Global Epidemiology 6 (2023) 100114

Contents lists available at ScienceDirect

Global Epidemiology

journal homepage: www.sciencedirect.com/journal/global-epidemiology

# What is an exposure-response curve?

Louis Anthony Cox Jr [*]

*Cox Associates, Entanglement, University of Colorado, United States of America*

ARTICLE INFO

*Keywords:*
Exposure-response curve
Partial dependence plot (PDP)
Accumulated local effects (ALE) plot
Individual conditional expectation (ICE) plot
Causal artificial intelligence (CAI)

ABSTRACT

Exposure-response curves are among the most widely used tools of quantitative health risk assessment. However, we propose that exactly what they mean is usually left ambiguous, making it impossible to answer such fundamental questions as whether and by how much reducing exposure by a stated amount would change average population risks and distributions of individual risks. Recent concepts and computational methods from causal artificial intelligence (CAI) and machine learning (ML) can be applied to clarify what an exposure-response curve means; what other variables are held fixed (and at what levels) in estimating it; and how much inter-individual variability there is around population average exposure-response curves. These advances in conceptual clarity and practical computational methods not only enable epidemiologists and risk analysis practitioners to better quantify population and individual exposure-response curves but also challenge them to specify exactly what exposure-response relationships they seek to quantify and communicate to risk managers and how to use the resulting information to improve risk management decisions.

## Introduction: What does an exposure-response curve mean?

Exposure-response curves are among the most widely used tools of quantitative health risk assessment. Debates over their shapes, especially at low exposure concentration levels, have occupied countless journal articles and thousands of hours of deliberation in regulatory risk assessments and policy-making. Whether the expected number of adverse health effects per unit of exposure to a substance is estimated to be a linear, sublinear, supralinear, or threshold function may determine how much exposure to it is allowed. Acknowledging the enormous effort and large-scale collaborations that have gone into preparing such exposure-response curves, this paper turns to a fundamental interpretive question: *What does an exposure-response curve mean?* We propose that, despite their widespread acceptance in peer-reviewed reports and articles and use in informing applied public health risk assessments and regulatory scientific processes, what exposure-response functions mean is usually left importantly ambiguous at a fundamental conceptual and definitional level. They leave unanswered such fundamental questions as

- How, if at all, would a proposed change in exposure change *average* population risk? (This is different from quantifying the estimated level of population risk for different observed or estimated levels of exposure, as discussed next.)
- How would it change *individual* risks?
- What is the *distribution* of individual exposure-response curves around a population average exposure-response curve?
- What *factors* are assumed to be held fixed in preparing an exposure-response curve? At what levels are they assumed to be held fixed? Are these assumptions realistic?
- In reality, how much do other factors (including unmeasured ones) differ for different observed levels of exposure? How do these differences affect the exposure-response curve?
- By how much would interventions that reduce pollution change other causally relevant variables (e.g., temperature, income, co-pollutants, co-morbidities, etc.) that also affect the response in an exposure-response curve?

Without answers to these questions, it is (or should be) hard for the recipients of an exposure-response curve to guess what it might mean or how, if at all, it should be used to inform risk management decisions.

The following sections seek to clarify these ambiguities and present methods for resolving them. To do so, they draw on concepts and methods that have been developed in causal artificial intelligence (CAI) and machine learning (ML) to support reasoning about how some variables depend on others and how changing some variables causes the probability distributions of others to change.

[*] Corresponding author.
*E-mail address:* tony@cox-associates.com.

https://doi.org/10.1016/j.gloepi.2023.100114
Received 6 May 2023; Received in revised form 7 June 2023; Accepted 12 June 2023
Available online 15 June 2023

## Exposure-response regression curves describe responses at different observed exposures

A curve showing the conditional expected value of observed responses for each observed level of an exposure variable is called a *regression curve*. Regression curves can be estimated from observational data on exposures, response rates, and covariates by well-developed techniques. These include regression models relating exposure to continuous or discrete response variables; non-parametric smoothing regression models (e.g., spline or LOESS curves) if appropriate parametric regression models are unknown; and proportional hazards and other survival data analysis methods if the response variable is the time until an event such as death or diagnosis with a disease [27]. Regression curves are widely treated as if they had a causal interpretation: that reducing exposure would reduce the risk of response as described by the regression curve. However, in general, a regression curve does not predict the interventional causal effects caused by interventions that change exposure, but rather it describes the average responses for different observed levels of exposure under the conditions for which data were collected (undisturbed by potential future interventions) [7,12,22,31].

### *A point of departure: Correlation vs. causality*

The interpretation of a regression-based exposure-response curve is usually straightforward: it shows the model-predicted average level of the response variable in a population for each level of the exposure variable. All such plots involve assumptions.These range from the assumption that definitions and measurements of variable are stable enough to make plots informative to the assumption that patterns relating values of variables in past data will persist in future data. Visual interpretations of plots are often subjective, as is the decision of what to plot when there are many variables. Exposure-response curves are most easily interpreted if the population consists of individuals with accurate individual-level exposure, covariate, and response data, but this ideal level of granularity is seldom available. Regression-based exposure-response data can also be developed from other types of data such as time series showing daily deaths in an exposed population.

The limitations of regression models are equally straightforward. A regression model does not explain *why* average population response rates are different for different levels of exposure. It does not explain or predict how or whether average population response rates would *change* in response to an intervention that changes exposure levels [22]. It is common practice to misinterpret the slope of a regression curve as describing the change in average response that would be caused per unit of change in exposure, but even under ideal conditions (correctly specified model, no errors-in-variables), the slope only describes the different average response rates *observed* at different levels of exposure. As demonstrated by Simpson's Paradox, this information about *differences* in observed response rates for different levels of exposure does not necessarily have any implications for how, if at all, changing exposure would *change* response rates (ibid). For example, average daily consumption of baby aspirin among people over 65 years old may be significantly positively associated with increased heart attack risk (because people at higher risk of heart disease, on doctor's orders, might increase their consumption of baby aspirin), and yet reducing such consumption might increase heart attack risk, so that change in exposure is negatively associated with change in heart attack risk.

It is widely appreciated among epidemiology theorists and statisticians that regression coefficients do not distinguish between correlation and causation or between direct and indirect (mediated) causal effects of some variables on others. For example, suppose that eq. (1) describes an observed exposure-response relationship (on some appropriate exposure and response scales) for a population.

$$Response = 100 + 0.5^{*}Exposure \tag{1}$$

How would reducing *Exposure* from an initial level of 100 to a final level of 0 change *Response*, assuming that no other relevant variables change? A common misinterpretation of exposure-response models among some practitioners is that regression eq. 1 implies that reducing *Exposure* from 100 to 0 would reduce *Response* from 150 to 100. In reality, model 1 has no implications for how or whether changing *Exposure* would change *Response*. For example, suppose that (perhaps unknown to the analyst) the underlying causal model relating past values of *Exposure* and *Response* is described by the pair of structural eqs. 2a and 2b with initial values of 100 for *Exposure* and *Poverty*. (The causal interpretation of a structural equation model is that changing an independent variable on its right side causes the dependent variable on the left to adjust to restore equality.) Eqs. 2a and 2b are chosen so that together they imply eq. 1. Then exogenously reducing *Exposure* from 100 to 0 (without changing *Poverty*, so that the historical relationship in 2a is superseded by the intervention) would *increase* risk (meaning the expected value of *Response* in the exposed population in eq. 2b) from 150 to 200.

$$Exposure = 1^{*}Poverty \tag{2a}$$

$$Response = 100 + Poverty - 0.5^{*}Exposure \tag{2b}$$

By contrast, if the underlying causal model has structural eq. 3 in place of eq. 2b, then reducing *Exposure* from 100 to 0 without changing *Poverty* would indeed *decrease* risk from 150 to 100:

$$Response = 100 + 0^{*}Poverty + 0.5^{*}Exposure \tag{3}$$

Finally, if it has eq. 4 instead of eq. 2b, then reducing *Exposure* from 100 to 0 without changing *Poverty* would have no effect on risk.

$$Response = 100 + 0.5^{*}Poverty + 0^{*}Exposure \tag{4}$$

Model 1 does not reveal which, if any, of these (or other observationally equivalent) underlying causal models is correct. No matter how well it fits past data, an empirical regression-based exposure-response model such as model 1 cannot provide a sound basis for predicting how or whether reducing *Exposure* would change *Response* [12,22,31]. The same is true for more sophisticated models; for example, as noted by Martinussen [20] for Cox proportional-hazards models, "the Cox hazard ratio is not causally interpretable as a hazard ratio unless there is no treatment effect or an untestable and unrealistic assumption holds."

### *Assumption-dependent causal interpretations of exposure-response regression models*

In response to such limitations, many investigators have proposed using simplifying assumptions to interpret regression coefficients causally. For example, it is often convenient to assume that there are no unmeasured confounders (such as *Poverty* in the preceding example). Likewise, it is often convenient to assume that any observed differences in response rates between two populations, or between response rates at two different times for the same population, or between the rates of changes in response rates in a more-exposed population compared to a less-exposed population, are caused solely by differences in exposures. A major limitation of such assumption-based causal interpretations is that the simplifying assumptions used to draw causal conclusions often prove to be mistaken [7,31]. Assuming that an empirical exposure-response model such as eq. 1 that describes observations can also be interpreted as a causal structural model such as eq. 3 to predict the effects of interventions simply assumes away, without resolving,the basic distinction between associational and causal relationships [22]. It does not overcome the fundamental methodological challenge that *differences* are not *changes* and that *seeing* is not *doing* (ibid). In practice, treating regression models and other associational models such as Global Burden of Disease (GBD) or population attributable fraction (PAF) models as if they were causal models that can be used to predict effects on responses of changing exposures sometimes leads to failed predictions and unsound policy advice, such as that a 70% reduction in particulate pollution in Ireland would cause detectable reductions in all-cause mortality rates [31] or that increasing consumption of certain vitamins would reduce lung cancer risk instead of increasing it [16].

### *Heterogeneity in individual risks*

Regression-based exposure-response curves describe *average* responses for different levels of exposure in a population rather than describing *distributions* of individual response risks. For example, suppose that an exposure-response model shows that, under current exposure conditions, there is a 10% probability of disease over a certain time interval. Does this mean that each individual has a 10% probability of disease, or that 10% of the population has a 100% probability of disease and the rest 0%, or that half the population has a 20% probability and the other half has zero, or something else? The exposure-response curve does not give an answer. Uncertainty bands around it are typically for the estimated *average* response at each level; they do not describe the *distribution* of individual risks around this average. Yet, the answer may matter to policymakers and the public. A smaller risk that is more broadly distributed in the population may raise different levels of concern and support for regulation than a larger risk focused in a smaller subset of the exposed population, especially if the subset is identifiable e.g., only active smokers, or only children with asthma, or only elderly people with COPD. Exposure-response curves omit such information. The limitations of using averages for decision-making are well documented [26]. Extending exposure-response curves to provide information about *distributions* of individual risks may be essential for well-informed decision-making.

### *Ambiguous regression coefficients: Inference vs. intervention*

The sign and magnitude of the slope of an exposure-response regression coefficient often depend on an investigator's selection of independent variables to include in the model [12]. For example, suppose that structural eqs. 5a and 5b describe how exposure, poverty, and response are related. Eq. 5a implies that *Poverty* = 2**Exposure*. Substituting 2**Exposure* for *Poverty* in eq. 5b yields eq. 5c. (Rewriting eq. (5a) as *Poverty* = 2**Exposure* would make for a simpler exposition, but the form *Exposure* = 0.5**Poverty* preserves the structural equation (causal) interpretation that causality flows from right to left, i.e., changing the right-hand side variable *Poverty* would change the left-hand side variable *Exposure*, but changing *Exposure* would not change *Poverty*.) Eq. 5c is the regression equation relating *Response* to *Exposure* when only those two variables are measured, i.e., when *Poverty* is an unmeasured (latent) variable or is simply not included in the model.

$$Exposure = 0.5^{*}Poverty \tag{5a}$$

$$Response = 100 + 0.5^{*}Poverty - 0.5^{*}Exposure \tag{5b}$$

$$Response = 100 + 0.5^{*}Exposure \tag{5c}$$

Thus, when only *Exposure* is included as an independent variable on the right side of the regression model as a predictor of *Response*, its regression coefficent is 0.5, and hence is positive (eq. 5c). But if both *Poverty* and *Exposure* are included as independent variables on the right side of the regression model for *Response*, the regression coefficent for *Exposure* becomes −0.5, and hence is negative (eq. 5b). This illustrates that the sign of a regression coefficient can depend on the selection of independent variables. The interpretation of the positive regression coefficient 0.5 for *Exposure* as a predictor of *Response* in eq. 5c is that when *Exposure* is observed to have a higher value, one can infer (via eq. 5a) that *Poverty* has a higher value, and can therefore predict (via eq. 5b) that *Response* will have a higher value. However, this inference has no implications for predicting the effects of an intervention that changes *Exposure* (without changing *Poverty*). Such an intervention acts through eq. 5b. It renders eq. 5a irrelevant. If *Exposure* is set to a new level by an exogenous intervention, it is no longer determined endogenously via eq. 5a, which can therefore be excised from the analysis – a form of "graph surgery" used to model the effects of interventions [18].

The regression coefficient for exposure in an exposure-response regression model such as eq. 1 typically reflects a mixture of inference and intervention effects and its sign and magnitude can depend on a modeler's choices about which other independent variables to include in the model (such as *Poverty* in this example) [12]. These considerations imply that regression coefficients are not suitable in general for predicting how or whether an intervention that changes exposure will affect the risk of response [20]. Regression models focus on creating curves that optimize measures of departure of data from a model rather than predicting the effects of potential future changes in exposure. This distinction is well-recognized in machine learning, which assesses the predictive accuracy of models using a train-and-test paradigm (e.g., using cross-validation rather than traditional statistical fit criteria such as AIC, BIC, or adjusted R-squared values) [17,27]. The desire to predict the effects of interventions has led to recent innovations in heterogeneous treatment effect (HTE) analysis, including causal tree and causal forest algorithms that modify standard machine learning (ML) algorithms to predict changes in risk due to different exposures, rather than estimating levels or differences in risk between differently exposed individuals [3,5,15,17]. The following sections explore how ML ideas can be applied to create exposure-response curves that have valid causal interpretations.

## Logistic regression vs. non-parametric exposure-responsecurves

To fix ideas concretely, we use a dataset in which mortality risk increases with exposure. To keep the discussion general and conceptual, we refer to the exposure variable simply as "exposure" without specifying units. (For the curious reader, however, the example dataset is for blood lead levels measured in μg/dL and mortality during the follow-up period for male non-smokers in the NHANES III dataset. It is described further by Cox [9].) Fig. 1 shows a logistic regression model fit to the exposure-response data.

A fundamental challenge in interpreting such a curve is the

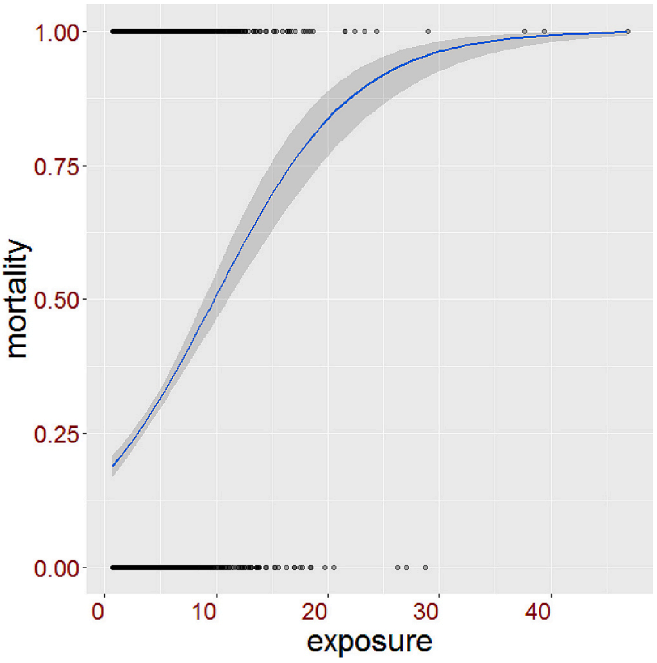

**Fig. 1.** A logistic regression curve for mortality vs. exposure.

possibility of confounding: perhaps people with higher levels of exposure tend to be older, or less educated, or to have lower incomes than people with lower levels of exposure, and perhaps such differences contribute to the differences in mortality risk estimated for different levels of exposure. (Indeed, all of these associations hold in this example.) A second fundamental challenge is that the best-fitting model in a specific parametric class of models (e.g., logistic regression) may not describe the data very accurately.

## Controlling for observed confounders with partial dependence plots (PDPs)

An *empirical* exposure-response curve shows the average response rates observed at different levels of exposure (presumably in different populations or subpopulations or at different times) without superimposing a model-based curve (such as the logistic regression curve in Fig. 1). If exposure levels are known and discrete and data is abundant enough to allow accurate estimates of response rates for each exposure level, then such a curve has the advantage over a regression model that it does not depend on modeling assumptions. It has the disadvantage that it does not control for the levels of other variables that may have different values for people with different exposures. This is unacceptable if we want to see how exposure alone affects mortality risk while holding other variables fixed. For example, Fig. 2 plots mean age against exposure (rounded to the nearest integer). It is clear that exposure is positively associated with age. A similar plot shows that exposure is also positively associated with mortality risk. To isolate the effect of exposure alone on mortality risk, therefore, it is necessary to control for potential confounding by age (and also by other confounders such as income and education, both of which are negatively associated with exposure and mortality risk). We would like to control for the levels of these other variables without making potentially invalid modeling assumptions such as that they affect mortality risk through a multivariate logistic regression function.

A practical solution is to use nonparametric ML methods to predict the average value of mortality probability as exposure alone changes, holding all other variables fixed at their current levels for each individual in the data set. This yields a *partial dependence plot* (PDP) [30]. Fig. 3 shows the PDP exposure-response curve for our example data using the NHANES III data for male non-smokers. The interpretation is that each point on this curve shows the predicted average value of the 0–1 *mortality* dependent variable (i.e., the average conditional probability of mortality during the follow-up period) that the individuals in the data set would have if exposure were set to each value on the x-axis and all other variables were kept at the values that they currently have. In this setting, the basic distinction between predicted *differences* and predicted *changes* in outcomes as exposure varies disappears if it can be assumed that the only explanation for differences in predicted mortality is corresponding differences in exposure (since all other independent variables are held fixed). This does not solve the problem of unobserved confounders, however. For example, if the exposure variable (blood lead level) indicates unreported smoking in this allegedly nonsmoking population, then the apparent effect of exposure on mortality risk may actually be due to the hidden confounder of smoking.

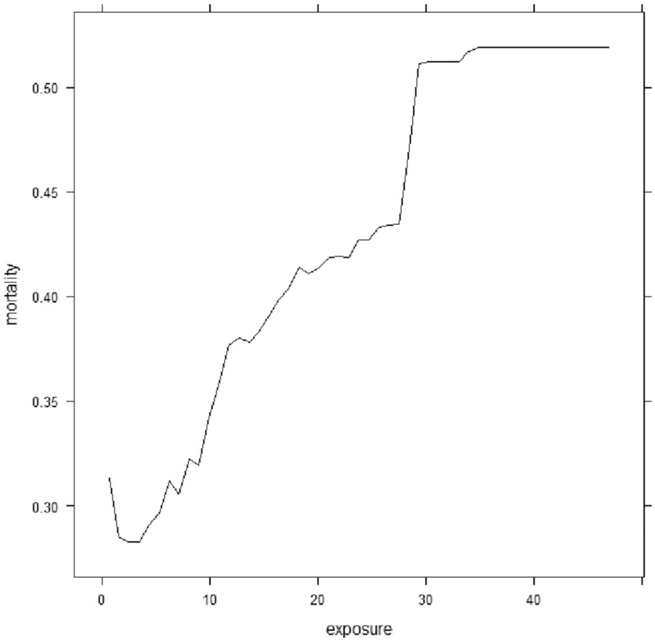

**Fig. 3.** Partial Dependence Plot (PDP) of mortality probability vs. exposure for male nonsmokers.

To create a PDP, we need an ML algorithm (or "model" in ML

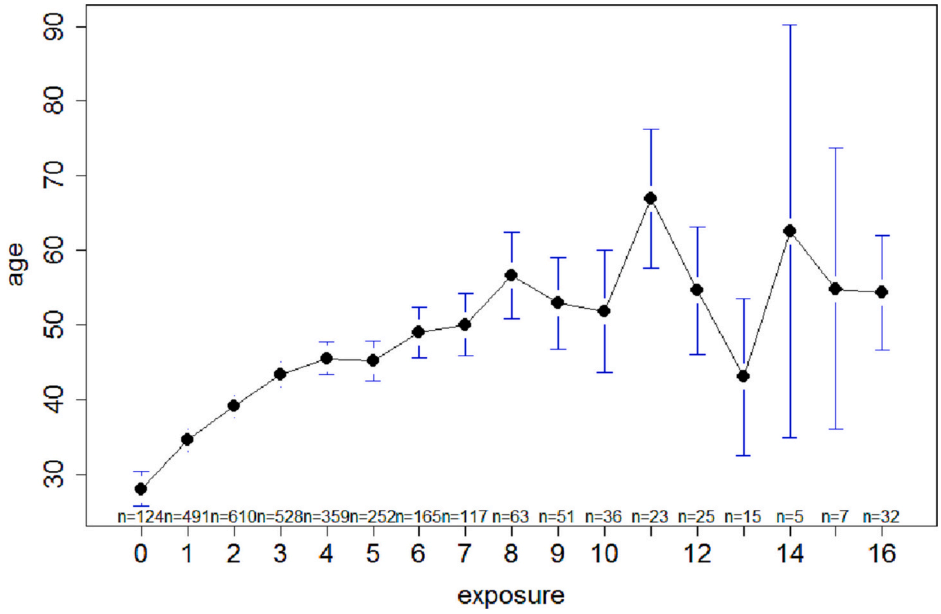

**Fig. 2.** Plot of mean age vs. exposure for the data in Fig. 1. (Exposures >15 are rounded to 16 to save space.)

terminology) that predicts the value of the dependent variable (here, *mortality*, coded as 1 for mortality and 0 otherwise) for each combination of values of the independent variables. There are many such algorithms in the modern ML toolkit. Fig. 3 uses the popular Random Forest algorithm [11] to predict average mortality risk for different values of exposure, given the values of the following covariates: *age*, *income*, *grade*, *married*, *Hispanic*, *Black*, *small.metro*, and *West*. Here, *married, Hispanic*, *Black*, *small.metro*, and *West* are 0–1 binary indicator variables (i.e., dummy variables) indicating whether each case is identified as married, Hispanic, Black, living in a small metropolitan area, and living in the Western US, respectively. (Sex and smoking status are excluded since we selected non-smoking men as the cases for which an exposure-response curve is to be estimated.) *Grade* indicates the highest grade completed.

The PDP is constructed as follows. Begin with a dataset organized so that each row contains the data for one individual (also called a "case" or "record" or "observation" or "data point") and each column contains the value of one variable (also called a "feature" or "attribute" or "field"). This layout of the data is called a *data frame*. Thus, the value in row *r* and column *c* is the value for individual *r* of variable *c* in the data frame. For each case, the values of the exposure variable (called *exposure* in our example), the response variable (the 0–1 *mortality* indicator in our example), and measured covariates (*age*, *income*, *grade*, *married*, *Hispanic*, *Black*, *small.metro*, and *West* in our example) are recorded for each case. Next, create a succession of modifications of the original data frame by changing just one column, the *exposure* variable, while leaving all other columns unchanged. The successive modified datasets contain successively increasing values of the exposure variable (which is set to have the same value for all cases in each modified dataset), beginning with the smallest value of *exposure* observed in the original dataset and increasing to the largest value. Each of these modified datasets generates one point on the PDP as follows: the point corresponding to a specific modified dataset plots the average predicted value of the response variable, *mortality* (averaged over all cases in the modified dataset) against the value of *exposure* for the modified dataset. Joining these points with line segments completes the PDP. The PDP can be viewed as generalizing the concept of an *average treatment effect* (ATE) for a binary exposure variable to continuous exposure variables [30].

To illustrate this process, Table 1 shows a simple hypothetical dataset with only 3 records (rows). For purposes of a simple exposition only, suppose that the probability of mortality is predicted from other variables using the linear regression model.

$$E(mortality \mid Exposure, age, income) = -0.20 + 0.003^{*}Exposure + 0.017^{*}age - 0.02^{*}income$$

(In practice, of course, linear regression is not appropriate for a binary dependent variable, but we use it to make the arithmetic easy to follow and verify. For all calculations shown later, we use random forest as a more flexible and realistic predictive model and include additional significant predictors of mortality risk such as *married* and *grade*. Readers unfamiliar with random forest can find an accessible introduction in Molnar [21]. Briefly, it is an ensemble method that averages predictions from several hundred non-parametric classification and regression (CART) trees.) Table 2 is identical to Table 1 except that the first column shows the model-predicted mortality probabilities instead of the 0–1 indicators of observed mortality; thus, for the first case we have.

$$E(mortality \mid Exposure = 5.0, age = 21, income = 0.6) \\ = -0.20 + 0.003^{*}5 + 0.017^{*}21 – 0.02^{*}0.6 = 0.16$$

Tables 3 and 4 show successive modified datasets with *Exposure* set to the smallest and second-smallest observed exposure values in the NHANES data, 0.7 and 1.0, respectively. The corresponding predicted mortality probabilities (i.e., predicted conditional expected values for *Mortality*, given the values of the independent variables) are slightly different for these two consecutive values of *Exposure*. Averaging the predicted mortality probabilities for all individuals for each level of *Exposure* gives the corresponding height of the PDP for that exposure level. Thus, for *Exposure* = 0.7, the PDP predicts an average mortality probability in the exposed population (with all individuals having *Exposure* = 0.7) of (0.147 + 0.584 + 0.863)/3 = 0.531; while for Exposure = 1.0, the PDP predicts a slightly higher average mortality probability of (0.148 + 0.585 + 0.864)/3 = 0.532. Continuing to increase the *Exposure* values and calculate the resulting average predicted mortality probabilities generates the entire PDP curve. The PDP curve in Fig. 3 is generated by this process using all 2903 cases in the NHANES dataset for lead-exposed non-smoking males and using random forest to predict conditional expected values of *Mortality* for each set of independent variable values.

The PDP in Fig. 3 solves the problem of controlling for observed potential confounders such as age, income, and grade by holding their values fixed: only *exposure* and resulting predicted *mortality* values vary. (It also avoids creating spurious dependencies between variables due to collider bias (ibid) as long as *mortality* is a possible effect, but not a possible cause, of the other variables.) The PDP reduces dependence on modeling assumptions by using non-parametric prediction methods, such as the Random Forest algorithm used to create Fig. 3. Thus, a PDP may provide a useful definition for an exposure-response curve.

## Describing interindividual heterogeneity in exposure-response functions: Individual conditional expectation (ICE) plots

The challenge of characterizing interindividual differences in exposure-response curves remains to be addressed. The sequence of modified datasets used to construct the PDP can be repurposed to provide a constructive solution to this challenge. Instead of averaging the predicted values of the dependent variable, *mortality*, for all individuals in each modified dataset to show how *average* mortality probability depends on exposure when all other variables are held fixed, as in a PDP plot, an alternative is to display the predicted values of *mortality* for each *individual* for the different levels of exposure. This yields an *individual conditional expectation (ICE) plot* [14,30], so called because the predicted values of mortality are its conditional expected values (as predicted by an ML model such as Random Forest) given the values of each individual's covariates, for different assumed levels of exposure. Fig. 4 shows an ICE plot for the same data used to create the PDP in Fig. 3. Each black curve is the estimated exposure-response curve for one individual; these are too numerous and too dense to see clearly, but their spread shows the wide range of inter-individual variability in exposure-response curves. Averaging these curves gives the PDP curve (indicated in red in Fig. 4) for the population of individuals.

The ICE plot shows the range of variability in individual response probabilities at different levels of exposure, but it can be quite difficult to read when there are many individuals (e.g., there are 2903 individual

**Table 1**
Original dataset.

| Mortality | Exposure | age | income | grade | married | Hispanic | Black | small.metro | West |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 5.0 | 21 | 0.6 | 12 | 0 | 1 | 0 | 0 | 1 |
| 0 | 7.3 | 50 | 3.4 | 12 | 1 | 1 | 0 | 0 | 1 |
| 1 | 9.3 | 65 | 2.2 | 3 | 1 | 1 | 0 | 0 | 1 |

**Table 2**
Predicted mortality probabilities for original dataset.

| Predicted Mortality | Exposure | age | income | grade | married | Hispanic | Black | small.metro | West |
|---|---|---|---|---|---|---|---|---|---|
| 0.16 | 5.0 | 21 | 0.6 | 12 | 0 | 1 | 0 | 0 | 1 |
| 0.60 | 7.3 | 50 | 3.4 | 12 | 1 | 1 | 0 | 0 | 1 |
| 0.89 | 9.3 | 65 | 2.2 | 3 | 1 | 1 | 0 | 0 | 1 |

**Table 3**
Predicted mortality probabilities for modified dataset with *Exposure* = 0.7.

| Predicted Mortality | Exposure | age | income | grade | married | Hispanic | Black | small.metro | West |
|---|---|---|---|---|---|---|---|---|---|
| 0.147 | 0.7 | 21 | 0.6 | 12 | 0 | 1 | 0 | 0 | 1 |
| 0.584 | 0.7 | 50 | 3.4 | 12 | 1 | 1 | 0 | 0 | 1 |
| 0.863 | 0.7 | 65 | 2.2 | 3 | 1 | 1 | 0 | 0 | 1 |

**Table 4**
Predicted mortality probabilities for modified dataset with *Exposure* = 1.0.

| Predicted Mortality | Exposure | age | income | grade | married | Hispanic | Black | small.metro | West |
|---|---|---|---|---|---|---|---|---|---|
| 0.148 | 1.0 | 21 | 0.6 | 12 | 0 | 1 | 0 | 0 | 1 |
| 0.585 | 1.0 | 50 | 3.4 | 12 | 1 | 1 | 0 | 0 | 1 |
| 0.864 | 1.0 | 65 | 2.2 | 3 | 1 | 1 | 0 | 0 | 1 |

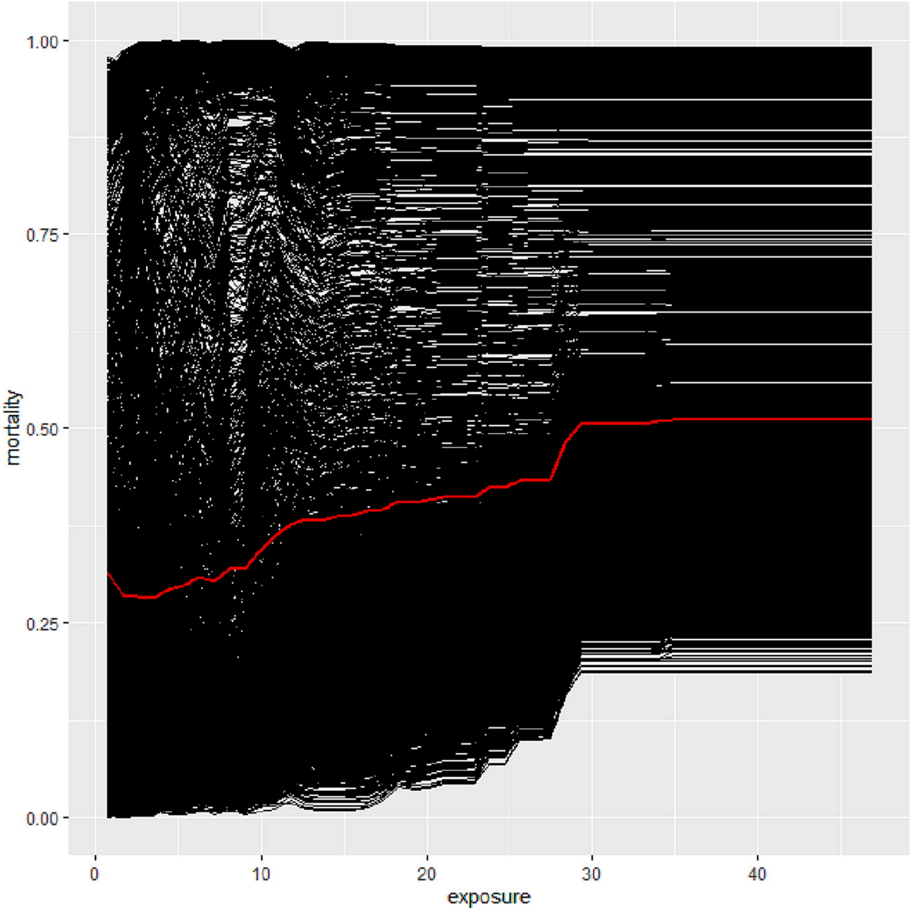

**Fig. 4.** ICE Plot for mortality probability vs. exposure for male nonsmokers.

curves in Fig. 4). To improve legibility, it is common practice to center all curves at a common value, such as the lowest value of exposure in the dataset, and to show only the differences in predicted values of mortality at each exposure level compared to their predicted values at this lowest (baseline) level of exposure. Fig. 5 shows such a *centered ICE plot* [14,21].

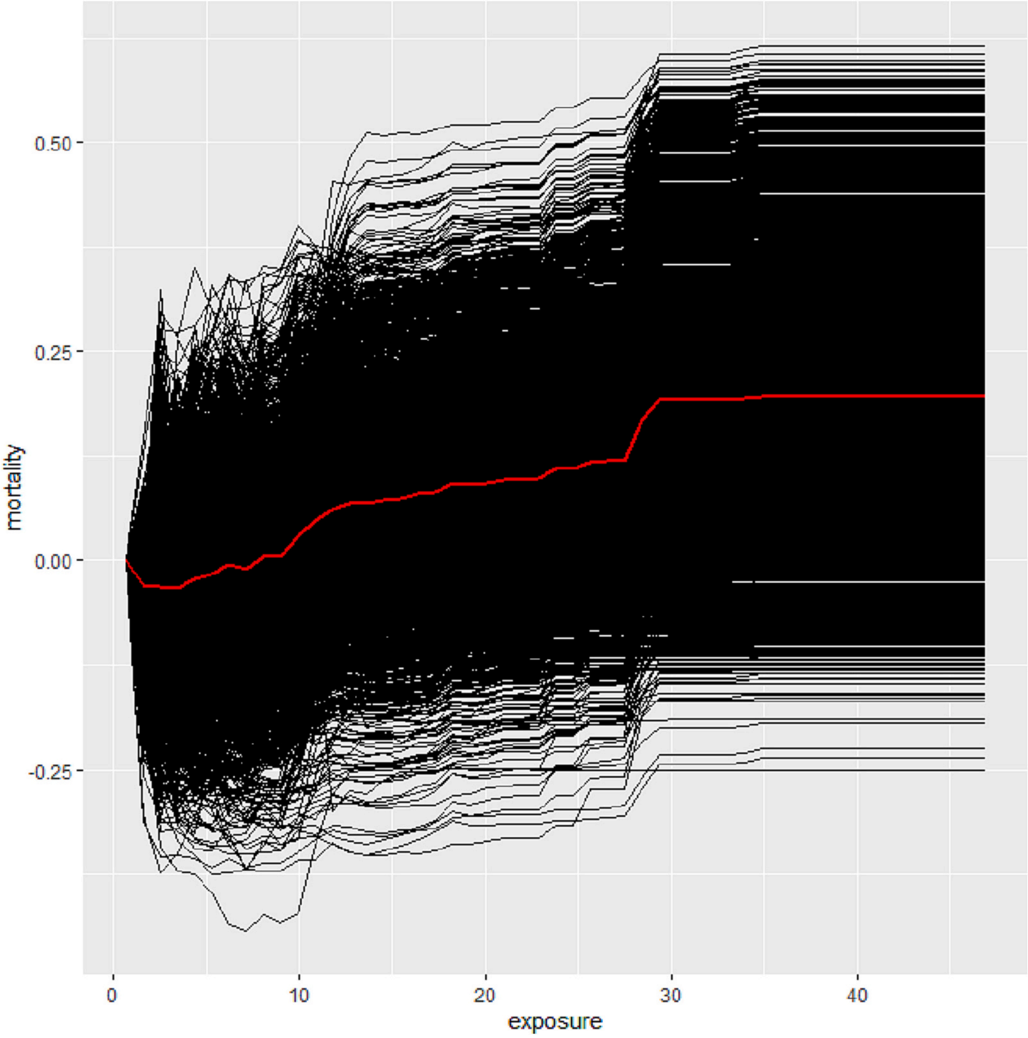

**Fig. 5.** Centered ICE Plot for mortality probability vs. exposure for male nonsmokers. The curves show the differences in predicted individual mortality probabilities for different levels of exposure compared to their predicted values for the lowest level of exposure.

The centered ICE plot shows that even though the average predicted mortality probability in the population of cases increases with exposure, individual predicted mortality probabilities vary with exposure in a variety of ways.

To further simplify and abstract from the wealth of detail in the ICE plot in Fig. 5, Fig. 6 uses k-means clustering to cluster the centered ICE curves into three clusters. The proportions of cases in each cluster are shown in the legend. (The y-axis (labeled “cluster yhat”) is a rescaled version of the differences in mortality probabilities shown on the y-axis in Fig. 5, expressed in terms of standard deviations, as is usual for k-means clustering; see Goldstein et al. [13] for details.) The clusters in Fig. 6 suggest that about 23% of the individuals in the dataset (green cluster) have individual exposure-response curves that increase with exposure over the range from 0 to about 10, while 70% have no significant response (red cluster) and the remaining cases (shown as 6% due to rounding) have smaller mortality risks at higher exposure levels (blue cluster).

What distinguishes members of the different clusters is not necessarily simple to identify and summarize when ICE curves are very variable over a wide range, as in Fig. 5, but further exploration shows that exposure is significantly positively ordinally correlated with predicted mortality risk for people with above-median age (Kendall’s tau = 0.17, *p*-value = 1.97E-14) but not for younger people (Kendall’s tau = 0.015, p-value = 0.49), especially for the youngest individuals with below-median income (Kendall’s tau = −0.056, p-value = 0.15 for people under 25, corresponding to the first quartile of the age distribution of cases with below-median incomes). However, any such clustering simplifies the complex reality shown in Fig. 5, which is that different individuals have very different exposure-response functions, depending on the levels of their other covariates.

## Data-informed counterfactuals: Two-dimensional partial dependence plots (2D-PDPs)

The construction of PDPs and ICE plots involves (i) varying one independent variable, exposure, while holding all others fixed at their current levels; and (ii) plotting how predicted average (for PDPs) or individual (for ICE plots) values of the dependent variable vary in response to the different levels of exposure, given the fixed levels of other covariates for the individual cases in the population for which the PDP or ICE plot is derived. These concepts can be extended to allow more than one independent variable to be varied while holding the rest fixed. In the *two-dimensional PDP* in Fig. 7, both *age* (y-axis) and *exposure* are systematically varied from their lowest to their highest observed values, holding all other variables fixed at their current value for each case. The plot shows the predicted value of *mortality* risk (yellow = high, blue = low) for each pair of *exposure* and *age* values. The white spaces to the upper right and lower right of the colored region correspond to

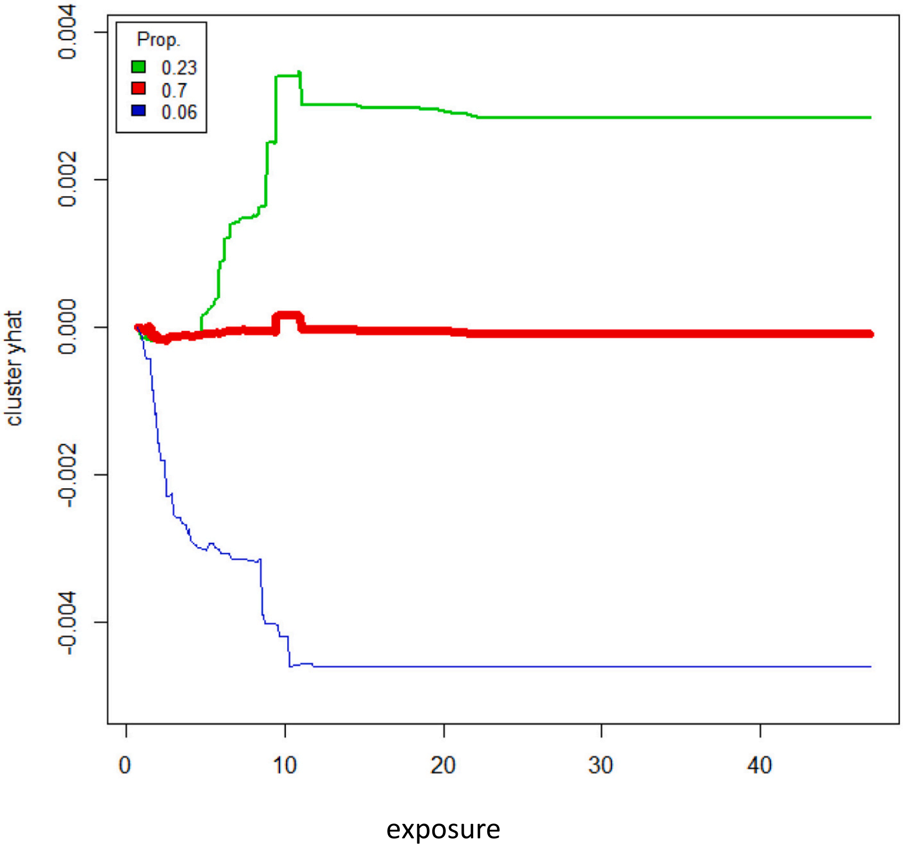

**Fig. 6.** Clustered ICE Plot showing 3 clusters of individual exposure-response curves.

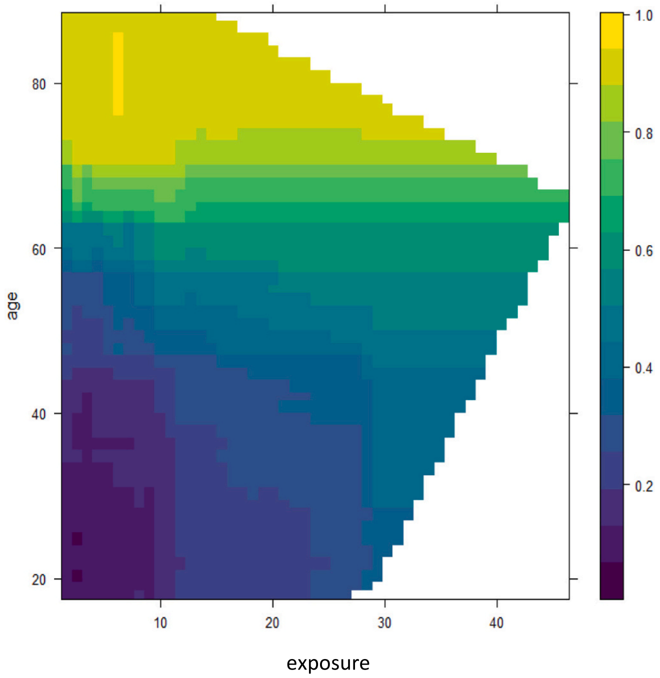

**Fig. 7.** 2D-PDP for *mortality* risk (false color scale, yellow = high, blue = low) vs. *age* and *exposure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)*

combinations of *age* and *exposure* that do not occur in this data set, and hence fall outside the colored data cloud region. For example, no one younger than 40 or older than 70 years old has *exposure* > 40.

Fig. 7 shows that the effect of *exposure* on *mortality* risk depends on *age*. For peple with *age* >70 years old, exposure has little or no relevance to mortality risk: the risk is high (yellow) for all levels of exposure, and is not higher at higher exposure levels. On the other hand, for people with *age* <36 (the median age), mortality risk, although relatively low, increases significantly with *exposure*, from <0.1 on the lower left (dark blue) to 0.3 or more on the lower right of the data cloud. This illustrates that the shape of an exposure-response PDP depends on the distribution of *age* (and other covariates) in the population for which the PDP is calculated.

To apply a PDP estimated for one population to a different population, the predicted individual-level risks (conditioned on values of covariates) that are averaged to form the PDP must be re-weighted to reflect the joint frequency distribution of covariates in the new population. This makes the interpretation of "integrated exposure-response functions" (e.g., [4]) based on data from multiple populations (often developed via international collaborations) problematic, as there is no population for which the resulting curve holds. Rather, different parts of the curve come from different populations in different countries, with different distributions of age, income, and other covariates.

Fig. 8 shows a 2D-PDP for the joint effects of *exposure* and *income* on mortality risk, other independent variables being held fixed. The highest risks occur at the lower right, for individuals with high exposure and low income. The white space in the upper right shows that the dataset contains no individuals who have both high income and high exposure. This implies that counterfactual predictions about what mortality risks would be for high-income, high-exposure individuals are not relevant to the observed real-world cases and are not supported by the data. This illustrates a difficulty with the key concept of "holding other variables

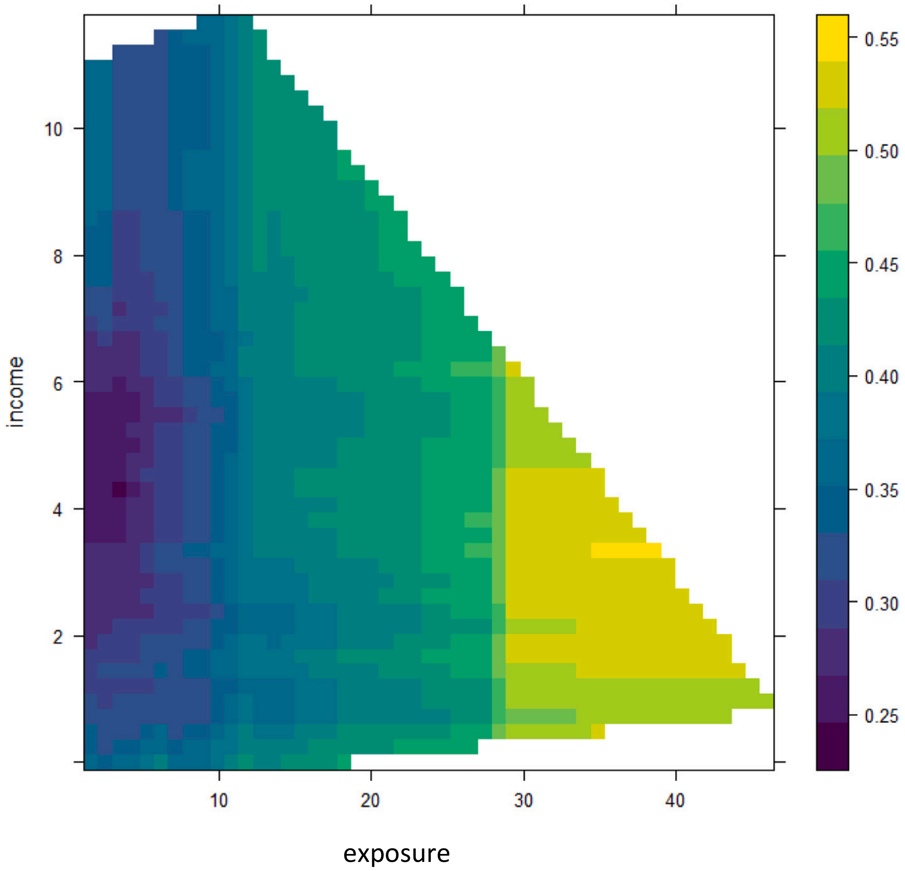

**Fig. 8.** 2D-PDP for *mortality* risk (false color scale, yellow = high, blue = low) vs. *income*and *exposure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)*

fixed" as exposure is varied: doing so may generate unrealistic counterfactual cases that dilute the realism and practical relevance of the resulting exposure-response PDP curves. For example, in Fig. 8, it is true that all individuals with *exposure* > 35 have relatively high mortality risks, but they also all have relatively low incomes. A PDP curve that includes mortality risks estimated for high-income individuals in its calculation of average mortality risk as a function of exposure therefore includes extrapolated risks for counterfactual scenarios of high-income, high-exposure individuals for which there are no corresponding observations in the data cloud. It is not clear that such extrapolations are correct or relevant for real populations. Hence, it is not clear that they should be included in calculating exposure-response curves that are meant to inform risk management decisions for real populations.

Two refinements of the PDP concept have been proposed to avoid including unrealistic or irrelevant counterfactual cases in the calculation of exposure-response functions. *Accumulated local effects* (ALE) plots [1] work entirely within the observed data cloud. They use observed combinations of variable values for real cases to estimate the change in average response for small changes around each level of exposure, holding other variables fixed. They do so by using only the subset of cases with similar levels of exposure to estimate the local change in risk from a small change in exposure around a given level. For example, in Fig. 8, only cases with exposures in the neighborhood of 35 would be used to estimate how mortality probability changes as exposure increases from 34 to 35 or from 35 to 36. This automatically selects cases with relevant (relatively low) incomes and excludes irrelevant counterfactual cases of individuals with high incomes and exposures in this range. The second approach is to drop the requirement that all other variables be held fixed as exposure changes. For example, causal graph models can be used to identify *adjustment sets* [29] of variables that must be held fixed to obtain estimates of direct or total causal effects of exposure on mortality (e.g., confounders), while other variables are left free to have their conditional distributions change realistically as exposure changes. In this approach, the conditional distribution of income could change (shifting downward) at higher levels of exposure, again avoiding the need to include unrealistic counterfactual cases in calculating the exposure-response curve.

## Discussion and conclusions: What do we *want* exposure-response curves to mean?

The previous sections have suggested that ideas and computational methods from modern causal artificial intelligence (CAI) and machine learning (ML) can help to clarify what an exposure-response curve means. These ideas and clarifications include the following:

1. A *predictive* exposure-response curve shows the conditional expected value of a response variable such as *mortality* predicted for different observed values of *exposure*, given the distribution of other independent variables (i.e., covariates) in a population.
2. A *predictive* exposure-response curve may be quite different from an *interventional* exposure-response curve that shows how the average value of the response variable would change if the values of exposure were changed [10,22].
3. The shape of a regression-based estimated predictive exposure-response curve typically reflects modeling assumptions as well as obervations. Non-parametric ML techniques including PDPs based on random forests or other ML algorithms can help to avoid the need for parametric modeling assumptions, although they still require adequate sample sizes and coverage of the combinations of exposures and other covariates for which predictions are to be made in order to produce accurate predictions. (By contrast, nonparametric *statistical*

methods often make their own strong modeling assumptions, such as that different conditional curves have the same shapes, are symmetric about their medians, etc.; such assumptions are not required for PDPs based on nonparametric ML methods.)

4. Figs. 7 and 8 illustrate (via white spaces outside the data cloud) the common occurrence that many logically possible combinations of exposure values with other covariate values do not occur in practice. Exposure-response curves should be accompanied by explanations of whether the average response probabilities they predict are based on observed combinations only (as in ALE plots) or whether they include hypothetical, perhaps unrealistic, combinations of exposure values with other covariate values that do not occur in the data (as in one-dimensional PDP plots).
5. The predictions from a (predictive or interventional) exposure-response curve depend in general on the values of other covariates for individuals in the population for which the curve is developed. Fig. 7 illustrates this point: how mortality risk depends on exposure depends on age. Therefore, averaging mortality risk over individuals for different assumed levels of exposure will produce very different population exposure-response curves depending on the frequency distribution of individual ages (and other covariates) in the population. Even if the dependent variable were age-specific hazard rate, other covariates (such as *grade*) would still modify the predictive relationship between exposure and response. Thus, any exposure-response curve is valid only for specific joint distributions of other causally relevant covariates that interact with exposure in predicting risk. These distributions are seldom stated explicitly when exposure-response curves are presented, but they are crucial parts of the definition of the curves.
6. In general, an exposure-response curve developed for one population does not apply to other populations with different distributions of covariates. (An exception occurs if there are no interactions among exposure and other covariates so that the effect of exposure can be assessed independently of age, income, education co-morbidities, co-exposure, or other covariates. But in our experience, such independence of effects seldom holds in practice.)
7. Integrated exposure-response functions (e.g., [4]) do not hold for any population. Rather, different parts are taken by averaging the predicted or observed levels of the response variable over individuals in different populations, typically with different types and levels of exposures and different distributions of covariates. Such a curve has no clear conceptual definition and does not apply to any specific population.
8. Even when the meaning of an exposure-response curve is clearly specified, e.g., as the predicted average value of response for each level of observed exposure in a specific population of individuals with known values of other covariates (assumed to remain fixed as exposure varies), the exposure-response curve is only an average of individual-level curves. This is illustrated in Figs. 5 and 6, which make explicit the inter-individual variability in exposure-response curves. The flaws of using averages in decision-making are well known [26], but how to use the more detailed information in ICE plots to improve risk management decision-making has not yet been much discussed in applied epidemiology, which still relies largely on population-level exposure-response curves.

These ideas emphasize different possible meanings for an exposure-response curve (e.g., regression model, PDP, ALE, descriptive model, predictive model, interventional model); some of their limitations; and the fact that the meanings of some exposure-response curves cannot be inferred from the curves themselves because the levels of other relevant covariates for individuals in the populations described by the curves are not specified.

Exposure-response curves are usually used to communicate information to inform health risk management decisions. It is therefore worth asking what information the users of exposure-response curves *want* or intend for them to communicate. Fig. 9 presents an exampleof why the intended meaning of such a curve requires clarification. Suppose that high concentrations of PM2.5 are found to predict both higher levels of chronic obstructive lung disease (COPD) in a population and also higher rates of lung cancer specifically among COPD patients, but not among other individuals.

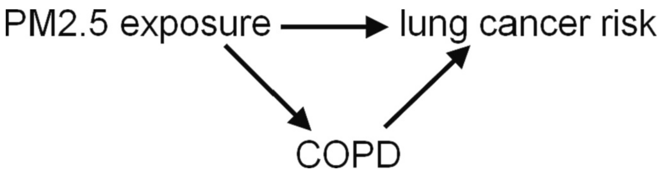

**Fig. 9.** Should levels of COPD be held fixed in quantifying the PM2.5-lung cancer risk exposure-response curve?

Now, suppose that we are asked to develop an exposure-response curve relating PM2.5 exposure to lung cancer risk to inform discussions about the health benefits of lowering PM2.5. Then we must decide what should be held fixed (and at what levels) in developing the requested exposure-response curve. Specifically, should COPD prevalence in the population be held fixed as PM2.5 exposures are varied across the horizontal axis of the exposure-response curve? In the terminology of mediation analysis, this would correspond to quantifying a controlled direct effect of PM2.5 on lung cancer risk [28]. If so, should COPD prevalence be fixed at today's levels? Doing so would maximize the estimated direct effect on lung cancers caused by reducing PM2.5 exposure by providing a relatively large pool of COPD cases for whom lung cancer risk can be reduced by reducing PM2.5. Or should COPD prevalence be fixed instead at the predicted future lower prevalence level for COPD anticipated if PM2.5 is reduced to a new, lower level? This would reduce the predicted direct benefit from reduced lung cancer risk among COPD cases while acknowledging an additional benefit from reduced COPD cases. In the terminology of mediation analysis, the *total effect* of reducing PM2.5 on lung cancer risk reflects both the *direct effect* caused by reducing lung cancer risk for COPD patients and also the *indirect effect* caused by reducing the number of COPD patients (ibid). How, if at all, should choices about which effect to present depend on how long it takes for reductions in PM2.5 to reduce COPD prevalence?

There are not correct or incorrect technical answers to these questions. Rather, they are questions about what the users of exposure-response curves want and intend them to mean – about what information they are meant to convey to recipients. Once a choice has been made about whether to model the PM2.5-mortality exposure-response curve by holding COPD prevalence fixed at current levels, or at some anticipated future level, or perhaps as changing over time in a population following a specified reduction in PM2.5, then the focus can shift to the technical challenges of quantifying the desired curve. The technical tools such as PDPs, ICE plots, ALE plots, and adjustment sets may then be deployed to quantify it. But the logically prior question of what it is that we *want* to quantify is seldom explicitly addressed in current presentations of exposure-response curves. What actually *was* quantified – what exactly the presented curves represent – is also seldom described at a level of detail that specifies the joint distribution of covariates averaged over in calculating the exposure-response curves. Yet, without such detailed information, it may be impossible to calculate how the presented exposure-response curve (or the underlying relative risks or other measures of association) must be modified to apply to a target population with a different joint distribution of variables – or even to the same population used to estimate the exposure-response curve following an intervention that reduces exposure [22]. A major contribution of the technical methods and diagrams we have discussed is to enable calculation of exposure-response curves that reflect the detailed characteristics (observed or assumed) of target populations, but they do not address which assumptions should be made about the counterfactual levels of other variables in calculating exposure-response curves.

In summary, the increasing availability of causal artificial intelligence (CAI) and machine learning algorithms is making it increasingly practical to calculate relatively sophisticated nonparametric population exposure-response curves, 2D-PDPs, and distributions of individual exposure-response curves from the kinds of data already used to estimate parametric regression exposure-response curves. Figs. 3-8 illustrate how modern CAI/ML methods can quantify not only nonparametric PDPs that control for observed confounders without requiring parametric modeling assumptions but also ICE plots that address inter-individual heterogeneity. With this increased computational capacity comes an increasing need to define clearly what we want to use it for. It is perhaps fair to suggest that the capacity to calculate various types of exposure-response curves currently outstrips clarity about what should be calculated and how the results should be used to inform and improve public health risk management decisions. The need for greater clarity on these points can perhaps be addressed by better philosophical and causal reasoning about the information required for more causally effective risk management decisions [8,19]. Meanwhile, continuing conceptual and algorithmic advances in CAI and ML (e.g., [1,13,30]) will continue to both improve our ability to quantify carefully defined exposure-response relationships and to challenge practitioners to carefully define the relationships that they want to quantify and communicate to policymakers and risk managers.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interestsor personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The author thanks George Maldonado for a close reading and very useful comments and an anonymous reviewer for suggestions that improved the final exposition and the discussion of the implications of this work. This work was supported by the author's employer, Cox Associates LLC. Cox Associates received funding in 2022-2023 from ExxonMobil Biomedical Sciences Inc. (EMBSI) to develop and apply statistical methods to quantify inter-individual variability in exposure-response relationships. Cox Associates also received funding in 2019-2022 from the US Department of Agriculture (USDA) to support research on statistical and machine learning methods of causal inference from observational data. The author thanks Saumitra Rege and Jeff Lewis of EMBSI for stimulating discussions of the challenges of quantifying risks in heterogeneous populations. The methods, examples, and applications discussed and conclusions reached are solely those of the author.

## References

[1] Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. J R I State Dent Soc 2020;82:869–1164.

[3] Bodory H, Busshoff H, Lechner M. High resolution treatment effects estimation: uncovering effect heterogeneities with the modified causal forest. Entropy. 2022 Jul 28;24(8):1039. https://doi.org/10.3390/e24081039. PMID: 36010703; PMCID: PMC9407165.

[4] Burnett RT, Pope 3rd CA, Ezzati M, Olives C, Lim SS, Mehta S, et al. An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. Environ Health Perspect 2014 Apr;122(4): 397–403. https://doi.org/10.1289/ehp.1307049.

[5] Cáceres A, González JR. Teff: estimation of treatment EFFects on transcriptomic data using causal random forest. Bioinformatics. 2022 May 26;38(11):3124–5. https://doi.org/10.1093/bioinformatics/btac269 [PMID: 35426914].

[7] Carone M, Dominici F, Sheppard L. In pursuit of evidence in air pollution epidemiology: the role of causally driven data science. Epidemiology 2020 Jan;31 (1):1–6. https://doi.org/10.1097/EDE.0000000000001090 [PMID: 31430263; PMCID: PMC6889002].

[8] Cox LA. Toward practical causal epidemiology. Glob Epidemiol 2021;(3). https://doi.org/10.1016/j.gloepi.2021.100065. Nov.

[9] Cox Jr LA. Using Bayesian networks to clarify interpretation of exposure-response regression coefficients: blood lead-mortality association as an example. Crit Rev Toxicol 2020 Aug;50(7):539–50. https://doi.org/10.1080/10408444.2020.1787329 [PMID: 32903110].

[10] Cox Jr LAT. Do causal concentration-response functions exist? A critical review of associational and causal relations between fine particulate matter and mortality. Crit Rev Toxicol 2017 Aug;47(7):603–31. https://doi.org/10.1080/10408444.2017.1311838 [PMID: 28657395].

[11] Denisko D, Hoffman MM. Classification and interaction in random forests. Proc Natl Acad Sci U S A 2018 Feb 20;115(8):1690–2. https://doi.org/10.1073/pnas.1800256115. PMID: 29440440; PMCID: PMC5828645.

[12] Dominici F, Greenstone M, Sunstein CR. Science and regulation. Particulate matter matters. Science 2014;344(6181):257–9. https://doi.org/10.1126/science.1247348.

[13] Goldstein A, Kapelner A, Bleich J, Pitkin E. Package ICEbox. https://cran.r-project.org/web/packages/ICEbox/ICEbox.pdf; 2022.

[14] Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 2015;24(1):44–65. https://doi.org/10.1080/10618600.2014.907095.

[15] Gong X, Hu M, Basu M, Zhao L. Heterogeneous treatment effect analysis based on machine-learning methodology. CPT Pharmacometrics Syst Pharmacol 2021 Nov; 10(11):1433–43. https://doi.org/10.1002/psp4.12715. Epub 2021 Oct 30. PMID: 34716669; PMCID: PMC8592515.

[16] Goodman GE, Thornquist MD, Balmes J, Cullen MR, Meyskens Jr FL, Omenn GS, et al. The Beta-carotene and retinol efficacy trial: incidence of lung cancer and cardiovascular disease mortality during 6-year follow-up after stopping beta-carotene and retinol supplements. J Natl Cancer Inst 2004;96(23):1743–50. https://doi.org/10.1093/jnci/djh320. Dec 1. PMID: 15572756.

[17] Huntington-Klein N. The effect: An introduction to research design and causality. Boca Raton, Florida: CRC Press; 2022.

[18] Jacobs B, Kissinger A, Zanasi F. Causal inference by string diagram surgery. In: Bojańczyk M, Simpson A, editors. Foundations of software science and computation structures. FoSSaCS 2019. Lecture notes in computer science. vol. 11425. Cham: Springer; 2019. https://doi.org/10.1007/978-3-030-17127-8_18.

[19] Maldonado G, Cox Jr LA. Causal reasoning in epidemiology: philosophy and logic. Glob Epidemiol 2020. https://doi.org/10.1016/j.gloepi.2020.100020.

[20] Martinussen T. Causality and the Cox regression model (March 1, 2022). Annu Rev Stat Appl 2022;9(1):249–59. Available at, https://doi.org/10.1146/annurev-statistics-040320-114441.

[21] Molnar C. Interpretable machine learning: a guide for making black box models explainable. 2nd ed. 2022. christophm.github.io/interpretable-ml-book/.

[22] Pearl J. Causal inference in statistics: an overview. Stat Surv 2009;3:96–146. https://doi.org/10.1214/09-SS057.

[26] Savage S. The flaw of averages: Why we underestimate risk in the face of uncertainty. Hoboken, New Jersey: John Wiley & Sons; 2009.

[27] Tse YK. Nonlife Actuarial Models: Theory, Methods and Evaluation. New York, New York: Cambridge University Press; 2009.

[28] VanderWeele T. Explanation in causal inference: Methods for mediation and interaction. United Kingdom: Oxford University Press; 2015.

[29] Witte J, Henckel L, Maathuis MH, Didelez V. On efficient adjustment in causal graphs. J. Mach. Learn. Res. 2020;21(1):246.

[30] Zhao Q, Hastie T. Causal interpretations of black-box models. J Bus Econ Stat 2019; 2019. https://doi.org/10.1080/07350015.2019.1624293. PMID: 33132490; PMCID: PMC7597863.

[31] Zigler CM, Dominici F. Point: clarifying policy evidence with potential-outcomes thinking—beyond exposure-response estimation in air pollution epidemiology. Am J Epidemiol 2014;180(12):1133–40. https://doi.org/10.1093/aje/kwu263.