Title

The P value plot does not provide evidence against air pollution hazards

Journal

Environmental Epidemiology, 6(2)

ISSN

2474-7882

Author

Hicks, Daniel J

Publication Date

2022

DOI

10.1097/ee9.0000000000000198

Peer reviewed

Review Article

ENVIRONMENTAL EPIDEMIOLOGY

OPEN

The *P* value plot does not provide evidence against air pollution hazards

Daniel J. Hicks

Background: A number of papers by Young and collaborators have criticized epidemiological studies and meta-analyses of air pollution hazards using a graphical method that the authors call a *P* value plot, claiming to find zero effects, heterogeneity, and *P* hacking. However, the *P* value plot method has not been validated in a peer-reviewed publication. The aim of this study was to investigate the statistical and evidentiary properties of this method.

Methods: A simulation was developed to create studies and meta-analyses with known real effects δ, integrating two quantifiable conceptions of evidence from the philosophy of science literature. The simulation and analysis is publicly available and automatically reproduced.

Results: In this simulation, the plot did not provide evidence for heterogeneity or *P* hacking with respect to any condition. Under the right conditions, the plot can provide evidence of zero effects; but these conditions are not satisfied in any actual use by Young and collaborators.

Conclusion: The *P* value plot does not provide evidence to support the skeptical claims about air pollution hazards made by Young and collaborators.

Keywords: *P* value plot; Simulation methods; *P* hacking; Air pollution

Introduction

In numerous recent works,[1–12] statistician Young and collaborators have criticized epidemiological studies and meta-analyses of the harmful effects of air pollution. These authors have used a graphical method they call a *P* value plot, claiming that this method reveals that zero effects, heterogeneity, and *P* hacking are widespread in environmental epidemiology.

Young and collaborators have drawn highly skeptical conclusions about the hazards of air pollution using their *P* value plot method, claiming that "causality of PM10/PM2.5 on heart attacks is not supported"[4] and "There is no convincing evidence of an effect of PM2.5 on all-cause mortality"[9]; Young has advocated that "regulation of PM2.5 should be abandoned altogether."[13] While recent papers in this body of work have not yet been highly cited in the academic literature, one paper by these authors[14] was cited in the scientific review of US EPA's Ozone Integrated Science Assessment[15]; this review was conducted while Young was serving on US EPA's Science Advisory Board.[16] So it is highly plausible that this body of work could be cited in the future.

However, Young and collaborators have provided only a minimal analysis of the statistical properties of the *P* value plot method, in a set of publicly available but non–peer-reviewed notes.[6] They have sometimes attempted to justify their method by citing plots of *P* values developed by other authors.[17,18] But these other plots are designed to answer different questions, are constructed in different ways, and have different statistical properties.

The aim of this study was to formally evaluate the evidentiary value of the *P* value plot method as used by Young and collaborators. Numerical simulations were chosen for accessibility, extensibility, and speed.

Methods

The P value plot

The *P* value plot is constructed from collections of *P* values, often extracted from the primary studies in a meta-analysis. Features of the plots are interpreted as indicating that there is no underlying effect,[1,3,6,8,9] a heterogeneous mixture of zero and nonzero effects,[4–7] or that the authors of the primary studies have engaged in *P* hacking.[7,12] (Young and collaborators often

University of California, Merced, Merced, California

The complete code and manuscript for this paper is available at https://github.com/dhicks/p_curve. The automatically-reproduced analysis can be viewed at https://dhicks.github.io/p_curve/.

SDC *Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.environepidem.com).*

**Corresponding Author. Address: University of California, 5200N Lake Road, University of California, Merced, Merced, CA 95343. E-mail: dhicks4@ucmerced.edu (D.J. Hicks).*

Environmental Epidemiology (2022) 6:e198

Received: 15 October 2021; Accepted 24 January 2022

Published online 18 February 2022

DOI: 10.1097/EE9.0000000000000198

What this study adds

This study uses a simulation approach to examine the statistical and evidentiary properties of the p-value plot, a graphical method that has been used to criticize air pollution epidemiology. These properties have not been examined in previous peer-reviewed publications. The results show that the method is incapable of providing evidence to support claims of p-hacking and statistical heterogeneity. While the method can produce evidence of a zero effect, the method only has this ability under certain conditions. These conditions are identified, and it is observed that the published criticisms do not satisfy these conditions.

refer to the replication crisis literature, but this literature is not immediately relevant to environmental epidemiology.[19])

Lack of exposition is a major initial challenge in evaluating this method. The method appears to rely almost entirely on visual inspection, with only vague characterizations of how features of the plots should be interpreted. These interpretations are rarely given a clear justification, and at least one key term in the exposition is used in a nonstandard way. Therefore, in this section, I develop a more precisely-defined and automatically reproducible set of methods. Because we would not expect less rigorous methods to provide better evidence, if my rigorous methods do not provide evidence to support the skeptical claims about air pollution made by Young and collaborators, then we should conclude that purely visual methods do not provide evidence either.

The *P* value plot does not appear to have been validated in any peer-reviewed studies. Young and collaborators do give references to two other graphical methods.[17,18] However, these other methods are substantially different from each other and the *P* value plot used by Young and collaborators. The *P* curve[18] is a histogram of *P* values below the conventional 0.05 threshold, analyzed in terms of its skew, whereas the *P* value plot contains points for each individual *P* value in a collection (including those above 0.05) and is analyzed in terms of a "hockey stick" shape, "gaps," slope, and linearity. Simulation methods have been used to validate the *P* curve.[18] Schweder and Spjøtvoll's[17] *P* value plot corresponds more closely to the *P* value plot used by Young and collaborators, and is also analyzed in terms of its slope; but the two slopes are not in 1-1 correspondence and Schweder and Spjøtvoll[17] assume the *P* values come from tests of different hypotheses rather than replications of a test of a single hypothesis. (Supplement; http://links.lww.com/EE/A178, provides formal definitions for all three plots, shows examples using simulated data for a range of real effects that include heterogeneous cases, and explains the differences in detail.) So citations to these other graphical methods are insufficient to validate the method used by Young and collaborators. Two works[6,12] include a handful of examples generated using simulated data with zero real effect $\delta = 0$. However, these studies do not report any simulations of cases where the real effect was nonzero or heterogeneous, do not report making the simulation code available anywhere for checking reproducibility or extending the analysis, and have not undergone peer review.

To formally define the *P* value plot, we begin with a set of N *P* values $\mathbb{P} = \{p_1, p_2, \cdots, p_N\}$, (nominally) produced by applying a given statistical hypothesis test to *N* replications of a given study design, each replication drawing samples of size n from the given population. This corresponds to the simplest case of meta-analysis. Thus, the *P* values in $\mathbb{P}$ are nominally samples from a single underlying distribution $p_i \sim P$. Note that, if the real effect is zero $\delta = 0$, then *P* is the uniform distribution on $[0,1]$.

Next, let $\text{rank}_{\text{asc}}(p_i)$ be the (1-indexed) ascending rank of $p_i \in \mathbb{P}$, i.e., $\text{rank}_{\text{asc}}(p_i)$ is the number of *P* values $p_j \in \mathbb{P}$ less than or equal to p_i. The smallest *P* value has ascending rank 1, and the largest *P* value has ascending rank *N*. Without loss of generality, if $\mathbb{P}$ is already in ascending order $p_1 < p_2 < \cdots < p_N$, then $\text{rank}_{\text{asc}}(p_i) = i$. The *P* value plot is the graph (i, p_i). Note that this is equivalent to a rescaled QQ-plot of $\mathbb{P}$ against the uniform distribution, with the theoretical quantiles $q_i = \frac{i}{N} = \frac{\text{rank}_{\text{asc}}(p_i)}{N}$.

Young and collaborators explain their interpretation of the plot as follows: "Evaluation of a p-value plot follows a logical path. Is the p-value plot [sic] homogeneous? If the points roughly [sic] on a 45-degree, they are homogeneous and consistent with randomness; a lessor slope with all points roughly on a line indicates a consistent effect even if some of the individual p-values are not considered statistically significant. If the effects differ, one from another, beyond chance, then the effects are heterogeneous."[4]

A "45-degree line" typically refers to the graph of an identity function or the line $y = x$, which forms a 45-degree angle with respect to both the x and y axes. This interpretation makes sense for the equivalent QQ-plot, where both axes are on the scale [0,1]; here, a slope of 1 indicates that the underlying distribution *P* is uniformly distributed, which in turn indicates that the real effect is zero. Strictly speaking, this interpretation does not make sense for the *P* value plot as defined, where the x axis (rank) is on the scale [1,*N*] and the y axis (*P* value) is on the scale [0,1]. I will assume that a "45-degree line" typically means that the slope of the equivalent QQ-plot is 1. It appears that slopes are only evaluated visually; there are no reports of fitting regression models or using any other quantitative methods to measure slopes.

There are frequent claims that the *P* value plot contains a "hockey stick" or "bilinear" pattern,[4–7,12] which has "small *P* values to the lower left … and then points ascending in a roughly 45-degree line."[7] In this context, "45-degree line" seems to mean that the right-hand side of the plot is linear, even if it does not have a slope of 1. This nonlinear "hockey stick" pattern is taken to indicate some combination of heterogeneous effects, *P* hacking, and researcher misconduct. On two occasions, a formal test for nonlinearity was conducted by comparing a linear and quadratic regression using an *F* test.[4,6] More often there is no explanation of how the "hockey stick" pattern was determined to be present or absent.

Simulation design

Each run of the simulation comprises *N* studies, collected together as though for a single meta-analysis. To facilitate interpretation, each study is based on a two-sample *t* test. Two samples, each of size n, are drawn from Gaussian distributions with means $\mu_1 = 0$ and $\mu_2 = \delta$, respectively, and common standard deviation $\sigma = 1$. (This data-generating process is just slightly more complicated for the heterogeneous case; see below.) δ corresponds to the true effect size as measured by Cohen's $d = (\mu_2 - \mu_1)/\sigma$.[20]

Parameter settings can be systematically varied to compare, e.g., different effect sizes, and multiple runs in each condition allow us to analyze the statistical properties of the *P* value plot within and across conditions. For the primary analysis, seven different effect sizes are used—corresponding to conventional thresholds for zero, very small, small, moderate, large, and very large effects,[20] and a "mixed" or heterogeneous condition. All studies have the same sample size, n = 26. Compared with the convention of 80% power, this sample size makes the studies severely underpowered to detect the very small (power 3%) and small effects (11%), somewhat underpowered for the moderate effect (42%), adequately powered for the large effect (81%), and overpowered for the very large effect (99%). Each condition is simulated with 500 runs. The primary conditions examined in the current study are summarized in Table 1.

In the "mixed" or heterogeneous condition, subsets of the population have different responses to the intervention or exposure. The author uses a {0, 0.8} mixture, meaning that one subpopulation has a zero effect or no response to the intervention and the other has a strong response of 0.8. These values were chosen to create a mixture that is less likely to look like a homogeneous case; a condition with a {0.3, 0.5} mixture would be difficult to distinguish from a homogeneous 0.4 condition.

When given this kind of mixture, in the simulation, an individual study draws its sample from one of the subpopulations, selected uniformly at random. So, in expectation, half of the studies will find a zero effect and half a strong effect, though in any given set of studies there will be variation in the ratios of the two subpopulations. For simplicity, the simulation currently does not support a continuous mixture.

The simulation does not have a way to represent *P* hacking, publication bias, or researcher misconduct; however, this means that all conditions represent cases in which these factors are absent.

Table 1.

Parameter values used in the current simulation study.

Parameter	Meaning	Value(s)
δ	Real effect size	0, 0.01, 0.2, 0.5, 0.8, 1.2, {0, 0.8}
σ	SD of samples	1
n	Study sample size	26
N	Number of studies in each run	20

The real effect size of {0, 0.8} indicates the "mixed" or heterogeneous condition, in which half the population have no response and half the population have a strong response. Five hundred runs in each combination of parameters are used for the results reported in the table.

After generating the data, plots are constructed and analyzed (i) visually for the "hockey stick" pattern, and (ii–iv) using computationally reproducible analyses of gaps, slopes, and linearity.

To assess "gaps" in the plot (ii), I calculate the largest difference in consecutive *P* values $\max(p_{i+1} - p_i)$. I classify a *P* value plot as "gappy" if this largest difference is greater than a threshold value 0.125. That is, if there is at least one visual gap of at least 12.5% in the *P* value plot, the plot is considered "gappy." This threshold was chosen to capture the sense of a visual "discontinuity" in the sequence of *P* values. A continuous measure, size of the largest gap, is reported in the Supplement; http://links.lww.com/EE/A178.

Slopes can be calculated using a simple univariate linear regression. If the slope of the QQ-plot is approximately 1, this indicates that the underlying distribution of *P* values *P* is uniform, which in turn indicates that the zero hypothesis is true. Because the slope of the *P* value plot is *N* times the slope of the QQ-plot, analysis of the QQ-plot is simpler than analyzing the *P* value plot directly.

I assess the fitted slope in several ways, to judge whether it is "approximately 1" (iii). First, I consider simply whether the slope is in the range 1 ± 0.1. This range was chosen to capture the sense of visual difference from 1. Next, I apply three hypothesis tests: a *t* test against the null hypothesis that the slope is exactly 1; an equivalence test, the TOST (two one-sided tests) procedure,[21] against the null that the slope is outside the range 1 ± 0.1; and a Kolmogorov-Smirnov test (KS-test) against the null of the uniform distribution. I use the conventional 0.05 threshold for statistical significance for all three tests. When the estimate is not statistically significant, the tests as implemented in the simulation accept the null hypothesis. While strictly problematic, this approach simplifies the presentation and analysis of results and aligns with the way hypothesis tests are often used in practice. Because we are interested in these tests independently—as though the other analyses were not conducted—there is no need to correct for multiple comparisons. Distributions of slopes are also reported in the Supplement; http://links.lww.com/EE/A178.

Finally I evaluate the linearity of the plot (iv). Linearity of a plot can be tested by fitting two regression models—one linear, one quadratic—and selecting the model with the better fit. I use AIC (Akaike Information Criterion) and an *F* test for model selection. I use the conventional 0.05 threshold for statistical significance in the *F* test, and accept the null (linear model) when the *F* test is not significant. As with slope, for simplicity, I analyze the QQ-plot rather than the *P* value plot directly. In the Supplement (http://links.lww.com/EE/A178), area under the curve (AUC) is reported as a continuous measure of linearity, where nonlinear QQ-plots plots have lower AUC.

Measuring evidence

I use two quantifiable conceptions of evidence that have been widely discussed in the philosophy of science literature, severity, and likehood ratios. For brevity, only severity analysis results are reported here; likelihood ratio results are included in the Supplement; http://links.lww.com/EE/A178.

The severity conception of evidence is associated with philosopher of science Deborah Mayo's reconceptualization of frequentist hypothesis testing.[22,23] Young provided a positive blurb for Mayo's 2018 book,[23] stating that "Her severity requirement [sic] demands that the scientist provide a sharp question and related data. Absent that, the observer should withhold judgment or outright reject." In one work,[12] Young and collaborators cite Mayo[23] multiple times, and repeatedly characterize the *P* value plot as a "severe test."

The author uses Mayo's weak severity criterion:

> One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false. If data *x* agree with a claim [*H*] but the method used is practically guaranteed to find such agreement, and had little or no capability of finding flaws with [*H*] even if they exist, then we have bad evidence, no test (BENT).[23]

On this conception of evidence, a test or analytical method *T* with observed output *t* can provide evidence supporting a target hypothesis *H* only if *T* would have given a different output if *H* were false. (Note that the test *T* can give any kind of output: T = *t* might mean that an test statistic is greater than or equal to some value, or that a plot has some purely qualitative visual feature.) Hypothesis testing assesses this counterfactual condition using the *P* value. In its most general form, the *P* value can be defined as $P = \mathrm{pr}(T{=}t|\neg H)$, where the role of $\neg H$ ("not H," the logical negation of the hypothesis of interest) is played by a null hypothesis H_0. (Throughout this article, I distinguish a zero hypothesis—that some effect is exactly zero—from a null hypothesis—the alternate or rival hypothesis used to calculate a *P* value. For instance, the TOST procedure as used in this study has the null hypothesis that the true slope is in the set {0.9,1.1}.) A "small" *P* value indicates that the counterfactual is probably true, that is, if *H* were false then *T* would probably have given a different output. On the other hand, a "large" *P* value indicates that the test "is practically guaranteed" to produce this output, and so in this case by the weak severity principle "we have bad evidence, no test."

Severity can also be evaluated qualitatively when a *P* value cannot be calculated. Consider visual features of plots, such as the "hockey stick" pattern that is interpreted as evidence of heterogeneity. It is not clear how to quantitatively determine whether this visual pattern is present in a given plot. However, if this pattern is qualitatively common in homogeneous cases, then the weak severity principle implies that this visual pattern does not provide evidence of heterogeneity.

The severity conception of evidence can be applied to the skeptical claims about air pollution hazards as follows. The claims *H* are the zero hypothesis $\delta = 0$, mixture hypothesis $\delta = \{0, 0.8\}$, and the hypothesis that researchers have engaged in *P* hacking or other questionable research practices. The method or test *T* is the *P* value plot; the outputs *t* are detailed above: (i) the "hockey stick" pattern; (ii) "gaps" in the plot; (iii) slope of approximately 1 (on the QQ-plot); and (iv) nonlinearity inferences. Table 2 summarizes the outputs examined in the current study.

In the current simulation study, there are *P* values at multiple levels. First, there are the *P* values produced by the *t* tests in the primary studies, the *N* replications bundled together and plotted in a single *P* value plot. I will call these primary *P* values. Second, some of the tests conducted using the *P* value plot are themselves statistical hypothesis tests, such as the *F* test of linearity. These tests involve meta-level *P* values. Meta-level *P* values are tracked by the simulation, but not reported directly here. Instead, meta-level *P* values are compared to the conventional 0.05 threshold, and this comparison is used to produce a meta-level test output such as "non-linear." Finally, to assess the validity of the *P* value plot method, we use the simulation results to estimate the probability of observing a certain meta-level test output value—such as "non-linear"—when the true effect

Table 2.

Outputs of the *P* value plot examined using the simulation.

	Output	Determined using	Taken as evidence for
i	"Hockey stick"	Visual inspection	Mixed effect
ii	"Gaps"	Visual inspection Largest gap >0.125	*P* hacking or other problems
iii	Slope ≈1	Range 1 ± 0.1 *T* test not statistically significant TOST test statistically significant KS-test not statistically significant	Zero effect
iv	Nonlinearity	AIC: quadratic *F* test: statistical significant	Mixed effect

"Outputs" are features of plots that are taken as evidence for critical assessments of air pollution epidemiological studies. The "determined using" column indicates how these outputs are identified as present/absent in the current study.

size satisfies a given null hypothesis—such as $\delta = 0.5$. Because the resulting *P* value is being used to assess the validity of the *P* value plot method, I will call it a validation *P* value. When the validation *P* value is "large" (greater than the conventional 0.05), the weak severity criterion implies that this particular use of the *P* value plot does not provide evidence for the target hypothesis with respect to the given null hypothesis.

To generate validation *P* values (or qualitatively assess severity for the visual analysis), we need to specify the null hypothesis that plays the role of $\neg H$. I will use each of the following:

a. very small: $\delta = .01$.
b. small effect: $\delta = 0.2$
c. moderate effect: $\delta = 0.5$
d. strong effect: $\delta = 0.8$
e. very strong effect: $\delta = 1.2$
f. greater than zero: $a \vee \cdots \vee e$
g. mixed effect: $\delta = \{0, 0.8\}$ for $H : \delta = 0$ and vice-versa (i.e., the other skeptical hypothesis)
h. any other effect: $f \vee g$ (any of the nonzero effects or the other skeptical hypothesis)

In each case, when the validation *P* value is "large" $P > 0.05$, the simulation results indicate that this test output is common in the null hypothesis case, and so the weak severity criterion implies that the *P* value plot would not provide evidence to support the skeptical claim made by Young and collaborators.

Reproducibility

The simulation, analysis, and outputs (figures and tables) are publicly available and automatically reproduced. Code is available at https://github.com/dhicks/p_curve and the automatically reproduced analysis can be viewed at https://dhicks.github.io/p_curve/.

The simulation and analysis were both written in R version 4.1.0[24] and make extensive use of the tidyverse suite of packages, version 1.3.1.[25] The TOSTER package version 0.3.4[26] was used to conduct the TOST analysis. Because the software on the virtual machine used to automatically reproduce the analysis is updated each time the analysis is re-run, software versions reported online may be different from those reported here.

Results

Figure 1 shows 35 examples of the *P* value plot across the seven conditions, and Figure 2 shows the *P* value plot across all runs of the simulation.

There are substantial qualitative differences within effect sizes as well as similarity across consecutive effect sizes. Except for the very large (overpowered) effect size, there tends to be both statistically significant and insignificant primary *P* values. Larger effect sizes have more statistically significant primary results, resulting in a series of small primary *P* values that gradually bend up. However, the top row of Figure 1 and the first panels of Figure 2 indicates that even zero effects can look nonlinear. Comparing the composite plots (Figure 2) for the mixed condition and the moderate effects condition, on average the mixed condition tends to produce a sharper bend upwards. However, Figure 1 indicates that an individual moderate effects plot can have a sharp bend (index 290) and a mixed effects plot can have a gradual bend (index 66).

Visual analysis

I focus on two visual patterns that are frequently discussed in the critiques of air pollution epidemiology: (i) the "hockey stick" pattern; (ii) "gaps" in the plot.

The "hockey stick" pattern is taken to be evidence of heterogeneity. The "hockey stick" comprises a more-or-less flat series of small p-values on the left (supposedly corresponding to the mixture component with a real effect) and a second series of steeply increasing *P* values on the right (supposedly corresponding to the mixture component with zero effect). This pattern is visible in all of the example plots for the moderate, strong, and mixed effects (rows 4, 5, and 7), and arguably also with small effects (row 3). Because the "hockey stick" pattern appears in plots where there is only a single homogeneous effect, the weak severity criterion implies that the hockey stick pattern does not provide evidence of a mixed or heterogeneous case.

Visual "gaps" in the plot are taken to be evidence of *P* hacking, publication bias, or other questionable researcher practices.[2,6,7] These gaps are common in Figure 1, across all conditions except the very strong effect (where almost all values are below 0.05). Note that the simulation does not include *P* hacking, publication bias, or other questionable researcher practices. Thus the weak severity criterion implies that gaps in the plot do not provide evidence of *P* hacking, publication bias, or other questionable researcher practices.

Computationally reproducible analyses

Figure 3 shows the results of the severity analysis as validation *P* values; see the Supplement (http://links.lww.com/EE/A178) for a table version of these results. By the weak severity criterion, when the results of the severity analysis are greater than 0.05 (above the dashed line), the test output does not provide evidence in support of the target hypothesis.

Unspecified problems of *P* hacking, publication bias, or other questionable research practices are supposedly supported by gaps in the plot. The validation *P* value for the presence of these gaps (output ii-gap) is greater than 0.25 for every null hypothesis except the very strong effect, indicating that gaps are

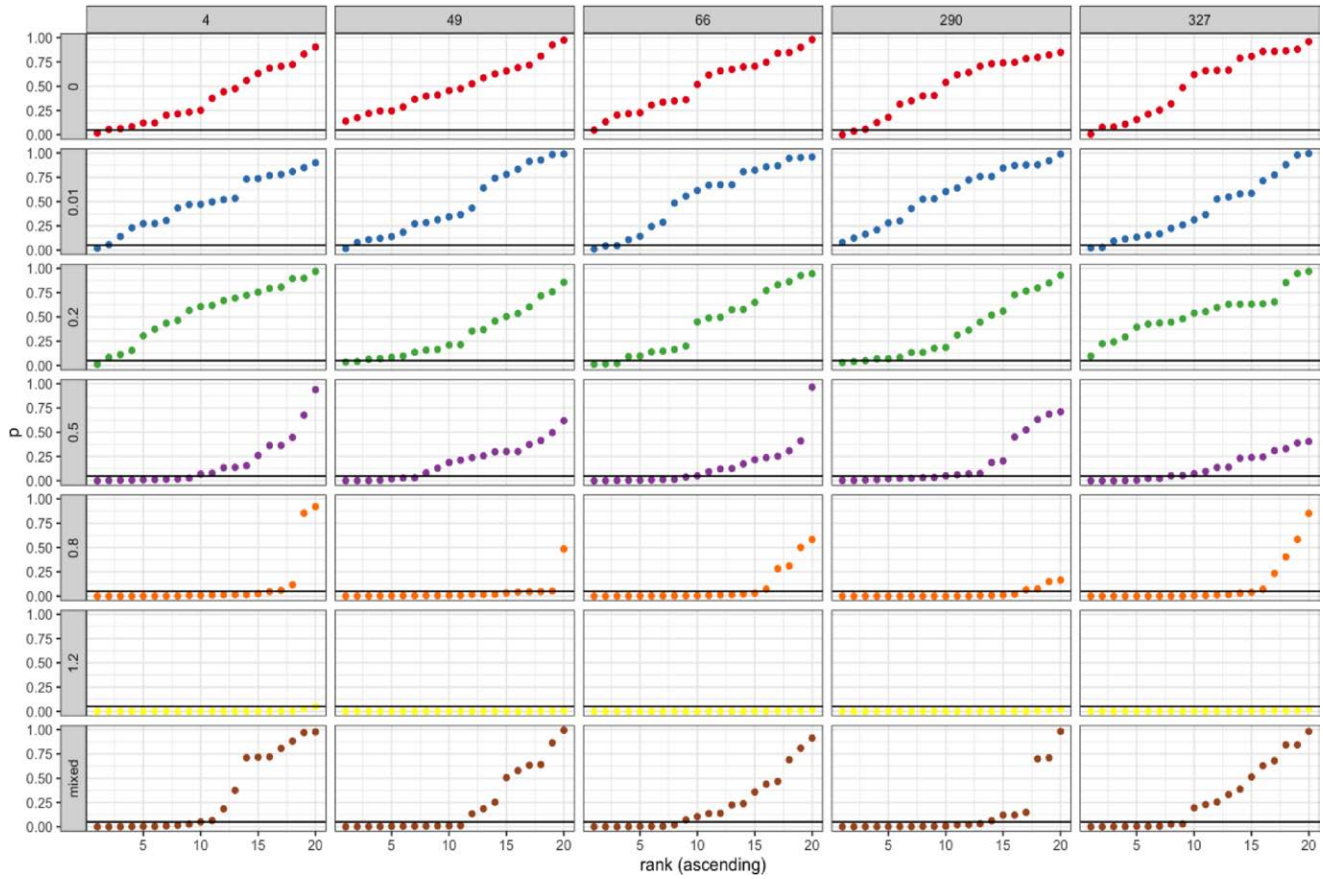

Figure 1. Examples of the *P* value plot. Drawn at random from the simulation results. Rows and colors correspond to conditions or real effects (δ), from zero (0) to very strong (1.2) and a mixed condition δ = {0.0,0.8}. Columns correspond to indices for the simulation runs that produced these results, and are not meaningful. Each point corresponds to a single *P* value in the meta-analysis (simulation run); the *x* axis is the ascending rank of the *P* value in the set, and the *y* axis is the *P* value itself.

quite common. In all conditions except the large and very large effects, the majority of *P* value plots are "gappy." Thus, gaps in the plot do not provide evidence of *P* hacking, publication bias, or other questionable researcher practices. The distribution of the size of the largest gap across each real effect size is reported in the Supplement; http://links.lww.com/EE/A178. Most plots have gaps larger than 0.125 across all conditions except the very strong effect.

The zero effect hypothesis is supposedly supported by a slope of approximately 1 on the QQ-plot. All methods are severe against large and very large effects; only the TOST and KS tests are severe against moderate effects; and no methods are severe against very small or small effects. The TOST test also may be severe against the greater-than-zero and non-zero hypotheses. So the "45-degree line" may or may not provide evidence for zero effects, depending on the particular null hypothesis being tested and particular test used.

This means that, to determine whether the "45-degree line" provides evidence for zero effects, the analyst must be explicit about both the test method used and the null hypothesis being tested. A slope of 0.92 or statistically significant TOST test (meta-level *P* value less than 0.05), for example, will not provide evidence if the null hypothesis is a very small effect. In addition, while these results recommend using the TOST test, this test requires an explicit equivalence interval, the range of values that are "approximately 1." But Young and collaborators are not explicit in any of these requisite ways. They do not report calculating a slope, using a regression line, or other method, much less conducting some further analysis of that slope. I take it that the lack of detail is a good reason to think that they have simply assessed the slope visually. A visual assessment will be less sensitive than calculating a slope and determining whether it is in the range 1 ± 0.1, which can only provide evidence against a relatively large null hypothesis. So in cases where a small effect is a live possibility, insofar as the *P* value plot is interpreted using visual judgment alone, the "45-degree line" does not provide evidence of a zero effect. The distribution of slopes across each real effect size is reported in the Supplement; http://links.lww.com/EE/A178. Slopes in the range 1 ± 0.1 are common for very small, small, and mixed effects.

The mixed-effect hypothesis, or heterogeneity, is supposedly supported by nonlinearity. Two quantified versions of this output are examined here, comparing linear and quadratic regressions using AIC (iv-AIC) and an *F* test (iv-*F*). The AIC and *F* test evaluations of nonlinearity do not provide severe tests against any of the alternative conditions. That is, neither of the tests of linearity considered here provides evidence for heterogeneity. In the Supplement; http://links.lww.com/EE/A178, the area under the curve (AUC) of the QQ-plot is used as a continuous measure of nonlinearity, and distributions of AUC values are reported across each real effect size. Lower AUC values indicate further deviation from the line $x = y$. The AUC distribution for the mixed effect overlaps substantially with the distributions for all other effect sizes except very large.

Discussion

This simulation analysis finds that the *P* value plot does not provide evidence for heterogeneity or *P* hacking based on the "hockey stick" shape, "gaps" in the plot, or AIC or *F* tests

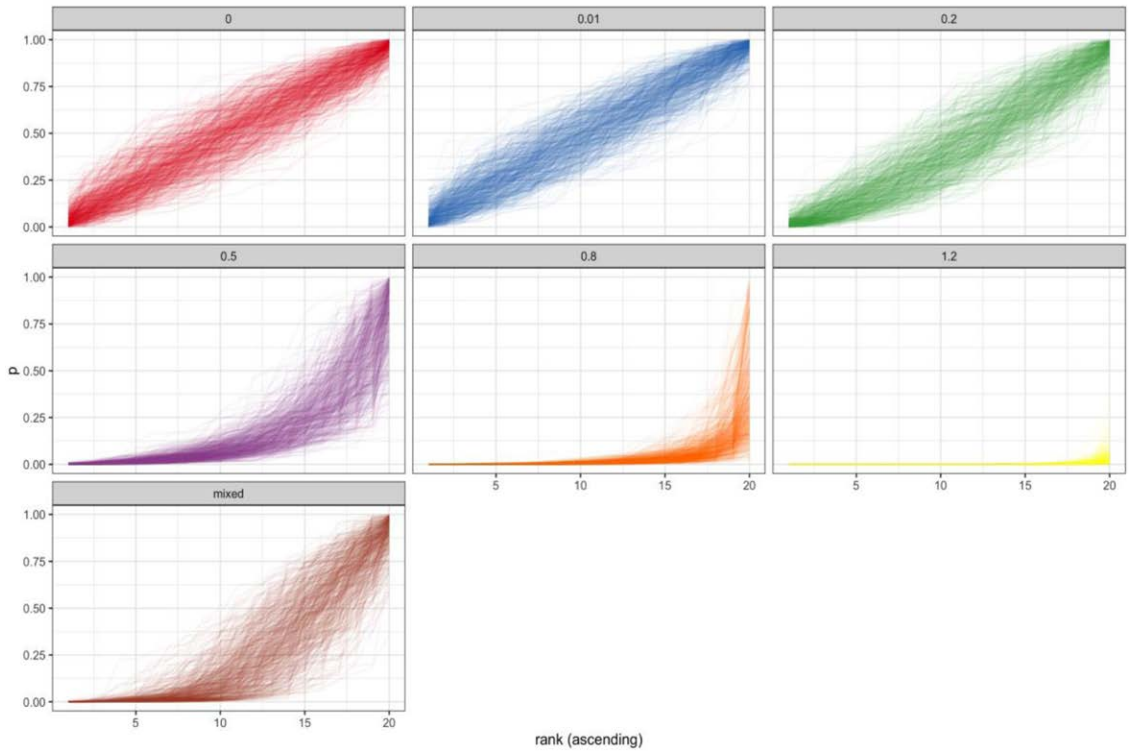

Figure 2. Composite of the *P* value plot. Each individual line is the *P* value plot for a single run of the simulation; all simulation runs are shown here. Panels correspond to conditions or real effects (δ), from zero (0) to strong (1.2) and a mixed condition δ = {0.0, 0.8}.

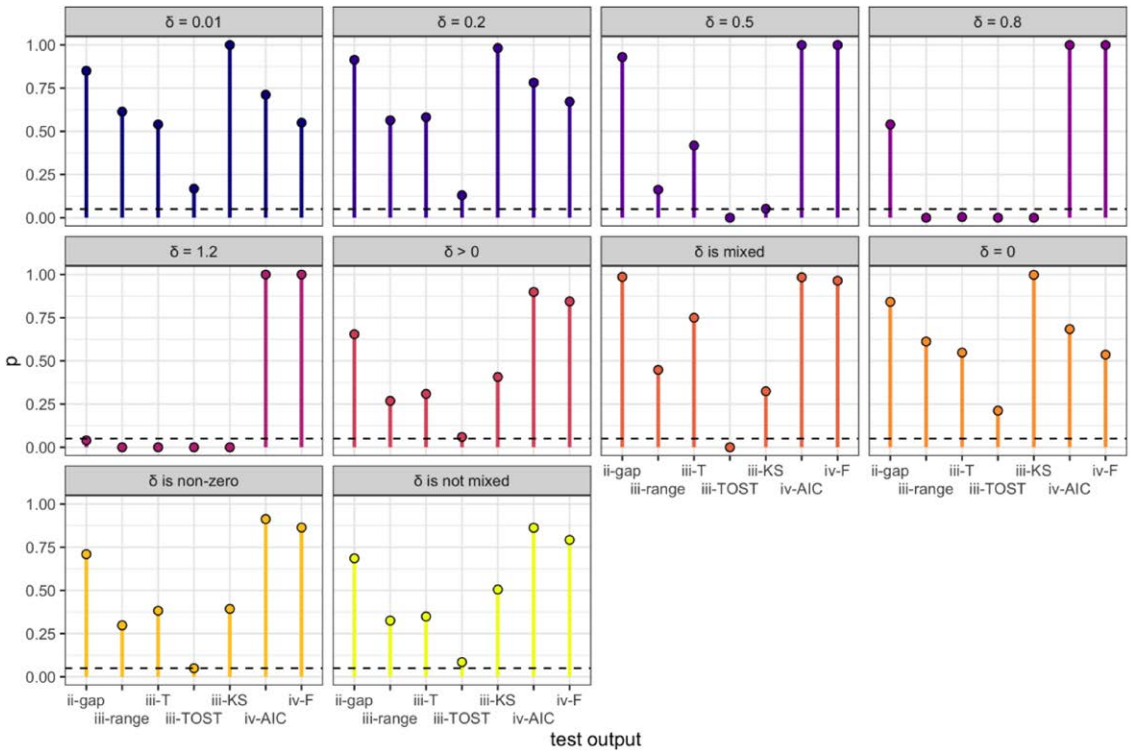

Figure 3. Results of the severity analysis for outputs (ii) gaps in the plot, (iii) slope of 1, and (iv) nonlinearity. Severity analysis results are reported as validation *P* values: small values (conventionally <0.05) indicate a severe test with respect to the null hypothesis. Panels correspond to null hypotheses, and *y* axis values correspond to the severity assessment (as a validation *P* value) for the output with respect to the given null hypothesis. The dashed line indicates *P* = 0.05; points below this line indicate severe tests.

of nonlinearity. The method can provide evidence for zero effects based on a slope of 1, depending on what rival or null hypothesis is considered and how the plot is analyzed. In general, producing evidence for zero effects against small effects requires using the TOST approach; visual inspection alone will not be sufficient. This approach requires setting an explicit range of values within which the slope is considered "approximately 1."

In the meta-analyses criticized by Young and collaborators, the estimated short-term effects of air pollution are small or very small on a relative risk scale. For example, Nawrot et al[27] estimated the effect for air pollution (increase of 10 μg/m^3 PM_{10}) on nonfatal myocardial infarction to be 1.02 (95% confidence interval [CI] = 1.01–1.02); the point estimates for six pollutants reported by Mustafic et al[28] (increase of 10 μg/m^3 for all except carbon monoxide) were all in the range 1.003–1.048; Liu et al[29] estimated effects for PM_{10} (increase of 10 μg/m^3) on all-cause mortality of 1.044 (95% CI = 1.039, 1.050); and Orellano et al[30] estimated effects for $PM_{2.5}$ on all-cause morality of 1.0065 (95% CI = 1.0044, 1.0086). A precise conversion from risk ratios to Cohen's *d* is beyond the scope of this article. However, using a rule of thumb that the risk ratio is approximately equal to the log odds when the outcome is rare[31] and the conversion factor $\sqrt{3}/\pi$ between the log odds and Cohen's *d*, a risk ratio of 1.05 is roughly equivalent to $d = 0.03$, which is only slightly larger than the very small effect condition examined here. (An additional analysis in the automatically reproduced analysis document examines conditions with a very small real effect sizes of $\delta = 0.05$ and varying power to detect these very small effects.) Visual inspection of the *P* value plot alone is incapable of producing evidence against very small epidemiological effects of air pollution.

More often, Young and collaborators have claimed to find evidence of heterogeneity, *P* hacking, and publication bias.[4,5,7] The simulation results indicate that the *P* value plot is incapable of providing evidence for any of these claims, using either visual inspection or either of the quantitative approaches examined here. The features that Young and collaborators point to—the "hockey stick" shape, "gaps," nonlinearity—are readily produced by moderate and stronger effects, and can even appear in zero and very small effect conditions.

All together, the *P* value plot method cannot support the skeptical claims about air pollution epidemiology made by Young and collaborators.

Conflicts of interest statement

The author declares that they have no conflicts of interest with regard to the content of this report.

References

1. Young SS, Bang H, Oktay K. Cereal-induced gender selection? Most likely a multiple testing false positive. *Proc Biol Sci*. 2009;276:1211–1212.
2. Young SS, Xia JQ. Assessing geographic heterogeneity and variable importance in an air pollution data set. Statistical Analy Data Mining [Internet]. 2013;6:375–386. Available at: http://doi.wiley.com/10.1002/sam.11202. Accessed 26 June 2020.
3. You C, Lin DKJ, Young SS. PM2.5 and ozone, indicators of air quality, and acute deaths in California, 2004–2007. Regulatory Toxicology and Pharmacology [Internet]. 2018;96:190–196. Available at: http://www.sciencedirect.com/science/article/pii/S0273230018301430. Accessed 26 July 2020.
4. Young SS, Acharjee MK, Das K. The reliability of an environmental epidemiology meta-analysis, a case study. *Regul Toxicol Pharmacol*. 2019;102:47–52.
5. Young SS, Kindzierski W. Ambient air pollution and mortality in 652 cities. *N Engl J Med*. 2019;381:2073.
6. Young SS, Kindzierski W. Combined background information for meta-analysis evaluation [Internet]. 2019. Available at: http://arxiv.org/abs/1808.04408. Accessed 22 January 2020.
7. Stanley Young S, Kindzierski WB. Evaluation of a meta-analysis of air quality and heart attacks, a case study. *Crit Rev Toxicol*. 2019;49:85–94.
8. Young SS. Re: extended mortality follow-up of a cohort of 25,460 workers exposed to acrylonitrile. *Am J Epidemiol*. 2020;189:360–361.
9. Young SS, Kindzierski W. PM2.5 and all-cause mortality [Internet]. 2020 [cited 2020 Dec 9]. Available from: http://arxiv.org/abs/2011.00353
10. Young SS, Kindzierski W. Particulate Matter Exposure and Lung Cancer: A Review of two Meta-Analysis Studies [Internet]. 2020. Available at: http://arxiv.org/abs/2011.02399. Accessed 6 September 2021.
11. Kindzierski W, Young S, Meyer T, Dunn J. Evaluation of a meta-analysis of ambient air quality as a risk factor for asthma exacerbation. *J Respir*. 2021; 1:173–196.
12. Young SS, Kindzierski W, Randall D. Shifting Sands: Unsound Science and Unsafe Regulation [Internet]. National Association of Scholars. 2021. Available at: https://www.nas.org/reports/shifting-sands-report-i. Accessed 9 September 2021.
13. Young SS. Suggestions for EPA [Internet]. 2017. Available at: https://www.regulations.gov/document?D=EPA-HQ-OA-2017-0190-36647. Accessed 28 July 2020.
14. Young SS, Smith RL, Lopiano KK. Air quality and acute deaths in California, 2000–2012. Regulatory Toxicology and Pharmacology [Internet]. 2017;88:173–184. Available at: http://www.sciencedirect.com/science/article/pii/S0273230017301538. Accessed 24 January 2020.
15. US EPA Clean Air Scientific Advisory Committee. Review of the EPA's Integrated Science Assessment for Ozone and Related Photochemical Oxidants [Internet]. 2020. Available at: https://yosemite.epa.gov/sab/sabproduct.nsf/0/6b29a4de74ff843985258485005f18ca!OpenDocument&TableRow=2.3#2. Accessed 27 July 2020.
16. US EPA. Members of the Science Advisory Board [Internet]. 2017. Available at: https://web.archive.org/web/20171201172820/https://yosemite.epa.gov/sab/sabpeople.nsf/WebExternalCommitteeRosters?OpenView&committee=BOARD&secondname=Science%20Advisory%20Board. Accessed 31 July 2020.
17. Schweder T, Spjøtvoll E. Plots of P-values to evaluate many tests simultaneously. *Biometrika*. 1982;69:493–502.
18. Simonsohn U, Nelson LD, Simmons JP. P-curve: a key to the file-drawer. *J Exp Psychol Gen*. 2014;143:534–547.
19. Hicks DJ. Open science, the replication crisis, and environmental public health. Accountability in Research [Internet]. 2021. Available at: https://doi.org/10.1080/08989621.2021.1962713. Accessed 31 July 2021.
20. Sawilowsky S. New effect size rules of thumb. *J Modern Appl Stat Methods*. [Internet]. 2009;8. Available at: https://digitalcommons.wayne.edu/jmasm/vol8/iss2/26.
21. Lakens D, Scheel AM, Isager PM. Equivalence Testing for Psychological Research: A Tutorial: Advances in Methods and Practices in Psychological Science [Internet]. 2018. Cited 17 October 2020. Available at: https://journals.sagepub.com/doi/10.1177/2515245918770963.
22. Mayo D. *Error and the Growth of Experimental Knowledge*. University of Chicago Press; 1996:493 (Science and its conceptual foundations).
23. Mayo DG. Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars [Internet]. Cambridge University Press; 2018:503.
24. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing; 2018. Available at: https://www.r-project.org/. Accessed 27 August 2018.
25. Wickham H, RStudio. Tidyverse: Easily Install and Load the 'Tidyverse' [Internet]. 2019. Available at: https://CRAN.R-project.org/package=tidyverse. Accessed 31 July 2020.
26. Lakens D. TOSTER: Two One-Sided Tests (TOST) Equivalence Testing [Internet]. 2018. Available at: https://CRAN.R-project.org/package=TOSTER. Accessed 11 December 2020.
27. Nawrot TS, Perez L, Künzli N, Munters E, Nemery B. Public health importance of triggers of myocardial infarction: a comparative risk assessment. *Lancet*. 2011;377:732–740.
28. Mustafic H, Jabre P, Caussin C, et al. Main air pollutants and myocardial infarction: a systematic review and meta-analysis. *JAMA*. 2012;307:713–721.
29. Liu C, Chen R, Sera F, et al. Ambient particulate air pollution and daily mortality in 652 cities. *N Engl J Med*. 2019;381:705–715.
30. Orellano P, Reynoso J, Quaranta N, Bardach A, Ciapponi A. Short-term exposure to particulate matter (PM10 and PM2.5), nitrogen dioxide (NO2), and ozone (O3) and all-cause and cause-specific mortality: systematic review and meta-analysis. *Environ Int*. 2020;142:105876.
31. Viera AJ. Odds ratios and risk ratios: what's the difference and why does it matter? *South Med J*. 2008;101:730–734.