Title: Robust inference and errors in studies of wildlife control

Authors: Adrian Treves[1*], Igor Khorozyan[2]

Affiliations:

[1]University of Wisconsin, Madison, USA.

[2]Independent consultant, Göttingen, Germany.

*Corresponding author. Email: atreves@wisc.edu

Abstract: Randomized, controlled trials (RCT) are seen as the strongest basis for causal inference, but their strengths of inference and error rates relative to other study have never been quantified in wildlife control and rarely in other ecological fields. We simulate common study designs from simple correlation to RCT with crossover design. We report rates of false positive, false negative, and over-estimation of treatment effects for five common study designs under various confounding interactions and effect sizes. We find non-randomized study designs mostly unreliable and that randomized designs with suitable safeguards against biases have much lower error rates. One implication is that virtually all studies of lethal predator control interventions appear unreliable. Generally, applied fields can benefit from more robust designs against the common confounding effects we simulated.

Main Text: Identifying the cause of a phenomenon often holds the key to developing an effective intervention to interrupt the cause-and-effect connections or improve outcomes. The stakes increase whenever an intervention risks counter-productive effects on the target or side-effects for another valued entity. Therefore, scientific and public scrutiny of outcomes rather than intentions is intensifying in many applied fields [1]. For example, as societies attach more value to wild animals, scrutiny has intensified for interventions aimed at controls intended to protect human interests from wild animals. Recognition of ineffective or counter-productive effects of lethal wildlife control has exposed an alternative to the traditional hypothesis that removing wild animals, e.g., killing gray wolves (*Canis lupus*), might prevent damage to assets or resources [2]. The more recent hypothesis predicts that removing wild animals might exacerbate the losses of property or threats to safety resources [2]. Hence, the field of wildlife control has become increasingly introspective about robust study designs to evaluate the effectiveness of interventions [2-5]. Resolving these uncertainties about wildlife control interventions would advance the fields of human-animal interactions and ethics, including subfields of biodiversity conservation, agricultural or other property protection, and animal welfare. Other applied fields whose interventions may backfire might also benefit from such introspection.

Quantifying the strengths of inference across study designs

Most investigators advocate the so-called ‘gold-standard’ of randomized, controlled trials (RCT) without biases [6-8]. Yet the urgency of problems may rule against using RCT, exposing tension between swift action and well-informed action [9]. Moreover, RCT can also be infeasible or

opposed by interest groups [10, 11], let alone higher standard designs with crossover (within-subject analysis including the reversal of treatment and control conditions for all subjects) and other blinding steps to avoid research and publication biases [2]. Therefore, evaluations of the effectiveness of interventions in many fields often rely on lower standards of evidence than RCT [1, 11, 12]. Drawing inferences from studies with less robust designs than RCT is the norm in studies of wildlife or ecosystems [3, 11, 13], including our field of wildlife control [2-5]. Approximately 75% of studies in one review of North American and European wildlife control interventions [5], and an unquantified majority of studies in global reviews of wildlife control [3, 14, 15] were non-randomized. Lower standard study designs produce weaker inference because they lack random assignment of treatments and controls or even strict observational controls.

Employing the convenient shorthand and ranking RCT as the gold-standard, we refer to the platinum-standard for crossover designs defined as above, and we hypothesize that one could improve the strength of inference in RCT by employing a within-subjects before-and-after intervention[(rBACI, for "before-after-control" impact or intervention, depending on how the authors name it [2, 5, 16, 17]]. When non-randomized, we refer to nBACI or the 'silver standard'.

The lowest standard in this study is the 'bronze standard' of simple correlation, which compares different doses of intervention and outcomes. This so-called bronze-standard lacks within-subjects comparisons so it introduces additional confounding variables of pre-existing differences between subjects. Therefore, some authors [2, 5] predicted that the gold-standard and higher would outperform the silver- and bronze-standards in strength of inference by a factor of two or more. They further predicted that nBACI would outperform simple correlations and rBACI would outperform RCT, but did not estimate by how much [2].

However, randomized designs are not free of concerns [6]. Murtaugh [17] simulated how temporal autocorrelations confounded the interpretation of a treatment effect. Among the concerns, false positive rates (FPR, inferring a treatment effect when none exists) figure prominently, e.g., electric fences are routinely deemed effective in wildlife control when the evidence is fairly weak [4]. FPR are usually under-estimated due to confusion with p-values which do not tell us how often a test or intervention will fail [8, 18]. Also, "new discoveries" in which the null hypothesis of no effect of an intervention is rejected, under the traditional $p=0.05$ threshold for statistical significance, have been producing high levels of spurious findings that fail replication attempts, whether or not they use randomized study designs [1]. A short-term remedy might be to lower the threshold for significance to $p=0.005$ for new discoveries. But more importantly, Benjamin et al. [1] urge all applied fields to strengthen inference through more robust study designs with safeguards against research and publication biases.

Simulations to quantify error rates

Here we quantify error rates to compare five study designs and their strengths of inference about the effectiveness of lethal wildlife control interventions, following [11, 12]. The simulations in [12] revealed that sample size and study design interact in a complex fashion to influence the probability of detecting true effects on population density change. Here we extend that study by holding sample size constant and investigating two sources of confounding effects. First, we investigate the influence of background interactions arising from correlations between baseline state and intervention (i.e., in our context, property loss and wildlife removal), which is analogous to self-selection or treatment bias. This is a very common interaction in our subfield.

Second, we investigate the confounding effect of correlation between baseline property loss and subsequent property loss in the absence of intervention (temporal autocorrelation). Third, we extend [8, 11, 12] by measuring error rates in simulations of study designs that use Pearson correlation coefficients when treatment effects vary in size and stochasticity. We use simple simulations that expose the rates of Type I errors, Type II errors. and spurious correlations in which the direction of the sign of correlation is reversed when compared to the true direction of the cause and effect. We calculate FPRs and over-estimation bias.

Our approach applies generally to many or all fields that investigate systems characterized by the baseline-intervention-outcome or state-stimulus-reaction causal relationships, including so-called natural experiments. Our simulations model only three parameters and their itneractions: (1) loss of asset or resource prior to intervention, analogous to the baseline/state; (2) removal of wildlife, shortly after time t, analogous to the intervention/stimulus; and (3) loss after intervention, analogous to the outcome/reaction.

Methods

All variable names and definitions are presented in SM Table S1 along with definitions of study designs and models.

To test the traditional wildlife control hypothesis (negative effect of treatment) and more recent hypothesis (positive effect of treatment), we simulated losses of property such as the number of domestic animals L t lost at time t, followed by the intervention as people removed W wild animals, and then we simulated losses in the next time step (L t+1). To simulate crossover designs, we added W at time t + 1 resulting in L t+2. We modeled all W and L as independent, normally distributed random, real numbers from zero to one inclusive, hereafter R. We varied background interactions (B) to mimic potential conditions in the real world (see **Credibility of models** below).

<u>Estimating Type I and II error rates</u>

Type I errors create false positives (we infer an effect of treatment when none exists) and Type II errors lead to false negatives (we infer no treatment effect when one exists). We simulated separately for each type of error. Separately with new iterations of simulations, we examined extreme Type I error when the sign of correlation was reversed over the true sign of correlation. In that simulation, we also examined extreme overestimation of treatment effects by >2SD above a positive mean treatment effect or >2SD below a negative mean treatment effect.

In step one, we set T = 0 for no treatment effect (W x T) and assigned B = 0, -1.16, +1.16, -2.32, or +2.32. We combined different background interactions for Models 0-8 to estimate rates of Type I errors (Table 1, Panels A–D). We set the coefficients empirically to yield an average Pearson r = 0.50 (n=1000 replicates, 10 iterations) so there would be an equal space in either tail for errors. We simulated 200 sets of 20 correlation coefficients with n=50 replicates each (400 iterations per scenario) for each of the 9 model permutations (3600 iterations per scenario-model).

In step two, we repeated the same number of independent simulations as in step one. We simulated cause-and-effect relationships between W and L t+1 (i.e., we set T = ±0.58, Table 1, Panels E-H), to estimate rates of Type II errors (Table 1, Panels E–H).

For step three, we estimated false positive rates (FPR) following [8] as Type I error rate/[Type I error rate + (1- Type II error rate)] using data from Table 1 to construct Table 2.

In step four, we produced five new independent simulations (400 iterations each) to investigate variations of the Type I error in which the lack of a treatment effect changed from a constant T = 0 to a normally distributed random variable centered on zero but with more or less variability per subject from -0.5 to +0.5, -1 to +1, -2 to +2, -4 to +4, and finally -8 to +8. Operationally, we created that random T by subtracting two random numbers of equal magnitude from each other for every replicate. This is analogous to a treatment effect that varies by subject (see Credibility of models below). We estimated Type I error rates again as above. We modeled with a generalized linear mixed model those error rates with four predictors (study design, variable treatment effect for each replicate, background interactions from Models 3 and 4, and the direction of the Type I error (i.e., whether a spurious significant result emerged for a positive or a negative correlation).

In steps five and six, we explored the extreme Type II errors. We ran seven simulations independent of those above (400 iterations each). For sign reversal, we counted the number of correlation coefficients that had an opposite sign as the real correlation regardless of the magnitude. In step 5, for extreme errors we repeated the procedure in steps 1-2 but counted the number of treatment effect size estimates that exceeded the mean +2SD for a positive treatment effect or fell below the mean -2SD for a negative treatment effect. For both steps 3 and 4, temporal autocorrelation (B) varied from -2.32 to +2.32 independently of study design. We estimated mean and standard deviations of error rates in both steps (Figs. 1 and 2).

In all steps, we chose deterministic and probabilistic scenarios in preference to empirical domestic animal loss rates from the literature, because the latter would include unmeasured background interactions and unreported treatment (e.g., poaching), which would undermine our effort at measuring the odds of Type I and II errors.

Credibility of models

Background interactions simulate common situations in wildlife control. A positive correlation between W and L t (Models 1 and 2, Table S1) mimics a common background interaction in which people kill more predators if losses were high in the past [19]. Probably uncommon, a negative correlation between W and L t mimics when people kill fewer predators after high losses, e.g., when people and wildlife separate spatially after high losses [20, 21]. A positive correlation between L t and L t+1 (Models 3 and 4, Table S1) without intervention mimics a common temporal autocorrelation, in which sites with high losses one year have high losses the next year [22, 23]. Possibly less common, a negative temporal autocorrelation mimics cyclical patterns of damage in non-sequential years. For example, when wild food availability influences bear damage to crops and human foods, one may see a negative temporal autocorrelation of losses from year to year [24, 25]. Or, if predators switch from domestic to wild prey selection based on their relative scarcity or vulnerability varying over time, we can see prey switching from season to season that might produce negative autocorrelations of losses in sequential time steps [26-29].

These first four background interactions create univariate permutations. In the last four bivariate permutations (Models 5–8, Table S1), we simulated both sets of interactions occurring simultaneously in a two by two matrix of positive or negative interactions. For step four, when we varied the treatment effect size in every replicate, we mimicked a situation in which the same dose had variable effects on different replicates. For example, an individual predator may respond differently than its neighbor or the composition of social groups may affect how the survivors respond to removal of a group member, e.g., removing alpha individuals from a wolf pack is expected to have different effects than removing subordinate adults or pups from a pack, and even packs experiencing the same removal of dominant breeders might have different effects depending on timing and availability of replacement breeders [30]. Hence, the same dose (W) could have different treatment effect (T) depending on the idiosyncrasies of different replicates. Similarly, some individual predators might be attracted or repelled by vacancies left by removals of other predators [31]. Alternately, any of the individuals involved might respond differently to lethal treatments. Theory provides five potential explanations for why the traditional hypothesis may fail [31]. In brief, the wrong predators may be killed, e.g., [32]; the survivors may prey on livestock that are mor predictable than wild prey after the predators' social group has been disrupted, e.g., pack hunting carnivores that rely on teamwork to hunt or reproduce successfully, e.g., [33]; more immigrants may replace fewer residents that were killed, e.g., [34]; smaller-bodied predator species at higher densities may refill the vacancies left by larger, scarcer predator species that died, e.g., [35]; or humans and domestic animals may change their behavior after lethal intervention. When we consider the entire set of actors, predators, humans, and domestic animals, one can imagine interindividual differences in response to lethal interventions. For example, some bold and tolerant individuals might explore wilder habitat after predator removal while others might continue to avoid those areas [31]. In short, the same treatment of different actors could result in diametrically opposed consequences even though the treatment did have an effect on a subset of replicates. Despite different effects on different subjects, across replicates, the general effect of treatment approximates zero so we estimated Type I error rates.

Analysis

We calculated Pearson's correlation coefficient r in JMP Pro V15.0.0 (SAS 2019). Pearson's r is easily interpretable, dimensionless, and suitable for normally distributed, random variables [36]. With normally distributed response variables like L and change in L, Pearson's r is unbiased, normal (Anderson-Darling test $A = 0.78$, $p = 0.05$ and $A = 0.37$, $p = 0.38$, respectively). We calculated r in 20 batches of 50 replicates (analogous to independent sites or populations), a larger sample size than most studies of wildlife control. We used the Pearson r standard critical value of $|r| = 0.273$ (two-tailed test at alpha=0.05, n=50 calculated from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/table-of-critical-values-pearson-correlation/, accessed 5 July 2023) in 400 iterations of each combination of scenarios (Table S1) for a total of 108,000 independent combinations. We calculated 400 correlations per simulation (108 scenarios in Table 1, 25 scenarios for the mixed model of Type I errors, and 35 scenarios for extreme Type II errors) for a total of 67,200 Pearson r values including 50 independent replicates each. There were fewer scenarios for randomized designs because the background interactions of L t correlated with W were eliminated by random assignment procedures (Table S1).

We involved neither animals nor human subjects in this research.

Results

False Positive Rates (FPR)

As predicted in [2], study designs differed noticeably in Type I and II error rates (Table 1) and therefore, in FPR (Table 2). As predicted by [8], FPRs exceeded Type I error rates based on p values in 93% (100/108) of our simulations (Table 2). None of the scenarios had FPR <1%. Therefore, we echo calls for lowering the statistical threshold for new discoveries [1].

The lowest FPR was 3.9% for rBACI when there were no background interactions (Table 2). In 8 scenarios, the FPR was 5.0% or less (4 scenarios with rBACI and 4 with crossover). Although rBACI had two of the lowest FPR (Table 2), it was outperformed by crossover when we introduced temporal autocorrelation in either direction, i.e., background interaction between B due to correlation between L t and L t+1. Indeed, crossover designs had a lower average FPR across 12 scenarios (6.1%, SD 1.4%) than RCT (6.4%, SD 1.0%) and rBACI (6.5%, SD 2.6%). Although these differences in FPR among randomized designs are small, the case for crossover design strengthened as we explain next.

We used a generalized linear mixed equation to model the interactions between confounding effects and study design on Type I error rates when treatment effects were centered on zero, but random in each replicate, i.e., no treatment effect in general (see Methods for examples of when this might arise). The mixed model revealed significant fixed effects only for study design ($df=4$, $F=78$, $p<0.00001$) and variable treatment effect for each replicate ($df=1$, $F=31$, $p<0.0001$); neither direction of error ($df=1$, $F=0.2$, $p=0.62$) nor the magnitude of temporal autocorrelation ($df=6$, $F=1$, $p=0.44$) were predictive of error. Also, study design and variable treatment effect for each replicate interacted significantly to predict the Type I error ($df=4$, $F = 64$, $p<0.0001$). Crossover performed best, because RCT and rBACI were somewhat vulnerable to randomly varying treatment effects (0.8% higher error rates), probably because the crossover design exposes each replicate to both control (treatment T = 0) and treatment (T varies randomly around zero) conditions. Because Type I error rates contribute to FPR directly, crossover design (platinum-standard) provided a stronger inference than the other study designs we tested [2].

Given FPR >1% seem risky to us, we recommend lowering the threshold for significance even when randomized designs are employed. Our results also corroborate prior cautions to measure and account for temporal autocorrelation [17]. We believe that temporal autocorrelation is a common condition in our field because of the widespread and frequent reports of 'hot spots' of damage by wild animals year after year [22, 37-40].

By comparison to the randomized study designs, we cannot recommend simple correlation or nBACI (bronze- and silver-standard, respectively) because their FPR ranged from 5.2-42% and 5.8-88%, respectively (Table 2). Negative temporal autocorrelation (Model 4) made these designs particularly vulnerable with FPR two to three times higher than for positive temporal autocorrelation. The highest FPR arose in Models 5-8 (Table 2). Although nBACI was somewhat resistant to Models 5 and 8 when the background interactions were strong (2.32), nBACI failed in most cases, including several with only one background interaction (Table 2). Although simple correlations yielded consistent FPR of 5-12.5% when we introduced only one background interaction, their FPR rose above 20% whenever we included two background interactions.

Although one might be tempted to look at a few low Type I error rates in Table 1 for simple correlation and nBACI, and declare these study designs viable in many circumstances, the FPR

in Table 2 warn against such confidence. Also, with FPR for simple correlation averaging 16% (SD 12%) and nBACI averaging 29% (SD 25%), in the absence of good evidence about background interactions, one should not credit these study designs. Indeed, in many experimental situations, particularly under field conditions surrounding wildlife control, researchers will have little or no evidence to dismiss background interactions. Even when such evidence for background interactions is robust and well-accounted in the analyses, few researchers in our field can build a sample size of 50 on which our simulations depend. Therefore, FPR values in Table 2 are likely under-estimates of what others will encounter with smaller samples, imperfect randomization, variable treatment effect for each replicate, deviations from the assumptions of Pearson correlations, and measurement error [8].

Severe Type II errors: overestimation and sign reversal

Some of the simulated Type II error rates were very high (Table 1), which by itself may not raise concern because Type II error conservatively leads us to infer no effect when one exists in reality. However, reporting the opposite sign of correlation than the real direction of correlation when a treatment is effective would be an extreme form of Type II error that merits concern (Fig. 1). Also, when we overestimate the real effect substantially (e.g., >2SD above a positive mean or below a negative mean), exaggerated claims about treatment effectiveness can mislead users, payers, and distributors of that treatment (Fig. 2). As temporal autocorrelation increased, the rate of sign reversal increased and simple correlation was more strongly affected than nBACI (Fig. 1). The converse was true for overestimation error, which declined among the non-randomized study designs. Simple correlation was less prone to these errors than nBACI (Fig. 2).

Compared to randomized designs, the rates of sign reversal for simple correlation and nBACI were higher (8% and 0.8% respectively; only simple correlation differed significantly from every other design, each t-test pairwise comparison $p<0.0001$) than randomized designs (RCT – 0.09%, rBACI – 2%, crossover – 0.08%, which did not differ among randomized designs, Welch test unequal variances, F ratio = 2, $p=0.15$).

Similarly, non-randomized designs had higher rates of overestimating treatment effect sizes (8% for simple correlation and 31% for nBACI), which differed significantly from randomized designs ($p<0.0001$ for each pairwise comparison with nBACI, $p<0.009$ for pairwise comparisons of simple correlation to each randomized design). Also, randomized study designs were statistically different in rates of overestimation error (RCT – 0.2%, rBACI – 1%, crossover – 3%, F ratio = 31, $p<0.0001$).

In sum, our predictions of the relative strength of inference among study designs were only partly supported [2]. The predicted difference between simple correlation (bronze-standard) and nBACI (silver-standard) held for sign reversal (Fig. 1), but not for overestimation bias (Fig. 2) or most FPR (Table 2). Similarly, the so-called gold+ of rBACI compared to gold-standard RCT did not play out as we predicted [2]. Yet, our predictions about crossover design (platinum-standard) producing stronger inference than RCT and rBACI (gold-standards) were supported. Therefore, we revised our first hypotheses [2] by producing a schematic graph of relative strengths of inference estimated for five study designs (Fig. 3).

Discussion

Some public authorities may not test treatments with randomized, controlled experiments because they perceive intervening as infeasible or impractical, perhaps in part because they

believe the treatments will be popular and the placebo controls will be unpopular, e.g., [41]. Therefore, authorities may prefer to intervene in ways they consider less controversial, such as treating all subjects or serving the loudest complainants [[5], see webpanel 1]. Such steps that lead to non-randomized study designs risk backfiring or wasting time and resources.

When subjects are self-selected (self-selection bias), vulnerable subjects receive higher doses (treatment bias), or baseline conditions affect outcomes and not just treatments (e.g., temporal autocorrelation), we can expect high false positive rates (FPR, Table 2), especially for non-random before-and-after comparisons of interventions (nBACI). When background interactions are strong, FPR rise sharply (Table 2). When both sets of background interactions coincide, we estimated that wrong conclusions would be drawn in 18–42% of simple correlation studies and even more variably in 8–88% of nBACI (Table 2). Also, when temporal autocorrelation is present, the results of non-randomized study designs will produce additional errors even if the study is designed to minimize false positives. Non-randomized designs pose a considerable risk of the reversal of the sign of correlation, which can substantially mislead researchers and practitioners about the treatment effect (Fig. 1). If sign reversal does not occur, overestimation of treatment effects is also possible (Fig. 2). These compounding errors associated with non-randomized study designs can be visualized as a hierarchy of study designs (Fig. 3).

Overall, the compounding errors weigh heavily against non-randomized designs (Fig. 3). Unlike randomized designs, non-randomized designs produce errors asymmetrical with regard to positive or negative background interactions (Figs. 1, 2). Namely, positive temporal autocorrelations produced more sign reversal errors and fewer overestimation errors in non-randomized designs than did negative temporal autocorrelations. That asymmetry would tend to confuse the direction of the treatment effect more often when outcomes correlate positively to baseline conditions (Fig. 1); that situation is common in our subfield where hot spots of wildlife damage recur annually (*SM*).

Regrettably, predator control has been dominated by unreliable, non-randomized studies. Hence, predictably, there is no scientific consensus about the effects of predator control on subsequent domestic animal losses, particularly in case of lethal treatments [3, 14, 15]. For example, non-randomized study designs have produced equivocal results for lethal control including recurrent findings of counter-productive increases in domestic animal losses following killing gray wolves [42, 43], bears (*Ursus* spp.) [25, 44, 45] and cougars (*Puma concolor*) [46, 47]. Theory provides five potential explanations for why the traditional hypothesis may fail, cf. [31] and described with references in Methods. In brief, the wrong predators may be killed; the survivors' behaviors may change if they relied on group-mates that were killed; immigrants of the same species or smaller-bodied predatory species may refill in greater numbers the vacancies left after killing; or survivors of any species may change behavior after predators are removed.

Even well-financed RCT across broad areas may be hard to interpret, e.g., U.K.-funded RCT of badger (*Meles meles*) killing to prevent bovine tuberculosis documented variable effects of this intervention that can be difficult to detect [48-53].Even methods considered politically unpalatable but highly effective, such as poisoning red foxes (*Vulpes* vulpes) in Australia to protect sheep, when tested with RCT prove highly variable in effect [54]. The latter research team concluded from an RCT that poisoning foxes wasted much effort and was ineffective because it produced very slight decreases in lamb mortality. Despite these doubts, lethal methods are rarely subjected to RCT. Most randomized studies of predator control have been conducted

on non-lethal methods to prevent predators from damaging property [41, 55, 56]. An analogy would be to ignore experiments on handgun control [57] while subjecting pepper spray to robust RCT. Moreover, in the absence of scientific consensus the historical intervention of killing predators continues unabated despite years of criticism [5, 48].

The resilience of lethal treatments in policy circles may reflect a perceptual bias of "cherry picking" arising from the adoption of a few effective cases and the dismissal of more numerous ineffective cases [33, 42, 43, 58]. Our mixed models show that treatments that help some replicates and harm others will raise FPR with worrying frequency in non-randomized studies. In addition, animal killing may fall into another perceptual bias because either humans cannot recognize individual animals, some of which are culprits and some of which are not [32, 33], or some persons may claim a lethal treatment has succeeded because the death of a competitor might have been their primary goal regardless of its culpability.

If a non-randomized design is analyzed in spite of our cautions above, researchers should account for potential self-selection bias, treatment bias, and temporal autocorrelation. For example, lethal wildlife control studies should measure (a) killing and property losses before that killing occurred, and (b) property losses from year to year in the absence of intervention [17, 43]. The absence of intervention includes unplanned or unregulated interventions by the people participating or using the same areas. This is a very difficult hurdle to overcome without strict control of participant actions because predator killing can still be present as an illicit behavior and hushed up [59-61]. Therefore, we suggest randomized designs in smaller, well-controlled sites are likely to be more feasible than strict control over potentially confounding variables across entire landscapes. Even for randomized designs, we counsel care because FPR does not diminish to zero. To lower the risk of FPR, we recommend the platinum-standard crossover design RCT (all subjects receive both treatment and placebo in random order), lowering the significance threshold [1], and other safeguards against bias [2].

A common argument for drawing inference from non-randomized studies has been that experts can infer accurately despite confounding variables [17]. For example, expert-based adaptive managers claim they can intervene, learn, and revise without exacerbating the problems at hand and without exposing hypotheses to experimental test [62, 63]. That argument depends on learning correctly. The counter-argument is that biased designs and lower standards hinder learning with false information and can produce inferences diametrically opposed to the actual effect of interventions [6, 64]. Our results of sign reversal in treatment effects support that concern. Therefore, prioritizing randomized designs for urgent and important policy decisions may avoid the age-old problem that haste makes waste. The reasoning here provides a guide to donors, regulators, and the public to anticipate situations in which RCT becomes a prerequisite for reliable inference and sound policy.

Acknowledgments: We thank RJ Treves for statistical advice.

Data and materials availability: For scripts and a full spreadsheet with 1000 rows of data for a single iteration of each simulation, see https://faculty.nelson.wisc.edu/treves/data_archives/Simulate_study_designs_scripts_data_archive.zip, accessed 21 October 2023.

References

[1] D. Benjamin *et al.*, "Redefine statistical significance," *Nature Human Behaviour,* vol. 2, pp. 6–10, 2018.

[2] A. Treves, M. Krofel, O. Ohrens, and L. M. Van Eeden, "Predator control needs a standard of unbiased randomized experiments with cross-over design," *Frontiers in Ecology and Evolution,* vol. 7 pp. 402-413, 2019, doi: 10.3389/fevo.2019.00462.

[3] I. Khorozyan, "Defining practical and robust study designs for interventions targeted at terrestrial mammalian predators," *Conserv. Biol.,* vol. 36, p. e13805, 2022, doi: 10.1111/cobi.13805.

[4] I. Khorozyan, "Dealing with false positive risk as an indicator of misperceived effectiveness of conservation interventions," *PLoS One,* vol. 16, no. 5, p. e0255784, 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0255784.

[5] A. Treves, M. Krofel, and J. McManus, "Predator control should not be a shot in the dark," *Front. Ecol. Environ.,* vol. 14, pp. 380-388, 2016.

[6] J. P. Ioannidis, "Why most published research findings are false," *PLOS Medicine,* vol. 2, no. 8, p. e124, 2005.

[7] M. Baker and K. Brandon, "Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help," *Nature,* vol. 533, pp. 452-454, 2016. [Online]. Available: https://www.nature.com/articles/533452a.

[8] D. Colquhoun, "An investigation of the false discovery rate and the misinterpretation of p-values," *Royal Society Open Science,* vol. 1, p. 140216, 2014. [Online]. Available: http://dx.doi.org/10.1098/rsos.140216.

[9] N. Oreskes, *Why Trust Science?* Princeton, NJ: Princeton University Press, 2019.

[10] J. R. Platt, "Strong inference," *Science,* vol. 146, pp. 347–353, 1964.

[11] A. P. Christie *et al.*, "Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences," *Nature Communications,* vol. 11, p. 6377, 2020. [Online]. Available: https://doi.org/10.1038/s41467-020-20142-y | www.nature.com/naturecommunications.

[12] A. P. Christie, T. Amano, P. A. Martin, G. E. Shackelford, B. I. Simmons, and W. J. Sutherlan, "Simple study designs in ecology produce inaccurate estimates of biodiversity responses," *J. Appl. Ecol.,* vol. 56, pp. 2742–2754, 2019, doi: 10.1111/1365-2664.13499.

[13] T. J. Clark and M. Hebblewhite, "Predator control may not increase ungulate populations in the future: A formal meta-analysis," *J. Appl. Ecol.,* vol. 58, no. 4, pp. 812-824, 2021, doi: 10.1111/1365-2664.13810.

[14] A. Eklund, J. V. López-Bao, M. Tourani, G. Chapron, and J. Frank, "Limited evidence on the effectiveness of interventions to reduce livestock predation by large carnivores," *Scientific Reports,* vol. 7, pp. 2097 | DOI:10.1038/s41598-017-02323-w, 2017.

[15] L. M. van Eeden *et al.*, "Carnivore conservation needs evidence-based livestock protection," *PLoS Biol.,* vol. l 16, no. 9, p. e2005577, 2018, doi: e2005577.

[16] A. J. Underwood, "Beyond BACI: The detection of environmental impacts on populations in the real, but variable, world," *J. Exp. Mar. Biol. Ecol.,* vol. 161, pp. 145-178, 1992.

[17] P. A. Murtaugh, "On rejection rates of paired intervention analysis," *Ecology,* vol. 83, pp. 1752–1761, 2002.
[18] D. Colquhoun, "The reproducibility of research and the misinterpretation of p-values," *Royal Society Open Science,* vol. 4, p. 171085, 2017. [Online]. Available: http://dx.doi.org/10.1098/rsos.171085.
[19] M. R. Conover, "Effect of hunting and trapping on wildlife damage " *Wildl. Soc. Bull.,* vol. 29, no. 2, pp. 521-532, 2001.
[20] J. Knight, *Waiting for Wolves in Japan*. Oxford: Oxford University Press, 2003.
[21] L. Naughton-Treves, "Whose animals? A history of property rights to wildlife in Toro, western Uganda," *Land Degradation and Development,* vol. 10, pp. 311-328, 1999.
[22] A. Treves, K. A. Martin, A. P. Wydeven, and J. E. Wiedenhoeft, "Forecasting environmental hazards and the application of risk maps to predator attacks on livestock," *Bioscience,* vol. 61, pp. 451-458, 2011.
[23] J. R. B. Miller, "Mapping attack hotspots to mitigate human–carnivore conflict: approaches and applications of spatial predation risk modeling," *Biodivers. Conserv.,* vol. 24, no. 12, pp. 2887-2911, 2015.
[24] D. L. Garshelis, "Nuisance bear activity and management in Minnesota," in *Bear - People Conflicts - Proceedings of a Symposium on Management Strategies*, M. Bromley Ed. Yellowknife, Canada: Northwest Territories Department of Renewable Resources, 1989, pp. 169-180.
[25] J. M. Northrup, E. J. Howe, J. Inglis, E. Newton, M. E. Obbard, B. Pond, and D. Potter, "Experimental test of the efficacy of hunting for controlling human–wildlife conflict," *J. Wildl. Manage.,* vol. 87, no. 3, p. e22363, 2022, doi: 10.1002/jwmg.22363.
[26] A. Janeiro-Otero, T. M. Newsome, L. M. Van Eeden, W. J. Ripple, and C. F. Dormann, "Grey wolf (Canis lupus) predation on livestock in relation to prey availability," *Biol. Conserv.,* vol. 243, 2020, doi: 10.1016/j.biocon.2020.108433.
[27] I. Khorozyan, A. Ghoddousi, M. Soofi, and M. Waltert, "Big cats kill more livestock when wild prey reaches a minimum threshold," *Biol. Conserv.,* vol. 192, pp. 268–275, 2015.
[28] I. Laporte, T. B. Muhly, J. A. Pitt, M. Alexander, and M. Musiani, "Effects of wolves on elk and cattle behaviors: implications for livestock production and wolf conservation," *PLoS One,* vol. 5, no. 8, p. e11954, 2010, doi: 10.1371/journal.pone.0011954.
[29] J. Odden, I. Herfindal, J. D. C. Linnell, and R. Andersen, "Vulnerability of domestic sheep to lynx depredation in relation to roe deer density " *J. Wildl. Manage.,* vol. 72, no. 1, pp. 276-282, 2008.
[30] S. M. Brainerd *et al.*, "The effects of breeder loss on wolves," *J. Wildl. Manage.,* vol. 72, no. 1, pp. 89-98, 2008.
[31] L. Elbroch and A. Treves, "Why might removing carnivores maintain or increase risks for domestic animals? ," *Biol. Conserv.,* vol. 283, p. 110106, 2023, doi: 10.1016/j.biocon.2023.110106.
[32] L. Plumer, T. n. Talvi, P. Männil, and U. Saarma, "Assessing the roles of wolves and dogs in livestock predation and suggestions for mitigating human-wildlife conflict and conservation of wolves," *Conserv. Genet.,* vol. 19, pp. 665–672, 2018. [Online]. Available: https://doi.org/10.1007/s10592-017-1045-4

[33] F. F. Knowlton, E. M. Gese, and M. M. Jaeger, "Coyote depredation control: An interface between biology and management," *Journal of Range Management,* vol. 52, pp. 398-412., 1999.
[34] H. S. Cooley, R. B. Wielgus, H. S. Robinson, G. M. Koehler, and B. T. Maletzke, "Does hunting regulate cougar populations? A test of the compensatory mortality hypothesis," *Ecology,* vol. 90, pp. 2913-2921, 2009.
[35] K. R. Crooks and M. E. Soulé, "Mesopredator release and avifaunal extinctions in a fragmented system," *Nature,* vol. 400, pp. 563-566, 1999.
[36] Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," ed. https://osf.io/ezcuj/, 2015, p. https://osf.io/ezcuj/.
[37] L. Naughton-Treves, "Predicting patterns of crop damage by wildlife around Kibale National Park, Uganda," *Conserv. Biol.,* vol. 12, pp. 156-168, 1998.
[38] K. K. Karanth, A. M. Gopalaswamy, P. K. Prasad, and S. Dasgupta, "Patterns of human-wildlife conflicts and compensation: insights from Western Ghats protected areas," *Biol. Conserv.,* vol. 166, pp. 175-185, 2013.
[39] J. R. B. Miller, Y. V. Jhala, J. Jena, and O. J. Schmitz, "Landscape-scale accessibility of livestock to tigers: implications of spatial grain for modeling predation risk to mitigate human–carnivore conflict," *Ecology and Evolution,* vol. 5, no. 6, pp. 1354-1367, 2015.
[40] A. Treves and M. F. Rabenhorst, "Risk map for wolf threats to livestock still predictive 5 years after construction," *PLoS One,* vol. 12, no. 6, p. e0180043, 2017, doi: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0180043.
[41] O. Ohrens, C. Bonacic, and A. Treves, "Non-lethal defense of livestock against predators: Flashing lights deter puma attacks in Chile.," *Front. Ecol. Environ.,* vol. 17, no. 1, pp. 32-38, 2019, doi: 10.1002/fee.1952.
[42] O. Grente, "Présentation des objectifs et de la méthodologie de la thèse sur l'efficacité des tirs de loup et la gestion adaptative du loup, menée conjointement par l'ONCFS et le CEFE," OFB, Direction de la recherche et de l'appui scientifique, Unité Prédateurs, animaux déprédateurs et exotiques, Équipe Loup et Lynx, Gières, France, 2021.
[43] F. J. Santiago-Avila, A. M. Cornman, and A. Treves, "Killing wolves to prevent predation on livestock may protect one farm but harm neighbors," *PLoS One,* vol. 13, no. 1, p. e0189729 2018, doi: /10.1371/journal.pone.0189729.
[44] I. Khorozyan and M. Waltert, "Variation and conservation implications of the effectiveness of anti-bear interventions," *Scientific Reports,* vol. 10, p. 15341, 2020, doi: 10.1098/rsos.190826.
[45] D. L. Garshelis, K. V. Noyce, and V. St-Louis, "Population reduction by hunting helps control human-wildlife conflicts for a species that is a conservation success story," *PLoS One,* vol. 15, no. 8, p. e0237274, 2020, doi: 10.1371/journal.pone.0237274.
[46] J. W. Laundré and C. Papouchis, "The elephant in the room: What can we learn from California regarding the use of sport hunting of pumas (Puma concolor) as a management tool?," *PLoS One,* vol. 15, no. 2, p. e0224638, 2020, doi: https://doi.org/10.1371/journal.pone.0224638.
[47] K. Peebles, R. B. Wielgus, B. T. Maletzke, and M. E. Swanson, "effects of remedial sport hunting on cougar complaints and livestock depredations.," *PLoS One,* vol. 8, no. 11, p. e79713, 2013.

[48] J. Bielby, F. Vial, R. Woodroffe, and C. A. Donnelly, "Localised badger culling increases risk of herd breakdown on nearby, not focal, land," *PLoS One,* vol. 11, no. 10, p. e0164618, 2016, doi: 10.1371/journal.pone.0164618.
[49] C. Donnelly and R. Woodroffe, "Reduce uncertainty in UK badger culling," *Nature,* vol. 485, p. 582, 2012.
[50] C. Donnelly *et al.*, "Positive and negative effects of widespread badger culling on tuberculosis in cattle," *Nature,* vol. 439, pp. 843-846, 2006.
[51] C. Donnelly *et al.*, "Impacts of widespread badger culling on cattle tuberculosis: concluding analyses from a large-scale field trial.," *International Journal of Infectious Diseases,* vol. 11, pp. 300-308, 2007.
[52] H. C. J. Godfray *et al.*, "A restatement of the natural science evidence base relevant to the control of bovine tuberculosis in Great Britain," *Proceedings of the Royal Society B,* vol. 280, no. 1768, p. 20131634, 2013.
[53] F. Vial and C. Donnelly, "Localized reactive badger culling increases risk of bovine tuberculosis in nearby cattle herds," *Biol. Lett.,* vol. 8, pp. 50-53, 2012.
[54] C. Greentree, G. Saunders, L. McLeod, and J. Hone, "Lamb predation and fox control in south-eastern Australia," *J. Appl. Ecol.,* vol. 37, pp. 935-943, 2000.
[55] C. G. Radford, J. W. McNutt, T. Rogers, B. Maslen, and N. R. Jordan, "Artificial eyespots on cattle reduce predation by large carnivores," *Communications Biology Nature,* vol. 3, p. 430, 2020. [Online]. Available: https://doi.org/10.1038/s42003-020-01156-0 | www.nature.com/com.
[56] P. Fersterer, D. Nolte, G. Ziegltrum, and H. Gossow, "Effect of feeding stations on the home ranges of American black bears in western Washingto," *Ursus,* vol. 12, pp. 51-53, 2001.
[57] A. I. Leshner and V. J. Dzau, "Good gun policy needs research," *Science,* vol. 359, no. 6381, p. 1195, 2018.
[58] M. Krofel, R. Černe, and K. Jerina, "Effectiveness of wolf (Canis lupus) culling as a measure to reduce livestock depredations," *Acta Silvae et Ligni,* vol. 95, pp. 11-22, 2011.
[59] O. Liberg, G. Chapron, P. Wabakken, H. C. Pedersen, N. T. Hobbs, and H. Sand, "Shoot, shovel and shut up: cryptic poaching slows restoration of a large carnivore in Europe," *Proceedings of the Royal Society of London Series B,* vol. 270, pp. 91-98, 2012.
[60] G. Chapron and A. Treves, "Correction to 'Blood does not buy goodwill: allowing culling increases poaching of a large carnivore'," *Proceedings of the Royal Society B,* vol. 283, no. 1845, p. 20162577, 2016.
[61] F. J. Santiago-Ávila and A. Treves, "Poaching of protected wolves fluctuated seasonally and with non-wolf hunting," *Scientific Reports,* vol. 12, p. e1738 2022, doi: 10.1038/s41598-022-05679-w.
[62] J. Hone, V. A. Drake, and C. J. Krebs, "the effort–outcomes relationship in applied ecology: evaluation and implications," *Bioscience,* vol. 67, pp. 845–852, 2017.
[63] N. Salafsky, R. Margoluis, K. Redford, and J. G. Robinson, "Improving the practice of conservation: a conceptual framework and research agenda for conservation science," *Conserv. Biol.,* vol. 16, no. 6, pp. 1469-1479, 2002.
[64] J. González-González, E. Dorsey-Treviño, N. Alvarez-Villalobos, F. Barrera-Flores, A. Díaz González-Colmenero, C. Quintanilla-Sánchez, and e. al., "Trustworthiness of randomized trials in endocrinology—A systematic survey," *PLoS One,* vol. 14, no. 2, p. e0212360, 2019, doi: https://doi.org/10.1371/journal.pone.0212360.

accessed 19 August 2023.

Author contributions:

Conceptualization: AT

Methodology: AT, IK

Investigation: AT, IK

Visualization: AT

Funding acquisition: AT

Project administration: AT

Supervision: AT

Writing – original draft: AT

Writing – review & editing: AT, IK

Additional Information

Funding: AT acknowledges the receipt of a fellowship from the OECD Co-operative Research Programme: Sustainable Agricultural and Food Systems in 2022.

Competing interests: The authors declare no competing interest but readers can judge for themselves by accessing a full statement of AT's potentially competing interests at http://faculty.nelson.wisc.edu/treves/archive_BAS/funding.pdf, accessed 13 August 2023, with a complete CV at http://faculty.nelson.wisc.edu/treves/archive_BAS/Treves_vita_latest.pdf, accessed 13 August 2023.

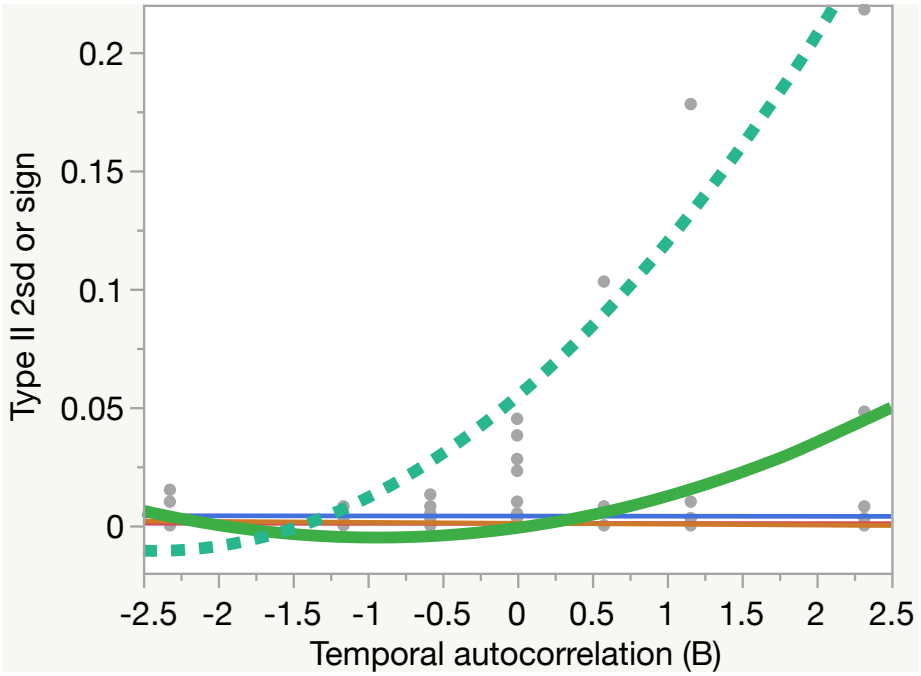

Figure 1. Severe Type II error resulting in reversal of the sign of correlation, in relation to temporal autocorrelation between L t and L t+1 (B). We present a curve fit by second-order ordinary least squares regression for visualization purposes only for each study design (dashed green = simple correlation, solid, thick green = nBACI, gold = RCT, purple = rBACI, red = crossover). The x-axis presents varying levels of temporal autocorrelation from Models 3 and 4 (Table S1). The y-axis presents the frequency of reversal of the true sign of correlation to the opposite sign estimated from 400 iterations of each combination of study design and value of B.

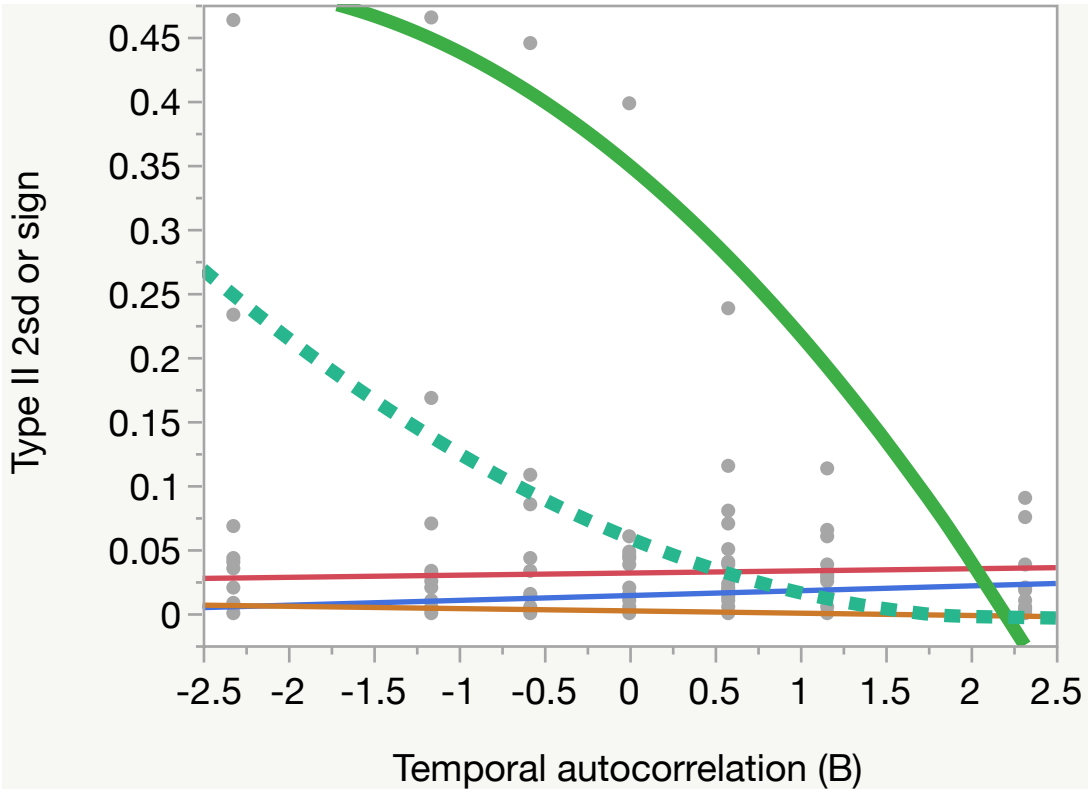

Figure 2. Overestimation of treatment effect in relation to temporal autocorrelation between L t and L t+1 (B). We present a curve fit by second-order ordinary least squares regression for visualization purposes only for each study design (dashed green = simple correlation, solid, thick green = nBACI, gold = RCT, purple = rBACI, red = crossover). The x-axis presents varying levels of temporal autocorrelation from Models 3 and 4 (Table S1). The y-axis presents the frequency of overestimation of treatment effect >2 SD below and above the mean estimated from 400 iterations per data point. Simulations are the same as in Fig. 1.

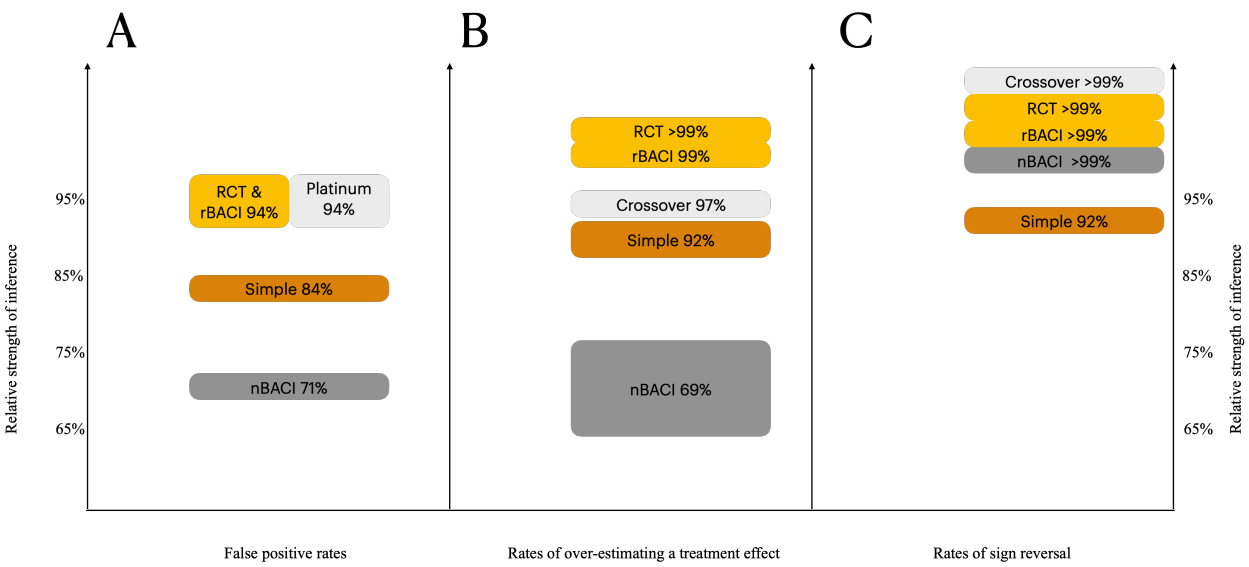

Figure 3. Relative strength of inference (100% - mean error rate) for crossover (platinum), RCT (gold), rBACI (gold), nBACI (silver), and simple correlation (bronze). The height of polygons is scaled to the 95% CI within each panel: (**A**) False positive rates, (**B**) Rates of overestimating as treatment effect, and (**C**) Rate of sign reversal. Side-by-side bars (e.g., panel A platinum and gold standards indicate identical mean and 95% CI but stacked bars indicate means were not identical (e.g., Panel C).

Table 1. Error rates estimated with and without background interactions: (**A)** B=1.16, (**B**) B=2.32, (**C**, **D**) T = 0 for Type I error; (**E–H)** are set to T = 0.58 x W.

	Simple correlation	nBACI	RCT †	rBACI †	Crossover design †	Simple correlation	nBACI	RCT †	rBACI †	Crossover design †
Models	**A. Background interactions 1.16**					**B. Background interactions 2.32**				
	C. Type I errors					**D. Type I errors**				
0	0.053	0.053	0.055	0.040	0.068	0.053	0.053	0.055	0.040	0.068
1	0.055	0.515				0.068	0.745			
2	0.068	0.548				0.060	0.718			
3	0.050	0.050	0.075	0.060	0.043	0.038	0.045	0.053	0.048	0.035
4	0.045	0.075	0.063	0.083	0.050	0.058	0.070	0.053	0.055	0.060
5	0.225	0.145				0.405	0.105			

6	0.223	0.595				0.435	0.743			
7	0.240	0.615				0.448	0.760			
8	0.218	0.158				0.455	0.088			
Models	**E. Type II errors, positive treatment**					**F. Type II errors, positive treatment**				
0	0.025	0.185	0.000	0.023	0.193	0.025	0.185	0.000	0.023	0.193
1	0.005	0.385				0.000	0.010			
2	0.000	0.000				0.000	0.000			
3	0.245	0.195	0.020	0.020	0.190	0.515	0.475	0.350	0.238	0.203
4	0.190	0.410	0.030	0.238	0.200	0.595	0.710	0.340	0.505	0.165
5	0.005	0.135				0.000	0.000			
6	0.210	0.915				0.365	0.890			
7	0.000	0.000				0.000	0.000			
8	0.190	0.000				0.350	0.015			
Models	**G. Type II errors, negative treatment**					**H. Type II errors, negative treatment**				
0	0.030	0.195	0.000	0.015	0.188	0.030	0.195	0.000	0.015	0.188
1	0.000	0.000				0.000	0.000			
2	0.000	0.440				0.000	0.005			
3	0.255	0.220	0.033	0.030	0.185	0.640	0.500	0.275	0.173	0.205
4	0.205	0.435	0.018	0.215	0.208	0.575	0.715	0.338	0.505	0.245
5	0.180	0.005				0.385	0.025			
6	0.005	0.005				0.005	0.005			
7	0.180	0.890				0.370	0.895			
8	0.000	0.075				0.000	0.000			

† Blank cells reflect that random assignment eliminates a correlation between W and L t.

Table 2. False positive rates (FPR) estimated from Type I and II error rates in Table 1 with background interactions: (**A**) B = 1.16 (**B**) B = 2.32, (**C**) positive treatment effect, (**D**) negative treatment effect.

	False positive rates (FPR) %									
Models	Simple correlation	nBACI	RCT †	rBACI †	Crossover design†	Simple correlation	nBACI	RCT †	rBACI †	Crossover design†
	A. Background interactions 1.16					**B. Background interactions 2.32**				
	C. Positive treatment ††									
0	5.2	6.1	5.2	3.9	7.8	5.2	6.1	5.2	3.9	7.8
1	5.5	45.6				6.4	42.9			
2	6.4	35.4				5.7	41.8			
3	6.2	5.8	7.1	5.8	5.0	7.3	7.9	7.5	5.9	4.2
4	5.3	11.3	6.1	9.8	5.9	12.5	19.4	7.4	10.0	6.7
5	18.4	14.4				28.8	9.5			
6	22.0	87.5				40.7	87.1			
7	19.4	38.1				30.9	43.2			
8	21.2	13.6				41.2	8.2			
	D. Negative treatment effect ††									
0	5.2	6.2	5.2	3.9	7.7	5.2	6.2	5.2	3.9	7.7
1	5.2	34.0				6.6	42.7			
2	6.4	49.5				5.7	41.9			
3	6.3	6.0	7.2	5.8	5.0	9.5	8.3	6.8	5.5	4.2
4	5.4	11.7	6.0	9.6	5.9	12.0	19.7	7.4	10.0	7.4
5	21.5	12.7				39.7	9.7			
6	18.3	37.4				30.4	42.8			
7	22.6	84.8				41.6	87.9			
8	17.9	14.6				31.3	8.1			
Minimum	**5.2**	**5.8**	**5.2**	**3.9**	**5.0**	**4.35.2**	**6.1**	**5.2**	**3.9**	**4.2**
95% CI of mean	**9–15**	**17–41**	**5–7**	**4–8**	**5–7**	**13–27**	**18–42**	**6–8**	**5–9**	**5–7**

† Blank cells reflect that random assignment eliminates a correlation between W and L t.

†† Simulated positive treatments may produce different FPR than negative treatments.