead>
<tr>
<th>Lag in Days</th>
<th>Median</th>
<th>% &gt; 0.20 or &lt; -0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days</td>
<td>0.08</td>
<td>15%</td>
</tr>
<tr>
<td>60 days</td>
<td>0.07</td>
<td>4%</td>
</tr>
<tr>
<td>90 days</td>
<td>0.07</td>
<td>5%</td>
</tr>
<tr>
<td>360 days</td>
<td>0.10</td>
<td>13%</td>
</tr>
</tbody>
</table>

Table 3: Adjusted odds ratio for fatal and nonfatal self-harm associated with within-community variation in violence, California, 2005-2013

<table>
<thead>
<tr>
<th>Self-harm type</th>
<th>Odds Ratio</th>
</tr>
</thead>
</table>
Fatal
1.005 (0.997, 1.013)

Nonfatal
1.005 (1.002, 1.007)

Analysis restricted to CPUMAs with absolute value violence residual autocorrelation less than 0.2. Within-community variation in violence was defined as the difference between the observed rate and the Kalman smoothed rate of community violence, across 30-day time units spanning 2005 to 2013.

In the positive control analysis, higher-than-expected disemployment (4 per 10,000 versus 0, or approximately the 80th percentile versus the median) was positively associated with fatal self-harm (OR 1.015, 95% CI: 1.003, 1.027) but not nonfatal self-harm (OR 1.000, 95% CI: 0.997, 1.003).

Appendix Figures 5-8 present results by demographic subgroup. Although there was little discernable variation by subgroup, higher-than expected levels of community violence were associated with elevated odds (though imprecise) for fatal self-harm among those aged 15-24 and sub-urban residents and for nonfatal self-harm among those aged 40-54 and urban residents. There was little meaningful variation by race/ethnicity or gender.

**Figure 5: Adjusted odds ratio for the association of within-community variation in violence with fatal and nonfatal self-harm, by age group**

Within-community variation in violence was defined as the difference between the observed rate and the Kalman smoothed rate of community violence, across 30-day time units spanning 2005 to 2013.
Figure 6: Adjusted odds ratio for the association of within-community variation in violence with fatal and nonfatal self-harm, by gender

Within-community variation in violence was defined as the difference between the observed rate and the Kalman smoothed rate of community violence, across 30-day time units spanning 2005 to 2013.
Figure 7: Adjusted odds ratio for the association of within-community variation in violence with fatal and nonfatal self-harm, by race/ethnicity

Within-community variation in violence was defined as the difference between the observed rate and the Kalman smoothed rate of community violence, across 30-day time units spanning 2005 to 2013. Race groups are non-Hispanic unless otherwise noted.
Within-community variation in violence was defined as the difference between the observed rate and the Kalman smoothed rate of community violence, across 30-day time units spanning 2005 to 2013. Urbanized Area (Metropolitan): densely settled core with at least 50,000 residents. Urban Cluster (Micropolitan): densely settled core with at least 10,000 residents and less than 50,000 residents. Small town/rural: densely settled area with less than 10,000 residents or rural area.

Extended discussion

To our knowledge, this is the first study to assess whether increases from expected levels of community violence were associated with greater fatal and nonfatal self-harm.\textsuperscript{11} We utilized comprehensive population-based data from California and a case-crossover design, which caters to brief exposures and transient changes in risk for acute-onset outcomes like those in this study, enhances control of unmeasured individual confounders such as genetics and family history, and reduces concerns related to structural confounding and control-selection bias.

Results suggested that higher within-community variation in violence was not meaningfully associated with greater self-harm. This may be because we only assessed self-harm associated with single deviations from expected levels of community violence, and therefore did not capture the entire relationship between community violence and self-harm. Previous research suggests that overall levels of community violence are a risk factor for self-directed violence\textsuperscript{4-9,11} and are likely stronger determinants of self-harm than within-community variation. Higher-than-expected community violence is associated with greater anxiety or substance use.\textsuperscript{36} Our work suggests these symptoms may not translate to self-harm hospital visits or deaths, although mediation was not formally assessed. Larger or
different types of increases in community violence (e.g. mass shootings, wars, or terrorism) may produce bigger effects (for example\textsuperscript{37,38}).

The exposure measure in this study—30-day deviations from expected levels of community violence in the CPUMA of residence—may not be the optimal characterization. Previous research has identified strong associations between long-term CPUMA-level community violence and self-harm,\textsuperscript{30} but the most salient time frame and geographic scope for elevated risk remain uncertain. This is an area for future research. In addition, our measure of community violence only captures events sufficiently serious to result in deaths or hospital visits. Other types of community violence (e.g. assaults not resulting in hospital visits) may also be relevant to self-harm. Crime data can capture incidents not resulting in hospital visits or deaths, but may contain patterns that are artefacts of inconsistent reporting practices between jurisdictions and over time.\textsuperscript{39,40} Victimization surveys are also available, and capture perceptions of violence, but only include those who have not died from self-harm. Surveys may also introduce same-source bias, in which error in self-report of both community violence and self-harm may be associated, for example due to respondent temperament. Our objective measure of community violence is likely to be strongly correlated with frequency of experiences of direct injury and witnessing violence reported by residents,\textsuperscript{41,42} but avoids this concern.

The economic downturns positive control was associated with fatal self-harm. This finding is reassuring, because it suggests that social contextual stressors at the CPUMA-month level are relevant to suicide risk for this study population and can be captured with our study design and analytic approach. The finding that economic downturns were not associated with nonfatal self-harm was not surprising, as evidence that economic downturns are linked to nonfatal self-harm is less consistent, particularly for women who make the majority of nonfatal cases,\textsuperscript{43,44}

In using control periods after case occurrence, we must assume that past outcomes do not affect future exposure. This is a reasonable assumption because self-harm is uncommon and rarely publicized, and the exposure is characterized as unpredictable patterning in community violence. In addition, for fatal outcomes, using controls drawn from after case occurrence is a violation of the Study Base Principle\textsuperscript{33} because the case has died in the index time and thus is not eligible to become a case 30 days later. Nevertheless, for short-term exposures and outcomes, post-case exposure is a reasonable proxy for exposure in the study base\textsuperscript{45} the experience of the cases had they instead been controls , and excluding post-outcome referent periods may result in even greater confounding or selection bias.\textsuperscript{25–27} Concerns regarding the use of dead controls are also addressed by the ACS case-control sensitivity analysis, which found results consistent with the main analysis. Using longer time units for exposure assessment (e.g. 45- or 60-day) may reduce concerns about carryover between cases and controls, but may also make estimates more vulnerable to unmeasured time-varying confounders such as individual distress or hopelessness, regional closing of a major employment center, or other compositional or structural changes of communities. Explicit measurement and control for these factors, along with more research on the critical time periods of increased risk for self-harm, would address these concerns in future research.

Higher-than-expected levels of community violence were associated with a slightly elevated odds of fatal and nonfatal self-harm in some subgroups considered in this study. Despite the large sample size, it is difficult to distinguish between differences that are substantive versus those due to variation in sample sizes and hence, precision, of different subgroup estimates. If the lack of effect measure modification is real, this finding is surprising, because self-harm has diverse causes and forms in
different subgroups.\textsuperscript{1,13,14} It may be that these subgroups have different levels of predisposing risk factors, but respond similarly to stressors such as within-community increases in violence.

This study builds on\textsuperscript{36} and adds nuance to\textsuperscript{4–11} previous research on the harms of community violence by characterizing exposure as a dynamic “state” risk factor (i.e. assessing whether within-community fluctuations in violence risk state affect self-harm), rather than a static “status” risk factor (i.e., assessing whether between-community variation in violence affect self-harm).\textsuperscript{46} This is an important distinction, because understanding risk states is critical for moving beyond risk assessment to understanding and achieving risk reduction.\textsuperscript{46} Our findings suggest that 30-day deviations from expected levels of community violence were not relevant to self-harm in the study population, but other forms of “state”-type variation (e.g. mass shootings or level shifts caused by interventions) may still be important. Future research examining the impacts of violence prevention programs aiming to limit increases in community violence (e.g. Los Angeles Summer Night Lights\textsuperscript{47} or summer employment programs\textsuperscript{48}) may provide more conclusive evidence.

The methodology used in this study provides a template for future research. We used large, existing databases from California to study social ecological drivers of self-harm, an outcome for which previous research has been limited by small sample sizes. Recent increases in the size, scope, and availability of large health data facilitate epidemiologic studies that combine different data sources in efficient ways and leverage the high degree of geographic and temporal precision available in these data. This study is one application in which such data are particularly useful—the case of population-based case-crossover studies with transient ecological exposures. Our findings suggest that within-community variation in violence is not meaningfully associated with higher levels of self-harm. Future research should further assess the effects of cumulative community violence, critical time periods of increased risk of self-harm, and the impacts of violence prevention programs and policies.

References


**R Code**

```r
# Code
```

```
# R Code

## Load packages used throughout

```r
library(forecast)
library(KFAS)
library(doBy)
library(readstata13)
library(reshape2)
library(foreign)
library(DataCombine)
library(data.table)
library(multcomp) # for glht()
library(survival) # for clogit
```

## Create controls from cases

```r
rm(list=ls())
load(file="cases.rdata")

# Controls are cases but 30 days prior and 30 days after
controls1 <- controls2 <- cases

# month prior
controls1$day <- as.numeric(format(controls1$date,"%d"))
controls1$date <- controls1$date - 30
controls1$ctype <- -1

# year after
controls2$day <- as.numeric(format(controls2$date,"%d"))
controls2$date <- controls2$date + 30
controls2$ctype <- 1

# Combine
controls <- rbind(controls1, controls2)

summary(controls$date)
table(is.na(controls$date)) # no missingness
head(controls)
tail(controls)

# update month and year; remove day
controls$year <- as.numeric(as.character(format(controls$date,"%Y")))
controls$month <- as.numeric(as.character(format(controls$date,"%m")))
controls <- controls[,names(controls) != "day"]

summary(controls)
table(table(controls$id))

# Save
save(controls, file="controls.rdata")

# Combine
cases$case <- 1
controls$case <- 0
cases$ctype <- 0

matched.data <- rbind(cases, controls)
```
```r
code here...
```
head(data)
summary(data)

###
# Loop through each index day (the day in Jan 2014 on which the series is based - i.e. index_day is the unique identifier for the time series)
###
all_results <- NULL
for (index_day in 1:30) {
  print(paste0("**************************************** Index Day: ", index_day))
  
  ## Assign 30-day unit numbers according to index day
  temp <- data
  temp$time_unit_num <- NA
  
  # Initialize
  end_date <- as.Date(paste0("1/", index_day, "/2014"), ",%m/%d/%Y")
  
  for (i in max_number_time_units:1) {
    if (i %% 10 == 1) print(paste0('time_unit_num: ', i))
    # Take the end date and backtrack 30 days to identify the start date
    start_date <- end_date-30
    # Assign time unit number
    temp$time_unit_num[temp$date>start_date & temp$date<=end_date] <- i
    # Set start date to end date to increment backwards in time and initialize for the next loop
    end_date <- start_date
  }
  
  ## Create numerator data, merge, and clean
  temp$month_calendar <- as.numeric(format(temp$date, ",%m")
  nums <- summaryBy(count + month_calendar + year ~ time_unit_num + cpuma, FUN=c(sum, mean, min, max), data=temp)
  nums$month_calendar <- nums$year <- NA
  
  # If the year is uniform across time units, use it:
  nums$year[ nums$year.min==nums$year.max] <- nums$year.min[ nums$year.min==nums$year.max]
  # If the year is not uniform across time units, take the year that is dominant:
  nums$year[ nums$year.min!=nums$year.max] <- round(nums$year.mean[ nums$year.min!=nums$year.max])
  
  # If the year is uniform across time units, take the month that is dominant
  nums$month_calendar[ nums$year.min==nums$year.max] <- round(nums$month_calendar.mean[ nums$year.min==nums$year.max])
  # If the year is not uniform across time units, take the month that is dominant, but not by averaging.
  nums$month_calendar[ nums$year.min!=nums$year.max & nums$month_calendar.min==1 &
                      nums$month_calendar.max==12 & nums$month_calendar.mean>=6.5] <- 12
  nums$month_calendar[ nums$year.min!=nums$year.max & nums$month_calendar.min==1 &
                      nums$month_calendar.max==12 & nums$month_calendar.mean< 6.5] <- 1

  table(nums$year)
  table(nums$month_calendar)

  nums <- nums[, c('cpuma', 'time_unit_num', 'count.sum', 'month_calendar', 'year')]
  names(nums) <- c('cpuma', 'time_unit_num', 'count', 'month_calendar', 'year_calendar')

  # Calendar year and calendar month here are just approximations to be used for merging on the denominator data.
# Make sure dataset will be square
set <- expand.grid(cpuma = sort(unique(nums$cpuma)), time_unit_num = 1:max_number_time_units)
an <- merge(nums, set, by = c('cpuma','time_unit_num'), all=T)

# Merge numerators and denominators
an <- merge(an, denom, by=c('cpuma','year_calendar','month_calendar'), all.x=T)

# Make rates
an$rate <- an$count / an$pop * 100000

# Drop any (first and/or last) 'time units' which are partial - these cannot be used.
if (index_day %in% 1:12) an <- an[!an$time_unit_num %in% c(1,2, max_number_time_units),]
if (index_day %in% 13:30) an <- an[!an$time_unit_num %in% c(1,   max_number_time_units),]

## Apply Kalman Filter
an <- an[order(an$cpuma, an$time_unit_num),]

# Fit Kalman and identify shocks and residuals
an$residual <- an$shock <- NA
for (i in sort(unique(an$cpuma))) {
  print(paste('cpuma: ',i))
v <- as.ts(an$rate[an$cpuma==i], frequency=12)
res <- kfs.out(v)
an$residual[an$cpuma==i] <- res[[1]]
an$shock   [an$cpuma==i] <- res[[2]]
}

an$index_day <- index_day

# Store results
all_results[[index_day]] <- an
}

final <- do.call(rbind, all_results)
write.csv(final, "Kalman_residuals_day-specific_30-day_forcedSeasonal.csv", row.names=F)

#####################################################
# Merge exposure onto outcome data
#####################################################
rm(list=ls())

# Bring in matched data
load(file="matched.data.rdata")
summary(matched.data)

# Bring in data
exp <- read.csv("Kalman_residuals_day-specific_30-day_forcedSeasonal.csv", stringsAsFactors=F)

# Unique identifiers
dim(exp)
dim(unique(exp[,c('cpuma','index_day','time_unit_num')]))

# In outcome data, create variables for merging
# i.e. index_day is the unique identifier for the time series
max_number_time_units <- 111

for (index_day in 1:30) {
end_date <- as.Date(paste0("1/",index_day,"/2014"), "%m/%d/%Y")

for (i in max_number_time_units:1) {
  start_date <- end_date-30
}
# If the event date falls on the last day of this time window such that event date = end_date,
# assign this index day and time unit number
matched.data$time_unit_num[matched.data$date==end_date] <- i
matched.data$index_day[matched.data$date==end_date] <- index_day

# Set start date to end date to increment backwards in time and initialize for the next loop
end_date <- start_date
}
}

# If a self-harm event occurred on May 25, the "same time unit" would be May 11 - May 25. This is what we want for exposure.
# If a self-harm event occurred on May 1, the "same time unit" would be April 17 - May 1.
# So we do not want to lag 1 time unit.
dim(matched.data)
data <- merge(matched.data, exp, by=c('cpuma', 'index_day', 'time_unit_num'), all.x=T)
dim(data)

save(data, file="matched.data.exposure.forcedSeasonal.rdata")

# Merge community-level covariates of varying geographies and time units
rm(list=ls())
# Analysis data
load(file="matched.data.exposure.forcedSeasonal.rdata")

summary(data)

###
# Yearly cpuma-level covs
###
# Covariates
covs <- read.csv("compiled.covs.csv", stringsAsFactors = F)

# Exposures will go back up to 30 days, so in theory covariates should be at least 30 days lagged.
# For each event, assing data from the same year.
data$covs_year <- data$year
table(data$year, data$covs_year)

# Merge
dim(data)
data <- merge(data, covs, by=c("cpuma", "covs_year"), all.x=T)
dim(data)

###
# Monthly covs
###
# Unemployment
unemp <- NULL
for (y in 2005:2013) {
temp <- read.csv(paste0("BLS ", y, ", County Unemployment.csv"), stringsAsFactors = F)
temp <- temp[temp$desc_short=="Unemployment Rate", c('county', 'year', 'month', 'monthly_unemp')]
unemp <- rbind(unemp, temp)
}

# Precip
precip <- read.csv("CA-PRISM-county-pcp.txt")

# Temp
avgt <- read.csv("CA-PRISM-county-avgt.txt")
dim(unemp);dim(precip)
monthly_covs <- merge(unemp, precip, by=c("county","year","month"), all=T)
dim(monthly_covs)
summary(monthly_covs)

dim(monthly_covs)
monthly_covs <- merge(monthly_covs, avgt, by=c("county","year","month"), all=T)
dim(monthly_covs)
summary(monthly_covs)

## Crosswalk county to 2010 puma
cross <- read.csv("E:/Jen/Data/geocoding and denominators/denominators/8 - CPUMA/draft/county_to_PUMA2010.csv", stringsAsFactors = F)
dim(monthly_covs)
monthly_covs <- merge(monthly_covs, cross, by="county", all.x=T)
dim(monthly_covs) # should have more here because we've expanded each county row into multiple puma10 rows

# Checks
head(monthly_covs)
summary(monthly_covs)

# Check for 0 allocation factors
head(monthly_covs[monthly_covs$afact==0,]) # none

# Check for duplicates
nrow(monthly_covs)
nrow(unique(monthly_covs[,c('year','month','PUMA10')]))

monthly_covs <- monthly_covs[order(monthly_covs$PUMA10, monthly_covs$year, monthly_covs$month),]
head(monthly_covs[duplicated(monthly_covs[,c('year','month','PUMA10')]),])
tail(monthly_covs[duplicated(monthly_covs[,c('year','month','PUMA10')]),])

# Collapse across duplicates from different counties allocating to the same puma
monthly_covs2 <- data.table(monthly_covs)
monthly_unemp <- monthly_covs2[,list(monthly_unemp = weighted.mean(monthly_unemp, pop10)),by=c("PUMA10","year","month")]
avgt <- monthly_covs2[,list(avgt = weighted.mean(avgt, pop10)),by=c("PUMA10","year","month")]
precip <- monthly_covs2[,list(precip = weighted.mean(precip, pop10)),by=c("PUMA10","year","month")]
dim(monthly_unemp); dim(avgt)
monthly_covs3 <- merge(monthly_unemp, avgt, by=c("PUMA10","year","month"))
dim(monthly_covs3)
monthly_covs3 <- merge(monthly_covs3, precip, by=c("PUMA10","year","month"))
dim(monthly_covs3)
monthly_covs3 <- as.data.frame(monthly_covs3)
nrow(monthly_covs3)
nrow(unique(monthly_covs3[,c('year','month','PUMA10')])) # good
head(monthly_covs3)
summary(monthly_covs3)
rm(avgt, cross, monthly_covs, monthly_covs2, monthly_unemp, precip, temp, unemp, y)

## Crosswalk 2010 puma to cpuma
cross <- read.csv("E:/Jen/Data/geocoding and denominators/denominators/8 - CPUMA/draft/CPUMA0010_PUMA2010_components.csv", stringsAsFactors = F)
dim(monthly_covs3)
monthly_covs3 <- merge(monthly_covs3, cross, by="PUMA10", all.x=T)
dim(monthly_covs3) # Same number of rows.

# Check for duplicates
nrow(monthly_covs3)
nrow(unique(monthly_covs3[,c('year','month','CPUMA0010')]))

monthly_covs3 <- monthly_covs3[order(monthly_covs3$CPUMA0010, monthly_covs3$year, monthly_covs3$month),]

head(monthly_covs3[duplicated(monthly_covs3[,c('year','month','CPUMA0010')])])
tail(monthly_covs3[duplicated(monthly_covs3[,c('year','month','CPUMA0010')])])

# Collapse across duplicates from different counties allocating to the same puma
monthly_covs3 <- summaryBy(monthly_unemp + avgt + precip ~ year + month + CPUMA0010, 
data=monthly_covs3, FUN = function(x) mean(x, na.rm=T))

names(monthly_covs3) <- gsub("\function(x) mean(x, na.rm = T)","", names(monthly_covs3), fixed=T)

head(monthly_covs3)
summary(monthly_covs3)

table(monthly_covs3$year, monthly_covs3$month)

monthly_covs3 <- monthly_covs3[order(monthly_covs3$CPUMA0010, monthly_covs3$year, monthly_covs3$month),]

nrow(monthly_covs3)
nrow(unique(monthly_covs3[,c('year','month','CPUMA0010')]))

length(unique(monthly_covs3$CPUMA0010))

rm(cross)

## Merge

# Create merge vars. Exposures will go back up to 30 days, so in theory covariates should be at least 30 days lagged.
# For monthly, assign the covariate from 1 month prior.
# Assign the same month for Jan 2005 since we don't have covs for 2004.

# Assign lagged t for merging
monthly_covs3$t_covs <- (monthly_covs3$year-2005)*12 + monthly_covs3$month
names(monthly_covs3)[names(monthly_covs3)=="CPUMA0010"] <- "cpuma"

data$t_covs <- (data$year-2005)*12 + data$month - 1
data$t_covs[data$t_covs<0] <- 1
dim(data)
data <- merge(data, monthly_covs3, by=c('cpuma','t_covs'), all.x=T)
dim(data)

rm(monthly_covs3)

head(data)
summary(data)

## Save
save(data, file="matched.data.exposure.forcedSeasonal.covs.rdata")

#############################################################
# Analyze
#############################################################

rm(list=ls())
fatal <- F
fat <- ifelse(fatal, "fatal","nonfatal")

# Load data
load(file="matched.data.exposure.forcedSeasonal.covs.rdata")
Subset to relevant data
fatal vs. nonfatal
dim(data)
data <- data[data$fatal==as.numeric(fatal),]
dim(data)

Covariates
# No need to adjust for any individual covariates - time-invariant ones are controlled by the
design,
# and time-varying ones (insurance type, years in US, education, marital status, PLS, veteran,
etc.)
# are still only have measured at one time point.
# No need to adjust for time-invariant community-level confounders - controlled by design.
# Don’t adjust for year - that would be a practical positivity violation and we are already
removing the temporal trend in the exposure.

# make sure matching strata is a factor
data$id <- as.factor(data$id)

# Bring in names of final community-level covariates - use continuous ones.
# Categorical ones from this file are year, race, month, none of which we want to control for.
load(file="final.covs.rdata")

# Keep only variables we're using
data <- data[,c("fatal","year", "month", "cpuma", "date", "agegrp", "sex",
"case","id","residual", "final.covs.cont")]

# exposure
exposure.names <- "residual"

# Confirm that there is a case and 1-2 controls in each stratum
table(table(data$id, data$case))
table(table(data$id))

## Formulas
# include at least linear exposure as min formula:
min.formula <- as.formula(paste0("case ~ strata(id) + ", paste0(exposure.names, collapse="+")))

unadj.formula <- as.formula(paste0("case ~ ",
paste0(c(exposure.names), collapse="+"), " + ", # linear terms
"I(" , paste0(c(exposure.names), collapse=" ^2) + I("), "^2) + " , #
quadratic terms
"I(" , paste0(c(exposure.names), collapse=" ^3) + I("), "^3) +
strata(id")") )

adj.formula <- as.formula(paste0("case ~ ",
paste0(c(exposure.names, final.covs.cont), collapse="+"), " + ", # linear
"I(" , paste0(c(exposure.names, final.covs.cont), collapse=" ^2) + I("), "^2) + " , #
quadratic
"I(" , paste0(c(exposure.names, final.covs.cont), collapse=" ^3) + I("), "^3) +
strata(id")") )

## Run mode
fit.step <- step(clogit(min.formula, data=data), scope = adj.formula, direction = "forward")
summary(fit.step)
save(fit.step, file =
paste0("clogit.overall.both.forcedSeasonal.modelObj.p3.step",".fat",".rdata"))

# Calculate parameters using glht to combine terms
results <- data.frame(est = NA, lower = NA, upper = NA, n=nrow(data))
k <- rep(0,length(names(coef(fit.step))))
k[ names(coef(fit.step)) %in% c('residual', 'I(residual^2)', 'I(residual^3)') ] <- 1
k <- matrix(k,1)
results[, c('est','lower','upper')] <- exp(confint(glht(fit.step, linfct=k))$confint)[,1:3]
round(results, 3)

# Save results
write.csv(results, paste0("clogit.overall.both.forcedSeasonal.p3.step.", fat, ".csv"), row.names=F)

########################################################################
# END
########################################################################