PLOS ONE

RESEARCH ARTICLE

A simulation-based assessment of the ability to detect thresholds in chronic risk concentration-response functions in the presence of exposure measurement error

Garrett Glasgow[1]*, **Bharat Ramkrishnan**[2], **Anne E. Smith**[2]

1 NERA Economic Consulting, San Francisco, California, United States of America, **2** NERA Economic Consulting, Washington DC, District of Columbia, United States of America

* garrett.glasgow@nera.com

Citation: Glasgow G, Ramkrishnan B, Smith AE (2022) A simulation-based assessment of the ability to detect thresholds in chronic risk concentration-response functions in the presence of exposure measurement error. PLoS ONE 17(3): e0264833. https://doi.org/10.1371/journal.pone.0264833

Editor: Asif Qureshi, Indian Institute of Technology Hyderabad, INDIA

Received: June 14, 2021

Accepted: February 17, 2022

Published: March 11, 2022

Data Availability Statement: All simulated cohort data are available from Dataverse at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/75YFSJ.

Funding: GG, BR, and AES were funded by the Texas Commission on Environmental Quality (TCEQ), WO 582-14-40698-10. URL: https://www.tceq.texas.gov/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

An important question when setting appropriate air quality standards for fine particulate matter ($PM_{2.5}$) is whether there exists a "threshold" in the concentration-response (C-R) function, such that $PM_{2.5}$ levels below this threshold are not expected to produce adverse health effects. We hypothesize that measurement error may affect the recognition of a threshold in long-term cohort epidemiological studies. This study conducts what is, to the best of our knowledge, the first simulation of the effects of measurement error on the statistical models commonly employed in long-term cohort studies. We test the degree to which classical-type measurement error, such as differences between the true population-weighted exposure level to a pollutant and the observed measures of that pollutant, affects the ability to statistically detect a C-R threshold. The results demonstrate that measurement error can obscure the existence of a threshold in a cohort study's C-R function for health risks from chronic exposures. With increased measurement error the ability to statistically detect a C-R threshold decreases, and both the estimated location of the C-R threshold and the estimated hazard ratio associated with $PM_{2.5}$ are attenuated. This result has clear implications for determining appropriate air quality standards for pollutants.

Introduction

Numerous epidemiological studies over the past several decades have found a statistical association between $PM_{2.5}$ (atmospheric particulate matter with a diameter of 2.5 micrometers or less) and mortality, with the strongest effects usually reported from chronic exposure studies that compare survival outcomes of cohorts in different communities with differing ambient concentrations of $PM_{2.5}$ [1–7]. Moving beyond these studies, determining how the relationship between $PM_{2.5}$ and mortality might change as exposure changes is a fundamental challenge important to determining whether and to what extent tightening air quality standards for $PM_{2.5}$ will result in public health benefits. Such determinations are complicated by the presence of potential exposure measurement error, as has been suggested by other authors [8,9].

Competing interests: The authors have declared that no competing interests exist.

The major cohort epidemiological analyses done to date have involved people who were enrolled at the beginning of a study, with data such as health markers and demographic information gathered at enrollment. The cohort directors followed the study subjects over time and marked when certain health events occurred, such as a medical diagnosis or death. Study authors also estimated the subjects' air pollution exposure and conducted statistical analyses (often Cox proportional hazards modeling) to determine whether there was an association between the pollutant exposure and the health effect in the specific cohort of people. The exposure estimation is rarely at the personal level; exposure is often assigned at a community level, meaning that all members of one community (such as a census tract, a zip code, or a whole city) are assigned the same estimate of pollution exposure. Recently air quality modeling methods have been used to develop much more refined locational exposure estimates for epidemiological studies, but also without personal exposure data.

Setting aside thorny questions of determining whether the relationship between $PM_{2.5}$ and mortality is causal, the "shape" of the concentration-response (C-R) function underlying that association is an important consideration when setting an adequately-protective ambient air quality standard for $PM_{2.5}$. A C-R relationship's shape can determine whether changes in $PM_{2.5}$ at relatively low concentrations will have the same, greater, or lesser effects on health risk as changes at concentrations that contributed to an observed association. Although shape can have many forms, a common shape of interest to policy makers is whether there may be an effect "threshold" such that $PM_{2.5}$ levels below this threshold are not expected to produce adverse health effects.

A majority of the previous research on the relationship between long-term $PM_{2.5}$ exposure and mortality in U.S.-based cohorts has not detected such a threshold. There are at least three possible explanations for these findings. The first is that there is in fact no threshold in the C-R function, and any concentration of $PM_{2.5}$ will produce an adverse health effect. A second possible explanation is that a C-R threshold does exist, but the historical $PM_{2.5}$ concentrations in these studies have almost always been above the C-R threshold. For example, while the U.S. population-weighted annual average $PM_{2.5}$ concentration in 2016 was 9.0 $\mu g/m^3$ [10], less than 10 percent of the $PM_{2.5}$ concentrations measured from 1999–2000 in the American Cancer Society (ACS) study fell below this level [3]. A C-R threshold near the bottom of the range measured by these studies might be difficult to detect, even with quite reliable exposure estimates. A third possibility, and the subject of this paper, is that a C-R threshold does exist even within the range of $PM_{2.5}$ concentrations observed in the study, but it is obscured by measurement error in the estimated $PM_{2.5}$ concentrations.

There are two major types of measurement error that could affect the $PM_{2.5}$ exposure estimate: Berkson-type error and classical-type error [11]. Berkson-type error typically arises from differences between each individual's personal $PM_{2.5}$ exposure and the community-average personal exposure as measured by a $PM_{2.5}$ monitor or estimated by a model. With Berkson-type error the true $PM_{2.5}$ exposure for each individual is randomly distributed around the measured $PM_{2.5}$ value, but the true population-weighted mean $PM_{2.5}$ value is accurately measured and will thus not lead to bias in estimates of the C-R function, although it is expected to increase the variance of the estimate. Classical-type error will arise if there are differences between the true population-weighted exposure level and the observed measures of $PM_{2.5}$. It is well known that classical-type measurement error can lead to biased estimates of a C-R function's parameters [e.g., 12–18]. In summarizing previous research on this topic, Rhomberg et al. [9] conclude "[t]he majority of the literature, both theoretical and experimentally based, indicates that measurement error results in the masking of a threshold, and that even when exposure measurement error is not large enough to completely linearize a truly threshold exposure-response relationship, bias still exists." This study adds to that literature by

conducting what is, to the best of our knowledge, the first simulation of the effects of measurement error on the statistical models commonly employed in long-term cohort studies.

One common approach to testing for thresholds in C-R functions is to estimate a statistical model relating a pollutant to mortality, assuming the pollutant has no effect on mortality below a specified threshold, and a linear effect above the threshold such that increasing levels of the pollutant increase the risk of mortality. The fit of the threshold model is then tested against the fit of a no-threshold model [e.g., 19,20]. The statistical model used to test for a threshold in this paper is the Cox proportional hazards model, a commonly-employed model in the study of long-term cohort studies that assesses the relationship between a risk factor (such as air pollution) and survival time. Another approach to testing for thresholds in C-R functions is to estimate a non-parametric (spline) regression of the relative risk of mortality on the exposure range of the pollutant, and examine the resulting plot for evidence of a threshold [e.g., 21]. We examine the effect of measurement error on both approaches.

We note that simulation is an important supplement to standard epidemiological investigations that rely on observed rather than simulated evidence. In standard epidemiology studies with observed cohort data the true mean $PM_{2.5}$ value for each community and the true shape of the C-R relationships are unknown to epidemiological researchers, so it is not possible to determine the effect of classical measurement error on the ability to detect a C-R threshold. In this study we test the degree to which measurement error affects the ability to statistically detect a C-R threshold under a variety of different conditions (different C-R thresholds, levels of measurement error, and hazard ratios), using a large, simulated cohort. We also test the degree to which estimates of the level of a true C-R threshold and of the true hazard ratio associated with $PM_{2.5}$ are affected by differing amounts of classical measurement error.

Methods

Our simulations are described in detail below. A flow chart representing the structure of our simulations is presented in Fig 1.

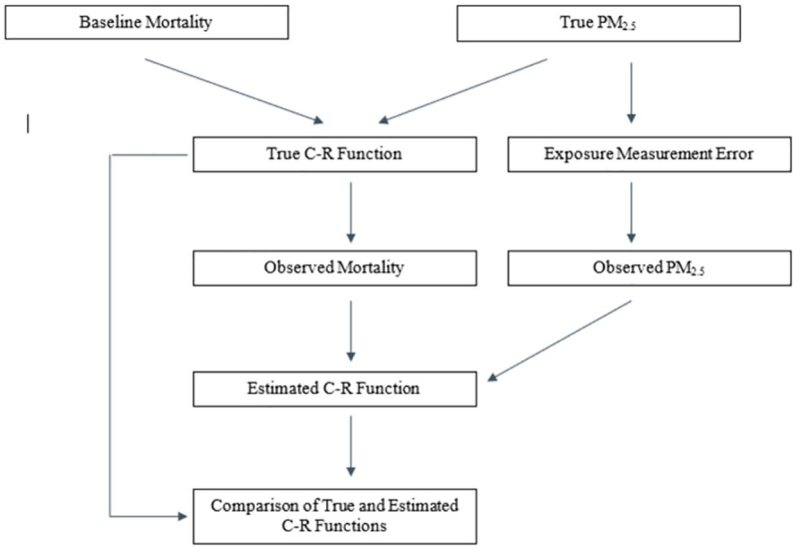

Fig 1. Flow chart of simulation structure.

https://doi.org/10.1371/journal.pone.0264833.g001

Generating the simulated cohorts

We generated a cohort of simulated individuals that was tracked for 20 years, from 2000 to 2020. Our simulated cohort consists of populations from 100 different hypothetical cities, each with 20,000 simulated individuals, for a total cohort size of 2 million individuals.

The baseline mortality rate for the simulated individuals in our cohort was calculated from cohort life tables compiled by the US Social Security Administration [22]. These life tables give the probability of mortality at each age based on birth year and sex, with birth year reported in 10-year increments from 1900 to 2100. We used linear interpolation to assign mortality probabilities for birth years that fell between those in the life table.

For each simulated individual, a birth year was calculated by subtracting the age assigned to that individual from the first calendar year of the simulation (2000), and the probability of mortality for each simulated individual in each year of the simulation was assigned based on the mortality probabilities calculated above.

For each simulated individual in each year of the simulation, mortality was determined by a random draw from a uniform distribution in the range of 0 to 1. If the random draw was less than the probability of mortality in that year for a simulated individual, that individual was recorded as dying in that year.

To clarify the influence of measurement error on the detection of a threshold in a C-R function, we limited the cohort we generated to only males aged 60 at the start of the simulation. We did this to eliminate variation in the baseline mortality rate based on age and sex that would otherwise make the detection of a threshold more difficult. Variation in age and sex could also be introduced into the cohort and is an area for future study.

$PM_{2.5}$ levels and the C-R functions

To create a dataset with $PM_{2.5}$ conditions for 100 hypothetical cities that would be generally consistent with those in the U.S., we used the weighted annual mean of daily $PM_{2.5}$ concentrations from all 229 core-based statistical areas (CBSAs) in the EPA's AQS database from 2000 to 2016, with the weighted annual mean calculated as the average of the quarterly averages of the 24-hour values [23]. We used linear extrapolation to extend the data for each CBSA out to 2020 to create a twenty-year data set. We organized the data into quartiles, and randomly selected 25 sets of $PM_{2.5}$ values from each quartile, which were then randomly assigned to the 100 hypothetical cities. For this analysis, we have assumed no time-variation in $PM_{2.5}$ concentrations and have used the simple average $PM_{2.5}$ concentration from the 2000–2020 period of actual monitored values, plus a random draw from a uniform distribution bounded between -1 and 1 $\mu g/m^3$ to arrive at the $PM_{2.5}$ concentrations for the hypothetical cities. These $PM_{2.5}$ values were then used in our simulations as the "true" population-weighted average exposure. The additional effect of time-varying $PM_{2.5}$ on simulation results is an area for future study.

In each hypothetical city in each year, true $PM_{2.5}$ concentration was assumed to influence the probability of mortality. Specifically, the probability of mortality (P_{ijt}) for simulated individual i in city j at time t was calculated as:

$$P_{ijt} = B_{ijt} \times h^{PM_{jt}} \quad (1)$$

where B_{ijt} is the baseline probability of mortality for simulated individual i in city j at time t, PM_{jt} is the $PM_{2.5}$ level in city j at time t, and h is the hazard ratio (HR) related to exposure to $PM_{2.5}$. This equation defines a linear C-R function on the log hazard scale. Further, the entire risk from $PM_{2.5}$ exposure in this C-R function occurs in the same year as exposure (i.e., there are no lags or cumulative effects to complicate the detection of the true C-R shape).

In our simulations we test C-R functions with well-defined (i.e., "hockey-stick"), population-wide thresholds, such that $PM_{2.5}$ below the threshold has no effect on mortality, and $PM_{2.5}$ above the threshold has a linear effect. To do this we subtract the threshold from the $PM_{2.5}$ level in each city and year, setting "effective" $PM_{2.5}$ to zero if this calculation results in a negative number. This modified measure of $PM_{2.5}$ is then substituted in for PM_{jt} in the equation above. This produces a true C-R function that is zero below the threshold, and linear above the threshold.

We ran simulations for a wide range of alternative C-Rs. We examined three alternative "true" $PM_{2.5}$ threshold levels: 7 $\mu g/m^3$, 8.5 $\mu g/m^3$, and 9.5 $\mu g/m^3$. These thresholds correspond approximately to the 15^{th}, 40^{th}, and 55^{th} percentiles of our assumed true $PM_{2.5}$ exposure levels and were selected to test various threshold levels approximately at or below the mean level of $PM_{2.5}$. In our simulations we considered all combinations of these three C-R thresholds with five alternative "true" HRs above the threshold (1.0025, 1.005, 1.01, 1.02, and 1.05 per $\mu g/m^3$), for a total of 15 different C-Rs.

Measurement error

The measurement error we examine in our simulations allows the "observed" average $PM_{2.5}$ exposure assigned to a hypothetical city to deviate from the "true" average $PM_{2.5}$ exposure experienced by the individuals in these cities. This type of measurement error could arise if, for example, the monitoring location used to measure $PM_{2.5}$ levels does not reflect the "true" population-weighted average $PM_{2.5}$ levels experienced by the population. This is a type of classical measurement error (as opposed to Berkson measurement error, under which individual exposures vary randomly around the population average exposure).

The "observed" $PM_{2.5}$ measures for each hypothetical city were created by adding a random draw to the city's "true" $PM_{2.5}$ level. The random draws came from a truncated normal distribution with bounds at ± 4 $\mu g/m^3$. Increasing amounts of measurement error were created by increasing the standard error on this truncated normal distribution. We considered draws of measurement error from distributions with standard deviations of 1, 2, and 4 $\mu g/m^3$.

For each standard deviation, we drew 100 different sets of "observed" $PM_{2.5}$ values for each of the hypothetical cities in our simulations. We then ran 100 versions of the epidemiological models described below for each simulated cohort, once for each of the 100 different sets of "observed" $PM_{2.5}$ values. The same sets of "observed $PM_{2.5}$" were used for each set of epidemiological models tested below. Thus, any variation in the ability to detect a C-R threshold across different "true" C-R functions is not due to variation in the measurement errors used from test to test.

Empirical tests

In our simulations we examined all combinations of true C-Rs and measurement error. There were three "true" $PM_{2.5}$ thresholds (7, 8.5, and 9.5 $\mu g/m^3$), five "true" HRs (1.0025, 1.005, 1.01, 1.02, and 1.05 per $\mu g/m^3$), and three levels of measurement error (distributed as a truncated normal with standard deviations of 1, 2, and 4 $\mu g/m^3$), for a total of 45 combinations of variables. Each of these combinations was represented by a separate simulated set of cohort survival outcome data, on which we ran epidemiological models to estimate the underlying true C-R relationship. For each of these 45 simulated cohort survival outcome datasets, we ran 100 models, one for each of the 100 different sets of "observed" $PM_{2.5}$ values described above.

As a preliminary examination we first fit splines to the simulated data. The dependent variable in our splines was relative risk as calculated from the "observed" mortality in each city, with relative risk defined as 1 for the lowest "observed" level of $PM_{2.5}$. For each of the 45

combinations of threshold, HR, and measurement error, natural cubic splines with four degrees of freedom were estimated on the relative risk and examined for evidence that thresholds in a C-R function would be more difficult to detect as measurement error increases.

Next, for each of these combinations, we ran a series of Cox proportional hazard (PH) models to test for our ability to detect a C-R threshold in the face of measurement error. For each "true" threshold, we created a new $PM_{2.5}$ measure by subtracting the "true" threshold from "observed" $PM_{2.5}$, as described above. This measure of $PM_{2.5}$ was then used to estimate a Cox PH model. For each combination of variables, we also estimated a model that assumed there was no C-R threshold.

We then compared the statistical fits of these models. For each combination of "true" threshold, HR, and level of measurement error, both a model that assumed the true threshold and a no-threshold model were estimated for each of the 100 values of "observed" $PM_{2.5}$. The threshold model and the no-threshold model are not nested, so statistical tests such as the likelihood-ratio test cannot be applied. Thus, we followed the suggestion of Jerrett et al. [20,24] and calculated a test statistic of 2 times the difference in the log-likelihoods between the threshold model and the no-threshold model *(2 × (LL(threshold)–LL(no threshold))*, which was then compared to the likelihood-penalty function applied by the Akaike information criteria (AIC) and Bayesian information criteria (BIC) for one additional parameter, which for the AIC is 2 and for the BIC is ln(n), with n defined as either the number of deaths observed in the data or the number of individuals in the data. A test statistic greater than any one of these values could lead one to conclude that the threshold model is a better fit than the no-threshold model. Some of these tests are more stringent than others (i.e., are less likely to reject the non-threshold hypothesis). The simulation results presented below use the most stringent test standard proposed by Jerrett et al. [24] and compare the difference in the log-likelihoods to the natural log of the number of individuals in the data. Simulation results based on the other test standards proposed by Jerrett et al. [24] are presented in the Supporting Information (SI).

We also examine the estimated C-R threshold levels (based on the best fitting model) for each combination of "true" threshold, HR, and measurement error for each of the 100 values of "observed" $PM_{2.5}$. For each simulated dataset a grid search approach was used to test the fit of various candidate thresholds. We examined alternative threshold estimates in the range ± 4 $\mu g/m^3$ around the "true" $PM_{2.5}$ threshold value, incremented by 1 $\mu g/m^3$. For example, if the "true" threshold was 8.5 $\mu g/m^3$, we examined models that assumed thresholds ranging between 4.5 and 12.5 $\mu g/m^3$, incremented by 1 $\mu g/m^3$. For each potential threshold, we created a new $PM_{2.5}$ measure by subtracting the potential threshold from the "observed" $PM_{2.5}$. This measure of $PM_{2.5}$ was then used to estimate a Cox PH model. The threshold from the model with the best log-likelihood was then reported as the best-fitting threshold.

We also examined the estimated HRs for each combination of "true" threshold, HR, and measurement error.

Results

Threshold detection with splines

Examination of splines estimated on each combination of threshold, HR, and measurement error indicates that as measurement error increases, the existence of a threshold becomes more difficult to determine, the estimated threshold location decreases, and the HR becomes attenuated.

Here we discuss splines estimated on the data produced by simulations with a threshold of 8.5 $\mu g/m^3$ and an HR of 1.05. Fig 2 presents four sets of splines estimates (each estimated with 4 degrees of freedom). One scenario assumes no measurement error, while the other three

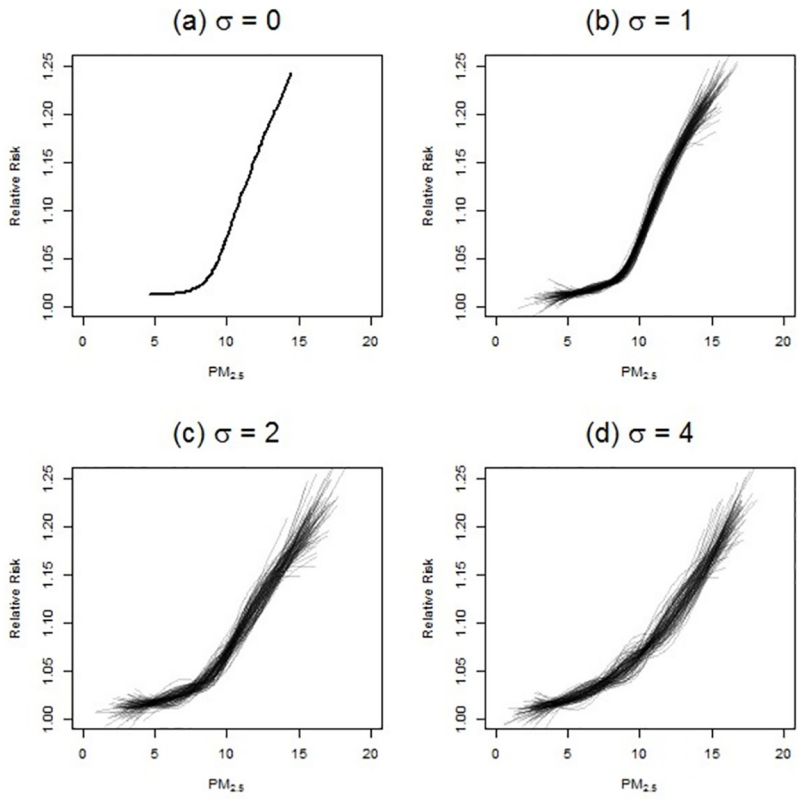

Fig 2. Spline estimates of same "true" C-R function under varying amounts measurement error.

https://doi.org/10.1371/journal.pone.0264833.g002

examine the three levels of measurement error (σ = 1, 2, or 4 μg/m^3). Each panel in Fig 2 presents 100 splines estimated using one of 100 separately and randomly generated sets of city-specific "observed" values of $PM_{2.5}$ for a particular level of measurement error: σ = 0, 1, 2, or 4 μg/m^3. In all 100 iterations, the true C-R function and the true city-specific $PM_{2.5}$ values were the same, with the identical dataset on cohort survival outcomes being used for each spline; only the draws of the measurement error differ across the 100 estimated splines and cause the differences from spline to spline in each panel.

The presence of a threshold in the C-R function at a $PM_{2.5}$ concentration of 8.5 μg/m^3 is clearly apparent in panel (a) of Fig 2, when there is no measurement error. As measurement error increases, both the apparent level of the threshold and strength of the relationship between $PM_{2.5}$ and mortality are attenuated towards zero. Although some degree of "sub-linear" shape is evident when viewing all 100 splines overlaid in a single panel, many of the individual splines have no threshold-like shape at all; some even take the opposite, "supra-linear" shape over parts of the C-R function. However, it is misleading to rely on a comparison across 100 draws of measurement error to understand the "model uncertainty" created by measurement error: any actual epidemiological study faces the equivalent of a single draw, with a single resulting C-R estimate. The potential degree of deviation of the shape of that C-R estimate from the true underlying shape can be much greater than the aggregated patterns in Fig 2.

All else being equal, such model uncertainty increases with increasing measurement error, but it can be exacerbated or moderated under different locations of the true threshold (relative

to the observed range of concentrations) and different levels of the true HR. Figures equivalent to Fig 2 for every combination of threshold and HR are presented in the SI (S1 to S15 Figs). We note that in some of these cases the evidence of a nonlinear shape dissipates across the 100 measurement error draws (e.g., S3 Fig). We also note that in certain cases, noise in the cohort simulation itself (due to the mortality outcomes being determined by random draws) is sometimes large enough that the splines estimated on the cohort outcomes are unable to detect the presence of a threshold in the true C-R function even when such a threshold exists (e.g., S11 Fig).

Measurement error affects the ability of splines to detect the true shape of the underlying C-R relationship by shifting each observed value of $PM_{2.5}$ horizontally in either a positive or negative direction, thus "flattening out" the relationship between $PM_{2.5}$ and mortality, resulting in non-zero estimates of risk below the C-R threshold. This phenomenon also affects the ability of parametric statistical methods used in cohort studies to produce reliable estimates of the true C-R shape, as we show below.

Parametric statistical tests for thresholds

Table 1 presents the results from testing the fit of Cox PH models that assume the correct threshold compared to the fit of a no-threshold model for each level of measurement error (σ = 1, 2, or 4 $\mu g/m^3$) for the five different HRs considered in the simulations. Results for each of the three C-R thresholds we consider are presented in separate columns. The values in each row indicate the number of times (out of 100 sets of city-specific measurement error assignments) the Cox PH model that assumed the true threshold fit the data better than the no-threshold Cox PH model. The test statistic used is based on 2 times the difference in the log-likelihoods between the threshold and no-threshold models being larger than the natural log of the number of individuals in the data ($2 \times \Delta LL > ln(n)$).

Table 1. Rejection of the no C-R threshold model under varying amounts of measurement error.

	Threshold = 7	Threshold = 8.5	Threshold = 9.5
HR = 1.0025			
σ = 1	0	0	0
σ = 2	0	0	0
σ = 4	0	0	0
HR = 1.005			
σ = 1	13	0	0
σ = 2	4	2	0
σ = 4	5	0	0
HR = 1.01			
σ = 1	1	99	67
σ = 2	3	71	27
σ = 4	3	37	18
HR = 1.02			
σ = 1	96	100	100
σ = 2	67	100	98
σ = 4	43	82	91
HR = 1.05			
σ = 1	100	100	100
σ = 2	94	100	100
σ = 4	79	100	100

https://doi.org/10.1371/journal.pone.0264833.t001

Three patterns are apparent in Table 1. First, the ability to detect a threshold increases as the HR increases. Higher hazard ratios lead to a more rapid increase in risk once the C-R threshold is crossed, making it easier to distinguish the boundary between zero and positive risk. For HRs less than 1.01 there was little detection of the true threshold at all, even for the highest threshold level tested.

Second, the ability to detect a threshold tends to increase when the threshold is located higher in the range of true $PM_{2.5}$ exposures experienced by the cohort. Recall that the thresholds of 7, 8.5, and 9.5 μg/m^3 correspond approximately to the 15th, 40th, and 55th percentiles of the true $PM_{2.5}$ exposure levels, respectively. Our simulations find that for HRs of 1.01 or higher (the HR levels for which a meaningful fraction of the simulations does detect the threshold), as the true threshold increases, the Cox PH model that assumes a threshold is more likely to detect it, especially when the threshold is increased from 7 to 8.5 μg/m^3. This pattern does not seem to hold for the lower HRs.

Third, for the higher HR and threshold levels, the ability to detect a threshold decreases as measurement error increases (as σ grows larger). In all cases, the number of times the threshold model fit the data better than the no-threshold model decreased as the amount of measurement error increased.

Simulation results based on the other test standards proposed by Jerrett et al. [24] are presented in the SI (S1 and S2 Tables). These alternative results support the conclusions we describe here.

Estimated threshold levels

Tables 2–4 present the results from estimating the location of the C-R threshold based on the best-fitting Cox PH model across a range of candidate thresholds and across the 100 different

Table 2. Best-fitting C-R threshold level under varying amounts of measurement error, true threshold = 7.

	Potential C-R Threshold Tested for Goodness of Fit							
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**
HR = 1.0025								
σ = 1	1		4	7	20	26	33	9
σ = 2	7	3	5	14	16	22	11	14
σ = 4	11	9	17	11	9	14	9	11
HR = 1.005								
σ = 1				1	23	63	13	
σ = 2	1	2	5	18	24	31	18	1
σ = 4	6	6	6	12	24	19	15	7
HR = 1.01								
σ = 1	2	2	10	50	35	1		
σ = 2	20	18	28	17	11	6		
σ = 4	28	20	17	8	13	8	5	
HR = 1.02								
σ = 1			1	11	83	5		
σ = 2		4	17	36	32	8	3	
σ = 4	9	17	22	16	19	14	3	
HR = 1.05								
σ = 1				24	75	1		
σ = 2			26	49	19	5	1	
σ = 4	6	29	25	13	17	6	4	

https://doi.org/10.1371/journal.pone.0264833.t002

Table 3. Best-fitting C-R threshold level under varying amounts of measurement error, true threshold = 8.5.

	Potential C-R Threshold Tested for Goodness of Fit							
	4.5	**5.5**	**6.5**	**7.5**	**8.5**	**9.5**	**10.5**	**11.5**
HR = 1.0025								
σ = 1		3		3	24	48	20	2
σ = 2	4	5	5	12	15	29	15	13
σ = 4	5	4	7	8	16	12	19	10
HR = 1.005								
σ = 1	3	1	2	25	53	15	1	
σ = 2	5	10	18	19	25	13	7	2
σ = 4	19	6	11	19	20	5	10	4
HR = 1.01								
σ = 1				1	63	34	2	
σ = 2		1	3	25	53	15	3	
σ = 4	6	7	14	25	27	13	3	5
HR = 1.02								
σ = 1				23	77			
σ = 2			14	56	26	4		
σ = 4	3	11	31	29	16	10		
HR = 1.05								
σ = 1				21	79			
σ = 2			7	67	22	3	1	
σ = 4		6	29	40	21	3		1

https://doi.org/10.1371/journal.pone.0264833.t003

Table 4. Best-fitting C-R threshold level under varying amounts of measurement error, true threshold = 9.5.

	Potential C-R Threshold Tested for Goodness of Fit							
	5.5	**6.5**	**7.5**	**8.5**	**9.5**	**10.5**	**11.5**	**12.5**
HR = 1.0025								
σ = 1	56	11	12	14	5		1	1
σ = 2	39	11	13	23	9	2	2	1
σ = 4	41	12	13	9	7	8	5	3
HR = 1.005								
σ = 1	2	8	44	40	6			
σ = 2	26	16	18	23	14	2	1	
σ = 4	30	12	11	22	5	11	5	4
HR = 1.01								
σ = 1			5	50	43	2		
σ = 2	8	6	29	40	12	4	1	
σ = 4	12	12	26	24	19	3	3	1
HR = 1.02								
σ = 1				17	81	2		
σ = 2		1	15	50	25	9		
σ = 4		10	23	34	15	8	10	
HR = 1.05								
σ = 1				11	88	1		
σ = 2			8	62	25	5		
σ = 4		3	20	48	12	10	6	1

https://doi.org/10.1371/journal.pone.0264833.t004

sets of "observed" values of $PM_{2.5}$. For the three "true" C-R threshold levels that we analyzed (each in a separate table) we present results for each combination of HR and level of measurement error. The values in each row indicate the number of times out of 100 each potential threshold (as identified by the column headers) was found to produce the best-fitting Cox PH model. As there were 100 separate sets of observed $PM_{2.5}$ modeled for each row, the numerical values in each row sum to 100.

It is apparent in examining Tables 2–4 that measurement error makes detection of the true C-R threshold difficult. The effect of measurement error on the ability to detect a threshold is most clear when considering the higher HRs in the simulation (HR = 1.02 or 1.05). There the best-fitting Cox PH model indicates the "true" threshold in a majority of the 100 tests for the lowest level of measurement error ($\sigma = 1$), and underestimates of the threshold become more common as measurement error increases. For lower HRs the patterns become less clear, and for very low HRs overestimates of the threshold are frequent. However, for most combinations of HR, threshold, and levels of measurement error we examine, underestimates of the threshold are more common than overestimates.

Estimated hazard ratios

Fig 3 considers the HRs estimated by the best-fitting Cox PH model across the 100 different "observed" values of $PM_{2.5}$. Each panel of Fig 3 presents the results for one hazard ratio, which is indicated by the solid vertical line on the panel. For each combination of "true" threshold

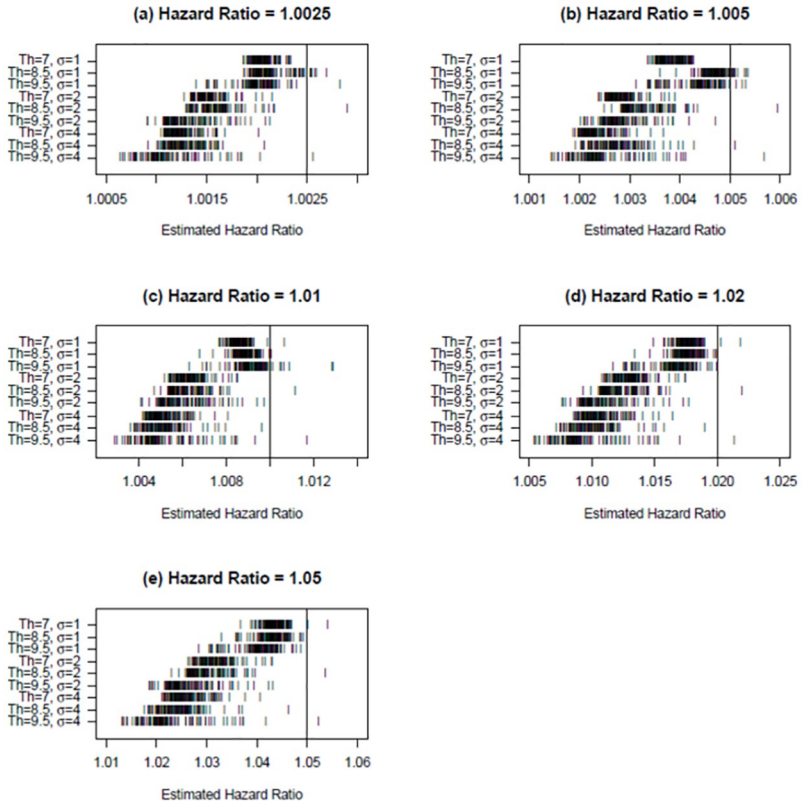

Fig 3. Attenuation in hazard ratios under varying amounts of measurement error and different "true" C-R thresholds.

https://doi.org/10.1371/journal.pone.0264833.g003

Table 5. True hazard ratio above upper limit of estimated hazard ratio, under varying amounts of measurement error and different "true" C-R thresholds.

	Threshold = 7	Threshold = 8.5	Threshold = 9.5
HR = 1.0025			
σ = 1	0	0	0
σ = 2	86	51	58
σ = 4	95	91	100
HR = 1.005			
σ = 1	72	2	20
σ = 2	99	91	94
σ = 4	100	96	90
HR = 1.01			
σ = 1	85	42	19
σ = 2	100	99	95
σ = 4	100	99	95
HR = 1.02			
σ = 1	98	90	95
σ = 2	100	100	100
σ = 4	100	99	99
HR = 1.05			
σ = 1	99	98	99
σ = 2	100	100	100
σ = 4	100	100	99

https://doi.org/10.1371/journal.pone.0264833.t005

and level of measurement error (identified on the vertical axis of the panel), each tick mark in a panel indicates an estimated hazard ratio.

Two patterns are apparent when examining Fig 3. First, as expected, the estimated HRs become more attenuated as the level of measurement error increases. This pattern holds true regardless of which HR or threshold is considered. Second, the amount of attenuation related to measurement error tends to be higher as the threshold increases, especially at higher levels of measurement error.

In most cases, the 95 percent confidence interval on the HRs presented in Fig 3 would reject the "true" HR as too high. These results are presented in Table 5. Each numerical result indicates the number of times the upper limit of the 95 percent confidence interval on the HR in the best-fitting Cox PH model falls below (fails to cover) the "true" HR assumed in that simulation. Table 5 shows that the best-fitting Cox PH model is more likely to reject the "true" HR as too high as measurement error increases across all three "true" thresholds, especially for larger values of the "true" HR.

Discussion

In this study we generated simulated cohorts with C-R functions for $PM_{2.5}$ that varied both in the location of a C-R threshold and the HR associated with $PM_{2.5}$. We found that as measurement error increased, our ability to statistically detect a threshold decreased, and both the estimated location of the threshold and the estimated HR were attenuated. Beyond these general observations, examination of nonlinear splines revealed that measurement error could lead to a variety of different estimated shapes for the C-R function for the same underlying "true" C-R function. Since the true level of measurement error is unknown, this introduces considerable model uncertainty into the estimated shapes of C-R functions from epidemiological studies using observed data.

The results of our study are consistent with other simulation-based studies that have assessed how measurement error affects estimates of the shape of a pollutant's C-R relationship. Some studies have demonstrated that measurement error leads to attenuated estimates of the slope of a C-R function, without reference to the shape of the C-R function [12,13]. Other simulation-based studies have also examined the shape of the C-R function and have demonstrated that exposure measurement error can "linearize" and "flatten" estimates of a segmented linear, or "hockey stick" C-R function such as the ones we examine here [8,9,14–18]. As in our study, these studies find that exposure measurement error leads to (1) attenuated estimates of the slope of a C-R function, (2) estimates that suggest a C-R function is linear when it in fact has a threshold, and (3) underestimates of the location of a threshold in a C-R function. To the best of our knowledge, our study is the first simulation-based study to test for the magnitude of these effects using the methods commonly employed in cohort studies of the relationship between long-term $PM_{2.5}$ exposure and mortality. We do this by simulating prospective cohort survival data and applying Cox proportional hazard and spline regression methods to assess C-R shape. As our study is focused on the methods commonly employed in cohort studies, it does not consider statistical approaches that might better address concerns with measurement error, such as Bayesian methods that incorporate knowledge or beliefs about the measurement error process [25].

Our study does not consider other sources of uncertainty and variability that might also affect the ability to detect a threshold in a C-R function, such as Berkson measurement error, confounding variables, differences in the properties of $PM_{2.5}$ in different locations, or the effects of pollutants other than $PM_{2.5}$. For instance, Moolgavkar et al. [26] simulated a prospective cohort and applied a Cox PH model to study the potential for spurious small associations to be detected (or masked) if the largest risk factors (e.g., smoking) violate the proportional hazards assumption, and found that inadequate control for strong, time-dependent confounding produces unreliable estimates of risk. Our simulations show that even with a single risk factor and very "clean" data that meets the proportional hazards assumption, measurement error alone can make estimates of the shape of the C-R function unreliable. The difficulties we note in our paper would probably be greatly exacerbated if we were to also consider confounding covariates, or the other sources of uncertainty and variability listed above) [26]. The simulations we describe here could be extended to cover these and other issues.

Most previous epidemiological research on the relationship between $PM_{2.5}$ and mortality that has considered the possibility of a C-R threshold has not detected a statistically significant threshold, and risk analyses based on these epidemiological results then often assume that no such threshold exists. Our results show that the inability of previous research to detect a C-R threshold may be due to measurement error, rather than the nonexistence of such a threshold. Put another way, "because of the prevalence of exposure measurement error in epidemiology data and lack of reliable error-mitigating techniques, conclusions about the linearity of the exposure-response curve must be examined carefully and treated with some scepticism" [9]. This has obvious implications for determining appropriate air quality standards, since most policy makers have relied on risk analyses that have assumed no C-R thresholds exist.

Conclusions

To the best of our knowledge, this is the first simulation-based study to examine the effect of classical-type measurement error in pollutant exposure on estimates of the shape of a pollutant's C-R function using simulations of prospective cohort data, and then applying the statistical models commonly employed in long-term cohort studies of the relationship between long-term $PM_{2.5}$ exposure and mortality. The results of our study demonstrate that exposure

measurement error obscures the existence of a threshold in the C-R function when such a threshold in fact exists and leads to attenuated estimates of both the estimated location of the C-R threshold and the estimated hazard ratio associated with $PM_{2.5}$. These results have clear implications for determining appropriate air quality standards for pollutants. The extent of measurement error in estimates of pollutant exposure should be more carefully quantified, and its potential effects on uncertainty in the shape of the C-R functions merits consideration by policy makers when setting air quality standards.

Supporting information

S1 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 7 μg/m^3, hazard ratio 1.0025).
(TIF)

S2 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 7 μg/m^3, hazard ratio 1.005).
(TIF)

S3 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 7 μg/m^3, hazard ratio 1.01).
(TIF)

S4 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 7 μg/m^3, hazard ratio 1.02).
(TIF)

S5 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 7 μg/m^3, hazard ratio 1.05).
(TIF)

S6 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 8.5 μg/m^3, hazard ratio 1.0025).
(TIF)

S7 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 8.5 μg/m^3, hazard ratio 1.005).
(TIF)

S8 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 8.5 μg/m^3, hazard ratio 1.01).
(TIF)

S9 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 8.5 μg/m^3, hazard ratio 1.02).
(TIF)

S10 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error threshold 8.5 μg/m^3, hazard ratio 1.05).
(TIF)

S11 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 9.5 μg/m^3, hazard ratio 1.0025).
(TIF)

S12 Fig. Spline Estimates of same "true" C-R function under varying amounts of measurement error (threshold 9.5 μg/m^3, hazard ratio 1.005).
(TIF)

S13 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 9.5 μg/m^3, hazard ratio 1.01).
(TIF)

S14 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 9.5 μg/m^3, hazard ratio 1.02).
(TIF)

S15 Fig. Spline estimates of same "true" C-R function under varying amounts of measurement error (threshold 9.5 μg/m^3, hazard ratio 1.05).
(TIF)

S1 Table. Rejection of the no C-R threshold model under varying amounts of measurement error ($2 \times \Delta LL > 2$).
(PDF)

S2 Table. Rejection of the no C-R threshold model under varying amounts of measurement error ($2 \times \Delta LL > \ln(nevents)$).
(PDF)

Author Contributions

Conceptualization: Garrett Glasgow, Anne E. Smith.

Formal analysis: Garrett Glasgow, Bharat Ramkrishnan, Anne E. Smith.

Funding acquisition: Garrett Glasgow, Anne E. Smith.

Investigation: Garrett Glasgow, Anne E. Smith.

Methodology: Garrett Glasgow, Bharat Ramkrishnan, Anne E. Smith.

Project administration: Garrett Glasgow, Anne E. Smith.

Supervision: Garrett Glasgow, Anne E. Smith.

Validation: Garrett Glasgow, Bharat Ramkrishnan, Anne E. Smith.

Visualization: Garrett Glasgow, Bharat Ramkrishnan.

Writing – original draft: Garrett Glasgow, Bharat Ramkrishnan, Anne E. Smith.

Writing – review & editing: Garrett Glasgow, Bharat Ramkrishnan, Anne E. Smith.

References

1. Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, et al. 1993. An association between air pollution and mortality in six US cities. N Engl J Med. 329: 1753–1759. https://doi.org/10.1056/NEJM199312093292401 PMID: 8179653.
2. Eftim S, Samet J, Janes H, McDermott A, Dominici F. 2008. Fine particulate matter and mortality: A comparison of the Six Cities and American Cancer Society cohorts with a Medicare cohort. Epidemiology. 19: 209–216. https://doi.org/10.1097/EDE.0b013e3181632c09 PMID: 18223484.
3. Krewski D, Jerrett, M, Burnett RT, Ma R, Hughes E, Shi Y, et al. 2009. Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality. HEI Research Report 140. Health Effects Institute: Boston, MA. 19627030.

4. Laden F, Schwartz J, Speizer FE, Dockery DW. 2006. Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard Six Cities study. Am J Respir Crit Care Med. 173: 667–672. https://doi.org/10.1164/rccm.200503-443OC PMID: 16424447.

5. Lepeule J, Laden F, Dockery D, Schwartz J. 2012. Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009. Environ Health Perspect. 120: 965–970. https://doi.org/10.1289/ehp.1104660 PMID: 22456598.

6. Pope CA, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, et al. 1995. Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. Am J Respir Crit Care Med. 151: 669–674. https://doi.org/10.1164/ajrccm/151.3_Pt_1.669 PMID: 7881654.

7. Pope CA, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, et al. 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA. 287: 1132–1141. https://doi.org/10.1001/jama.287.9.1132 PMID: 11879110.

8. Brauer M, Brumm J, Vedal S, Petkau AJ. 2002. Exposure misclassification and threshold concentrations in time series analyses of air pollution health effects. Risk Anal. 22: 1183–93. https://doi.org/10.1111/1539-6924.00282 PMID: 12530788.

9. Rhomberg LR, Chandalia JK, Long CM, Goodman JE. 2011. Measurement error in environmental epidemiology and the shape of exposure-response curves. Crit. Rev. Toxicol. 41: 651–671. https://doi.org/10.3109/10408444.2011.563420 PMID: 21823979.

10. Apte JS, Brauer M, Cohen AJ, Ezzati M, Pope CA. 2018. Ambient $PM_{2.5}$ reduces global and regional life expectancy. Environ. Sci. Technol. Lett. 5: 546–551. https://doi.org/10.1021/acs.estlett.8b00360

11. Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, Dockery D, et al. 2000. Exposure measurement error in time-series studies of air pollution: Concepts and consequences. Environ Health Perspect. 108: 419–426. https://doi.org/10.1289/ehp.00108419 PMID: 10811568.

12. Butland BK, Armstrong B, Atkinson RW, Wilkinson P, Heal MR, Doherty RM, et al. 2013. Measurement error in time-series analysis: A simulation study comparing modelled and monitored data. BMC Medical Research Methodology. 13: 136–147. https://doi.org/10.1186/1471-2288-13-136 PMID: 24219031.

13. Dionisio KL, Chang HH, Baxter LK. 2016. A simulation study to quantify the impacts of exposure measurement error on air pollution health risk estimates in copollutant time-series models. Environmental Health. 15: 114–123. https://doi.org/10.1186/s12940-016-0186-0 PMID: 27884187.

14. Cox LA. 2018. Effects of exposure estimation errors on estimated exposure-response relations for $PM_{2.5}$. Environ Res. 164: 636–646. https://doi.org/10.1016/j.envres.2018.03.038 PMID: 29627760.

15. Küchenhoff H, Carroll RJ. 1997. Segmented regression with errors in predictors: Semi-parametric and parametric methods. Stat Med. 16: 169–188. PMID: 9004390.

16. Lipfert FW, Wyzga R. 1996. The effects of exposure error on environmental epidemiology. In: Proceedings of the Second Colloquium on Particulate Air Pollution and Human Mortality and Morbidity, pp. 295–302.

17. Watt M, Godden D, Cherrie J, Seaton A. 1995. Individual exposure to particulate air pollution and its relevance to thresholds for health effects: a study of traffic wardens. Occup. Environ. Med. 52: 790–792. https://doi.org/10.1136/oem.52.12.790 PMID: 8563840.

18. Yoshimura I. 1990. The effect of measurement error on the dose-response curve. Environ. Health Perspect. 87: 173–178. https://doi.org/10.1289/ehp.9087173 PMID: 2269223.

19. Daniels MJ, Dominici F, Samet JM, Zeger SL. 2000. Estimating particulate matter-mortality dose-response curves and threshold levels: An analysis of daily time-series for the 20 largest US cities. Am J Epidemiol. 152: 397–406. https://doi.org/10.1093/aje/152.5.397 PMID: 10981451.

20. Jerrett M, Burnett RT, Pope CA, Ito K, Thurston G, Krewski D, et al. 2009. Long-term ozone exposure and mortality. New England Journal of Medicine. 360: 1085–1095. https://doi.org/10.1056/NEJMoa0803894 PMID: 19279340.

21. Schwartz J, Zanobetti A. 2000. Using meta-smoothing to estimate dose-response trends across multiple studies, with application to air pollution and daily death. Epidemiology. 11: 666–672. https://doi.org/10.1097/00001648-200011000-00009 PMID: 11055627.

22. Bell FC, Miller ML. 2005. Life tables for the United States Social Security Area 1900–2100, Actuarial Study 120. Social Security Administration, Office of the Chief Actuary, SSA Pub. No. 11–11536.

23. US EPA. 2018. Air quality trends by city, 1990–2016. Air Quality System (AQS) database [Last accessed 12 June 2018]. https://www.epa.gov/air-trends/air-quality-cities-and-counties.

24. Jerrett M, Burnett RT, Pope CA, Ito K, Thurston G, Krewski D, et al. 2014. Explanation and interpretation of threshold model presented in "Long-term ozone exposure and mortality." US EPA memorandum "Response to Comments Regarding the Potential Use of a Threshold Model in Estimating the Mortality Risks from Long-term Exposure to Ozone in the Health Risk and Exposure Assessment for Ozone,

Second External Review Draft," April 28, 2014 [Last accessed 9 August 2019]. https://www3.epa.gov/ttn/naaqs/standards/ozone/data/20140428responsetocomments.pdf.

25. Edwards JK, Keil AP. 2017. Measurement error and environmental epidemiology: A policy perspective. Curr Environ Health Rep. 4: 79–88. https://doi.org/10.1007/s40572-017-0125-4 PMID: 28138941.

26. Moolgavkar SH, Chang ET, Watson HN, Lau EC. 2018. An assessment of the Cox proportional hazards regression model for epidemiologic studies. Risk Anal. 38: 777–794. https://doi.org/10.1111/risa.12865 PMID: 29168991.