Residential exposure to petroleum refining and stroke in the southern United States

Honghyok Kim\textsuperscript{1,7,∗}, Natalia Festa\textsuperscript{2,3}, Kate Burrows\textsuperscript{1}, Dae Cheol Kim\textsuperscript{5}, Thomas M Gill\textsuperscript{6,8} and Michelle L Bell\textsuperscript{1,8}

\textsuperscript{1}School of the Environment, Yale University, New Haven, CT 06511, United States of America
\textsuperscript{2}Veterans Affairs (VA) Office of Academic Affiliations through the VA/National Clinician Scholars Program and Yale University, New Haven, CT, United States of America
\textsuperscript{3}National Clinician Scholars Program, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, United States of America
\textsuperscript{4}The Institute at Brown University for Environment and Society, Providence, RI, United States of America
\textsuperscript{5}Graduate School of Public Health, Seoul National University, Seoul, Republic of Korea
\textsuperscript{6}Department of Internal Medicine, Yale School of Medicine, New Haven, CT, United States of America
\textsuperscript{7}Contributed equally to this manuscript.
\textsuperscript{8}Senior authors.
∗Author to whom any correspondence should be addressed.
E-mail: honghyok.kim@yale.edu

Keywords: stroke belt, oil industry, petroleum refinery, environmental pollution, environmental justice, small-area variation in stroke prevalence

Abstract

The southern United States (US) sustains a disproportionate burden of incident stroke and associated mortality, compared to other parts of the US. A large proportion of this risk remains unexplained. Petroleum production and refining (PPR) is concentrated within this region and emits multiple pollutants implicated in stroke pathogenesis. The relationship between residential PPR exposure and stroke has not been studied. We aimed to investigate the census tract-level association between residential PPR exposure and stroke prevalence for adults ($\geq 18$ years) in seven southern US states in 2018. We conducted spatial distance- and generalized propensity score-matched analysis that adjusts for sociodemographic factors, health behavioral factors, and unmeasured spatial confounding. PPR was measured as inverse-distance weighted averages of petroleum production within 2.5 km or 5 km from refineries, which was strongly correlated with measured levels of sulfur dioxide, a byproduct of PPR. The prevalence of self-reported stroke ranged from 0.4% to 12.7% for all the census tracts of the seven states. People with low socioeconomic status and of Hispanic ethnicity resided closer to petroleum refineries. The non-Hispanic Black population was exposed to higher PPR, while the non-Hispanic White population was exposed to lower PPR. Residential PPR exposure was significantly associated with stroke prevalence. One standard deviation increase in PPR within 5 km from refineries was associated with 0.22 (95% confidence interval: 0.09, 0.34) percentage point increase in stroke prevalence. PPR explained 5.6% (2.4, 8.9) of stroke prevalence in the exposed areas. These values differed by states: 1.1% (0.5, 1.7) in Alabama to 11.7% (4.9, 18.6) in Mississippi, and by census tract-level: 0.08% (0.03, 0.13) to 25.3% (10.6, 40.0). PPR is associated with self-reported stroke prevalence, suggesting possible links between pollutants emitted from refineries and stroke. The increased prevalence due to PPR may differ by sociodemographic factors.

1. Introduction

Globally, stroke affects an estimated 7.8 million adults and of neurological diseases, has the highest absolute disability adjusted life-years [1]. Its direct and indirect costs are estimated at $103.5 billion yr$^{-1}$ in the United States (US) [2]. Despite national declines in age-standardized stroke incidence and mortality in the US, geographic disparities in stroke outcomes have persisted [3]. Residence within the southern US
confers an increased risk of stroke incidence and mortality, which is most concentrated in the Lower Mississippi River Basin (LMRB) [3–5]. These risks are further heightened for residents of socioeconomically disadvantaged neighborhoods, racial/ethnic minority subpopulations, and rural residents [3, 6].

The mechanisms underpinning geographic disparities in stroke outcomes remain only partly elucidated [7]. Despite a presumptive contribution by regional differences in the prevalence and management of vascular risk factors, residents of the southern US have similar Framingham stroke risk scores to residents of the other geographic areas [8, 9]. Socioeconomic conditions are estimated to explain a substantial share of this excess geographic burden [10, 11]. Many novel contextual factors have been studied as explanatory mechanisms for entrenched geographic and sociodemographic disparities, but have not fully explained such differences [12, 13]. Environmental contributors to regional and small-area geographic disparities in stroke outcomes remain understudied.

The geographic concentration of economic sectors, and their associated byproducts, is an underexplored, plausible risk factor for stroke. Byproducts of petroleum production and refining (PPR) include a mixture of pollutants that may impact the quality of adjacent air, soil, and potable water in residential areas (figures 1 and S1) [14]. In particular, air pollution from refineries includes pollutants that have been implicated in stroke pathogenesis, such as particulate matter (PM), sulfur dioxide (SO$_2$), nitrogen oxides (NO$_x$), polycyclic aromatic hydrocarbons (PAHs), and volatile organic compounds (VOCs) [15–17]. Approximately two-thirds of PPR in the US takes place within the third Petroleum Administration for Defense District (PADD-3) [18]. The states comprising PADD-3 (Alabama, Arkansas, Louisiana, Mississippi, New Mexico, and Texas) subsume the LMRB and overlap other states with elevated stroke incidence and mortality. We hypothesize that residential exposure to the byproducts of PPR may explain a proportion of entrenched regional and small-area disparities in stroke outcomes.
stroke risk. To the best of our knowledge, prior epidemiological studies have not evaluated the association between residential exposure to PPR and stroke outcomes.

In this study, we investigated possible links between residential PPR exposure, due to petroleum refineries and stroke in the southern US (figure 1).

2. Methods

2.1. Study population

We restricted our study population to states within PADD-3, the region responsible for the majority of PPR in the US [18]. These states include Alabama, Mississippi, Louisiana, Arkansas, Texas, and New Mexico. We also included Oklahoma (PADD-2) due to a concentration of PPR along its border with Texas.

2.2. Health outcomes

We obtained census tract-level stroke prevalence for the year 2018, using the Centers for Disease Control and Prevention (CDC) PLACES dataset. We evaluated the prevalence of self-reported stroke within each census tract among respondents aged \( \geq 18 \) years who report ever having been told by a doctor, nurse, or other health professionals that they have had a stroke. These prevalence estimates were originally derived from the Behavioral Risk Factor Surveillance System, which samples from a non-institutionalized population.

2.3. Classification of residential exposure to PPR

We obtained geocoded addresses for 59 petroleum refineries and their average production capacities for the years 2015–2017 using data from the US Energy Information Administration. We used two exposure metrics: (a) proximity to refineries and (b) inverse squared distance-weighted petroleum production. We assumed that these proxy measures could be related to multiple plausible exposure mechanisms, including air pollution mixtures, noise, water and soil contamination, and potential occupational exposures (if residents of refinery-adjacent municipalities are also employed within this setting) (figure 1).

The proximity to refineries was dichotomized based on pre-specified distances (\(<2.5 \text{ km}, 2.5–5 \text{ km}, \) or \(<5 \text{ km})\) from the centroid of a census tract to a petroleum refinery. We specified a consistent control group for each threshold distance, consisting of census tracts located at distances \(>5 \text{ km}\) and \(<11.1 \text{ km}\) from a refinery. The distance thresholds were specified \(a \ priori\) and informed by measured levels of \(\text{SO}_2\) near refineries and their patterns of decay with distance from smokestacks [19]. Selection of the control group was informed by visual inspection on spatial correlation of stroke prevalence using semi-variograms (figure S2) and by the distribution of stroke prevalence over proximity to refineries to adjust for unmeasured spatial confounding by spatial matching because census tracts within 11.1 km from a refinery have more spatial similarity (see section 2.6). If multiple refineries were located within the pre-specified distances, proximity was based on the closest refinery. All census tracts within 2.5 km or 5 km from a petroleum refinery included residential areas.

Although our distance-based measure of proximity to petroleum refineries is parsimonious, it does not account for the actual burden of pollution from refineries. Because the overall production capacity of a refinery may correspond to a proportionately higher burden of emissions, we also considered the potential petroleum production of each refinery, as a dimension of residential exposure. We defined production capacity using the reported atmospheric distillation capacity of each facility. We obtained average annual petroleum production for each PADD corresponding to states within our study area (PADDs 2 and 3). We estimated actual petroleum production for each refinery by approximating its fraction of annual oil production as proportionate to the share of the reported atmospheric distillation capacity represented by each facility, within its respective PADD. We applied an inverse distance weighting method with a pre-specified buffer distance from the centroid of census tracts (\(<2.5 \text{ km}\) and \(<5 \text{ km})\). We calculated an inverse squared distance-weighted average of actual petroleum production as a continuous exposure metric for each census tract as follows:

\[
X_c = \frac{\sum_{r \in R} \frac{\text{APP}_r \times \frac{1}{D_{c,r}^2}}{\sum_{r \in R} \frac{1}{D_{c,r}^2}}}{\sum_{r \in R} \frac{1}{D_{c,r}^2}}
\]

where \(X_c\) is the inverse squared distance-weighted average of actual petroleum production (barrels per day) at a census tract \(c\). \(D_{c,r}\) is the distance between the centroid of census tract \(c\) and a refinery \(r\). \(R\) is a set of refineries within a pre-specified buffer distance from the centroid of census tract \(c\). We calculated \(X_c\) values corresponding to each buffer distance threshold.

2.4. Specification of buffer distances for exposure classification

As mentioned above, we selected distance thresholds \(a \ priori\) based on patterns of decay with distance from smokestacks [19]. We further considered measured levels of ambient \(\text{SO}_2\) when establishing \(a \ priori\) buffer distance thresholds. Because crude petroleum contains a large amount of sulfur, \(\text{SO}_2\) is an expected byproduct of PPR. \(\text{SO}_2\) may be coupled with other pollutants generated from refineries, such as fine particulates, black carbon, \(\text{NO}_x\), \(\text{PAH}\), and \(\text{VOCs}\) [17]. Depending on the pollutant mixture and decay radius of its constituent pollutant classes, secondary pollutants can also be formed [17]. In part, because \(\text{SO}_2\) has been regulated for several decades as a Criteria Air Pollutant under the US Clean Air Act, its emission from automotive and residential activities is limited.
This implies that inverse squared distance-weighted APP may be correlated with SO$_2$ (as a rough marker of air pollutants generated from refineries) if petroleum production is related to the burden of air pollution, as the level from other sources is far less. We obtained measurements of SO$_2$ for 2015–2017 from the Environmental Protection Agency (EPA) Air Quality System (AQS). Figure 2 displays many SO$_2$ monitoring sites adjacent to refinery complexes. We estimated correlations between SO$_2$ and an inverse squared distance-weighted APP at AQS sites. $X_{AQS}$ is the inverse squared distance-weighted average of actual petroleum production (barrels per day) at a monitoring site AQS:

$$
X_{AQS} = \frac{\sum_{r \in R} \text{APP}_r \times \frac{1}{D_{AQS,r}}}{\sum_{r \in R} \frac{1}{D_{AQS,r}}} .
$$

$D_{AQS,r}$ is the distance between a monitoring site AQS and a refinery $r$. $R$ is a set of refineries within a pre-specified buffer distance from a monitoring site AQS. We calculated $X_{AQS}$ values corresponding to five buffer distance thresholds: $\leq 2.5$ km, $\leq 5$ km, $\leq 10$ km, $\leq 25$ km, and $\leq 50$ km. The correlation between SO$_2$ and $X_{AQS}$ was highest at buffer distances of 2.5 km and 5 km (Spearman’s correlation: 0.34, 0.26, respectively). Correlation decreased with the farther buffer distances ($\leq 10$ km, $\leq 25$ km, $\leq 50$ km) (figure S3). These correlations suggest that 2.5 km and 5 km buffer distances are more appropriate for exposure metrics.

2.5. Potential stroke risk factors and confounders

We considered a large set of potential confounders within each census tract, including sociodemographic characteristics and vascular risk factors. We obtained 5 years estimates from the American Community Survey (2014–2018) for age and sex distributions within each census tract. We considered the percentage of the population aged 18–19, 20–24, 25–44, 45–64, 65–84, and $\geq 85$ years, as well as the percentage of adult males and females ($\geq 18$ years). We also considered race and ethnicity, median household income, and percentages of the population who live under the federal poverty line or whose highest educational attainment is less than a high-school diploma. Data on self-reported current smoking, self-reported diabetes, self-reported hypertension, which are major vascular risk factors within the Revised Framingham Stroke Risk Score [20], and self-reported obesity [21], were obtained from CDC PLACES.

2.6. Statistical analysis

To evaluate cross-sectional census tract-level association between residential exposure to PPR and stroke prevalence, we conducted two analyses. The first analysis was spatial proximity- and propensity score (PS)-matching analysis for the proximity to refineries. The second analysis was spatial proximity- and generalized PS (GPS)-matching analysis for the weighted average of actual petroleum production ($X_c$). These two analyses were performed for each buffer distance threshold ($<2.5$ km, 2.5–5 km, and $<5$ km for the first analysis; and $<2.5$ km, and $<5$ km for the second analysis). All census tracts for which the centroid was within the buffer distance from a refinery site were included in an exposure group. Census tracts for which the centroid was $\geq 5$ km and $<11.1$ km from a refinery site was included in a control group.

The two analyses were conducted using a double-matching method to adjust for unmeasured spatial confounders and measured confounders. Unmeasured spatial confounding may be concerning because the entrenched stroke risk in our study population was not fully explained by many known risk factors (e.g. sociodemographic factors, behavioral factors, structural racism, and access to healthcare). For both analyses, the first matching was one-to-$n$ distance-matching, which matches one exposed census tract to unexposed census tract(s) by their spatial distance, which may adjust for unmeasured spatial confounding to some extent. We matched each exposed census tract to unexposed census tract(s) using a radius of 11.1 km (0.1°).

For the second round of matching, we conducted one-to-one nearest neighbor matching with replacement (NNWR) by PS (for the first analysis) or GPS (for the second analysis) in each of the distance-matched strata. These dual matching techniques resulted in one-to-one matched pairs of exposed and unexposed census tracts. We estimated PS for proximity to refineries. To estimate GPS, we estimated conditional GPS (CGPS) for $X_c$ conditional on proximity to refineries, and then multiplied PS and CGPS. This is because $X_c$ has both binary (i.e. if a refinery exists within a prespecified buffer distance of the tract) and continuous (i.e. if petroleum production is higher or lower) exposure dimensions, such that $X_c$ has a bimodal distribution. Conventional regression approaches for GPS for a bimodal distribution of exposure may lead to biased estimation, specifically in settings for unmeasured spatial confounding, which may be the case such for our exposure metric, $X_c$. The method for the second analysis is referred to as CGPS-based spatial matching (CGPS$\text{sm}$) [22] and statistical simulations showed that CGPS$\text{sm}$ can adjust for unmeasured spatial confounding [22]. The target estimand of the first and second analyses is the average treatment effect in the treated [22].

Generalized additive models (GAMs) with spatial smoother (i.e. thin-plane spline) were used to build the PS and CGPS models to augment adjustment for unmeasured spatial confounding. Variables with
Figure 2. Census tract-level self-reported stroke prevalence, locations of petroleum refineries, and locations of SO\textsubscript{2} monitoring stations.
their standardized mean difference (SMD) $\geq \pm 0.25$ were added to the PS model. The CGPS model was fitted using only exposed census tracts of the original dataset; this model was used to estimate CGPS for a one-to-\(n\) distanced matched sample. Variables that were correlated with \(X_c \geq \pm 0.25\) in the exposed census tracts were added to the CGPS model [22]. GPS for \(X_c = w > 0\) in each distance-matched stratum was estimated. Covariate balance was checked using SMD (figures S4 and S5) because our analysis is for one exposed census tract-to-one unexposed census tract matched samples [22].

We generated 1000 bootstrapped samples from a distance- and (G)PS-matched sample to estimate a standard error for the association between PPR and stroke prevalence. As sensitivity analyses, we conducted one-to-one nearest neighbor matching without replacement (NNWoR) and one-to-one nearest neighbor caliper matching with/without replacement (NNCWR/NNCWoR) for the second round of matching.

Figures S4 and S5 show SMD for covariates by NNWR/NNWoR/NNCWR/NNCWoR. Covariates were balanced at the cut-off value of $\pm 0.25$ [23]. Covariates were balanced the most by NNWR; most of SMDs were $\leq \pm 0.1$. We tested non-linear associations using a thin-plane spline in GAM.

We estimated how much of the variation in stroke prevalence might be explained by residential exposure to PPR. For each exposed census tract, the number of self-reported prevalent cases explained by residential exposure to PPR was estimated as $\hat{\beta}X_c \times \text{Pop}_c / 100$, where $\hat{\beta}$ is a coefficient reflecting a percentage point increase in self-reported stroke prevalence (%) per one unit increase in $X_c$ and Pop$_c$ is the population aged $\geq 18$ years The percentage of prevalence explained was estimated as $\hat{\beta}X_c / \text{Prev}_c \times 100$, where Prev$_c$ is stroke prevalence (%).

We conducted all statistical analyses using R software 3.5.3 with mgcv, xgboost, and CGPSpatialmatch packages [22].

3. Results

Population characteristics, by distance to refineries, are presented in table 1. Census tracts nearer to refineries had lower educational attainment and median household income, as well as higher rates of poverty and Hispanic persons. Populations residing within 2.5 km of refineries had the highest proportion of residents whose educational attainment was less than high school (25.5%). Individuals identified as Hispanic were the most likely to reside within 2.5 km of refineries (39.2%). The prevalence of current smoking and of obesity were higher in census tracts nearer to refineries. Figure S6 presents correlations between sociodemographic variables and $X_c$ within 5 km from refineries, suggesting that the non-Hispanic Black population was exposed to higher PPR, while the non-Hispanic White population was exposed to lower PPR. This trend was similar for $X_c$ within 2.5 km from refineries (results not shown).

Figure 2 presents stroke prevalence across census tracts. The prevalence of self-reported stroke ranged from 0.4% to 12.7% for all the census tracts of the seven states. For the subset of census tracts located within 11.1 km from refineries, the prevalence ranged from 0.7% to 12.7%.

Table 2 presents the association between proximity to refineries and self-reported stroke prevalence. We did not observe statistically significant associations. Table 2 also shows the association between the weighted average of actual petroleum production (i.e. $X_c$) and self-reported stroke prevalence. We found positive associations between $X_c$ at each of the two prespecified buffer distances and stroke, but only one of these associations was statistically significant. One standard deviation (SD) increase in $X_c$ with the 2.5 km buffer distance was associated with a 0.23 (95% confidence intervals: −0.02, 0.48) percentage point increase in self-reported stroke prevalence. One SD increase in $X_c$ within the 5 km buffer distance was associated with a 0.22 (0.09, 0.34) percentage point increase in stroke prevalence. We did not find evidence of non-linear associations between $X_c$ and self-reported stroke prevalence.

Table 3 presents self-reported stroke prevalence potentially explained by residential exposure within 5 km of PPR across seven states. For all seven states, the percentage of potentially explained geographic variation in stroke was 5.6 (2.4, 8.9) and the number of prevalent cases potentially explained was 2200 (921, 3478). The percentage was the highest in Mississippi, followed by Texas, Louisiana, New Mexico, Oklahoma, Arkansas, and Alabama. The number of cases was the highest in Texas, followed by Louisiana, Oklahoma, Mississippi, Alabama, Arkansas, and New Mexico. At the census tract-level, the variation in stroke prevalence potentially explained by PPR ranged from 0.08% (0.03, 0.13) to 23.3% (10.6, 40.0) (figure 3).

Tables S1 and S2 present sensitivity analyses for the associations between proximity to refineries, the weighted average of actual petroleum production, and self-reported stroke prevalence. The association between the weighted average of actual petroleum production with a buffer distance of 5 km and self-reported stroke prevalence was robust although covariates were less balanced by the matching techniques other than the main analysis (i.e. NNWR) (figure S5(B)). The other associations (i.e. the weighted average of actual petroleum production with a buffer distance of 2.5 km, the proximity to refineries) were less consistent. The other matching techniques showed...
Table 1. Distribution of census tract-level stroke prevalence, potential stroke risk factors, and potential confounders by proximity to petroleum refineries.

| Number of census tracts<sup>c</sup> | <2.5 km<sup>a</sup> | P-value<sup>b</sup> | 2.5–5 km<sup>a</sup> | P-value<sup>b</sup> | ≤5 km<sup>a</sup> | P-value<sup>b</sup> | 5–11.1 km<sup>a</sup> | P-value<sup>b</sup> |
|-----------------------------------|-------------------|------------------|------------------|------------------|----------------|------------------|----------------|------------------|
| **Sex (%)**                       |                   |                  |                  |                  |                |                  |                |                  |
| Males                             | 48.4 (4.6)        | 0.466            | 48.2 (5.5)       | 0.555            | 42.2           | 0.410            | 47.9 (5.9)      |                  |
| Females                           | 51.6 (4.6)        | 0.466            | 51.8 (5.5)       | 0.555            | 51.8 (5.3)     | 0.410            | 52.1 (5.9)      |                  |
| **Age, %**                         |                   |                  |                  |                  |                |                  |                |                  |
| 18–19 years                       | 4.0 (2.7)         | 0.556            | 3.6 (3.2)        | 0.743            | 3.7 (3.1)      | 0.997            | 3.7 (4.9)       |                  |
| 20–24 years                       | 10.1 (3.2)        | 0.922            | 10.0 (4.8)       | 0.755            | 10.1 (4.5)     | 0.755            | 10.2 (6.0)      |                  |
| 25–44 years                       | 37.4 (6.5)        | 0.139            | 35.9 (6.5)       | 0.760            | 36.2 (6.5)     | 0.679            | 36.0 (7.8)      |                  |
| 45–64 years                       | 32.1 (4.4)        | 0.742            | 32.8 (5.4)       | 0.035            | 32.6 (5.2)     | 0.053            | 31.8 (6.5)      |                  |
| 65–84 years                       | 14.4 (4.9)        | 0.020            | 15.6 (5.3)       | 0.411            | 15.3 (5.2)     | 0.087            | 15.9 (5.7)      |                  |
| 85+ years                         | 2.0 (1.5)         | 0.148            | 2.1 (1.9)        | 0.133            | 2.1 (1.8)      | 0.058            | 2.3 (1.9)       |                  |
| **Race/ethnicity, %**             |                   |                  |                  |                  |                |                  |                |                  |
| Hispanic                          | 39.2 (36.0)       | 0.043            | 36.0 (35.8)      | 0.053            | 36.8 (35.8)    | 0.014            | 31.0 (33.8)     |                  |
| Non-Hispanic White                | 28.6 (26.6)       | 0.255            | 31.4 (28.0)      | 0.636            | 30.7 (27.6)    | 0.372            | 32.4 (28.5)     |                  |
| Non-Hispanic Black                | 28.5 (32.6)       | 0.363            | 28.4 (32.8)      | 0.147            | 28.5 (32.7)    | 0.109            | 32.0 (32.6)     |                  |
| **Socioeconomic conditions**      |                   |                  |                  |                  |                |                  |                |                  |
| Educational attainment of less than high school, % | 25.5 (10.7) | <0.001 | 22.4 (12.8) | 0.001 | 23.2 (12.4) | <0.001 | 19.5 (11.7) |                  |
| Under the federal poverty line, % | 22.1 (9.9)        | 0.265            | 20.1 (10.0)      | 0.661            | 20.6 (10.0)    | 0.892            | 20.5 (11.9)     |                  |
| Have health insurance, %          | 73.1 (11.9)       | <0.001           | 76.6 (11.9)      | 0.002            | 75.7 (12.0)    | <0.001           | 79.2 (10.6)     |                  |
| Median household income, $         | 39 882.7 (15 095.5) | 0.066 | 44 326.9 (28.4) | 0.147 | 28.5 (32.7) | 0.109 | 32.0 (32.6) |                  |
| **Others**                        |                   |                  |                  |                  |                |                  |                |                  |
| Prevalence of current smoking, %  | 23.2 (5.6)        | <0.001           | 21.1 (4.9)       | 0.108            | 21.6 (5.2)     | <0.001           | 20.5 (5.3)      |                  |
| Prevalence of obesity, %          | 41.6 (7.2)        | <0.001           | 40.8 (6.8)       | 0.032            | 41.0 (6.9)     | 0.005            | 39.7 (6.6)      |                  |
| Prevalence of diabetes, %         | 15.9 (4.3)        | 0.415            | 15.8 (4.7)       | 0.387            | 15.8 (4.6)     | 0.276            | 15.4 (5.1)      |                  |
| Prevalence of hypertension, %     | 38.3 (7.8)        | 0.756            | 38.7 (8.2)       | 0.919            | 38.6 (8.1)     | 0.959            | 38.6 (8.4)      |                  |
| Stroke prevalence, %              | 4.4 (1.6)         | 0.798            | 4.4 (1.6)        | 0.708            | 4.4 (1.6)      | 0.670            | 4.4 (1.7)       |                  |

<sup>a</sup> Mean (standard deviation).  
<sup>b</sup> P-value for the difference with respect to 5–11.1 km.  
<sup>c</sup> Distribution of variables are shown regardless of whether they were used in matched analyses. Distributions of variables used in matched analyses are presented in figures S4 and S5 as standardized mean differences.

Table 2. Percentage point increase in self-reported stroke associated with proximity to petroleum refineries and percentage point increase in self-reported stroke by a one standard deviation (SD) increase<sup>a</sup> in the weighted average of actual petroleum production.

| Exposure | Percentage point increase (95% CI) |
|----------|-----------------------------------|
| Proximity to refineries |                     |
| <2.5 km  | 0.08 (−0.30, 0.46)                |
| 2.5–5 km | 0.22 (0.00, 0.45)                 |
| ≤5 km    | 0.10 (−0.08, 0.28)                |
| 5–11.1 km|                     |
| The weighted average of actual petroleum production by buffer distance |                     |
| 2.5 km   | 0.23 (−0.02, 0.48)                |
| 5 km     | 0.22 (0.09, 0.34)                 |

<sup>a</sup> SD was calculated for only exposed groups (i.e. Xc > 0); a one SD for Xc with 2.5 km buffer distance was 59 805 barrels per day; a one SD for Xc with 5 km buffer distance was 61 619 barrels per day.

Table 3. Self-reported stroke prevalence and number of prevalent cases explained by weighted average of actual petroleum production within 5 km distance from refineries by seven states, based on results of association between PPR and stroke.

| State       | % of prevalence explained | Prevalent cases explained |
|-------------|---------------------------|----------------------------|
| Alabama     | 1.1 (0.5, 1.7)            | 17.8 (7.4, 28.1)           |
| Arkansas    | 2.0 (0.8, 3.1)            | 14.9 (6.2, 23.5)           |
| Louisiana   | 5.9 (2.5, 9.3)            | 602.6 (252.3, 952.9)       |
| Mississippi | 11.7 (4.9, 18.6)          | 31.8 (13.3, 50.3)          |
| New Mexico  | 3.8 (1.6, 6.0)            | 12.0 (5.0, 18.9)           |
| Oklahoma    | 2.9 (1.2, 4.6)            | 94.6 (39.6, 149.5)         |
| Texas       | 6.8 (2.9, 10.8)           | 1426.1 (597.2, 2255.0)     |
| All seven states | 5.6 (2.4, 8.9) | 2199.8 (921.0, 3478.2)     |

4. Discussion

This study investigated the association between residential exposure to PPR and stroke prevalence in the southern US. Our results demonstrated positive associations between residential exposure to PPR within 5 km from petroleum refineries and self-reported stroke prevalence. Our results support the hypothesis statistically significant association for 2.5–5 km to refineries (table S1) while some covariates were less balanced (figure S4(B)).
that differential exposures to the pollutant mixtures generated from PPR may plausibly contribute to geographical disparities in stroke risk.

Our descriptive analysis (table 1) shows that the prevalence of stroke and risk factors of stroke were distributed differently across distances from census tracts to refineries. For example, the prevalence of stroke, hypertension and diabetes were similar between census tracts located <2.5 km from refineries, census tracts located at ≥2.5 km and ≤5 km from refineries, and census tracts located at >5 km and ≤11.1 km from refineries, while census tracts nearer to refineries had a higher prevalence of obesity. These health conditions are interrelated, and
sociodemographic factors and smoking may also contribute to these distributions. When adjusting for sociodemographic factors, smoking, hypertension, diabetes, obesity, and unmeasured spatial confounding, we found that census tracts within 5 km from refineries may have a higher prevalence of stroke than census tracts located at >5 km and ≤11.1 km from refineries (table 2).

The EPA regulates several classes of the air pollutants emitted by PPR under the Clean Air Act. The association between residential exposure to PPR and self-reported stroke, however, may support further study of emissions standards in communities with differential exposure. Indeed, our preliminary finding of state-level variation in the prevalence of stroke explained by PPR exposure merits further investigation with respect to differences in Clean Air Act State Implementation Plans. PPR is concentrated within the southeastern US and emits multiple airborne pollutant classes that have been implicated in stroke pathogenesis, including PM, NOₓ, SO₂, PAH, and VOCs [15–17]. These constituent pollutants may travel sufficient distances to expose residents of adjacent communities [17, 19]. Effluent may contain hydrocarbons and related derivatives, affecting the quality of adjacent soil and potable water supplies [14].

The difference between the results by the inverse distance squared weighted actual petroleum production (i.e. X_air) and those with the proximity to refineries suggests that the former may better represent the nature of residential exposure to pollutants generated from petroleum refineries. The degree of pollution depends on many factors such as fuels used to operate petroleum refining, pollution control measures, treatment systems, chemical compositions of crude oils, and meteorological factors, among others. Although our study could not consider each of these factors in our exposure classification, our results suggest that the capacity of petroleum refineries may be a rough proxy of residential exposure to PPR. The moderate correlation between X_AQS and SO₂ at 2.5 km and 5 km buffer distances may support the use of our exposure metrics as a proxy for at least air pollution. (American Meteorological Society/EPA) (AERMOD), an EPA’s recommended regulatory dispersion model, revealed that SO₂ and VOCs are dispersed over 10 km in Texas City [19] and SO₂ may be coupled with air pollutants such as VOCs and particles smaller than PM₁.₅ (e.g. secondary particles) close to oil refineries [17], although X_AQS is less correlated with SO₂ at 10 km buffer distance in our study.

We did not find significant associations between residential exposure to PPR within 2.5 km from refineries and self-reported stroke prevalence. A smaller sample of census tracts for the 2.5 km might have contributed to the null finding with higher standard errors. That said, all sensitivity analyses showed positive association for the 2.5 km buffer (one result is significant) and the strength of association was comparable to that of the association for the 5 km buffer (table S2). This suggests that census tracts within 5 km from refineries are at increased risk of stroke.

The stroke prevalence potentially explained by PPR was different over census tracts and states. Several factors may have contributed to this difference. For example, PPR exposure due to the degree of emissions from refineries, characteristics of refineries (e.g. size, control measures), state environmental regulations, profiles of stroke risk factors, and population size and characteristics may differ.

We note strengths of this study. To the best of our knowledge, this study is the first to investigate the association between residential exposure to PPR and prevalent stroke in the southern US. We considered residential exposure as a function of both distance and refinery capacity. Through conducting our analysis at the census tract level, this study contributes to the emerging literature on small area variation in stroke outcomes. In addition to the geographic patterning of PPR, its association with sociodemographic factors may be relevant to understanding and addressing entrenched sociodemographic disparities in stroke outcomes.

Our study has several limitations. First, our cross-sectional study design does not permit us to establish temporal (or causal) relationships between the exposure and outcome. Second, self-reported disease status may be a poor surrogate for administratively verifiable diagnoses [24]. Third, because it was not possible to isolate possible mixtures of pollutants in air, noise, soil, and water, we could not consider profiles of actual pollutants and exposure histories. Our exposure metrics were approximate proxies for multiple potential routes of exposure to pollution mixtures generated from PPR. We were unable to account for the route(s), magnitude, and duration of individual exposures. Our exposure definition assumes that each refinery is responsible for a respective share of annual production and emissions that is proportionate to its documented atmospheric distillation capacity. Our exposure definition does not capture corollary major determinants of emissions such as complex refining processes, pollution control measures, fuels used, the quality of crude oil, and temporal fluctuations in operating and idle capacity, among others. Residential exposure to air pollution is further determined by smokestack attributes (e.g. height, diameter), topology, and meteorology, all of which were not considered in our study. While our visual inspection of annual wind pattern data obtained from the National Oceanic and Atmospheric Administration supported the theory that air pollutants might have been propelled to residential areas, wind directionality was not evenly distributed, such that inverse squared distance weighing might have a greater impact on stroke risk in census tracts closer to refineries.
have attenuated the strength of associations. This possible underestimation is because residents with higher exposure were likely assigned lower exposure estimates and vice versa. It was not feasible to evaluate potential occupational exposures. Fourth, the nature of our data dictated artificial dichotomization of vascular risk factors with a dose-response relationship to stroke risk, such as cigarette use. Fifth, since the distance between census tracts and refineries was based on the geographic centroid of census tracts and point coordinates of petroleum refineries, some residents in spatially large census tracts, who lived within a pre-specified distance from refineries, were not included in the main analysis. Such census tracts are generally where residents are not evenly distributed due to topological variation, such as mountainous terrain. Lastly, we did not evaluate buffer distances other than 2.5 km and 5 km. We categorized census tracts within 2.5 km or 5 km from refineries as exposed groups and census tracts located at >5 km and ≤11.1 km from refineries as the control group. There is a trade-off between expanding buffer distances to increase the sample size of those defined as exposed to fully capture actual exposed census tracts and reducing buffer distances to achieve adjustment for unmeasured spatial confounding and reduce potential exposure misclassification. If these census tracts categorized as the control group had been actually exposed to PPR, the association identified might be underestimated. Further studies are needed to better understand the implications of different buffer distances regarding potential links between PPR exposure and stroke outcomes and to further explore the various potential pathways (e.g. air pollution, water quality) through which PPR may impact health.

5. Conclusion

Our cross-sectional ecological study, which is a research based at the population level without individual-level data, investigated the association between residential exposure to PPR and self-reported stroke prevalence in the southern US. Although emissions of hazardous air pollutants and effluent from petroleum refineries are regulated under the US federal law, our study suggests that residential proximity to petroleum refineries with higher PPR may increase stroke risk. Longitudinal studies with incidence-based outcome measures and more detailed exposure assessments are needed to establish a temporal and potentially causal relationship between PPR exposure and stroke risk. This line of research may improve understanding of the public health externalities sustained by refinery-adjacent communities and has the potential to inform both public health and environmental regulatory interventions to mitigate the potential health risks conferred by PPR exposure.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/HonghyokKim/Stroke_PPR.

Acknowledgments

Dr Honghyok Kim and Dr Michelle Bell were supported by Assistance Agreement No. RD835871 awarded by the US Environmental Protection Agency to Yale University and were also supported by the National Institute on Minority Health and Health Disparities of the National Institutes of Health under Award Number R01MD012769. Dr Honghyok Kim was further supported by Basic Science Research Program through the National Research of Korea (NRF) funded by the Ministry of Education (2021R1A6A3A14039711). Dr Natalia Festa was supported by VA Office of Academic Affiliations through the VA/National Clinician Scholars Program and Yale University. Yale National Clinician Scholars Program and by CTSA Grant Number TL1 TR001864 from the National Center for Advancing Translational Science (NCATS), a component of the National Institutes of Health (NIH). She is currently supported by T32AG019134. Dr Gill is supported by the Yale Claude D Pepper Older Americans Independence Center (P30AG021342).

This document has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The contents of this manuscript do not represent the views of the US Department of Veterans Affairs or the US Government.

Conflict of interests

Mr Dae Cheol Kim is a full-time employee at Hyundai Oilbank. His contribution was made solely under his program at Graduate School of Public Health in Seoul National University, which is not related to the company. The other authors declare there is no conflict of interests.

ORCID iDs

Honghyok Kim  https://orcid.org/0000-0001-9636-3428
Kate Burrows  https://orcid.org/0000-0002-0530-6776
Michelle L Bell  https://orcid.org/0000-0002-3965-1359
References

[1] Feigin V L et al 2021 Burden of neurological disorders across the US from 1990–2017: a global burden of disease study JAMA Neurol. 78 165–76
[2] Girotra T, Lekoumbou A, Bishu K G and Orbiagile B 2020 A contemporary and comprehensive analysis of the costs of stroke in the United States J. Neurol. Sci. 410 116643
[3] Orbiagile B et al 2013 Forecasting the future of stroke in the United States: a policy statement from the American Heart Association and American Stroke Association Stroke 44 2361–75
[4] Tassone E C, Waller L A and Casper M L 2009 Small-area racial disparity in stroke mortality: an application of Bayesian spatial hierarchical modeling Epidemiology 20 234
[5] Wadley V G, Unverzagt F W, McGuire L C, Moy C S, Go R, Kissela B, McClure L A, Crowe M, Howard V J and Howard G 2011 Incident cognitive impairment is elevated in the stroke belt: the REGARDS study Ann. Neurol. 70 229–36
[6] Howard G, Kleindorfer D O, Cushman M, Long D L, Jasne A, Judd S E, Higginbotham J C and Howard V J 2017 Contributors to the excess stroke mortality in rural areas in the United States Stroke 48 1773–8
[7] Glymour M M, Avendaño M and Berkman L F 2007 Is the ‘stroke belt’ worn from childhood? Risk of first stroke and state of residence in childhood and adulthood Stroke 38 2415–21
[8] Cushman M, Cantrell R A, McClure L A, Howard G, Prineas R J, Moy C S, Temple E M and Howard V J 2008 Estimated 10-year stroke risk by region and race in the United States Ann. Neurol. 64 507–13
[9] Howard G and Howard V J 2020 Twenty years of progress toward understanding the stroke belt Stroke 51 742–90
[10] Howard V J, McClure L A, Kleindorfer D O, Cunningham S A, Thrift A G, Diez Roux A V and Howard G 2016 Neighborhood socioeconomic index and stroke incidence in a national cohort of blacks and whites Neurology 87 2340–7
[11] Liao Y, Greenland K J, Croft J B, Keenan N L and Giles W H 2009 Factors explaining excess stroke prevalence in the US stroke belt Stroke 40 3336–41
[12] Howard G 1999 Why do we have a stroke belt in the southeastern United States? A review of unlikely and uninvestigated potential causes Am. J. Med. Sci. 317 160–7
[13] Karp D N, Wolff C S, Wiebe D J, Branas C C, Carr B G and Mullen M T 2016 Reassessing the stroke belt: using small area spatial statistics to identify clusters of high stroke mortality in the United States Stroke 47 1399–42
[14] Radeysh L, Tussupova K, Klemes J J and Persson K M 2021 Oil refinery and water pollution in the context of sustainable development: developing and developed countries J. Clean. Prod. 302 126987
[15] Zhang H-W, Kok V C, Chuang S-C, Tseng C-H, Lin C-T, Li T-C, Sung F-G, Wen C P, Hsung C A and Hsu C Y 2019 Long-term ambient hydrocarbons exposure and incidence of ischemic stroke Plas One 14 e0225363
[16] Ljungman P L and Mittleman M A 2014 Ambient air pollution and stroke Stroke 45 3734–41
[17] Ragothaman A and Anderson W A 2017 Air quality impacts of petroleum refining and petrochemical industries Environments 4 66
[18] United States Energy Information Administration (USEIA) 2022a Drilling productivity report (available at: www.eia.gov/petroleum/drilling/) (Accessed 3 February 2022)
[19] Chen J A et al 2012 Sulfur dioxide and volatile organic compound exposure to a community in Texas City, Texas evaluated using AERMOD and empirical monitoring data Ann. J. Environ. Sci. 8 622–32
[20] Dufouil C et al 2017 Revised Framingham stroke risk profile to reflect temporal trends Circulation 135 1145–59
[21] Kernan W N, Inzucchi S E, Sawan C, Macko R F and Furie K L 2013 Obesity: a stubbornly obvious target for stroke prevention Stroke 44 278–86
[22] Kim H and Bell M 2022 Adjustment for unmeasured spatial confounding in settings of continuous exposure conditional on the binary exposure status: conditional generalized propensity score-based spatial matching (https://doi.org/10.48550/arXiv.2202.00814)
[23] Stuart E A, Lee B K and Leacy F P 2013 Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research J. Clin. Epidemiol. 66 584–90
[24] Wu C-S, Lai M-S, Gau S S-F, Wang S-C and Tsai H-J 2014 Concordance between patient self-reports and claims data on clinical diagnoses, medication use, and health system utilization in Taiwan PLoS One 9 e112257
[25] United States Environmental Protection Agency (USEPA) 2015 AP 42, 5th edn, vol I ch 5: petroleum industry (available at: www.epa.gov/sites/production/files/2020-09/documents/5.1_petroleum_refining.pdf) (Accessed 3 February 2022)
[26] Garcia-Gonzales D A, Shonkoff S B, Hays J and Jerrett M 2019 Hazardous air pollutants associated with upstream oil and natural gas development: a critical synthesis of current peer-reviewed literature Annu. Rev. Public Health 40 283–304
[27] Carnegie Endowment for International Peace 2022 Oil-climate index (available at: http://oci.carnegieendowment.org/Supply-Chain/) (Accessed 3 February 2022)
[28] United States Energy Information Administration (USEIA) 2022 Refining crude oil (available at: www.eia.gov/energyexplained/oil-and-petroleum-products/refining-crude-oil-the-refining-process.php) (Accessed 3 February 2022)