# Shifting Sands: Using *P*-Value Plotting to Test Reproducibility of Research

S. Stanley Young, Ph.D.
Warren B. Kindzierski, Ph.D.

Governments issue regulations with tremendous impact on the economy and on health practices. Just the Clean Air regulations of the Environmental Protection Agency (EPA) cost Americans at least $120 billion per year. The EPA's pollution regulations constitute a large proportion of the total regulatory power available to the federal government.[1]

These regulations are based on published scientific research. But there is a crisis of irreproducibility (falseness) of research claims in science. Many claims show the "Bunnies in the Sky" phenomenon: If you look long and hard enough at the clouds, you will see something. The way data analysis is done in many observational population studies explains why statistically significant comparisons do not stand up. A false positive claim can become an established narrative. That is the basis for many regulations that have major impacts on business and industrial/agricultural activities.

Today bad (irreproducible) science has crowded out good (reproducible) science in the literature.[2] Bad science permits governments and policymakers to develop policies and regulations in many areas—including the environment, health, and the workplace—with no reliable evidence of public benefit or even with actual harm and in many cases wasteful expenditures. As part of the National Association of Scholars Shifting Sands Project, we investigated the impact and role of the irreproducibility crisis in scientific research and how it affects different areas of government policy.[3]

### *P*-Value Plots

Many physicians may be familiar with risk ratios – e.g., relative risks (RRs) and odds ratios (ORs) along with confidence intervals (CIs). These statistics are common measures used in numerous areas of research in medicine, environmental epidemiology, nutritional epidemiology, and elsewhere. Many physicians may also be familiar with a forest plot, which graphically shows several RRs (or ORs) along with their CIs in one plot. A forest plot is commonly used in meta-analysis of studies addressing a specific research question—e.g., does more of factor A cause more of outcome B?

The *p*-value is a complementary statistical measure of a confidence interval, CI. That is to say, a *p*-value can be calculated directly from a RR, or OR, given the CI. The *p*-value plot is an analogy to the forest plot, except that it can better show the heterogeneity (dissimilarity) of a data set being combined in meta-analysis compared to the forest plot.

It is important to convey to the reader that we do not use *p*-values and the *p*-value plot to make statements about probabilities or outcomes. We use *p*-values and the plot only to test for the irreproducibility (falseness) of a research claim. This is an informal inference process by which we make a generalization based on the characteristics (pattern) of the *p*-values in the plot without using formal statistical procedures or methods.

More specifically, the *p*-value plot provides a simple visual insight into possible inappropriateness (incorrectness) of assumptions of the statistical model used in a meta-analysis—in our case, a cause-effect model. There is no disputing the pattern of *p*-values in a *p*-value plot for a true null (no effect or association) versus a true effect (real effect or association).

We take the test statistics presented in observational studies (i.e., the RRs, or ORs, and CIs), convert them into *p*-values, and display them in the plot according to rank, smallest to largest. Our experience is that many meta-analysis research claims using observational population data are irreproducible (i.e., false). This is something which can be shown in a *p*-value plot. We can use the plot to independently confirm, disprove, or identify ambiguity (uncertainty) in a meta-analysis finding and research claim.

In a simulated series of 100 experiments designed to confirm or reject some unspecified null hypothesis, the black dots in Figure 1-A are *p*-values, presented in chronological order from left to right. These *p*-values were simulated using a pseudo-random number generator; a uniform distribution on the interval [0, 1] was specified. The results for experiments involving A and B that have little or no (null) association form a scatter plot.

If the results are ordered by the *p*-value rank, smallest to largest (Figure 1-B), they form nearly a straight line at a 45-degree angle. An upward sloping quasi-linear curve, appearing when *p*-values are plotted in ascending order, is a sort of fingerprint. Whenever it appears, we can be certain that the null hypothesis has not been rejected, i.e., there is no effect.[1]

Under the null hypothesis (i.e., hypothesis of no effect or association between two variables), a set of *p*-values will be distributed uniformly over the interval 0 to 1 regardless of sample size.[4] In a *p*-value plot, these *p*-values—plotted according to rank—will present as a nearly straight line at a 45-degree angle. This fingerprint implies a data set exhibiting randomness (chance findings) and not a real effect or association.

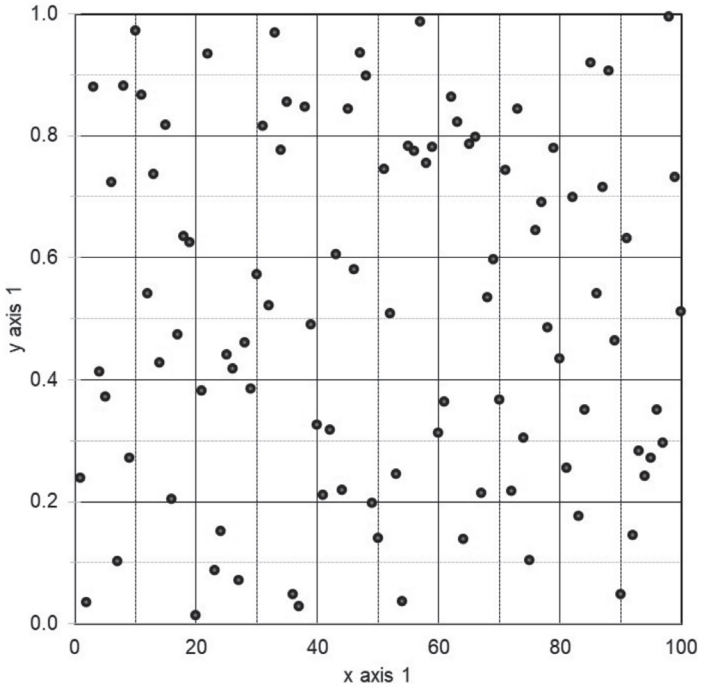


**Figure 1A.** Plot of *p*-values (y-axis) obtained chronologically in 100 simulated experiments testing an association between two variables that have a null association.

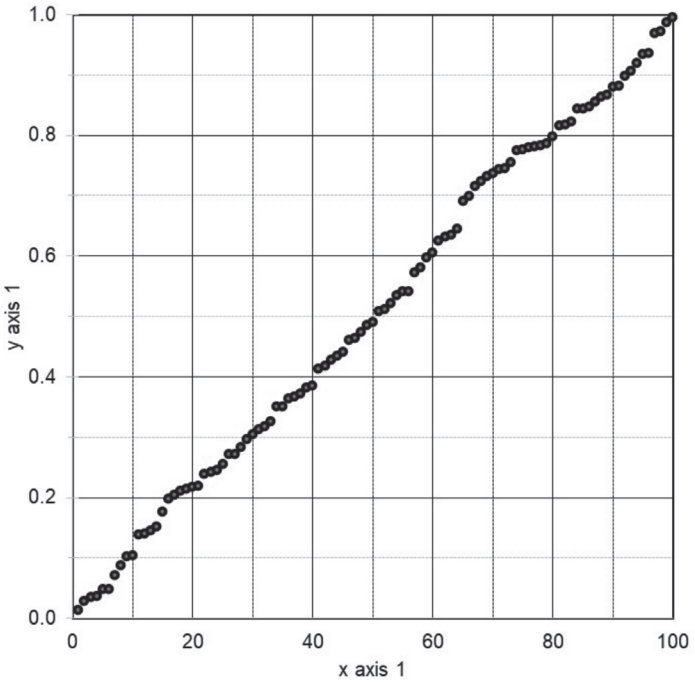


**Figure 1B.** Rank-ordered plot of *p*-values, from smallest to largest, in a series of experiments testing an association between variables that have a null association.

Although not shown here, a plot resembling Figure 1B was obtained in testing a published claim that "women who eat breakfast cereal are more likely to have a boy child." The results of 262 questions (hypothesis tests) asked in a food frequency questionnaire data set plotted as a nearly straight line at a 45-degree angle—implying that the claim is false.[5]

Examples of plots from meta-analysis data sets from literature based on good (reproducible) science claims are shown in Figure 2. The meta-analysis used small observational data sets of selected cancers in petroleum refinery workers.[6] Twelve studies of chronic myeloid leukemia risk (median RR = 1.01) showed a true null (left plot). Ten studies of mesothelioma risk (median RR = 2.47) showed a true association (right plot).

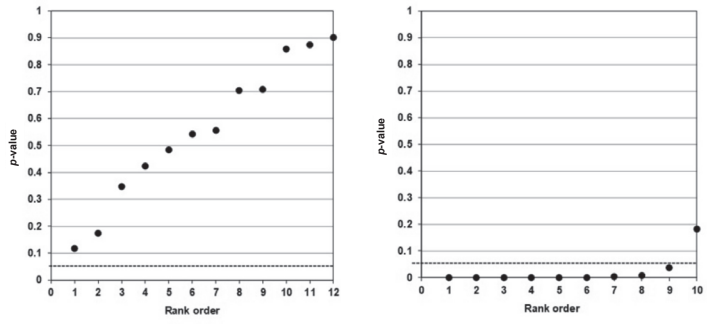


**Figure 2.** *P*-value plots for cancer risk in petroleum refinery workers

Figure 3 shows *p*-value plots for some large data sets.[1] In a meta-analysis of 69 studies of long-term exercise training and morbidity and mortality (left plot), no association was found (median RR = 0.8). In a meta-analysis of 102 studies of smoking and squamous cell carcinoma of the lung (right plot), a clear association was found (median RR = 12.6).

A RR greater than 2 to 3 is generally large enough to rule out bias and confounding. This benchmark—i.e., RR greater than 2 to 3—is an important feature of the Bradford Hill criteria because it supports proof of causation and not random effects.

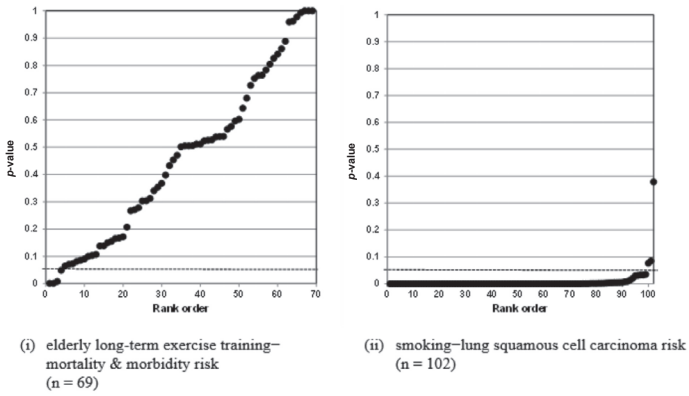


**Figure 3.** *P*-value plots for two large data sets

## Biases Leading to Irreproducibility (Falseness) of Research Claims

Observational studies are affected by many sources of bias leading to false research claims. Publications generally discuss transparent biases, but seldom mention the following hidden biases:

Researchers have enormous flexibility to manipulate their data selection and analysis to get results they want. Examples include selective design, selective use of data, selective analyses, and selective reporting of results to show an effect from a risk factor. With these types of deceptive practices, the effects reported may be nothing more than random (chance) outcomes or due to confounding.

Confirmation bias is the tendency to focus on selectively generating and reporting results in line with expectations or a favored explanation.

Omitted variable bias is purposely excluding true confounding variables in an analysis.

Multiple testing multiple modeling (MTMM) bias involves using data sets with a computer to test multiple outcomes, multiple predictors, different population subgroups, or multiple statistical cause-effect models without statistical correction. This increases the likelihood of making a type I (false positive) error. Observational studies routinely perform MTMM statistical tests on a data set. One-in-twenty results could be "significant" (a false positive) even when the null hypothesis is true. Other terms for this type of bias in published literature are data dredging, fishing expeditions, multiplicity, or multiple comparisons.

*P*-hacking means analyzing data sets using multiple testing to achieve outcomes that fit the desires of researchers, allowing them to find nominally statistically significant results even when studying a non-existent effect or association.

HARKing—hypothesizing after results are known—is another way to achieve outcomes that fit the expectations or desires of researchers.

Publication bias prevents the publication of studies that do not conform to the preferences of journals and editors. This type of peer review process is compromised as the studies allowed to be published report effects that are unreliable and irreproducible (false).

In observational studies, which are non-randomized, it is far too common that researchers do what they want and then report what they want in a paper submitted for publication. All that is needed is that they carefully write up a "tight scientific story" to support their research claim.

**Estimating the SearchSpace**

Performing many tests on a data set allows researchers to "search out," "select," and "report" partial results to fit a narrative. One component of *Shifting Sands*[1] was to identify and report the typical number of possible statistical hypothesis tests performed in observational studies AND whether they make any corrections for this bias. (They never do.) It is typical in observational studies to perform thousands of statistical hypothesis tests. In essence, these are all "research questions" being asked of a data set.

If a data set contains "O" outcomes and "P" predictors, O × P is the possible number of hypotheses at play. A covariate "C" (e.g., age, weight, height, sex, etc.) can be included as a yes/no adjustment (correction) to see how it may confound each of the O × P hypotheses. The number of possible confounding conditions = $2^C$.

The SearchSpace = possible number of hypotheses (research questions) tested in a study can be approximated = O × P × $2^C$.

**Studies of Small Particulate Matter**

Particles of less than 2.5 microns in diameter suspended in air ($PM_{2.5}$) have been the subject of intense regulatory concern. Much of this arose from the early environmental movement and targeting by the Environmental Protection Agency (EPA) of small particles in outdoor air as harmful, toxic, and lethal to the public. In our first Shifting Sands report, we investigated portions of the questionable research claims underlying the EPA's $PM_{2.5}$ regulations—the regulations based upon research claiming that $PM_{2.5}$ has a deleterious effect on human health.[1]

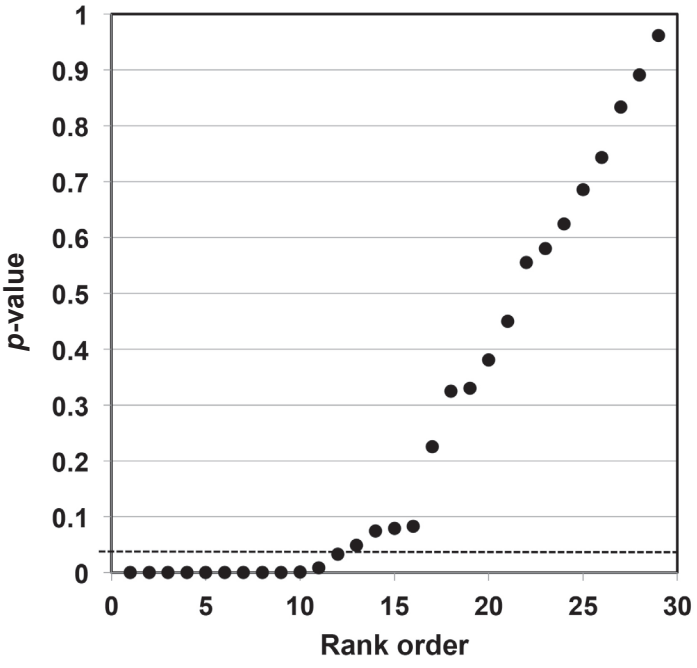


**Figure 4.** *P*-value plot for studies of $PM_{2.5}$ exposure and all-cause mortality

Figure 4 concerns testing a claim that there is "evidence of a positive association between short-term exposure to $PM_{2.5}$ and all-cause mortality." This is from a 2020 meta-analysis of 29 observational studies.[1] The *p*-values present as two lines (bilinear or like a hockey stick). The median RR = 1.02 and is well less than the 2–3 benchmark discussed previously. The pattern clearly departs from the two distinct sample distributions shown earlier, one for true nulls (45-degree line) and the other for true associations (a shallow slope << 1).

A bilinear *p*-value plot strongly suggests a two-component (i.e., dissimilar) mixture of results—some *p*-hacked or manipulated, and some showing a true null or no-effect result. The research claim should be regarded as uncertain (and unproven) because bias cannot be ruled out. The result is consistent with the work of James Enstrom, who found that $PM_{2.5}$ does not cause mortality.[7]

The results of testing a claim of an association between $PM_{2.5}$ and heart attacks are shown in Figure 5. This is from a 2012 meta-analysis of 13 observational studies.[8] The *p*-values again present as bilinear. The median RR for the 13 studies = 1.01 and is well less than the 2–3 benchmark. The bilinear plots suggest two distinct sample distributions. This research claim can be regarded as uncertain (and unproven) because bias cannot be ruled out.

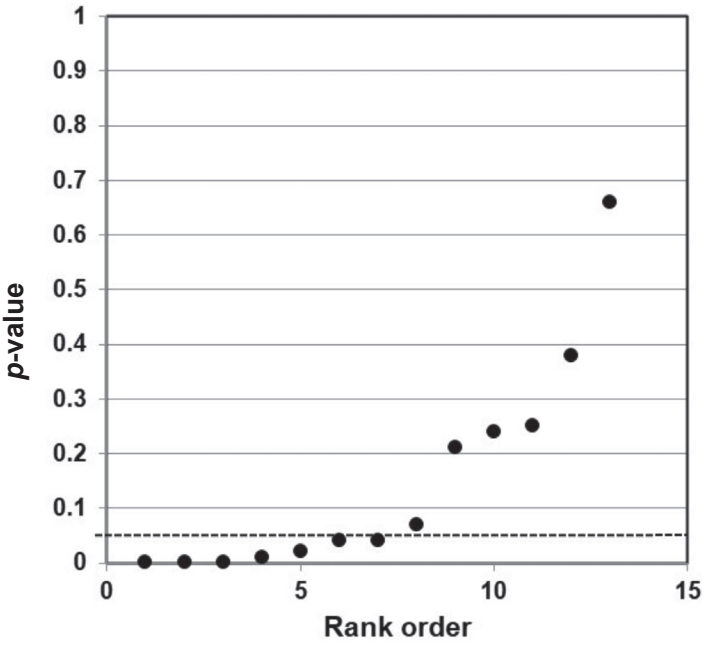


**Figure 5.** *P*-value plot for studies of $PM_{2.5}$ exposure and heart attacks

The results for case-control studies of short-term exposure to $PM_{2.5}$ and asthma attacks are shown in Figure 6. In a 2015 meta-analysis of 37 observational studies,[9] the *p*-values again present as bilinear (most are nulls), with a median OR = 1.02, and is well less than the 2–3 benchmark. Thus, this research claim can be regarded as uncertain and unproven, and bias cannot be ruled out.

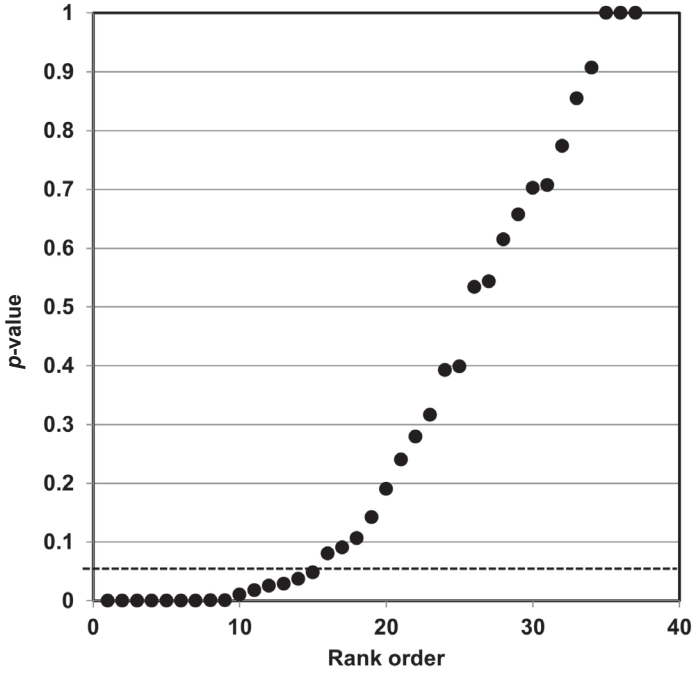


**Figure 6.** *P*-value plot for studies of $PM_{2.5}$ exposure and asthma attacks

Looking at long-term exposure to $PM_{2.5}$ and oxides of nitrogen and development of asthma, Figure 7 displays results of a 2013 meta-analysis of 18 observational studies.[10] The open circles are for $PM_{2.5}$; solid circles are for NOx. The *p*-values again present as bilinear. The median RR for the 18 studies = 1.18 and is well less than the 2–3 benchmark. Again, this research claim cannot be substantiated.

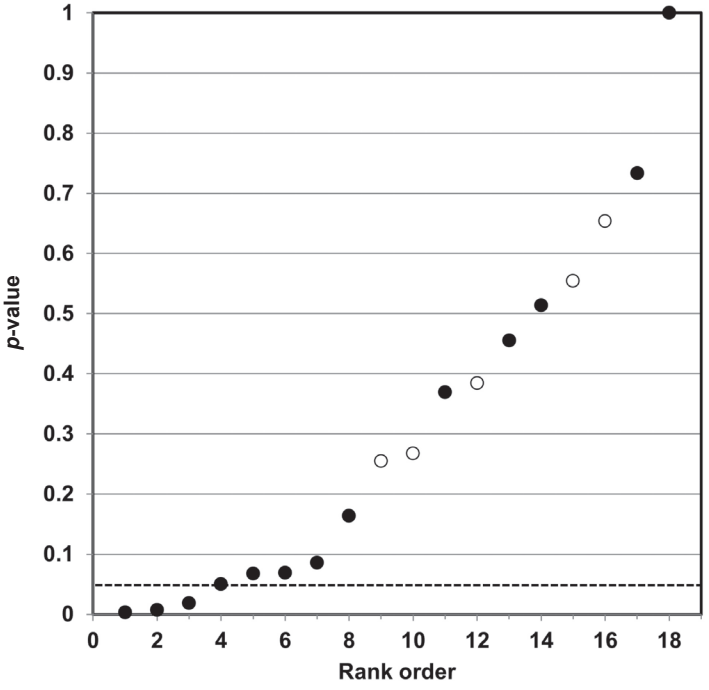


**Figure 7.** *P*-value plot for studies of long-term exposure to $PM_{2.5}$ exposure and development of asthma

To investigate the multiple testing multiple modeling (MTMM) bias, we need to consider the SearchSpace. Three examples are provided: The median estimate of SearchSpace is 12,288 (n=34) for short-term exposure air quality–heart attack studies;[8] 15,360 (n=17) for short-term exposure air quality–asthma attack studies;[9] and 13,824 (n=19) for long-term exposure to $PM_{2.5}$ and oxides of nitrogen.[10]

A typical number of hypothesis tests conducted in these types of studies would be in excess of 12,000. Under ideal conditions, a MTMM analysis is set up to claim "statistical significance" 5 percent of the time. Any researcher having 12,000 research questions expects $0.05 \times 12{,}000 = 600$ statistically significant (but false positive) results from which to choose to write and publish a paper. And that is how irreproducible air quality/health effect studies can be published.

Basically, what can happen here that that researchers use MTMM—perform thousands of hypothesis tests with the data set—and then selectively report only those results that are statistically significant; something that is more easily publishable even though it is likely nothing more than a false-positive finding.

## Nutritional Epidemiology

In our second report, we looked at irreproducible research in the field of nutritional epidemiology, which informs much of the Food and Drug Administration's (FDA) nutrition policy.[11] Cohort studies on diet start with hundreds to thousands of subjects that are followed over an extended period—often many years. Researchers measure subjects' initial attributes, e.g., dietary habits using food frequency questionnaires (FFQs), and then track and collect health results over a period of time. These are important because the FDA uses these results to formulate its recommendations and regulations.

Cohort studies make it easy for researchers to publish multiple papers using the same data set. These studies engage in multiple testing and produce large numbers of false positive results. This provides ample opportunities to fish for spurious (false) correlations that are the product of randomness.

We looked at two types of claims. One is that red meat causes disease. The other is that soy protein is beneficial for heart health.

A food frequency questionnaire lists various foods (e.g., breads or savory biscuits, such as white bread, brown bread, whole-wheat bread, cream crackers, cheese biscuits; cereals such as porridge or corn flakes) and the frequency eaten over a period of time (less than once/month, 1-3/month, once/week, 2-4/week, 4-6/week, once/day, 2-3/day, 4-5/day, 6 or more/day), based on memory.[12] After a lapse of time, the subjects report their health symptoms. The researchers can then make statistical comparisons between different food or beverage items and health symptoms.

We examined a 2019 meta-analysis looking at 30 health outcomes from 105 nutritional observational studies and tested the research claim that "low- or very-low-certainty evidence suggests that dietary patterns with less red and processed meat intake may result in very small reductions in adverse cardiometabolic and cancer outcomes."[13] We tested six health outcomes studied most frequently in the base study papers used in the meta-analysis with *p*-value plots (Figure 8), and we estimated the SearchSpace of the studies used.

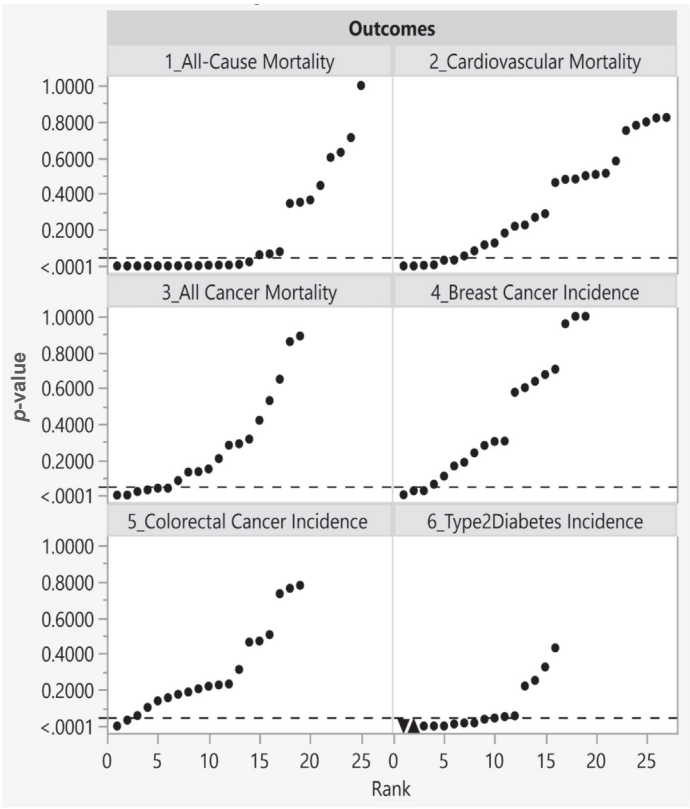


**Figure 8.** *P*-value plots of health outcomes related to intake of red and processed meat

In Figure 8, all-cause mortality (1) and all cancer mortality (3) appear bilinear; hence their red meat research claims can be taken as uncertain and unproven. Cardiovascular mortality (2), breast cancer incidence (4), and colorectal cancer incidence (5) plot as near 45-degree lines, supporting no associations with red meat. One might think Type 2 diabetes incidence (6) shows a real association, as most *p*-values are <0.05. However, the two smallest *p*-values, shown as triangles, have conflicting results. The first, smallest *p*-value, is for a decrease of effect (more red meat leads to lower Type 2 diabetes incidence), and the second smallest and the other *p*-values are for an increase of effect (more red meat leads to more Type 2 diabetes incidence). How can results like this imply a "true association"?

To look at the MTMM bias issue, we randomly selected 15 papers (14%) from the list of 105 studies for SearchSpace determination (number of hypothesis tests) (Table 1). A study with the median number of hypotheses included about 50 foods and three outcomes (diseases) and used 15 foods as predictors and nine covariates in the comparisons. Such a study had more than 20,000 possible hypothesis tests.

Any researcher having this number of research questions expects 0.05 × 20,000 = 1,000 statistically significant (but false positive) randomly occurring results from which to choose to write and publish a paper. That is the way to publish irreproducible red meat/health effect studies. These meta-analysis researchers were honest as they too questioned an association between red meat and the health outcomes they investigated.

**Table 1.** Multiple testing and multiple modeling bias in meta-analysis base study papers on red meat

| *Statistic* | *Foods* | *Outcomes* | *Predictors* | *Covariates* | *SearchSpace* |
|---|---|---|---|---|---|
| minimum | 12 | 1 | 3 | 3 | 176 |
| 25th %ile | 40 | 1 | 8 | 7 | 1,728 |
| *median* | *51* | *3* | *15* | *9* | *20,736* |
| 75th %ile | 129 | 5 | 25 | 11 | 331,776 |
| maximum | 280 | 32 | 51 | 17 | 20,054,016 |
| mean | 93 | 5 | 18 | 9 | 1,451,216 |

Several years ago, the FDA was considering whether to revoke a soy protein heart health claim. A 2019 meta-analysis used FDA's data set of 46 randomized controlled trials (RCTs) looking at soy protein intake and lipid markers (LDL cholesterol and other cholesterol markers) as surrogates for cardiovascular disease (CVD) risk reduction.[14] The study claimed that "soy protein lowers LDL [the bad] cholesterol by a small but significant amount."

These 46 RCTs provided 50 data sets with RRs and CIs for our evaluation. We focused on one aspect of their meta-analysis: LDL cholesterol as a surrogate for CVD risk reduction.[11] We also estimated possible numbers of statistical hypotheses tested in the RCTs (SearchSpace).

The *p*-values in the plot (Figure 9) present as bilinear. Most (38 of 50) are nulls. The median RR = 0.97 and is well less than the 2–3 benchmark. Thus, this research claim can be regarded as uncertain and unproven.

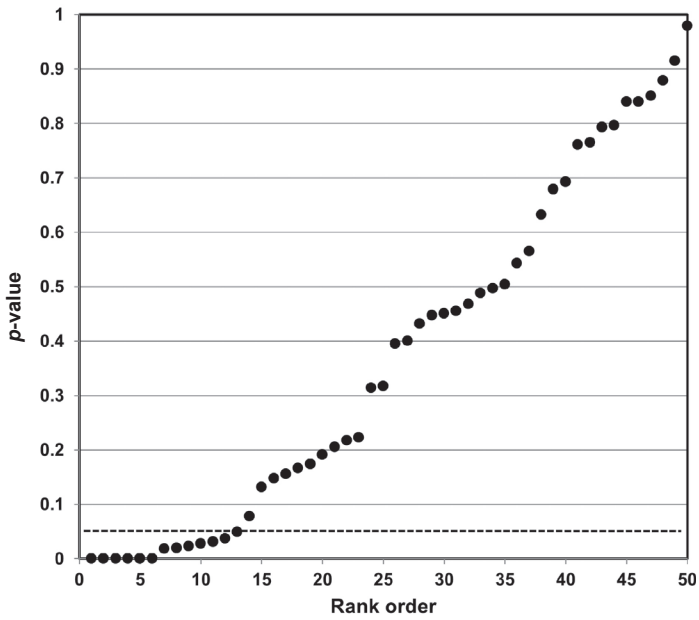


**Figure 9.** *P*-value plots of health outcomes related to intake of red and processed meat

To evaluate the MTMM bias, nine papers from the 46 (20%) referenced by Blanco Mejia et al.[14] were randomly selected for *SearchSpace* counting (Table 2).[11] The median *SearchSpace* was 24—much smaller than for observational studies (~20,000). Researchers who conduct RCTs usually, but not always, pre-select a single outcome variable and a single treatment variable (predictor), and randomization typically leads to fewer covariates. Overall, the weak effect (median RR = 0.97), predominance of null findings (38 *p*-values > 0.05), and the lack of a distinct sample distribution for a true association (shallow

slope << 1) do not support the claim that soy protein lowers LDL cholesterol by a small but significant amount. It appeared that the FDA was on the right track here, but COVID intervened, and no decision was made.

**Table 2.** SearchSpace evaluation for studies on soy protein and cardiovascular disease

| 1st Author | Year | Outcomes | Predictors | Covariates | SearchSpace |
|---|---|---|---|---|---|
| Bakhit | 1994 | 8 | 5 | 3 | 320 |
| Chen | 2006 | 9 | 1 | 4 | 144 |
| Hori | 2001 | 20 | 1 | 0 | 20 |
| Jenkins | 2000 | 14 | 1 | 0 | 14 |
| Ma | 2005 | 20 | 6 | 0 | 120 |
| Mangano | 2013 | 6 | 3 | 0 | 18 |
| Takatsuka | 2000 | 10 | 2 | 0 | 20 |
| Van Horn | 2001 | 3 | 4 | 1 | 24 |
| Wong | 1998 | 14 | 4 | 3 | 448 |

## COVID Non-pharmaceutical Interventions

In our third report, we looked at scientific research behind the public health bureaucrats' mishandling of the federal government's response to the COVID-19 pandemic.[15] Many research studies have been published concerning the response to COVID, but there are concerns about their reproducibility. We reviewed meta-analyses of lockdowns and masking.

A 2022 meta-analysis led by Steve Hanke at the Johns Hopkins Institute for Applied Economics[16] triggered enormous controversy because its finding of no benefit did not fit the lockdown narrative being promoted by public health bureaucrats. A *p*-value plot of statistics from 20 data sets is shown in Figure 10. The *p*-values comprise mostly null associations (14), with 5 or 6 possible non-null associations, one of which could be a chance, false positive association. The plot best fits a sample distribution for a true null association between two variables. This is consistent with Hanke's research claim of no benefit.

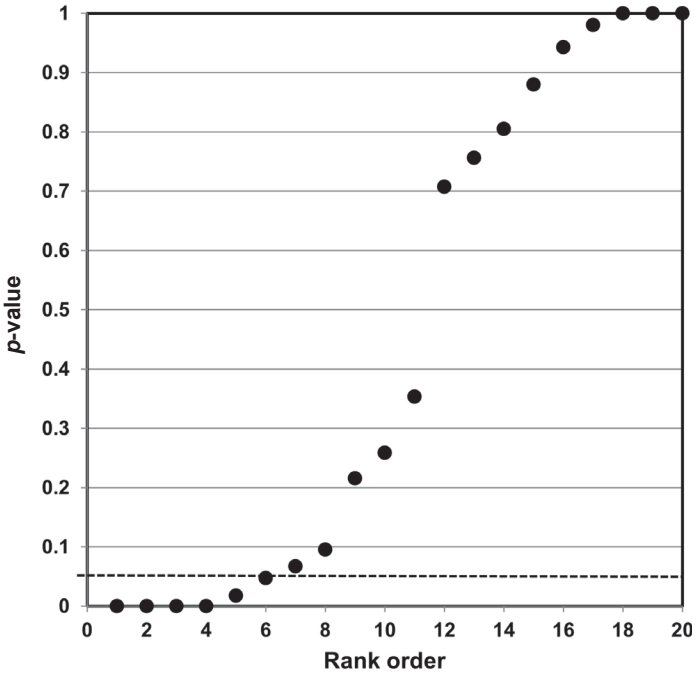


**Figure 10.** *P*-value plot of effect of COVID lockdowns and mortality

The effect of lockdowns on domestic violence was examined in a 2021 meta-analysis by Piquero et al., who concluded the results showed a "moderate to strong increase in domestic violence incidents between pre- and post-lockdown periods."[17]

Our *p*-value plot (Figure 11) is consistent with the expected behavior of a distinct sample distribution for a true effect. The plot has 13 of 17 *p*-values <0.05 (although not shown, 8 < 0.001), and only 4 null associations. It supports the researchers' claim. It is also consistent with the extensive literature on the association of isolation with domestic violence.

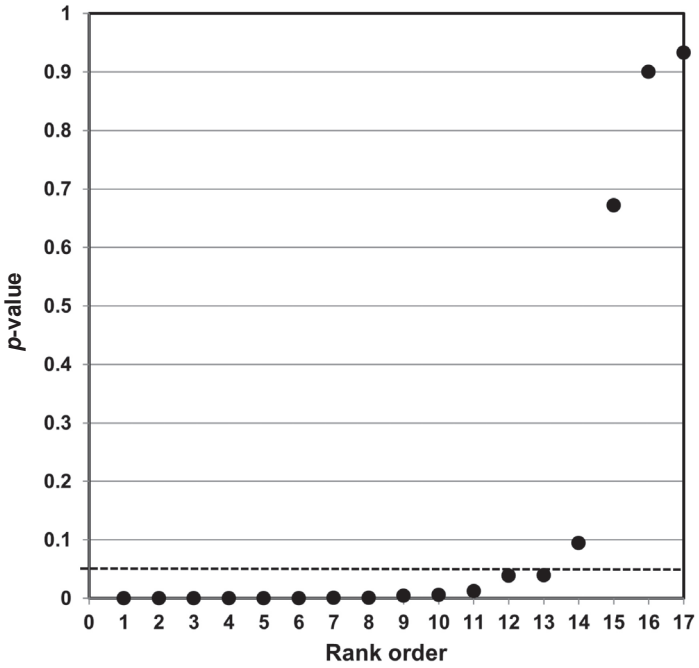


**Figure 11.** *P*-value plot of association of lockdowns and domestic violence

Medical mask use in public was being promoted by public health bureaucrats in response to COVID-19. Concerning masks, we evaluated meta-analyses and systematic reviews.[18] We did three literature searches, specifically looking at use of medical masks for viral infections over the period Jan 1, 2020, to Dec 7, 2022, a 35-month period. These included the Cochrane Review library, a medical literature search through the National Institutes of Health PubMed data base, and a general internet search. There are a large number of observational studies, which claim that masks work, but we disregarded them because of excessive bias.[18]

We looked only at meta-analyses of randomized trials (RCTs). We also looked closely at the studies that were combined in the meta-analysis to establish whether they included self-reported information of viral symptoms. Such information is very unreliable.[18] We eliminated such studies. We also rejected meta-analyses that had too heavy reliance on self-reported symptoms.

Medical masks were never intended to stop respiratory infections.[18] They are ineffective at blocking the aerosols that contain the virus. Plus, within 30 to 60 minutes one exhales so many aerosols that the mask is plugged up, so that air uptake from above, sides, and below is increased.

A 2020 meta-analysis by Cochrane Review researchers met

our selection criteria.[19] The *p*-value plot for this meta-analysis (Figure 12) clearly shows results aligned at a 45-degree slope, agreeing with the researchers' claim of a null association (no benefit to using medical masks in public).

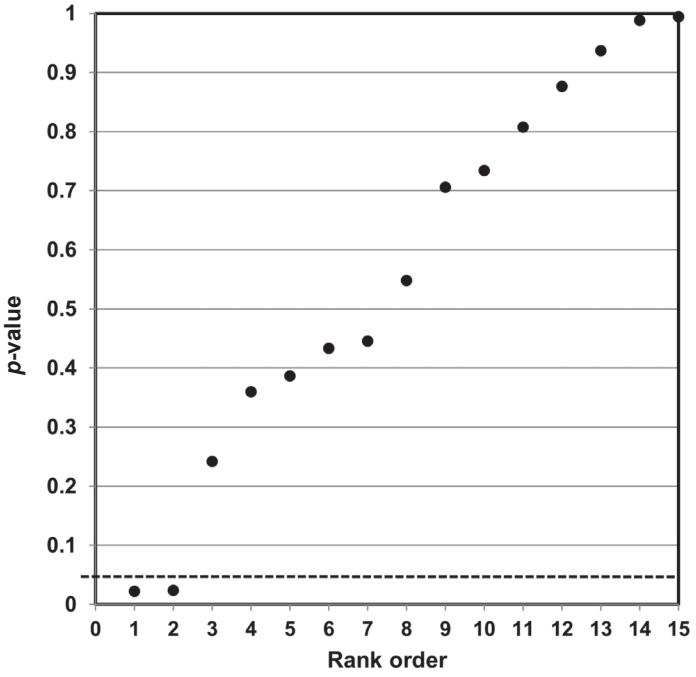


**Figure 12.** *P*-value plot of Cochrane Review's 2020 meta-analysis on masking and respiratory illness

Another study, which got little attention, was a systematic review published by the Cato Institute.[20] This was retrieved by an internet search from what is called the "gray literature." The plot clearly exhibits *p*-values aligned at 45 degrees (Figure 13), in agreement with the research claim of no benefit to using medical masks in public.

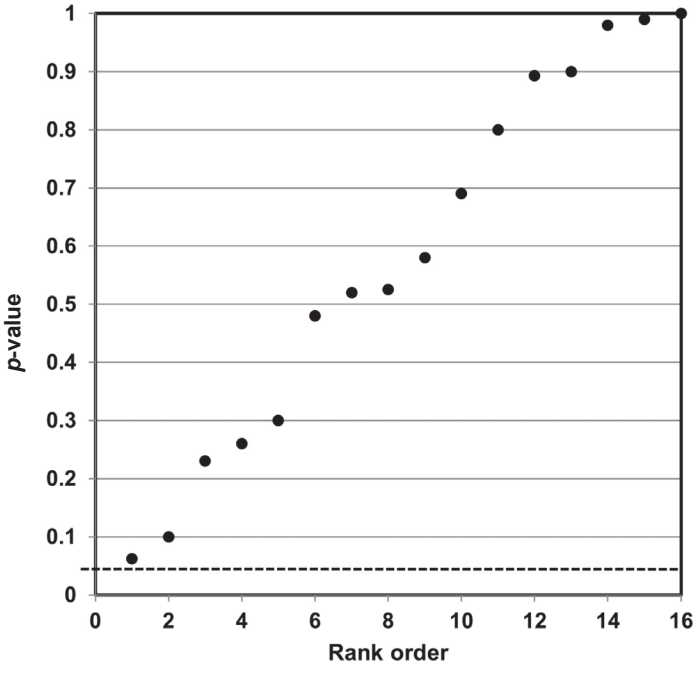


**Figure 13.** *P*-value plot on the relationship between masking and respiratory illness from Cato Institute systematic review.

A summary of the results of our reviews of meta-analyses (MA) and systematic reviews (SR)[20] is in Table 3. Those labeled "unable to determine"[21,22] in the fourth column relied heavily on self-reported information. In the two studies labeled "no,"[23,24] we were not able to reproduce the researchers' claims that masks work. In two studies from the general medical literature,[25,26] we were able to confirm the researchers' conclusions of no benefit to using medical masks in public.

**Table 3.** Comparison of meta-analysis claims to *p*-value plot results

| *Study detail* | *Research claim* | *Finding of p-value plot* | *Research claim supported?* |
|---|---|---|---|
| *Cochrane review literature*: | | | |
| Jefferson et al. (2020) MA | no significant benefit | null (no) effect | yes |
| *Medical research literature*: | | | |
| Aggarwal et al. (2020) MA | no significant benefit | insufficient data to examine | unable to determine |
| Xiao et al. (2020) MA | " | null effect | yes |
| Nanda et al. (2021) MA | " | null effect | yes |
| Tran et al. (2021) MA | benefit to mask use | finding is uncertain | no |
| Kim et al. (2022) MA | " | null effect | no |
| Ollila et al. (2022) MA | " | insufficient data to examine | unable to determine |
| *Gray literature*: | | | |
| Liu et al. (2021) SR | no significant benefit | null effect | yes |

## Race and Gender Implicit Bias

In our fourth report, we looked at the reproducibility of science behind the implicit bias theory.[27] Evaluating studies of bias is highly challenging compared to other areas. There are three ways to characterize bias in the ways groups of people look at a topic, such as wealth, race, or gender. One way is to observe groups of people and their behavior (i.e., real-world observations). The second is explicit bias, which is characterized by surveys and questionnaires taken by groups of people. The third way is implicit or subconscious bias—implicit bias.

Implicit bias is established by a computer-based test referred to as the Implicit Association Test (or IAT) developed by researchers at Harvard.[28] It is a speed response test. This is extremely important as it forms the scientific framework of Diversity, Equity, and Inclusion (DEI). Researchers attempt to show that implicit bias is real by correlating it with real-world observations. There is, however, a lot of science, which is not getting much attention, showing that implicit bias is not correlated with behavior and may not be meaningful at all.[29] We examined racial bias and gender bias research claims in our fourth report.

With regard to racial implicit bias, we looked at a meta-analysis examining whether results from the IAT correlates with real-world microbehaviors, that is, measures of verbal and subtle verbal behavior. The plot in Figure 14[30] clearly departs from one that would show a true association or effect. It shows 83 data sets. The 50 represented by circles have positive associations between the IAT test score and real-world measures, and 30 represented by downward triangles had negative associations. This meta-analysis was done by Philip Tetlock's group at the University of Pennsylvania, very capable researchers, who questioned implicit bias as being a reliable measure of racial bias.[31] Our plot confirmed that IAT results were poor predictors of microbehaviors.

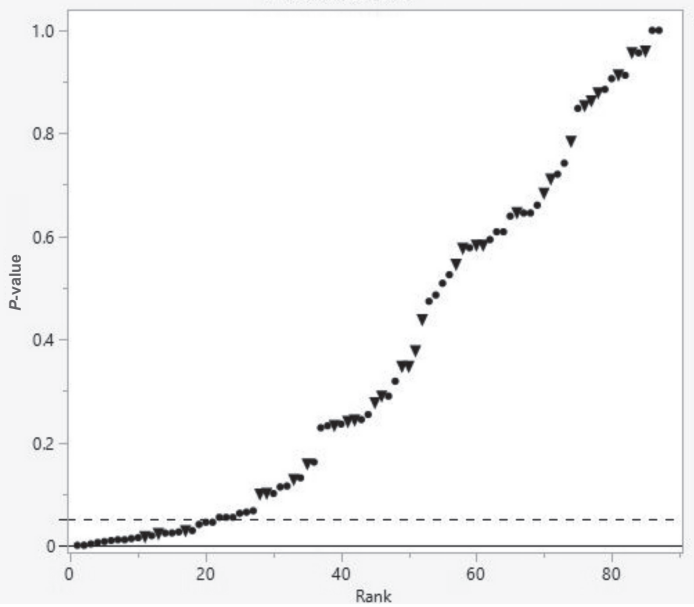


**Figure 14.** *P*-value plot testing association of racial implicit bias and real-world microbehaviors

The gender Implicit Association Test (or gIAT) is used to measure gender implicit bias. In our report,[27] we examined a 2019 meta-analysis inferring implicit bias of males towards females published by Benedek Kurdi's group out of Harvard.[32,33] Gender bias as measured by the gIAT refers to gender attitude, gender stereotypes, and gender identity. Criterion (real-world) measures of female and male behavior in Kurdi's meta-analysis included policy preference expressions, resource allocation, academic performance, subtle nonverbal behaviors, performance on interference tasks like the Stroop task, and criminal sentencing decisions.

There were 27 studies in the Kurdi meta-analysis that correlated gIAT test results with criterion measures. The 27 correlations plotted in Figure 15 show an irregular pattern, and

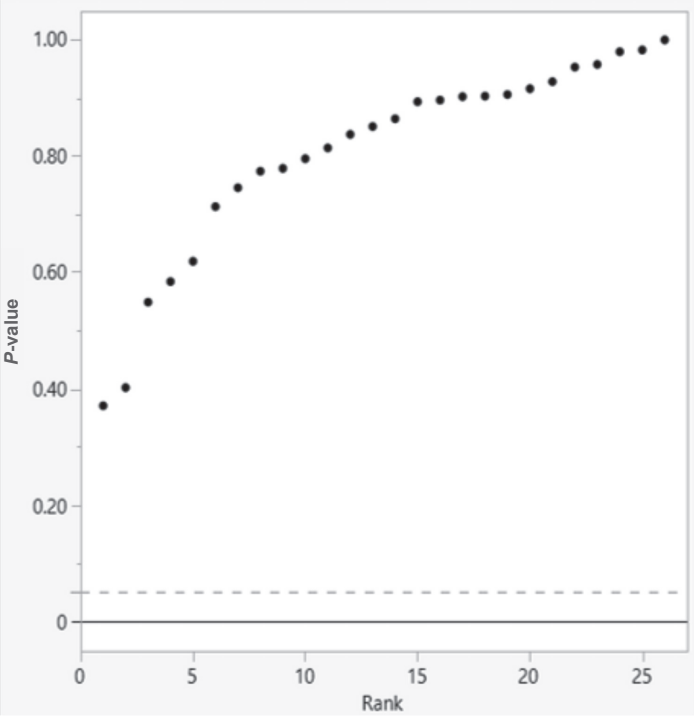


**Figure 15.** *P*-value plot testing correlation of implicit gender with criterion measurements

all *p*-values are insignificant (i.e., > .05). Given these results, we made numerous email attempts to contact Kurdi, but he was non-responsive to our attempts to discuss these results. Their research claim that measures of gender implicit bias correlate with criterion measures is unproven.

## Conclusions

Bad science has crowded out good science in the literature. Bad science permits governments and policymakers to develop policies and regulations in many areas—including the environment, health, and the workplace—with no public benefit or with actual harm.

Our work with the National Association Scholars Shifting Sands Project using *P*-value plots and SearchSpace counting are practical ways to help expose the statistical shortcomings and irreproducibility (falseness) of bad science, and also to identify and support good science.

**S. Stanley Young, Ph.D.,** is CEO of CGStat in Raleigh, NC, and Director of the National Association of Scholars' Shifting Sands Project. **Warren B. Kindzierski, Ph.D.,** now retired, was a professor of public health at the University of Alberta, Edmonton, Alberta. Contact: Warren B. Kindzierski, email: warrenk@ualberta.ca.

**Acknowledgment:** This paper is based on a talk at the 42nd annual meeting of Doctors for Disaster Preparedness in El Paso, Texas, July 7, 2024. Videorecording available at: https://www.youtube.com/watch?v=78sTkKrJ0bA&t=2877s.

## REFERENCES

1. Young SS, Kindzierski W, Randall D. Shifting Sands: Unsound Science and Unsafe Regulations. Report #1: Keeping Count of Government Science: *P*-Value Plotting, *P*-Hacking, and PM2.5 Regulation. National Association of Scholars; 2021. Available at: https://www.nas.org/storage/app/media/Reports/Shifting%20Sands%20Report%20I/Shifting_Sands_pt1.pdf. Accessed Oct 14, 2024.
2. Smaldino PE, McElreath R. The natural selection of bad science. *R Soc Open Sci* 2016;3(9):160384. Available at: https://doi.org/10.1098/rsos.160384. Accessed Oct 28, 2024.
3. National Association of Scholars (NAS). Shifting Sands: Keeping Count of Government Science Series; 2024. Available at: https://www.nas.org/report-series/shifting-sands-keeping-count-of-government-science. Accessed Oct 26, 2024.
4. Hung HMJ, O'Neill RT, Bauer P, Kohne K. The behavior of the *p*-value when the alternative hypothesis is true. *Biometrics* 1997;53:11-22. Available at: https://doi.org/10.2307/2533093. Accessed Nov 13, 2024.
5. Young SS, Bang H, Oktay K. Cereal-induced gender selection? Most likely a multiple testing false positive. *Proc R Soc B: Bio Sci* 2009;276(1660):1211-1212. Available at: https://doi.org/10.1098/rspb.2008.1405. Accessed Oct 28, 2024.
6. Schnatter AR, Chen M, DeVilbiss EA, Lewis RJ, Gallagher EM. Systematic review and meta-analysis of selected cancers in petroleum refinery workers. *J Occup Environ Med* 2018;60(7):e329-e342. Available at: https://doi.org/10.1097/JOM.0000000000001336. Accessed Oct 29, 2024.
7. Enstrom JE. Scientific distortions in fine particulate matter epidemiology. *J Am Phys Surg* 2018;23:15-19.
8. Young SS, Kindzierski WB. Evaluation of a meta-analysis of air quality and heart attacks, a case study. *Crit Rev Toxicol* 2019;49(1):85–94. Available at: https://doi.org/10.1080/10408444.2019.1576587. Accessed Oct 19, 2024.
9. 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(3):173-196. Available at: https://doi.org/10.3390/jor1030017. Accessed Oct 19, 2024.
10. Young SS, Cheng K-C, Chen JH, Chen S-C, Kindzierski WB. Reliability of meta-analysis of an association between ambient air quality and development of asthma later in life. *IJSP* 2022;11(2):61-76. Available at: https://doi.org/10.5539/ijsp.v11n2p61. Accessed Oct 29, 2024.
11. Young SS, Kindzierski W, Randall D. Shifting Sands: Unsound Science and Unsafe Regulations. Report #2: Flimsy Food Findings, Food Frequency Questionnaires, False Positives, and Fallacious Procedures in Nutritional Epidemiology. National Association of Scholars; 2022. Available at: https://www.nas.org/reports/shifting-sands-report-ii?f=true. Accessed Oct 26, 2024.

12. Mulligan AA, Luben RN, Bhaniani A, et al. A new tool for converting food frequency questionnaire data into nutrient and food group values: FETA research methods and availability. *BMJ Open* 2014;4:e004503. Available at: https://doi.org/10.1136/bmjopen-2013-004503. Accessed Oct 29, 2024.
13. Young SS, Kindzierski WB. Statistical reliability of a diet-disease association meta-analysis. *IJSP* 2022;11(3):40-50. Available at: https://doi.org/10.5539/ijsp.v11n3p40. Accessed Oct 29, 2024.
14. Blanco Mejia S, Messina M, Li SS, et al. A meta-analysis of 46 studies identified by the FDA demonstrates that soy protein decreases circulating LDL and 9 total cholesterol concentrations in adults *J Nutrition* 2019;149(6): 968–981. Available at: https://doi.org/10.1093/jn/nxz020. Accessed Oct 19, 2024.
15. Young SS, Kindzierski W, Randall D. Shifting Sands: Unsound Science and Unsafe Regulations. Report #3: The Confounded Errors of Public Health Policy Response to the COVID-19 Pandemic. National Association of Scholars; 2023. Available at: https://www.nas.org/reports/shifting-sands-report-iii?f=true. Accessed Oct 26, 2024.
16. Herby J, Jonung L, Hanke SH. A literature review and meta-analysis of the effects of lockdowns on COVID-19 mortality. Studies in Applied Economics No. 200. Johns Hopkins Institute for Applied Economics; January 2022. Available at: https://sites.krieger.jhu.edu/iae/files/2022/01/A-Literature-Review-and-Meta-Analysis-of-the-Effects-of-Lockdowns-on-COVID-19-Mortality.pdf. Accessed Oct 20, 2024.
17. Piquero AR, Jennings WG, Jemison E, Kaukinen C, Knaul FM. Domestic violence during the COVID-19 pandemic: evidence from a systematic review and meta-analysis. *J Criminal Justice* 2021;74:101806. Available at: https://doi.org/10.1016/j.jcrimjus.2021.101806. Accessed Oct 29, 2024.
18. Young SS, Kindzierski WN. Statistical reproducibility of meta-analysis for medical mask use in community settings to prevent airborne respiratory virus infection. *IJSP* 2023;12(4):1-18. Available at: https://doi.org/10.5539/ijsp.v12n4p1. Accessed Oct 20, 2024.
19. Jefferson T, Del Mar CB, Dooley L, et al. Physical interventions to interrupt or reduce the spread of respiratory viruses. Cochrane Database of Systematic Reviews 2020;11(11): CD006207. Available at: http://doi.org/10.1002/14651858.CD006207.pub5. Accessed Oct 20, 2024.
20. Liu IT, Prasad V, Darrow JJ. Evidence for community cloth face masking to limit the spread of SARS-CoV-2: A critical review. Cato Working Paper No. 64. Washington, D.C.: Cato Institute; 2021. Available at: https://www.cato.org/sites/cato.org/files/2021-11/working-paper-64.pdf. Accessed Oct 21, 2024.
21. Aggarwal N, Dwarakanathan V, Gautam N, et al. Facemasks for prevention of viral respiratory infections in community settings: A systematic review and meta-analysis. *Indian J Pub Health* 2020;64(S192–S200). Available at: https://doi.org/10.4103/ijph.IJPH_470_20. Accessed Oct 21, 2024.
22. Ollila HM, Partinen M, Koskela J, et al. (2022). Face masks to prevent transmission of respiratory infections: Systematic review and meta-analysis of randomized controlled trials on face mask use. *PLoS One* 2022;17(12): e0271517. Available at: https://doi.org/10.1371/journal.pone.0271517. Accessed Oct 21, 2024.
23. Tran TQ, Mostafa EM, Tawfik GM, et al. Efficacy of face masks against respiratory infectious diseases: a systematic review and network analysis of randomized-controlled trials. *J Breath Res* 2021;15:047102. Available at: https://doi.org/10.1088/1752-7163/ac1ea5. Accessed Oct 21, 2024.
24. Kim MS, Seong D, Li H, et al. Comparative effectiveness of N95, surgical or medical, and non-medical facemasks in protection against respiratory virus infection: a systematic review and network meta-analysis. *Rev Med Virol* 2022;32(5):e2336. Available at: https://doi.org/10.1002/rmv.2336. Accessed Oct 21, 2024.
25. Xiao J, Shiu EYC, Gao H, et al. Nonpharmaceutical measures for pandemic influenza in nonhealthcare settings—personal protective and environmental measures. *Emerg Infect Dis* 2020;26(5):967–975. Available at: https://doi.org/10.3201/eid2605.190994. Accessed Oct 21, 2024.
26. Nanda A, Hung I, Kwong A, et al. Efficacy of surgical masks or cloth masks in the prevention of viral transmission: Systematic review, meta-analysis, and proposal for future trial. *J Evid Based Med* 2021;14(2):97–111. Available at: https://doi.org/10.1111/jebm.12424. Accessed Oct 21, 2024.
27. Young SS, Kindzierski W, Randall D. Shifting Sands: Unsound Science and Unsafe Regulations. Report #4: Zombie Psychology, Implicit Bias Theory, and the Implicit Association Test. National Association of Scholars; 2024. Available at: https://www.nas.org/reports/shifting-sands-report-iv?f=true. Accessed Oct 26, 2024.
28. Greenwald AG, McGhee DE, Schwartz JLK. Measuring individual differences in implicit cognition: The implicit association test. *J Per Soc Psychol* 1998;74:1464-1480. Available at: https://doi.org/10.1037//0022-3514.74.6.1464. Accessed Oct 28, 2024.
29. Jussim L, Thulin E, Fish J, Wright JD. Articles critical of the IAT and Implicit Bias. Open Science Foundation. OSFHome; 2023. Available at: https://osf.io/74whk/. Accessed Oct 26, 2024.
30. Young SS, Kindzierski WB. Reproducibility of Implicit Association Test (IAT)—Case study of meta-analysis of racial bias research claims. *IJSP* 2024;13(2):33-45. Available at: https://ccsenet.org/journal/index.php/ijsp/article/view/0/50266. Accessed Oct 29, 2024.
31. Oswald F, Mitchell G, Blanton H, Jaccard J, Tetlock PE. Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *J Personality Social Psychol* 2013;105:171-192. Available at: https://doi.org/10.1037/a0032734. Accessed Oct 22, 2024.
32. Kurdi B, Seitchik AE, Axt JR, et al. Relationship between the Implicit Association Test and intergroup behavior: a meta-analysis. *Am Psychologist* 2019;74(5):569–586. Available at: http://doi.org/10.1037/amp0000364. Accessed Oct 22, 2024.
33. Kurdi B, Banaji MR. Relationship between the Implicit Association Test and explicit measures of intergroup cognition: Data from the meta-analysis by Kurdi et al. (2018). Semantic Scholar; Mar 18, 2019. Available at: https://doi.org/10.17605/OSF.IO/K7QX5. Accessed Oct 22, 2024.

**"INFORMATION HEALS"**

*neutralresearcher.substack.com*

A digital database
for physicians and patients
containing information on
***hard to find and***
***difficult to debate***
**controversial medical topics.**

***Misinformation is a disease.***
***Information is a cure.***

**SCAN ME**