medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

ORIGINAL ARTICLE

Estimating excess mortality during heat waves using a causal inference framework and non-randomized time series

Yasmine Mabene[1*] | Lauren Mock[2*] | Luke Miratrix[3†] | Young Lee[4†] | Marie-Abèle Bind[5,6†]

[1]Department of Computer Science, Stanford University, Stanford, CA, 94305, USA

[2]Department of Biostatistics, Harvard T.H. Chan School of Public Heath, Boston, MA, 02115, USA

[3]Harvard University Graduate School of Education, Cambridge, MA, 02138, USA

[4]Department of Statistics and Data Science, National University of Singapore, Singapore

[5]Biostatistics Center, Massachusetts General Hospital, Boston, MA, 02114, USA

[6]Department of Medicine, Harvard Medical School, Boston, MA, 02115, USA

Correspondence
Yasmine Mabene, Department of Computer Science, Stanford University, Stanford, CA, 94305, USA
Email: ymabene@stanford.edu

Funding information
Office of the Director, National Institutes of Health, Grant/Award Number: DP5OD021412. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Researchers must often rely on observational data to draw conclusions about the causal relationship between an exposure and an outcome. This is common in environmental epidemiology research, where the assignment to an exposure can be unethical or impossible. Consider a setting in which we seek to estimate excess mortality during heat waves, but are unable to randomize individuals to moderate or extreme heat. We propose to follow a causal framework for non-randomized time series that uses what we call *historical time series matching* to construct hypothetical randomized experiments, in which background covariates are balanced. We applied this framework and assessed the following causal question: Are daily mortality counts among elderly Americans higher during hot versus warm summer time periods? We addressed this causal question by drawing frequentist, Fisherian, and Bayesian inferences for five American cities. We estimated the average treatment effect of heat on three-day mortality with 95% Fisherian intervals to be 0.5 [-3.6, 4.4] deaths in Pittsburgh, 3.7 [0.7, 6.6] deaths in Seattle, 12.5 [6.1, 18.9] deaths in Chicago,

*† Equally contributing authors.

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

8.5 [1.4, 15.6] deaths in Los Angeles, and 20.3 [11.2, 29.5] deaths in New York.

KEYWORDS
Environmental epidemiology, climate change, heat waves, Fisherian inference, Bayesian inference, time series

1 | INTRODUCTION

Global climate change is a substantial threat to human health [1]. In the United States, the average temperature has increased between 1.3 and 1.9°F since record-keeping began in 1895, with a total projected increase of 3°F to 10°F by the end of this century [2]. In addition to increases in average temperature, extreme heat events have become increasingly frequent and intense. The Intergovernmental Panel on Climate Change attributes increases in heat-related mortality to these heat events [3], which are the leading cause of weather-related mortality in the United States [4]. Extreme heat can interfere with the body's temperature regulation, resulting in the exacerbation of chronic conditions and the onset of several new health complications, leading to morbidity and mortality [2]. Although extreme heat events are becoming more widespread, the human health impacts of rising temperatures are not evenly distributed. Extreme heat events disproportionately affect the elderly population and marginalized groups [1,2,5,6]. Those living in urban areas are especially vulnerable due to the urban heat island effect, which explains why dense cities are hotter than surrounding rural areas [7,8]. Additionally, the synergistic health effects of high temperatures and air pollution further contribute to elevated risks in these areas [4].

Previous epidemiological studies have identified a J-, V-, or U-shaped association between temperature and mortality [9,10]. Several studies have linked rising temperatures with increased risks for natural, cardiovascular, and respiratory mortality [5,11,12], and others indicate that daily mortality counts increase during heat waves [13,14].

Determining the temperature-mortality relationship from medical records may underestimate the number of temperature related deaths due to lack of consistent diagnoses [2]. More commonly, temperature-mortality studies utilize time series and case-crossover methods [15]. The time series approach involves comparing mortality outcomes at different exposure levels for a set of intervals over a period of time. These studies frequently incorporate distributed lag nonlinear models and generalized additive models to investigate the relationship between temperature and mortality [16–18]. Typically, smoothing functions such as natural splines are used to control for potential confounding factors including seasonality and humidity. However, existing correlations between nonlinear functions or small regression coefficients can lead to biased estimates in these models [19]. Additionally, *a posteriori* selection for the parametric form of the functions in these models inflates type I errors and results in undercoverage in confidence intervals [20–23].

The case-crossover design resembles a case-control study, but cases serve as their own reference, allowing for the control of time-invariant, individual-level confounding factors. When examining temperature-mortality relationships, case-crossover studies often rely on estimated coefficients from conditional logistic regression models to estimate causal effects [24,25]. However, these models can produce misleading findings as a result of selection bias and overlap bias [24,26–28]. Estimates based on maximum likelihood estimation may also be heavily biased for studies with small sample sizes or with high or low prevalence risk factors [24].

To estimate the causal effect of hot temperatures on mortality, we rely on the Neyman-Rubin Causal Model framework, in which we depart from traditional association modeling, and instead design the observational study prior to its analysis [29,30]. We obtain daily measurements of mortality, meteorological variables, and air pollution concentrations

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

in five major cities in the United States. We conceptualize a plausible, hypothetical randomized experiment in which three-day time periods are randomly assigned to be exposed or unexposed to extreme heat. We then identify comparable exposed and unexposed time periods with respect to certain covariates selected *a priori*, implement a maximum bipartite matching algorithm to identify the best 1:1 pairs of exposed and unexposed time periods, and conduct our analysis on these matched pairs [31–34]. By following this framework, we follow a preferred method for determining causal effects [35], limiting the bias that occurs due to confounding in traditional time series approaches [36]. Additionally, the randomization-based inference approach does not depend on asymptotic distributional assumptions [37,38].

After designing this experiment, we assess the temperature-mortality relationship with frequentist, Fisherian, and Bayesian approaches. Our study provides estimates and uncertainty intervals for the causal effect of extreme temperatures on mortality from each approach. Furthermore, the causal inference framework implemented here can be utilized in other observational studies to estimate the effects of environmental exposures on health outcomes [32,33,39].

2 | METHODS: A CAUSAL INFERENCE FRAMEWORK

We obtained daily counts of mortality, maximum and minimum temperatures, dew point temperature, relative humidity, and air pollution concentrations (i.e., O_3, NO_2, CO, SO_2, PM_{10}, and $PM_{2.5}$) between 1987 and 2000 in five American cities: Pittsburgh, Seattle, Chicago, Los Angeles, and New York. These data were collected through the National Morbidity, Mortality, and Air Pollution Study (NMMAPS). Total daily mortality counts in the NMMAPS dataset are grouped into three age categories: under age 65, between ages 65 and 74, and 75 years or older. As the elderly are much more susceptible to extreme heat events than younger people [6], the study population was limited to those aged 65 and older. The number of deaths due to certain specific causes was available but not incorporated into this study, which focuses on all-cause mortality. Other characteristics of those who died each day were unavailable. The study's time period was limited to days in June, July, August, and September; all days during these months will be referred to as "summer" days. Of the 8,540 summer days between 1987 and 2000 across five cities, the total mortality count among those 65 and older was unavailable for 11 days, and the maximum temperature was unavailable for one day. These days were removed from the analysis. We followed a causal inference framework [30,33,34,38,39], adapted for observational time series [32].

2.1 | Stage 1: Conceptual Intervention

We seek to assess the following causal question: Are daily mortality counts among elderly Americans higher during hot versus warm summer time periods? Because we cannot conduct an experiment that randomly assigns summer days to be either hot or warm, we instead set up a plausible, *hypothetical* randomized experiment and consider the Neyman-Rubin Causal Model [29,30,32,39,40]. In this analysis, we estimate effect sizes for each city separately to account for city-specific differences in age structure, weather patterns, air pollution, and cooling infrastructure.

Because heat events are typically defined by multiple consecutive days with higher-than-usual temperatures, we investigate three-day time periods. We refer to each rolling three-day period comprised of Day 1, Day 2, and Day 3 as an "experimental unit." We refer to the days preceding each experimental unit as lag days, counting backwards; the day before an experimental unit is denoted by Lag 1, two days before the experimental unit by Lag 2, etc.

Consider a period of three consecutive summer days denoted by i (where $i \in \{1, ..., N\}$) in city c. For any day i in city c, we denote the maximum daily temperature by $\text{temp.max}_{i,c}$ and a dichotomous, city-specific treatment by $W_{i,c}$. We define a "treated" or hot time period in a given city c to be three consecutive summer days, for which $\text{temp.max}_{i,c}$

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

is above that city's summer median daily maximum temperature. We define a "control" or warm time period in a given city c to be three consecutive summer days, for which temp.max$_{i,c}$ is below or equal to that same threshold. Any other three-day period is considered to have an undefined treatment and is excluded from our hypothetical randomized experiment. See Figure 1 for examples of control, treated, and undefined periods.

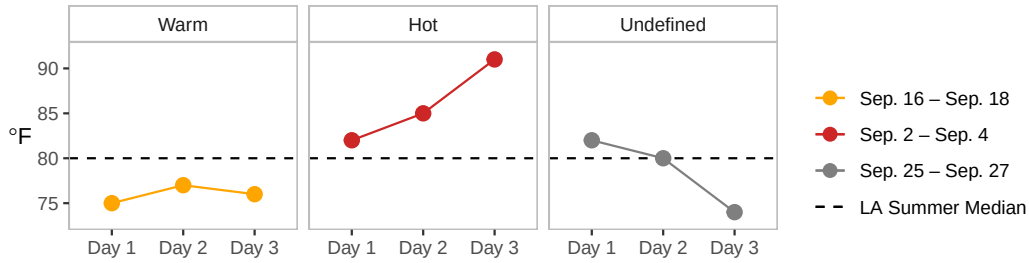


FIGURE 1 Examples of warm (control), hot (treated), and undefined three-day periods in 1990 in Los Angeles as defined in Equation 1.

Our dichotomous treatment indicator, $W_{i,c}$, is then defined in the following way for each rolling three-day period i in city c:

$$W_{i,c} = \begin{cases} 0 & \text{if temp.max}_{\text{i,c}} \leq \text{summer median (temp.max}_{i,c}) \text{ on Day 1, Day 2, and Day 3} \\ 1 & \text{if temp.max}_{\text{i,c}} > \text{summer median (temp.max}_{\text{i, c}}) \text{ on Day 1, Day 2, and Day 3.} \end{cases} \quad (1)$$

The median daily maximum temperatures (temp$_{\text{max}}$) across summer days in each city are reported in Table 2. While Pittsburgh, Chicago, Los Angeles and New York have similar medians around 80°F, Seattle has a notably lower median of 72°F.

$Y_{i,c}$, our outcome of interest, is the total number of deaths among people aged 65 and older in a given city c during a three-day period i. If a three-day period is a control unit (warm), then the treatment indicator $W_{i,c} = 0$, and the observed number of deaths is $Y_{i,c}(W_{i,c} = 0)$. Similarly, if the three-day period is a treated unit (hot), then $W_{i,c} = 1$, and the observed number of deaths is $Y_{i,c}(W_{i,c} = 1)$.

2.2 | Stage 2: Design of the observational study

In the next stage of this framework, we seek to obtain a subset of the data, for which we can assume that treatment assignment is unconfounded given the measured covariates, $X_{i,c}$; i.e., $P(W_{i,c} \mid Y_{i,c}(W_{i,c} = 0), Y_{i,c}(W_{i,c} = 1), X_{i,c}) = P(W_{i,c} \mid X_{i,c})$ [41]. In other words, we seek to identify time periods for which the treatment assignment is independent of the joint potential outcomes, conditional on a set of observed confounders [37].

Here, we use matching to approximate the unconfoundedness assumption. We identify groups of treated and control units with similar lag days, meaning that the days preceding the treated and control three-day periods are similar with respect to observed potential confounders. For each potential confounder, we set the maximum discrepancy allowed between treated and control periods in order to consider them comparable. These thresholds are listed in Table 5. Stricter matching requirements lead to better covariate balance between treated and control units, but fewer matched pairs to evaluate our causal question.

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

We then employ a maximum bipartite matching algorithm[32–34] to identify the most similar pairs of treated and control units among all those that meet the requirements set in Table 5. Our study population is therefore limited to the pairs of treated and control units that are selected through this matching process[31–34]. This causal inference step, which we call "historical time series matching," ensures that the days preceding matched treated and control units are as similar as possible, thus mimicking a randomized experiment. This algorithm accounts for mortality differences by year, season, and day of the week, as well as potential lagged effects of temperature, humidity, and pollutants on mortality. For example, on August 8th, 9th, and 10th 1987, the daily maximum temperatures in Chicago were each above Chicago's summer median maximum temperature (80°F), making these three days a "treated" time period. We aim to match this treated period with a corresponding control period—three consecutive summer days that were each below or equal to 80°F in Chicago. This control period must span the same three days of the week (i.e., Saturday, Sunday, and Monday) as the treated period and also fall within three weeks of the treated period. This limits potential matching control periods to July 18th—20th, July 25th—27th, and August 1st—3rd 1987 in the three weeks preceding the treated period, and August 15th—17th, August 22nd—24th, and August 29th—31st 1987 in the three weeks following the treated period. In addition, we require NO_2 and O_3 concentrations to be within 5 $\mu g/m^3$ on the single days preceding the three-day treated and control periods (Lag 1). If none of the six potential control periods meet these criteria, the treated period cannot be matched. If a single control period fits these criteria, it is matched with the treated period. If multiple control periods fit these criteria, the maximum bipartite matching algorithm identifies the control period that is most similar to the treated period, and only the control period is matched with the treated period. We then limit our analysis to only those treated and control units that were matched, with equal numbers of treated and control units.

2.3 | Stage 3: Statistical Analysis

After designing the observational study in this hypothetical randomized experiment, we apply three inferential frameworks to assess our causal question: frequentist (Neyman), randomization-based (Fisher), and Bayesian (Rubin). In each framework, we assume that the stable unit treatment value assumption (SUTVA)[42] holds, which means the treatment assignments for the experimental units cannot affect the observed outcomes for any given unit and each experimental unit can only receive its given treatment (i.e., hot vs. warm).

2.3.1 | Frequentist inference

Neymanian inference

Estimands: For the i^{th} time period in a given city c, the unit-level causal treatment effect on the additive scale, $\tau_{i,c}$, is defined as follows:

$$\tau_{i,c} = Y_{i,c}(W_{i,c} = 1) - Y_{i,c}(W_{i,c} = 0). \tag{2}$$

However, we are unable to observe both $Y_{i,c}(W_{i,c} = 1)$ and $Y_{i,c}(W_{i,c} = 0)$ for a single period; this is the fundamental problem of causal inference[43]. Recall that if a period is hot, the number of deaths we observe during that period is $Y_{i,c}(W_{i,c} = 1)$; if it is warm, we observe $Y_{i,c}(W_{i,c} = 0)$. For this reason, we instead estimate the average difference between the number of deaths we would have observed had time period i in city c been hot (treated) and the number of deaths we would have observed had the same time period i in city c been just warm (control), i.e., the average treatment effect in each city, which we denote by τ_c. τ_c is also the average difference between the potential

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

outcomes $Y_{i,c}(W_{i,c}=1)$ and $Y_{i,c}(W_{i,c}=0)$ across all periods i ($i \in \{1,...,N\}$) in a city c:

$$\tau_c = \frac{1}{N_c}\sum_{i=1}^{N_c}\left(Y_{i,c}(W_{i,c}=1) - Y_{i,c}(W_{i,c}=0)\right) = \bar{Y}_c(1) - \bar{Y}_c(0). \tag{3}$$

Point estimate: The estimated average treatment effect in city c, $\hat{\tau}_c$, is the difference between the average number of deaths during hot periods and the average number of deaths during warm periods:

$$\hat{\tau}_c = \frac{1}{N_{1,c}}\sum_{i:W_{i,c}=1} Y_{i,c}(W_{i,c}=1) - \frac{1}{N_{0,c}}\sum_{i:W_{i,c}=0} Y_{i,c}(W_{i,c}=0) = \bar{y}^{\text{obs}}_{1,c} - \bar{y}^{\text{obs}}_{0,c}, \tag{4}$$

where $N_{1,c}$ is the number of matched treated periods in city c and $N_{0,c}$ is the number of matched control periods in city c. Because we use one-to-one matching, $N_{1,c} = N_{0,c} = N_c/2$, where N_c is the total number of matched periods in city c.

For the k^{th} pair of treated and control time periods (i.e., $k \in \{1,...,N_c/2\}$), let $d_{k,c}$ denote the observed city-specific difference between the number of deaths during the treated time period (denoted by $y^{\text{obs}}_{k,1,c}$) and the control time period (denoted by $y^{\text{obs}}_{k,0,c}$):

$$d_{k,c} = y^{\text{obs}}_{k,1,c} - y^{\text{obs}}_{k,0,c}.$$

Using the above notations, the estimated average treatment effect $\hat{\tau}_c$ can also be expressed as $\bar{d}_c$, the average of $d_{1,c},...,d_{N_c/2}$:

$$\hat{\tau}_c = \frac{1}{\frac{N_c}{2}}\sum_{k=1}^{\frac{N_c}{2}} d_{k,c}. \tag{5}$$

Sampling variance estimator: Previous papers have determined the variance properties of $\hat{\tau}_c$ [44–46], allowing us to construct uncertainty intervals for this estimate. Let $s^2_{d,c}$, the sample variance of $d_{1,c},...,d_{k,c}$ be defined in the following way:

$$s^2_{d,c} = \frac{1}{\frac{N_c}{2}-1}\sum_{k=1}^{\frac{N_c}{2}}(d_{k,c}-\bar{d}_c)^2. \tag{6}$$

If the treatment effect is the same for each pair k, an unbiased estimate of $\text{var}(\hat{\tau}_c)$ is:

$$\widehat{\text{var}}(\hat{\tau}_c)_{\text{paired}} = \frac{s^2_{d,c}}{\frac{N_c}{2}}. \tag{7}$$

Note that if the treatment effect is not the same for each pair, $\frac{s^2_{d,c}}{\frac{N_c}{2}}$ is a conservative estimate of $\widehat{\text{var}}(\hat{\tau}_c)_{\text{paired}}$. Assuming a constant unit-level treatment effect, we can also estimate the variance without considering pairing:

$$\widehat{\text{var}}(\hat{\tau}_c)_{\text{unpaired}} = \frac{s^2_{1,c}}{N_{1,c}} + \frac{s^2_{0,c}}{N_{0,c}}, \tag{8}$$

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

where $s^2_{1,c} = \frac{1}{N_{1,c}-1} \sum_{i:W_{i,c}=1} \left(Y_{i,c}(W_{i,c}=1) - \bar{y}^{\text{obs}}_{1,c}\right)^2$ and $s^2_{0,c} = \frac{1}{N_{0,c}-1} \sum^{N_c}_{i:W_{i,c}=0} \left(Y_{i,c}(W_{i,c}=0) - \bar{y}^{\text{obs}}_{0,c}\right)^2$.

Uncertainty interval: We estimate 95% confidence intervals for τ_c as:

$$\hat{\tau} \pm 1.96 \sqrt{\widehat{\text{var}}(\hat{\tau}_c)}. \tag{9}$$

Poisson Regression

Historically, there has been a trend toward employing regression analysis to uncover causal effects. During this time, researchers began to shift their focus away from meticulous planning and thoughtful consideration, favoring instead the quicker and more straightforward solutions offered by statistical methods. These methods involved regressing an observed health outcome on an environmental exposure of interest and background covariates (see discussions in [47]). One alternative frequentist approach is to use Poisson regression, which models the count of deaths over a three-day period. For each city, we fit three separate Poisson regression models:

Model 1. A model with only treatment as an explanatory variable:

$$\log(E[Y_{i,c}]) = \beta_{0,c} + \beta_{1,c} W_{i,c}. \tag{10}$$

Model 2. A model using treatment and pair as explanatory variables:

$$\log(E[Y_{i,c}]) = \beta_{0,c} + \beta_{1,c} W_{i,c} + \sum_{k=1}^{\frac{N_c}{2}-1} \beta_{k,c} I[\text{pair}_i = k] \tag{11}$$

Model 3. A model using treatment, pair, month, as well as O_3 and NO_2 concentrations on the day preceding the three-day period (Lag 1) as explanatory variables:

$$\begin{aligned}\log(E[Y_{i,c}]) = \beta_{0,c} + \beta_{1,c} W_{i,c} + \sum_{k=1}^{N_c/2-1} \beta_{k,c} I[\text{pair}_i = k] + \sum_{m=1}^{3} \beta_{\text{month},c} I\left[\text{month}_{i,c} = m\right] + \\ \beta_{\text{O}_{3,\text{lag1}},c}(\text{O}_{3,\text{lag1},i,c}) + \beta_{\text{NO}_{2,\text{lag1}},c}(\text{NO}_{2,\text{lag1},i,c})\end{aligned} \tag{12}$$

We estimate 95% confidence intervals for the risk ratios, $e^{\beta_{1,c}}$, using the estimated Poisson regression coefficients and standard errors $\exp\left(\hat{\beta}_{1,c} \pm 1.96 \sqrt{\widehat{\text{var}}(\hat{\beta}_{1,c})}\right)$.

2.3.2 | Fisherian inference

Fisher-exact p-value: We also use a Fisherian approach to test city-specific sharp null hypotheses ($H_{0,c}$) stating that the three-day city-specific mortality is the same during hot and warm time periods. Using this framework, we proceed with a stochastic proof by contradiction of the city-specific sharp null hypotheses [48,49], which are defined as the following:

$$H_{0,c} : Y_{i,c}(W_{i,c}=1) - Y_{i,c}(W_{i,c}=0) = 0. \tag{13}$$

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Thus, we assume that the observed and unobserved potential outcomes for each three-day period in a given city c are the same and seek a stochastic, empirical contradiction.

As our test statistic, we choose $\hat{\tau}_c$, the estimated average treatment effect in city c defined in Equation (4) in the Neymanian inference section. We construct the null randomization distribution, i.e., the set of test statistics calculated under $H_{0,c}$ for all possible treatment allocations in the hypothetical city-specific experiment. Because the experimental units (i.e., the sets of three-day hot and warm days in a given city c) are paired using one-to-one matching, we assume a paired-randomization design, such that there are $2^{\frac{N_c}{2}}$ possible treatment allocations, where $N_c/2$ is the number of matched pairs for each city. Running all possible treatment allocations is quite computationally intensive, thus we approximate the null randomization distribution using 100,000 treatment allocations.

Using the null randomization distribution, we calculate the one-sided Fisher-exact p-value, which is defined as the following:

$$\text{p-value} = \frac{1}{N_{\text{rand}}}\left(\sum_{\text{rep}=1}^{N_{\text{rand}}} \mathrm{I}(\hat{\tau_c}^{\text{rep}} \geq \hat{\tau_c}^{\text{obs}})\right) \tag{14}$$

where N_{rand} is the total number of calculated test statistics under $H_{0,c}$ (here, $100,000$), $\hat{\tau_c}^{rep}$ is the estimated average treatment effect in city c after randomization, and I is the indicator function (i.e., $\mathrm{I}(x) = 1$ if x is true, 0 otherwise). The Fisher-exact p-value is the proportion of test statistics as or more extreme than the observed test statistic, $\hat{\tau}_c^{obs}$.

Fisherian intervals: To determine the range of effect sizes consistent with our data, we compute Fisherian intervals (FIs) for a common additive treatment effect [38]. Unlike (Neymanian) confidence intervals, which provide estimated uncertainty of average treatment effects, Fisherian intervals consist of a range of city-specific *unit*-level treatment effects $\tau_{i,c}$ that are assumed constant [32]. Similar to Fisher-exact p-values, Fisherian intervals do not rely on asymptotic assumptions.

Additive scale: Fisher intervals consist of a range of effect sizes, τ_{diff}, for which the associated sharp hypothesis, stating that the unit-level causal effect is equal to τ_{diff}, is not rejected. To obtain Fisherian intervals, we generate null randomization distributions and calculate Fisher-exact p-values for a set of sharp hypotheses contrasting the unit-level potential outcomes on the additive scale [41]. We use the following hypothesis to calculate Fisherian intervals for $\tau_{i,c}$ defined in Equation 2:

$$Y_i(W_i = 1) - Y_i(W_i = 0) = \tau_{\text{diff}}. \tag{15}$$

Multiplicative scale: Additionally, we use the above hypothesis to estimate Fisherian intervals for the unit-level risk ratios. We use the exponent of the estimated coefficient from the Poisson regression model, i.e., $\hat{\beta_{1,c}}$ from model 1 (Equation 10), as our test statistic. Using $\alpha = 0.05$ as our significance level, we retrieve two effect sizes such that the Fisher-exact p-values of their respective null randomization distributions equal $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$. These two effect sizes form the bounds of the Fisherian interval.

Multiple comparison adjustment: We adopt the randomization-based procedure outlined by Lee et al. [50] and implemented in [33] to adjust our Fisher-exact p-values for multiple testing, using the following steps:

1. Construct a null randomization distribution with $100,000$ treatment allocations for each of the five cities.

2. Treat each of the $100,000$ test statistics as an "observed test statistic." Compute unadjusted randomization-based p-values by calculating the proportion of test statistics as or more extreme than the "observed test statistic."

3. The previous step results in five sets of $100,000$ unadjusted randomization-based p-values from which we

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

retrieve the smallest p-value across all of the cities for each of the 100, 000 iterations.

4. Next, construct a distribution from these minimum p-values. This process captures the joint distribution of the test statistics and rescales hypothetical test statistics to hypothetical p-values of the same $0-1$ scale. For each city, calculate the adjusted p-value, i.e., the proportion of p-values in this final distribution that are smaller than its unadjusted Fisher-exact p-value.

According to [50], randomization-based adjusted p-values account for possible correlations among test statistics and can be more powerful than adjusted Bonferroni p-values. Just as when calculating Fisher-exact p-values, this procedure relies on the assignment mechanism from the original experiment [51], and does not require assumptions about the underlying distribution of the data.

2.3.3 | Bayesian inference

We use an alternative approach that formally treats the estimation of causal effects as a missing data problem [52] and follow the general approach outlined in [53]. This method imputes the missing vector of potential outcomes, i.e., $\mathbf{Y}_c^{\text{mis}} = \{Y_{i,c}(W_{i,c} = 0) \text{ if } W_{i,c}^{\text{obs}} = 1 \text{ and } Y_{i,c}(W_{i,c} = 1) \text{ if } W_{i,c}^{\text{obs}} = 0\}$. The Bayesian perspective considers the observed values $\mathbf{Y}_c^{\text{obs}}, \mathbf{X}_c^{\text{obs}}, \mathbf{W}_c^{\text{obs}}$ to be a realization of random variables and $\mathbf{Y}_c^{\text{mis}}$ to be unobserved random variables. With this approach, the distribution of the missing potential outcomes can be evaluated by computing the posterior $\mathbf{Y}_c^{\text{mis}} \mid \mathbf{Y}_c^{\text{obs}}, \mathbf{W}_c^{\text{obs}}, \mathbf{X}_c^{\text{obs}}$. For our purpose of modeling the number of deaths, we propose the following independent count potential outcomes

$$Y_{i,c}(W_{i,c} = 0) \mid \gamma_c^{[0]} \sim \text{Poisson}(\mu_{i,c}^{[0]}), \quad Y_{i,c}(W_{i,c} = 1) \mid \gamma_c^{[1]} \sim \text{Poisson}(\mu_{i,c}^{[1]}), \tag{16}$$

where Poisson denotes the Poisson distribution, and

$$\mu_i^{[0]} := \exp(\mathbf{x}_{i,c}^\top \gamma_c^{[0]}), \quad \mu_i^{[1]} := \exp(\mathbf{x}_{i,c}^\top \gamma_c^{[1]}), \tag{17}$$

where the covariates $\mathbf{x}_{i,c} = (1, x_{i,1,c}, \ldots, x_{i,H,c})$ are the $(H+1)$-dimensional vector for each day i for city c and let $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top)$. X represents the covariates that go into the model, such as temperature and humidity. We also have $\gamma_c^{[0]}$ and $\gamma_c^{[1]} \in \mathbb{R}^{H+1}$. Concretely, we have

$$\gamma_c^{[0]} = (\gamma_{0,c}^{[0]}, \gamma_{1,c}^{[0]}, \ldots, \gamma_{h,c}^{[0]}), \quad \text{and} \quad \gamma_c^{[1]} = (\gamma_{0,c}^{[1]}, \gamma_{1,c}^{[1]}, \ldots, \gamma_{h,c}^{[1]}). \tag{18}$$

We further assume a prior distribution for γ_c:

$$\gamma_c := \begin{pmatrix} \gamma_c^{[0]} \\ \gamma_c^{[1]} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}_{2(H+1)}, \sigma_\gamma^2 \cdot \mathbf{I}_{2(H+1)}\right) \tag{19}$$

with σ_γ^2 being a fixed positive number, $\mathcal{N}$ denoting the normal distribution, and $\gamma_c \in \mathbb{R}^{2(H+1)}$. Drawing Bayesian inferences for causal effects requires the specification of an assignment mechanism, that is, a probabilistic model for how experimental units are allocated to the treatment assignment given the potential outcomes and covariates. Let $\mathbf{W}_c$ denote the N_c-vector of treatment assignments with i^{th} element $W_{i,c}$. As before, $N_{0,c} = \sum_{i=1}^{N_c}(1 - W_{i,c})$ and $N_{1,c} = \sum_{i=1}^{N_c} W_{i,c}$ as the number of units assigned to the control and active treatment respectively, with $N_{0,c} + N_{1,c} = N_c$. We also assume a paired-randomized assignment design (see Section 4.6, pages 52-53 in [41]). Using the following

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

identity

$$Y_{i,c}^{\text{mis}} = (1 - W_{i,c}) \cdot Y_{i,c}(W_{i,c} = 1) + W_{i,c} \cdot Y_{i,c}(W_{i,c} = 0),$$

the conditional distribution of $Y_{i,c}^{\text{mis}}$ given $\mathbf{Y}_c^{\text{obs}}, \mathbf{W}_c^{\text{obs}}, \mathbf{X}_c^{\text{obs}}$ can be expressed as:

$$Y_{i,c}^{\text{mis}} \mid \mathbf{Y}_c^{\text{obs}}, \mathbf{W}_c^{\text{obs}}, \mathbf{X}_c^{\text{obs}}, \gamma_c \sim \text{Poisson}(\mu_{i,c}^{\text{mis}}) \tag{20}$$

where $\mu_{i,c}^{\text{mis}} := \exp(\xi_{i,c}^{\text{mis}})$ with $\xi_{i,c}^{\text{mis}} := \left((1 - W_{i,c}) \cdot \mathbf{x}_{i,c}^{\top}\gamma_c^{[1]} + W_{i,c} \cdot \mathbf{x}_{i,c}^{\top}\gamma_c^{[0]}\right)$. We immediately see from (20) that the conditional distribution of $Y_{i,c}^{\text{mis}}$ given $\mathbf{Y}_c^{\text{obs}}, \mathbf{W}_c^{\text{obs}}, \mathbf{X}_c^{\text{obs}}$ is simply equal to the marginal distribution $Y_{i,c}^{\text{mis}}$ given $\mathbf{W}_c^{\text{obs}}, \mathbf{X}_c^{\text{obs}}, \gamma_c$ due to the independence of the potential outcomes. We now proceed to compute the conditional distribution of $\mathbf{Y}^{\text{obs}}$ given $\mathbf{W}_c^{\text{obs}}, \mathbf{X}_c^{\text{obs}}, \gamma_c$. It follows that

$$Y_{i,c}^{\text{obs}} \mid \mathbf{W}_c^{\text{obs}}, \mathbf{X}_c^{\text{obs}}, \gamma_c \sim \text{Poisson}(\mu_{i,c}^{\text{obs}}) \tag{21}$$

where $\mu_{i,c}^{\text{obs}} := \exp(\xi_{i,c}^{\text{obs}})$ with $\xi_{i,c}^{\text{obs}} = (1 - W_{i,c}) \cdot \mathbf{x}_{i,c}^{\top}\gamma_c^{[0]} + W_{i,c} \cdot \mathbf{x}_{i,c}^{\top}\gamma_c^{[1]}$. We estimate the average treatment effect (ATE) defined in Equation (3) by repeatedly filling in the missing potential outcomes, which factors in the uncertainty in the imputation. Let $\hat{Y}_{i,c}^{\text{mis}}$ be the imputed value corresponding to $Y_{i,c}^{\text{mis}}$, then the ATE in (3) can be estimated as

$$\widehat{\text{ATE}}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \left((2W_{i,c} - 1)(Y_{i,c}^{\text{obs}} - \hat{Y}_{i,c}^{\text{mis}})\right). \tag{22}$$

In particular, given a sample of γ_c (drawn using the simulation methods outlined in Section 8.5 [41]), we substitute these values into the conditional distribution of $\mathbf{Y}_c^{\text{mis}}$ to independently impute all the missing potential outcomes, that is, draw $\mathbf{Y}_c^{\text{mis}}$ from a Poisson distribution by substituting γ_c using Equation (20). Substituting the observed and imputed missing potential outcomes into (22) yields an estimate for ATE based on the r^{th} imputation, $\hat{\tau}_c^{(r)}$ for city c:

$$\widehat{\text{ATE}}_c^{(r)} = \frac{1}{N_c} \sum_{i=1}^{N_c} \left((2W_{i,c} - 1) \cdot (Y_{i,c}^{\text{obs}} - \hat{Y}_{i,c}^{(r)})\right), \tag{23}$$

we have used the identity in Equation (20). To derive the full posterior distribution of the ATE, this procedure is repeated R_c times and the average and variance of the imputed estimators $\widehat{\text{ATE}}_c^{(1)}, \widehat{\text{ATE}}_c^{(2)}, \ldots, \widehat{\text{ATE}}_c^{(R_c)}$ are

$$\frac{1}{R_c} \sum_{r=1}^{R_c} \widehat{\text{ATE}}_c^{(r)} =: \overline{\text{ATE}}_c, \qquad \frac{1}{R_c - 1} \sum_{r=1}^{R_c} (\widehat{\text{ATE}}_c^{(r)} - \overline{\text{ATE}}_c)^2. \tag{24}$$

The estimated average treatment effect using the Bayesian approach can be found in Figure 5.

Heuristic Connections to Poisson Regression with Interaction Terms. In this section, we compare the traditional Poisson regression framework with a Bayesian approach to estimating the average treatment effect. The methodology closely follows the exposition presented in Chapters 7 and 8 of [41], with an important extension: we adapt the analysis to accommodate count outcomes [53], rather than the continuous outcomes primarily discussed in their work. From the standard frequentist perspective, the unknown parameters are treated as fixed but unknown quantities, and inference is conditioned upon them. In contrast, the Bayesian framework treats both the missing and observed potential

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

outcomes as random variables—unobserved and observed, respectively—conditional on the model parameters.

Given the specification of the joint distributions $Y_{i,c}(W_{i,c}=0)$ and $Y_{i,c}(W_{i,c}=1)$ in Equation (16), we recall that the average treatment effect is defined as the difference between the two expected values, restated here:

$$\tau_c = \mathbb{E}\left[Y_{i,c}(W_{i,c}=1) - Y_{i,c}(W_{i,c}=0), \middle| , \theta_c\right]. \tag{25}$$

By expanding the terms and leveraging the binary nature of $W_{i,c}$, we can re-express μ_{obs} from Equation (21) as follows:

$$\mu_{\text{obs}} = \exp\left(\gamma_{0,c}^{[0]} + \sum_{h=1}^{H} \gamma_{h,c}^{[0]} \cdot x_{i,h,c} + (\gamma_{0,c}^{[1]} - \gamma_{0,c}^{[0]}) \cdot W_{i,c} + \sum_{h=1}^{H} (\gamma_{h,c}^{[1]} - \gamma_{h,c}^{[0]}) \cdot W_{i,c} x_{i,h,c}\right). \tag{26}$$

From this, we observe that the average treatment effect (ATE) in Equation (25) can be expressed as:

$$\tau_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \left(\mu_{i,c}^{[1]} - \mu_{i,c}^{[0]}\right). \tag{27}$$

Now let us consider the regression function specified as $Y_{i,c}^{\text{obs}} \mid \lambda_{i,c} \sim \text{Poisson}(\lambda_{i,c})$, with $\lambda_{i,c}$ taking the form:

$$\lambda_{i,c} = \exp\left(\alpha_c + \sum_{h=1}^{H} \kappa_{h,c} \cdot x_{i,h,c} + \sum_{h=1}^{H} \eta_{h,c} \cdot W_i x_{i,h,c} + \gamma_c \cdot W_{i,c}\right) \tag{28}$$

where α_c, γ_c, $\kappa_{h,c}$, and $\eta_{h,c}$ denote the regression parameters for city c. This specification resembles Equation (2) in Section 2.3.1, but the mean structure corresponds to that of a Poisson regression model. By comparing Equations (26) and (28), we observe that the inferential problem reduces to estimating the regression coefficients via maximum likelihood estimation for the parameters α_c, γ_c, $\kappa_{h,c}$, and $\eta_{h,c}$. Once these estimates are obtained, simple transformations can be applied to derive the average treatment effect. Moreover, we note that maximum likelihood estimation under the frequentist framework yields two sets of equations: one corresponding to the estimated control parameters, and another capturing the estimated differences between the treatment and control parameters.
In particular, we have

$$\text{Control parameters:} \quad \begin{cases} \hat{\alpha}_c = \gamma_{0,c}^{[0]}, \\ \hat{\kappa}_h = \gamma_{h,c}^{[0]}, \quad h = 1, \ldots, H \end{cases} \tag{29}$$

$$\text{Difference of treatment and control parameters:} \quad \begin{cases} \hat{\gamma}_c = \gamma_{0,c}^{[1]} - \gamma_{0,c}^{[0]}, \\ \hat{\eta}_{h,c} = \gamma_{h,c}^{[1]} - \gamma_{h,c}^{[0]}, \quad h = 1, \ldots, H \end{cases} \tag{30}$$

With these, $\gamma_{h,c}^{[1]}$ and $\gamma_{0,c}^{[1]}$ can be constructed from the control parameters and their corresponding treatment-control differences via $\gamma_{h,c}^{[1]} = \hat{\eta}_{h,c} + \hat{\kappa}_h$ and $\gamma_{0,c}^{[1]} = \hat{\gamma}_c + \hat{\alpha}_c$, respectively. Equipped with these inferred parameters, we can then compute the ATE in Equation (27) as a function of $\mu^{[1]}i, c$ and $\mu^{[0]}i, c$ through equations (17) and (18).

2.4 | Illustrative Example with $H = 3$ Covariates

To concretize the derivations in Section 2.3.1, we present a toy example where the number of covariates is set to $H = 3$. This example demonstrates how the treatment group parameters can be recovered from estimated control

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

parameters and their corresponding treatment-control differences. We then compute the average treatment effect (ATE) accordingly.

Example Data.

Consider two days with the following covariate values in Table 1.

Estimated Parameters.

Assume that the following coefficients have been estimated:

- Control group parameters:

$$\hat{\alpha}_c = 0.20, \quad \hat{\kappa}_{1,c} = 0.30, \quad \hat{\kappa}_{2,c} = -0.10, \quad \hat{\kappa}_{3,c} = 0.05$$

- Differences between treatment and control:

$$\hat{\gamma}_c = 0.10, \quad \hat{\eta}_{1,c} = 0.20, \quad \hat{\eta}_{2,c} = 0.10, \quad \hat{\eta}_{3,c} = -0.05$$

The treatment group parameters are then constructed as follows:

$$\gamma_{0,c}^{[1]} = \hat{\gamma}_c + \hat{\alpha}_c = 0.30, \quad \gamma_{h,c}^{[1]} = \hat{\eta}_{h,c} + \hat{\kappa}_{h,c}, \quad h = 1, 2, 3$$

$$\Rightarrow \gamma_{1,c}^{[1]} = 0.50, \quad \gamma_{2,c}^{[1]} = 0.00, \quad \gamma_{3,c}^{[1]} = 0.00$$

Expected Outcomes.

Using Equations (17) and (18), we compute the expected outcomes under control and treatment:

- Individual 1:

$$\mu_{1,c}^{[0]} = \exp(0.2 + 0.3 \cdot 1.0 - 0.1 \cdot 0.5 + 0.05 \cdot (-1.0)) = \exp(0.4) \approx 1.4918$$

$$\mu_{1,c}^{[1]} = \exp(0.3 + 0.5 \cdot 1.0) = \exp(0.8) \approx 2.2255$$

- Individual 2:

$$\mu_{2,c}^{[0]} = \exp(0.2 - 0.1 \cdot 1.0 + 0.05 \cdot 0.5) = \exp(0.125) \approx 1.1331$$

$$\mu_{2,c}^{[1]} = \exp(0.3) \approx 1.3499$$

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Average Treatment Effect.

Applying Equation (27), the ATE is given by:

$$\tau_c = \frac{1}{2}\left[(2.2255 - 1.4918) + (1.3499 - 1.1331)\right] = \frac{0.7337 + 0.2168}{2} = 0.4753$$

Thus, the estimated average treatment effect in this illustrative case is $\tau_c \approx 0.4753$.

We defer the theoretical computation of the asymptotic precision of our Poisson regression estimators, as well as the exploration of constraints on precision gains from additional covariates, to future work, following the approaches in [54] and [55]. In the recent paper by [56], a causal interpretation of the hazard ratio in randomized clinical trials is provided. Future research may explore connections between the "time series" version of our formulation and theirs.

2.5 | Stage 4: Interpretation of results

An objective of this causal inference framework is to reduce confounding and draw conclusions not just about the association between heat and mortality, but to be able to assess this causal relationship. We assume that our hypothetical randomized experiment is plausible and that we have accounted for all confounding by matching only on observed covariates [37].

It is also important to note that matched treated and control time periods have similar preceding days, but we do not require the three-day treatment periods themselves to be similar. Warm temperatures lead to higher concentrations of many potentially dangerous air pollutants including ground-level ozone, making secondary air pollutants such as ozone possible mediators between heat and mortality [57]. We then interpret the effect of heat on mortality to include any potential mediators between heat and mortality, such as secondary air pollutants. Finally, these results apply only to the time periods that were selected through the matching process and may not extend to all days in the study period.

2.6 | Comparison to a traditional analysis of unmatched data

We also compared the results from this causal inference framework to a more traditional analysis; for each city, we fit a Poisson regression model on the full, unmatched dataset. These models have the same form as those in Model 3 (Equation 12) and serve as a point of comparison for the models fit on the matched dataset. We note that these models typically allow for a more complex relationship between exposure to heat and mortality, often with a generalized additive model (GAM) or spline [58,59]. We use the binary exposure defined in Equation 1 for comparison across models.

3 | RESULTS

3.1 | Balance assessment

Here, we present the results of the matching process with covariate balance diagnostics. Note that since we have designed a hypothetical randomized experiment, we seek to identify pairs of treated and control periods with similar lag days, but we do not require days within the three-day experiment to be similar. We seek to balance only the lagged covariates.

First, we must ensure that the matching process leaves us with a reasonable number of pairs in each city for the analysis stage. Table 4 displays the number of 1) control (warm) and treated (hot) units identified, 2) the number of

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

matched pairs of control and treated units, and 3) the percentage of treated units for which a corresponding control unit was available. For the majority of treated units identified in the full, unmatched data, no suitable control units are available for matching within the constraints set in Table 5.

Next, we investigate covariate balance; the goal is that all covariates are balanced in the treated and control groups after matching to minimize potential confounding. Figure 2 is a Love plot, which displays potential relationships between background covariates and the treatment assignment [60]. For both the full, unmatched data and the matched subset, we calculate the standardized difference in covariate means as follows for each covariate X_i.

$$\text{Standardized mean difference} = \frac{\frac{1}{K}\sum_{i=1}^{\frac{N_c}{2}}(X_i[W_i=1]) - \frac{1}{K}\sum_{i=1}^{K}(X_i[W_i=0])}{\sqrt{0.5(var(X_i[W_i=1]) + var(X_i[W_i=0]))}}. \tag{31}$$

If a covariate is perfectly balanced between treated and control periods, then the covariate has no association with treatment assignment, and the standardized mean difference is 0. We see this with the variable year, which is matched exactly in each pair. A positive standardized mean difference indicates that a covariate is positively associated with assignment to treatment, and a negative standardized mean difference indicates that it is positively associated with assignment to control. We see in Figure 2 that the standardized differences in covariate means were greatly reduced after matching.

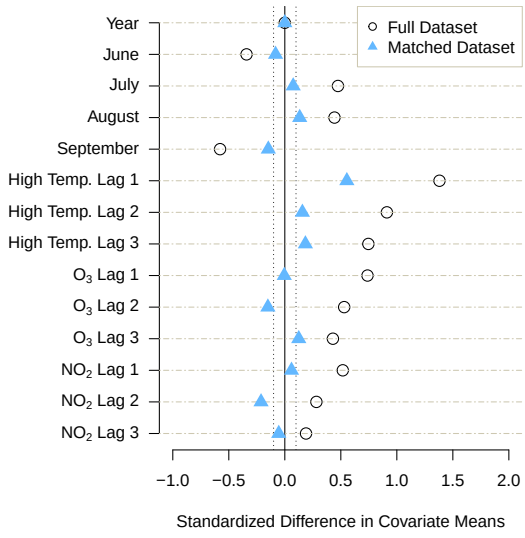


FIGURE 2 Love plot with the standardized differences in covariate means for several potential confounders; see Equation 31 for this calculation. Standardized mean difference values near zero indicate covariate balance between treated and control units, whereas values that are farther from zero indicate that covariates are associated with treatment assignment. Through matching, we aim to reduce these standardized mean differences to be between -0.1 and 0.1 [61], marked here with vertical dashed lines.

We can also inspect density plots to assess covariate balance in continuous covariates that may exhibit complex associations with the treatment assignment. We see in Figure 1 (appendix) that in the full, unmatched data, periods

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

in the middle of the summer were most likely to be hot, while periods in the beginning and end of the summer were most likely to be warm, as expected. However, in the matched dataset, the distribution of hot and warm paired units is more similar throughout the summer, reducing potential confounding by variables with seasonal variations.

We can also visualize these hypothetical randomized experiments by plotting several potential confounders leading up to and during each three-day experimental unit. Figure 3 displays the mean values of six covariates over each three-day treated or control experimental unit, as well as the three days preceding these experimental units. All six covariates display similar patterns. Lag days preceding treated periods have similar covariate values to those preceding control periods. These covariate values then diverge during the three-day experimental units when they are unconstrained by the matching process, thus imitating randomization of treatment assignment on Day 1.

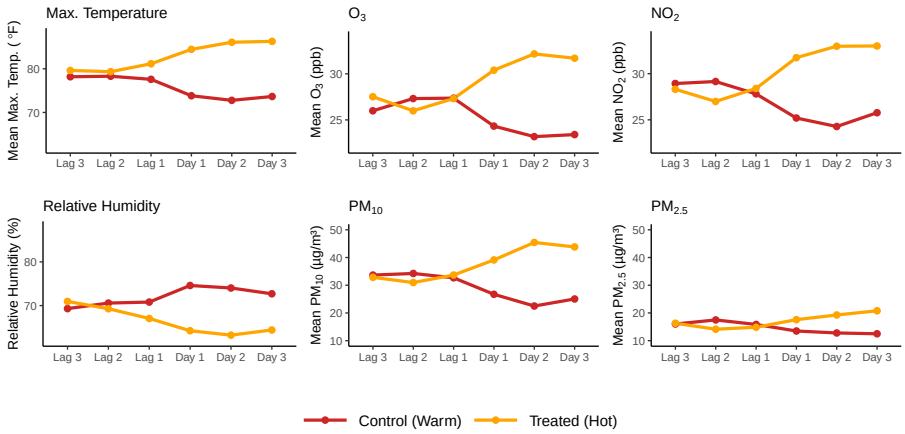


FIGURE 3 This figure displays the mean values of six covariates in matched treated and control periods over time across all five cities. The first three points in each time series correspond to lag days, or the three days preceding each three-day experimental unit. The last three points in each time series correspond to the three-day treated and control periods in the matched dataset. Note that $PM_{2.5}$ and PM_{10} were not recorded daily during this time period, so this figure includes only matched periods for which particulate matter concentrations were available. Missingness is assumed to be independent of the exposure assignment.

3.2 | Three-day mortality during warm versus hot periods

After assessing covariate balance in the matched dataset, we compare mortality during hot versus warm time periods. Figure 4 displays box plots comparing the distribution of three-day mortality counts during matched control and treated periods. In the matched dataset, the median number of deaths during treated periods is greater than the median number of deaths during control periods in all five cities.

We display inference results on two scales: 1) the additive scale, as a measure of the difference between the number of deaths during three-day treated and control periods, and 2) the multiplicative scale, as a measure of the ratio between the number of deaths during three-day treated and control periods.

The Neymanian and Bayesian inference approaches described here estimate the average treatment effect τ_c and uncertainty intervals on the additive scale, whereas the traditional frequentist approach of Poisson regression estimates the risk ratio associated with treatment. We also estimated uncertainty intervals for both the estimated

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

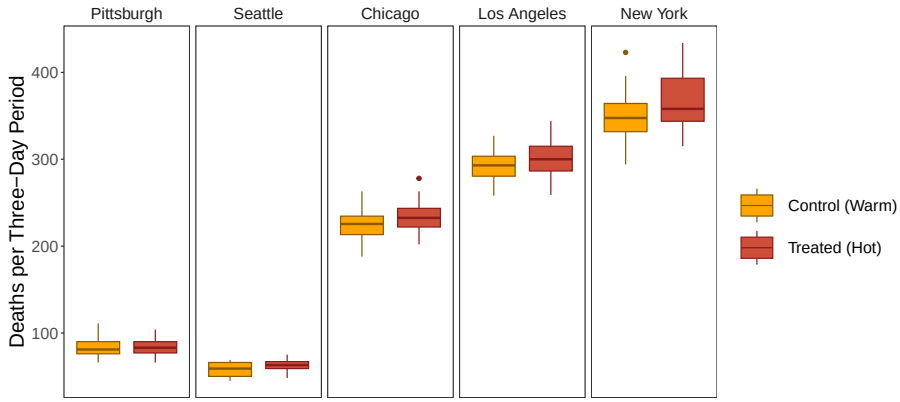


FIGURE 4 Box plots comparing the distribution of three-day mortality (the total number of deaths over a three-day time period) during matched treated versus control periods.

average treatment effects on the additive scale and on the multiplicative scale with the Fisherian approach. All inference results on the additive scale are displayed in Figure 5; inference results on the multiplicative scale are displayed in Figure 6.

Because the additive scale is dependent on city-specific population, we have included population-adjusted estimates of average treatment effect for each city in Table 3. This table also includes Fisherian p-values for the estimates in each city, as defined in Equation 14.

After matching, we observe more deaths during hot time periods compared to warm time periods among the elderly living in Seattle $(3.7, [95\%\text{FI: } .7, 6.6])$, Chicago $(12.5, [95\%\text{FI: } 6.1, 18.9])$, Los Angeles $(8.5, [95\%\text{FI: } 1.4, 15.6])$, and New York $(20.3, [95\%\text{FI: } 11.2, 29.5])$, with unadjusted Fisherian p-values $\approx 0.0078, 0.00005, 0.00848$, and < 0.00001, respectively. We retrieve similar results from the Neymanian and Bayesian analyses, which suggests that hotter summer days are associated with higher mortality than cooler summer days in these cities. In Pittsburgh, we could not reject the sharp null hypothesis (Fisherian p-value ≈ 0.42).

The risk ratios obtained from the Poisson regression models align with the results from the other approaches. We found the largest risk ratio in Seattle: $1.076[95\%\text{CI: } 0.999, 1.16]$, indicating a 7.6% increase in mortality rates for hot versus warm time periods. This was followed by New York, Chicago and Los Angeles: $1.057[95\%\text{CI: } 1.023, 1.092]$, $1.051[95\%\text{CI: } 1.019, 1.084]$, $1.039[95\%\text{CI: } 1.013, 1.065]$ respectively. For the cities in which we found increases in the deaths (i.e., Seattle, Chicago, New York, and Los Angeles), the risk ratios were larger when using the Poisson regression model on the historical times series matched data compared to those from the Poisson regression model on the unmatched data (i.e., traditional analysis). This finding indicates that the traditional analysis may underestimate the effect of extreme temperatures.

4 | DISCUSSION

4.1 | Our results

Our findings indicate meaningful increases in mortality can be attributed to heat in four of the five cities analyzed. This is in agreement with existing literature that has reported on increased mortality due to heat-related episodes.

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

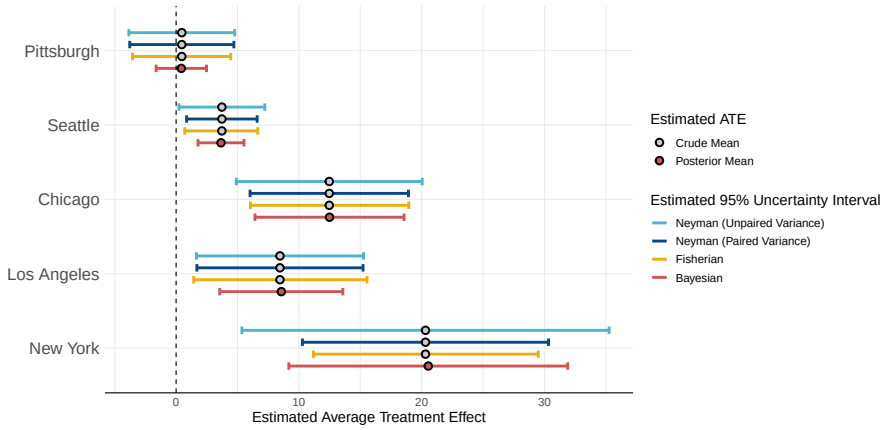


FIGURE 5 Estimated average treatment effects and estimated 95% uncertainty intervals comparing the total number of deaths during three-day treated and control periods. The Neymanian estimate of the average treatment effect is simply the observed crude mean difference between three-day mortality during treated and control units (Equation 4), while the Bayesian estimate is the posterior mean. See Table 3 (appendix) for exact numerical values displayed here.

Previous studies conducted on the National Morbidity and Mortality Air Pollution Study (NMMAPS) have estimated increases in mortality risks related to heat waves nationally to be 3%[95% Posterior Interval (PI): 2.4%, 3.6%] and 3.74%[95% PI :2.29, 5.22%] [59,62]. Similarly, a county-specific analysis conducted in California found approximately 4% increases in non-accidental mortality in Los Angeles per 10°F, while an 18-year study on New York found that heat waves increased mortality by 6.5% [95% CI: 2.0%, 11.2%] [63,64].

4.2 | Drawing different types of causal inference

The uncertainty intervals obtained from the frequentist, Fisherian, and Bayesian approaches heavily overlap, indicating similar effect sizes across the three modes of inference. These approaches rely on different assumptions, so their alignment is reassuring. Additionally, three different analysts implemented the different modes of inference separately, highlighting the robust nature of this causal inference analysis and increasing confidence in our results.

Together, the three approaches provide slightly distinct interpretations of the results. Neymanian inference produces an estimated average treatment effect, which we interpret as the estimated average additional number of deaths during a three-day time period had it been hot instead of warm. For example, we observe an estimated average treatment effect of 12.5 deaths in Chicago, indicating that on average, a hot three-day time period will have 12.5 [95%CI: 6, 18.9] more deaths than a warm three-day time period, holding all other observed covariates constant. Similar to Neymanian Inference, the Bayesian models also provide estimates of the average treatment effect. In Chicago, this is estimated to be an increase of 12.495 [95%CI: 6.427, 18.564] more deaths. The intervals obtained from the Fisherian approach have a slightly different interpretation in that they estimate the uncertainty in the unit-level treatment effect consistent with the data as opposed to the average treatment effect. The Poisson regression models provide us with risk ratios; for example, a risk ratio of 1.029 from Model 1 for Los Angeles indicates 2.9%[95%CI: 1.008, 1.051] more deaths during three-day treated time periods compared to control time periods, holding all other observed co-

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

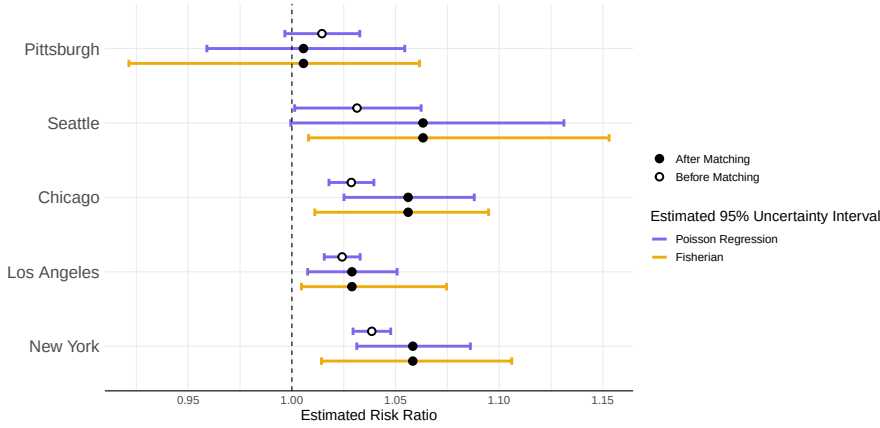


FIGURE 6 Estimated risk ratios and estimated 95% uncertainty intervals comparing the total number of deaths during three-day treated and control periods. The point estimates represented by open circles were obtained using all three-day periods in the full, unmatched dataset with defined treatment, before matching treated and control periods and eliminating those without corresponding matches, adjusting for potential observed confounders (Model 3). The point estimates represented by closed circles were obtained using the subset of matched treated and control units. Although these point estimates were obtained for Models 1, 2, and 3, only results from Model 1 (the unadjusted model) are displayed here. Fisherian intervals were also obtained for these estimated risk ratios from the matched data. See Table 6 for all exact numerical values displayed in this figure and the results from Models 2 and 3, which were fit on the matched data and were adjusted for certain covariates.

variates constant.

The Neymanian approach assumes a constant additive treatment effect, i.e., the effect of heat is the same across all matched treated and control periods. In order to calculate 95% confidence intervals (CIs), we also assume that the estimated average treatment effect is normally distributed. The Fisherian method allows us to draw inferences without making any asymptotic assumptions. This is particularly useful for small sample sizes and when data do not follow a normal distribution. The Bayesian approach, like the Neymanian approach, estimates the average treatment effect. Here, we specify a model for the joint distribution of $(Y_{i,c}(W_{i,c} = 0), Y_{i,c}(W_{i,c} = 1))$ and rely on a large sample assumption approximation.

The standard frequentist approach of Poisson regression specifies a model and assumes that mortality counts, conditional on all covariates in the model, follow the Poisson distribution. Poisson regression allows us to estimate risk ratios, and requires additional assumptions to estimate an effect size on the additive scale. All approaches presented here assume independence between time periods, although we expect some temporal correlations within each city.

The interpretability of each approach is also an important difference. One benefit of presenting results on the additive scale, i.e., as the estimated average treatment effect or unit-level effect, is that the results may be more interpretable than risk ratios since they quantify the actual number of deaths associated with a certain exposure. We are able to convert results on the multiplicative scale to the additive scale, but only with certain assumptions. This makes the Neymanian, Fisherian, and Bayesian approaches potentially more useful in a policy setting than the frequentist Poisson regression approach.

Had we forgone the matching process entirely and only fit adjusted Poisson regression models on the full, unmatched dataset, it appears as though we may have underestimated the effect of heat on mortality in Seattle, Chicago,

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Los Angeles, and New York–the four cities in which the effect of heat appears to be significant (Figure 5)–pointing to the importance of matching, although we note that the study population after matching is different.

By constructing a matched dataset prior to our statistical analysis, we mimic a randomized experiment and limit confounding bias common in environmental time series studies. Many of these biases arise from the selection methods for the functions that address confounding in these models. Other biases arise due to unbalanced covariates in the treatment and control groups[65]. This study's causal inference framework presents a method for addressing these issues by separating the design from the statistical analysis and is not unique to any particular exposure. This framework can easily be extended to additional environmental health applications where observational data is heavily relied on, or to other applications in time series at large.

4.3 | Limitations

Although historical time series matching successfully improves observed covariate balance between treatment groups, there are limitations to our approach. Matching requires reducing a dataset down to a smaller subset, which may not be representative of the full population. When we define a dichotomous treatment (Equation 1) and set allowable discrepancies for matching criteria (Table 5), we must ensure that this remaining subset is large enough to draw meaningful conclusions. It is important to note that these results apply only to this subset of matched data in these five cities. Appendix Figure 1 displays covariate differences between the full, unmatched dataset and the matched subset analyzed here.

We also used each city's summer median maximum temperature to define our dichotomous treatment assignment; we defined a hot time period to be three consecutive days for which the maximum temperature was above the city's summer median. If we were interested in studying more extreme heat–for example, daily temperature above 90°F, we would require cities experiencing such high temperatures. Additionally, using reported city-wide temperature measurements does not account for neighborhood-level variations in temperature or allow us to evaluate specific impacts on microclimates.

Although previous studies have identified associations between subgroups of the population and increased risk of heat-related mortality[66], our data set did not have individual-level (e.g., demographic) information. This information could be particularly helpful for informing policies to mitigate heat-related mortality in susceptible, low-income communities (e.g., programs to expand access to solar-powered air conditioning units in these communities). Additionally, we do not differentiate between mortality caused directly and indirectly by heat. Short-term mortality displacement can also occur when extreme heat expedites time of death in individuals with pre-existing conditions and short life expectancies. This can be seen in a forward shift in the mortality rate followed by a reduction of deaths in the weeks after an extreme heat event. Current literature reports discordant results on the extent to which mortality displacement contributes to heat-induced mortality[67,68,69]. Hence, this question presents a future avenue for exploration.

4.4 | Relevance of Results

Although not explored within our analysis, our results highlight the importance of government policies and programs that prevent the impact of extreme heat events, which are becoming increasingly common with anthropogenic climate change. Potential strategies include the designation of cooling centers during heat events and programs providing transportation to these locations, which can reduce the risk of overheating, particularly for the susceptible population (e.g., elderly)[70]. Adopting coordinated heat warning systems can ensure that vulnerable populations have information on upcoming heat events, as well as potential risks and actions to protect themselves[71]. Future studies could adopt

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

our causal inference framework to evaluate the effects of specific interventions, such as studying random assignments of cooling centers in certain locations to provide further insight on the efficacy of these interventions in reducing heat-related mortality.

It is also important to recognize that the impact of heat extremes can be influenced by spatial factors inherent to the design of a city. Microclimates in urban areas often have higher air temperatures compared to surrounding rural areas due to numerous factors, including lack of vegetation, air pollution, and low reflective materials used in infrastructure [72]. Increasing the amount of green space in urban areas, as well as the usage of high-albedo materials in houses and buildings to reflect solar radiation can be important in reducing temperatures in cities [73]. Finally, policies that limit greenhouse gas emissions can mitigate rising temperatures, while those that reduce air pollution can limit the exacerbative effects of these pollutants to minimize heat-related mortality [74].

references

[1] Watts N, Adger WN, Agnolucci P, et al. Health and climate change: policy responses to protect public health. *The Lancet* 2015; 386(10006): 1861–1914.

[2] Saforim S, Hawkins M, Hess H, Kinney S, St.Juliana . *The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment*. US Global Change Research Program . 2016.

[3] Allen M, Dube O, Solecki W, et al. Global Warming of 1.5 C.. tech. rep., 2019.

[4] Luber G, McGeehin M. Climate change and extreme heat events. *American journal of preventive medicine* 2008; 35(5): 429–435.

[5] Huang W, Kan H, Kovats S. The impact of the 2003 heat wave on mortality in Shanghai, China. *Science of the total environment* 2010; 408(11): 2418–2420.

[6] Gamble JL, Hurley BJ, Schultz PA, Jaglom WS, Krishnan N, Harris M. Climate change and older Americans: state of the science. *Environmental health perspectives* 2013; 121(1): 15–22.

[7] Stone B, Hess JJ, Frumkin H. Urban form and extreme heat events: are sprawling cities more vulnerable to climate change than compact cities?. *Environmental health perspectives* 2010; 118(10): 1425–1428.

[8] Clarke JF. Some effects of the urban structure on heat mortality. *Environmental research* 1972; 5(1): 93–104.

[9] Song X, Wang S, Hu Y, et al. Impact of ambient temperature on morbidity and mortality: an overview of reviews. *Science of the Total Environment* 2017; 586: 241–254.

[10] Yu W, Mengersen K, Wang X, et al. Daily average temperature and mortality among the elderly: a meta-analysis and systematic review of epidemiological evidence. *International journal of biometeorology* 2012; 56(4): 569–581.

[11] Rocklöv J, Forsberg B. The effect of high ambient temperature on the elderly population in three regions of Sweden. *International journal of environmental research and public health* 2010; 7(6): 2607–2619.

[12] Stafoggia M, Forastiere F, Michelozzi P, Perucci CA. Summer temperature-related mortality: effect modification by previous winter mortality. *Epidemiology* 2009: 575–583.

[13] D'Ippoliti D, Michelozzi P, Marino C, et al. The impact of heat waves on mortality in 9 European cities: results from the EuroHEAT project. *Environmental Health* 2010; 9(1): 1–9.

[14] Lee DG, Kim KR, Kim J, et al. Effects of heat waves on daily excess mortality in 14 Korean cities during the past 20 years (1991–2010): an application of the spatial synoptic classification approach. *International Journal of Biometeorology* 2017; 62(4): 575–583.

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[15] Basu R. High ambient temperature and mortality: a review of epidemiologic studies from 2001 to 2008. *Environmental health* 2009; 8(1): 1–13.

[16] Goodman PG, Dockery DW, Clancy L. Cause-specific mortality and the extended effects of particulate pollution and temperature exposure. *Environmental health perspectives* 2004; 112(2): 179–185.

[17] Saucy A, Ragettli MS, Vienneau D, et al. The role of extreme temperature in cause-specific acute cardiovascular mortality in Switzerland: A case-crossover study. *Science of The Total Environment* 2021; 790: 147958.

[18] Yu W, Xu R, Ye T, et al. Temperature-mortality association during and before the COVID-19 pandemic in Italy: A nationwide time-stratified case-crossover study. *Urban climate* 2021; 39: 100948.

[19] Dominici F, McDermott A, Zeger SL, Samet JM. On the use of generalized additive models in time-series studies of air pollution and health. *American journal of epidemiology* 2002; 156(3): 193–203.

[20] Gasparrini A. Modeling exposure–lag–response associations with distributed lag non-linear models. *Statistics in medicine* 2014; 33(5): 881–899.

[21] Sylvestre MP, Abrahamowicz M. Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in medicine* 2009; 28(27): 3437–3453.

[22] Gasparrini A, Scheipl F, Armstrong B, Kenward MG. A penalized framework for distributed lag non-linear models. *Biometrics* 2017; 73(3): 938–948.

[23] Hurvich CM, Tsai C. The impact of model selection on inference in linear regression. *The American Statistician* 1990; 44(3): 214–217.

[24] Doerken S, Mockenhaupt M, Naldi L, Schumacher M, Sekula P. The case-crossover design via penalized regression. *BMC medical research methodology* 2016; 16(1): 1–9.

[25] Armstrong BG, Gasparrini A, Tobias A. Conditional Poisson models: a flexible alternative to conditional logistic case cross-over analysis. *BMC medical research methodology* 2014; 14(1): 1–6.

[26] Lumley T, Levy D. Bias in the case–crossover design: implications for studies of air pollution. *Environmetrics: The official journal of the International Environmetrics Society* 2000; 11(6): 689–704.

[27] Wang X, Wang S, Kindzierski W. Eliminating systematic bias from case-crossover designs. *Statistical Methods in Medical Research* 2019; 28(10-11): 3100–3111.

[28] Bateson TF, Schwartz J. Selection bias and confounding in case-crossover analyses of environmental time-series data. *Epidemiology* 2001; 12(6): 654–661.

[29] Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 2005; 100(469): 322–331.

[30] Bind MAC, Rubin DB. Bridging observational studies and randomized experiments by embedding the former in the latter. *Statistical methods in medical research* 2019; 28(7): 1958–1978.

[31] Micali S, Vazirani V. An o (v| v| c| e|) algorithm for finding maximum matching in general graphs. *21st Annual Symposium on Foundations of Computer Science* 1980: 17–27.

[32] Zabrocki L, Leroutier M, Bind MA. Estimating the Local Air Pollution Impacts of Cruise Traffic: A Principled Approach for Observational Data. *Improving the Design of Studies on the Acute Health Effects of Air Pollution* 2021: 41.

[33] Sommer AJ, Peters A, Rommel M, et al. A randomization-based causal inference framework for uncovering environmental exposure effects on human gut microbiota. *PLOS Computational Biology* 2022; 18(5): e1010044.

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[34] Sommer A, Leray E, Lee Y, Bind MA. Assessing environmental epidemiology questions in practice with a causal inference pipeline: An investigation of the air pollution-multiple sclerosis relapses relationship. *Statistics in Medicine* 2021; 40(6): 1321–1335.

[35] Cochran WG, Chambers SP. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)* 1965; 128(2): 234–266.

[36] Rubin DB. For objective causal inference, design trumps analysis. *The annals of applied statistics* 2008; 2(3): 808–840.

[37] Rubin DB. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* 1991: 1213–1234.

[38] Rosenbaum PR, Rosenbaum P, Briskman . *Design of observational studies*. 10. Springer . 2010.

[39] Sommer A, Lee M, Bind MA. Comparing apples to apples: an environmental criminology analysis of the effects of heat and rain on violent crimes in Boston. *Palgrave communications* 2018; 4(1): 1–10.

[40] Splawa-Neyman J. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles.. *Statistical Science* 1990; 5(4): 465–72.

[41] Imbens GW, Rubin DB. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press . 2015.

[42] Rubin DB. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American statistical association* 1980; 75(371): 591–593.

[43] Holland PW. Statistics and Causal Inference. *Journal of the American Statistical Association* 1986; 81(396): 945–960.

[44] Gadbury G. Randomization inference and bias of standard errors. *The American Statistician* 2001; 55: 310–313.

[45] Imai K. Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine* 2008; 27(24): 4857–4873.

[46] Zhao A, Ding P, Mukerjee R, Dasgupta T. Randomization-based causal inference from split-plot designs.. *The Annals of Statistics* 2022; 46(5): 1876–1903.

[47] Bind MA. Causal Modeling in Environmental Health. *Annual Review of Public Health* 2019; 40(1): 23-43.

[48] Fisher RA. Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* 1925; 22: 700 - 725.

[49] Pearce S. Introduction to Fisher (1925) statistical methods for research workers. In: Springer. 1992 (pp. 59–65).

[50] Lee JJ, Forastiere L, Miratrix L, Pillai NS. More powerful multiple testing in randomized experiments with non-compliance. *Statistica Sinica* 2017: 1319–1345.

[51] Bind MA, Rubin D. When possible, report a Fisher-exact P value and display its underlying null randomization distribution. *Proceedings of the National Academy of Sciences* 2020; 117(32): 19151–19158.

[52] Rubin DB. Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics* 1978; 6: 34-58.

[53] Lee Y, Bergsma WP, Bind MA. Bayesian causal inference for count potential outcomes. *arXiv preprint arXiv:2008.03271* 2020.

[54] Lin W. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics* 2013; 7(1): 295 – 318.

[55] Wager S, Du W, Taylor J, Tibshirani RJ. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences* 2016; 113(45): 12673-12678.

medRxiv preprint doi: https://doi.org/10.64898/2026.10.13.26365467; this version posted October 14, 2026.
The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[56] Fay MP, Li F. Causal interpretation of the hazard ratio in randomized clinical trials. *Clinical Trials* 2024; 21(5): 623–635.

[57] Alari A, Chen C, Schwarz L, Hansen K, Chaix B, Benmarhnia T. The role of ozone as a mediator in the relation between heat waves and mortality in 15 French urban agglomerations. *American Journal of Epidemiology* 2023. kwad032http://dx.doi.org/10.1093/aje/kwad032 doi: 10.1093/aje/kwad032

[58] Gasparrini A, Armstrong B. The impact of heat waves on mortality. *Epidemiology (Cambridge, Mass.)* 2011; 22(1): 68–73. http://dx.doi.org/10.1097/EDE.0b013e3181fdcd99 doi: 10.1097/EDE.0b013e3181fdcd99

[59] Anderson BG, Bell ML. Weather-related mortality: how heat, cold, and heat waves affect mortality in the United States. *Epidemiology (Cambridge, Mass.)* 2009; 20(2): 205.

[60] Ahmed A, Husain A, Love T, et al. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. *European Heart Journal* 2006; 27(12): 1431–1439.

[61] Stuart A. Matching Methods for Causal Inference: A Review and a Look Forward.. *Statistical Science* 2010; 25(1): 1–21.

[62] Anderson GB, Bell ML. Heat waves in the United States: mortality risk during heat waves and effect modification by heat wave characteristics in 43 US communities. *Environmental health perspectives* 2011; 119(2): 210–218.

[63] Basu R, Feng WY, Ostro BD. Characterizing temperature and mortality in nine California counties. *Epidemiology* 2008: 138–145.

[64] Anderson B, Bell M. Heat Waves and Mortality in New York, NY. *Epidemiology* 2011; 22(1): S20.

[65] Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* 2001; 2: 169–188.

[66] Uejio CK, Wilhelmi OV, Golden JS, Mills DM, Gulino SP, Samenow JP. Intra-urban societal vulnerability to extreme heat: the role of heat exposure and the built environment, socioeconomics, and neighborhood stability. *Health & place* 2011; 17(2): 498–507.

[67] Baccini M, Biggeri A, Accetta G, et al. Heat effects on mortality in 15 European cities. *Epidemiology* 2008: 711–719.

[68] Basu R, Malig B. High ambient temperature and mortality in California: exploring the roles of age, disease, and mortality displacement. *Environmental research* 2011; 111(8): 1286–1292.

[69] Le Tertre A, Lefranc A, Eilstein D, et al. Impact of the 2003 heatwave on all-cause mortality in 9 French cities. *Epidemiology* 2006: 75–79.

[70] Widerynski S, Schramm PJ, Conlon KC, et al. Use of cooling centers to prevent heat-related illness: summary of evidence and strategies for implementation. tech. rep., Center for Disease Control and Prevention; 2017.

[71] Atmospheric Programs oUSEPAO. *Excessive heat events guidebook*. US Environmental Protection Agency, Office of Atmospheric Programs . 2006.

[72] Kleerekoper L, Van Esch M, Salcedo TB. How to make a city climate-proof, addressing the urban heat island effect. *Resources, Conservation and Recycling* 2012; 64: 30–38.

[73] Deilami K, Kamruzzaman M, Liu Y. Urban heat island effect: A systematic review of spatio-temporal factors, data, methods, and mitigation measures. *International journal of applied earth observation and geoinformation* 2018; 67: 30–42.

[74] Stieb DM, Judek S, Burnett RT. Meta-analysis of time-series studies of air pollution and mortality: effects of gases and particles and the influence of cause of death, age, and season. *Journal of the Air & Waste Management Association* 2002; 52(4): 470–484.

[75] Bureau UC. 1990 Census of Population: General Population Characteristics. Section: Government.