Check for updates

Exposures and behavioural responses to wildfire smoke

Marshall Burke [1,2,3 ✉], Sam Heft-Neal[2], Jessica Li [2], Anne Driscoll[2], Patrick Baylis[4], Matthieu Stigler [2], Joakim A. Weill [5], Jennifer A. Burney [6], Jeff Wen [1], Marissa L. Childs [7] and Carlos F. Gould [1]

Pollution from wildfires constitutes a growing source of poor air quality globally. To protect health, governments largely rely on citizens to limit their own wildfire smoke exposures, but the effectiveness of this strategy is hard to observe. Using data from private pollution sensors, cell phones, social media posts and internet search activity, we find that during large wildfire smoke events, individuals in wealthy locations increasingly search for information about air quality and health protection, stay at home more and are unhappier. Residents of lower-income neighbourhoods exhibit similar patterns in searches for air quality information but not for health protection, spend less time at home and have more muted sentiment responses. During smoke events, indoor particulate matter (PM$_{2.5}$) concentrations often remain 3–4× above health-based guidelines and vary by 20× between neighbouring households. Our results suggest that policy reliance on self-protection to mitigate smoke health risks will have modest and unequal benefits.

A large body of scientific evidence documents how environmental exposures can substantially shape human outcomes. For instance, poor air quality is estimated to kill millions of people per year[1], warming temperatures lead to more crime and less economic output[2], and exposure to lead reduces educational achievement[3]. Effects can vary substantially across subgroups: air pollution is more harmful to health in poorer US counties[4], warming temperatures have more negative effects on economic output in already hot locations[5] and lead exposures appear to have larger effects in already disadvantaged households[6]. Standard models of behaviour in both public health and economics suggest that the magnitude and variation of these effects in part reflect choices that individuals make or are unable to make regarding their exposures, and that the choice sets available to individuals are in turn reflective of individuals' knowledge, circumstances and preferences[7,8]. Unfortunately, these decision-making components, as well as their behavioural outcomes, are typically hard to observe at scale. This makes it difficult to understand why a given environmental exposure generates the effect it does, why this effect might differ across groups, and whether and how policy should respond. A lack of data on behaviours and exposures also makes the evaluation of existing policy approaches challenging, which is particularly problematic for the large number of environmental settings—from wildfires to heat waves to hurricanes—in which current policy largely relies on individuals to protect themselves.

Here we show how combining traditional sensor and survey data with information from non-traditional distributed sensors— including data from private outdoor and indoor pollution sensors, cell phones, social media posts and internet search activity—can generate population-scale insights into people's knowledge, preferences and choices regarding a changing environment, and into how economic circumstances shape their choice set. We focus on understanding responses to wildfire smoke, a rapidly growing

environmental stressor throughout much of the United States and internationally. The annual area burned by wildfires in the United States has more than doubled in recent decades, as a result of a century of fire suppression and a warming climate that has left the resulting abundant fuel much more flammable[9]. This increase in fire activity has led to substantial increases in average smoke exposure across the continental United States, potentially reversing decades of improvements in air quality[10]. Absent substantial intervention, these trends are expected to continue and perhaps accelerate in a warming climate[11–13].

A growing literature has begun to document myriad health impacts of ambient wildfire smoke exposure[14–18]. As with impacts of other environmental stressors, the magnitude of these health impacts may depend on behaviours and individual-specific exposures that are often poorly measured and understood. In particular, recent wildfire case studies suggest that individuals vary in their knowledge and beliefs about their own exposures and about the risks that these exposures pose to their health[19,20]. Documented heterogeneity in the impacts of wildfire smoke exposure and exposure to other particulates[4,18,21–23] suggests that both socio-economic circumstances and previous exposures could also constrain behaviour in important ways. Understanding which features matter most is consequential for policy design: impacts driven by a lack of awareness of one's exposure, for instance, call for different interventions than impacts driven by an inability to protect oneself from a known exposure. Such understanding is particularly important for wildfire, given that current policy approaches to risk mitigation focus on the private provision of protection—that is, asking individuals to stay indoors, limit infiltration and purchase protective technologies[24].

To better understand exposures, behavioural responses and outcomes in the face of rapidly changing wildfire risk, we first develop measures of daily exposure to ambient particulate matter (PM$_{2.5}$) from wildfire smoke, as this exposure itself is not directly measured
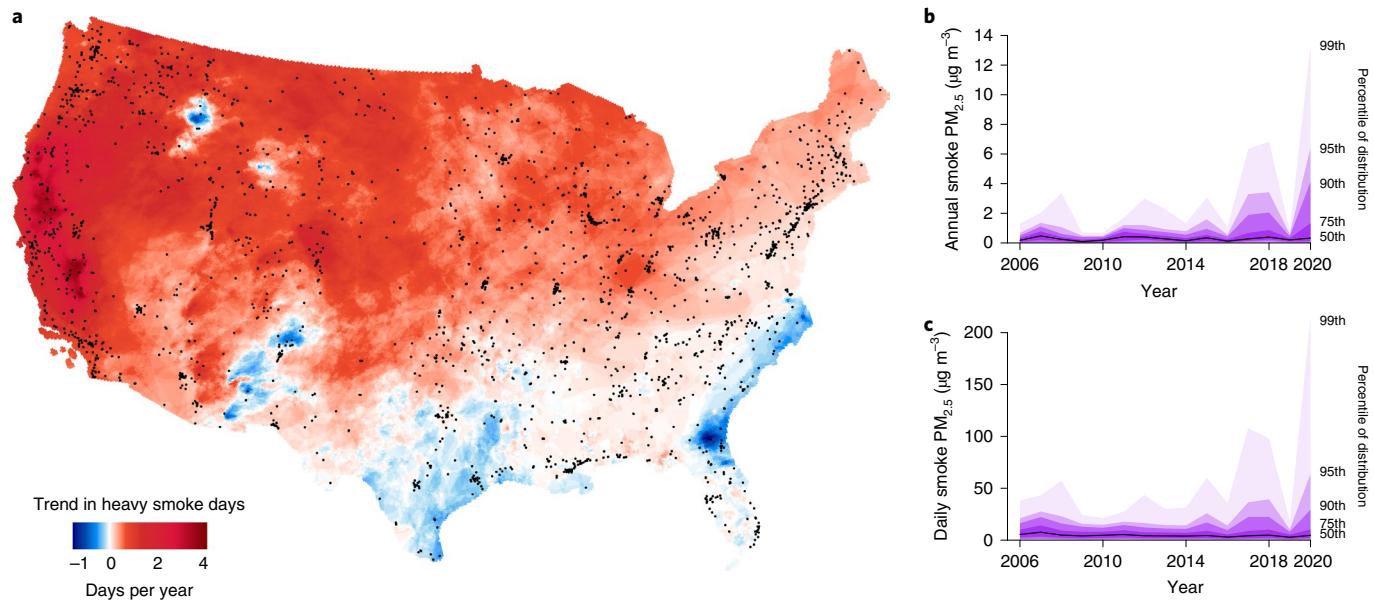
Fig. 1 | Trends in smoke exposure across the United States. a, Measurements from satellites indicate rapidly growing exposure to 'heavy' smoke plumes across much of the United States. The estimates are shown on a 10 km grid[63] and indicate the estimated annual increase between 2011 and 2020 in the number of days with smoke plumes that National Oceanic and Atmospheric Administration (NOAA) analysts designate as 'heavy', their densest plume classification. The dots indicate EPA ground-based pollution monitors. **b**, Distribution of estimated annual average smoke $PM_{2.5}$ across EPA ground monitors reporting in each year. **c**, Distribution of estimated daily smoke $PM_{2.5}$ across the same monitors. Recent increases in extreme annual exposure are being driven by increases in extreme daily exposure. Source for map in **a**: US Census Bureau.

by existing pollution sensors. To do so, we combine long time series of regulatory ground monitor data on $PM_{2.5}$ concentrations from US Environmental Protection Agency (EPA) pollution monitors with satellite-derived estimates of smoke exposure. To isolate wildfire-smoke-derived $PM_{2.5}$ from other sources of $PM_{2.5}$, we define smoke $PM_{2.5}$ as location-, month- and period-specific anomalous $PM_{2.5}$ on days in which satellites indicated that smoke was overhead (Methods). The resulting ambient smoke $PM_{2.5}$ exposure data cover 782 US counties that contain ~74% of the US population (Extended Data Fig. 1), and these data display wide spatial and temporal variability (Fig. 1). Satellite-based counts of the annual number of dense smoke plumes overhead have trended upwards in the past decade throughout most of the continental United States, particularly in the West (Fig. 1a), which we estimate have helped drive rapid increases in estimated annual smoke $PM_{2.5}$ across the country and in days with extreme $PM_{2.5}$ due to smoke (Fig. 1b,c).

We merge these daily wildfire-smoke-derived data with multiple high-frequency datasets that are measured at the population scale and shed light on individuals' knowledge, beliefs and behaviours during wildfires. To study individuals' awareness of their exposure, we first analyse location-specific variation in search query behaviour related to smoke exposure. We use public data on specific search queries (for example, 'air quality') from Google Trends, which provides normalized data on search term popularity at the weekly level across 'designated market areas' (DMAs) (roughly, metro areas; Methods). We interpret a purposeful search for information related to smoke exposure as evidence that an individual is aware he or she is being exposed and that his or her exposure level is worth learning about—what we call 'salience'.

Second, we study individuals' preferences and sentiments regarding wildfire smoke exposure. Such preferences underlie standard theoretical models of choice behaviour but are hard to observe directly and at scale. Past work has shown that social media behaviour can be a sensitive and accurate tool for understanding individuals' preferences and sentiments towards what is happening

around them[25], including a changing environment[26,27]. Following this earlier work, we analyse ~1.7 billion georeferenced Twitter updates ('tweets') posted since 2016 using natural language processing algorithms that extract information on the sentiment revealed in each tweet[28] (Methods). This approach has been validated at the population scale against self-reported measures of emotional state[25] and complements earlier work that used Twitter to directly measure wildfire activity[29] and infer smoke concentrations[30].

Third, we again use Google search queries to study whether individuals sought information regarding specific health-protective actions, analysing item-specific search terms such as 'air filter', 'air purifier' and 'smoke mask'. While we cannot observe whether individuals eventually purchased these items except in the case of PurpleAir monitors (as described below), such search behaviour can be interpreted, at a minimum, as evidence of an individual's belief that health-protective options exist. Evidence from other settings suggests that search activity is predictive of future behaviour, including consumer purchases[31,32].

Fourth, we use smartphone-derived location data to study whether individuals altered their physical movements during periods of smoke exposure. Short-term migration in response to other environmental stress (such as hurricanes) is common and is a plausible avenue by which individuals or households could seek to limit exposure to wildfire smoke. We study both the share of people estimated to be completely at home and the share estimated to be completely away from their homes, on days or weeks of smoke exposure.

We combine each measure with our smoke $PM_{2.5}$ data and analyse the effect of smoke on each outcome using panel fixed effects estimators that exploit local temporal variation in both exposures and outcomes. While long-term exposure to wildfire smoke shows clear spatial patterns and temporal trends (Fig. 1), local-level variation in daily exposure is highly random, and panel estimators—which are commonly employed in related environmental settings[2]—plausibly isolate the impact of variation in smoke exposure from other time-invariant and time-varying factors that could be correlated

with both smoke exposure and outcomes, including potential confounding from COVID-19 (Methods). To ensure that we are measuring the impact of wildfire smoke and not simply proximity to wildfire itself, we develop measures of distance to the nearest active wildfire and analyse whether responses differ by fire proximity. The unit of observation in these analyses is either the county-day (for sentiment and mobility analyses) or the metro area-week (for salience and health protection measures).

Finally, we analyse how ambient outdoor smoke $PM_{2.5}$ infiltrates into the indoor home environment and whether behaviours and circumstances shape this infiltration. Understanding indoor concentrations is critical, as individuals in the United States spend the vast majority of their time indoors. Using data from the American Time Use Survey, we calculate that Americans on average spend >70% of their time indoors at home, with higher shares for lower-income and elderly individuals and overall shares trending up over time (Extended Data Fig. 2). Personal integrated exposure to variation in ambient exposure is then probably substantially mediated by characteristics of home and work environments that are hard to observe[33–35], and these differences could in turn affect outcomes[36–39]. If socio-economic or demographic variables shape indoor environments in ways that affect exposures, as has been hypothesized[37], then exposure levels or policy choices that appear equitable on the basis of traditional outdoor measures could obscure large disparities in realized exposures.

We assemble and harmonize hourly data from 1,520 indoor PurpleAir air pollution monitors that individuals have put in their single-family homes across the United States and use nearby outdoor PurpleAir monitors to construct outdoor $PM_{2.5}$ concentrations at each home (Methods). To estimate infiltration, we use distributed lag or lagged-dependent variable panel regression to estimate the marginal increase in indoor $PM_{2.5}$ when outdoor $PM_{2.5}$ increases by one unit (that is, $\partial IndoorPM_{2.5}/\partial OutdoorPM_{2.5}$), controlling flexibly for time of day, day of week and month of sample (Methods). We estimate models that pool all indoor monitors as well as monitor-specific models, and we study how infiltration differs as a function of household and neighbourhood characteristics. Our approach complements recent work using PurpleAir to study infiltration generally[40] and during wildfires specifically[33], though our estimation approach offers advantages relative to the latter such as robustness to indoor pollution sources and to diurnal patterns in infiltration-relevant behaviours (Methods).

The timeliness and granularity of passive distributed sensor data need to be weighed against their potential non-representativeness, as the latter can bias population-scale inferences. Our search data and mobility data are probably our most representative, as the vast majority of Americans use the internet regularly and most own and use smartphones. Twitter users are less representative on average, but Twitter-derived sentiment measures have been shown to validate well against population emotional state, and related work shows that the response of sentiment to environmental stress mirrors that measured in representative survey data. The PurpleAir data are the least representative of our datasets, with wealthier and more educated households more likely to own monitors; however, as discussed below, socio-economic and demographic information does not appear strongly predictive of infiltration. See the Methods for a more detailed discussion of sample representativeness.

Results
Increases in wildfire-derived ambient $PM_{2.5}$ exposure lead to an increase in popularity for air-quality-related search terms, with even small increases above zero exposure appearing salient (Fig. 2a; $P < 0.001$; linear model effect size, 0.689; 95% confidence interval (CI), 0.503, 0.874). The results are robust to alternate air-quality-related search terms and to using analogous search terms in Spanish (Supplementary Table 1), are not driven by

proximity to active wildfires, and are robust to the inclusion of weather controls or alternate fixed effects (Supplementary Tables 2 and 3). Placebo search terms plausibly unrelated to smoke exposure do not respond to exposure, and search for smoke-related terms does not respond to variation in $PM_{2.5}$ on non-smoke days (Supplementary Table 4). Our results are consistent with interview-based evidence finding that individuals who perceived they were being exposed to smoke often used internet-based sources to confirm their perceptions[19], although we cannot easily distinguish whether individuals recognized that they were being exposed to smoke $PM_{2.5}$ specifically or just to poor air quality from any source (Supplementary Table 3). The salience of ambient exposure at low levels is also somewhat reassuring given recent evidence of health impacts for sensitive populations even at very low levels of ambient exposure[18].

We find that exposure to ambient smoke $PM_{2.5}$ makes people unhappier, as measured in an automated sentiment analysis of five years of tweets across the United States. Expressed sentiment in tweets declines roughly linearly above smoke $PM_{2.5}$ exposures of ~20 µg m^{-3} (Fig. 2b; linear effect size of 100 µg m-3 increase, −0.0087; 95% CI, −0.0108, −0.0067; $P < 0.001$). A very bad smoke day (average smoke $PM_{2.5}$ concentration of 100 µg m^{-3}) is associated with a 0.03 decline in sentiment in the non-linear model shown in Fig. 2b, which is equivalent to a roughly 0.2-standard-deviation decline relative to the overall sample standard deviation. For context, the average difference between tweet sentiment on Wednesdays and Saturdays (respectively, the lowest- and highest-sentiment days of the week in our data) is about 0.007 in our data; one day of very bad smoke (100 µg m^{-3}) is thus about four times worse in sentiment terms than replacing an average Saturday with an average Wednesday.

Negative effects of smoke on sentiment could occur through a variety of channels, including from fear or anxiety about proximate fires themselves or about what the fires represent (for example, a changing climate), from unhappiness due to disruption in normal activities (for example, school closure or inability to recreate), or from anticipation or experience of negative health impacts. While we cannot distinguish the latter channels in our data, the effects of smoke on sentiment are not driven by proximity to active wildfire and are robust to temperature and rainfall controls and to alternate fixed effects (Supplementary Table 5). These results are consistent with a broader literature documenting the negative psychological effects of air pollution exposure[41].

Exposure to smoke $PM_{2.5}$ increases search activity related to protective behaviour. Searches for technologies known to help limit exposure, including 'air filter', 'air purifier', 'smoke mask' and 'purple air', all increase on days in which smoke exposure is higher (Fig. 2c and Supplementary Table 6; $P < 0.001$; linear model effect size, 0.453; 95% CI, 0.3, 0.606). Some search queries in Spanish ('purificador de aire') respond similarly, although others ('filtro de aire') do not (Supplementary Table 6).

Finally, smoke $PM_{2.5}$ exposure on average causes more people to not leave their homes, with immediate increases at low levels of exposure that flatten off at high levels (Fig. 2d). A day of smoke exposure above 50 µg m^{-3} leads to a roughly three-percentage-point increase in the proportion of people fully at home ($P < 0.001$; linear effect size, 0.023; 95% CI, 0.016, 0.031), which corresponds to about a 10% increase above the mean. Smoke $PM_{2.5}$ exposure has a limited effect on the proportion of people fully away from their homes at low exposure levels but an increasing effect at higher exposure levels (Extended Data Fig. 3). Both results can be interpreted as protective behaviour: during heavy smoke days, many individuals shelter in their homes, and some leave the area when exposure gets severe. Both results are robust to controls and are not driven by proximity to active fires; the effect on the percentage of people at home is less robust to the addition of more stringent time controls (Supplementary Table 7).
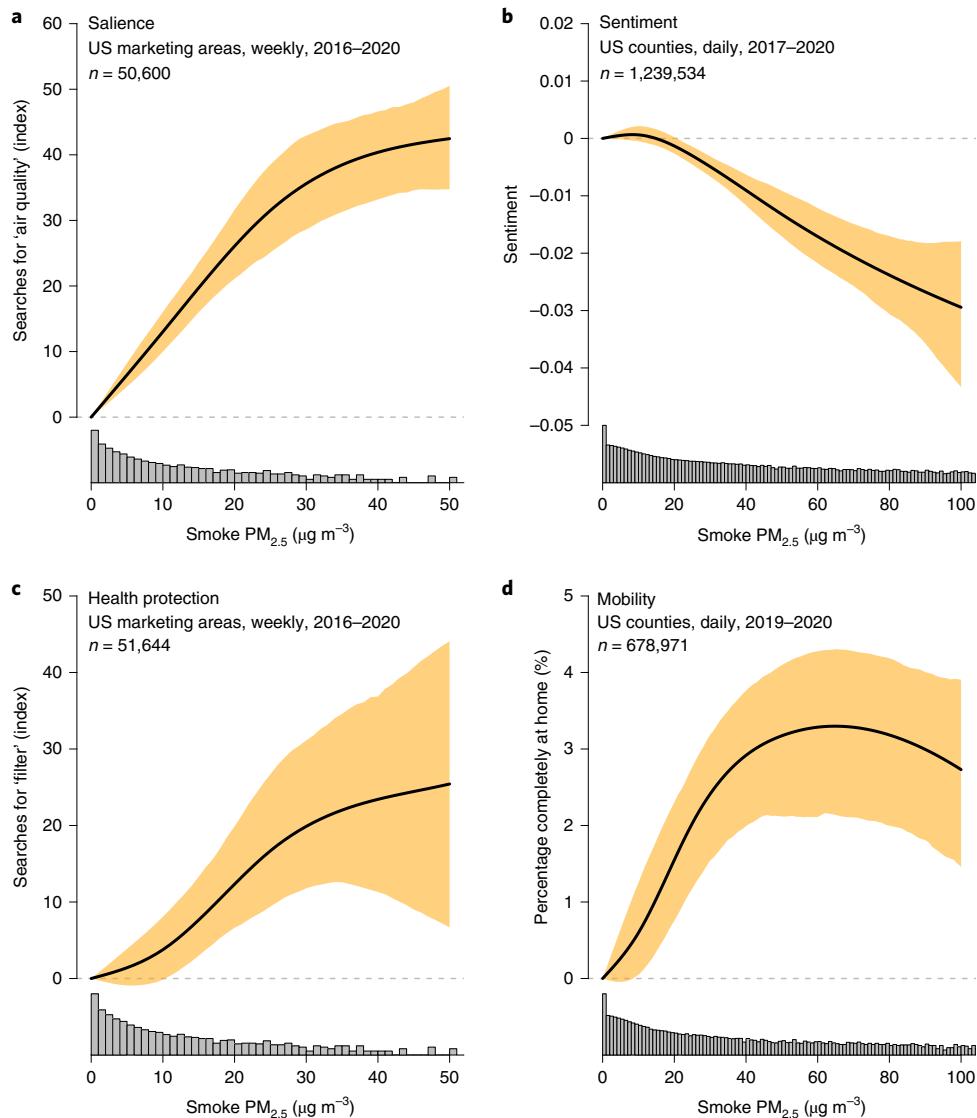
Fig. 2 | Behavioural responses to wildfire smoke exposure. a, Searches for 'air quality' on Google at the level of US designated marketing area by week, 2016–2020. The search index is normalized such that zero indicates no searches and 100 indicates the maximum number of weekly searches over the period. **b,** Average sentiment in tweets at the county-day level, 2017–2020. **c,** Searches for 'air filter' on Google at the level of US marketing area by week, 2016–2020. The search index is normalized as in **a. d,** Percentage of mobile phones estimated to be completely at home on a given day at the US county level, 2019–2020. The black lines are regression point estimates from spline fits conditional on fixed effects, and the shaded areas show bootstrapped 95% CIs. The number of observations in each regression is shown in the upper left corner of each panel. The histograms at the bottom show the log distribution of smoke $PM_{2.5}$ exposure in each sample.

Exposure and response heterogeneity. Individuals are likely to respond to environmental exposures in different ways, either because their personal exposure varies or because, for a given exposure, their knowledge of that exposure or their willingness or ability to respond to it differs. We explore heterogeneous exposures and responses to wildfire smoke as a function of socio-economic status (as measured by locality-specific median household income) and variation in average previous exposure to wildfire or other $PM_{2.5}$. Previous literature suggests that both could moderate behavioural responses to environmental stress through a variety of mechanisms, including through differential access to information about exposure risk or differential ability, motivation or knowledge of how to take protective action[8,20,42].

Consistent with earlier work[10], but in stark contrast to strong socio-economic and ethnic/racial gradients in exposure to other key pollutants in the United States[43,44], we find that exposure to both average and acute smoke $PM_{2.5}$ is largely uncorrelated with income in the United States (Extended Data Fig. 4). We also find no differences in salience of smoke exposure between lower- and higher-income counties, with similar responses of search query activity to a day of heavy smoke across income levels (Fig. 3a; $P = 0.85$ on linear interaction).

Other behavioural measures show strong income gradients. For sentiment, wealthier counties respond much more negatively to a heavy smoke day than less wealthy counties (Fig. 3b; $P = 0.001$; effect size on linear interaction, −0.001; 95% CI, −0.002, −0.001). This finding is not driven by average differences in sentiment between more and less wealthy counties, by higher overall variation in sentiment in wealthier versus poorer counties (temporal variation in sentiment is lower in wealthy counties than in less-wealthy counties in our sample) or by differences in average exposure to smoke $PM_{2.5}$ or other sources of $PM_{2.5}$ (Supplementary Table 9).
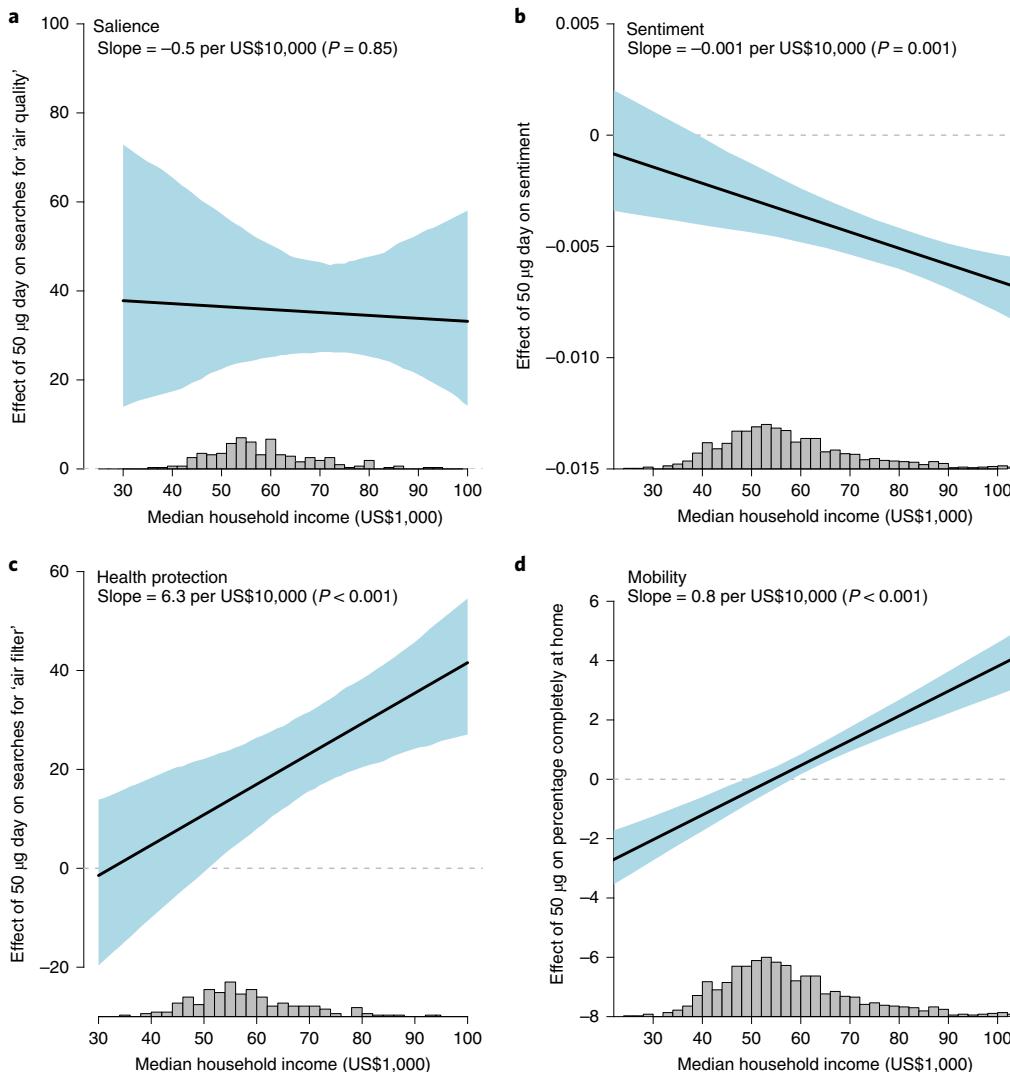
Fig. 3 | Smoke salience does not differ, but other responses do differ, among more and less wealthy populations. a, The effect of a heavy smoke exposure (50 μg smoke $PM_{2.5}$ on that day) on searches for 'air quality' on Google do not differ by income. **b**, A heavy smoke day has a stronger negative effect on sentiment among wealthy populations. **c**, Wealthier populations are substantially more likely to search for 'air filter' on Google during a heavy smoke day. **d**, Wealthier populations are more likely to be completely at home during a day of heavy smoke. The black lines show the regression point estimates of the marginal effect of a 50 μg smoke day on the outcome at different income levels, and the shaded areas show bootstrapped 95% CIs. The slope of each line is reported in the upper left with the corresponding *P* value based on a two-sided *t*-test, and the histograms at the bottom show the distribution of county incomes in each sample. The sample in each panel is same as in the corresponding panel in Fig. 2.

These results are consistent with a similar analysis in China, which also showed larger negative sentiment responses to air pollution in higher-income cities[45].

Search activity related to protective behaviour is also substantially higher in wealthier counties (Fig. 3c; *P* < 0.001; effect size on linear interaction, 0.013; 95% CI, 0.006, 0.019) and is not statistically different from zero in roughly the bottom third of the county income distribution. Finally, populations in wealthier counties are also substantially more likely to remain fully at home during a day of heavy wildfire smoke exposure than lower-income populations (Fig. 3d; *P* < 0.001; effect size on linear interaction, 0.002; 95% CI, 0.001, 0.002). These results are robust to more or less restrictive time controls (Extended Data Fig. 3c). We find no meaningful difference across income groups in the proportion of individuals fully away from their houses during days of heavy smoke exposure (Extended Data Fig. 3d).

Why do wealthier locations respond differently to smoke exposure? The measured differences do not appear to reflect differences in exposure information or in overall internet activity, given the consistent response of air-quality-related searches across income groups. Rather, the responses are consistent with lower incomes constraining choice sets and behaviours, including less flexibility in working from home, fewer resources with which to consider purchasing protective technology and (regarding the sentiment results) having other more pressing matters to worry about.

We find that behavioural measures are also affected by previous experience with smoke and with other $PM_{2.5}$ sources. An additional smoke day was less salient in locations with higher previous exposure to smoke $PM_{2.5}$, and people in locations with higher average $PM_{2.5}$ exposure prior to our study period showed smaller declines in sentiment during an additional high smoke day and fewer searches related to health-protective behaviour, but were more likely to stay at home when smoke $PM_{2.5}$ was high (Supplementary Tables 8 and 9). These results are consistent with individuals adapting their behaviour and beliefs on the basis of repeated exposure—for example, through investments in health-protective technologies.

Smoke PM$_{2.5}$ infiltration into indoor environments. We find that census tracts with PurpleAir monitors tend to be wealthier on average than tracts without monitors (Extended Data Fig. 5), a finding consistent with other analyses[46] and with the income-differentiated search activity for 'purple air' and related health-protective technologies found above. Nevertheless, the average income of locations owning indoor monitors varies by roughly 10× across locations, enabling an exploration of the role of income and other demographic factors in shaping exposures among a population with identical access to information on their exposures.

Using a pooled model, we estimate that a 1 µg m^{-3} increase in outdoor PM$_{2.5}$ is associated with a 0.145 (95% CI, 0.135, 0.153; $P < 0.001$) µg m^{-3} increase in indoor PM$_{2.5}$ over the next six hours. The estimates are robust to alternate regression models and alternate corrections to the monitor data (Extended Data Fig. 6) and are comparable in magnitude to recently published estimates[40]. Estimated infiltration is substantially lower during periods of high outdoor PM$_{2.5}$, and responses differ during smoke periods (Fig. 4a). When no smoke is present, at median outdoor PM$_{2.5}$ concentrations (6 µg m^{-3}), infiltration declines by 0.0281 for each 10 µg m^{-3} increase in outdoor PM$_{2.5}$ (95% CI, −0.02925, −0.02810; $P < 0.001$). However, when smoke is present, infiltration declines by only 0.0209 for each 10 µg m^{-3} increase in outdoor PM$_{2.5}$ (95% CI, −0.02141, −0.02043; $P < 0.001$). Earlier findings of lower infiltration on smoke days[33] were probably capturing the effect of overall high PM$_{2.5}$ rather than the effect of smoke-derived PM$_{2.5}$ specifically.

Consistent with our other behavioural measures, declining infiltration at high outdoor PM$_{2.5}$ levels suggests that salient ambient exposures induce behavioural responses, which could include closing windows or doors and/or using mechanical filtration. However, in contrast with our other behavioural measures, we find only a modest relationship between neighbourhood average income and infiltration, with households in much wealthier census blocks experiencing only slightly lower average infiltration than households in areas with one quarter the average income regardless of whether the PM$_{2.5}$ was smoke-derived (Fig. 4b; linear interaction effect size, −0.012 µg m^{-3} indoor PM$_{2.5}$ per additional 1 µg m^{-3} outdoor PM$_{2.5}$ for each US$100,000; 95% CI, −0.030, 0.006; $P = 0.180$).

To further explore predictors of infiltration, we estimate infiltration separately for each of the 1,520 indoor monitors in our dataset, match each monitor to a wide range of house- and neighbourhood-specific socio-economic, demographic, environmental and housing covariates, and fit flexible machine learning models relating infiltration to these covariates (Methods). Consistent with other work[40], we find many-fold differences in household-specific infiltration rates (mean = 0.19, s.d. = 0.16; Fig. 4c), and we confirm using a Bayesian hierarchical model that this variation is largely due to 'true' underlying variation between households rather than to sampling noise in household-level estimates (Extended Data Fig. 7a). The estimates are only modestly correlated with traditional indoor/outdoor ratio estimates (Extended Data Fig. 8), perhaps due to the difficulty in accounting for indoor sources of emissions or diurnal behavioural patterns in the traditional indoor/outdoor approach (Supplementary Information).

While racial/ethnic, socio-economic, environmental and housing variables are associated with infiltration on held-out data, their individual explanatory power is very modest, and our rich set of predictors and flexible models are surprisingly poor predictors of overall variation in infiltration, explaining only ~5% of variation across indoor monitors in our data (Extended Data Fig. 7b,c). This lack of predictive ability of socio-economic factors is also apparent on individual smoke days, where even among relatively socio-economically advantaged households, very similar outdoor PM$_{2.5}$ concentrations during a given smoke day are associated with widely varying indoor PM$_{2.5}$ concentrations (Extended Data Fig. 9).

To further investigate the differential influence of behaviour versus housing characteristics (and associated socio-economic factors), we re-estimated infiltration for individual households during periods when windows were likely to be closed and indoor filtration not running (Methods). While average infiltration during these periods was relatively similar to infiltration during all periods (Extended Data Fig. 7d), infiltration varies much more strongly with both income (Extended Data Fig. 7e) and housing age (Extended Data Fig. 7f) under these conditions. Taken together, and consistent with previous smaller-scale work[39], our results indicate that the poor explanatory power of socio-economic and housing characteristics is driven not by poor measurement of these characteristics but by the dominant effect of idiosyncratic household-specific behaviours that are not correlated with these characteristics.

Finally, using indoor monitors across the Bay Area and data prior to August 2020, we divide monitors into low (bottom quartile) and high (top quartile) infiltration groups (Fig. 4d) and study outdoor and indoor PM$_{2.5}$ levels across these groups during the extreme wildfire smoke event that the area experienced in August–September 2020. High- and low-infiltration households experienced nearly identical daily outdoor concentrations during the many-week event (Fig. 4e), but these ambient levels led to starkly different indoor concentrations. On the worst smoke days, daily average indoor concentrations across all high-infiltration homes exceeded 65 µg m^{-3}, and in some houses they exceeded 100 µg m^{-3}, well above the World Health Organization 24-hour PM$_{2.5}$ exposure guideline of 15 µg m^{-3}. Low-infiltration households were on average able to maintain indoor PM$_{2.5}$ concentrations near 5 µg m^{-3}. Across the duration of the smoke event, daily mean indoor concentrations were on average 3.5× higher in the highest quartile versus the lowest quartile of infiltration households. Differences were even larger when looking across all Bay Area monitors: households with average outdoor PM$_{2.5}$ levels within 5 µg m^{-3} of each other experienced >20× differences in average indoor PM$_{2.5}$ concentrations during the smoke event (Extended Data Fig. 10).

Discussion

A growing literature documents the large and often disparate impacts of wildfire smoke on a range of health outcomes[14–18,21,22]. Our results show how non-traditional sensor data can provide policy-relevant insight into why the magnitude and incidence of these impacts might vary. Multiple lines of evidence indicate that awareness of smoke concentrations does not appear to be a primary constraint on individual behaviour in the face of wildfire smoke exposure: even small increases in ambient exposure cause individuals to seek air quality information, become unhappier and stay in their homes. But while awareness appears to be broadly shared, it does not lead to adequate health protection. Even among populations that own indoor monitors and thus have access to accurate, real-time measures of their indoor concentrations, information is not enough to limit dangerous indoor exposures to these pollutants. This suggests that policies targeting information provision about smoke are insufficient, and perhaps not central, to enabling protective behaviour.

Socio-economic status is not correlated with outdoor smoke levels but does appear to mediate behavioural responses to such pollution. Wealthier households in our sample can more easily stay home, are more likely to seek information on protective technology and are more likely to own indoor pollution monitors. Such differential behaviour is consistent with a broader literature that shows how socio-economic status constrains households' abilities to invest in environmental quality and health protection[8,47]. Yet, at least in our sample of monitor-owning households, income is only weakly correlated with the infiltration of ambient smoke into indoor environments, and we observe many households in wealthy neighbourhoods experiencing exceedingly high levels of indoor smoke exposure.

Our results suggest that this is probably because present infiltration rates are dominated by actions such as opening windows and
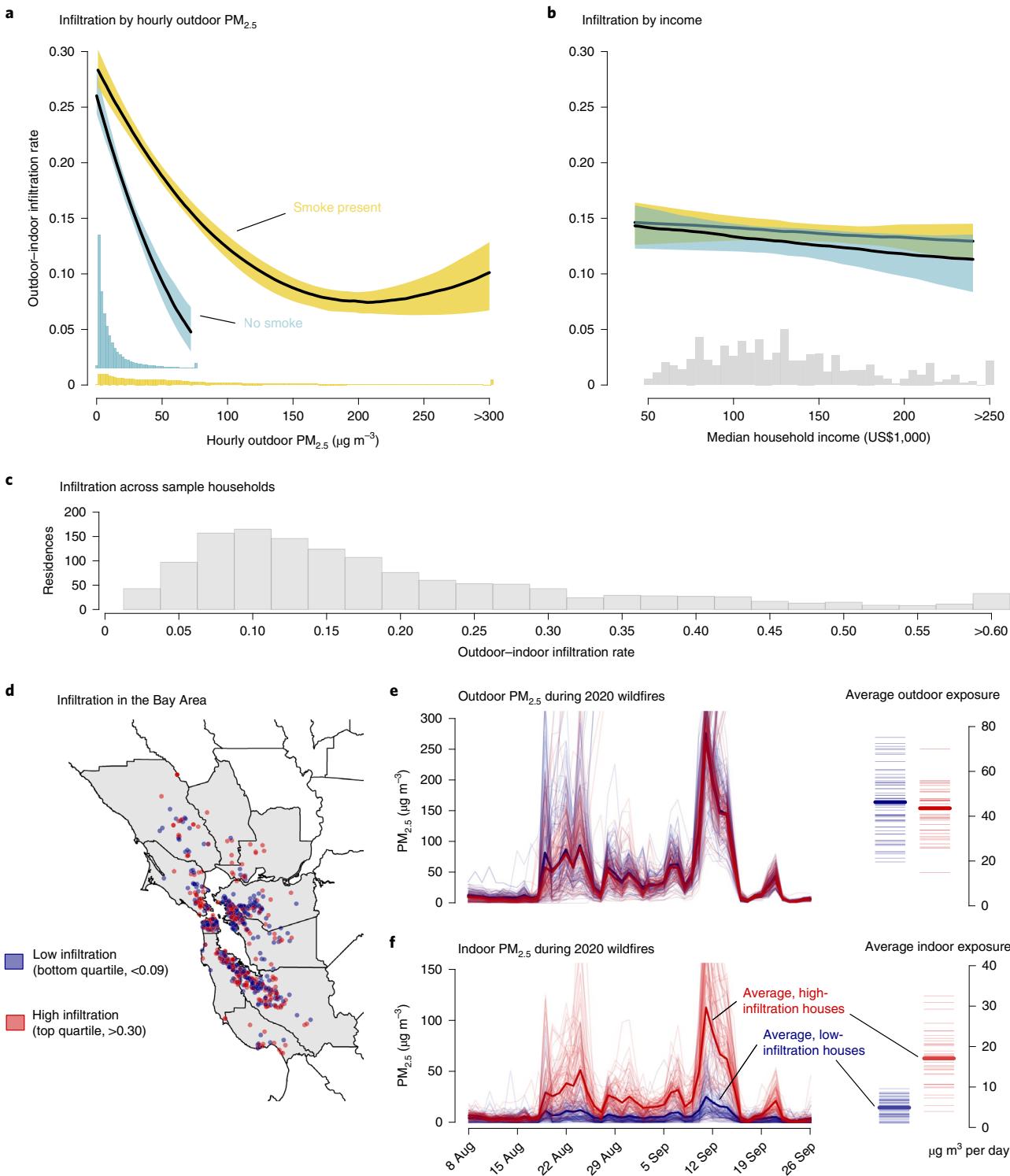
Fig. 4 | Infiltration rates decline strongly with outdoor PM$_{2.5}$ concentration during both fire and non-fire periods, but not with income, and they differ greatly across households, resulting in extreme differences in indoor exposure during wildfires. a, Infiltration rates, measured as the integrated hourly change in indoor PM$_{2.5}$ per unit increase in outdoor PM$_{2.5}$, are lower during periods of high outdoor PM$_{2.5}$ but decline more quickly when PM$_{2.5}$ comes from other sources. The lines are regression point estimates, and the shaded areas are bootstrapped 95% CIs. **b**, Infiltration rates are only weakly correlated with census-block median income; the lines, shading and units are as in **a**. **c**, The distribution of infiltration rates estimated separately for each household in the sample is wide. **d**, Residents from the highest and lowest infiltration quartiles in the Bay Area[64] are mapped and coloured according to group, showing that the groups are geographically intermixed. **e**, Daily outdoor PM$_{2.5}$ concentrations during the unprecedented August–September 2020 wildfire smoke event were highly similar between high- and low-infiltration households. The transparent lines show concentrations at each residence, and the thick lines show the averages within each group. The left panel shows daily mean concentrations, and the right panel shows averages across all days during the event. **f**, Same as **e** but for indoor PM$_{2.5}$ concentrations, showing very large (>100 μg) daily differences during peak outdoor concentrations (left) and 3.5× differences in average indoor exposure between high- and low-infiltration groups (right).

doors, not housing materials or quality that might be reflected in prices. Infiltration patterns thus point to the importance of behaviour that remains unobserved, a fact that is both encouraging and troubling. If simple but difficult-to-observe behaviours such as closing windows and doors explain the vast majority of variation in smoke infiltration, then reducing infiltration at the population scale could be much easier in theory than if infiltration was largely determined by income or housing quality, as changing these latter factors requires addressing deeper societal problems of inequality and structural racism. Nevertheless, a key limitation of our infiltration analysis is our reliance on a convenience sample of households who own PurpleAir monitors, who are overwhelmingly Californian and higher-income. Better measurement of indoor air quality and infiltration in lower-income households, and in households around the country, remains a critical research priority.

Current policy approaches to addressing smoke exposure focus on behavioural recommendations to stay at home and close windows and doors[24], but our results suggest that these policies alone are difficult to comply with and may still be inadequate: many households' indoor environments remain highly exposed, and our mobility results suggest that adherence might be difficult for lower-income households. If such behaviours are indeed hard to adopt, then the policy approach of promoting private provision of protection could be biased against disadvantaged groups. This policy approach also stands in stark contrast to the approach of public provision of protection used for other sources of $PM_{2.5}$, which has sought to reduce emissions of pollutants at their source and which has successfully reduced overall ambient exposure inequalities[48]. Further understanding the variation and causes of the behaviours that can protect indoor environments will be key to designing policy that can both lower indoor concentrations and not disadvantage certain groups.

Methods

All of the data used in our study are either from public sources where individuals are posting public statements and/or consenting to have their location tracked (Twitter and PurpleAir), or from spatially and temporally aggregated data with no available personally identifying information (search trends data and mobility data).

Estimating ground $PM_{2.5}$ concentrations from smoke. We develop a generic, tractable method for estimating ground $PM_{2.5}$ attributable to smoke at the daily level. The method requires a credible estimate of whether there is smoke in the air on a given day, and a daily time series of $PM_{2.5}$ from which location- and period-specific anomalies can be constructed. In principle, any available (accurate) daily $PM_{2.5}$ estimates could be used, including recent promising machine-learning-based efforts at generating high-resolution gridded time series of $PM_{2.5}$ concentrations[49,50]. However, existing gridded data are not available for recent years, so we instead use station-based daily $PM_{2.5}$ measures from the network of thousands of EPA stations across the continental United States (Fig. 1a).

To construct our daily measures of smoke $PM_{2.5}$, we define PM_{idmy} as the $PM_{2.5}$ concentration recorded by the EPA monitor at location i on day d, month m and year y. From this time series, we construct location- and month-specific anomalies $PManom_{idmy} = PM_{idmy} - \overline{PM}_{imy}$, where $\overline{PM}_{imy}$ is the monthly median $PM_{2.5}$ on non-smoke days at that location, and where median is defined over the three years surrounding the year of interest. We use a three-year moving median to account for the long-term declining trend in $PM_{2.5}$ across most of the United States driven by non-wildfire causes[10]. So, for example, a $PM_{2.5}$ anomaly for the Redwood City, California, EPA station on 10 January 2019 is calculated as the value on 10 January 2019 minus the median $PM_{2.5}$ value on all January days in 2018, 2019 and 2020 in Redwood City when smoke was not overhead. Our measure of whether smoke was overhead, $plume_{idmy}$, is derived from the NOAA Hazard Mapping System (HMS) satellite estimates of smoke plume boundaries. We define $plume_{idmy} = 1$ if there was a smoke plume of any thickness over location i during any time on day d, and zero otherwise. We estimate that having a smoke plume of any thickness overhead increases daily $PM_{2.5}$ concentrations at EPA reference monitors by an average of $4.0\,\mu g\,m^{-3}$, and the effect becomes stronger as plume density increases from light to medium to heavy (Supplementary Table 10). Finally, from these data we can construct $SmokePM_{idmy} = PManom_{idmy} \times plume_{idmy}$. $SmokePM_{idmy}$ will thus equal zero when there is no plume overhead and will equal the anomaly value when there is smoke overhead. Our approach thus provides a continuous measure of smoke exposure intensity. We note that our approach is unaffected by an overhead smoke plume that does not mix down to the surface; in that case, ground $PM_{2.5}$ anomalies would be zero, and so no smoke $PM_{2.5}$ would be assigned.

Our approach is similar to recent work[51] using interpolated station data and plumes to estimate smoke $PM_{2.5}$. However, given the high spatial variation in smoke exposure and the often large distance between EPA stations, we chose not to interpolate EPA stations. To confirm that just one or a handful of monitoring stations in a given county or metro area can adequately represent temporal variation in smoke exposure in that area, we computed the pairwise correlations between time series of smoke observations in each pair of stations in our data, restricting to stations with at least 1,000 days of data (yielding >85,000 pairwise combinations). We then studied correlation in smoke $PM_{2.5}$ between stations as a function of distance between stations (Extended Data Fig. 1b). Counties in our sample (our main unit of analysis) have an average width of 55 km, and metro areas (used in the Google Trends data, described below) have an average width of 228 km; these widths represent the upper bound on an individual's distance from a monitor in our data, and average distances are probably much smaller given that monitors are purposely placed in populated locations. Median correlations in smoke $PM_{2.5}$ variation are on average $r = 0.84$ and $r = 0.63$ at these distances, suggesting that data from an individual point in a county/metro area is reasonably highly representative of variation elsewhere in the county/metro area. We emphasize that our statistical models exploit this location-specific temporal variation in smoke $PM_{2.5}$, which is unlikely to be affected by spatial bias or unrepresentativeness in average pollution values at some stations[52,53]. Any remaining non-systematic measurement error due to distance from monitors will attenuate our estimated effects of smoke towards zero[54].

Measuring salience and health-protective behaviour. We measure salience and health-protective behaviour using public search query data from Google Trends. The data are accessed using the R package gtrendsR version 1.4.8.9000[55] and are provided as location-, term- and period-normalized indices ranging from 0 to 100, where 0 is the lowest search volume for that term in that location during the chosen period, and 100 is the highest search volume. The data are available at the DMA level (referred to as 'metro' areas by Google Trends), which are geographic regions encompassing television media markets as defined by Nielsen.

We study searches in both English and Spanish, which together are the primary languages spoken by 92% of US households[56]. We use weekly data on DMAs, the native spatial resolution of the public Trends data, between January 2016 and December 2020, and analyse data on terms related to smoke exposure (including 'air quality', 'smoke' and 'wildfire smoke').

Measuring sentiment. We measure online sentiment for a county-day using the text of Twitter posts ('tweets') created in that county on that day. Specifically, we collect nearly all of the geolocated tweets for the continental United States between December 2016 and February 2021 through the Twitter Streaming API, in accordance with the terms and conditions laid out in Twitter's Developer Agreement (https://developer.twitter.com/en/developer-terms/agreement). Per the agreement, the authors cannot make individual tweets available publicly. To compute sentiment for each tweet, we apply the VADER sentiment analysis model[28], a natural language processing algorithm tuned specifically for estimating sentiment from online language. We take the average of the 'compound' scores (ranging between −1 and 1) computed by VADER for all tweets in a county-day as our measurement of sentiment. Our approach builds on the computation of expressed sentiment described in ref. [26]. Readers may refer to that article for additional details on the general approach to collecting and processing tweets for use in empirical analysis. On average, the mean sentiment for a county-day is 0.17, computed from 455.4 tweets.

Measuring mobility. We assembled a daily dataset of mobility measures at the county level collected between January 2019 and December 2020, the period over which mobility data were made available to researchers by SafeGraph. These data measure the aggregate activity of anonymized device signals, or 'pings', at the census block group level. Signals are collected from smartphones, not all cell phones. We focus on two measures constructed from these anonymized signals: the percentage of individuals completely at home on that day, and the percentage of individuals completely away from home on that day. We construct the 'completely away from home' variable by counting the percentage of devices on a given day that were not observed in their respective home location. SafeGraph assigns a home location to each device on the basis of its mobility pattern observed over the previous six weeks. We aggregate these data to the county-day level by taking means weighted by the number of devices in each census block. The data processing details are discussed further in ref. [57].

Measuring distance to fire. To distinguish the effects of exposure to wildfire smoke from potentially correlated effects of being near an active wildfire, we develop daily measures of proximity to active wildfires and test whether the effects of smoke we uncover on outcomes might instead be the direct effects of proximity to fire. We compute 'distance to fire' as the population-weighted average distance from 10 km grid cell centroids within a county to their nearest NOAA HMS fire point(s) and as the distance from a DMA centroid to the nearest fire cluster. Building on earlier work[10], 'fire clusters' are constructed by buffering each HMS fire point by 3 km square and taking the union of existing overlapping squares over a given day and the previous three days, and distance to fire cluster is set to 0 if the

active fire cluster is inside the DMA on that day. This does not mean that $10\,km^2$ are burning, but within that $10\,km^2$ there are multiple fire points over a three-day period, representing an active and potentially growing fire. We emphasize that our goal in this analysis is not to test the independent effect of proximity to wildfire on these outcomes, but to understand whether we're actually isolating smoke impacts or conflating them with fire proximity.

Estimating ambient smoke impact. We combine the above behavioural measures with our smoke $PM_{2.5}$ estimates and analyse their correspondence using panel fixed effects estimators, with the goal of isolating the impact of variation in smoke exposure from other time-invariant and time-varying factors that could be correlated with both smoke exposure and outcomes. Specifically, we estimate econometric models of the form:

$$y_{isdmy} = f(\text{SmokePM}_{isdmy}) + Z_{isdmy} + \alpha_{im} + \eta_d + \varepsilon_{isdmy} \qquad (1)$$

where y_{isdmt} is outcome of interest in unit i, state s, day d, month m and year y; SmokePM_{ismy} is our smoke $PM_{2.5}$ measure on the same day and location; and Z_{ismy} are additional time-varying controls. Our preferred model includes a location-by-month fixed effect α_{im} to account for local seasonality in either outcomes or exposures (for example, one intercept for each of the 12 months in Santa Clara County, California) and a day-of-sample fixed effect η_d (for example, a dummy for 1 January 2016 and so on) to account for common trends or shocks to outcomes or exposures on a given day. Our date fixed effect implicitly also accounts for any average differences in outcomes between weekends and weekdays. We estimate $f()$ using either linear models or more flexible cubic splines to capture potential nonlinearities. In all analyses using search query data, Twitter data or mobility data, smoke $PM_{2.5}$ is measured using EPA station data, as described above.

In these models, the effect of smoke exposure on outcome y is estimated by relating, for example, outcomes in Santa Clara County on 30 August 2020 versus 1 September 2020 to differences in smoke exposure on those days, after accounting for any common difference across counties in exposure or outcomes between the two days, and any average differences in smoke exposure or outcomes in August versus September in Santa Clara County. A confounding variable would have to be a local time-trending unobservable correlated with both smoke exposure and the outcome. Possible candidates include weather variables and the presence of an active wildfire nearby, and we additionally control for these variables (Z_{isdmy} in equation (1)) in robustness tests, or split the sample between locations nearby and further from an active wildfire.

Another potential threat to identification is the COVID-19 pandemic, which near the end of our sample period had demonstrated effects on mobility[58] and sentiment[59], and probably enhanced awareness about the importance of air filtration[60]; 2020 was also a year of severe smoke exposure throughout much of the US West. Because we exploit daily variation in smoke exposure over time at particular locations, and because such variation depends largely on stochastic factors such as exactly where fires ignite and which way the wind is blowing, we believe that daily variation in COVID-19 outcomes or behaviours is unlikely to be spuriously correlated with wildfire smoke exposures. However, to further address this confounding risk, we test robustness to even more stringent time controls, including county-by-month-of-sample fixed effects and state-by-day-of-sample fixed effects; these further account for any state-specific differences or trends in COVID-19 severity and/or policy intervention that happened to coincide with wildfire risk. We note that any changes in our observed behavioural outcomes due to wildfire-specific effects on health outcomes, including wildfire's potential effects on COVID-19 itself[17], are not confounding and would constitute part of the overall 'effect' that we wish to understand.

To study whether the effects of smoke on outcomes vary across locations, we interact smoke exposure with time-invariant covariates of interest:

$$y_{isdmy} = \beta_1 \text{SmokePM}_{isdmy} + \beta_2 \text{SmokePM}_{isdmy} \times X_i + Z_{isdmy} + \alpha_{im} + \eta_d + \varepsilon_{isdmy} \qquad (2)$$

where X_i in our analysis includes median household income, average previous exposure to $PM_{2.5}$ and average smoke $PM_{2.5}$ exposure, included either individually or jointly. Because our analysis is at the county level, and because some covariates (particularly income) could vary substantially within counties, the heterogeneous treatment effects estimated on county data with equation (2) could understate the true underlying heterogeneity in responses to smoke exposure.

Our approach does not allow us to estimate whether individuals respond to smoke $PM_{2.5}$ differently than they do other sources of $PM_{2.5}$. Unlike for smoke $PM_{2.5}$, we do not have a research design that can isolate plausibly exogenous variation in other sources of $PM_{2.5}$. For instance, if traffic is an important daily driver of non-smoke $PM_{2.5}$ in a given location, and traffic volume is correlated with a booming economy, an analysis of the impact of non-smoke $PM_{2.5}$ on any of our outcomes would struggle to separate the impact of the $PM_{2.5}$ itself from the impact of the activity that generated the $PM_{2.5}$. Even if people were unhappy about high $PM_{2.5}$ levels and would otherwise stay home, an analysis could easily find that both sentiment and mobility were higher on high-$PM_{2.5}$ days, as people enjoyed their trips to the office. This confounding is unlikely to be a problem for smoke $PM_{2.5}$, however, as day-to-day variation in smoke exposure (conditional on our controls) is plausibly random.

Measuring indoor and outdoor household $PM_{2.5}$ using PurpleAir. To estimate household infiltration of outdoor $PM_{2.5}$ into indoor environments, we utilize data collected by low-cost PurpleAir monitors. Raw ten-minute observations were downloaded from the PurpleAir servers (available at https://thingspeak.com/) via JSON in accordance with PurpleAir terms and conditions. Data were downloaded from the earliest available date through the end of 2020 or the last available date, whichever is earlier for all available indoor and outdoor PurpleAir monitors in the contiguous United States. Data quality checks were implemented following the procedures utilized in recent studies[38,40] to produce hourly indoor and outdoor $PM_{2.5}$ concentrations. We then followed existing literature and used multiple approaches to estimate $PM_{2.5}$ concentrations from the cleaned PurpleAir data (Supplementary Information).

Hourly ambient exposures were estimated at each indoor monitor site by first identifying all outdoor monitors within 5 km and then taking the inverse distance weighted average of hourly $PM_{2.5}$ concentration across the (up to) ten nearest monitors. Monitors with less than 720 non-missing hourly indoor and outdoor $PM_{2.5}$ measurements (that is, 30 days of hourly data) were excluded from the analysis.

Finally, indoor PurpleAir monitors are deployed in many different types of buildings. We used a combination of information from monitor labels and manual checking of geolocations to determine which buildings with indoor PurpleAir monitors were single-family residences. All other types of buildings were removed from the sample. In total, there were 1,520 indoor monitors reporting in our sample of single-family residences.

Estimating infiltration rates. To estimate the average indoor infiltration rate, which we define as the increase in indoor $PM_{2.5}$ concentration per unit increase in local outdoor $PM_{2.5}$ concentration (that is, $\partial \text{IndoorPM}_{2.5}/\partial \text{OutdoorPM}_{2.5}$), we estimate a regression at the monitor-hour level. Namely, for each residence i in hour h on day-of-week d and month-of-sample m, we estimate how indoor $PM_{2.5}$ varies with contemporaneous and previous hour measurements of outdoor $PM_{2.5}$:

$$\text{PM}^{in}_{ihdm} = \sum_{k=0}^{6} \beta_k \text{PM}^{out}_{i;h-k;dm} + \gamma_i + \delta_h + \eta_d + \theta_m + \varepsilon_{ihdm} \qquad (3)$$

To isolate the contribution of outdoor $PM_{2.5}$ to indoor $PM_{2.5}$ from other time-varying $PM_{2.5}$ sources (most notably, indoor-sourced $PM_{2.5}$), we use fixed effects to flexibly control for time invariant differences across households (γ_i), monthly trends in $PM_{2.5}$ over the sample (θ_m) and household-specific average variation in $PM_{2.5}$ within the day (δ_h). Day-of-week fixed effects (η_d) control for differences in patterns across weekdays and between weekdays and weekends.

We include six lags here (outdoor $PM_{2.5}$ at each of the previous six hours) to account for lingering effects of outdoor concentrations in previous hours on contemporaneous indoor concentrations, although the results are robust to the inclusion of additional lags. From this regression, we derive an estimate for outdoor–indoor infiltration by calculating the cumulative effect of a $1\,\mu g\,m^{-3}$ increase in outdoor concentrations on indoor concentrations:

$$\text{infiltration} = \sum_{k=0}^{6} \beta_k \qquad (4)$$

To assess the importance of modelling structure, we re-estimated equation (3) with four different lag structures: a distributed lag model with lags for outdoor $PM_{2.5}$ only (shown above), a lagged dependent variable model with a lag for indoor $PM_{2.5}$ only, a model with both indoor and outdoor $PM_{2.5}$ lags, and finally a model with no lag terms (Supplementary Table 11). Infiltration rate estimates derived from each of the models are highly similar (Extended Data Fig. 6), and models with more than six lags produce indistinguishable estimates of infiltration rates.

To examine heterogeneity in infiltration rates across hourly outdoor pollution levels and by smoke presence, we first estimate a nonlinear version of equation (3). Namely, we model indoor $PM_{2.5}$ as a fourth-degree polynomial of outdoor $PM_{2.5}$ (and its lags) and interact it with a dummy variable indicating whether smoke was present. The smoke dummy S_{it} is defined as 1 when a NOAA HMS plume reported a smoke plume of any density over the PurpleAir monitor on that day and 0 otherwise, where t indexes day of sample and all hours within a given day are assigned the same value for the smoke dummy:

$$
\begin{aligned}
\text{PM}^{in}_{ihdmt} = &\left[\sum_{k=0}^{6} \beta_k \text{PM}^{out}_{i;h-k;dmt} + \sum_{k=0}^{6} \alpha_k \left(\text{PM}^{out}_{i;h-k;dmt}\right)^2 + \right. \\
&\left. \sum_{k=0}^{6} \nu_k \left(\text{PM}^{out}_{i;h-k;dmt}\right)^3 + \sum_{k=0}^{6} \lambda_k \left(\text{PM}^{out}_{i;h-k;dmt}\right)^4 \right] + \\
&\left[\sum_{k=0}^{6} \beta_{k;s} \text{PM}^{out}_{i;h-k;dmt} + \sum_{k=0}^{6} \alpha_{k;s} \left(\text{PM}^{out}_{i;h-k;dmt}\right)^2 + \right. \\
&\left. \sum_{k=0}^{6} \nu_{k;s} \left(\text{PM}^{out}_{i;h-k;dmt}\right)^3 + \sum_{k=0}^{6} \lambda_{k;s} \left(\text{PM}^{out}_{i;h-k;dmt}\right)^4 \right] \times S_{it} + \\
&\gamma_i + \delta_h + \eta_d + \theta_m + \varepsilon_{ihdmt}
\end{aligned}
\qquad (5)
$$

To measure the infiltration rate, we then calculate the derivative of indoor $PM_{2.5}$ with respect to outdoor $PM_{2.5}$ estimated in equation (5) and use the estimated regression coefficients (β, α, ν, and λ) to evaluate across the 1st–99th percentile of observed hourly outdoor $PM_{2.5}$ concentrations as well as the indicator for whether or not smoke was present. The responses are plotted in Fig. 4a.

We also estimate infiltration rates as a function of median census tract income and smoke by estimating equation (3) with additional income and income-by-smoke interaction terms:

$$PM_{ihdmt}^{in} = \left[\sum_{k=0}^{6} \beta_k PM_{i;h-k;dmt}^{out}\right] + \left[\sum_{k=0}^{6} \beta_{k;i} PM_{i;h-k;dmt}^{out}\right] \times income_i +$$

$$\left[\sum_{k=0}^{6} \beta_{k;i;s} PM_{i;h-k;dmt}^{out}\right] \times income_i \times S_{it} + \gamma_i + \delta_h + \eta_d + \theta_m + \varepsilon_{ihdmt}$$

(6)

The median income data come from the American Community Survey. Each indoor monitor was matched to a census tract, and median income was pulled for the most recent available year and updated to 2020 US dollars. We then similarly calculated the derivative of indoor $PM_{2.5}$ with respect to outdoor $PM_{2.5}$ and evaluated across the 1st–99th percentile of observed $PM_{2.5}$ concentrations. The responses are plotted separately for smoke and non-smoke periods in Fig. 4b.

Finally, for each indoor monitor, we estimated a separate distributed lag model analogous to the pooled model in equation (3):

$$PM_{ihdm}^{in} = \sum_{k=0}^{6} \beta_{ik} PM_{i;h-k;dm}^{out} + \delta_h + \eta_m + \theta_d + \varepsilon_{ihdm}$$

(7)

where $PM_{2.5}$ at indoor monitor i in hour h on day-of-week d and month-of-sample m is modelled as a function of outdoor $PM_{2.5}$ in that location in the contemporaneous period and for each of the previous six hours. Our estimate of the overall infiltration rate for each monitor, which we denote β_i, is then the sum of coefficients over time from the regression for that monitor (that is, $\beta_i = \sum_{k=0}^{6} \beta_{ik}$).

Understanding variation in household infiltration rates. Monitor-specific estimates suggest large variation in infiltration across households (Fig. 4c), consistent with earlier work[40]. However, since monitor-specific infiltration values are themselves estimates from data, the observed variation across monitors could reflect 'true' underlying heterogeneity in infiltration or could simply reflect sampling variation (or some combination of the two).

To distinguish sampling variation from underlying heterogeneity, we estimate a Bayesian hierarchical model[61,62] that models monitor-specific infiltration estimates as being distributed normally about true monitor-specific infiltration values with estimated monitor-specific sampling variance $\hat{\beta}_i \sim N(\beta_i, s.\hat{e}._i^2)$ and true monitor-specific infiltration values as drawn from an underlying normal distribution with unknown mean and variance $\beta_i \sim N(\beta, \sigma^2)$.

We then train flexible machine-learning-based models to predict monitor-specific infiltration rates from matched covariates (Supplementary Information). We divide our sample into a 75% training dataset and a 25% held-out test dataset, splitting train and test within 13 disjoint geographic regions covering the contiguous United States to ensure a geographically balanced split. We train random forest and gradient boosted trees models with manually tuned forest and boosting hyperparameters, respectively, and tree parameters tuned using random search with threefold cross-validation repeated five times. We conduct tuning and training for each method of matching monitors and CoreLogic houses for robustness.

We report performance statistics (R^2) on held-out test data and compute the marginal effect of each predictor by evaluating the predicted effect in the test data of moving from the 5th percentile to the 95th percentile of the predictor, with all other variables fixed at their mean values. We repeat this evaluation for both random forest and gradient boosted trees models, and for all four ways of spatially matching to housing characteristics.

Understanding the representativeness of the study samples. See the Supplementary Information for a discussion of sample representativeness across our multiple datasets.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The data to replicate all the results in the main text and supplementary material are available at https://github.com/echolab-stanford/wildfire-exposure-behavior-public.

Code availability
The code to replicate all the results in the main text and supplementary material is available at https://github.com/echolab-stanford/wildfire-exposure-behavior-public.

References
1. Landrigan, P. J. et al. The Lancet Commission on pollution and health. *Lancet* **391**, 462–512 (2018).
2. Carleton, T. A. & Hsiang, S. M. Social and economic impacts of climate. *Science* **353**, aad9837 (2016).
3. Aizer, A., Currie, J., Simon, P. & Vivier, P. Do low levels of blood lead reduce children's future test scores? *Am. Econ. J. Appl. Econ.* **10**, 307–41 (2018).
4. Deryugina, T., Miller, N., Molitor, D. & Reif, J. Geographic and socioeconomic heterogeneity in the benefits of reducing air pollution in the United States. *Environ. Energy Policy Econ.* **2**, 157–189 (2021).
5. Burke, M., Hsiang, S. M. & Miguel, E. Global non-linear effect of temperature on economic production. *Nature* **527**, 235–239 (2015).
6. Grönqvist, H., Nilsson, J. P. & Robling, P.-O. Understanding how low levels of early lead exposure affect children's life trajectories. *J. Polit. Econ.* **128**, 3376–3433 (2020).
7. US Department of Health and Human Services *Theory at a Glance: A Guide for Health Promotion Practice* (National Cancer Institute, 2005).
8. Greenstone, M. & Jack, B. K. Envirodevonomics: a research agenda for an emerging field. *J. Econ. Lit.* **53**, 5–42 (2015).
9. Abatzoglou, J. T. & Williams, A. P. Impact of anthropogenic climate change on wildfire across western US forests. *Proc. Natl Acad. Sci. USA* **113**, 11770–11775 (2016).
10. Burke, M. et al. The changing risk and burden of wildfire in the United States. *Proc. Natl Acad. Sci. USA* **118**, e2011048118 (2021).
11. Hurteau, M. D., Westerling, A. L., Wiedinmyer, C. & Bryant, B. P. Projected effects of climate and development on California wildfire emissions through 2100. *Environ. Sci. Technol.* **48**, 2298–2304 (2014).
12. Liu, J. C. et al. Particulate air pollution from wildfires in the western US under climate change. *Climatic Change* **138**, 655–666 (2016).
13. Goss, M. et al. Climate change is increasing the likelihood of extreme autumn wildfire conditions across California. *Environ. Res. Lett.* **15**, 094016 (2020).
14. Reid, C. E. et al. Critical review of health impacts of wildfire smoke exposure. *Environ. Health Perspect.* **124**, 1334–1343 (2016).
15. Cascio, W. E. Wildland fire smoke and human health. *Sci. Total Environ.* **624**, 586–595 (2018).
16. Xu, R. et al. Wildfires, global climate change, and human health. *N. Engl. J. Med.* **383**, 2173–2181 (2020).
17. Zhou, X. et al. Excess of COVID-19 cases and deaths due to fine particulate matter exposure during the 2020 wildfires in the United States. *Sci. Adv.* **7**, eabi8789 (2021).
18. Heft-Neal, S., Driscoll, A., Yang, W., Shaw, G. & Burke, M. Associations between wildfire smoke exposure during pregnancy and risk of preterm birth in California. *Environ. Res.* **203**, 111872 (2021).
19. Santana, F. N., Gonzalez, D. J. & Wong-Parodi, G. Psychological factors and social processes influencing wildfire smoke protective behavior: insights from a case study in Northern California. *Clim. Risk Manage.* **34**, 100351 (2021).
20. Rappold, A. et al. Smoke Sense initiative leverages citizen science to address the growing wildfire-related public health problem. *GeoHealth* **3**, 443–457 (2019).
21. Reid, C. E. et al. Differential respiratory health effects from the 2008 northern California wildfires: a spatiotemporal approach. *Environ. Res.* **150**, 227–235 (2016).
22. Kondo, M. C. et al. Meta-analysis of heterogeneity in the effects of wildfire smoke exposure on respiratory health in North America. *Int. J. Environ. Res. Public Health* **16**, 960 (2019).
23. Wen, J. & Burke, M. Wildfire smoke exposure worsens learning outcomes. Preprint at *EarthArXiv* https://doi.org/10.31223/X52H06 (2021).
24. *Wildfire Smoke: A Guide for Public Health Officials, 2019 Revision* (US Environmental Protection Agency, 2019).
25. Pellert, M., Metzler, H., Matzenberger, M. & Garcia, D. Validating daily social media macroscopes of emotions. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2108.07646 (2021).
26. Baylis, P. Temperature and temperament: evidence from Twitter. *J. Public Econ.* **184**, 104161 (2020).
27. Baylis, P. et al. Weather impacts expressed sentiment. *PLoS ONE* **13**, e0195750 (2018).
28. Hutto, C. & Gilbert, E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. *Proc. Int. AAAI Conf. Web Soc. Media* **8**, 216–225 (2014).
29. Wang, Z., Ye, X. & Tsou, M.-H. Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Nat. Hazards* **83**, 523–540 (2016).
30. Sachdeva, S., McCaffrey, S. & Locke, D. Social media approaches to modeling wildfire smoke dispersion: spatiotemporal and social scientific investigations. *Inform. Commun. Soc.* **20**, 1146–1161 (2017).

31. Choi, H. & Varian, H. Predicting the present with Google Trends. *Econ. Rec.* **88**, 2–9 (2012).

32. Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. & Watts, D. J. Predicting consumer behavior with web search. *Proc. Natl Acad. Sci. USA* **107**, 17486–17490 (2010).

33. Liang, Y. et al. Wildfire smoke impacts on indoor air quality assessed using crowdsourced data in California. *Proc. Natl Acad. Sci. USA* **118**, e2106478118 (2021).

34. Miller, K. A. et al. Estimating ambient-origin $PM_{2.5}$ exposure for epidemiology: observations, prediction, and validation using personal sampling in the Multi-Ethnic Study of Atherosclerosis. *J. Expo. Sci. Environ. Epidemiol.* **29**, 227–237 (2019).

35. Shrestha, P. M. et al. Impact of outdoor air pollution on indoor air quality in low-income homes during wildfire seasons. *Int. J. Environ. Res. Public Health* **16**, 3535 (2019).

36. Uejio, C. et al. Summer indoor heat exposure and respiratory and cardiovascular distress calls in New York City, NY, US. *Indoor Air* **26**, 594–604 (2016).

37. Ferguson, L. et al. Exposure to indoor air pollution across socio-economic groups in high-income countries: a scoping review of the literature and a modelling methodology. *Environ. Int.* **143**, 105748 (2020).

38. Bi, J., Wallace, L. A., Sarnat, J. A. & Liu, Y. Characterizing outdoor infiltration and indoor contribution of $PM_{2.5}$ with citizen-based low-cost monitoring data. *Environ. Pollut.* **276**, 116763 (2021).

39. Allen, R. W. et al. Modeling the residential infiltration of outdoor $PM_{2.5}$ in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Environ. Health Perspect.* **120**, 824–830 (2012).

40. Krebs, B., Burney, J., Zivin, J. G. & Neidell, M. Using crowd-sourced data to assess the temporal and spatial relationship between indoor and outdoor particulate matter. *Environ. Sci. Technol.* **55**, 6107–6115 (2021).

41. Lu, J. G. Air pollution: a systematic review of its psychological, economic, and social effects. *Curr. Opin. Psychol.* **32**, 52–65 (2020).

42. Rappold, A. G. et al. Cardio-respiratory outcomes associated with exposure to wildfire smoke are modified by measures of community health. *Environ. Health* **11**, 71 (2012).

43. Brulle, R. J. & Pellow, D. N. Environmental justice: human health and environmental inequalities. *Annu. Rev. Public Health* **27**, 103–124 (2006).

44. Hajat, A. et al. Air pollution and individual and neighborhood socioeconomic status: evidence from the Multi-Ethnic Study of Atherosclerosis (MESA). *Environ. Health Perspect.* **121**, 1325–1333 (2013).

45. Zheng, S., Wang, J., Sun, C., Zhang, X. & Kahn, M. E. Air pollution lowers Chinese urbanites' expressed happiness on social media. *Nat. Hum. Behav.* **3**, 237–243 (2019).

46. deSouza, P. & Kinney, P. L. On the distribution of low-cost $PM_{2.5}$ sensors in the US: demographic and air quality associations. *J. Expo. Sci. Environ. Epidemiol.* **31**, 514–524 (2021).

47. Sun, C., Kahn, M. E. & Zheng, S. Self-protection investment exacerbates air pollution exposure inequality in urban China. *Ecol. Econ.* **131**, 468–474 (2017).

48. Currie, J., Voorheis, J. & Walker, R. What caused racial disparities in particulate exposure to fall? New evidence from the Clean Air Act and satellite-based measures of air quality. Preprint at *National Bureau of Economic Research* https://doi.org/10.3386/w26659 (2020).

49. Di, Q. et al. An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution. *Environ. Int.* **130**, 104909 (2019).

50. Reid, C. E., Considine, E. M., Maestas, M. M. & Li, G. Daily $PM_{2.5}$ concentration estimates by county, zip code, and census tract in 11 western states 2008–2018. *Sci. Data* **8**, 112 (2021).

51. O'Dell, K., Ford, B., Fischer, E. V. & Pierce, J. R. Contribution of wildland-fire smoke to US $PM_{2.5}$ and its influence on recent trends. *Environ. Sci. Technol.* **53**, 1797–1804 (2019).

52. Grainger, C. & Schreiber, A. Discrimination in ambient air pollution monitoring? *AEA Pap. Proc.* **109**, 277–282 (2019).

53. Fowlie, M., Rubin, E. & Walker, R. Bringing satellite-based air quality estimates down to earth. *AEA Pap. Proc.* **109**, 283–288 (2019).

54. Wooldridge, J. M. *Introductory Econometrics: A Modern Approach* (Cengage Learning, 2015).

55. Massicotte, P. & Eddelbuettel, D. gtrendsR, R package version 1.4.8.9000 https://github.com/PMassicotte/gtrendsR (2021).

56. *Characteristics of People by Language Spoken at Home 2019: 2015–2019 American Community Survey 5-Year Estimates* (US Census Bureau, 2020).

57. Squire, R. F. Measuring and correcting sampling bias in SafeGraph patterns for more accurate demographic analysis. *Safegraph* https://www.safegraph.com/blog/measuring-and-correcting-sampling-bias-for-accurate-demographic-analysis (2019).

58. Weill, J. A., Stigler, M., Deschenes, O. & Springborn, M. R. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proc. Natl Acad. Sci. USA* **117**, 19658–19660 (2020).

59. Valdez, D., Ten Thij, M., Bathina, K., Rutter, L. A. & Bollen, J. Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of Twitter data. *J. Med. Internet Res.* **22**, e21418 (2020).

60. Morawska, L. et al. How can airborne transmission of COVID-19 indoors be minimised? *Environ. Int.* **142**, 105832 (2020).

61. Meager, R. Understanding the average impact of microcredit expansions: a Bayesian hierarchical analysis of seven randomized experiments. *Am. Econ. J. Appl. Econ.* **11**, 57–91 (2019).

62. Vivalt, E. How much can we generalize from impact evaluations? *J. Eur. Econ. Assoc.* **18**, 3045–3089 (2020).

63. *2019 Tiger/Line Shapefiles* (US Census Bureau, 2019).

64. *GADM Data* (GADM, 2018).

Acknowledgements

Author contributions

All authors contributed to the conception and design of the study. S.H.-N., J.L., A.D., J.W. and M.L.C. constructed the smoke data. J.L. and A.D. constructed the Google search data. P.B. constructed the Twitter dataset. M.S. and J.A.W. constructed the mobility dataset. S.H.-N. constructed the PurpleAir dataset. M.B. and S.H.-N. led the econometric analysis. All authors contributed to analysing the results and writing the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended Data Fig. 1 | Counties included in analyses that use EPA pollution monitors, and correlation in smoke PM$_{2.5}$ between monitor pairs as a function of distance between monitors. a. Counties in red are those with EPA pollution monitors from which we construct smoke PM$_{2.5}$ measures for the behavioral analyses. **b** Colors depict a heatmap of the 85,102 pairwise correlations, with lighter colors depicting areas with more data and shown in legend at right; solid black line is the median correlation at each distance. Sample is restricted to stations with at least 1000 days of data. Mean width of counties in our data is 55 km, and mean width of metro areas is 228 km. Source for **a**: US Census Bureau.
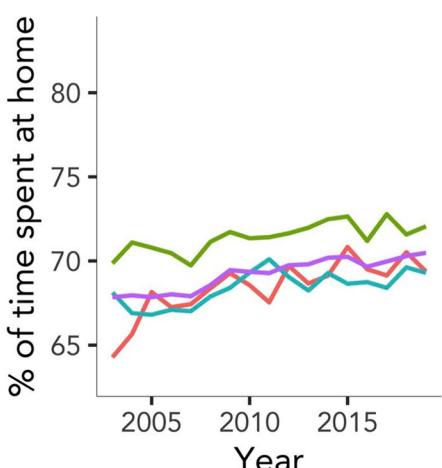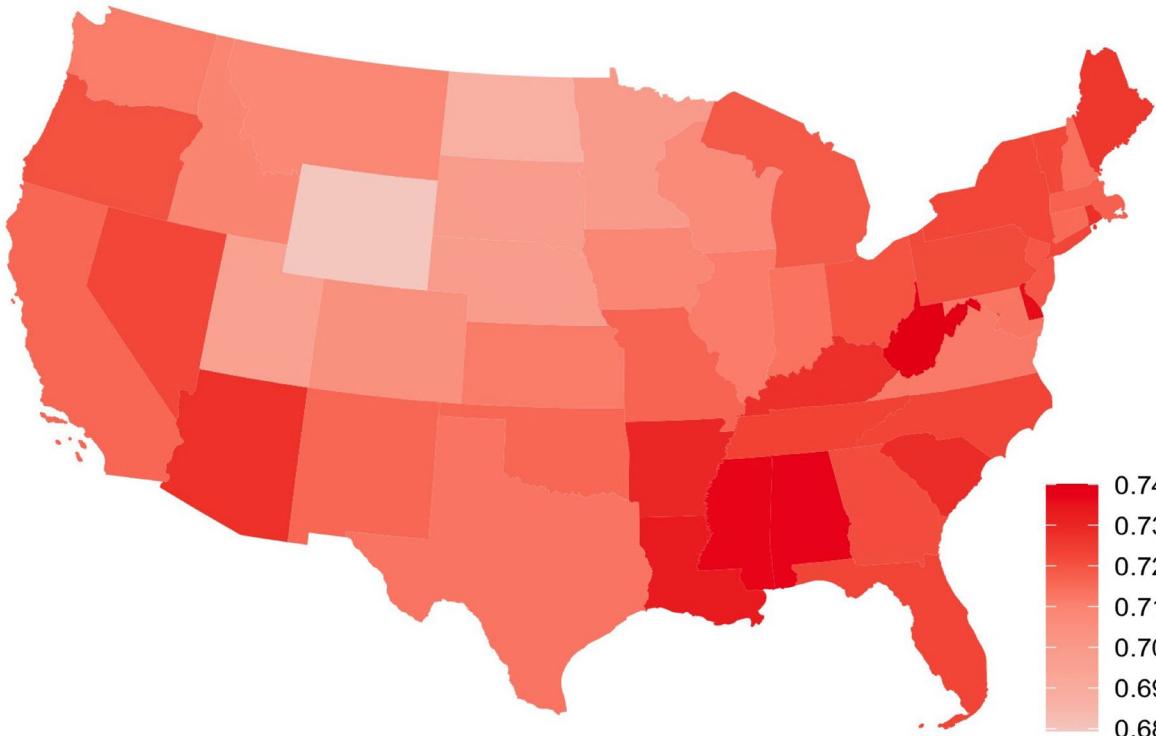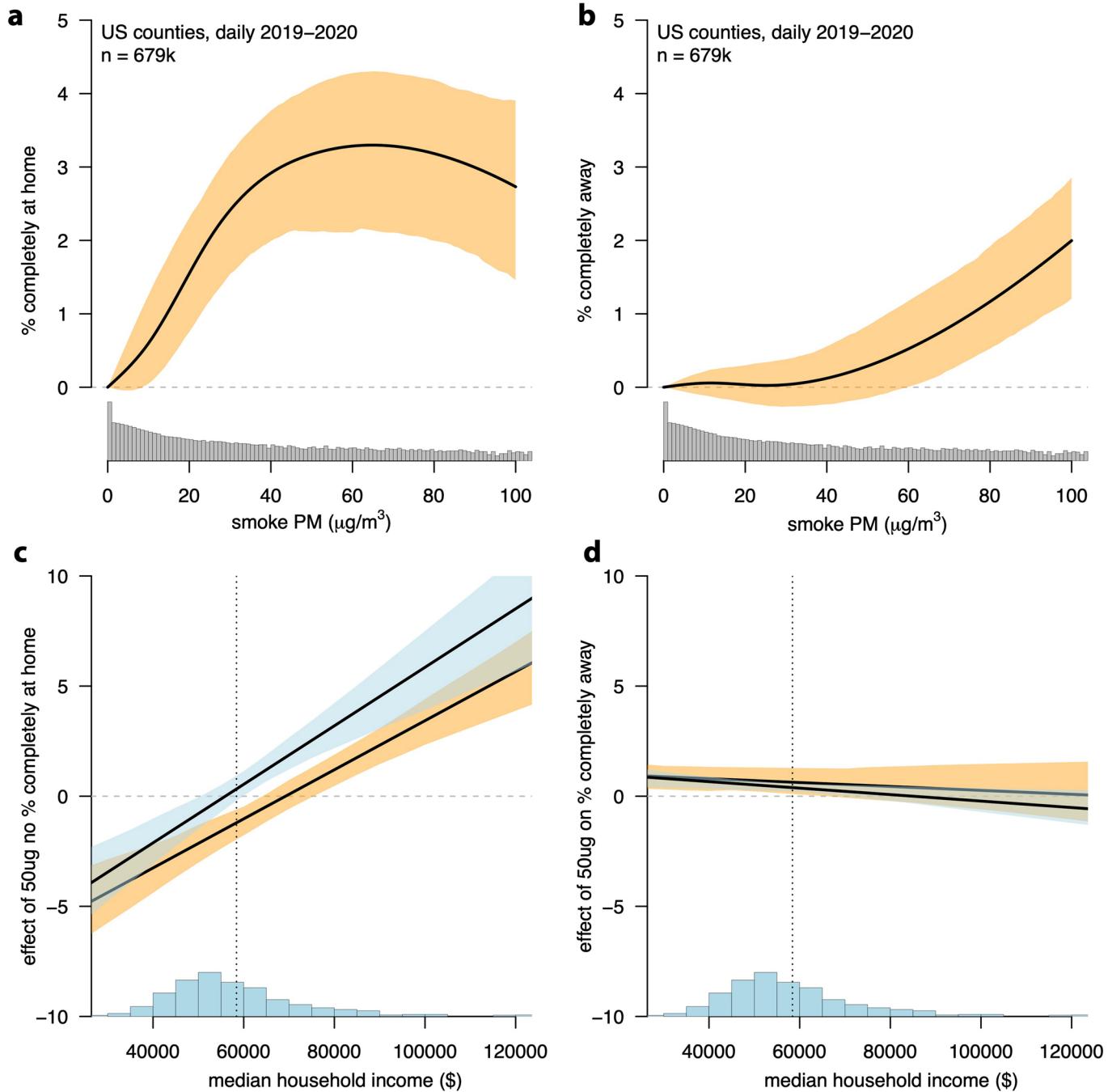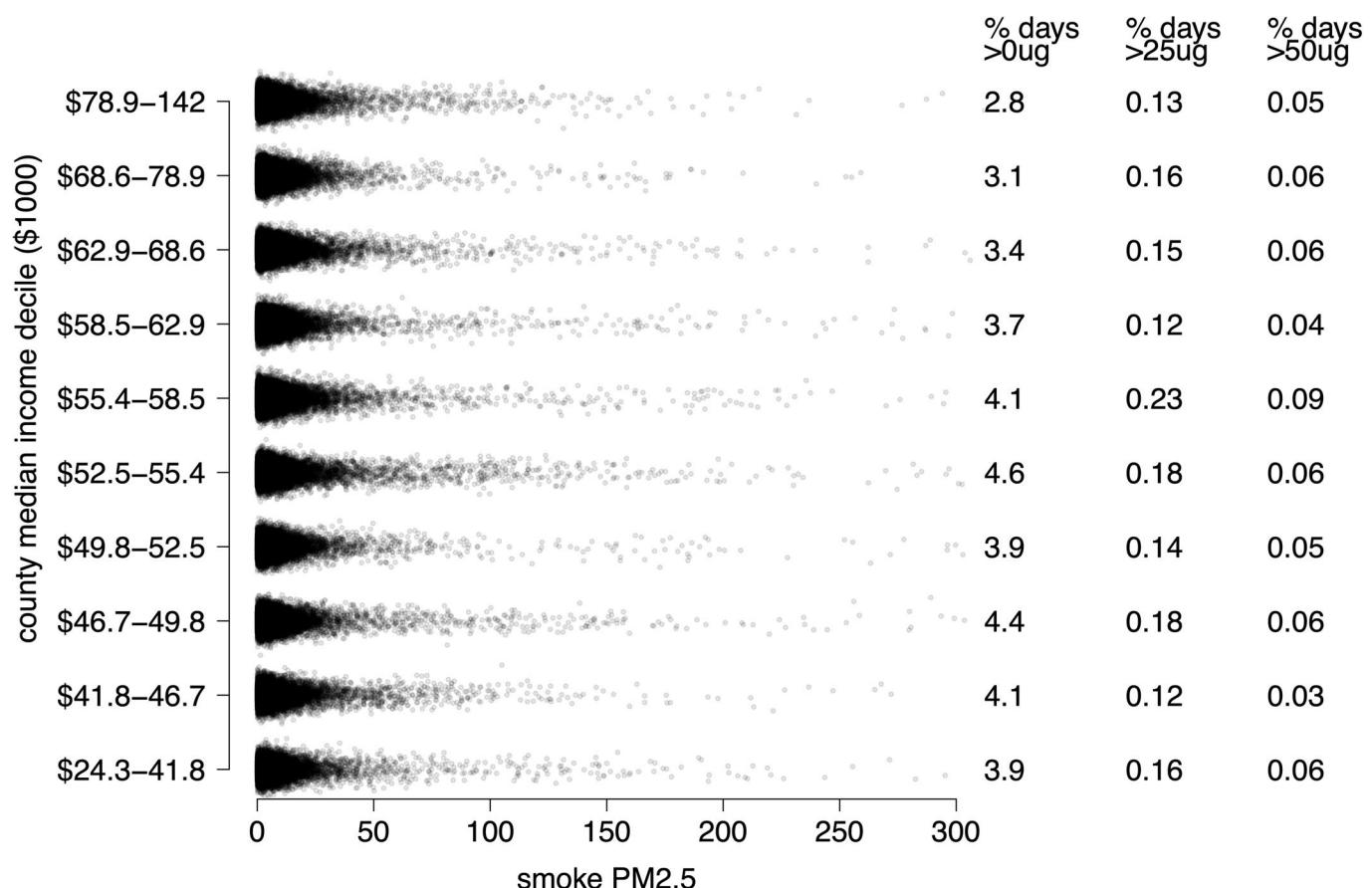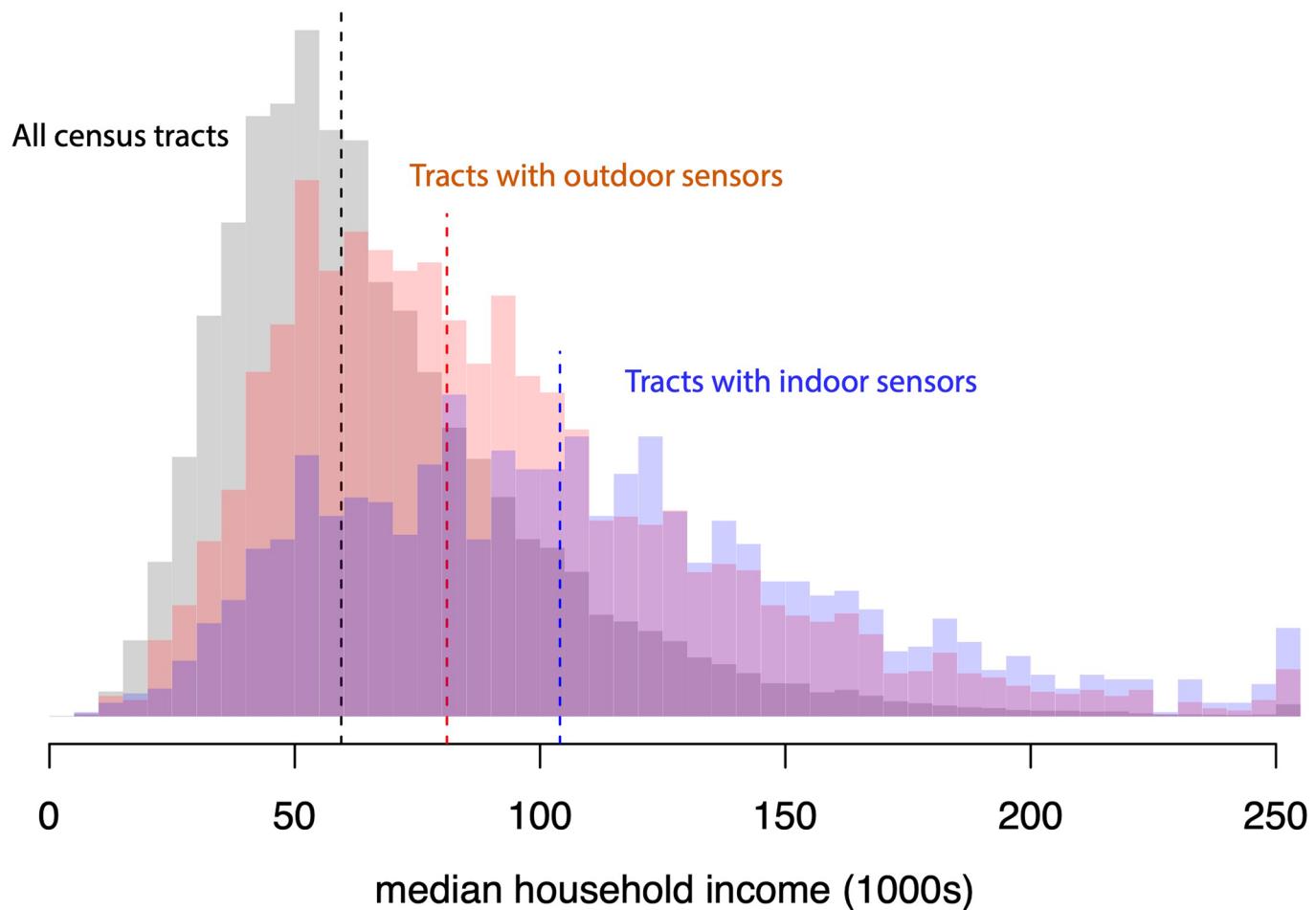
Extended Data Fig. 2 | Time spent indoors at home in America. Data are from repeated rounds of the American Time Use Survey. Top panels show data by average income, age, season, and race/ethnicity. Bottom map shows averages by state across survey rounds. Source for **e**: US Census Bureau.
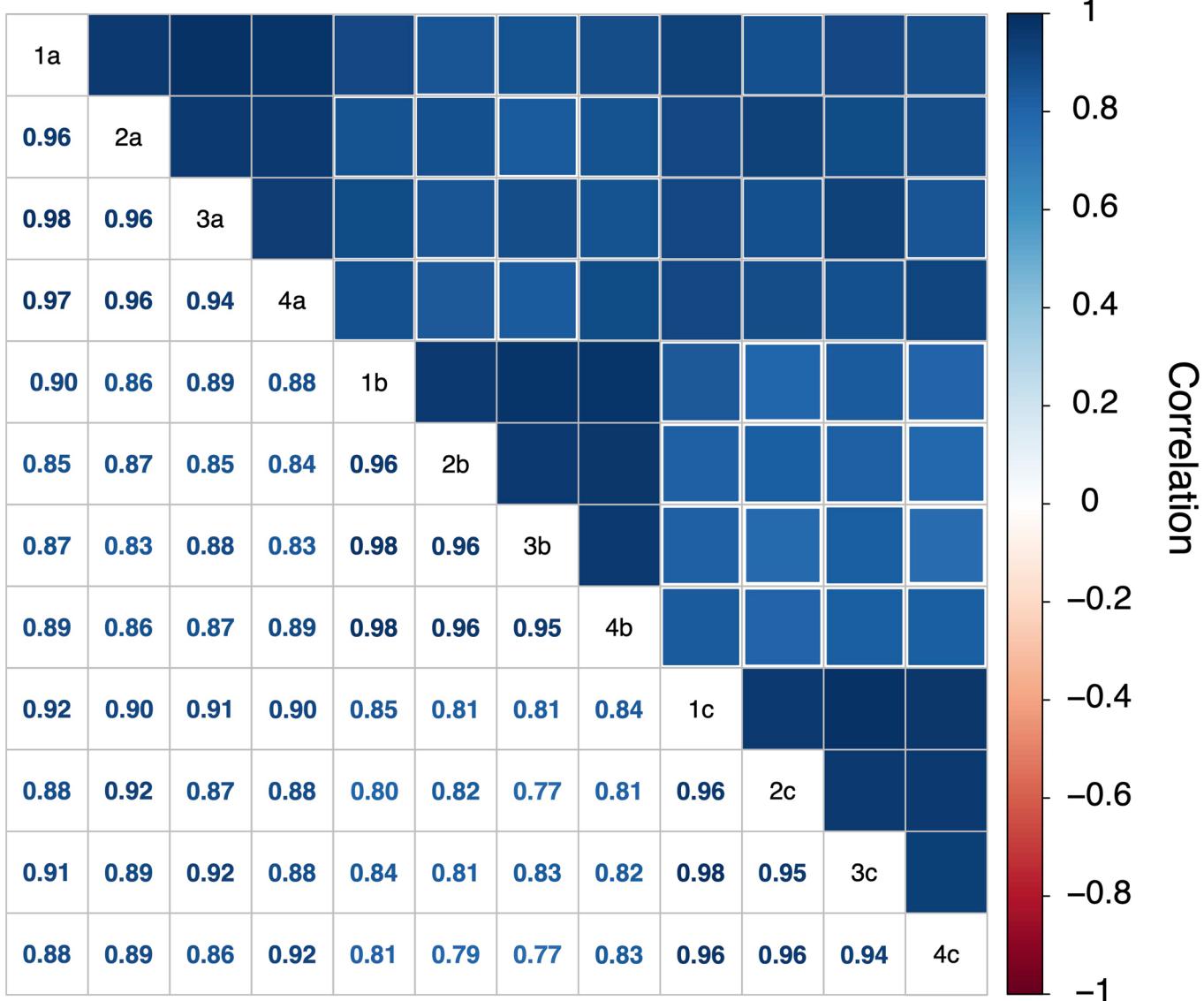
Extended Data Fig. 3 | Effect of smoke PM$_{2.5}$ on different mobility measures, and heterogeneity by income. a Percent of mobile phones estimated to be completely at home on a given day at the US county level, 2019–2020. Black lines are regression point estimates from spline fits conditional on fixed effects, with shaded areas showing bootstrapped 95% confidence intervals. Number of observations in each regression is shown in upper left corner of each panel. Histograms at the bottom show the log distribution of smoke PM$_{2.5}$ exposure in each sample. **b** Same but for % fully away from home on that day. **c-d** Effect of smoke PM$_{2.5}$ on mobility as a function of income. Lines show the marginal effect of a heavy smoke exposure (50ug smoke PM$_{2.5}$ on that day) on percent of individuals completely at home on that day (c) or completely away from home (d), as a function of median household income in that county. Colors represent models run with either date fixed effects (blue) or state-by-date fixed effects (orange). Dark lines show regression point estimates, shaded area the bootstrapped 95% CI.

Extended Data Fig. 4 | Exposure to average and acute smoke PM$_{2.5}$ at the county level does not differ systematically by income. Daily smoke PM$_{2.5}$ exposures by income decile across US counties, 2006-2020. Dots represent daily observations where smoke PM$_{2.5}$ was non-zero. Plot is truncated at 300ug for clarity; not plotted are 71 days (0.001% of the sample) in which smoke PM$_{2.5}$ exceeded 300. Statistics at right show the percent of observations across the study period with daily smoke PM$_{2.5}$ observations above the listed value.

Extended Data Fig. 5 | Higher income US census tracts are more likely to have PurpleAir monitors. Grey bars show the distribution of tract-level median household income across all US census tracts in the contiguous US, red bars the income in tracts with at least one outdoor PurpleAir sensor, and blue bars the income in tracts with at least one indoor PurpleAir sensor. Vertical lines give the median of each distribution.
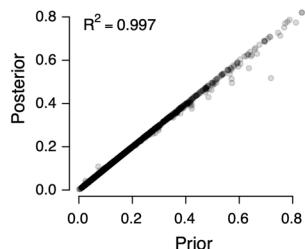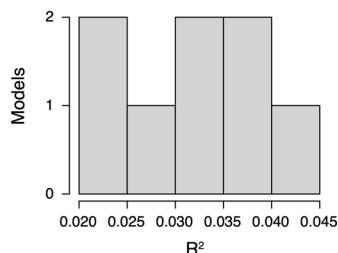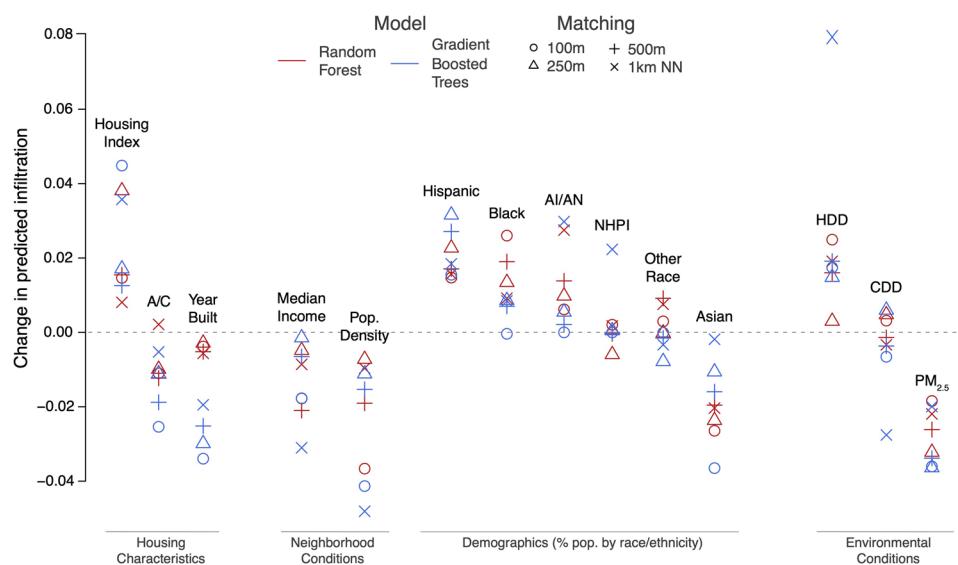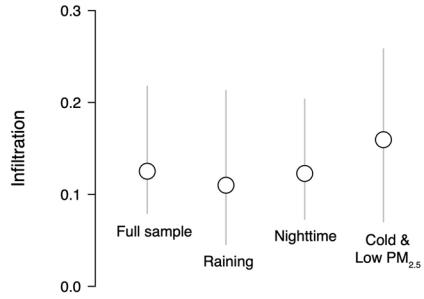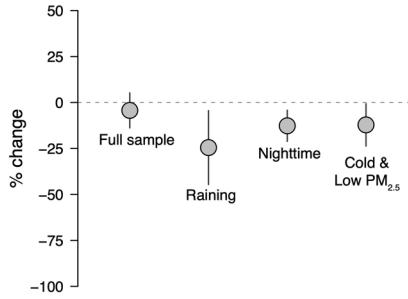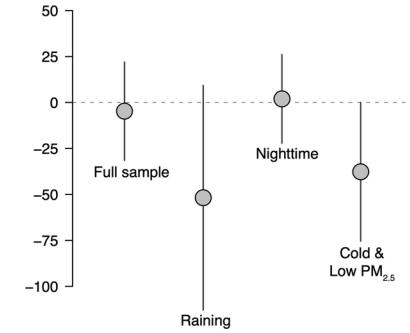
Model

1. Distributed Lag
2. Lagged Dependent Variable
3. No Lags
4. Both Dependent and Independent Variable Lags
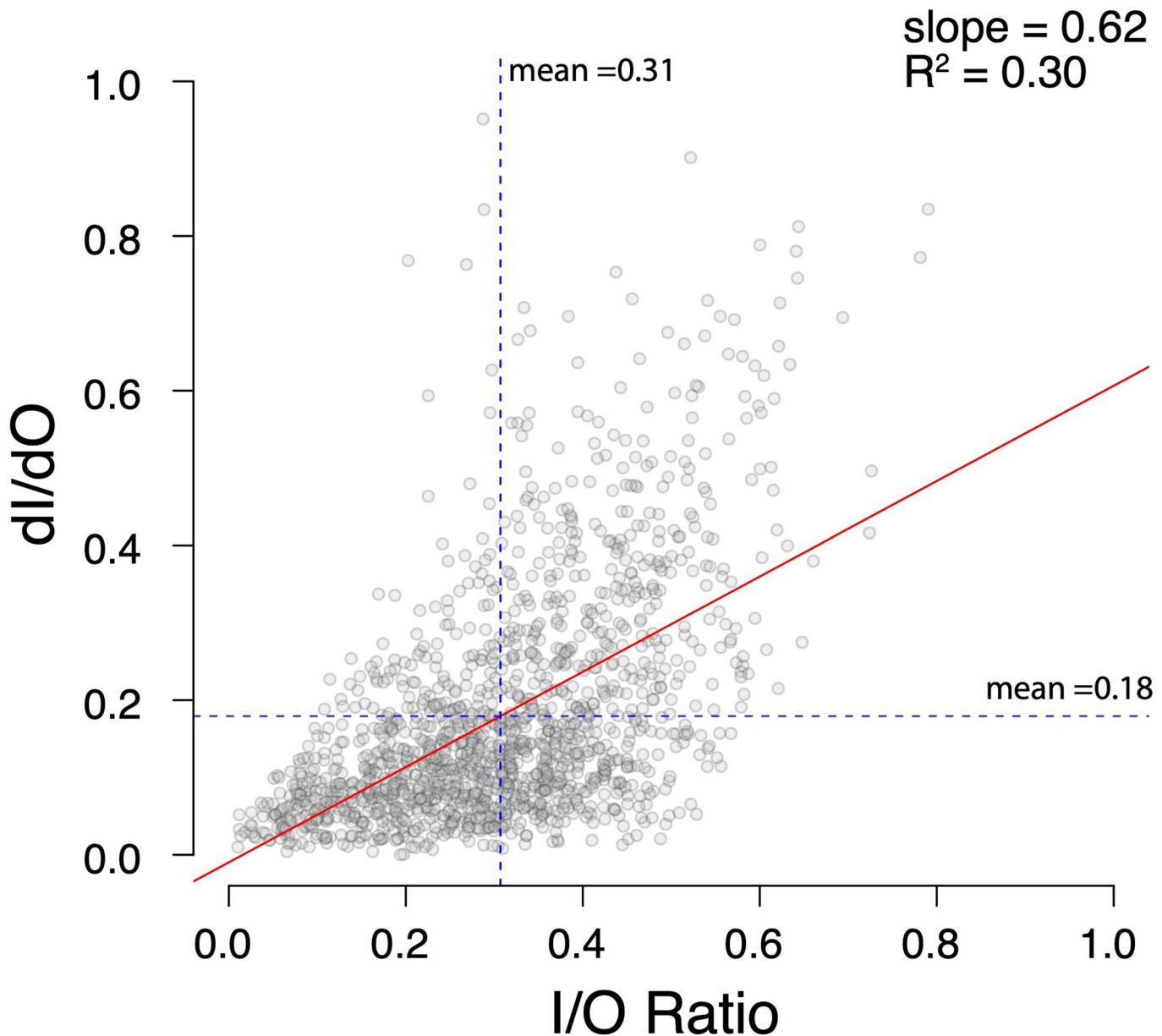
PM$_{2.5}$ Measurements

a. Derived from particle counts
b. Barkjohn et al 2021 corrected
c. No correction

Extended Data Fig. 6 | Infiltration estimates are highly correlated across alternate statistical models and methods of deriving PM$_{2.5}$ concentrations from Purple Air data. Correlation between infiltration estimates from statistical models with different lag structures and different PM$_{2.5}$ concentration estimates (see Supplementary Table 11 for details). 1a is our preferred specification presented in the main results.
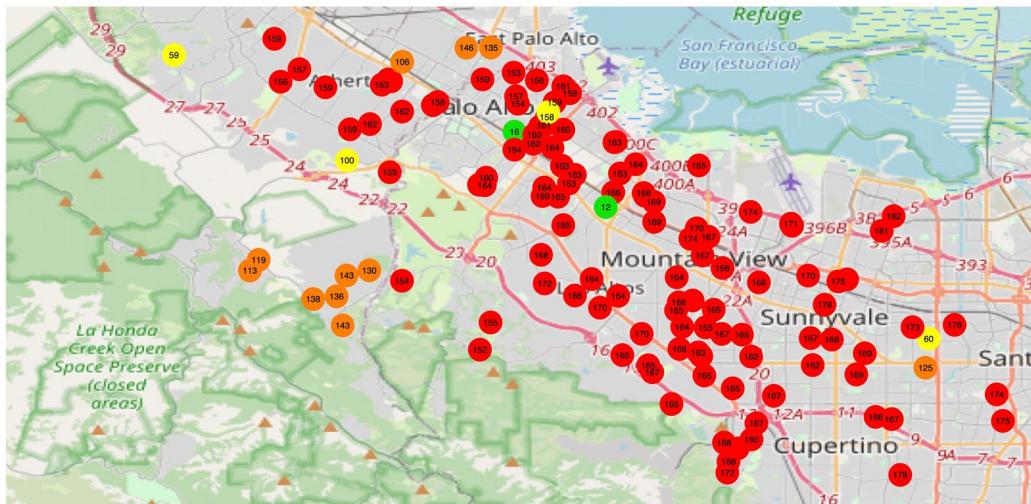
a. Raw vs Bayesian hierarchical estimates

c. Explainers of household level infiltration



b. Explained variation in infiltration

**d. Inter–quartile range of infiltration estimates**

e. % change in infiltration per + $100K income

**f. % change in infiltration if home built after 2008**

Extended Data Fig. 7 | Understanding variation in household-specific infiltration estimates. a Posterior estimates of monitor-specific infiltration rates from a Bayesian hierarchical model are very similar to "raw" estimates from our monitor-specific time-series regressions, indicating that true heterogeneity rather than sampling noise is what is driving observed differences in estimated infiltration. **b** Ability of random forest (RF) or gradient boosted trees (GBT) model to explain variation (r2) in infiltration across monitors remains low; models use predictors in (**c**). **c** For each predictor, we calculate the effect on infiltration of moving from the 5th to the 95th percentile of that predictor in the test dataset, holding the other predictors constant at their average value in the test dataset; estimates are shown for RF and GBT models and for four alternate spatial buffers used to construct housing predictors. Housing Index is constructed by averaging standardized values of home value, number of stories, number of baths, number of bedrooms, height, and area. A/C measures the inverse distance weighted proportion of matched CoreLogic houses that have air conditioning. Median Income is the median household income in the Census tract population. Race variables (i.e. all demographic covariates except Hispanic) are measured among the non-Hispanic/Latino population. AI/AN stands for American Indian and Alaska Native. NHPI stands for Native Hawaiian and other Pacific Islander. HDD and CDD stand for heating degree days and cooling degree days, respectively. **d** Range of household-level infiltration estimates for the full sample and for sub-samples when behavior (ie opening/closing of doors, use of air purifier) is expected to matter less: when it's raining, nighttime, and periods when it is cold ($<10^{\circ}$C) and low $PM_{2.5}$ ($<30\mu g/m^3$). **e-f** during periods when behavioral factors are more likely minimized, infiltration varies more strongly with income and housing age.

Extended Data Fig. 8 | Monitor-specific infiltration estimates using indoor/outdoor ratios versus regression-based approaches. I/O estimates are only modestly correlated with our preferred regression-based estimates that measure the marginal effect on indoor $PM_{2.5}$ concentrations of a unit increase in outdoor concentrations. For each monitor I/O ratio was calculated across all observations with hourly indoor $PM_{2.5}$ less or equal to outdoor $PM_{2.5}$. $\delta I/\delta O$ was estimated as described in Methods.

Outdoor Monitors
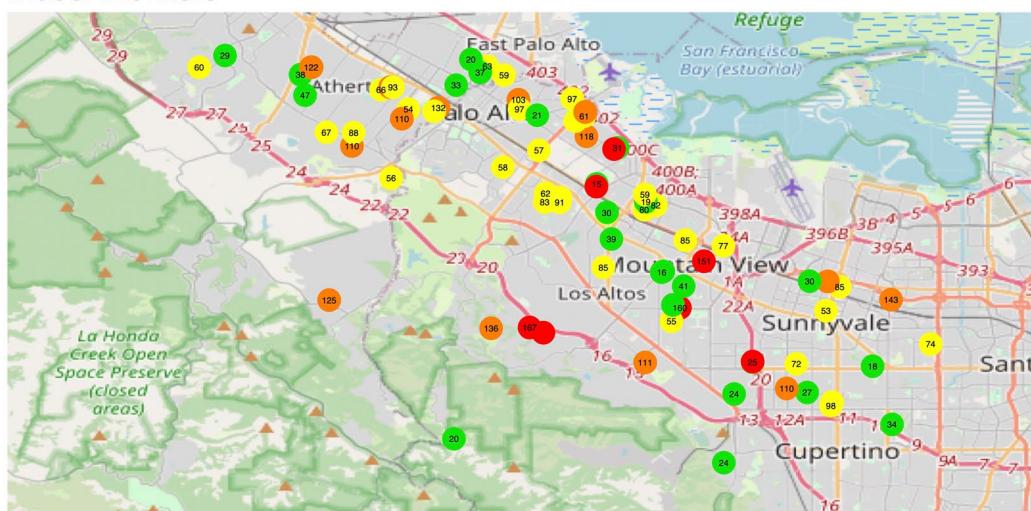


Indoor Monitors
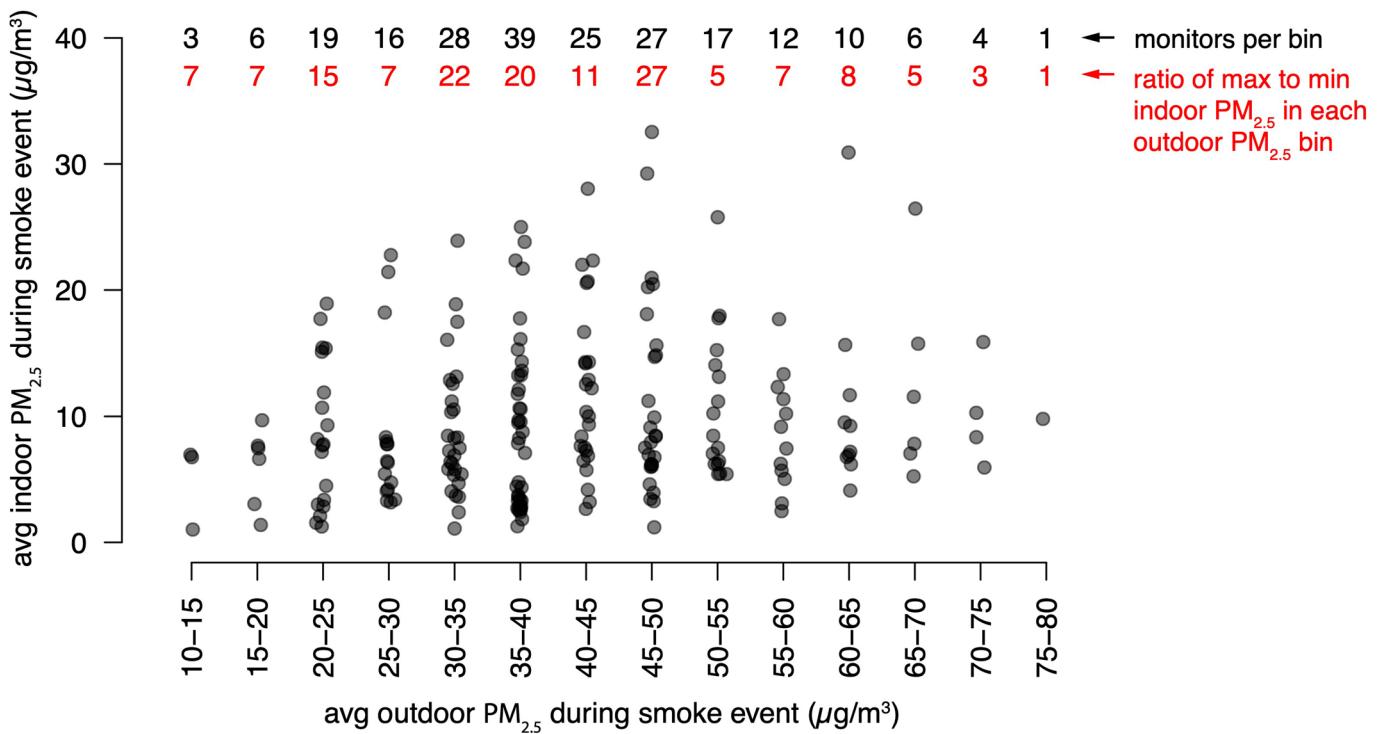
August 19, 2020 1PM
AQI Category

- 🟢 Good
- 🟡 Moderate
- 🟠 Unhealthy for Sensitive Groups
- 🔴 Unhealthy
- 🟣 Very unhealthy
- ⚫ Hazardous

Extended Data Fig. 9 | Outdoor and indoor PM$_{2.5}$ concentrations on a smoke day in CA. Very similar outdoor PM$_{2.5}$ concentrations during a smoke event on Aug 20th, 2021 over a high-income area of the peninsular Bay Area were associated with widely varying contemporaneous indoor PM$_{2.5}$ concentrations. © OpenStreetMap contributors.

Extended Data Fig. 10 | Variation in indoor PM$_{2.5}$ across monitors with similar outdoor PM$_{2.5}$ during the Aug/Sep 2020 smoke event in the Bay Area.
Each dot is average outdoor PM$_{2.5}$ and average indoor PM$_{2.5}$ for an individual monitor in the Bay Area over the Aug/Sep 2020 smoke event, with monitors grouped into 5 μg/m^3 bins based on outdoor exposure. Numbers at top show the number of monitors in each bin (black), and the ratio of maximum to minimum indoor PM$_{2.5}$ across monitors within each outdoor PM$_{2.5}$ bin (red). Monitors with average outdoor PM$_{2.5}$ exposures within 5 μg/m^3 of each other experienced > 20x differences in indoor PM$_{2.5}$ exposures.

nature portfolio

Corresponding author(s): Marshall Burke

Last updated by author(s): May 6, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All data collection code was written in R v4.0 by the research team.
Data analysis	All analysis code was written in R v4.0 by the research team. All code are available at https://github.com/echolab-stanford/wildfire-exposure-behavior-public

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data to replicate all results in the main text and supplementary material are available at https://github.com/echolab-stanford/wildfire-exposure-behavior-public

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences   ☒ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative observational study
Research sample	Research sample changes based on the analysis, but is clearly described in the methods for each dataset. Datasets include a random sample of all individuals using Google search, the universe of individuals using Twitter, all individuals with smartphones, or individuals with indoor air pollution monitors.
Sampling strategy	Samples were convenience samples of all individuals using google search, twitter, mobile phones, or indoor air pollution monitors, as described in paper methods.
Data collection	Data were collected via computer.
Timing	Sample periods roughly range from 2016-2021, with the mobile phone data covering only the 2019-2021 period.
Data exclusions	No data were excluded.
Non-participation	NA
Randomization	NA

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	Antibodies
☒	Eukaryotic cell lines
☒	Palaeontology and archaeology
☒	Animals and other organisms
☒	Human research participants
☒	Clinical data
☒	Dual use research of concern

Methods

n/a	Involved in the study
☒	ChIP-seq
☒	Flow cytometry
☒	MRI-based neuroimaging