Fragility Metrics as Legal Standards in Gun Control Research

Thomas F. Heston, MD

Clinical Assistant Professor, University of Washington School of Medicine, Seattle, Washington, USA

Clinical Associate Professor, Elson S. Floyd College of Medicine, Washington State University, Spokane, Washington, USA

ORCID: https://orcid.org/0000-0002-5655-2512

Abstract

Empirical claims about gun policy often enter courtrooms and legislative hearings framed through p-values and confidence intervals. Those conventions were not designed to answer the legal questions that matter most: whether a proposition is more likely than not to be helpful, and how stable that conclusion is under small perturbations of the data. This article proposes two diagnostics for firearm policy evidence: the Percent Fragility Index and the Risk Quotient. The Percent Fragility Index reports the minimum percentage of observed outcomes that would have to change to reverse a study's statistical significance under conventional criteria for p-values and confidence intervals. This helps determine how confident we can be that a p-value < 0.05 is meaningful. The Risk Quotient completely removes any reliance on p-values, and instead reports the minimum percentage of outcomes that would have to change to eliminate any claimed benefit of the law. Together these measures translate statistical uncertainty into legally relevant information about reliability and probative value, giving courts and policymakers clearer guidance for admissibility and for the weight to assign to contested studies.

Keywords: fragility metrics, statistical significance, gun policy, legal evidence, percent fragility index, risk quotient, expert testimony, Daubert standard, evidence reliability, firearm regulation

Introduction

Disputes over firearm regulation repeatedly devolve into arguments about whether a study is statistically significant (1). This binary classification is ill suited to adjudication. Significance does not tell judges or juries whether a proposition meets the preponderance standard, nor does it reveal whether the result would collapse if a handful of observations were different. Recent legal scholarship has urged a move away from bright-line significance thresholds toward standards that foreground evidentiary strength and credibility (2). Building on that call, this Article develops a compact, two-metric framework for the assessment of gun policy studies.

The central claim is simple. When legal actors rely on empirical studies to justify or invalidate firearm regulations, they should ask two questions. First, what minimum fraction of outcomes would need to change to eliminate any asserted "therapeutic" benefit of the law? That is the Risk Quotient (RQ). Second, what minimum fraction would flip the study's statistical significance based on p-values? That is the Percent Fragility Index (PFI). Preponderance remains a legal judgment informed—but not dictated—by these diagnostics. Applied together, they convert technical outputs into decision-relevant information without importing a new set of opaque models.

Law's Standards and Statistical Practice

Evidence law links admissibility to relevance and reliability (3). In practice, however, litigation over empirical studies frequently treats a p-value below 0.05 as proof of both (4). That approach misfires on two fronts. It confuses a convention of scientific communication with the legal standard of persuasion, and it ignores the stability of the finding itself. A study may report a p-value of 0.04 and still rest on a knife's edge: a single outcome reversal could flip the conclusion. Conversely, a study may narrowly miss the 0.05 threshold yet strongly indicate that the policy effect is more probable than not. Statistically significant findings are not always robust, and statistically insignificant findings can be meaningful in many cases (5).

Gun policy research magnifies these tensions. Outcomes such as homicide, suicide, and injury are relatively rare in many samples; policy interventions are heterogeneous; quasi-experimental designs are common. In that environment, small shifts in event counts or modeling choices can generate large swings in nominal significance. Courts and agencies need tools that surface these features directly, rather than relying on categorical labels that hide them.

The Percent Fragility Index

PFI addresses the stability problem. Developed for biomedical research (6), the PFI can be adapted for legal applications. For a two-arm study with a binary outcome, PFI is defined as the smallest proportion of all participants whose outcome status would need to change to move an inferential conclusion across a prespecified threshold, typically the customary significance boundary of p = 0.05. Expressed as a percentage, PFI normalizes across study sizes and communicates fragility in plain language. A PFI of 1.8% means that if fewer than two percent of observed outcomes were different, the result would no longer meet the chosen threshold. For a statistically significant study, this means that fewer than 2% of observed outcomes would flip the study from significant ($p < 0.05$) to non-significant ($p > 0.05$).

PFI does not require endorsement of any single significance threshold, and can be computed against whatever alpha level the study itself used (0.05, 0.01, etc.), making it a diagnostic that adapts to the study's own inferential choices rather than imposing new standards. The metric does not claim that the p-value threshold is itself optimal; it reports how close the study stands to that line. Because it is scale-free and interpretable without technical background, PFI allows judges to ask a concrete question: does this conclusion depend on a handful of outcomes, or would it survive modest perturbations consistent with ordinary measurement error or model uncertainty?

A low PFI value indicates the study is more fragile to small changes in outcomes, while a high PFI value indicates the findings are more robust and stable.

The Risk Quotient

RQ expresses distance from neutrality on a 0–1 scale. Defined as the normalized sum of absolute residuals between the observed contingency table and the neutrality table with fixed margins, RQ equals the minimum fraction of outcomes that would need to change to eliminate the observed effect. Because it is scale-free and bounded, RQ allows direct comparison across studies and extends beyond 2×2 designs to multi-arm and tabulated continuous outcomes when the neutrality table is explicitly specified. In court, RQ translates technical departures from relative risk = 1 into a plain-language statement about how much the data would have to change for there to be no difference. For example, an RQ of 0.12 would be interpreted as meaning that if 12% of outcomes changed, then any claimed benefit would be entirely eliminated.

A low RQ value indicates there is very little benefit of the policy even if all the data was correct, while a high RQ value indicates a substantial policy benefit that would require major data changes to eliminate.

A Working Hypothetical

Consider a study evaluating whether a concealed-carry policy increases aggravated assaults relative to a matched set of jurisdictions. Suppose the reported estimate yields statistical significance at the $p < 0.05$ level. The proponent of the policy effect asserts that this meets the relevant burden; the opponent argues that the study is too brittle to bear legal weight.

PFI provides immediate clarity. If the PFI is 2.6%, then flipping the status of less than 3% of outcomes would make the study non-significant and therefore below the relevant burden advocated by the proponent. In this case, the PFI would provide numerically valid support for the opponent's view that the study is too brittle to bear legal weight. In a dataset populated by administrative records subject to ordinary coding error, late reporting, or model misspecification, a PFI of under 5% may be unacceptable for dispositive legal use.

RQ completes the picture. If the same analysis yields RQ of 0.04, then only four percent of outcomes would need to change to completely eliminate the effect of the policy. This magnitude is minimal even if the study was statistically significant by conventional standards. A court could permissibly conclude that, although the estimate crosses a significance boundary, it neither reflects a material departure from the null nor rests on a stable foundation. The RQ provides numeric clarity, and objective justification.

The low PFI of 2.6% indicates the study is fragile, and supports the opponent's "too brittle" argument. The low RQ of 0.04 indicates a minimal policy benefit even if data is correct. Conclusion: the study is statistically significant but not meaningful or stable

Now vary the facts. Suppose a red-flag law evaluation yields a non-significant p-value of 0.06 but a PFI of 1.2% and RQ=0.18. In this case, the low PFI indicates that the study findings are fragile. A 1.2% change in outcomes could change the analysis such that the study would be considered statistically significant ($p < 0.05$). The RQ of 0.18 indicates a large effect of the policy, even though $p > 0.05$. The RQ shows that 18% of all outcomes would need to change before the benefit of the policy would be eliminated. Together, the PFI and RQ suggest a meaningful, positive effect of the law, even though traditional statistical analyses suggest non-significant benefits. A judge confronting cross-motions could acknowledge that the conventional threshold is narrowly missed and highly fragile, while also recognizing that the study shows a large material benefit. The probative value would be substantial even if the result is not stamped "significant."

In this second example, the low PFI of 1.2% indicates fragility of the non-significant finding ($p > 0.05$ could easily flip to significant, $p < 0.05$). The high RQ of 0.18 indicates substantial policy benefit despite $p > 0.05$. Conclusion: study findings are non-significant but meaningful and the law appears to offer substantial benefit.

Institutional Pathways

Courts can incorporate PFI and RQ without rewriting doctrine. Under Rule 702 and Daubert, judges already assess methodological reliability (7). Fragility is one facet of reliability. A requirement that experts disclose the PFI of their main results, computed against their own inferential choices, would reveal the fragility of their conclusions to the court. Likewise, RQ states the minimum fraction of outcomes that would need to change to eliminate the effect of the policy, law, or other intervention. It is evidentiary magnitude, not a probability threshold. Courts may consider that requirement alongside the applicable burden when assigning weight.

Agencies and legislatures have parallel needs. When agencies assemble records to justify firearm rules, the question is not whether every study achieves nominal significance, but whether the totality of the record makes the policy more likely than not beneficial and whether that conclusion is resilient (8). Requiring disclosure of PFI and RQ in agency reports would structure the record around the right questions. Legislatures can also demand these disclosures as a condition of publicly funded gun violence research, improving transparency without dictating outcomes.

Objections and Limits

Two objections recur. The first is that any scalar synopsis oversimplifies complex studies (9). That concern is valid but misdirected. PFI and RQ do not replace design appraisal, causal identification, or external validity analyses. They complement those inquiries by illuminating two decision-critical features that p-values and confidence intervals poorly convey. A judge can, and should, still consider whether a study's identification strategy is credible; PFI and RQ simply ensure that the strength and stability of the numerical claim are not obscured.

The second objection is that fragility depends on contestable choices, such as the alpha level or the outcome definition (10). That is true of all inferential reporting. The appropriate response is disclosure, not disregard. Experts should compute PFI against the threshold they themselves adopt and report sensitivity to reasonable alternatives. For the RQ, specify the neutrality table with fixed margins and disclose sensitivity to reasonable tabulation or modeling choices. Transparency exposes arbitrariness when present.

Finally, the scope conditions of the two diagnostics should be candidly stated. PFI is defined for binary outcomes and two-arm comparisons, which fits many but not all firearm policy evaluations. Extensions exist for multi-arm or time-series designs, but where they are inappropriate the metric should not be forced. RQ depends on an explicitly specified neutrality table with fixed margins; for continuous data, any tabulation assumptions should be disclosed so that the court can evaluate their reasonableness.

Conclusion

Legal standards ask questions that conventional statistical summaries do not answer. In gun policy litigation and lawmaking, the right questions are whether a claim is more likely than not and whether that conclusion is stable. The Risk Quotient and the Percent Fragility Index answer those questions directly. They reorganize expert testimony around the legal burdens that matter, enabling courts and policymakers to distinguish sturdy findings from brittle ones and to assign weight accordingly. In a domain as consequential and contentious as firearm regulation, that recalibration is overdue.

Declarations

Submission status: Original methodological work; not submitted elsewhere.

Data availability: All data is provided in the manuscript.

Funding: None.

Conflicts of interest: None declared.

Bibliography

1. Smart R, Morral AR, Murphy JP, Jose R, Charbonneau A, Smucker S. The science of gun policy: A critical synthesis of research evidence on the effects of gun policies in the united states, fourth edition. Rand Health Q. 2024 Dec 10;12(1):3.

2. Filippini T, Vinceti SR. The role of statistical significance testing in public law and health risk assessment. J Prev Med Hyg. 2022 Mar;63(1):E161–5.

3. Albright TD. A scientist's take on scientific evidence in the courtroom. Proc Natl Acad Sci USA. 2023 Oct 10;120(41):e2301839120.

4. Lytsy P, Hartman M, Pingel R. Misinterpretations of P-values and statistical tests persists among researchers and professionals working with statistics and epidemiology. Ups J Med Sci. 2022 Aug 4;127.

5. Grabowski B. "P < 0.05" might not mean what you think: american statistical association clarifies P values. J Natl Cancer Inst. 2016 Aug 10;108(8).

6. Heston TF. Redefining significance: robustness and percent fragility indices in biomedical research. Stats. 2024 Jun 17;7(2):537-548. GSM 2024 h5-index 16.

7. Daubert v. Merrell Dow Pharmaceuticals, Inc. | 509 U.S. 579 (1993) | Justia U.S. Supreme Court Center [Internet]. [cited 2025 Sep 26]. Available from: https://supreme.justia.com/cases/federal/us/509/579/

8. Department of Commerce v. New York | 588 U.S. ___ (2019) | Justia U.S. Supreme Court Center [Internet]. [cited 2025 Sep 26]. Available from: https://supreme.justia.com/cases/federal/us/588/18-966/

9. Tarpey T, Petkova E, Ciarleglio A, Ogden RT. Extracting scalar measures from functional data with applications to placebo response. Stat Interface. 2021;14(3):255–65.

10. Lin L, Chu H. Assessing and visualizing fragility of clinical results with binary outcomes in R using the fragility package. PLoS ONE. 2022 Jun 1;17(6):e0268754.