Statistical Significance, Narrative, and the Scholastic Fallacy: How Ritualized Statistics Exaggerate Social Science Theories

Dianshi Moses Li[1], Hongyang Leon Zhang[2], Qinru Ruby Ju[1*]

1 Centre for Empirical Legal Studies; Faculty of Law; University of Macau; Macau SAR, 999078, China.

2 School of Criminal Justice, Zhongnan University of Economics and Law; Wuhan, 430073, China.

*Corresponding author. Email: qinrujv@126.com

Keywords

statistical significance; scholastic fallacy; theory fetishism, Null Hypothesis Significance Testing, importance scores, effect size, type of causes, causal inference

Abstract

The social sciences are facing a credibility crisis driven by two interrelated tendencies: an overreliance on abstract theorizing ("theory fetishism") and the ritualistic application of Null Hypothesis Significance Testing (NHST), which often lends false authority to inconsequential p-values. Through a satirical 'Health Communication Criminology' thought experiment and real-world examples from psychology, nutrition, and criminology, we illustrate how theory fetishism encourages post hoc rationalizations, while NHST, model fit indices, and machine-learning-derived 'importance scores' create an illusion of certainty. These practices routinely exaggerate the significance of findings while accounting for minimal variance at the population level. To address these issues, we propose two correctives: (1) integrating effect size and prevalence using population attributable fractions to better assess societal impact; and (2) interpreting the proportion-mediated statistic (T_p) as an upper limit—not definitive evidence—of mediation. We further identify common analytical pitfalls, such as comparing coefficients across different models, indiscriminate use of matching and mediation analyses, and conflating interpretability with causality. A path forward requires rigorous causal designs, transparent reporting of uncertainty, and a commitment to epistemic humility—essential steps toward improving replicability and generating evidence that is truly meaningful for policy and real-world outcomes.

Research Article

Submitted: 1 May 2025 / Accepted: 3 May 2025 / Published online: 3 May 2025
Trans. Soc. 2025. 1(2): 39–62
https://doi.org/10.63336/TransSoc.28
Online ISSN: 3079-8310

1. Introduction

Imagine a scholarly conference where a new discipline called "health communication criminology" (HCC) is the hot topic. In one presentation, a theorist passionately explains how public health messaging campaigns inadvertently control crime waves through a complex chain of psychological reactions. He claims that HCC is grounded in the logic of welfare economics, viewing public safety as a public good and crime as a problem of negative externalities. According to classical welfare economics, such externalities lead to suboptimal outcomes for society unless they are somehow internalized by the actor causing the harm (Coase, 1960). HCC proposes to internalize these external costs through targeted health communication aimed at potential offenders. The idea is that by emphasizing the personal health risks of criminal behaviors, most pointedly by highlighting how drug use damages one's health, those at risk of offending might recalibrate their decisions. If a drug user, for instance, becomes acutely aware that using drugs will harm their health, they may factor this personal cost into their decision-making. Hence, the crime rate triggered by drug use would be controlled on a social level. The audience nods at the elegant theory. The audience readily embraces the elegance of this model. Next, the presenter unveils a regression table—tiny coefficients marked with asterisks for $p < .05$—and proclaims the grand theory empirically validated. Enthralled by the compelling narrative, the room erupts in applause.

This scenario sounds farcical, but it reflects real tendencies in social science. Researchers too often fall in love with theory-driven storytelling, crafting elaborate narratives and confirming them with dubious statistical methods. What looks like logical rigor on paper turns out to be harshly transgressive in practice. The HCC thought experiment thus exposes a broader pathology of social science: the temptation to theories for the sake of theorizing, elevating conceptual neatness above lived reality. It is a caricature, of course—no real scholars advocate HCC as a discipline—but its very extravagance highlights a real intellectual hazard.

This hazard has been described by Pierre Bourdieu as the "scholastic fallacy." Bourdieu notes that scholars, by operating in sheltered intellectual environments, often mistake their abstract models for how the world actually works (Bourdieu, 1990, 2000). In academia's leisure of thought, one may assume that social agents have the same capacity to analyze and follow theoretical logic as researchers. This fallacy "induces [one] to think that agents involved in action, in practice, in life, think, know and see as someone who has the leisure to think" (Bourdieu, 1990). Such a scholastic illusion assumes that society can be engineered to conform with surgical precision to an elegant model of incentives and deterrence. In reality, it cannot—and attempts to do so are not only destined to fail but may also cause harm, as policies born of armchair logic often misfire amid the complex dynamics of real communities.

Crucially, the scholastic fallacy is not unique to our fictitious HCC. It is a pervasive risk across the social sciences wherever theory outruns evidence. Grand theoretical frameworks in sociology, economics, psychology, and beyond can foster an excess of theoretical consistency that is achieved only by overlooking the gap between analytical assumptions and on-the-ground realities. The danger is that researchers obsessed with a particular theory may cherry-pick or distort empirical observations to fit the model, rather than allowing observations to challenge the model's premises. In extreme cases, entire academic subfields can drift into self-referential exercises, polishing the wheels of theory even as they lose traction with reality. The widespread nature of this tendency demands a rigorous reexamination of our research ethos. Are we privileging theoretical romance at the expense of practical relevance? The irony inherent in HCC forces us to confront these uncomfortable questions. In the sections that follow, we explore how a fetishization of theory fuels these scholarly temptations and how dubious statistical practices cloak them in a veneer of scientific credibility.

2. Theory Fetishism and the Scholastic Fallacy

Social scientists have long been obsessed with theory. The ability to develop comprehensive explanations of human behavior and society has often been seen as the pinnacle of academic achievement. This cultural bias elevates theory to an almost sacrosanct status—the fetishism of theory—where developing or adding to "the theory" becomes an end in itself. In many disciplines, top journals and tenure committees send an implicit message: without a novel theoretical contribution, empirical research is of limited value (Hambrick, 2007). The result is that the academic ecosystem will reward complexity over clarity and novelty over accuracy. Researchers may feel pressured to frame their findings within grand theoretical narratives, even if those narratives do not match current data or phenomenon (Duffee et al., 2015). In fact, many scholars have criticized grand theories, general theories and theory-centrism (Duffee et al., 2015; Hambrick, 2007; Howard & Freilich, 2007; Mills, 2000; X. Zhao et al., 2024). As management scholar Donald Hambrick has argued, an "undue emphasis on the development of theory" can come at the expense of real-world insight (Hambrick, 2007). In other words, strong theory has sometimes eclipsed practical significance as the coin of the realm in research excellence (Hambrick, 2007).

This theory-driven proclivity is often well-intentioned. Theories, after all, are meant to generalize knowledge and deepen understanding; a good theory can unify isolated facts into a coherent picture. The danger lies in what we might call the over-romanticization of theory—a belief that elegant theory reveals truth with a capital "T", and that pursuing theoretical beauty is akin to pursuing scientific virtue. Under this spell, scholars might disregard evidence that doesn't fit, or risk further abstraction in search of theoretical perfection. C. Wright Mills warned of this in his critique of "Grand Theory," describing how mid-20th-century sociology became

enamored with highly abstract formulations (Mills, 2000). Such Grand Theories, Mills argued, floated in a conceptual stratosphere, impressively intricate yet disconnected from the urgent "troubles" of people's everyday lives (Mills, 2000). Mills's admonition was that theory should be a tool, not an idol to understand society, not a self-justifying intellectual exercise.

When theory becomes an idol, the Scholastic Fallacy is its ever-faithful attendant. The fetishization of theory creates conditions ripe for this fallacy to flourish. Scholars so wedded to a particular theoretical lens may unconsciously project its assumptions onto their subjects, presuming a coherence and applicability that reality often fails to uphold. Bourdieu's critique goes straight to the heart of this problem: academics can forget that their research subjects do not live in a world of stylized variables and ceteris paribus conditions (Bourdieu, 2000). Actual human beings juggle multiple, often conflicting pressures; they make decisions in contexts of uncertainty, emotion, and constraint—contexts that theory can only partially capture. However, a theorist insulated in academia might presume agents follow the logic that their theory stipulates. Our HCC vignette satirizes this tendency: its architects treat ordinary citizens as if they were emotionless, textbook utility maximizers—automatically adjusting their behavior to punitive health laws. It is a form of intellectual hubris, or as Bourdieu noted, a kind of epistemic ethnocentrism—viewing the social world through the peculiar values and dispositions of scholars (Bourdieu, 2000).

A key consequence of theory fetishism is that empirical research becomes subservient to theory rather than its judge. Rather than letting theory emerge from the data, researchers too often prune or tweak observations to fit the theory. This theory-first mindset creates blind spots, leading scholars to overlook phenomena that are important but theoretically inconvenient—a failure of social science to matter in practice, as Flyvbjerg (2001) warns when he contrasts lofty "episteme" with grounded "phronesis." Consider, for example, how some experimental psychology studies have been faulted in hindsight for building flashy theories on narrow lab findings—only to see them collapse when tested on broader samples or in real-world settings. The disconnect between theory and empirical reality thus manifests both subtly and overtly. Practitioners and policymakers, encountering these "elegant irrelevances," often lament ivory-tower research that, however conceptually sophisticated, proves impractical for addressing real problems (Flyvbjerg, 2001; Hambrick, 2007).

Returning to our HCC parable, we see these problems magnified: the fictional scholars weave health-communication and criminological theories so seamlessly that their models appear flawless on paper. Their zeal for theory drives them to absurd extremes, blind to its ethical and practical shortcomings—a hallmark of the scholastic fallacy. In service of this theoretical "purity," inconvenient realities—poverty, mistrust of authorities, cultural norms, mental-health challenges—are conveniently sidelined. Although satirical, this scenario rings uncomfortably

true: when researchers oversimplify assumptions and idealize conditions, even the most rigorous theory can crumble against the messy complexities of real life.

In sum, theory fetishism and the scholastic fallacy are intertwined challenges. Theory fetishism elevates abstract models to a status of prestige beyond their practical utility, while the scholastic fallacy reflects the cognitive bias born of inhabiting those theories too exclusively. Together, they remind us that theoretical sophistication is no substitute for wisdom. Robust theory remains indispensable for advancing knowledge, but it must stay open to empirical evidence, respectful of the complexity of human experience, and sensitive to context. Social scientists should continually ask themselves: Are we studying the world as it truly is, or only as our theories presume it to be? That distinction can determine whether our research enlightens or misleads. Yet even beyond theory, researchers can become captivated by compelling narratives—so next, we turn to the allure and pitfalls of storytelling in social science.

3. The Allure of Novel Narratives

Social scientists, like everyone else, are storytellers at heart. The allure of a compelling narrative often tempts researchers to cast every finding as part of a sweeping, romantic saga. We yearn for explanations with a clear beginning, middle, and end—a coherent arc that neatly wraps disparate phenomena into a satisfying whole. In academia, this often translates into overvaluing novelty: journals preferentially publish surprising, counter-intuitive results that weave an exciting tale (Nosek et al., 2012; Wang et al., 2017). The pressure to "tell a good story" can be so intense that researchers may unintentionally drift from cautious empiricism into embellished narrative. As Nicholas (2008) warned, with the narrative fallacy, our minds are apt to mistake a compelling story for reality, especially when the story confirms what we want to believe.

One consequence is the rise of Hypothesizing After the Results are Known (HARKing),where scholars retroactively craft hypotheses to fit their findings (Kerr, 1998). Rather than presenting findings as provisional and contingent, researchers often retroactively weave a narrative that makes the outcomes appear predestined by theory. The logic is made to appear seamless: "Of course we expected this result all along!" That tidy narrative is woven post hoc, smoothing over anomalies and uncertainties to reveal a polished arc of cause and effect. Such romanticized narratives may embellish a study's logical arc, but they come at a steep price: they obscure the exploratory nature of the research and sacrifice intellectual honesty. Findings that do not conform to the favored storyline are downplayed or omitted, whereas those that bolster it are overstated. This curated narrative may make for an engaging tale, but it remains perilously detached from the messy truth (Collaboration, 2015; Ioannidis, 2005; Munafo et al., 2017; Wooditch et al., 2020).

The social sciences have seen entire subfields propelled by narrative appeal. One notorious example is the rise and fall of food-behavior scientist Brian Wansink. For years, Wansink published striking studies suggesting that subtle environmental cues dramatically influence how much we eat – for instance, that serving food on smaller plates or using catchy dish names can nudge people toward healthier choices (Wansink, 1996, 2004; Wansink et al., 2005). These studies were celebrated in top journals and covered enthusiastically in the media because they offered intuitive, reassuring solutions to public health challenges. Yet behind the attractive narratives, science was less solid. An investigation found widespread data irregularities in Wansink's work, including impossible data values and statistical inconsistencies across papers (Anaya et al., 2017; Bauchner, 2018; van der Zee et al., 2017). Ultimately, many of these publications were retracted once it became clear that the results stemmed from questionable research practices and analytical flexibility rather than from rigorous experimentation (Dahlberg, 2018). In this case, the seductive power of the hypothesis ("small changes can have big effects") delayed the recognition of fundamental flaws. By the time the truth came to light, an entire generation of research—and the beliefs it shaped—rested on precarious foundations.

Another telling example is the "power posing" phenomenon popularized by social psychologist Amy Cuddy and her colleagues. In 2010, a study claimed that adopting a high-power pose (like standing expansively with hands on hips) for just two minutes could significantly boost confidence and even alter hormone levels, thereby leading to more successful outcomes (Carney et al., 2010; Cuddy et al., 2018). This finding offered an irresistibly simple narrative: a one-step "life hack" to instant empowerment. The idea captured public imagination – Cuddy's TED talk on power posing garnered millions of views – and the psychology community was initially excited by it as well. However, the empirical support for this story quickly began to crumble. However, a larger, more rigorously conducted replication study (Ranehill et al., 2015; Simmons & Simonsohn, 2017), failed to reproduce the purported hormonal and behavioral effects, casting serious doubt on the original findings. In the years that followed, the "power pose" effect came under intense scrutiny and debate—and confidence in its validity largely dissipated. Notably, one of the original authors eventually publicly conceded that the effects were likely nonexistent (Carney, 2016). What began as a groundbreaking narrative about body language's power to empower individuals ultimately became a high-profile example of the replication crisis, reminding us that bold claims demand equally robust evidence.

Narrative isn't the problem—well-crafted stories can reveal hidden social patterns and captivate audiences—but its overuse and abuse strip away the finer distinctions that rigorous analysis demands. In research environments driven by the pursuit of novelty and romanticized narratives, unexpected results are often dismissed as "noise," removed from the dataset or omitted entirely, leaving only the "successful" findings that fit the story. Researchers may unconsciously fall in love with their hypotheses, seeking confirmatory evidence and rationalizing contradictions to

keep the story intact. Fueled by the allure of novel discoveries, this confirmation bias yields fragile scholarship—conclusions that may feel compelling but collapse under replication (Smaldino & McElreath, 2016). Consequently, scholars were led down dead ends chasing ideas that ultimately unraveled, while laypeople were sold interventions that never delivered.

This highlights a serious danger: the seductive pull of novel narratives can derail scientific progress when dazzling stories overshadow careful verification. While compelling hypotheses deserve investigation, they earn their keep only through rigorous, reproducible evidence—not mere rhetorical flair. The hard lessons from Wansink, Cuddy, and similar cases remind us that every scientific narrative must be continually tested and questioned to guard against wishful thinking (Dahlberg, 2018; Ranehill et al., 2015). Beyond the twin temptations of theory and storytelling lies a third peril: conclusions that seem empirically grounded can conceal a different fallacy when statistical methods are misapplied.

Yet beyond the seduction of sweeping theories and captivating stories, social science faces another potent source of error: the improper use of statistical methods. In particular, our field's near-ritualistic emphasis on "significance" ($p < .05$) can create a misleading veneer of importance—a phenomenon we call the NHST illusion. Compounding this issue, two further misapplications commonly arise: first, overreliance on model-fit indices (such as R^2) and standardized coefficients as proxies for causal importance; and second, a tendency to treat feature-importance metrics in machine learning (e.g., from random forests) as though they reveal genuine causal drivers rather than correlational patterns. Together, these misuses contribute to a broader set of "statistical pitfalls," which can bestow unwarranted credibility on theory-driven narratives and overshadow genuine practical significance. In the following sections, we examine how each of these pitfalls operates in practice—and how researchers can guard against the false certainty they impart.

4. Significance Without Substance: The NHST Illusion

If romanticized narratives are one pillar of shaky science, the misuse of statistics is another. Chief among these is our field's almost religious devotion to p-values and Null Hypothesis Significance Testing (NHST) – an obsession often yielding "significance without substance." Among social scientists, a $p < .05$ is often regarded as the ticket to publication and taken as confirmation that a finding is "real." Such p-value worship conflates statistical significance with substantive importance, as if merely surpassing an arbitrary cutoff transforms a tentative result into immutable truth (McCloskey & Ziliak, 2008). This self-perpetuating illusion—one that has long pervaded researchers' minds—drives the relentless hunt for significant p-values while null findings (often more informative) are conveniently brushed aside. (Nosek et al., 2012).

But what does merely crossing the $p < .05$ threshold actually guarantee? On its own, it offers little practical insight. A p-value indicates the probability of observing the data (or something more extreme) if the null hypothesis were true – nothing more (Wasserstein & and Lazar, 2016). It tells us nothing about effect magnitude, the substantive importance of a result, or the theoretical and causal significance of a hypothesis (Cohen, 1994). A finding can be statistically significant yet have a minuscule effect size or be one among a hundred spurious comparisons. Likewise, an important real-world effect might not reach arbitrary "significance" in a small sample. Elevating the $p < .05$ threshold to the status of an infallible criterion for truth is a misplaced veneration that has misguided both academic journals and tenure-review committees.

To understand the NHST illusion, we must distinguish between statistical significance, effect size, and causal importance. Statistical significance is a fickle indicator; it flags when an observed effect is unlikely due to chance given a model, but it conflates sample size with meaning. With a large enough sample, even trivial effects will produce tiny p-values (Lin et al., 2013). Conversely, studies that lack adequate power may dismiss genuinely important effects as "non-significant."

By contrast, effect size directly quantifies the magnitude of a relationship or difference—whether expressed as a correlation coefficient, a mean difference, or an odds ratio—providing context that p-values alone cannot. Effect size speaks to practical significance: whether an effect is small, moderate, or large in the real-world units of the problem (Cohen, 1994). An intervention that reduces crime by 0.5% may be statistically significant in a huge sample ($p < .01$) but practically trivial, whereas one that cuts crime by 30% might fail to achieve $p < .05$ in a tiny pilot study yet obviously has substantive importance. By obsessing over p-values, researchers often lose sight of these distinctions. They herald a "significant" finding even when the effect is so small it would hardly matter if true, or they ignore a potentially major effect because it didn't meet the magic number in one study. This is the NHST illusion: confusing statistical detectability with actual significance in the colloquial sense (Gigerenzer, 2004).

Crucially, neither p-value nor effect size alone guarantees causal relevance. Causal relevance asks: Is this factor a key driver of outcomes in a broader sense? Social phenomena usually have multiple causes interwoven in complex ways. A factor can have a modest effect size yet be a critical piece of a causal puzzle (especially if it's widespread), or it can have a large effect in isolation but operate so rarely that it contributes little to overall outcomes. Here we encounter a commonly neglected aspect: prevalence. The impact of a cause in the population depends on how common that cause is. This concept, well-known in epidemiology as population attributable risk, is often overlooked in social science inference.

Geoffrey Rose famously illustrated that a moderate risk factor affecting many people can cause more total harm than a high-risk factor affecting only a few (Rose, 2001). In our fictional HCC

example, imagine researchers find that teenagers with a rare metabolic disorder are twice as likely to engage in violent crime in adulthood (a relative risk that sounds large and yields p = .03 in their study). Excited by the result, they declare a key breakthrough – "At long last, significance reveals a clue to the causes of crime." But how substantive is this clue? If the metabolic disorder exists in only 0.1% of the population, its prevalence is so low that it might account for a vanishingly small fraction of crimes overall. Eliminating this disorder would barely dent the crime rate. Meanwhile, another factor – say, chronic exposure to childhood violence – might only raise the risk of later crime by 20% (a smaller relative effect), but if that exposure is widespread in 30% of the population, its population-level impact is far greater. Yet this latter factor might be ignored or deemed "insignificant" in a single study because it didn't hit $p < .05$ or because its effect per person seems modest. The fixation on NHST and narrow effect metrics blinds us to the bigger picture of causation. Without considering how prevalent a cause is and how it interacts with other conditions, we end up with findings that are statistically significant but lack real-world substance.

Across social-science research, summary metrics—from R^2 and standardized betas to algorithmic importance scores and even the interpretability of "transparent" models——are often mistaken for evidence of causal impact. Yet unless we rigorously interrogate underlying assumptions, model complexity, and potential confounders, these measures impart a deceptive veneer of certainty that obscures rather than illuminates the mechanisms we aim to understand.

4.1 Misuse of R^2 and Standardized Betas

Researchers often overemphasize measures like R^2 (the coefficient of determination) and standardized beta coefficients as markers of importance, without understanding their limitations. R^2 represents the proportion of variance in the outcome explained by a model – a handy goodness-of-fit metric, yet it is routinely misused. In social science studies, an R^2 of 0.50 is sometimes praised as if "50% of the behavior is explained," and an R^2 of 0.05 is dismissed as trivial. However, both attitudes can be misleading. A high R^2 can be achieved by including many predictors or even irrelevant ones in complex models, especially with overfitting; it doesn't guarantee that the model's factors are causal or important individually (Achen, 1982). Conversely, a low R^2 doesn't mean the predictors lack significance – it might reflect the inherent unpredictability of the outcome or the presence of unmeasured influences. By fixating on R^2, researchers can be incentivized to add variables until the model's fit improves, even if those variables contribute little understanding. For example, an eager HCC analyst might throw a kitchen sink of health-related predictors (diet, exercise, sleep habits, etc.) into a regression to boost R^2 for predicting crime rates. The final model might sport an impressive R^2, but the effect of each variable is tiny and multicollinear, and the narrative becomes that "we can explain X% of crime by health factors" – a potentially spurious claim built on aggregate fit rather than clear

causation.

4.2 Misuse of Importance Scores

Social scientists have increasingly used machine learning models like Random Forests to identify "important" predictors of behavior. In criminology, for example, researchers often report top-ranked features from a Random Forests as if they were key risk factors (Berk, 2008, 2012; Berk et al., 2009; Berk et al., 2016). But Random Forest's importance scores are purely predictive measures – they reflect how often a feature was used in constructing the model, not evidence of a causal effect. For instance, Wallace et al. (2023) show that a non-causal variable highly correlated with true risk factors can appear to be "important" simply because the model splits on it often. As they note, model-based importance values are "inherently associative and non-causal" and depend on which features and algorithm are used. In practice, this means Random Forests importance rankings (e.g., age at release or gang affiliation in a recidivism model) only highlight predictive correlations. Presenting them as if they pinpoint the true "drivers" of crime is misleading, since omitted variables or confounders may actually be the real culprits.

More sophisticated interpretability techniques – permutation importance, SHAP values, partial dependence plots, etc. also quantify predictive contribution, but they do not overcome the correlation–causation gap (Altmann et al., 2010; Berk, 2008; Li et al., 2025; Van den Broeck et al., 2022; X. Zhao et al., 2024). Permutation importance (and methods like Boruta) simply measures how shuffling a feature degrades model accuracy, still within the same dataset and without addressing confounding. Likewise, SHAP values attribute a model's predictions to features using Shapley-value logic, but as its authors emphasize, SHAP "makes transparent the correlations picked up by predictive ML models. But making correlations transparent does not make them causal!" In other words, SHAP can show that a feature contributes to predictions, but it says nothing about what would happen under a hypothetical intervention. Both practitioners and documentation warn that these tools can be misinterpreted. For example, the SHAP developers explicitly caution that one must still build causal models with proper assumptions to answer "what-if" questions. Thus even cutting-edge feature-attribution metrics remain associative: they can help explain what the model learned, but not why the data behaves as they do (Albini et al., 2022).

Cynthia Rudin (2019) famously argues that high-stakes decisions should use inherently interpretable models rather than explained black boxes. Rudin correctly points out that in predictive tasks – such as recidivism forecasting – we make no assumption that the model reflects the true data-generating process, and that "the importance of the variables in the model does not reflect a causal relationship". However, her proposed solution (prefer simpler transparent models) does not address the deeper issue facing social science. In our field, the

goal is to explain human behavior, which ultimately requires causal inference. A sparse or rule-based model may be transparent, but it still encodes associations unless it is carefully designed around a causal framework. Rudin's stance sidelines causal thinking in favor of interpretability alone (Ghassemi et al., 2021; Han et al., 2023; X. Zhao et al., 2024).

In contrast, recent social-science perspectives emphasize that machine learning must be integrated with causal methods. Computational tools should be combined with "conventional social science approaches" to build theory. Relying on predictive importance for theory or policy can mislead: it may seem to identify a "key factor," when in fact that factor is only correlated with unmeasured drivers. In criminology, using importance scores as causal evidence can distort our understanding of crime etiology and yield ineffective or unjust interventions. We must resist the temptation to equate model interpretability with explanation. Instead, social scientists should focus on designing studies (quasi-experiments, longitudinal designs, structural models, etc.) that can isolate causal effects, using ML only as a predictive aid. In sum, interpretability tools can illuminate models, but causal inference, not plain interpretability, must remain the ultimate aim in social science theory and policy (Breiman, 2001; Ghassemi et al., 2021).

5. Pitfalls in Statistical Practice

Even when causal questions are central, common statistical practices can undermine inference through several interrelated pitfalls. First, researchers often improperly compare regression coefficients across models or groups without formal statistical tests. It is common for authors to assert that one effect is "larger" than another simply because its coefficient is significant ($p < .05$) while the other's is not. However, This is statistically invalid – "the difference between 'significant' and 'not significant' is not itself statistically significant" (Gelman & and Stern, 2006; Hayes, 2022). In other words, one cannot assume two coefficients differ simply because one has a $p < .05$ and the other doesn't. Unfortunately, such errors are widespread: a survey of neuroscience papers found that over half misinterpreted differences in this way (Nieuwenhuis et al., 2011). Proper practice involves explicitly testing interactions or differences between coefficients within a unified model rather than casually comparing separate results.

A related mistake arises from conducting separate analyses for subgroups instead of modeling interactions directly, leading to invalid comparisons. A rigorous approach is to incorporate interaction terms within a single analytical framework to statistically test differences across groups. Failing to do so has repeatedly generated flawed conclusions in social sciences and health communication studies. For instance, rather than splitting a dataset by high and low socioeconomic status and conducting separate regressions, researchers should examine whether socioeconomic status significantly moderates effects within a single regression model

(Nieuwenhuis et al., 2011).

Beyond issues of improper model comparison, reliance on advanced causal inference tools without sufficient critical scrutiny—what might be termed "statistical ritualism"—poses additional risks. Powerful techniques such as propensity score matching (PSM) and mediation analysis are frequently used uncritically, treated as magic solutions to resolve confounding or establish causality. Yet each method relies on strong and often untestable assumptions. For instance, PSM assumes comprehensive measurement of all confounders and can paradoxically increase bias if used incorrectly (King & Nielsen, 2019). King and Nielsen (2019) demonstrate, inappropriate use of PSM can exacerbate imbalance between treated and control groups, defeating its original purpose. Nevertheless, matching is sometimes applied mechanically, without proper diagnostics for overlap or balance.

Complexity is often misconstrued as validity—a prime example of statistical-method worship that clouds sound judgment. The pitfall here believes that sophisticated models can substitute for careful research design. No amount of statistical refinement can fully compensate for fundamental issues like unmeasured bias, poor data quality, or violations of assumptions (Freedman, 1991; Freese & Kevern, 2013; Mahoney et al., 2013; Smith, 2013). As David Freedman famously argued, a regression (or any model) is only as good as the subject-matter reasoning and data behind it - otherwise, one is just "adding variables and stirring," a recipe for spurious results.

Another trendy tool susceptible to misuse is causal mediation analysis. Mediation analysis aims to decompose effects (e.g., "How much of a violence prevention program's effect on crime is mediated by improved mental health?"). But causal mediation requires stringent conditions: there must be no unaccounted confounders between the mediator and outcome, among other assumptions (MacKinnon et al., 2007). In practice, these conditions are often unmet. Yet it is tempting for researchers to run mediation models on observational data and proclaim insights about mechanisms. This can be highly misleading – an apparent mediator effect might vanish or reverse if even slight hidden bias exists (Pearl, 2014). Without randomized experiments or strong quasi-experimental designs, causal mediation results are speculative at best. However, the literature contains many such claims of "partial mediation" or "indirect-only mediation" that likely reflect statistical artifacts rather than true causal pathways. The broader lesson is that statistical complexity does not guarantee causal credibility (Bollen & Pearl, 2013; Cinelli & Hazlett, 2019; Freese & Kevern, 2013; Han et al., 2023; Hayes, 2022; Jiang et al., 2021; MacKinnon et al., 2007; Pearl, 2009; Pearl et al., 2016; Yarkoni, 2022; Zhao et al., 2010). Researchers must rigorously verify that the necessary assumptions are met and conduct sensitivity analyses; without these safeguards, sophisticated mediation or structural equation models may merely lend a false veneer of rigor to unreliable conclusions (Chernozhukov et al.,

2018; D'Amour, 2021; Gelman et al., 2021; Müller et al., 2023).

Lastly, the cult of statistical significance and overconfidence in quantitative output deserves critical mention. Quantitative sophistication can seduce researchers into believing the results per se. But p-values, significance stars, and fancy model coefficients are not the voice of truth – they are tools that require correct use and context (Chin et al., 2023; Kerr, 1998; Lösel, 2018; Masicampo & Lalande, 2012; Munafo et al., 2017; Perignon et al., 2019; So et al., 2023; Wooditch et al., 2020). In criminology and health communication studies, an overemphasis on achieving significant results or maximizing fit can overshadow substantive reasoning (Austin, 2003; Chin et al., 2023; Gao et al., 2024; Kühberger et al., 2014; Lösel, 2018; McNeeley & Warner, 2015; Wooditch et al., 2020). We see examples of researchers "tuning" models with numerous controls or polynomial terms to reach significance, forgetting that such practices risk overfitting and obscuring the real signal. We also see blind faith in software output: if a structural model converges or a machine learning algorithm ranks predictors, the results are taken as factual patterns rather than hypotheses to be vetted. These statistics-as-supremacy mindset mistakes analytical output for ground reality. It ignores that models are simplifications – all models are wrong, some are useful, as the aphorism goes (Breiman, 2001). They are only as good as their alignment with data-generating processes. When scholars abdicate substantive interpretation to the algorithm's complexity, they commit what might be called methodological fatalism: assuming that the complexity itself confers legitimacy (Han et al., 2023; Smith, 2013; X. Zhao et al., 2024; Y. J. Zhao et al., 2024).

6. Toward a More Grounded Social Science

To move toward a more grounded social science, researchers must adopt metrics that clearly link statistical findings to real-world impact and causal clarity—two promising approaches are the Population Attributable Fraction (PAF) and the percent contribution to total effect ($\boldsymbol{T_p}$).

6.1 Population Attributable Fraction (PAF)

The Population Attributable Fraction (PAF) originates in epidemiology as the share of cases in a population that can be linked to a risk factor (Levin et al., 2021; Nemati et al., 2023; Poole, 2015). In Levin's classic lung-cancer example, he estimated what fraction of lung cancer cases were "attributable" to cigarette smoking. PAF thus directly asks: What fraction of the total disease burden would vanish if a cause were removed? As Ferguson et al. (2024) note, "PAF… is a commonly used metric in epidemiology" that measures "the importance of a risk factor in causing disease" and helps target interventions. By construction, PAF blends effect size (e.g., a risk ratio) with exposure prevalence. It answers a fundamentally causal question – in a counterfactual world with the exposure eliminated – making it ideal for assessing societal impact rather than mere statistical significance.

However, it should be noted that although the term “attribution” implies a causal relationship, many PAFs based on observed data do not have a clear causal relationship (Peterson et al., 2023). Like other causal indicators, PAF is also based on many important assumptions. For instance, it might still strongly assume that there is no bias in the research design and data analysis. In particular, the estimated effects have been adjusted for all confounding factors. It is also assumed that eliminating exposure does not affect other risk factors, that is, the stable unit-treatment variable assumption or SUTVA, which makes many methodologists feel tormented and painful (Dawid, 2000; Hong & Raudenbush, 2013; Smith, 2013). Furthermore, there are many improvements and variations in PAF. For more information, it is recommended that readers refer to Poole (2015).

Mathematically, for a binary exposure PAF can be written in terms of population prevalence p and relative risk:

$$PAF = \frac{p(RR-1)}{1+p(RR-1)}$$

Equivalently, PAF can be expressed as the difference between the observed incidence and the incidence if all were unexposed, divided by the observed incidence (i.e., the proportion of cases preventable). In cohort data, one often uses Miettinen’s form:

$$PAF = \frac{p_{c(RR-1)}}{1+p_{c(RR-1)}}$$

where p_c is the fraction of cases exposed (Hozawa, 2011). Intuitively, PAF is the fraction of current cases that would not occur if the exposure were eliminated (Ferguson et al., 2024). For example, if smoking has a high RR and half the population smokes, PAF might be large, meaning most lung cancers are “attributable” to smoking. WHO emphasizes that PAF “provides a bridge between the size of a risk factor… and the prevalence of that risk factor” in a population , making it a clear gauge of public-health relevance. In short, PAF quantifies how common and dangerous a cause must be to generate disease burden.

PAF has been used to quantify disease burdens and related costs across many domains (Ferguson et al., 2024; Peterson et al., 2023). For example, large cohort studies in Japan found that modifiable CVD risk factors explained most heart-disease deaths: nonoptimal blood pressure accounted for ~47% of cardiovascular mortality in middle-aged adults (and 26% in the elderly), while smoking explained ~34% of deaths in middle-aged men (Hozawa, 2011). In combination, hypertension and smoking explained 57% of CVD deaths in younger Japanese men (Hozawa, 2011). Likewise, PAF calculations for cancer show that substantial fractions of cases are preventable. One international analysis estimated that ~32.6% of all cancers in 2020 (Iran data) were attributable to known modifiable risks (with smoking 15%, obesity 5%,

infections a few percent each) (Nemati et al., 2023). These examples illustrate how PAF translates risk-factor effects into actual disease impact at the population level.

PAF is also used to assess economic and productivity costs of health risks. For instance, Xiong et al. (2024) applied PAFs from the Global Burden of Disease to a cost-of-illness model in Shanghai. They estimated that in 2020, about US$7.9 billion of societal costs (healthcare plus productivity losses) were attributable to modifiable risk factors, with most loss from cancer and cardiovascular disease. (Notably, ~68% of that burden was due to lost productivity.) In public-health economics, PAF thus allows one to assign portions of healthcare spending and workforce absenteeism to specific causes (e.g. smoking or obesity). This approach is widespread: WHO, CDC, and many national studies compute PAFs to allocate costs and prioritize prevention.

For social scientists, PAF offers a valuable perspective: it bridges the gap between statistical effect and real-world impact. A risk factor may have a large effect (high RR) but be very rare, yielding a small PAF (and vice versa). PAF forces us to consider both magnitude and prevalence. For example, a modest average effect that applies to everyone (high prevalence) can produce a bigger PAF than a strong effect in a tiny subgroup. This contrasts with much social research that reports standardized effect sizes or p-values without assessing how many people are affected. PAF remedies this by framing results in population terms: "If this factor were eliminated, X% of outcomes would be prevented."

Hypothetical examples illustrate PAF's logic in social contexts. Suppose a large fraction of police officers experience chronic burnout (high p) and that burnout modestly increases the risk of misconduct or missed workdays. The PAF would estimate the fraction of crime-related costs (or reduced policing productivity) attributable to burnout. Even a moderate risk ratio could yield a sizable PAF if burnout is common. Similarly, if a prevalent form of parental stress slightly raises the chance of child health problems, PAF quantifies what proportion of pediatric health outcomes is "explained" by that stress. In each case, PAF answers a policy question: How much could we reduce the problem by removing this cause (Suh & Luthar, 2020)?

6.2 Percent contribution to total effect (T_p)

In causal mediation analysis, researchers often quantify how much of an exposure's total effect on an outcome operates through an intermediary variable (mediator). A common metric is the total proportion mediated (T_p) (sometimes simply "proportion mediated", Zhao et al, unpublished article named it T_p), defined as the total indirect effect divided by the total effect. In a simple linear mediation (with one mediator), T_p can be formalized as follows:

$$T_p = \frac{indirect\ effect\ (via\ specific\ mediator)}{total\ effect}$$

T_p often expressed as a percentage. In other words, T_p quantifies the fraction of the overall effect of X on Y that passes through the specified mediator (Alwin & Hauser, 1975; Barreto & Ellemers, 2005; Ditlevsen et al., 2005; Freedman, 2001; Freedman et al., 1992; Huang et al., 2004; MacKinnon, 1994; Mackinnon & Dwyer, 1993; MacKinnon et al., 1995; Sobel, 1982; Tofighi et al., 2009; Zhu et al., 2025).

Despite its intuitive appeal, T_p has well-recognized limitations in practice. These include: 1) statistical instability: The ratio in T_p can be highly variable. Unless sample sizes and effect sizes are large, T_p estimates tend to be imprecise or biased (MacKinnon et al., 1995; Tofighi et al., 2009). Simulation studies show that T_p often requires very large samples to be estimated reliably, and confidence intervals for T_p can be wide or undefined when effects are small; 2) competitive (inconsistent) mediation: When the direct effect of X on Y and the indirect effect via the mediator have opposite signs (so-called "inconsistent" or "competitive" mediation), the proportion mediated can exceed 100% or even be negative, defying straightforward interpretation. For example, if the indirect path is positive but the direct path is negative (or vice versa), T_p can be negative or >1, which undermines the idea of "what portion" is mediated (Hayes, 2022; MacKinnon, 1994; Mackinnon & Dwyer, 1993). In such cases, T_p is not a meaningful fraction and can be misleading; 3) unmodeled pathways: The standard T_p formula assumes that all mediating pathways are included in the model. In reality, omitted mediators or unmeasured confounders will generally end up in the "direct effect." Thus, the direct effect term can absorb other causal pathways not explicitly modeled. This means T_p may overestimate the mediator's true share of the effect: the actual causal influence of the focal mediator is bounded above by T_p. In practice, if there are important omitted paths, T_p will be inflated relative to the mediator's real importance.

Because of these issues, authors have cautioned that T_p should be interpreted with care (MacKinnon et al., 1995; Tofighi et al., 2009). It can be a rough gauge of mediation in simple cases, but it is not a definitive proof of causal relevance.

Given the limitations above, we suggest a shift in perspective: view T_p not as an exact effect-size measure but as an *upper* bound on a mediator's importance. In this view, T_p tells us how large the mediated share could be (if all assumptions held and no other mediators existed), but the actual importance of the mediator is likely no greater than T_p. This framing has two useful implications:

1) A large T_p (e.g. > 20%) suggests that if the model were correct and complete, most of the effect flows through the mediator. This is suggestive evidence that the mediator is important, but by itself it is not sufficient to prove causation. One must still rule out alternative pathways and verify assumptions.
2) In contrast, a very small T_p (e.g. <5–10%) almost certainly indicates the mediator plays a

negligible causal role. If only a few percent of the total effect can possibly go through the mediator, then even perfect mediation would be too weak to be substantively important. In other words, T_p <5% reliably implies limited causal importance of the mediator.

To our knowledge, no prior work has explicitly characterized T_p as a bound in this manner; we believe this represents a novel conceptual contribution of the present Perspective. By treating T_p as a maximum possible share, we avoid misinterpretation: a mediator with T_p = 90% is potentially very important but not proven so, whereas T_p = 2% means the mediator's role is trivial regardless of other assumptions. In summary, T_p remains useful as a rough indicator when large, but its limitations mean it really only provides an upper bound on what a mediator might explain.

7. Conclusion

Our hypothetical HCC thought experiment was not merely whimsical – it was a cautionary tale. That scenario vividly illustrated the pitfalls of questionable social science: overzealous theorizing detached from reality (the scholastic fallacy), romanticizing convenient narratives despite flimsy evidence, and creatively misusing statistics to lend those narratives undue credibility. The HCC saga shows how an investigator, by stringing together impressive-sounding constructs and scattering p-values like confetti, can conjure the illusion of a profound discovery where none exists. Regrettably, this parody mirrors real-world practice: when researchers fall in love with their own models and stories, they start molding reality to fit their ideas rather than the reverse. The result is a scholarly mirage—an internally coherent, data-backed narrative that collapses the moment it meets the complexities of the real world.

The costs of such misadventures are far from academic. In the real realm of policy and public opinion, a seductive but false finding can do genuine damage, misdirecting resources, entrenching stereotypes, or offering a false sense of security. To avoid these harms, a cultural shift in social science is imperative. When interpreting data, we should value candor over elegance and skepticism over certainty. As the statistician David Freedman admonished, no amount of sophisticated modeling can substitute for solid empirical grounding – ultimately, one must put "shoe leather" to the pavement and verify that statistical patterns reflect causal reality (Freedman, 1991). Encouragingly, some disciplines have begun moving in this direction. The past two decades have seen a "credibility revolution" in fields like economics, forcing researchers to confront the challenge of separating correlation from causation with far greater rigor. Emulating this mindset across all social sciences – by emphasizing robust research designs, transparent methodologies, and meaningful effect size reporting – can help inoculate us against the allure of convenient narratives. It means favoring careful causal inference and real-world validation over the facile allure of novelty or technical wizardry.

In closing, we argue for a more grounded, humble, and substantively engaged social science. This means building a scholarly culture that continually asks: What do our findings really mean in human terms? Are we sure that X causes Y, and how much does it truly matter? Answering these questions requires resisting the pull of academic glamour and instead embracing the hard work of credible causal discovery. It requires seeing statistics as a tool subordinate to reality, not a magic wand that creates reality. By doing so, social scientists can avoid the twin pitfalls of analytical overreach and narrative excess. The reward will be a discipline that accumulates genuine knowledge about society: knowledge that may be less flamboyant but is far more useful for understanding and improving the world we all inhabit. Such a down-to-earth approach may lack the romance of an HCC-like fairy tale, but it is the surest route toward a more honest and impactful social science.

References

Albini, E., Long, J., Dervovic, D., & Magazzeni, D. (2022). Counterfactual shapley additive explanations. Proceedings of the 2022 ACM conference on fairness, accountability, and transparency,

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Alwin, D. F., & Hauser, R. M. (1975). The Decomposition of Effects in Path Analysis. *American Sociological Review*, *40*(1), 37–47. https://doi.org/10.2307/2094445

Anaya, J., van der Zee, T., & Brown, N. (2017). *Statistical infarction: A postmortem of the Cornell Food and Brand Lab pizza publications* (2167-9843).

Austin, J. (2003). Why criminology is irrelevant. *Criminology & Public Policy*, *2*(3), 557–564.

Barreto, M., & Ellemers, N. (2005). The burden of benevolent sexism: How it contributes to the maintenance of gender inequalities. *European Journal of Social Psychology*, *35*(5), 633–642.

Bauchner, H. (2018). Expression of Concern: Wansink B, Cheney MM. Super Bowls: Serving Bowl Size and Food Consumption. JAMA. 2005;293(14):1727-1728. *jama*, *319*(18), 1869–1869. https://doi.org/10.1001/jama.2018.4908

Berk, R. (2008). *Statistical learning from a regression perspective* (Vol. 14). Springer.

Berk, R. (2012). *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

Berk, R., Sherman, L., Barnes, G., Kurtz, E., & Ahlman, L. (2009). Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. *Journal of the Royal Statistical Society Series a-Statistics in Society*, *172*(1), 191–211. https://doi.org/10.1111/j.1467-985X.2008.00556.x

Berk, R. A., Sorenson, S. B., & Barnes, G. (2016). Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions. *Journal of empirical legal studies*, *13*(1), 94–115. https://doi.org/10.1111/jels.12098

Bollen, K. A., & Pearl, J. (2013). Eight Myths About Causality and Structural Equation Models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301–328). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_15

Bourdieu, P. (1990). The logic of practice. *Polity*.

Bourdieu, P. (2000). *Pascalian meditations*. Stanford University Press.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199–231.

Carney, D. (2016). *My position on "Power Poses"*. https://faculty.haas.berkeley.edu/dana_carney/pdf_My%20position%20on%20power%20poses.pdf

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power Posing:Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance. *Psychological science*, *21*(10), 1363–1368. https://doi.org/10.1177/0956797610383437

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Chin, J. M., Pickett, J. T., Vazire, S., & Holcombe, A. O. (2023). Questionable research practices and open science in quantitative criminology. *Journal of Quantitative Criminology*, *39*(1), 21–51.

Cinelli, C., & Hazlett, C. (2019). Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 39–67. https://doi.org/10.1111/rssb.12348

Coase, R. (1960). The Problem of Social Cost. *The Journal of Law and Economics*, *3*, 1–44.

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Cuddy, A. J. C., Schultz, S. J., & Fosse, N. E. (2018). P-Curving a More Comprehensive Body of Research on Postural Feedback Reveals Clear Evidential Value for Power-Posing Effects: Reply to Simmons and Simonsohn (2017). *Psychological science*, *29*(4), 656–666. https://doi.org/10.1177/0956797617746749

D'Amour, A. (2021). Revisiting Rashomon: A Comment on" The Two Cultures". *Observational Studies*, *7*(1), 59–63.

Dahlberg, B. (2018). Cornell food researcher's downfall raises larger questions for science. The Salt,

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American statistical association*, *95*(450), 407–424.

Ditlevsen, S., Christensen, U., Lynch, J., Damsgaard, M. T., & Keiding, N. (2005). The mediation proportion: a structural equation approach for estimating the proportion of exposure effect on outcome explained by an intermediate variable. *Epidemiology*, *16*(1), 114–120.

Duffee, D. E., Worden, A. P., & Maguire, E. R. (2015). Directions for theory and theorizing in criminal justice. In *Criminal justice theory* (pp. 425–457). Routledge.

Ferguson, J., Alvarez, A., Mulligan, M., Judge, C., & O'Donnell, M. (2024). Bias assessment

and correction for Levin's population attributable fraction in the presence of confounding. *European journal of epidemiology*, *39*(2), 111–119.

Flyvbjerg, B. (2001). *Making social science matter: Why social inquiry fails and how it can succeed again*. Cambridge university press.

Freedman, D. A. (1991). Statistical Models and Shoe Leather. *Sociological Methodology*, *21*, 291–313. https://doi.org/10.2307/270939

Freedman, L. S. (2001). Confidence intervals and statistical power of the 'Validation'ratio for surrogate or intermediate endpoints. *Journal of Statistical Planning and Inference*, *96*(1), 143–153.

Freedman, L. S., Graubard, B. I., & Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, *11*(2), 167–178.

Freese, J., & Kevern, J. A. (2013). Types of Causes. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 27–41). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_3

Gao, Q., Li, D., Wang, Y., Zhao, C., Li, M., Xiao, J., Kang, Y., Lin, H., & Wang, N. (2024). Analysis of intestinal flora and cognitive function in maintenance hemodialysis patients using combined 16S ribosome DNA and shotgun metagenome sequencing. *Aging Clin Exp Res*, *36*(1), 28. https://doi.org/10.1007/s40520-023-02645-y

Gelman, A., & and Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Gelman, A., Hill, J., & Vehtari, A. (2021). *Regression and other stories*. Cambridge University Press.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/https://doi.org/10.1016/j.socec.2004.09.033

Hambrick, D. C. (2007). The field of management's devotion to theory: Too much of a good thing? *Academy of Management Journal*, *50*(6), 1346–1352.

Han, T., Zhang, L., Zhao, X., & Deng, K. (2023). Total-effect Test May Erroneously Reject So-called" Full" or" Complete" Mediation. *arXiv preprint arXiv:2309.08910*.

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. The Guilford Press.

Hong, G., & Raudenbush, S. W. (2013). Heterogeneous Agents, Social Interactions, and Causal Inference. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 331–352). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_16

Howard, G. J., & Freilich, J. D. (2007). Durkheim's comparative method and criminal justice theory. *Criminal justice theory*, 51–69.

Hozawa, A. (2011). Attributable Fractions of Risk Factors for Cardiovascular Diseases. *Journal of Epidemiology*, *21*(2), 81–86. https://doi.org/10.2188/jea.JE20100081

Huang, B., Sivaganesan, S., Succop, P., & Goodman, E. (2004). Statistical assessment of

mediational effects for logistic mediational models. *Statistics in Medicine*, *23*(17), 2713–2728.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jiang, Y. K., Zhao, X. S., Zhu, L. X., Liu, J. S., & Deng, K. (2021). Total-Effect Test Is Superfluous for Establishing Complementary Mediation. *Statistica Sinica*, *31*(4), 1961–1983. https://doi.org/10.5705/ss.202019.0150

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Pers Soc Psychol Rev*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

King, G., & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, *27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PloS one*, *9*(9), e105825.

Levin, R., Chao, D. L., Wenger, E. A., & Proctor, J. L. (2021). Insights into population behavior during the COVID-19 pandemic from cell phone mobility data and manifold learning. *Nature computational science*, *1*(9), 588–+. https://doi.org/10.1038/s43588-021-00125-9

Li, D. M., Wang, Y. D., & Gao, Q. (2025). A Visual Denoising Model Based on Vision Transformer and Image Groups. *Transformative Society*, *1*(1), 1–17. https://doi.org/10.63336/TransSoc.18

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information systems research*, *24*(4), 906–917.

Lösel, F. (2018). Evidence comes by replication, but needs differentiation: the reproducibility issue in science and its relevance for criminology. *Journal of experimental criminology*, *14*(3), 257–278.

MacKinnon, D. P. (1994). Analysis of mediating variables. *Scientific methods for prevention intervention research*, *139*, 127.

Mackinnon, D. P., & Dwyer, J. H. (1993). Estimating Mediated Effects in Prevention Studies. *Evaluation review*, *17*(2), 144–158. https://doi.org/10.1177/0193841x9301700202

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annu Rev Psychol*, *58*, 593–614. https://doi.org/10.1146/annurev.psych.58.110405.085542

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*(1), 41–62.

Mahoney, J., Goertz, G., & Ragin, C. C. (2013). Causal Models and Counterfactuals. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 75–90). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_5

Masicampo, E., & Lalande, D. R. (2012). A peculiar prevalence of p values just below. 05. *Quarterly journal of experimental psychology*, *65*(11), 2271–2279.

McCloskey, D. N., & Ziliak, S. T. (2008). Signifying nothing: reply to Hoover and Siegler. *Journal of Economic Methodology*, *15*(1), 39–55. https://doi.org/10.1080/13501780801913413

McNeeley, S., & Warner, J. J. (2015). Replication in criminology: A necessary practice.

European journal of criminology, *12*(5), 581–597. https://doi.org/10.1177/1477370815578197

Mills, C. W. (2000). *The sociological imagination*. Oxford University Press.

Müller, S., Toborek, V., Beckh, K., Jakobs, M., Bauckhage, C., & Welke, P. (2023). An Empirical Evaluation of the Rashomon Effect in Explainable Machine Learning. Joint European Conference on Machine Learning and Knowledge Discovery in Databases,

Munafo, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nat Hum Behav*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Nemati, S., Mohebbi, E., Toorang, F., Hadji, M., Hosseini, B., Saeedi, E., Abdi, S., Nahvijou, A., Kamangar, F., Roshandel, G., Ghanbari Motlagh, A., Pourshams, A., Poustchi, H., Haghdoost, A. A., Najafi, F., Sheikh, M., Malekzadeh, R., & Zendehdel, K. (2023). Population attributable proportion and number of cancer cases attributed to potentially modifiable risk factors in Iran in 2020. *International Journal of Cancer*, *153*(10), 1758–1765. https://doi.org/https://doi.org/10.1002/ijc.34659

Nicholas, N. (2008). The black swan: the impact of the highly improbable. *Journal of the Management Training Institut*, *36*(3), 56.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, *14*(9), 1105–1107.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia:II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146. https://doi.org/10.1214/09-Ss057

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological methods*, *19*(4), 459.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Perignon, C., Gadouche, K., Hurlin, C., Silberman, R., & Debonnel, E. (2019). Certify reproducibility with confidential data. *Science*, *365*(6449), 127–128. https://doi.org/10.1126/science.aaw2825

Peterson, C., Aslam, M. V., Niolon, P. H., Bacon, S., Bellis, M. A., Mercy, J. A., & Florence, C. (2023). Economic Burden of Health Conditions Associated With Adverse Childhood Experiences Among US Adults. *JAMA network open*, *6*(12), e2346323–e2346323. https://doi.org/10.1001/jamanetworkopen.2023.46323

Poole, C. (2015). A history of the population attributable fraction and related measures. *Annals of Epidemiology*, *25*(3), 147–154. https://doi.org/https://doi.org/10.1016/j.annepidem.2014.11.015

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the Robustness of Power Posing:No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women. *Psychological science*, *26*(5), 653–

656. https://doi.org/10.1177/0956797614553946
Rose, G. (2001). Sick individuals and sick populations. *International Journal of Epidemiology*, *30*(3), 427–432. https://doi.org/10.1093/ije/30.3.427
Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat Mach Intell*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
Simmons, J. P., & Simonsohn, U. (2017). Power Posing: P-Curving the Evidence. *Psychological science*, *28*(5), 687–693. https://doi.org/10.1177/0956797616658563
Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal society open science*, *3*(9), 160384.
Smith, H. L. (2013). Research Design: Toward a Realistic Role for Causal Analysis. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 45–73). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_4
So, E. M. Y., Yu, F. Q., Wang, B., & Haibe-Kains, B. (2023). Reusability report: Evaluating reproducibility and reusability of a fine-tuned model to predict drug response in cancer patient samples. *Nature Machine Intelligence*, *5*(7), 792–798. https://doi.org/10.1038/s42256-023-00688-4
Sobel, M. E. (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*, *13*, 290–312. https://doi.org/10.2307/270723
Suh, B., & Luthar, S. S. (2020). Parental aggravation may tell more about a child's mental/behavioral health than Adverse Childhood Experiences: Using the 2016 National Survey of Children's Health. *Child Abuse & Neglect*, *101*, 104330. https://doi.org/https://doi.org/10.1016/j.chiabu.2019.104330
Tofighi, D., MacKinnon, D. P., & Yoon, M. (2009). Covariances between regression coefficient estimates in a single mediator model. *British Journal of Mathematical and Statistical Psychology*, *62*(3), 457–484.
Van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2022). On the tractability of SHAP explanations. *Journal of Artificial Intelligence Research*, *74*, 851–886.
van der Zee, T., Anaya, J., & Brown, N. J. L. (2017). Statistical heartburn: an attempt to digest four pizza publications from the Cornell Food and Brand Lab. *BMC Nutrition*, *3*(1), 54. https://doi.org/10.1186/s40795-017-0167-x
Wallace, M. L., Mentch, L., Wheeler, B. J., Tapia, A. L., Richards, M., Zhou, S., Yi, L., Redline, S., & Buysse, D. J. (2023). Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. *BMC medical research methodology*, *23*(1), 144. https://doi.org/10.1186/s12874-023-01965-x
Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416–1436. https://doi.org/https://doi.org/10.1016/j.respol.2017.06.006
Wansink, B. (1996). Can Package Size Accelerate Usage Volume? *Journal of Marketing*, *60*(3), 1–14. https://doi.org/10.1177/002224299606000301
Wansink, B. (2004). Environmental factors that increase the food intake and consumption

volume of unknowing consumers. *Annu. Rev. Nutr.*, *24*(1), 455–479.
Wansink, B., Painter, J. E., & North, J. (2005). Bottomless Bowls: Why Visual Cues of Portion Size May Influence Intake. *Obesity Research*, *13*(1), 93–100. https://doi.org/https://doi.org/10.1038/oby.2005.12
Wasserstein, R. L., & and Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108
Wooditch, A., Fisher, R., Wu, X., & Johnson, N. J. (2020). P-value problems? An examination of evidential value in criminology. *Journal of Quantitative Criminology*, *36*, 305–328.
Xiong, X., Huo, Z., Zhou, Y., Bishai, D. M., Grépin, K. A., Clarke, P. M., Chen, C., Luo, L., & Quan, J. (2024). Economic costs attributable to modifiable risk factors: an analysis of 24 million urban residents in China. *BMC Medicine*, *22*(1), 549. https://doi.org/10.1186/s12916-024-03772-7
Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1, Article e1. https://doi.org/10.1017/S0140525X20001685
Zhao, X., Li, D. M., Lai, Z. Z., Liu, P. L., Ao, S. H., & You, F. (2024). Percentage Coefficient (bp)--Effect Size Analysis (Theory Paper 1). *arXiv preprint arXiv:2404.19495*.
Zhao, X. S., Lynch, J. G., & Chen, Q. M. (2010). Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research*, *37*(2), 197–206. https://doi.org/10.1086/651257
Zhao, Y. J., Feng, G. C., Li, D. M., Ao, S. H., Li, M. M., Tuo, Z. T., Huang, H., Deng, K., & Zhao, X. (2024). Liberal-Conservative Hierarchies of Intercoder Reliability Estimators. *arXiv preprint arXiv:2410.05291*. https://doi.org/https://doi.org/10.48550/arXiv.2410.05291
Zhu, Y., Xiao, Q. E., Ao, M. C., & Zhao, X. (2025). How eHealth use and cancer information-seeking influence older adults' acceptance of genetic testing: Mediating roles of PIGI and cancer worry. *Digital Health*, *11*, 20552076251317658. https://doi.org/10.1177/20552076251317658

8. Disclosure Statement

No potential conflict of interest was reported by the authors.

9. CRediT author statement

Dianshi Moses Li: Conceptualization; Methodology; Writing – Original Draft; Writing – Review & Editing; Project Administration

Hongyang Leon Zhang: Writing – Review & Editing

Qinru Ruby Ju: Conceptualization; Methodology; Writing – Review & Editing