Submitted May 29, 2026

Still Money for Nothing?
Two Decades of Empirical Evaluation of Conservation Investments

Alex Caruana [1], Joseph W. Bull [1], Paul J. Ferraro [2], Hannah S. Wauchope [3], Alec P. Christie [4] and Julia P. G. Jones [5,6]

[1] University of Oxford, Department of Biology
[2] Johns Hopkins University, Carey Business School and the Department of Environmental Health and Engineering
[3] University of Edinburgh, School of Geosciences
[4] Imperial College London, Centre for Environmental Policy
[5] Bangor University, School of Environmental and Natural Sciences
[6] University of Utrecht, Department of Biology

ABSTRACT

Twenty years ago, the landmark paper "Money for Nothing?" argued that biodiversity conservation relied too little on empirical evidence. It called for more evaluations of conservation effectiveness based on explicit counterfactuals, comparing observed outcomes with those that would likely have occurred in the absence of intervention. To assess progress towards this goal, we systematically reviewed the study designs used to evaluate one of the most widely implemented conservation interventions: protected areas. Across 614 studies published over the past two decades, half still relied on simple Before-After or Control-Impact designs that do not reliably support causal inferences, although their use has declined in recent years. The other half used more formal causal identification strategies, most commonly conditioning strategies that control for observed confounders. However, most of these studies lacked pre-protection outcome data, limiting their ability to address unobserved confounders. Because causal claims depend on causal assumptions, it is notable that few studies stated these assumptions explicitly, let alone interrogated their plausibility. Although the design of conservation impact evaluations has advanced substantially, much remains to be improved. Combining causal inference methods with expanding data streams from remote sensing and biodiversity monitoring offers a major opportunity to strengthen the evidence base for conservation.

Keywords: Impact evaluation — Counterfactual thinking — Conservation policy — Evidence-based conservation — Socio-ecological systems

1. INTRODUCTION

> *"For far too long, conservation scientists and practitioners have depended on intuition and anecdote to guide the design of conservation investments. . . . If we want to ensure that our limited resources make a difference, we must accept that testing hypotheses about what policies protect biological diversity requires the same scientific rigor and state-of-the-art methods that we invest in testing ecological hypotheses" (Ferraro and Pattanayak 2006, p. 0482)*

In the two decades since these words were written, interest in using better methods to understand the impacts of conservation programmes has grown dramatically (Baylis et al. 2016; Ferraro et al. 2019), in what has been called biodiversity conservation's 'causal revolution' (Jones and Shreedhar 2024). Yet the extent to which this growing interest has translated into practice remains unclear. Now is an appropriate moment to assess progress, identify persistent challenges, and consider the future of conservation impact evaluation.

Robust impact evaluation requires causal inference: determining whether an observed association between two variables (e.g., a conservation programme and deforestation) reflects a cause-and-effect relationship, and, if so, how large

Email: alexander.caruana@hertford.ox.ac.uk

that effect is. To infer causality from data, researchers use several frameworks (Correia et al. 2026), among which the potential outcomes framework is one of the most widely used (Rubin 1974). Under this framework, causality is inferred by comparing two states: the observed world where a treatment (e.g. the conservation programme) is implemented and a counterfactual world where it is not (or where a different treatment is implemented). The treatment effect is the difference between these two states.

Simple study designs, such as comparing outcomes before and after an intervention (Before-After) or comparing sites with and without an intervention (Control-Impact), have long been used to evaluate conservation interventions and can offer valuable insights (Christie et al. 2021; Ockendon et al. 2021). However, these simple designs rest on the heroic assumption that no variables are systematically associated with both treatment assignment and the outcome; i.e., there are no confounders. In practice, this assumption rarely holds. As a result, these designs generally cannot credibly estimate counterfactual outcomes, rendering them unreliable for causal inference (Christie et al. 2019).

Broadly, researchers can draw causal inferences using four main types of study designs. Throughout the paper, we refer to these designs as *causal identification strategies*: designs used to determine whether an intervention caused an observed outcome, rather than the outcome arising because of other factors. In other words, they are strategies for ruling out rival explanations that may mask or mimic a causal effect of a conservation intervention. These four strategies are not mutually exclusive and are often combined: Conditioning Strategies (e.g., matching; see Stuart (2010); Abadie and Gardeazabal (2003); Sills et al. (2015); Siegel and Dee (2025)), Natural Experiments; (e.g., instrumental variables; see Rosenzweig and Wolpin (2000); Imbens and Lemieux (2008)), Identification through Mechanisms (e.g., front-door criterion; see Pearl (1995, 2009)), and Experimental Control over Treatment Exposure (e.g., randomized controlled trial).

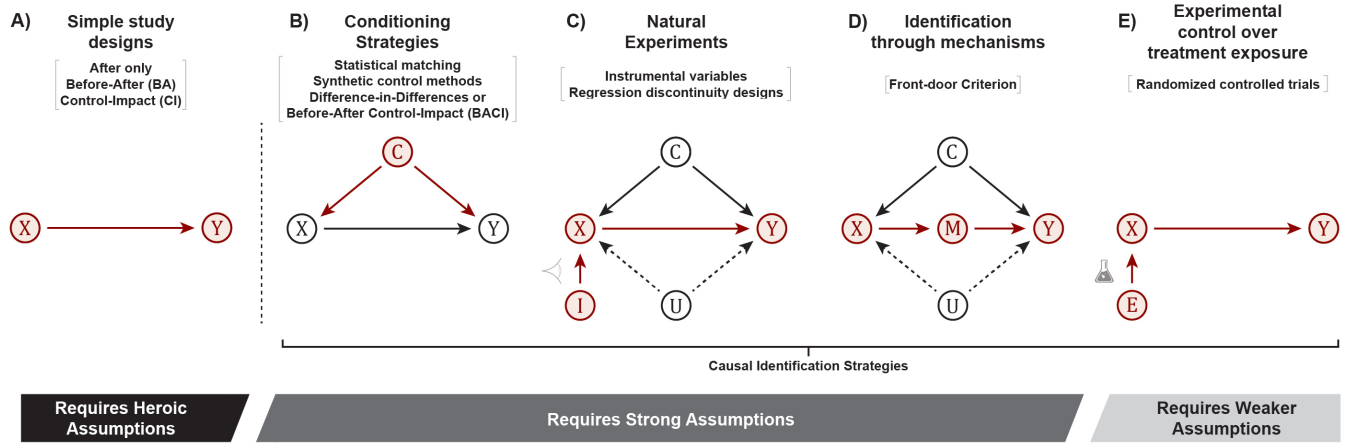


Figure 1. Study designs widely used in conservation impact evaluations and, in brackets, specific examples. To illustrate the assumptions underlying each design, we use Directed Acyclic Graphs (DAGs), where X represents the conservation intervention being evaluated, Y the outcome of interest, C the potential confounders that are explicitly controlled in the design, and U the potential uncontrolled confounders. Solid lines indicate possible causal relationships. Nodes highlighted in red indicate the key elements of each study design. (A) Simple study designs implicitly assume there are no confounders. (B) Conditioning strategies assume that treatment assignment is effectively random once the confounders are accounted for in the analysis (studies may acknowledge that uncontrolled confounders exist and test the sensitivity of results to them). (C) Natural experiments assume that there is an observable, unconfounded source of variation in the treatment (illustrated by I). (D) Identification through mechanisms (illustrated by M) assumes that the intermediate mechanisms M are observable and unconfounded; and (E) Experimental control over treatment exposure assumes that an experimenter (illustrated by E) has manipulated treatment exposure in a known, unconfounded way. Designs C and D allow one to estimate a causal effect even when U exists. This figure is adapted from Ferraro and Hanauer (2014).

Whichever study design is used, the credibility of the resulting conclusions depends on the plausibility of the underlying causal assumptions (Rubin 2005; Pearl 2009; Ferraro and Hanauer 2014). Although the terminology varies across disciplines, we consider four core assumptions: (a) Exchangeability; (b) Positivity; (c) Consistency; and (d) No Inter-

ference. Consistency and No Interference are sometimes grouped under the Stable Unit Treatment Value Assumption (SUTVA). We discuss these assumptions in greater detail in **Methods** and, in **Supplementary Information 1**, explain how our terminology maps onto related concepts used in different fields.

The credibility of study conclusions is also influenced by three other study features: causal reasoning, open science practices, and research integrity. We focus on two dimensions of causal reasoning: the data-generating process and the target causal effect. Readers can more easily evaluate the logic of a study design when the authors make the assumed data-generating process explicit through, for example, a theory of change or directed acyclic graph (Weiss 1995; Pearl 2009). Readers can better understand the population and conditions to which the study conclusions apply when the authors clearly define the causal effect being estimated, known as the estimand (Greifer and Stuart 2021; Barnard et al. 2024).

Readers have more trust in a study's conclusions when the authors use open science practices (Rosman et al. 2022; Song et al. 2022), such as pre-registration of analysis plans, which can reduce questionable research practices like selective reporting or p-hacking (Simonsohn et al. 2014) and presenting hypotheses developed after data analysis as if they had been specified in advance (i.e., Hypothesizing after results known or HARKing) (Kerr 1998). Questionable practices have been reported in the conservation-adjacent literature of ecology (Kimmel et al. 2023) and are encouraged by publication biases and donor preferences that favour large, statistically significant effects. Such practices can undermine the credibility of the evidence base (Parker et al. 2019; Pick et al. 2026).

Lastly, we focus on two dimensions of research integrity: conflicts of interest and ethics. When organisations or individuals involved in implementing an intervention also participate in its evaluation, the potential conflict of interest should be disclosed (Pynegar et al. 2025). When data are collected from people, authors should report how ethical issues were addressed and identify the authority that granted ethical approval (Brittain et al. 2020).

To document the evolution over the last twenty years of causal inference and impact evaluations in the conservation literature, we conduct a systematic review of studies estimating the impacts of one of the most common conservation interventions: protected areas. We include studies of terrestrial, marine and freshwater protected areas that report on ecological outcomes, socio-economic outcomes, human pressures, or combinations of the three.

Specifically, we ask 1) how study designs used to evaluate the impact of protected areas have changed over time, 2) how studies are distributed geographically and what kind of outcome measures are reported, 3) to what extent authors state, justify and interrogate the causal assumptions underlying their causal conclusions, and 4) to what extent studies report causal reasoning, open science practices, and research integrity measures that collectively increase confidence in their findings. We then consider emerging trends and discuss how conservation impact evaluations may develop over the next two decades.

2. METHODS

2.1. *Overview*

We aimed to identify all quantitative impact evaluations published since 2006 that estimate the causal effect of protected-area designation. Since our focus is on study designs rather than the outcomes themselves, the Population, Intervention, Comparison, Outcome framework recommended by the Collaboration for Environmental Evidence (Pullin et al. 2022) is not appropriate. Instead, to minimise bias, we adopted the Studies, Data, Methods, and Outcomes framework from the Cochrane Handbook for Systematic Reviews (Clarke et al. 2011; Munn et al. 2018). We also followed the Collaboration for Environmental Evidence guidelines (Pullin et al. 2022) and ROSES reporting standards (Haddaway et al. 2017a) to ensure search reliability and comprehensiveness (**Supplementary Information 2**). A detailed description of all methods is provided in **Supplementary Information 3**, with each stage transparently documented in Caruana et al. (2026).

2.2. *Search Protocol*

We adapted the search protocol from Langhammer et al. (2024), which assessed the global impact of conservation actions. The search string is based on the pressure-state-response model with an additional impact component.

Details on the search format and a comprehensive, replicable list of search strings are provided in **Supplementary Information 3, Section 1.1**. We conducted the search between July and August 2025 in two major electronic databases: Web of Science Core Collection and Scopus. We also conducted a separate grey literature search using ProQuest (for theses and dissertations) and 12 intergovernmental and non-profit organisation websites. Further details on search filters, institutional indexes, and the grey literature search are provided in **Supplementary Information 3, Section 1.2**.

In total, the database and the grey literature searches yielded 14,407 documents (**Figure 2**). After duplicate removal, 11,641 unique documents remained. We then screened titles and abstracts against our inclusion and exclusion criteria using ASReview, an open-sourced software that uses machine learning to reduce screening workload (Callaghan et al. 2024; van de Schoot et al. 2021; Quan et al. 2024).

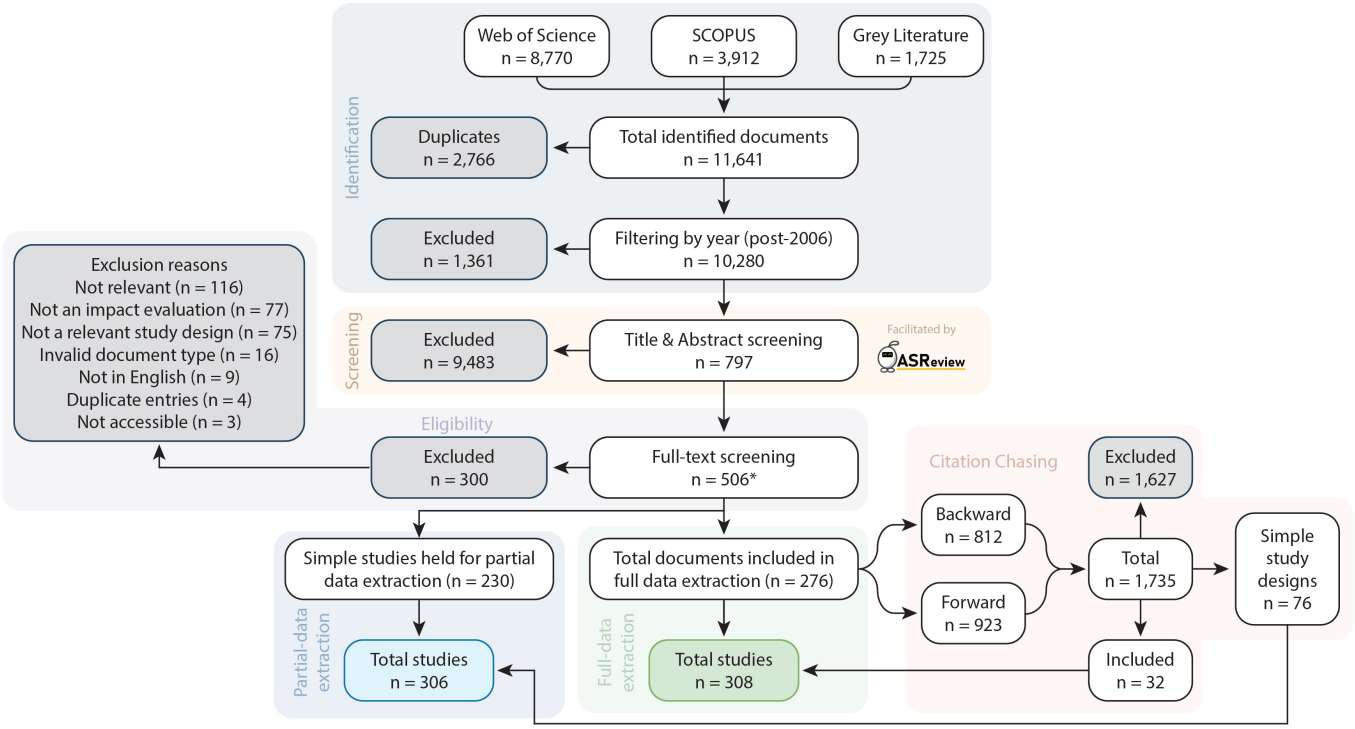


Figure 2. Summary of the systematic review search protocol presented using an adapted ROSES flow diagram (Haddaway et al. 2017b) Some retrieved documents were full PhD thesis; relevant chapters from these theses (n = 9) were extracted and added as separate entries in the review. We identified 306 studies using simple study designs, which were not taken to full data extraction. Full data were extracted from the 308 studies that reported one or more causal identification strategies.

Our inclusion and exclusion criteria were designed to capture any empirical impact evaluation of protected areas. During screening, studies that used simple study designs were retained for partial data extraction limited to the study design used. We did not assess whether these studies addressed causal assumptions because these designs do not make those assumptions explicit. Studies that used one of the remaining four causal identification strategies were retained for full data extraction so that we could assess whether, and how, causal assumptions were addressed. Details on these criteria, alongside the stopping heuristic, are provided in **Supplementary Information 3, Sections 2**. Title and abstract screening reduced the dataset to 506 documents, and full-text screening further reduced it to 276. Forward and backward citation chasing on 20 studies using Citationchaser (Haddaway et al. 2022) yielded 1,735 additional documents, of which 32 met inclusion criteria and were added to the dataset. The final dataset comprises 306 studies including simple study designs and 308 studies reporting one or more causal identification strategies, the latter of which were taken forward to full data extraction (**Supplementary Information 4**).

2.3. *Data Extraction*

For each included study, we recorded the years covered and the study design as described by the authors; where no design was explicitly stated, we inferred a broad category. We also recorded the main outcome variables analysed within each study (e.g., deforestation), alongside the datasets used. Based on this information, we classified each study as assessing ecological outcomes, socioeconomic outcomes, human pressures or some combination of the three.

To evaluate how well studies using a causal identification strategy addressed the four core assumptions underlying their causal claims (defined below), we developed a classification rubric with three components: (1) **Statement** of the assumption, (2) **Justification** of why it had been satisfied, and (3) **Interrogation** of its plausibility. A full explanation of the rubric is provided in **Supplementary Information 3, Section 3.3**. For each assumption, studies were assigned a score from Tier 1 to Tier 4. Tier 1 represents the aspirational benchmark, requiring all three components to be achieved. For each component that is missing, the study drops one tier, with Tier 4 studies missing all three components. Importantly, this rubric assesses the explicitness and comprehensiveness with which causal assumptions are reported; it does not measure overall research quality, nor does it evaluate the truth of the claims made. Although Tier 1 studies may, on average, be of higher quality than Tier 2 or 3 studies, this is not necessarily the case. We applied this rubric to all 308 documents taken to full data extraction.

Exchangeability means that the treated and control groups are interchangeable, such that the same outcomes would have been observed had the control group received the treatment or vice-versa. In idealized randomized experiments, this assumption is usually satisfied because the randomization balances both observed and unobserved confounders (Rosenbaum 2017). In conservation settings, however, exchangeability may be violated (Kimmel et al. 2021; Pynegar et al. 2025), which helps explain the common use of Conditioning Strategies as a causal identification strategy. Even so, the risk of an unobserved confounder remains, especially in studies using only post-treatment data. In the case of protected areas, for example, unobserved factors such as geological history may be correlated with protection and might also drive differences in species richness. To assess how robust the results are to unobserved confounders, researchers can use statistical methods like sensitivity testing (also referred to as *Quantitative bias analysis* in Epidemiology (Lash et al. 2014), including 'Rosenbaum's Bounds' (Rosenbaum et al. 2010), 'Sensemakr' (Cinelli and Hazlett 2020) and 'E-values' (VanderWeele and Ding 2017).

Positivity means that all units have a chance of being in treatment or control conditions. Randomized designs generally satisfy this assumption through random allocation of treatment, but observational studies face challenges (Hernán and Robins 2010). For instance, land tenure may restrict which sites could plausibly become protected, thereby limiting the availability of valid controls. When positivity is violated, conditioning on confounders is ineffective because the comparison sites are not true controls: they were never eligible for treatment.

Consistency means that the treatment is sufficiently well defined and uniformly applied across treated units (Hernán and Robins 2010). In the context of protected areas, this requires 'protection' to operate in broadly similar ways and with similar intensity across sites. In practice, this is rarely the case, as management regimes and implementation vary widely. This assumption is important when comparing effect sizes rather than simply the direction of effects – for example, when asking how much impact protected areas had, not just whether they had any impact at all. Although consistency cannot be tested directly, it remains important when interpreting heterogenous evidence across studies.

No interference means that the outcome of one unit is unaffected by the treatment status of other units. Violations of this assumption are commonly referred to as spillovers, which are plausible in many conservation settings (Ferraro et al. 2019; Pynegar et al. 2025). For example, protection at some sites may shift environmentally damaging behaviours to control sites where they would not have otherwise happened, making protection look more effective than it is. In contrast, protection at some sites could lead to reductions in damaging behaviours at nearby sites through, for example, wider enforcement of land use laws, which would make protection look less effective than it is.

We also recorded whether each study presented a Theory of Change or Directed Acyclic Graphs. Additionally, we recorded whether the study explicitly defined a causal estimand, such as the Average Treatment Effect (ATE) or the Average Treatment Effect on the Treated (ATT). Lastly, we noted whether the authors provided an archived data

and code to support computational reproducibility, whether a conflict of interest (competing interest) statement was included, and whether any ethics approval was reported. The full data extraction table and additional details on all extracted variables are provided in **Supplementary Information 3, Section 3**. All data extraction was conducted through Covidence (https://www.covidence.org/).

3. RESULTS

3.1. *Trends in study designs used over time*

Half of the 614 studies evaluating the impact of protected areas published over the past 20 years used simple study designs (i.e., Control-Impact, Before-After, and After-Only) (n = 306). Simple study designs were even more common among the studies that focused on marine and freshwater protected areas (**Supplementary Figures 7 - 9**). Nevertheless, in the last five years, the use of simple designs declined to about one-third of studies.

Among the studies that used a causal identification strategy, most used conditioning strategies (97.4%). Few studies used natural experiments (3.9%), and none used identification through mechanisms or experimental control over treatment. Experiments of environmental policy interventions, including in conservation, remain rare (Ferraro et al. 2023), likely because of technical and ethical constraints (Pynegar et al. 2025). We also sub-classified the causal identification studies based on whether they used pre-protection outcome data, which help researchers control for unobserved confounders (e.g., synthetic control methods or statistical matching with differences-in-differences) (Cook et al. 2008; Ferraro and Miranda 2017): only 40.9% did (**Figure 3**). This implies that only 20.5% of all 614 studies published in the last twenty years used the powerful combination of a causal identification strategy and data on pre-treatment outcome values.

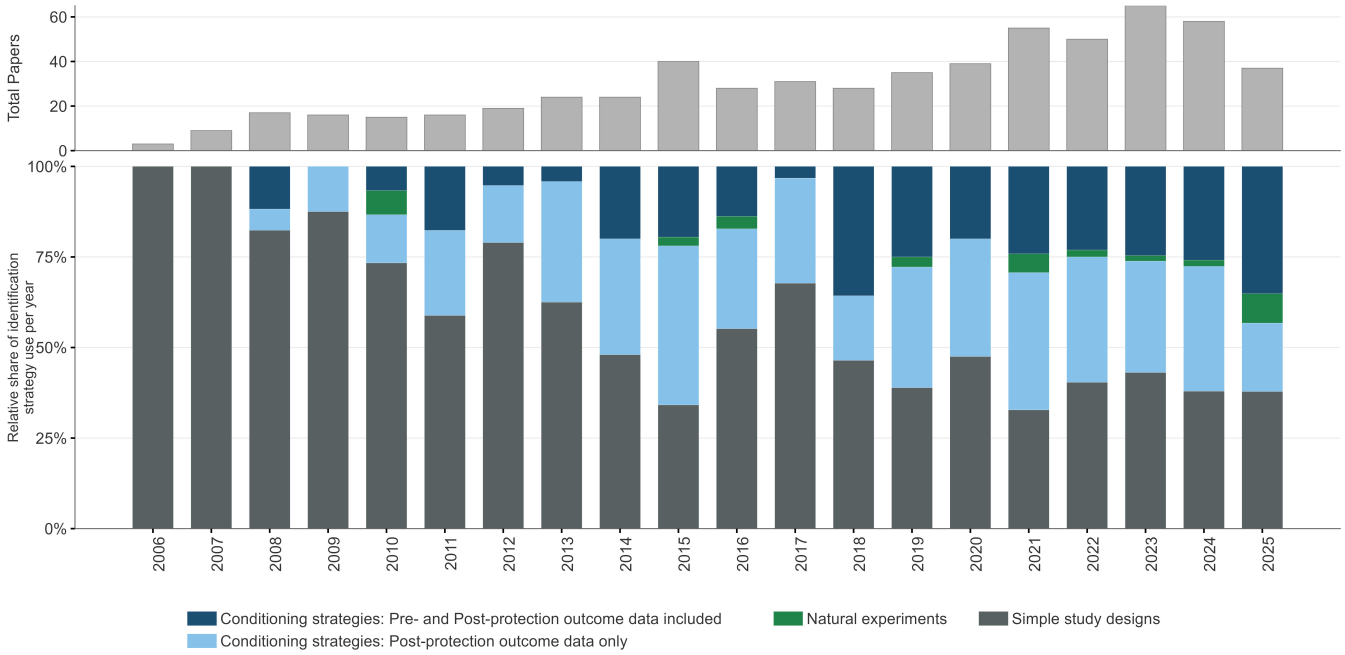


Figure 3. Shows the 20-year trend in published study designs used to evaluate the impacts of protected areas. About half (306) used simple study designs that are unreliable for causal inferences, although their use has decreased over time. The other half (308) used causal identification strategies, among which the conditioning strategies were most popular. Few studies used natural experiments, and none used identification through mechanisms or experimental control over treatment. Among the studies using a causal identification strategy, less than half used pre-protection outcome data that can help control for unobserved confounders.

3.2. *Geography and outcomes in impact evaluation using causal identification strategies*

The 308 studies using a causal identification strategy covered the globe, with China being the most frequently studied country (18.2%), followed by Brazil (10%) and the United States of America (5.1%) (**Figure 4:A**). Most studies focused on terrestrial protected areas (82.1%), with fewer examining marine (17.2%) and freshwater (0.7%) protected areas. Impact evaluations were most frequently conducted on a regional scale (37.3%), followed by national scale (29.9%) and individual protected areas (12.7%). In 145 studies (47.1%), we could identify the protected areas under study. Among these 4,106 protected areas, 2,322 (56.6%) were successfully linked to records within the World Database on Protected Areas (UNEP-WCMC and IUCN 2026)(http://protectedplanet.net/) (**Figure 4:B**). Most studies focused on ecological outcomes (52.3%), followed by socioeconomic outcomes (18.2%) and human pressures (12%). The remaining 54 studies looked at a combination of the three (17.5%). The most studied outcome variable for terrestrial protected areas was deforestation (42.3%) (**Figure 4:C**), while for marine protected areas it was biodiversity metrics (73.6%) (**Figure 4:D**). The full dataset, including a searchable Excel spreadsheet detailing all studies and identified protected areas, is available in **Supplementary Information 4**. Additional descriptive statistics are available in **Supplementary Information 5**.

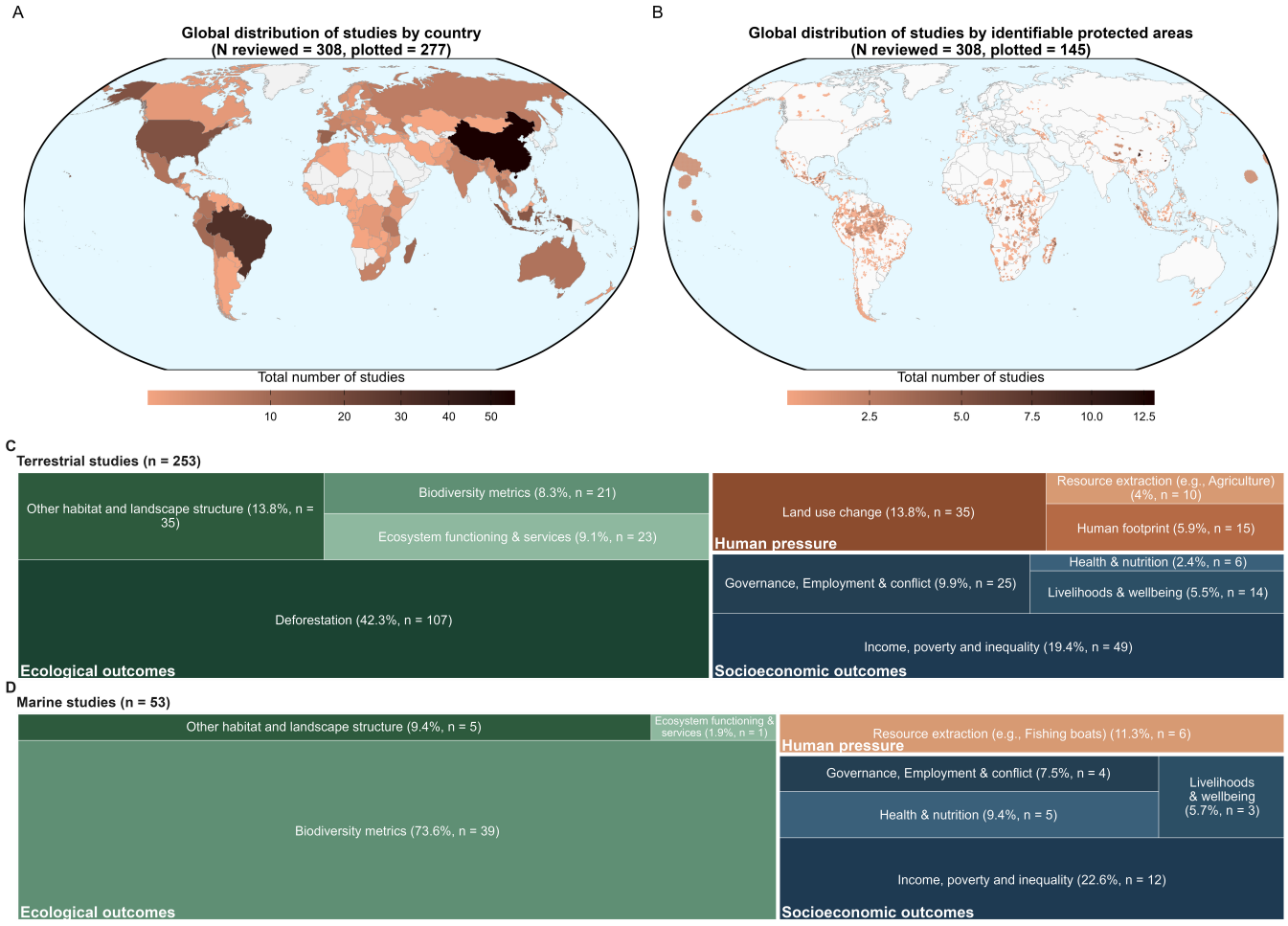


Figure 4. (A) Geographic distribution of impact evaluations by country, excluding those conducted on a global scale (n = 31). B) Spatial distribution of identifiable protected areas in the studies (n = 2,322) across 100 km hexagonal bins. Darker colours indicate areas with overlapping studies within the same protected areas. (C) Distribution of outcome variables assessed in terrestrial protected area studies (n = 253). (D) Distribution of outcome variables assessed in marine protected area studies (n = 53). In both panels, rectangle size is proportional to the number of studies assessing that outcome variable, grouped by broad classification (ecological outcomes, socioeconomic outcomes, and human pressures).

3.3. *Extent to which causal assumptions are explicitly stated, justified and interrogated*

Studies varied in terms of the clarity and comprehensiveness with which the causal assumptions that underpin causal claims are addressed (**Figure 5**). Recall that Tier 1 is the aspirational tier in which a study states, justifies and interrogates a core causal assumption. Although some studies achieved Tier 1 on one or more assumptions, no study achieved Tier 1 for all of them (**Supplementary Figure 10**). A breakdown of the tier classifications for each individual study is available in **Supplementary Information 5**. The descriptive results for each assumption are discussed below.

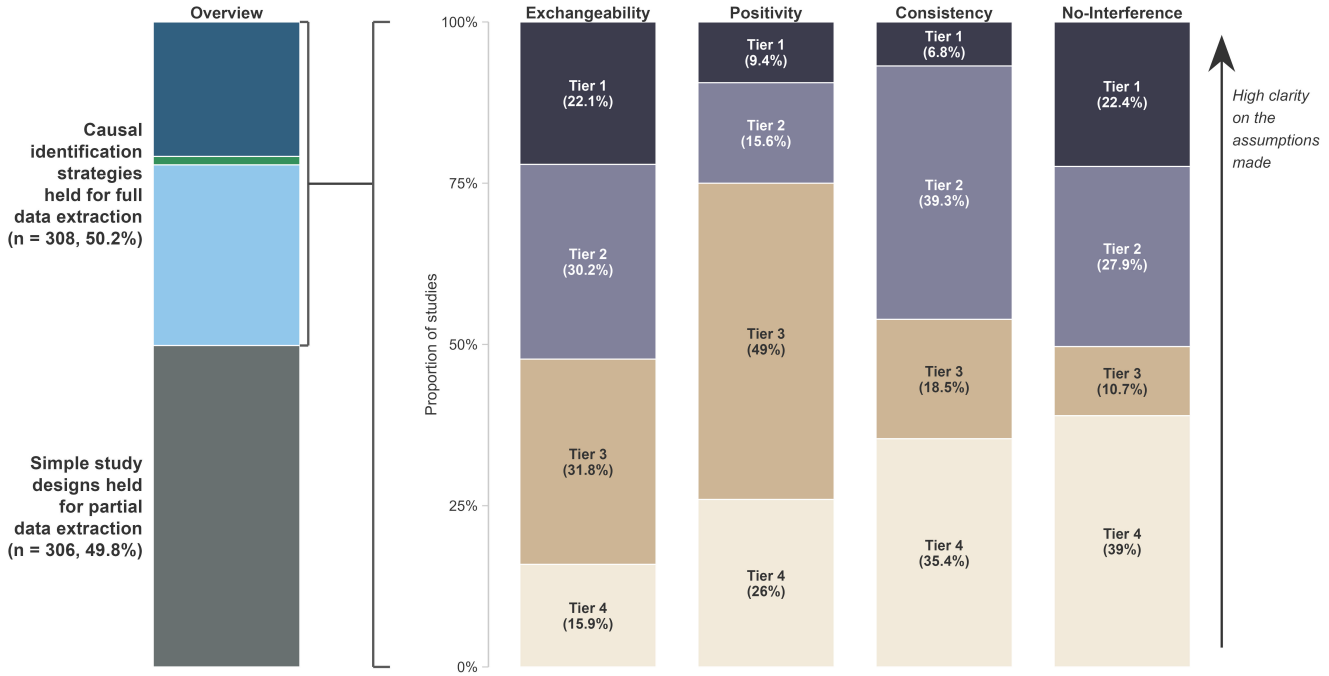


Figure 5. The extent to which studies using causal identification strategies (n = 308) state, justify and interrogate the four core causal assumptions that underpin causal claims. Studies that state, justify and interrogate the assumption are classified as Tier 1. For each of the three actions that is missing, the study drops one tier. The overview column represents a condensed version of Figure 3, with dark blue representing conditioning strategies that used pre-protection outcome data, light blue representing those that only used post-protection outcome data, green representing natural experiments, and grey representing simple study designs. Simple study designs are likely to be Tier 4 because these designs are not typically explicit about their assumptions.

A statement of the exchangeability assumption was most common, appearing in 257 studies (83.4%), followed by a justification in 138 studies (44.8%), and an interrogation in 93 studies (30.2%). Overall, 68 studies (22.1%) did all three actions (Tier 1). Forty-nine studies (15.9%) did none of them (Tier 4).

The positivity assumption was seldom stated explicitly (18.5%) or justified (15.6%). Instead, most studies addressed this assumption only implicitly through their choice of study design (49.0%). For instance, authors using statistical matching designs often implicitly try to satisfy positivity by discarding unmatched units. Only 29 studies (9.4%) were classified as Tier 1, whereas 81 studies (26.3%) were classified as Tier 4.

A statement of the consistency assumption was present in 201 studies (65.3%), followed by justification in 142 studies (46.1%) and interrogation in 112 studies (35.4%). Among the 112 studies that interrogated this assumption, the majority (n = 90) explored heterogeneity based on protected area classes, such as IUCN management categories or levels of regulatory strictness. Such analyses are important, but only partially interrogate the consistency assumption. To receive a Tier 1 designation, we required studies to move beyond these broad classes and address heterogeneity based on other protected area characteristics, such as governance quality, size, or biophysical context. Only 6.8% of studies did so. A total of 107 studies (34.7%) were in Tier 4.

Compared with the other three assumptions, the no-interference assumption had the highest proportion of studies in Tier 1 (22.4%) and Tier 4 (39.0%). In terms of individual components, 185 studies (60.1%) stated the assumption, 156 (50.6%) provided a justification, and 71 (23.1%) interrogated it.

3.4. *Reporting of causal reasoning, open science practises and research integrity measures*

Explicit theories of change or directed acyclic graphs were rarely presented, appearing in only 24 (7.8%) and 9 (2.9%) studies respectively. Estimands were specified in 121 studies (39.3%), with the Average Treatment Effect on the Treated (ATT) being the most reported (66.4%), followed by the Average Treatment Effect (ATE) (19.2%). Few studies reported open science practices. Only three studies (1.0%) were pre-registered, 40 studies (13.0%) provided archived code, 44 (14.3%) studies archived their datasets, and 31 studies (10.1%) provided both code and data. The remaining 201 studies (65.4%) either provided no information on the availability of code or data, or they reported that it was confidential. A single study (0.3%) included an explicit conflict-of-interest statement. Lastly, of the 308 studies reviewed, 34 collected human data in-situ (e.g., field surveys), of which only 8 (23.5%) explicitly reported ethics approval.

4. DISCUSSION

In the two decades since Ferraro and Pattanayak (2006) called for more, and better, empirical evaluations of conservation programmes, the field has changed substantially. Repeated calls for counterfactual thinking and better-designed evaluations (Miteva et al. 2012; Baylis et al. 2016; Butsic et al. 2017; Correia et al. 2026) have contributed to a marked shift in practice, mirroring transitions that occurred earlier in disciplines such as economics (Imbens 2024; Garg and Fetzer 2026). There is now a thriving community of conservation practitioners, scientists and policy makers engaged in conservation impact evaluation, and conservation organizations increasingly include staff dedicated to evaluating impacts (Jones and Shreedhar 2024; O'Garra et al. 2025). This momentum is reflected in initiatives such as the Society for Conservation Biology's Impact Evaluation Working Group (https://scb-impact.org/). This shift in practice is also beginning to influence policy, with impact evaluations increasingly cited in major international biodiversity assessments (IPBES 2019, 2025). In contrast, the 2005 Millenium Ecosystem Assessment had cited few such citations, reporting that (Millennium Ecosystem Assessment (MEA) 2005, p. 122) , "Few well-designed empirical analyses assess even the most common biodiversity conservation measures".

Our review shows more than half of all published studies on the effects of protected areas now use a formal causal identification strategy. These studies span a broad range of outcomes, including deforestation (Morgans et al. 2024; den Braber et al. 2024), land-use change (Gonçalves-Souza et al. 2021; Li et al. 2024), poverty (Keane et al. 2020), public health (Sheehan et al. 2024), and species populations (Wauchope et al. 2022; Santangeli et al. 2023). This is clear evidence of methodological progress.

However, while the rigour of conservation impact evaluations has increased substantially, considerable room for improvement remains. Some studies still rely on simple Before-After or Control-Impact designs, which require strong and often implausible assumptions. Among studies using a causal identification strategy, nearly 60% do not include pre-treatment data on outcomes, limiting their ability to account for unobserved confounders. Fewer than 10% of studies present causal reasoning through a theory of change or directed acyclic graph, and only 39.3% specify the causal effect they aim to estimate. Most strikingly, the assumptions on which causal claims depend are rarely acknowledged, let alone justified or interrogated.

4.1. *Barriers to well-designed evaluations*

To explain why well-designed evaluations of conservation programmes were not being conducted, Ferraro and Pattanayak (2006) identified several barriers. These barriers fall broadly into two categories: limited institutional knowledge about how to design impact evaluations, and the technical constraints associated with conducting them.

We see clear progress on institutional knowledge. Over the past two decades, conservation researchers, practitioners, and funders have developed a much stronger awareness of counterfactual thinking (Curzon and Kontoleon 2016; Jones and Shreedhar 2024). In recent years, more than half of the studies in our review used a causal identification strategy,

and similar approaches are now being applied to a wider range of conservation interventions beyond protected areas, including pay-to-release schemes (Booth et al. 2025), carbon and biodiversity offsets (zu Ermgassen et al. 2023; Swinfield et al. 2026), and rhino dehorning programmes (Kuiper et al. 2025). Conservation organisations and funders have also become more familiar with these approaches, with some funders now explicitly encouraging applicants to incorporate counterfactual thinking into their grant applications (Smith et al. 2026). Institutional understanding has not fully caught up with methodological best practice, but it is no longer the barrier it once was.

Technical barriers, by contrast, remain substantial. Data limitations continue to restrain evaluation designs. Ideally, studies would include data from before and after an intervention, for both treated and control sites. It is therefore unsurprising that many evaluations in our review focus on deforestation, for which suitable temporal and spatially explicit data for the globe became widely available in 2013 (Hansen et al. 2013). Initiatives such as the Global Biodiversity Information Facility (GBIF) are expanding the availability of biodiversity data (Lane and Edwards 2016; GBIF.org 2026), but limitations in temporal and spatial resolution remain a barrier. Time lags between intervention and measurable biodiversity responses pose an additional challenge. In many cases, the only practical solution is to evaluate intermediate outcomes within a well-articulated theory of change. Cost is another perceived barrier. Although well-designed impact evaluations can reduce costs by discouraging investments in ineffective programmes, they may still be viewed as expensive or difficult to integrate into implementation (Ferraro et al. 2023). Addressing this will require stronger collaborations among researchers, practitioners, and funders (Jones and Shreedhar 2024). Finally, concerns about reputational risk may also discourage engagement with impact evaluations (Asquith 2020), particularly when null or negative findings are possible. Such concerns can contribute to positivity bias in the conservation evidence base, in which null and negative findings are systematically under-reported (Redford and Taber 2000; Catalano et al. 2019; Chambers et al. 2022).

4.2. *Areas of improvement for conservation impact evaluations*

Most studies in our review rely exclusively on post-treatment outcome data. With only post-treatment outcome data, authors find it challenging to credibly demonstrate that treated and control units were comparable before treatment. Geldmann et al. (2025) illustrated this challenge when they showed that the results from a study evaluating the impact of protected areas on biodiversity with only post-protection data was highly vulnerable to bias from realistic, but uncontrolled, pre-existing differences between protected and unprotected sites (i.e., bias from unobserved confounders). We acknowledge that reliance on post-treatment data often reflects genuine data limitations, and we do not argue that such studies should not be conducted. However, researchers should be explicit about the limitations of these designs and cautious in interpreting their findings (Cook et al. 2008; Ferraro and Miranda 2017). Testing the sensitivity of results to unobserved confounders should also become more common in conservation impact evaluations (Cinelli and Hazlett 2020; Jones et al. 2022; Guizar-Coutiño et al. 2026). More broadly, researchers should interrogate all the assumptions on which their causal claims rest, a practice our review finds to be uncommon. As documented here and in recent reviews (Kimmel et al. 2021; Siegel and Dee 2025; Baylis et al. 2026; Correia et al. 2026), these assumptions are rarely stated, much less interrogated, despite being central to the validity of causal claims.

A second area for improvement concerns the specification of the causal effect being targeted for estimation, known as the estimand. More than half of the studies in our review do not clearly state the estimand. This may appear to be a minor reporting omission, but it has important implications. When studies do report an estimand, most report estimating an Average Treatment Effect on the Treated (ATT): the expected effect of protection on areas that were actually protected (i.e., not the expected effect on a randomly chosen site). Estimating the ATT requires weaker, and thus more credible, assumptions about counterfactual outcomes and is useful for understanding impacts on areas that have been protected. It does not, however, provide any information about the expected effects on areas not yet protected, and thus sheds little light on the potential impacts of expanding protected areas into new locations (Geldmann et al. 2025), a topic that is central to modern conservation debates. An estimand that would be useful for debates about protected area expansion would be the average treatment effect on the untreated (ATU): the expected effect of protection on areas that have not yet been protected (Greifer and Stuart 2021), for which estimation requires different causal assumptions than those invoked for estimating the ATT. When authors do not make their estimand explicit, they make it harder for others to evaluate their causal claims and they reduce the usefulness of their evidence.

4.3. *The future of conservation impact evaluations*

Addressing these limitations will require not only greater methodological awareness but also tools that enable more rigorous designs. Encouragingly, recent advances in both causal inference and Earth-observation science are creating new opportunities for impact evaluation in biodiversity conservation.

One promising development is causal machine learning. This rapidly growing interdisciplinary field builds on structural causal models and uses flexible algorithms to estimate treatment effects from large and complex datasets (Kaddour et al. 2025). Unlike conventional predictive machine learning (Esteva et al. 2019), causal machine learning is designed to estimate the effects of interventions and, when certain assumptions are satisfied, it can estimate individual-level treatment effects rather than only population- or sample-average estimands. These methods have gained traction in medicine under the banner of *precision medicine*, where individualised treatment effect estimation can help guide clinical decisions (Petito et al. 2020; Feuerriegel et al. 2024). Related ideas have only recently appeared in ecology, for example in the emerging concept of *precision ecology* (Spake et al. 2025). These developments are promising, but they should not be oversold. Causal machine learning does not eliminate the need for strong, untestable causal assumptions. Confounding from unobserved variables remains a threat, just as it does in more conventional approaches such as statistical matching. Researchers must therefore continue to state, justify, and interrogate the assumptions underlying their causal claims. That said, recent methodological advances have begun to make sensitivity analysis more feasible for treatment effects estimated using causal machine learning, which could provide a valuable path forward (Chernozhukov et al. 2021; Cinelli et al. 2026).

A second area of promise is Earth observation, particularly the emergence of geospatial foundation models. Remote sensing has long played an important role in conservation impact evaluations, but most conventional remote-sensing products are tailored to a single outcome, such as the Global Forest Change dataset for deforestation (Hansen et al. 2013) or NASA's MCD64A1 global burned area product (Giglio et al. 2018). Geospatial foundation models represent an important advance because they are trained on vast quantities of remote-sensing data and generate rich, multi-dimensional representations of the Earth's surface (Xiao et al. 2025). One example is TESSERA (Temporal Embeddings of Surface Spectra for Earth Representation and Analysis), an open-source model providing global coverage at 10-metre spatial resolution with 128-dimensional annual embeddings (Feng et al. 2025). Each pixel in the model carries 128 labels that jointly encode the characteristics of the vegetation and its phenology based on an unsupervised classification. These rich representations can be combined with field observations to model species distributions with high accuracy (Ball et al. 2026). In principle, they could therefore support the generation of annual biodiversity outcome layers at fine spatial resolution, helping to address one of the most persistent constraints on conservation impact evaluations.

Taken together, these advances point to an important next phase for the field. Better methods and richer data will not remove the need for careful study design, but they do expand what is possible. Both merit close attention from conservation researchers.

5. CONCLUSION

The causal revolution has firmly taken hold in biodiversity conservation. Causal inference is no longer an emerging trend in conservation science but rather a mainstream focus. The studies reviewed here are not without limitations, and improving the transparency of the core causal assumptions and the application of causal identification strategies remains an important priority. Yet the direction of travel is encouraging. Rapid advances in causal machine learning, the increasing availability of biodiversity data, and advances in earth-observation science are expanding the methodological toolkit available, promising more robust answers to questions about the effectiveness of interventions on a much wider range of outcomes. The quality of conservation investment decisions is better for this shift, and the foundations are in place to strengthen further the empirical evaluation of conservation interventions in the future.

ACKNOWLEDGMENTS

We want to thank Rachel Neugarten for sharing her private database of impact evaluations as it helped us develop our benchmark and refine our search strategies. J.P.G.J. thanks the Prince Bernhard Chair Foundation.

DATA AVAILABILITY

This manuscript's data, code and project files are all publicly available on Zenodo (Caruana et al. 2026).

AUTHOR CONTRIBUTIONS

A.C.: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Visualization, Writing – original draft, Writing – review & editing; J.W.B.: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing – review & editing. P.J.F.: Conceptualization, Investigation, Methodology, Validation, Writing – review & editing; H.S.W.: Investigation, Methodology, Validation, Visualization, Writing – review & editing; A.P.C.: Investigation, Methodology, Visualization, Writing – review & editing; J.P.G.J.: Conceptualization, Investigation, Methodology, Supervision, Validation, Visualization, Writing – review

CONFLICT OF INTEREST DECLARATION

We declare we have no competing interests.

FUNDING

A.C. is funded by the Clarendon Fund, Oxford-NaturalMotion Scholarship and the Hertford College Graduate Scholarship.

THEME

This article contributes to the upcoming theme issue *Causal inference: an interdisciplinary exchange of methodological approaches to enhance validity of ecology and evolutionary biology research* within the Philosophical Transactions of the Royal Society B.

REFERENCES

Abadie, A. and Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132. https://doi.org/10.1257/000282803321455188.

Asquith, N. (2020). Large-scale randomized control trials of incentive-based conservation: What have we learned? *World Development*, 127:104785. https://doi.org/10.1016/j.worlddev.2019.104785.

Ball, J. G., Wicklein, J. A., Feng, Z., Knezevic, J., Jaffer, S., Madhavapeddy, A., Atzberger, C., Dalponte, M., and Coomes, D. (2026). Geospatial foundation models enable data-efficient tree species mapping in temperate mountain forests. *bioRxiv*, pages 2026–02. https://doi.org/10.64898/2026.02.23.707022.

Barnard, M., Huling, J. D., and Wolfson, J. (2024). A Unified Framework for Causal Estimand Selection. *arXiv preprint arXiv:2410.12093*. https://doi.org/10.48550/arXiv.2410.12093.

Baylis, K., Garcia, A., and Heilmayr, R. (2026). Causal Inference for Biodiversity Conservation. *Review of Environmental Economics and Policy*, 20(1):000–000. https://doi.org/10.1086/739836.

Baylis, K., Honey-Rosés, J., Börner, J., Corbera, E., Ezzine-de Blas, D., Ferraro, P. J., Lapeyre, R., Persson, U. M., Pfaff, A., and Wunder, S. (2016). Mainstreaming Impact Evaluation in Nature Conservation. *Conservation Letters*, 9(1):58–64. https://doi.org/10.1111/conl.12180.

Booth, H., Pienkowski, T., Ramdlan, M. S., Naira, K. B., , M., Milner-Gulland, E. J., Adrianto, L., and Ferraro, P. J. (2025). Conservation impacts and hidden actions in a randomized controlled trial of a marine pay-to-release program. *Science Advances*, 11(17):eadr1000. https://doi.org/10.1126/sciadv.adr1000.

Brittain, S., Ibbett, H., de Lange, E., Dorward, L., Hoyte, S., Marino, A., Milner-Gulland, E. J., Newth, J., Rakotonarivo, S., Veríssimo, D., and Lewis, J. (2020). Ethical considerations when conservation research involves people. *Conservation Biology*, 34(4):925–933. https://doi.org/10.1111/cobi.13464.

Butsic, V., Lewis, D. J., Radeloff, V. C., Baumann, M., and Kuemmerle, T. (2017). Quasi-experimental methods enable stronger inferences from observational data in ecology. *Basic and Applied Ecology*, 19:1–10. https://doi.org/10.1016/j.baae.2017.01.005.

Callaghan, M., Müller-Hansen, F., Bond, M., Hamel, C., Devane, D., Kusa, W., O'Mara-Eves, A., Spijker, R., Stevenson, M., Stansfield, C., Thomas, J., and Minx, J. C. (2024). Computer-assisted screening in systematic evidence synthesis requires robust and well-evaluated stopping criteria. *Systematic Reviews*, 13(1):284. https://doi.org/10.1186/s13643-024-02699-7.

Caruana, A., Bull, J. W., Ferraro, P. J., Wauchope, H. S., Christie, A. P., and Jones, J. P. G. (2026). Still money for nothing? Two decades of empirical evaluation of conservation investments V1.0 [Dataset]. https://doi.org/10.5281/zenodo.20439985.

Catalano, A. S., Lyons-White, J., Mills, M. M., and Knight, A. T. (2019). Learning from published project failures in conservation. *Biological Conservation*, 238:108223. https://doi.org/10.1016/j.biocon.2019.108223.

Chambers, J. M., Massarella, K., and Fletcher, R. (2022). The right to fail? Problematizing failure discourse in international conservation. *World Development*, 150:105723. https://doi.org/10.1016/j.worlddev.2021.105723.

Chernozhukov, V., Cinelli, C., Newey, W. K., Sharma, A., and Syrgkanis, V. (2021). Long story short: Omitted variable bias in causal machine learning. *The Review of Economics and Statistics*, pages 1–45. https://doi.org/10.48550/arXiv.2112.13398.

Christie, A. P., Amano, T., Martin, P. A., Petrovan, S. O., Shackelford, G. E., Simmons, B. I., Smith, R. K., Williams, D. R., Wordley, C. F. R., and Sutherland, W. J. (2021). The challenge of biased evidence in conservation. *Conservation Biology*, 35(1):249–262. https://doi.org/10.1111/cobi.13577.

Christie, A. P., Amano, T., Martin, P. A., Shackelford, G. E., Simmons, B. I., and Sutherland, W. J. (2019). Simple study designs in ecology produce inaccurate estimates of biodiversity responses. *Journal of Applied Ecology*, 56(12):2742–2754. https://doi.org/10.1111/1365-2664.13499.

Cinelli, C., Chernozhukov, V., Syrgkanis, V., and Wang, J. (2026). Sensitivity Analysis for Debiased Machine Learning. https://carloscinelli.com/dml.sensemakr/index.html.

Cinelli, C. and Hazlett, C. (2020). 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.

Clarke, M., Oxman, A., Paulsen, E., Higgins, J., and Green, S. (2011). Appendix A: Guide to the contents of a Cochrane Methodology protocol and review. *Cochrane Handbook for systematic reviews of interventions*.

Cook, T. D., Shadish, W. R., and Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 27(4):724–750. https://doi.org/10.1002/pam.20375.

Correia, H. E., Dee, L. E., Byrnes, J. E. K., Fieberg, J. R., Fortin, M.-J., Glymour, C., Runge, J., Shipley, B., Shpitser, I., Siegel, K. J., Sugihara, G., von Holle, B., and Ferraro, P. J. (2026). Best practices for moving from correlation to causation in ecological research. *Nature Communications*, 17(1):1981. https://doi.org/10.1038/s41467-026-69878-z.

Curzon, H. F. and Kontoleon, A. (2016). From ignorance to evidence? The use of programme evaluation in conservation: Evidence from a Delphi survey of conservation experts. *Journal of Environmental Management*, 180:466–475. https://doi.org/10.1016/j.jenvman.2016.05.062.

den Braber, B., Oldekop, J. A., Devenish, K., Godar, J., Nolte, C., Schmoeller, M., and Evans, K. L. (2024). Socio-economic and environmental trade-offs in Amazonian protected areas and Indigenous territories revealed by assessing competing land uses. *Nature Ecology & Evolution*, 8(8):1482–1492. https://doi.org/10.1038/s41559-024-02458-w.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29. https://doi.org/10.1038/s41591-018-0316-z.

Feng, Z., Atzberger, C., Jaffer, S., Knezevic, J., Sormunen, S., Young, R., Lisaius, M. C., Immitzer, M., Jackson, T., and Ball, J. (2025). Tessera: Temporal embeddings of surface spectra for earth representation and analysis. *arXiv preprint arXiv:2506.20380*. https://doi.org/10.48550/arXiv.2506.20380.

Ferraro, P. J., Cherry, T. L., Shogren, J. F., Vossler, C. A., Cason, T. N., Flint, H. B., Hochard, J. P., Johansson-Stenman, O., Martinsson, P., Murphy, J. J., Newbold, S. C., Thunström, L., van Soest, D., van 't Veld, K., Dannenberg, A., Loewenstein, G. F., and van Boven, L. (2023). Create a culture of experiments in environmental programs. *Science*, 381(6659):735–737. https://doi.org/10.1126/science.adf7774.

Ferraro, P. J. and Hanauer, M. M. (2014). Advances in Measuring the Environmental and Social Impacts of Environmental Programs. *Annual Review of Environment and Resources*, 39(Volume 39, 2014):495–517. https://doi.org/10.1146/annurev-environ-101813-013230.

Ferraro, P. J. and Miranda, J. J. (2017). Panel Data Designs and Estimators as Substitutes for Randomized Controlled Trials in the Evaluation of Public Programs. *Journal of the Association of Environmental and Resource Economists*, 4(1):281–317. https://doi.org/10.1086/689868.

Ferraro, P. J. and Pattanayak, S. K. (2006). Money for Nothing? A Call for Empirical Evaluation of Biodiversity Conservation Investments. *PLOS Biology*, 4(4):e105. https://doi.org/10.1371/journal.pbio.0040105.

Ferraro, P. J., Sanchirico, J. N., and Smith, M. D. (2019). Causal inference in coupled human and natural systems. *Proceedings of the National Academy of Sciences*, 116(12):5311–5318. https://doi.org/10.1073/pnas.1805563115.

Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I. S., and van der Schaar, M. (2024). Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968. https://doi.org/10.1038/s41591-024-02902-1.

Garg, P. and Fetzer, T. (2026). Causal Claims in Economics. https://doi.org/10.48550/arXiv.2501.06873.

GBIF.org (2026). GBIF Home Page. https://www.gbif.org/.

Geldmann, J., Jones, J. P. G., Wauchope, H., and Ferraro, P. J. (2025). Causal claims, causal assumptions and protected area impact. *Nature*, 638(8052):E40–E41. https://doi.org/10.1038/s41586-024-08512-8.

Giglio, L., Boschetti, L., Roy, D. P., Humber, M. L., and Justice, C. O. (2018). The Collection 6 MODIS burned area mapping algorithm and product. *Remote Sensing of Environment*, 217:72–85. https://doi.org/10.1016/j.rse.2018.08.005.

Gonçalves-Souza, D., Vilela, B., Phalan, B., and Dobrovolski, R. (2021). The role of protected areas in maintaining natural vegetation in Brazil. *Science Advances*, 7(38):eabh2932. https://doi.org/10.1126/sciadv.abh2932.

Greifer, N. and Stuart, E. A. (2021). Choosing the causal estimand for propensity score analysis of observational studies. *arXiv preprint arXiv:2106.10577*. https://doi.org/10.48550/arXiv.2106.10577.

Guizar-Coutiño, A., Nicholson, G., Coomes, D., Ferraro, P. J., Swinfield, T., and Jones, J. P. (2026). Unobserved confounders cannot explain over-crediting in avoided deforestation carbon projects. *Nature Ecology & Evolution*, pages 1–11. https://doi.org/10.1038/s41559-026-03049-7.

Haddaway, N. R., Grainger, M. J., and Gray, C. T. (2022). Citationchaser: A tool for transparent and efficient forward and backward citation chasing in systematic searching. *Research Synthesis Methods*, 13(4):533–545. https://doi.org/10.1002/jrsm.1563.

Haddaway, N. R., Macura, B., Whaley, P., and Pullin, A. (2017a). ROSES for systematic map protocols. Version 1.0. https://www.roses-reporting.com/systematic-review-protocols.

Haddaway, N. R., Macura, B., Whaley, P., and Pullin, A. S. (2017b). ROSES Flow Diagram for Systematic Reviews. Version 1.0. https://doi.org/10.6084/m9.figshare.5897389.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., and Loveland, T. R. (2013). High-resolution global maps of 21st-century forest cover change. *science*, 342(6160):850–853. https://doi.org/10.1126/science.1244693.

Hernán, M. A. and Robins, J. M. (2010). *Causal inference: What If*. CRC Boca Raton, FL. https://miguelhernan.org/whatifbook.

Imbens, G. W. (2024). Causal inference in the social sciences. *Annual Review of Statistics and Its Application*, 11. https://doi.org/10.1146/annurev-statistics-033121-114601.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635. https://doi.org/10.1016/j.jeconom.2007.05.001.

IPBES (2019). *Global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services*. IPBES Secretariat, Bonn, Germany, brondizio, e. s., settele, j., diaz, s., ngo, h. t. edition. https://doi.org/10.5281/zenodo.6417333.

IPBES (2025). *IPBES transformative change assessment : Full report*. IPBES Secretariat, Bonn, Germany, o'brien, k., garibaldi, l., agrawal, a. edition. https://doi.org/10.5281/zenodo.17099472.

Jones, J. P. G., Barnes, M., Eklund, J., Ferraro, P. J., Geldmann, J., Oldekop, J. A., and Schleicher, J. (2022). Quantifying uncertainty about how interventions are assigned would improve impact evaluation in conservation: reply to Rasolofoson 2022. *Conservation Biology*, page e14007. https://doi.org/10.1111/cobi.14007.

Jones, J. P. G. and Shreedhar, G. (2024). The causal revolution in biodiversity conservation. *Nature Human Behaviour*, 8(7):1236–1239. https://doi.org/10.1038/s41562-024-01897-6.

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., and Silva, R. (2025). Causal Machine Learning: A Survey and Open Problems. *Found. Trends Optim.*, 9(1-2):1–247. https://doi.org/10.1561/2400000052.

Keane, A., Lund, J. F., Bluwstein, J., Burgess, N. D., Nielsen, M. R., and Homewood, K. (2020). Impact of Tanzania's Wildlife Management Areas on household wealth. *Nature Sustainability*, 3(3):226–233. https://doi.org/10.1038/s41893-019-0458-0.

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3):196–217. https://doi.org/10.1207/s15327957pspr0203_4.

Kimmel, K., Avolio, M. L., and Ferraro, P. J. (2023). Empirical evidence of widespread exaggeration bias and selective reporting in ecology. *Nature Ecology & Evolution*, 7(9):1525–1536. https://doi.org/10.1038/s41559-023-02144-3.

Kimmel, K., Dee, L. E., Avolio, M. L., and Ferraro, P. J. (2021). Causal assumptions and causal inference in ecological experiments. *Trends in Ecology & Evolution*, 36(12):1141–1152. https://doi.org/10.1016/j.tree.2021.08.008.

Kuiper, T., Haussmann, S., Whitfield, S., Polakow, D., Dreyer, C., Ferreira, S., Hofmeyr, M., Shaw, J., Bird, J., and Bourn, M. (2025). Dehorning reduces rhino poaching. *Science*, 388(6751):1075–1081. https://doi.org/10.1126/science.ado7490.

Lane, M. A. and Edwards, J. L. (2016). The global biodiversity information facility (GBIF). In *Biodiversity Databases*, pages 1–4. CRC Press. https://www.taylorfrancis.com/chapters/oa-edit/10.1201/9781439832547-1/global-biodiversity-information-facility-gbif-meredith-lane-james-edwards.

Langhammer, P. F., Bull, J. W., Bicknell, J. E., Oakley, J. L., Brown, M. H., Bruford, M. W., Butchart, S. H. M., Carr, J. A., Church, D., Cooney, R., Cutajar, S., Foden, W., Foster, M. N., Gascon, C., Geldmann, J., Genovesi, P., Hoffmann, M., Howard-McCombe, J., Lewis, T., Macfarlane, N. B. W., Melvin, Z. E., Merizalde, R. S., Morehouse, M. G., Pagad, S., Polidoro, B., Sechrest, W., Segelbacher, G., Smith, K. G., Steadman, J., Strongin, K., Williams, J., Woodley, S., and Brooks, T. M. (2024). The positive impact of conservation action. *Science*, 384(6694):453–458. https://doi.org/10.1126/science.adj6598.

Lash, T. L., Fox, M. P., MacLehose, R. F., Maldonado, G., McCandless, L. C., and Greenland, S. (2014). Good practices for quantitative bias analysis. *International Journal of Epidemiology*, 43(6):1969–1985. https://doi.org/10.1093/ije/dyu149.

Li, G., Fang, C., Watson, J. E. M., Sun, S., Qi, W., Wang, Z., and Liu, J. (2024). Mixed effectiveness of global protected areas in resisting habitat loss. *Nature Communications*, 15(1):8389. https://doi.org/10.1038/s41467-024-52693-9.

Millennium Ecosystem Assessment (MEA) (2005). *Ecosystems and human well-being: Policy responses: Findings of the responses working group*. Island Press, Washington (D. C.).

Miteva, D. A., Pattanayak, S. K., and Ferraro, P. J. (2012). Evaluation of biodiversity policy instruments: what works and what doesn't? *Oxford Review of Economic Policy*, 28(1):69–92. https://doi.org/10.1093/oxrep/grs009.

Morgans, C. L., Jago, S., Andayani, N., Linkie, M., Lo, M. G. Y., Mumbunan, S., St. John, F. A. V., Supriatna, J., Voigt, M., Winarni, N. L., Santika, T., and Struebig, M. J. (2024). Improving well-being and reducing deforestation in Indonesia's protected areas. *Conservation Letters*, 17(3):e13010. https://doi.org/10.1111/conl.13010.

Munn, Z., Stern, C., Aromataris, E., Lockwood, C., and Jordan, Z. (2018). What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Medical Research Methodology*, 18(1):5. https://doi.org/10.1186/s12874-017-0468-4.

Ockendon, N., Amano, T., Cadotte, M., Downey, H., Hancock, M. H., Thornton, A., Tinsley-Marshall, P., and Sutherland, W. J. (2021). Effectively integrating experiments into conservation practice. *Ecological Solutions and Evidence*, 2(2):e12069. https://doi.org/10.1002/2688-8319.12069.

O'Garra, T., Neugarten, R., Pynegar, E., and Eklund, J. (2025). Impact evaluation for conservation: Bridging research and practice. *Conservation Science and Practice*, 7(11):e70179. https://doi.org/10.1111/csp2.70179.

Parker, T., Fraser, H., and Nakagawa, S. (2019). Making conservation science more reliable with preregistration and registered reports. *Conservation Biology*, 33(4):747–750. https://doi.org/10.1111/cobi.13342.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688. https://doi.org/10.1093/biomet/82.4.669.

Pearl, J. (2009). *Causality*. Cambridge university press.

Petito, L. C., García-Albéniz, X., Logan, R. W., Howlader, N., Mariotto, A. B., Dahabreh, I. J., and Hernán, M. A. (2020). Estimates of overall survival in patients with cancer receiving different treatment regimens: emulating hypothetical target trials in the Surveillance, Epidemiology, and End Results (SEER)–Medicare linked database. *JAMA network open*, 3(3):e200452. https://doi.org/10.1001/jamanetworkopen.2020.0452.

Pick, J. L., Takola, E., Bairos-Novak, K. R., Ivimey-Cook, E. R., Morillo, D., Nakagawa, S., and Sharapi, J. (2026). Pre-registration and Registered Reports in Ecology and Evolutionary Biology: An Evidence-Based Appraisal by SORTEE. https://ecoevorxiv.org/repository/view/12377/.

Pullin, A. S., Frampton, G., Livoreil, B., and Petrokofsky, G. (2022). Guidelines and Standards for Evidence Synthesis in Environmental Management: Version 5.1. https://environmentalevidence.org/information-for-authors/.

Pynegar, E., Booth, H., Doulton, H., Ferraro, P. J., Mohamed, M., Rakotonarivo, O. S., and Jones, J. P. (2025). RCTs in the wild: Designing and implementing conservation programs as randomized control trials. *Conservation Science and Practice*, 7(11):e70029. https://doi.org/10.1111/csp2.70029.

Quan, Y., Tytko, T., and Hui, B. (2024). Utilizing ASReview in screening primary studies for meta-research in SLA: A step-by-step tutorial. *Research Methods in Applied Linguistics*, 3(1):100101. https://doi.org/10.1016/j.rmal.2024.100101.

Redford, K. H. and Taber, A. (2000). Writing the Wrongs: Developing a Safe-Fail Culture in Conservation. *Conservation Biology*, 14(6):1567–1568. https://doi.org/10.1111/j.1523-1739.2000.01461.x.

Rosenbaum, P. (2017). *Observation and experiment: An introduction to causal inference*. Harvard University Press.

Rosenbaum, P. R., Rosenbaum, P., and Briskman (2010). *Design of observational studies*, volume 10. Springer.

Rosenzweig, M. R. and Wolpin, K. I. (2000). Natural "Natural Experiments" in Economics. *Journal of Economic Literature*, 38(4):827–874. https://doi.org/10.1257/jel.38.4.827.

Rosman, T., Bosnjak, M., Silber, H., Koßmann, J., and Heycke, T. (2022). Open science and public trust in science: Results from two studies. *Public Understanding of Science*, 31(8):1046–1062. https://doi.org/10.1177/09636625221100686.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688. https://psycnet.apa.org/doi/10.1037/h0037350.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331. https://doi.org/10.1198/016214504000001880.

Santangeli, A., Weigel, B., Antão, L. H., Kaarlejärvi, E., Hällfors, M., Lehikoinen, A., Lindén, A., Salemaa, M., Tonteri, T., Merilä, P., Vuorio, K., Ovaskainen, O., Vanhatalo, J., Roslin, T., and Saastamoinen, M. (2023). Mixed effects of a national protected area network on terrestrial and freshwater biodiversity. *Nature Communications*, 14(1):5426. https://doi.org/10.1038/s41467-023-41073-4.

Sheehan, D., Mullan, K., West, T. A. P., and Semmens, E. O. (2024). Protecting Life and Lung: Protected Areas Affect Fine Particulate Matter and Respiratory Hospitalizations in the Brazilian Amazon Biome. *Environmental and Resource Economics*, 87(1):45–87. https://doi.org/10.1007/s10640-023-00813-2.

Siegel, K. and Dee, L. E. (2025). Foundations and Future Directions for Causal Inference in Ecological Research. *Ecology Letters*. https://doi.org/10.1111/ele.70053.

Sills, E. O., Herrera, D., Kirkpatrick, A. J., Jr, A. B., Dickson, R., Hall, S., Pattanayak, S., Shoch, D., Vedoveto, M., Young, L., and Pfaff, A. (2015). Estimating the Impacts of Local Policy Innovation: The Synthetic Control Method Applied to Tropical Deforestation. *PLOS ONE*, 10(7):e0132590. https://doi.org/10.1371/journal.pone.0132590.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534–547. https://doi.org/10.1037/a0033242.

Smith, R. K., Ockendon, N., Al-Fulaij, N., Badalotti, A., Beattie, R., Byrne, J., Cooke, S., de Zoeten, T., Frick, W. F., Funk, S. M., Gibbs, D., Gupta, L., Hendrix, T., Holmes, B., Johnstone, N., Kinghorn, J., Mickleburgh, S., Miller, F., Muir, M., Owens, J. R., Parks, D., Reilly-Pinion, V., Reid, H., Seely, K., Semelin, J., Shaw, T., Swetnam, R. D., Wheeler, L., and Sutherland, W. J. (2026). Role of funders in embedding tests in conservation practice. *Conservation Biology*, n/a(n/a):e70309. https://doi.org/10.1111/cobi.70309.

Song, H., Markowitz, D. M., and Taylor, S. H. (2022). Trusting on the shoulders of open giants? Open science increases trust in science for the public and academics. *Journal of Communication*, 72(4):497–510. https://doi.org/10.1093/joc/jqac017.

Spake, R., Jackson, E. E., Bullock, J. M., Gardner, E., Tipton, E., Grainger, M. J., and Doncaster, C. P. (2025). Precision ecology for targeted conservation action. *Nature Ecology & Evolution*, 9(7):1102–1111. https://doi.org/10.1038/s41559-025-02733-4.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1–21. https://doi.org/10.1214/09-STS313.

Swinfield, T., Williams, A., Coomes, D., Dales, M., Ferris, P., Guizar-Coutiño, A., Hartup, J., Holland, J., Jaffer, S., Jones, J. P. G., Lam, M. O. K., Keshav, S., Madhavapeddy, A., Toye-Scott, E., West, T. A. P., and Balmford, A. (2026). Learning lessons from over-crediting to ensure additionality in forest carbon credits. *Nature Communications*, 17(1):3944. https://doi.org/10.1038/s41467-026-71552-3.

UNEP-WCMC and IUCN (2026). Protected Planet: The World Database on Protected Areas (WDPA) and World Database on Other Effective Area-based Conservation Measures (WD-OECM). https://www.protectedplanet.net/en/about.

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., and Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133. https://doi.org/10.1038/s42256-020-00287-7.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of internal medicine*, 167(4):268–274. https://doi.org/10.7326/m16-2607.

Wauchope, H. S., Jones, J. P. G., Geldmann, J., Simmons, B. I., Amano, T., Blanco, D. E., Fuller, R. A., Johnston, A., Langendoen, T., Mundkur, T., Nagy, S., and Sutherland, W. J. (2022). Protected areas have a mixed impact on waterbirds, but management helps. *Nature*, 605(7908):103–107. https://doi.org/10.1038/s41586-022-04617-0.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. *New approaches to evaluating community initiatives: Concepts, methods, and contexts*, 1:65–92.

Xiao, A., Xuan, W., Wang, J., Huang, J., Tao, D., Lu, S., and Yokoya, N. (2025). Foundation models for remote sensing and earth observation: A survey. *IEEE Geoscience and Remote Sensing Magazine*. https://doi.org/10.48550/arXiv.2410.16602.

zu Ermgassen, S. O. S. E., Devenish, K., Simmons, B. A., Gordon, A., Jones, J. P. G., Maron, M., Schulte to Bühne, H., Sharma, R., Sonter, L. J., Strange, N., Ward, M., and Bull, J. W. (2023). Evaluating the impact of biodiversity offsetting on native vegetation. *Global Change Biology*, 29(15):4397–4411. https://doi.org/10.1111/gcb.16801.

Supplementary Information for

Still Money for Nothing?
Two Decades of Empirical Evaluation of Conservation Investments

SUPPLEMENTARY INFORMATION 1 – THE CORE CAUSAL ASSUMPTIONS

1. INTRODUCTION

> *"There is nothing is wrong with making assumptions; causal inference is impossible without making assumptions, and they are the strands that link statistics to science. It is the scientific quality of those assumptions, not their existence, that is critical." (Rubin 2005, p. 234)*

> *"Every claim invoking causal concepts must be traced to some premises that invoke such concepts; it cannot be derived or inferred from statistical associations alone" (Pearl 2003, p. 105)*

Beneath every counterfactual analysis lies a set of implicit core causal assumptions. Scientists from various fields have discussed these assumptions for decades, along with approaches to address potential violations. Understanding these assumptions is essential, as the validity of any identified causal effects depends on the assumption's plausibility (Rubin 2005). However, a persistent issue hindering both the understanding and adherence to these assumptions and, more broadly, the adoption of causal identification strategies, is the lack of a uniform terminology when referring to them. This problem has been noted not only in Conservation and Ecology (Siegel and Dee 2025; Correia et al. 2026) but also in Computer science (Neal 2020) and Statistical science (Cinelli et al. 2025).

2. CLARIFYING THE ASSUMPTIONS

We aimed to clarify the terminology used across various scientific fields to identify and define the core causal assumptions. Our goal was to improve the legibility of peer-reviewed literature from different disciplines. To accomplish this, we reviewed highly cited literature on causal inference, and the reference lists of several key conservation-related articles. Additionally, we examined the top search results on Web of Science and Google Scholar for the term *causal inference assumptions*. In total, we looked at 53 documents across 13 different scientific fields. This effort was not meant to be a systematic review, but a broad overview of the common terminologies synonymously referring to these core assumptions. From our analysis, we identified four core assumptions:

1. **Exchangeability Assumption**
   Exchangeability (also commonly referred to as *Ignorability*) means that the treated and control groups are interchangeable, such that the same outcomes would have been observed had the control group received the treatment or vice-versa. One might find reference to *conditional exchangeability*, which is specific to conditioning strategies and holds that treated and control groups are interchangeable only after conditioning on a set of observed covariates.

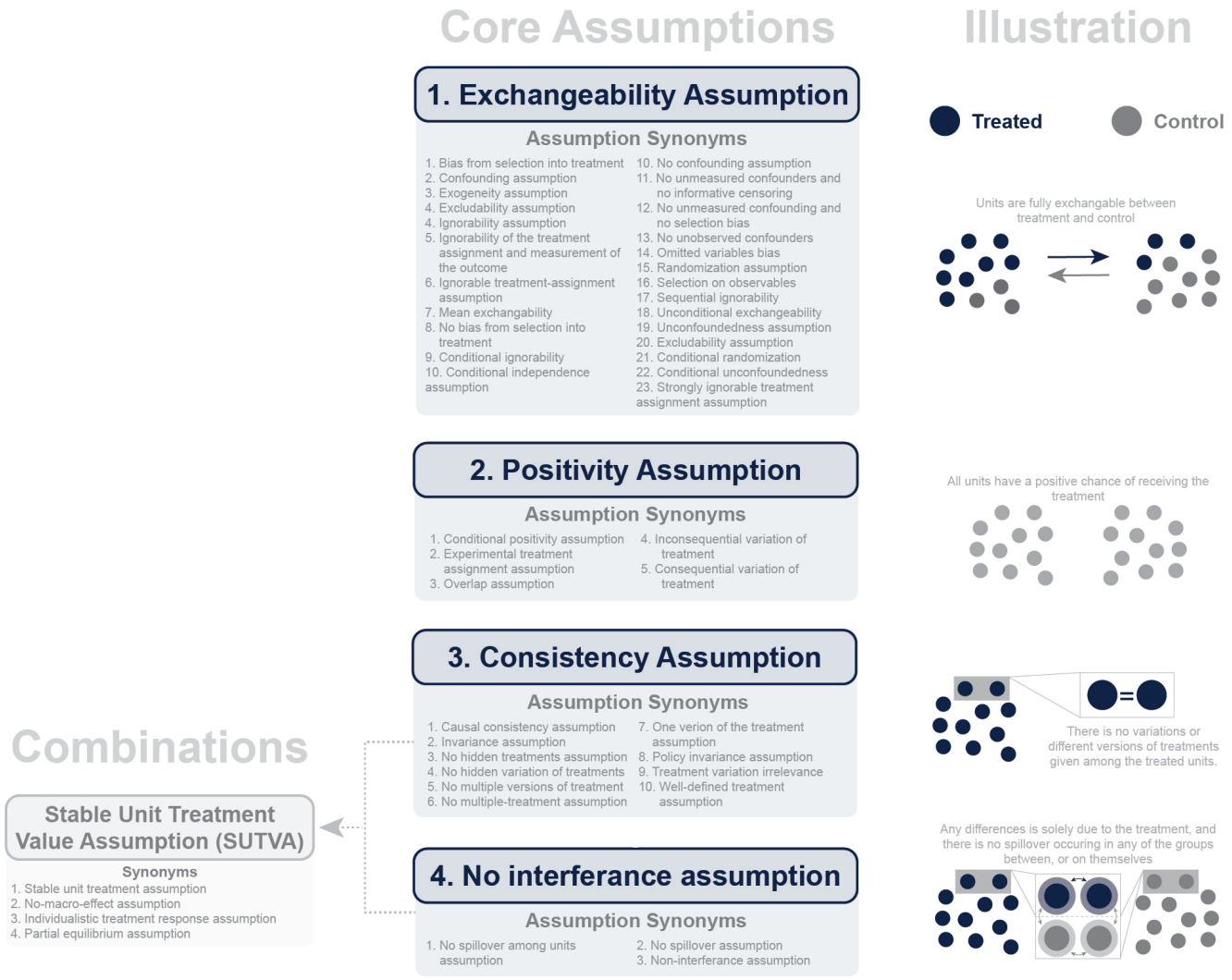
2. **Positivity Assumption**
   Positivity means that treated and control groups must be present throughout all subgroups that are defined by covariates (i.e., every unit has a non-zero probability of receiving the treatment). A deterministic treatment-assignment mechanism violates this assumption.

3. **Consistency Assumption**
   Consistency means that the treatment is sufficiently well defined and uniformly applied across treated units. This assumption holds if any differences in how the treatment was delivered do not meaningfully change the outcome - referred to as *inconsequential variation of treatment*. When those differences do matter, known as *consequential variation of treatment*, there are effectively multiple versions of the treatment producing different effects, and this assumption becomes harder to satisfy.

4. **No-Interference Assumption**
   No interference means that the outcome of one unit is unaffected by the treatment status of other units. Violations of this assumption are commonly referred to as spillovers.

Supplementary Figure 1 below illustrates all synonyms encountered in the literature for each core assumption, including any merged phrases. To minimise author subjectivity, we only identified synonymous terms based on direct

quotes from the literature. A comprehensive list of all identified synonyms, alongside the corresponding quotes and full citations is available at the end of the document.



Supplementary Figure 1. The core causal assumptions and the identified synonyms from 53 different peer-reviewed papers and books across 13 different fields of science, alongside a graphical illustration of the assumption's meaning. The merged phrases are also shown on the left.

3. VARIATION IN TERMINOLOGY

The terms *exchangeability* and *ignorability* are often treated as synonymous in causal inference, reflecting independent origins across disciplines. Rubin (1980) introduced ignorability within the potential outcomes framework in statistical science, describing instances where the treatment-assignment mechanism is independent of potential outcomes conditional on observed variables, essentially allowing researchers to *ignore* the treatment-assignment mechanism. In contrast, epidemiologists such as Greenland and Robins (1986) adopted *exchangeability* to indicate scenarios where treated and control groups, adjusted for covariates, exhibit comparable potential outcomes. This difference in terminology illustrates that ignorability emphasises statistical independence in the treatment-assignment mechanism, while exchangeability focuses on practical comparability in applied research contexts. Despite this divide, both fields centre on the core idea that there are no unobserved confounders (Greenland and Robins 2009). In **Supplementary Figure 1**, throughout the **Main document** and in **Supplementary information 1 - 5**, we adopt the phrase *Exchangeability assumption* for conservation impact evaluations, which explicitly centres the need for balanced comparison groups.

This is a critical consideration in environmental studies where unobserved confounding (e.g., site selection biases) often threatens validity.

4. COMBINED ASSUMPTIONS AND SYNONYMOUS TERMINOLOGY

In our search, we documented instances where core assumptions have been merged and referred to by a single term. The most common merged assumption was the *Stable Unit Treatment Value Assumption* (SUTVA), which combines the *consistency* and *no-interference* assumptions (Rubin 1980). We also identified the phrases *weak ignorability*, referring to a combination of *conditional exchangeability* and *positivity*, and *strong ignorability*, referring to *full exchangeability* and *positivity* (Rosenbaum and Rubin 1983). While this specificity might not be important in relation to Conservation Science's applied contexts, due to the likelihood of a conservation programme or intervention satisfying full exchangeability being low, it is essential to mention these terms so that readers are aware of their existence. More broadly, we identified 45 synonymous terms that conceptually refer to the same four core assumptions (**Supplementary Figure 1** and **Supplementary Tables 1 - 5**). We recognize there are subtle nuances and distinctions between these synonymous terms, however, this document is aimed at researchers and practitioners new to the field, for whom this additional terminology serves as an unnecessary barrier to adopting counterfactual thinking. These synonyms likely reflect different scientific fields independently developing their own approaches to causal inference, as discussed in **Section 3**, however, our sample did not fully support this explanation, as multiple synonyms for the same core assumptions were found cited within the same scientific field (**Supplementary Figures 2 - 5**).

5. THE TREATMENT-ASSIGNMENT MECHANISM

All the core assumptions illustrated in **Supplementary Figure 1** revolve around understanding *why* some units receive treatment while others do not. This process is known as the *treatment-assignment mechanism*, and it is a key component of the potential outcomes framework (Rubin 1974; Imbens and Rubin 2010). In specific causal identification strategies such as experimental control over treatment exposure (**Figure 1:E** in the **Main document**), the treatment assignment mechanism is clearly understood as it is within control of the researcher.

In observational studies, the treatment-assignment mechanism is considerably more complex than in experimental settings, which is precisely why explicitly specifying causal reasoning is so important, for example, through graphical depictions of theories of change or directed acyclic graphs (DAGs), and the causal estimands (all of which has been discussed in the **Main document**). This is particularly crucial for observational causal identification strategies (**Figure 1: B - D**, **Main document**). Following the protected areas examples, a researcher using conditioning strategies to satisfy the exchangeability assumption must carefully consider *how* and *why* a protected area was designated where they are, as this directly shapes which covariates need to be accounted for (Geldmann et al. 2025). Nevertheless, scholars across disciplines have noted that clearly articulating the treatment-assignment mechanism and its underlying assumptions does exposes one's work to criticism from reviewers and peers (Pearl 2003; Rubin 2005), creating an incentive to keep such reasoning implicit. However, if Conservation Science is to move towards producing evidence that reliably informs effective policy, the field must move beyond this concern and embrace transparency in causal reasoning as a standard practice.

6. UNOBSERVED CONFOUNDERS AND SENSITIVITY TESTING

The complexity of socio-ecological systems upon which biodiversity conservation operates in, makes it challenging to confidently assert that all potential confounding variables within observational studies have been conditioned for within a causal identification strategy. Some researchers attempt to mitigate this issue by increasing the number of covariates they include in their analysis. While this does lower the likelihood of an unobserved confounder going unaccounted for, it does not eliminate the possibility of its existence. Moreover, it also raises the risk of violating the positivity assumption, as this increased conditioning will inadvertently lower the quantity of potential units. This phenomenon is known as the *Positivity-Unconfoundness Trade-off* (with *Unconfoundness* referring to the *Exchangeability* assumption in **Supplementary Figure 1**) (Hernán and Robins 2010; Neal 2020; D'Amour et al. 2021; Vonk et al. 2023) or the 'Kitchen sink' approach (Bertomeu et al. 2016).

To address the vulnerability of an estimate being susceptible to unobserved confounders, statistical methods such as sensitivity testing, through *Rosenbaum's Bounds* (Rosenbaum 1987; Rosenbaum et al. 2010), *Oster's Bounds* (Oster

2019) and *Sensemakr* (Cinelli and Hazlett 2020) have been developed. These approaches help assess how sensitive the results are to the influence of unobserved confounding variables. The use of sensitivity testing in conservation impact evaluations remains sparse (Recall that only 30.2% of studies interrogated their exchangeability assumption within **Figure 5** in the **Main document**) and often unclear (Jones et al. 2022; Geldmann et al. 2025; Guizar-Coutiño et al. 2026). This issue is not unique to conservation; for instance, Cinelli and Hazlett (2020) found that only 6.25% of quantitative papers in top political science journals performed sensitivity testing to confirm the robustness of their results.

It is worth noting that numerous approaches exist for interrogating the causal assumptions, and these differ depending on the causal identification strategy used. For instance, Difference-in-Differences (synonymous with Before-After Control-Impact (BACI) in ecology) lends itself to *parallel trends* testing, which assesses whether treated and control groups followed the same trends prior to treatment (Gibson and Zimmerman 2021). Beyond this, other approaches such as *placebo tests* (also known as *falsification tests* (Pizer 2016; Kimmel et al. 2021), *tests for known effects* (Rosenbaum 1989), and *tests of unconfoundedness* (Imbens and Rubin 2015), exist (Eggers et al. 2024).

7. CONCLUDING REMARKS

Significant efforts have been made by scholars in impact evaluation to introduce these topics to a broad audience (Hernán and Robins 2010; Morgan and Winship 2014; Pearl et al. 2016; Pearl and Mackenzie 2018; Cunningham 2021), including ecology and conservation (Ferraro and Hanauer 2014; Kimmel et al. 2021; Correia et al. 2026; Siegel and Dee 2025; Baylis et al. 2026). Many individuals, including those without formal training in causal inference, may seek to apply these tools to better identify effective conservation actions or evaluate their impacts. However, they might be deterred due to the field's complex terminology and incomprehensibility. In this document, we outline the core causal assumptions that are relevant for all impact evaluations, along with their synonyms and merged phrases. Our goal was to facilitate understanding for readers engaging with more technical papers from other fields. The main takeaway is that while not all these assumptions are inherently testable, they are crucial, and so too is the treatment-assignment mechanism.

Supplementary Table 1. The collected definitions for the Exchangeability assumption

Synonym	Field	Direct Quote from the manuscript cited	Citation
Exchangeability	Epidemiology	"...that those individuals receiving the treatment should be considered as **exchangeable** (with respect to potential outcomes) with those not receiving the treatment and vice versa. That is, they should be identical on average for characteristics that may influence the outcome except for the treatment itself."	(Rehkopf et al. 2016)
	Epidemiology	"...**exchangeability** (ie, "**no unmeasured confounders and no informative censoring**," or "**ignorability of the treatment assignment and measurement of the outcome**"). The exchangeability assumptions are well known territory for epidemiologists and biostatisticians. Briefly, to be satisfied, these exchangeability assumptions that require exposed and unexposed subjects, and censored and uncensored subjects have equal distributions of potential outcomes, respectively. Indeed, the so-called fundamental problem of causal inference is directly linked to the first **exchangeability assumption**."	(Cole and Frangakis 2009)

Supplementary Table 1 *continued on next page*

Supplementary Table 1 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Epidemiology	"...A crucial assumption of causal inference is that (conditional on a set of variables) the treatment and control groups are **(conditionally) exchangeable** (Hernán & Robins, 2020). Confounding and selection bias (different from selection bias in econometrics) lead to a lack of **exchangeability** between the treatment and control groups (Hernán & Robins, 2020). **Exchangeability** implies that bias due to unobservable factors is almost eliminated."	(Chao and Yu 2023)
	Epidemiology	"...**Exchangeability** means that the risk of death in the white group would have been the same as the risk of death in the grey group had individuals in the white group received the treatment given to those in the grey group equivalently, **exchangeability** means that the counterfactual outcome and the actual treatment are independent. **Exchangeability** is expected to produce exchangeability. When the treated and the untreated are exchangeable, we sometimes say that treatment is exogenous, and thus exogeneity is commonly used as a synonym for exchangeability."	(Hernán and Robins 2010)
	Epidemiology	"...**Exchangeability assumption**, means that the exposed group (A = 1) and the unexposed group (A = 0) have the same distribution of potential outcome that would be observed if everyone was exposed (Y a=1) and, similarly, of Y a=0 . **Exchangeability** implies that the exposed versus the unexposed share equal distributions of outcome predictors, but such a condition is generally violated in observational studies.8–10"	(Shiba and Kawahara 2021)
	Computer Science	"...**Exchangeability**. Under randomization, the parameters A0 and B0 (and A1 and B1 as well) are outcomes of a random process and so can be treated as random variables. Successful randomization renders A0 and B0 **unconditionally exchangeable** in the usual probabilistic sense (Cornfield, 1976)"	(Greenland et al. 1999)
	Medicine	"...**Exchangeability**: individuals are exchangeable when counterfactual outcomes are independent of the actual exposures. To describe conditional independence, we use the symbol: . When Y a A, exchangeability is satisfied. In a randomized experiment, random allocation of participants to exposures will typically ensure exchangeability. In observational studies, however, exchangeability typically requires conditioning on a set of measured differences (L = l) such that Y a A — L = l. We say the counterfactual outcome of Y under exposure A = a is **conditionally independent of exposure assignment** given a set of measured confounders L = l."	(Bulbulia 2023)
	Epidemiology	"...The **exchangeability** (or **'no confounding'**) **assumption** requires that individuals who were exposed and unexposed have the same potential outcomes on average."	(Igelström et al. 2022)
	Epidemiology	"...Positivity, or the experimental treatment assignment assumption, is a necessary assumption for causal inference in observational data, along with consistency, **exchangeability** (i.e., **no unmeasured confounding and no selection bias**)"	(Westreich and Cole 2010)

Supplementary Table 1 *continued on next page*

Supplementary Table 1 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Epidemiology	"...Equivalence of response type may be thought of in terms of exchangeability of individuals: if the exposure states of the two individuals had been exchanged, the same data distribution would have resulted. Thus the seen as a partial **exchangeability assumption**: it says that if the exposure states were exchanged, the value observed for the incidence in the absence of exposure would have been the same. (Complete exchangeability—the same incidence-exposure relation if exposure states were exchanged)"	(Greenland and Robins 1986)
No bias from selection intro treatment	Epidemiology	"...**Confounding** or **omitted variable bias** or **bias from selection into treatment**: The key bias introduced by lack of randomization. This bias occurs when the association between treatment and outcome is partially attributable to the influence of a third factor that affects both the treatment and the outcome (e.g., parental education may influence both a child's own education and that child's later health; if not accounted for, parental education confounds the association between the child's education and subsequent health). This bias is often referred to as omitted variables bias because it is a problem when the common cause is omitted from a regression model. Selection bias in this context specifically refers to selection into treatment and is distinct from biases due to selection into the study sample, which is the phenomenon typically referred to as selection bias in epidemiology"	(Matthay et al. 2020)
Ignorable treatment-assignment assumption	Statistical Science	"...The assumption underlying this method is that assignment to treatment is associated only with observable preintervention variables, called the **ignorable treatment assignment assumption** or **selection on observables** (see Heckman and Robb 1985; Holland 1986; Rubin 1974, 1977, 1978). Although this is a strong assumption, we demonstrate that propensity score methods are an informative starting point, because they quickly reveal the extent of overlap in the treatment and comparison groups in terms of preintervention variables"	(Dehejia and Wahba 1999)
Randomization assumption	Epidemiology	"...**Exchangeability**, **ignorability**, **no confounding**, or **randomization assumption**: The assumption that which treatment an individual receives is unrelated to her potential outcomes if given any particular treatment. This assumption is violated for example if people who are likely to have good outcomes regardless of treatment are more likely to actually be treated. In the context of instrumental variables analysis, exchangeability is the assumption that the instrument does not have shared causes with the outcome"	(Matthay et al. 2020)
Mean exchangeability assumption	Statistical Science	"...Assumption 4 (**Mean exchangeability**). E[Y (a) — X = x, S = 1] = E[Y (a) — X = x] (mean exchangeability over trial participation), for all x and a = 0, 1. Another assumption can be found, relying on the transportability of treatment effect rather than the potential outcomes..."	(Colnet et al. 2024)
No unobserved confounders assumption	Computer Science	"...These methods are largely based on the **assumption** that there are **no unobserved confounders**, i. e., variables that affect both the treatment assignment and the outcome"	(Hatt and Feuerriegel 2024)

Supplementary Table 1 *continued on next page*

Supplementary Table 1 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Statistical Science	"...Observational research often seeks to estimate causal effects under a '**no-unobserved-confounding**' or '**ignorability**' (**conditional on observables**) assumption (see for example Rosenbaum and Rubin (1983a), Pearl (2009) and Imbens and Rubin (2015))."	(Cinelli and Hazlett 2020)
Exogeneity assumption	Statistical Science	"...Identification of the causal effect of a treatment T on an outcome Y in observational studies is typically based either on the **unconfoundedness assumption** (also called **selection on observables**, **exogeneity**, **ignorability**, see, e.g. de Luna and Johansson [1], Imbens and Wooldridge [2], Pearl [3]) or on the availability of an instrument."	(Luna and Johansson 2014)
	Economics	"...It excludes differences in Uj, and U0t arising from X dependence and allows us to focus on differences in outcomes solely attributable to D. While convenient, this assumption is overly strong. However, we stress that the **exogeneity assumption** in either cross-section or panel contexts is only a matter of convenience and is not strictly required. What is required for an interpretable definition of the "treatment on the treated" parameter is avoiding conditioning on X variables caused by D even holding YP = ((Y01,Y/0 ..... (YoT,Yvr)) fixed where YP is the vector of potential outcomes."	(Heckman et al. 1999)
Selection on observables assumption	Political Science	"...In this paper, we focus on **sequential ignorability**, which is, essentially, a **selection-on-observables assumption**: the researcher is able to choose a (time-varying) conditioning set to eliminate any unmeasured confounding."	(Blackwell and Glynn 2018)
Unconfoundedness Assumption	Statistical Science	"...The term **unconfoundedness** was coined by Rubin (1990a). It is sometimes referred to as the **conditional independence assumption** (Lechner, 2001; Angrist and Pischke, 2009). In the econometrics literature it is also closely related to the notion of **exogeneity** (Manski, Sandefur, McLanahan, and Powers, 1992), although formal definitions of exogeneity do not coincide with unconfoundedness (see Imbens, 2004, for some discussion)."	(Imbens and Rubin 2015)
	Economics	"...**Unconfoundedness**, also known as **selection on observables**, in which the treatment assignment mechanism depends only on observed covariates;..."	(D'Amour et al. 2021)
	Computer Science	"...The existence of unobserved confounders is equivalent to violation of the **unconfoundedness assumption** and is not testable. Controlling high-dimensional variables may make unconfoundedness more plausible but poses new challenges to propensity score estimation and confounder balancing."	(Kuang et al. 2020)

Supplementary Table 1 *continued on next page*

Supplementary Table 1 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Economics	"...**Unconfoundedness** and Common Support: One major strand of evaluation literature focuses on the estimation of treatment effects under the assumption that the treatment satisfies some form of exogeneity. Different versions of this assumption are referred to as **unconfoundedness** (Rosenbaum and Rubin, 1983b), **selection on observables** (Heckman and Robb, 1985) or **conditional independence assumption** (CIA) (Lechner, 1999). We will use these terms throughout the paper interchangeably."	(Caliendo and Kopeinig 2008)
Ignorability assumption	Computer Science	"...**Assuming ignorability** is like ignoring how people ended up selecting the treatment they selected and just assuming they were randomly assigned their treatment. We have seen that ignorability is extremely important (Equation 2.3), but how realistic of an assumption is it? In general, it is completely unrealistic because there is likely to be confounding in most data we observe (causal structure shown in Figure 2.1). However, we can make this assumption realistic by running randomized experiments, which force the treatment to not be caused by anything but a coin toss, so then we have the causal structure shown in Figure 2.2. We cover randomized experiments in greater depth in Chapter 5."	(Neal 2020)
	Computer Science	"...(**Ignorability/exchangeability**) In words, the potential outcomes under treatment are independent of treatment assignment. In this case, we can ignore how units ended up in the treatment or control group. Equivalently, the group that received treatment could have been exchanged with the group receiving control resulting in the same potential outcome. "	(Vonk et al. 2023)
	Statistical Science	"...Given the background variable, X , treatment assignment W is independent to the potential outcomes. The **ignorability assumption** is also named as **unconfoundedness assumption**. With this unconfoundedness assumption, for the units with the same background variable X , their treatment assignment can be viewed as random. The **ignorability assumption** is also named as the **unconfoundedness assumption**. Existing work overwhelmingly relies on the **unconfoundedness assumption** that all confounders can be measured. However, this assumption might be untenable in practice."	(Yao et al. 2021)
	Epidemiology	"...Assumption 3 (**Ignorability assumption** on trial participation). Y (0), Y (1) S — X. (Hotz et al., 2005; Stuart et al., 2011; Tipton, 2013; Hartman et al., 2015; Buchanan et al., 2018; Degtiar and Rose, 2022; Egami and Hartman, 2021b) A parallel can be made with the **strong ignorability condition** in causal inference with observational data (see Assumption S1 in Appendix), but applied to the sample selection rather than treatment assignment. In other words, these assumptions require to control for all covariates being shifted and predictive of Y . We call shifted covariates, all the baseline covariates along which the two populations – trial and target – do not follow the same distribution. A weaker version of Assumption 3 can be found in Dahabreh et al. (2019b, 2020a):"	(Colnet et al. 2024)

Supplementary Table 1 *continued on next page*

Supplementary Table 1 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Political Science	"...Formally, the **ignorability assumption** states that respondents' treatment status is independent of their potential outcomes Yi(0), Yi(1) Ti. However, in the UESD assignment to the treatment and control groups is not under the researchers' control and not at random. Treatment assignment is the result of a combination of the unexpected event and a set of decisions related to data collection taken by the fieldwork operative."	(Muñoz et al. 2020)
No specific phrasing used	Ecology	"...Assumption A1: **No unmeasured treatment–outcome confounders**, i.e. no unmeasured variables that influence both the treatment and the outcome. Assumption A2: **No unmeasured treatment–mediator confounders**, i.e. no unmeasured variables that influence both the treatment and the mediator. Assumption A3: **No unmeasured mediator–outcome confounders**, i.e. no unmeasured variables that influence both the mediator and the outcome. Assumption A4: **No mediator–outcome confounders** (measured or unmeasured) that are influenced by the treatment."	(Correia et al. 2025)
	Computer Science	"...**Conditional exchangeability** is the main assumption necessary for causal inference. Conditional exchangeability (Assumption 2.2) is a core assumption for causal inference and goes by many names. For example, the following are reasonably commonly used to refer to the same assumption: **un-confoundedness**, **conditional ignorability**, **no unobserved confounding**, **selection on observables**, **no omitted variable bias**, etc. We will use the name "**unconfoundedness**" a fair amount throughout this book. The main reason for moving from **exchangeability** (Assumption 2.1) to **conditional exchangeability** (Assumption 2.2) was that it seemed like a more realistic assumption. However, we often cannot know for certain if conditional exchangeability holds. There may be some unobserved confounders that are not part of , meaning conditional exchangeability is violated. Fortunately, that is not a problem in randomized experiment. Unfortunately, it is something that we must always be conscious of in observational data. Intuitively, the best thing we can do is to observe and fit as many covariates into as possible to try to ensure unconfoundedness."	(Neal 2020)
	Epidemiology	"...One may feel more confident that the groups are exchangeable conditional on a vector of covariates L (ie, within strata of the combinations of covariate values). This assumption is called **conditional exchangeability**, Y a ?? AjL (a = 0, 1), and the core of causal inference from observational data."	(Shiba and Kawahara 2021)
	Epidemiology	"...**Conditional exchangeability**, **conditional ignorability**, or **conditional randomization**: The assumption that **exchangeability**, **ignorability**, or randomization is fulfilled after controlling for a set of measured covariates. When this assumption is met, we say that the set of covariates—known as a sufficient set fulfills the backdoor criterion with respect to the treatment and outcome."	(Matthay et al. 2020)

Supplementary Table 1 *continued on next page*

Supplementary Table 1 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Medicine	"...**Conditional exchangeability**: **Exchangeability** means equal risks in treated and untreated groups if patients in the untreated group were treated, and vice versa. Observational data often violates exchangeability due to confounding and/or selection bias, and, therefore, causal inference requires the assumption that all confounders are measured and adjusted for, to achieve exchangeability conditional on these confounders."	(Smit et al. 2023)
	Epidemiology	"...While observational studies often rely on achieving **conditional exchangeability** (or '**no unmeasured confounding**'), which means that exchangeability holds after conditioning on some set of variables."	(Igelström et al. 2022)
No unmeasured confounding assumption	Epidemiology	"...The assumption of **no unmeasured confounding** is perhaps the strongest and also the least (statistically) controversial. Stated in terms of potential outcomes, it requires Y (0), Y (1) Z—X – that is, given the available information contained in covariates X, the treatment received carries no additional information on the individual's potential outcomes. This assumption is also referred to as conditional exchangeability or strong ignorability (Rosenbaum and Rubin, 1983)"	(Moodie and Stephens 2022)
Conditional unconfoundedness Assumption	Sociology	"...**assumption of (conditional) unconfoundedness** of treatment may be maintained and cases in which relevant confounders of treatment are unobserved."	(Gangl 2010)
Conditional independence assumption	Economics	"...The assumption that the assignment to treatments is ignorable conditional on attributes plays an important role in the applied statistic and econometric evaluation literature. Another term for it is **conditional independence assumption**."	(Lechner 2001)
	Economics	"...A weaker condition, called random assignment conditional on a covariate (Rubin 1977), is to assume that the assignment is independent (denoted by II) of the potential nontraining outcome conditional on the value of a covariate or attribute (**conditional independence assumption**, CIA)."	(Lechner 1999)
Conditional Ignorability	Computer Science	"...That means that treatment and control group are generally not exchangeable, but they become exchangeable when we condition on the confounding set. For that reason, **conditional ignorability** is also known as the unconfoundedness assumption. It is useful to adjust for confounding to reach **conditional ignorability** as long as the probability of receiving treatment and control remains strictly positive in each of the created sub-groups. The positivity assumption guarantees this is the case."	(Vonk et al. 2023)
Conditional Ignorability	Political Science	"...Under the "selection on observables" identification strategy, the analyst asserts that there is some set of covariates such that treatment assignment is random conditional on these covariates (Barnow, Cain, and Goldberger 1980). Under this assumption, there are no unobservable differences between the treated and control groups. This assumption has a number of different names, which include "conditional ignorability" and "no omitted variables"."	(Keele 2015)

Supplementary Table 1 *continued on next page*

Supplementary Table 1 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
Strong Ignorability	Computer Science	"...Both **conditional ignorability** and positivity together are called strong ignorability (Rosenbaum and Rubin 1983; Imbens and Rubin 2015)."	(Vonk et al. 2023)
	Economics	"...**Strong ignorability assumption** is a mix of two assumptions: (1) unconfoundedness of the treatment assignment, which states that conditional on observed covariates, assignment to the treatment is independent of potential outcomes, and (2) overlap or positivity of the treatment assignment, which assumes that the assignment to the treatment is probabilistic, that is the propensity score of the treatment is a probability strictly between zero and one."	(Rafieian 2023)
	Psychology	"...The often referenced **strong ignorability assumption** refers to the combination of an unconfounded and probabilistic assignment and is regularly used in connection with propensity scores techniques to estimate the causal effect (Rosenbaum & Rubin, 1983; Imbens & Rubin, 2015)."	(Rafieian 2023)
Weak Ignorability	Epidemiology	"...Rubin (1974, 1978) extended Neyman's theory for randomized experiments to observational studies. Rosenbaum and Rubin (1983) referred to the combination of exchangeability and positivity as **weak ignorability**, and to the combination of full exchangeability (see Technical Point 2.1) and positivity as **strong ignorability**"	(Hernán and Robins 2010)
Excludability assumption	Conservation Science	"...**Excludability**: the assumption that factors driving variation in the treatment variable have no causal link to the outcome variable except through their effects on variation in the treatment (the "treatment" is the causal variable of interest)."	(Ferraro et al. 2019)
	Ecology	"...**Excludability**: outcomes respond solely to a treatment itself and not to another causal pathway that is set in motion by the assignment of a treatment."	(Kimmel et al. 2021)
	Epidemiology	"...**Excludability**: To estimate the causal effect of the event one must assume that the timing of the survey interview does not affect the outcome through any other channel except for the event of interest."	(Muñoz et al. 2020)
	Epidemiology	"...**Excludability**, also known as the temporal stability assumption (Legewie, 2013), implies that the timing, as an instrument for the focal event's effect on the outcome, affects the outcome through no other channel than the event itself. This is conceptually similar to the exclusion restriction in instrumental variable estimation, since the exclusion restriction states that, conditional on covariates, the instrument (survey timing) cannot be correlated with the error term (e.g., other events) in the explanatory equation (Labrecque & Swanson, 2018; Stock & Watson, 2007)..."	(Nägel and Nivette 2023)

Supplementary Table 1 *continued on next page*

Supplementary Table 1 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Behavioural Science	"...**Excludability** is a core **assumption** of inference in experiments, and states that random assignment to treatment affects the outcome only through the variable of interest (25). Thus, when the excludability assumption is satisfied, the estimated treatment effect is attributable to the variable of interest..."	(Tappin et al. 2020)
	Epidemiology	"...**Excludability**: When we define two, and only two, potential outcomes based on whether the treatment is administered, we implicitly assume that the only relevant causal agent is receipt of the treatment. Because the point of an experiment is to isolate the causal effect of the treatment, our schedule of potential outcomes excludes from consideration factors other than the treatment. When conducting an experiment, therefore, we must define the treatment and distinguish it from other factors with which it may be correlated."	(Gerber and Green 2012)

Supplementary Table 2. The collected definitions for the Positivity assumption

Synonym	Field	Direct Quote from the manuscript cited	Citation
Positivity	Epidemiology	"...The **positivity assumption**—that in all covariate strata some individuals are treated while others are untreated."	(Rehkopf et al. 2016)
	Computer Science	"...**Positivity** is the condition that all subgroups of the data with different covariates have some probability of receiving any value of treatment. Another name for **positivity** is **overlap**. The intuition for this name is that we want the covariate distribution of the treatment group to overlap with the covariate distribution of the control group."	(Neal 2020)
	Statistical Science	"...If for some values of X, the treatment assignment is deterministic; then for these values, the outcomes of at least one treatment could never be observed. In this case, it would be unable and meaningless to estimate the treatment effect. To be more specific, suppose there are two treatments: Medicine A and Medicine B. Let's assume that patients with age greater than 60 are always assigned with medicine A, then it will be unable and meaningless to study the outcome of medicine B on those patients. In other words, the **positivity assumption** indicates the variability, which is important for treatment effect estimation. The ignorability and the **positivity assumptions** together are called Strong Ignorability or Strongly Ignorable Treatment Assignment."	(Yao et al. 2021)
	Epidemiology	"...The **positivity assumption** states that there is a nonzero (ie, positive) probability of receiving every level of exposure for every combination of values of exposure and confounders that occur among individuals in the population."	(Cole and Frangakis 2009)
	Epidemiology	"...Second, the **positivity assumption** means that both exposed and unexposed individuals need to be present in all sub- populations defined by the combinations of covariate values."	(Shiba and Kawahara 2021)

Supplementary Table 2 *continued on next page*

Supplementary Table 2 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Medicine	"...In the next section we will develop causal effects for the different subpopulations. In most settings we want to consider populations of individuals who have the possibility of receiving all treatment levels of interest. This restriction is referred to as the **positivity assumption**. It could be violated, for example, if the target population included women for whom breastfeeding is precluded (because of preexisting or pregnancy-related conditions)."	(Goetghebeur et al. 2020)
	Medicine	"...**Positivity**: To estimate a treatment's causal effect, one must compare treated and untreated patient data. This requires having both treated and untreated patients in all subgroups (or 'strata') defined by different confounder values. In other words, treatment and non-treatment should occur in all confounder strata with some positive probability."	(Smit et al. 2023)
	Medicine	"...**Positivity**: there is a non-zero probability of the exposure among the non-exposed and of non-exposure among the exposed. That is, the exposure cannot be deterministic."	(Bulbulia 2023)
	Epidemiology	"...The **positivity assumption** requires that every value of exposure was possible (ie, had a non-zero probability) for each individual at the time that exposure was assigned."	(Igelström et al. 2022)
	Psychology	"...The second **assumption** is **positivity**. This assumption requires that the probability that a given individual receives each level of treatment is positive for every combination of treatment and co-variates. This assumption eliminates illogical possibilities, such as men developing uterine cancer. As discussed below, this assumption has an empirical counterpart."	(Foster 2010)
Experimental treatment assignment assumption	Epidemiology	"...**Positivity**, or the **experimental treatment assignment assumptio**n, requires that there be both exposed and unexposed participants at every combination of the values of the observed confounders in the population under study. Positivity is essential for inference but is often overlooked in practice by epidemiologists."	(Westreich and Cole 2010)
Conditional positivity assumption	Epidemiology	"...Finally, the **assumption** of (**conditional**) **positivity** or **overlap** states that there exists no strata of X such that treatment assignment is uniquely determined (or indeed, some levels of treatment are impossible"	(Moodie and Stephens 2022)
Overlap assumption	Economics	"...**Overlap**, also known as **positivity** or **common support**, in which all units have a non-zero probability of assignment to each treatment condition."	(D'Amour et al. 2021)
	Statistical Science	"...Another currently active area of research is regression discontinuity designs, both sharp and fuzzy, where the **overlap assumption** regarding covariate distributions is not necessarily satisfied, but the extrapolation is limited."	(Imbens and Rubin 2015)

Supplementary Table 2 *continued on next page*

Supplementary Table 2 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Economics	"...**Assumption** 2. **Overlap**: It ensures that persons with the same X values have a positive probability of being both participants and nonparticipants (Heckman et al., 1999)."	(Caliendo and Kopeinig 2008)

Supplementary Table 3. The collected definitions for the Consistency assumption

Synonym	Field	Direct Quote from the manuscript cited	Citation
Consistency	Epidemiology	"...The **consistency assumption** requires that there are no two "flavors" or versions of treatment such that X=1 under both versions but the potential outcome for Y would be different under the alternative versions. This modest looking assumption—which in recent causal inference literature has been labeled the consistency assumption—is often overlooked. What consistency critically implies is that the exposure specified in the analysis must have enough precision that any variation within the exposure specification would not result in a different outcome."	(Rehkopf et al. 2016)
	Computer Science	"...**Consistency** is **the** assumption that the outcome we observe is actually the potential outcome under the observed treatment..."	(Neal 2020)
	Computer Science	"...Informally, the assumption forces one to unambiguously define treatment and tie the potential outcomes to the observed variables. Earliest claims for the use of this assumptions to 'simplify matters' date back to the seventies (Gibbard and Harper 1978), but have been formalized later by Robins (1986). Despite the fact that consistency can be derived from the definition of potential outcome variables (Malinsky et al. 2019) (which will be discussed in Sect. 6), scholars (VanderWeele 2009) propound the view that **consistency** is an **assumption** rather than a definition or axiom. Although this assumption is sometimes known as the no multiple-treatment assumption, some researchers draw a firm distinction between the two (VanderWeele and Hernan 2013). Consistency can be a strong assumption in the observational setting, but it is implicit in randomized controlled trials, because exposure to treatment is a result of experimental design (Cole and Frangakis 2009)"	(Vonk et al. 2023)
	Epidemiology	"...The **consistency assumption** is often stated such that an individual's potential outcome under her observed exposure history is precisely her observed outcome. Methods for causal inference require that the exposure is defined unambiguously..."	(Cole and Frangakis 2009)
	Epidemiology	"...**Consistency** means that the observed outcome for every treated individual equals her outcome if she had received treatment, and that the observed outcome for every untreated individual equals her outcome if she had remained untreated..."	(Hernán and Robins 2010)

Supplementary Table 3 *continued on next page*

Supplementary Table 3 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Epidemiology	"...The c**onsistency condition** has been widely discussed in the statistical and epidemiologic literature.16–19 In its simplest form, it is stated as Ya= Y for every individual with A= a. That is, among individuals who received treatment level A= a, their potential outcome Ya under treatment level a is equal to their observed outcome Y."	(Hernán 2016)
	Epidemiology	"...**Consistency**, also sometimes referred to as the **well-defined treatment assumption** or **treatment variation irrelevance**, states that any variations in the exposure of interest have no relevance to its impact on the outcome."	(Moodie and Stephens 2022)
	Epidemiology	"...First, the **consistency assumption** posits that exposure is sufficiently well-defined and does not have multiple "versions" that have different impacts on outcomes. Accessible introduction of this assumption is available elsewhere."	(Shiba and Kawahara 2021)
	Statistical Science	"...**Assumption** 1 (**Consistency**). Assumption 1 implies that the observed outcome is the potential outcome under the actual assigned treatment..."	(Colnet et al. 2024)
	Epidemiology	"...**Consistency**: The assumption that an individual's potential outcome setting treatment to a particular value is that person's actual outcome if s/he actually has that particular value of treatment. This could be violated if the outcome might depend on how treatment was delivered or some other variation in the meaning or content of the treatment. Some researchers consider consistency a truism rather than an assumption..."	(Matthay et al. 2020)
	Medicine	"...**Consistency**: Consistency assumes that the outcome for a given treatment will be the same, irrespective of the way treatments are 'assigned'..."	(Smit et al. 2023)
	Medicine	"...**Consistency**: an individual with observed exposure A has observed outcome Y equal to their counterfactual outcome Y A = Y a..."	(Bulbulia 2023)
	Epidemiology	"...The **consistency assumption** (unrelated to Bradford Hill's 'consistency' criterion 8) requires that the exposure is sufficiently well defined, so that each individual has one potential outcome for each level of the exposure.27 28 This assumption (sometimes called 'treatment-variation irrelevance') is violated if there are multiple different versions of the exposure (eg, dosages of a drug or reasons for becoming unemployed) with different causal effects..."	(Igelström et al. 2022)
	Psychology	"...A third **assumption** is **consistency** (Cole & Frangakis, 2008). (Economists label this policy invariance [Heckman & Vytlacil, 2007a].) This assumption implies that the outcome of treatment does not depend on the assignment mechanism. For example, this assumption means that the returns to enrolling a child in day care are the same for all regardless of the mix of incentives that led to that choice..."	(Foster 2010)

Supplementary Table 3 *continued on next page*

Supplementary Table 3 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
Invariance assumption	Statistical Science	"...Our discussion will rest on the following **assumption**. We assume the existence of a model that is i**nvariant under different experimental** or intervention setting..."	(Peters et al. 2016)
Causal consistency assumption	Medicine	"...The **assumption** of **causal consistency** relates the observed outcome to the potential outcomes. Consistency (at an individual level) means that Y(a) = Y when A = a, hence assuming consistency implies that the observed outcome in our data is the same as the potential outcome that would be realized in response to setting the treatment to the level of the exposure that was observed. This directly affects our interpretation of the estimated causal effect for the study population. It will also affect transportability to new settings in ways that may be hard to predict..."	(Goetghebeur et al. 2020)
No multiple versions of treatment	Ecology	"...**No multiple versions of treatment**: Each treatment condition has more than one version and thus each unit may have more than one potential outcome per treatment condition. 'No multiple versions of treatment' (a.k.a. '**no hidden treatments**') is one part of what statisticians call the SUTVA..."	(Kimmel et al. 2021)
No hidden variation of treatments	Psychology	"...The second part of SUTVA requires that there is **no hidden variation of treatments**. That is, all participants receive the same version of the treatment and control condition. For instance, if multiple versions of an online CBT intervention are (unintentionally) administered, then no causal claims with regard to a single specific CBT intervention can be made..."	(Steiner et al. 2023)
	Statistical Science	"...In any case, the depiction in Figure 1 requires assumptions for it to be adequate—in particular, SUTVA (stable unit treatment value assumption) (Rubin 1980), which comprises two subassumptions. First, it assumes that there is **no interference between units** (Cox 1958); that is, neither Y i(1) nor Yi(0) is affected by what action any other unit received. Second, it assumes that there are **no hidden versions of treatments**; no matter how unit i received treatment 1, the outcome that would be observed would be Yi(1) and similarly for treatment..."	(Rubin 1980)
	Ecology	"...Assumption A7: **No hidden variation (multiple versions) of treatment** or mediators..."	(Correia et al. 2025)
One version of the treatment assumption	Education	"...Under SUTVA, there is only **one version of the treatment** and control and the potential outcomes..."	(Chan 2024)

Supplementary Table 4. The collected definitions for the No-interference assumption

Synonym	Field	Direct Quote from the manuscript cited	Citation
No-interference	Ecology	"...When the potential outcomes of one experimental unit depends not only on its own treatment status but also on the treatment status of other units; in other words, when the treatment status of one unit affects the outcomes of other units. '**No interference**' is one part of what statisticians call the 'stable unit treatment value assumption' (SUTVA)."	(Kimmel et al. 2021)
	Ecology	"...Assumption A8: **No interference among units** (i.e. the treatment condition of one unit does not influence the mediator or outcome of other units)..."	(Correia et al. 2025)
	Conservation Science	"...**No interference**: The assumption that the outcome for a unit when exposed to a particular value of a treatment does not depend on the value of the treatment in any other units..."	(Ferraro et al. 2019)
	Computer Science	"...**No interference** means that my outcome is unaffected by anyone else's treatment. Rather, my outcome is only a function of my own treatment...."	(Neal 2020)
	Computer Science	"...The second assumption is known as the **no-interference assumption** (Cox 1958). It explicitly states that a potential outcome of a unit is not dependent on treatment received by other units. Interference is also known as spillover. In a randomized controlled trial the investigator can prevent causal spillover by designing the experiment such that different units do not interact..."	(Vonk et al. 2023)
	Medicine	"...**No interference** means that the impact of treatment on the outcome of individual i is not altered by other individuals being exposed or not. At first sight this is likely justified in our setting: one baby's weight typically does not change because another baby is being breastfed. In resource poor or closely confined settings this could, however, be challenged. For instance, interference would happen when a child is affected by the consequences of a reduced immune system of other children who were not breastfed and hence becomes more susceptible to infectious diseases which may impact their weight at 3 months. When the **assumption of no interference** is not met, the potential outcome definition becomes much more complex and involves the treatment assigned to other individuals..."	(Goetghebeur et al. 2020)
	Psychology	"...The **no interference assumption** requires that each participant's pair of potential outcomes is unaffected by the other participants' treatment exposure. This assumption is commonly violated if peer effects are present..."	(Steiner et al. 2023)

Supplementary Table 4 *continued on next page*

Supplementary Table 4 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Statistical Science	"...A fundamental assumption usually made in the potential outcomes approach to causal inference is that of no interference between individuals (Cox 1958), a critical component of the stable unit treatment value assumption (SUTVA) (Rubin 1980). Under the **no-interference assumption**, the potential outcomes of any individual are assumed to be unaffected by the treatment assignment of every other individual. However, in many settings, this assumption obviously does not hold. A classical example is given by the dependent happenings of infectious diseases (Ross 1916, p. 211), where whether one person becomes infected depends on who else in the population is vaccinated. In econometrics, a household's decision whether to move may be affected by whether their neighbors receive a housing voucher to move (Sobel 2006). In education, interventions given to certain students may affect other students in the same class (Rubin 1990; Rosenbaum 2007). Sobel (2006) and Rosenbaum (2007) gave several other examples where interference is likely..."	(Hudgens and Halloran 2008)
Non-interference assumption	Epidemiology	"...The **non-interference assumption** requires that an individual's potential outcomes (and hence the causal effect of the exposure for that individual) does not depend on the exposure status of anyone else.10 29 This assumption can be violated by 'spillover effects' of some exposures (eg, vaccination), where an individual's outcomes are affected by the exposure status of those around them..."	(Igelström et al. 2022)
	Epidemiology	"...For ease of presentation, the above discussion only briefly mentioned an assumption that plays an important role in the definition and estimation of causal effects. This assumption is sometimes dubbed the Stable Unit Treatment Value Assumption, or SUTVA, but we refer to it by a more accessible name, **non-interference**. The **non-interference assumption** cuts through this complexity by ignoring the potential outcomes that would arise if subject i were affected by the treatment of other subjects. Formally, we reduce the schedule of potential outcomes Y(d)¿ where d describes all of the treatments administered to all subjects, to a much simpler schedule Y.(d)y where d refers to the treatment administered to subject i..."	(Gerber and Green 2012)
	Education	"...The only identifying assumption for causal inference from an RCT that cannot be verified is the assumption that each unit's potential outcomes are independent of the assignment of other units to the treatment conditions.158 This "**non-interference**" or "**no-spillover**" assumption may be violated in some education research designs...."	(Elmendorf and Shanske 2018)

Supplementary Table 5. The collected definitions for the Stable Unit Treatment Value Assumption (SUTVA)

Synonym	Field	Direct Quote from the manuscript cited	Citation
Stable Unit Treatment Value Assumption (SUTVA)	Computer Science	"...You will also commonly see the **stable unit-treatment value assumption (SUTVA)** in the literature. **SUTVA** is satisfied if unit (individual) outcome is simply a function of unit's treatment. Therefore, **SUTVA** is a combination of consistency and no interference (and also deterministic potential outcomes). Active reading exercise: convince yourself that **SUTVA** is a combination of consistency and no inference..."	(Neal 2020)
	Conservation Science	"...No interference [also known as **stable unit treatment value assumption** or "no spillovers among units" assumption]:..."	(Ferraro et al. 2019)
	Conservation Science	"...A combination of both consistency and no-interference leads to the **stable unit-treatment value assumption** (**SUTVA**) (Rubin 1980). While interference is hard to restrain in the observational setting, in many causal inference applications the **stable unit-treatment value assumption** is implicitly adopted. Although a randomized control trial poses limitations on SUTVA violations, the strength of the randomized control trial lies in its implication of the ignorability assumption:..."	(Vonk et al. 2023)
	Statistical Science	"...Assumption 2.1 [**Stable Unit Treatment Value Assumption (SUTVA)**]. The potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes. This assumption emphasizes two points: The first point is the independence of each unit, that is, there are no interactions between units. In the context of the above illustrative example, one patient's outcome will not affect other patients' outcomes. The second point is the single version for each treatment. In the above example, Medicine A with different dosages are different treatments under the **SUTVA assumption**. **SUTVA** states that the potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes. This assumption mainly focuses on two aspects: (1) units are independent and identically distributed (i.i.d.); and (2) there only exists a single level for each treatment. An extensive literature exists on making causal inferences under **SUTVA**, but..."	(Yao et al. 2021)
	Statistical Science	"...In any case, the depiction in Figure 1 requires assumptions for it to be adequate—in particular, **SUTVA (stable unit treatment value assumption)** (Rubin 1980), which comprises two subassumptions. First, it assumes that there is no interference between units (Cox 1958); that is, neither Y i(1) nor Yi(0) is affected by what action any other unit received. Second, it assumes that there are no hidden versions of treatments; no matter how unit i received treatment 1, the outcome that would be observed would be Yi(1) and similarly for treatment 0..."	(Rubin 2005)

Supplementary Table 5 *continued on next page*

Supplementary Table 5 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Epidemiology	"...A type of independence assumption commonly made in the causal literature is known as the **Stable Unit Treatment Value Assumption (SUTVA)**; this requires consistency and further states that an individual's outcome is affected only by the treatment they receive, but not that of others (Cox, 1958; Rubin, 1980). This is also referred to as a lack of 'interference' or 'spill-over', and is most easily explained in terms of examples where it does not hold."	(Moodie and Stephens 2022)
	Epidemiology	"...**SUTVA** is the a priori **assumption** that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive. (p. 961). The purpose of the **SUTVA assumption** is to guarantee that a single value of the response will be observed for the participant under T and a single value of the response (possibly [hopefully] different) will be observed7 under C. Violations of **SUTVA** can also occur if there are hidden variants of an ostensibly well-defined treatment condition (e.g., different variants of the treatment are implemented at different sites in a multisite experiment) and the participants have different responses to the treatment variants."	(Hernán and Robins 2010)
	Education	"...Under **SUTVA**, there is only one version of the treatment and control and the potential outcomes Yi = ¬Y i(1), Yi(0) , i = 1, . . . , n, depend only on the treatment assigned to subject i and not on the treatment assigned to subject j for i j ."	(Chan 2024)
	Medicine	"...**Stable unit treatment value assumption (SUTVA)**. In words, this assumption says that the treatment affects only the subject taking the treatment and that there are not different versions of the treatment which have different effects (see [43, 44] for details)"	(Baiocchi et al. 2014)
	Epidemiology	"...**Stable unit treatment value assumption (SUTVA)**: The assumption that all versions of the treatment has the same effect (i.e., versions of the treatment with differences substantial enough to have different health effects are referred to as some other type of treatment), and each unit's outcomes are unaffected by the treatment values of other units."	(Matthay et al. 2020)
	Epidemiology	"...The consistency and non-interference assumptions together are sometimes known as the **stable unit treatment value assumption**."	(Igelström et al. 2022)
	Psychology	"...PO formulates two main sets of assumptions: the **stable-unit-treatment-value assumption (SUTVA)** with regard to the potential outcomes and assumptions about the assignment mechanism, the second core part of PO. **SUTVA** has two parts: the no interference and the no hidden variations in treatments assumption (Imbens & Rubin, 2015)."	(Steiner et al. 2023)
	Statistical Science	"...A fundamental assumption usually made in the potential outcomes approach to causal inference is that of no interference between individuals (Cox 1958), a critical component of the stable unit treatment value assumption (SUTVA) (Rubin 1980)."	(Hudgens and Halloran 2008)

Supplementary Table 5 *continued on next page*

Supplementary Table 5 *(continued)*

Synonym	Field	Direct Quote from the manuscript cited	Citation
	Psychology	"...The first assumption is **stable unit treatment value assumption (SUTVA**; Rubin, 1980).8 This assumption requires that one's counterfactual states (Y0,i and Y1,i) do not depend on the treatment status of other individuals. One can note that in the math above, there is no interference among individuals: An individual's outcome Yi does not depend on the treatment received by person j. As discussed below, many problems in developmental science may not fit this assumption."	(Foster 2010)
	Statistical Science	"...The assumption that such a representation is adequate may be called the **stable unit-treatment value assumption**..."	(Rubin 1980)
	Economics	"...Finally, to make the model's representation of outcomes adequate for causal analysis, the **stable-unit-treatment-value assumption (SUTVA)** has to be satisfied for all members of the population. Here, the most important implication is that the values of Yt and Ye do not depend on the treatment for units other than unit i (Rubin 1991)."	(Lechner 1999)
Stable unit treatment assumption	Medicine	"...I shall now assume for simplicity that the outcomes of individual i are independent of the outcomes of other individuals and their received treatment levels. This is referred to as the **stable-unit-treatment-assumption** [23]."	(Höfler 2005)
No-macro-effect assumption / Partial equilibrium assumption	Sociology	"...In most applications, the potential outcome model retains its tractability through the maintenance of a strong assumption known as the **stable unit treatment value assumption or SUTVA** (see Rubin 1980b, 1986). In economics, a version of this assumption is sometimes referred to as a **no-macro-effect** or **partial equilibrium assumption** (see Garfinkel, Manski, and Michalopoulos 1992, Heckman 2000, 2005, for the history of these ideas, and Manski and Garfinkel 1992 for examples). **SUTVA** is a much maligned acronym, and many others use different labels. Manski (2013a:S1), for example, has recently labeled the same assumption the "individualistic treatment response" assumption in order "to mark it as an assumption that restricts the form of treatment response functions."	(Morgan and Winship 2014)
Individualistic treatment response assumption	Economics	"...Rubin (1978) called it the **Stable Unit Treatment Value Assumption**. I call it **individualistic treatment response (ITR)**, to mark it as an assumption that restricts the form of treatment function..."	(Manski 2013)

Supplementary Table 6. The scientific fields, paper titles and total number of citations of each referenced document in **Supplementary Tables 1 - 5**.

Scientific Field	Citation	Paper Title	Total Citations
Epidemiology	Rehkopf et al. (2016)	The Consistency Assumption for Causal Inference in Social Epidemiology: When a Rose Is Not a Rose	217

Supplementary Table 6 *continued on next page*

Supplementary Table 6 *(continued)*

Scientific Field	Citation	Paper Title	Total Citations
	Cole and Frangakis (2009)	The consistency statement in causal inference: a definition or an assumption?	620
	Hernán and Robins (2010)	Causal Inference: What if?	6,355
	Hernán (2016)	Does water kill? A call for less causal causal inferences	287
	Moodie and Stephens (2022)	Causal inference: critical developments, past and future	17
	Shiba and Kawahara (2021)	Using propensity scores for causal inference: pitfalls and tips	188
	Igelström et al. (2022)	Causal inference and effect estimation using observational data	118
	Westreich and Cole (2010)	Invited Commentary: Positivity in Practice	533
	Chao and Yu (2023)	Causal inference using regression-based statistical control: Confusion in Econometrics	6
	Gerber and Green (2012)	Field Experiments: Design, Analysis, and Interpretation	2,354
	Greenland and Robins (1986)	Identifiability, Exchangeability, and Epidemiological Confounding	876
	Matthay et al. (2020)	Alternative causal inference methods in population health research: Evaluating tradeoffs and triangulating evidence	116
Computer Science	Neal (2020)	Introduction to Causal Inference	140
	Vonk et al. (2023)	Disentangling causality: assumptions in causal discovery and inference	46
	Greenland et al. (1999)	Confounding and Collapsibility in Causal Inference	1,303
	Hatt and Feuerriegel (2024)	Sequential Deconfounding for Causal Inference with Unobserved Confounders	54
	Kuang et al. (2020)	Causal inference	187
Statistical Science	Colnet et al. (2024)	Causal inference methods for combining randomized trials and observational studies: a review	292
	Peters et al. (2016)	Causal inference using invariant prediction: identification and confidence intervals	1,471
	Rubin (2005)	Causal Inference Using Potential Outcomes Design, Modeling, Decisions	3,692
	Yao et al. (2021)	A survey on causal inference	926
	Dehejia and Wahba (1999)	Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs	4,218
	Hudgens and Halloran (2008)	Toward Causal Inference With Interference	1,245

Supplementary Table 6 *continued on next page*

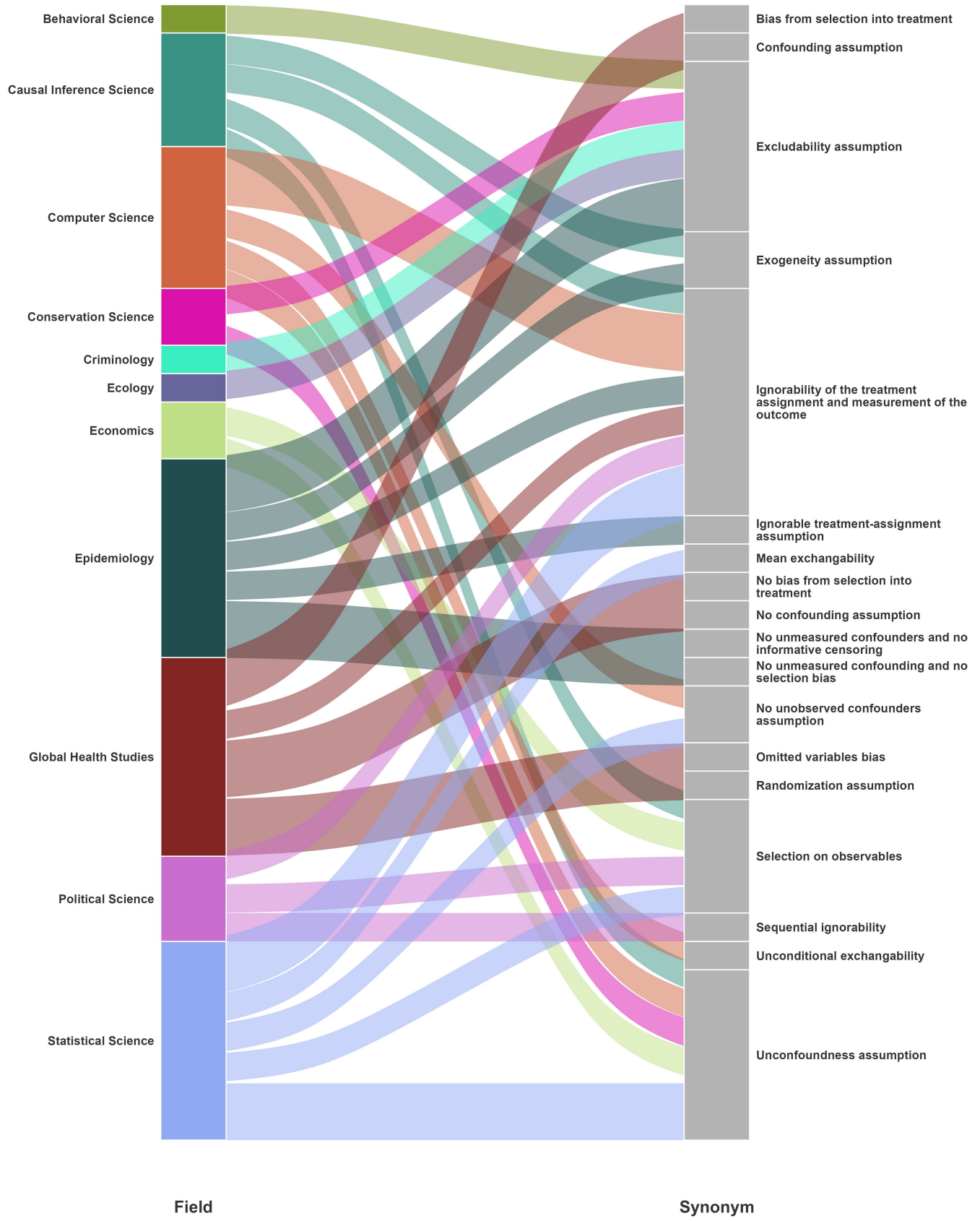
Supplementary Table 6 *(continued)*

Scientific Field	Citation	Paper Title	Total Citations
	Rubin (1980)	Randomization analysis of experimental data: The Fisher randomization test comment	2,671
	Cinelli and Hazlett (2020)	Making Sense of Sensitivity: Extending Omitted Variable Bias	1,404
	Imbens and Rubin (2015)	Causal Inference for Statistics, Social, and Biomedical Sciences	8,714
	Luna and Johansson (2014)	Testing for the Unconfoundedness Assumption Using an Instrumental Assumption	47
Medicine	Smit et al. (2023)	Causal inference using observational intensive care unit data: a scoping review and recommendations for future practice	32
	Bulbulia (2023)	A workflow for causal inference in cross-cultural psychology	25
	Goetghebeur et al. (2020)	Formulating causal questions and principled statistical answers	113
	Baiocchi et al. (2014)	Tutorial in Biostatistics: Instrumental Variable Methods for Causal Inference	0
	Höfler (2005)	Causal inference based on counterfactuals	386
Psychology	Foster (2010)	Causal inference and developmental psychology	260
	Steiner et al. (2023)	Frameworks for causal inference in psychological science	13
Ecology	Kimmel et al. (2021)	Causal assumptions and causal inference in ecological experiments	109
	Correia et al. (2025)	Designing causal mediation analyses to quantify intermediary processes in ecology	10
Education	Chan (2024)	Propensity score methods for causal inference and generalization	3
	Elmendorf and Shanske (2018)	Solving Problems No One Has Solved: Courts, Causal Inference, and the Right to Education	13
Economics	D'Amour et al. (2021)	Overlap in observational studies with high-dimensional covariates	348
	Lechner (2001)	Identification and estimation of causal effects of multiple treatments under the conditional independence assumption	1,154
	Rafieian (2023)	A Matrix Completion Solution to the Problem of Ignoring the Ignorability Assumption	3
	Manski (2013)	Identification of treatment response with social interactions	564
	Caliendo and Kopeinig (2008)	Some practical guidance for the implementation of propensity score matching	10,170
	Lechner (1999)	Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification	709
	Heckman et al. (1999)	Chapter 31 - The Economics and Econometrics of Active Labor Market Programs	5,124
Political Science	Blackwell and Glynn (2018)	How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables	193

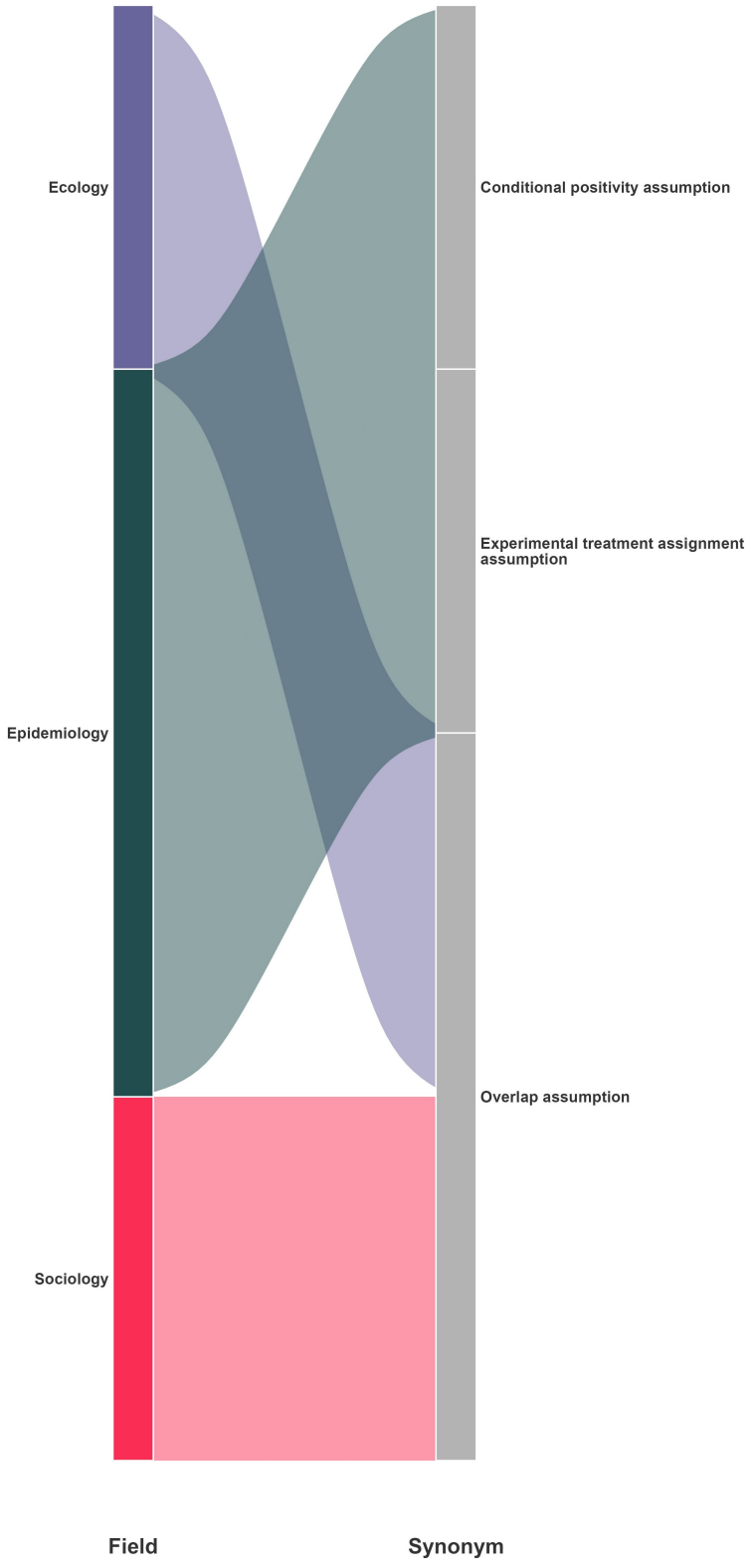
Supplementary Table 6 *continued on next page*

Supplementary Table 6 *(continued)*

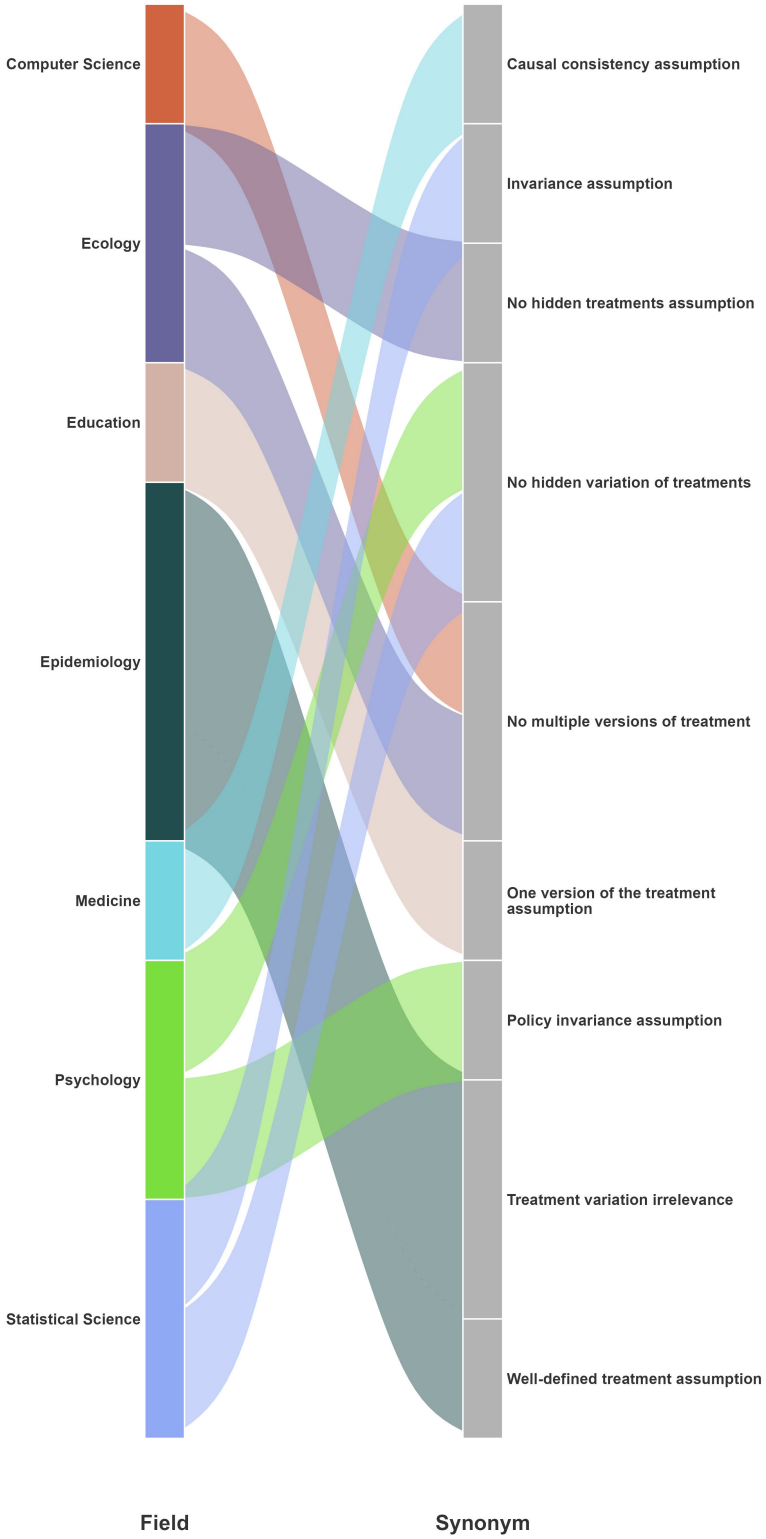
Scientific Field	Citation	Paper Title	Total Citations
	Muñoz et al. (2020)	Unexpected Event during Surveys Design: Promise and Pitfalls for Causal Inference	408
	Keele (2015)	The Statistics of Causal Inference: A View from Political Methodology	295
	Gangl (2010)	Causal Inference in Sociological Research	633
	Morgan and Winship (2014)	Counterfactuals and Causal Inference: Methods and Principles for Social Research	5,443
Conservation Science	Ferraro et al. (2019)	Causal inference in coupled human and natural systems	341
Behavioural Science	Tappin et al. (2020)	Thinking clearly about causal inferences of politically motivated reasoning: why paradigmatic study designs often undermine causal inference	208
Criminology	Nägel and Nivette (2023)	Unexpected events during survey design and trust in the police: a systematic review	32
		Total citations	**64,686**



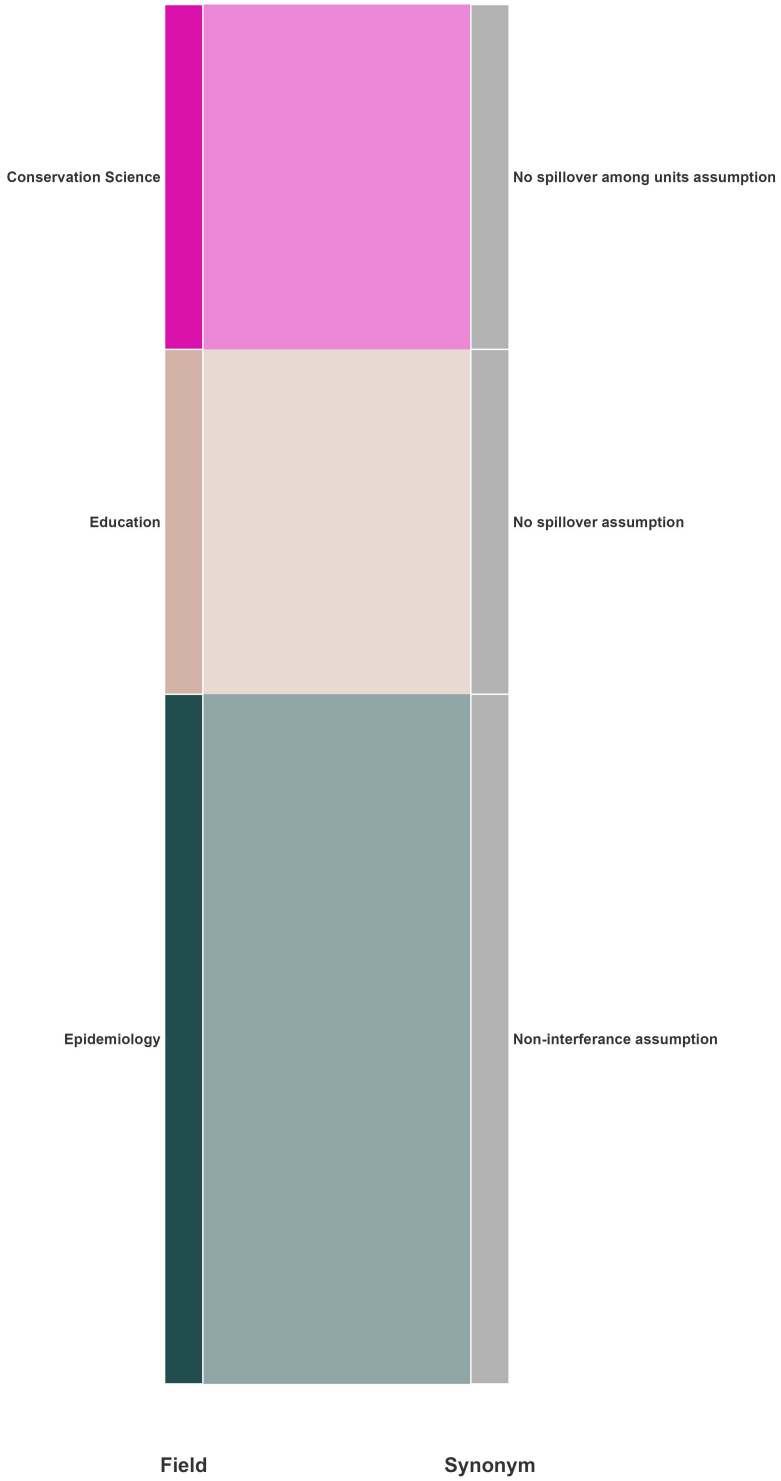
Supplementary Figure 2. Alluvial plot connecting the scientific fields in which the peer-reviewed scientific papers or books were published in (left) to the **Exchangeability assumption**.



Supplementary Figure 3. Alluvial plot connecting the scientific fields in which the peer-reviewed scientific papers or books were published in (left) to the **Positivity assumption**



Supplementary Figure 4. Alluvial plot connecting the scientific fields in which the peer-reviewed scientific papers or books were published in (left) to the **Consistency assumption**.



Supplementary Figure 5. Alluvial plot connecting the scientific fields in which the peer-reviewed scientific papers or books were published in (left) to the **No-Interference assumption**.

REFERENCES

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340. https://doi.org/10.1002/sim.6128.

Baylis, K., Garcia, A., and Heilmayr, R. (2026). Causal Inference for Biodiversity Conservation. *Review of Environmental Economics and Policy*, 20(1):000–000. https://doi.org/10.1086/739836.

Bertomeu, J., Beyer, A., and Taylor, D. J. (2016). From Casual to Causal Inference in Accounting Research: The Need for Theoretical Foundations. https://doi.org/10.2139/ssrn.2694105.

Blackwell, M. and Glynn, A. N. (2018). How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables. *American Political Science Review*, 112(4):1067–1082. https://doi.org/10.1017/S0003055418000357.

Bulbulia, J. A. (2023). A workflow for causal inference in cross-cultural psychology. *Religion, Brain & Behavior*, 13(3):291–306. _eprint: https://doi.org/10.1080/2153599X.2022.2070245.

Caliendo, M. and Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, 22(1):31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x.

Chan, W. (2024). Propensity score methods for causal inference and generalization. *Asia Pacific Education Review*, 25(3):647–662. https://doi.org/10.1007/s12564-023-09906-5.

Chao, F. and Yu, G. (2023). Causal inference using regression-based statistical control: Confusion in Econometrics. *Journal of Data and Information Science*, 8(1):21–28. https://doi.org/10.2478/jdis-2023-0006.

Cinelli, C., Feller, A., Imbens, G., Kennedy, E., Magliacane, S., and Zubizarreta, J. (2025). Challenges in Statistics: A Dozen Challenges in Causality and Causal Inference. *arXiv preprint arXiv:2508.17099*. https://doi.org/10.48550/arXiv.2508.17099.

Cinelli, C. and Hazlett, C. (2020). 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.

Cole, S. R. and Frangakis, C. E. (2009). The Consistency Statement in Causal Inference: A Definition or an Assumption? *Epidemiology*, 20(1):3. https://doi.org/10.1097/EDE.0b013e31818ef366.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2024). Causal Inference Methods for Combining Randomized Trials and Observational Studies: A Review. *Statistical Science*, 39(1):165–191. https://doi.org/10.1214/23-STS889.

Correia, H. E., Dee, L. E., Byrnes, J. E. K., Fieberg, J. R., Fortin, M.-J., Glymour, C., Runge, J., Shipley, B., Shpitser, I., Siegel, K. J., Sugihara, G., von Holle, B., and Ferraro, P. J. (2026). Best practices for moving from correlation to causation in ecological research. *Nature Communications*, 17(1):1981. https://doi.org/10.1038/s41467-026-69878-z.

Correia, H. E., Dee, L. E., and Ferraro, P. J. (2025). Designing causal mediation analyses to quantify intermediary processes in ecology. *Biological Reviews*, n/a(n/a). https://doi.org/10.1111/brv.70011.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale university press.

Dehejia, R. H. and Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, 94(448):1053–1062. https://doi.org/10.1080/01621459.1999.10473858.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654. https://doi.org/10.1016/j.jeconom.2019.10.014.

Eggers, A. C., Tuñón, G., and Dafoe, A. (2024). Placebo Tests for Causal Inference. *American Journal of Political Science*, 68(3):1106–1121. https://doi.org/10.1111/ajps.12818.

Elmendorf, C. S. and Shanske, D. (2018). Solving Problems No One Has Solved: Courts, Causal Inference, and the Right to Education. *University of Illinois Law Review*, 2018:693. https://dx.doi.org/10.2139/ssrn.2886754.

Ferraro, P. J. and Hanauer, M. M. (2014). Advances in Measuring the Environmental and Social Impacts of Environmental Programs. *Annual Review of Environment and Resources*, 39(Volume 39, 2014):495–517. https://doi.org/10.1146/annurev-environ-101813-013230.

Ferraro, P. J., Sanchirico, J. N., and Smith, M. D. (2019). Causal inference in coupled human and natural systems. *Proceedings of the National Academy of Sciences*, 116(12):5311–5318. https://doi.org/10.1073/pnas.1805563115.

Foster, E. M. (2010). Causal inference and developmental psychology. *Developmental Psychology*, 46(6):1454–1480. https://doi.org/10.1037/a0020204.

Gangl, M. (2010). Causal Inference in Sociological Research. *Annual Review of Sociology*, 36(Volume 36, 2010):21–47. https://doi.org/10.1146/annurev.soc.012809.102702.

Geldmann, J., Jones, J. P. G., Wauchope, H., and Ferraro, P. J. (2025). Causal claims, causal assumptions and protected area impact. *Nature*, 638(8052):E40–E41. https://doi.org/10.1038/s41586-024-08512-8.

Gerber, A. S. and Green, D. P. (2012). Field experiments: Design, analysis, and interpretation.

Gibson, L. and Zimmerman, F. (2021). Measuring the sensitivity of difference-in-difference estimates to the parallel trends assumption. *Research Methods in Medicine & Health Sciences*, 2(4):148–156. https://doi.org/10.1177/26320843211061306.

Goetghebeur, E., le Cessie, S., De Stavola, B., Moodie, E. E., Waernbaum, I., and Initiative, . o. t. t. g. C. I. T. o. t. S. (2020). Formulating causal questions and principled statistical answers. *Statistics in Medicine*, 39(30):4922–4948. https://doi.org/10.1002/sim.8741.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1):29–46. https://doi.org/10.1214/ss/1009211805.

Greenland, S. and Robins, J. M. (1986). Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology*, 15(3):413–419. https://doi.org/10.1093/ije/15.3.413.

Greenland, S. and Robins, J. M. (2009). Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations : EP+I*, 6:4. https://doi.org/10.1186/1742-5573-6-4.

Guizar-Coutiño, A., Nicholson, G., Coomes, D., Ferraro, P. J., Swinfield, T., and Jones, J. P. (2026). Unobserved confounders cannot explain over-crediting in avoided deforestation carbon projects. *Nature Ecology & Evolution*, pages 1–11. https://doi.org/10.1038/s41559-026-03049-7.

Hatt, T. and Feuerriegel, S. (2024). Sequential Deconfounding for Causal Inference with Unobserved Confounders. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, pages 934–956. PMLR. https://proceedings.mlr.press/v236/hatt24a.html.

Heckman, J. J., Lalonde, R. J., and Smith, J. A. (1999). Chapter 31 - The Economics and Econometrics of Active Labor Market Programs. In Ashenfelter, O. C. and Card, D., editors, *Handbook of Labor Economics*, volume 3, pages 1865–2097. Elsevier.

Hernán, M. A. (2016). Does water kill? A call for less casual causal inferences. *Annals of epidemiology*, 26(10):674–680. https://doi.org/10.1016/j.annepidem.2016.08.016.

Hernán, M. A. and Robins, J. M. (2010). *Causal inference: What If*. CRC Boca Raton, FL. https://miguelhernan.org/whatifbook.

Hudgens, M. G. and Halloran, M. E. (2008). Toward Causal Inference With Interference. *Journal of the American Statistical Association*, 103(482):832–842. https://doi.org/10.1198/016214508000000292.

Höfler, M. (2005). Causal inference based on counterfactuals. *BMC Medical Research Methodology*, 5(1):28. https://doi.org/10.1186/1471-2288-5-28.

Igelström, E., Craig, P., Lewsey, J., Lynch, J., Pearce, A., and Katikireddi, S. V. (2022). Causal inference and effect estimation using observational data. *J Epidemiol Community Health*, 76(11):960–966. https://doi.org/10.1136/jech-2022-219267.

Imbens, G. W. and Rubin, D. B. (2010). Rubin Causal Model. In Durlauf, S. N. and Blume, L. E., editors, *Microeconometrics*, pages 229–241. Palgrave Macmillan UK, London. https://doi.org/10.1057/9780230280816_28.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jones, J. P. G., Barnes, M., Eklund, J., Ferraro, P. J., Geldmann, J., Oldekop, J. A., and Schleicher, J. (2022). Quantifying uncertainty about how interventions are assigned would improve impact evaluation in conservation: reply to Rasolofoson 2022. *Conservation Biology*, page e14007. https://doi.org/10.1111/cobi.14007.

Keele, L. (2015). The Statistics of Causal Inference: A View from Political Methodology. *Political Analysis*, 23(3):313–335. https://www.jstor.org/stable/24573164.

Kimmel, K., Dee, L. E., Avolio, M. L., and Ferraro, P. J. (2021). Causal assumptions and causal inference in ecological experiments. *Trends in Ecology & Evolution*, 36(12):1141–1152. https://doi.org/10.1016/j.tree.2021.08.008.

Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., Huang, H., Ding, P., Miao, W., and Jiang, Z. (2020). Causal Inference. *Engineering*, 6(3):253–263. https://doi.org/10.1016/j.eng.2019.08.016.

Lechner, M. (1999). Earnings and Employment Effects of Continuous Gff-the-Job Training in East Germany After Unification. *Journal of Business & Economic Statistics*, 17(1):74–90. https://doi.org/10.1080/07350015.1999.10524798.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In Lechner, M. and Pfeiffer, F., editors, *Econometric Evaluation of Labour Market Policies*, pages 43–58, Heidelberg. Physica-Verlag HD. https://doi.org/10.1007/978-3-642-57615-7_3.

Luna, X. d. and Johansson, P. (2014). Testing for the Unconfoundedness Assumption Using an Instrumental Assumption. *Journal of Causal Inference*, 2(2):187–199. https://doi.org/10.1515/jci-2013-0011.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23. https://www.jstor.org/stable/23364965.

Matthay, E. C., Hagan, E., Gottlieb, L. M., Tan, M. L., Vlahov, D., Adler, N. E., and Glymour, M. M. (2020). Alternative causal inference methods in population health research: Evaluating tradeoffs and triangulating evidence. *SSM - Population Health*, 10:100526. https://doi.org/10.1016/j.ssmph.2019.100526.

Moodie, E. E. M. and Stephens, D. A. (2022). Causal inference: Critical developments, past and future. *Canadian Journal of Statistics*, 50(4):1299–1320. https://doi.org/10.1002/cjs.11718.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, Cambridge, 2 edition.

Muñoz, J., Falcó-Gimeno, A., and Hernández, E. (2020). Unexpected Event during Survey Design: Promise and Pitfalls for Causal Inference. *Political Analysis*, 28(2):186–206. https://doi.org/10.1017/pan.2019.27.

Neal, B. (2020). Introduction to causal inference. *Course Lecture Notes (draft)*. https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf.

Nägel, C. and Nivette, A. E. (2023). Unexpected events during survey design and trust in the police: a systematic review. *Journal of Experimental Criminology*, 19(4):891–917. https://doi.org/10.1007/s11292-022-09508-y.

Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics*, 37(2):187–204. https://doi.org/10.1080/07350015.2016.1227711.

Pearl, J. (2003). Statistics and causal inference: A review. *Test*, 12(2):281–345. https://doi.org/10.1007/BF02595718.

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

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal Inference by using Invariant Prediction: Identification and Confidence Intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012. https://doi.org/10.1111/rssb.12167.

Pizer, S. D. (2016). Falsification Testing of Instrumental Variables Methods for Comparative Effectiveness Research. *Health Services Research*, 51(2):790–811. https://doi.org/10.1111/1475-6773.12355.

Rafieian, O. (2023). A Matrix Completion Solution to the Problem of Ignoring the Ignorability Assumption. https://doi.org/10.2139/ssrn.4546996.

Rehkopf, D. H., Glymour, M. M., and Osypuk, T. L. (2016). The Consistency Assumption for Causal Inference in Social Epidemiology: When a Rose Is Not a Rose. *Current Epidemiology Reports*, 3(1):63–71. https://doi.org/10.1007/s40471-016-0069-5.

Rosenbaum, P. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26. https://doi.org/10.1093/biomet/74.1.13.

Rosenbaum, P. R. (1989). The Role of Known Effects in Observational Studies. *Biometrics*, 45(2):557–569. https://doi.org/10.2307/2531497.

Rosenbaum, P. R., Rosenbaum, P., and Briskman (2010). *Design of observational studies*, volume 10. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55. https://doi.org/10.1093/biomet/70.1.41.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688. https://psycnet.apa.org/doi/10.1037/h0037350.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371):591–593. https://doi.org/10.2307/2287653.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331. https://doi.org/10.1198/016214504000001880.

Shiba, K. and Kawahara, T. (2021). Using Propensity Scores for Causal Inference: Pitfalls and Tips. *Journal of Epidemiology*, 31(8):457–463. https://doi.org/10.2188/jea.JE20210145.

Siegel, K. and Dee, L. E. (2025). Foundations and Future Directions for Causal Inference in Ecological Research. *Ecology Letters*. https://doi.org/10.1111/ele.70053.

Smit, J. M., Krijthe, J. H., Kant, W. M. R., Labrecque, J. A., Komorowski, M., Gommers, D. a. M. P. J., van Bommel, J., Reinders, M. J. T., and van Genderen, M. E. (2023). Causal inference using observational intensive care unit data: a scoping review and recommendations for future practice. *npj Digital Medicine*, 6(1):1–11. https://doi.org/10.1038/s41746-023-00961-1.

Steiner, P. M., Shadish, W. R., and Sullivan, K. J. (2023). Frameworks for causal inference in psychological science. In *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics, Vol. 1, 2nd ed*, APA Handbooks in Psychology®, pages 23–56. American Psychological Association, Washington, DC, US. https://doi.org/10.1037/0000318-002.

Tappin, B. M., Pennycook, G., and Rand, D. G. (2020). Thinking clearly about causal inferences of politically motivated reasoning: why paradigmatic study designs often undermine causal inference. *Current Opinion in Behavioral Sciences*, 34:81–87. https://doi.org/10.1016/j.cobeha.2020.01.003.

Vonk, M. C., Malekovic, N., Bäck, T., and Kononova, A. V. (2023). Disentangling causality: assumptions in causal discovery and inference. *Artificial Intelligence Review*, 56(9):10613–10649. https://doi.org/10.1007/s10462-023-10411-9.

Westreich, D. and Cole, S. R. (2010). Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*, 171(6):674–677. https://doi.org/10.1093/aje/kwp436.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2021). A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46.

SUPPLEMENTARY INFORMATION 3 – FULL METHODOLOGY

1. ARTICLE SEARCH PROTOCOL

1.1. *Peer-reviewed search string*

We adopted and modified the search protocol from a study which aimed to assess the global impact of conservation actions (Langhammer et al. 2024). The search string is based on the pressure-state-response (PSR) model with an additional impact section. In our search format pressure/state refers to specific pressures on biodiversity such as deforestation and species populations, response refers to conservation actions, in our case, protected area establishment and impact refers to the specific study designs used (Morgan and Winship 2014; Ferraro and Hanauer 2014; Imbens and Rubin 2015; Pearl and Mackenzie 2018; Cunningham 2021). A comprehensive and replicable list of search strings used in this systematic review can be found in **Supplementary Table 7** & **8**.

Supplementary Table 7. Boolean search terms used in the Web of Science - Core Collection, SCOPUS and in the ProQuest Dissertations Thesis Global Database, adapted from Langhammer et al. (2024)

Intervention	Response	Pressure	Study design
Establishment and management of protected areas	"protected area*" OR "conservation area*" OR "nature reserve*" OR "reserve" OR "reserves" OR "site protection" OR "habitat protection" OR "land protection" OR "national park"	"biodiversity" OR "species population*" OR "fish population*" OR "key biodiversity area" OR "extinction risk" OR "deforestation" OR "forest loss" OR "nature" OR "habitat loss" OR "socioeconomic" Or "economic" OR "poverty" OR "biomass"	"directed acyclic graph" OR "Potential outcomes" OR "Rubin Causal Model" OR "Propensity score" OR "Match*" OR "Regression Discontinuity" OR "Instrumental Variable" OR "Panel Data" OR "difference-in-differences" OR "Synthetic control" OR "Before after control impact" OR "BACI" OR "control impact" OR "before after" OR "longitudinal" OR "interrupted time series" OR "Randomized control trial" OR "counterfactual" OR "causal" OR "random allocat" or "random assign" or "paired" OR "quasi-experimental" OR "experimental" OR "impact evaluation" OR "conservation outcome" Or "spillover" OR "robust evaluation*"

Supplementary Table 8. Complete Boolean search strings used in Web of Science and SCOPUS.

Intervention	Study design
Web of Science (search set to "ALL FIELDS")	(("protected area*" OR "conservation area*" OR "nature reserve*" OR "reserve" OR "reserves" OR "site protection" OR "habitat protection" OR "land protection" OR "national park") AND ("biodiversity" OR "species population*" OR "fish population*" OR "key biodiversity area" OR "extinction risk" OR "deforestation" OR "forest loss" OR "nature" OR "habitat loss" OR "socioeconomic" Or "economic" OR "poverty" OR "biomass") AND ("directed acyclic graph" OR "Potential outcomes" OR "Rubin Causal Model" OR "Propensity score" OR "Match*" OR "Regression Discontinuity" OR "Instrumental Variable" OR "Panel Data" OR "difference-in-differences" OR "Synthetic control" OR "Before after control impact" OR "BACI" OR "control impact" OR "before after" OR "longitudinal" OR "interrupted time series" OR "Randomized control trial" OR "counterfactual" OR "causal" OR "random allocat" or "random assign" or "paired" OR "quasi-experimental" OR "experimental" OR "impact evaluation" OR "conservation outcome" Or "spillover" OR "robust evaluation*"))
SCOPUS	TITLE-ABS-KEY ((protected area* OR conservation area* OR nature reserve* OR reserve OR reserves OR site protection OR habitat protection OR land protection OR national park) AND (biodiversity OR species population* OR fish population* OR key biodiversity area OR extinction risk OR deforestation OR forest loss OR nature OR habitat loss OR socioeconomic OR economic OR poverty OR biomass) AND (directed acyclic graph* OR Potential outcomes OR Rubin Causal Model OR Propensity score OR Matching OR Matched OR Regression Discontinuity OR Instrumental Variable* OR Panel Data OR difference-in-differences OR Synthetic control OR Before after control impact OR BACI OR control impact OR before after OR longitudinal OR interrupted time series OR Randomized control trial OR counterfactual* OR causal* OR random allocat* OR random assign* OR paired OR quasi-experimental OR experimental OR impact evaluation OR conservation outcome* OR spillover OR robust evaluation))

We conducted our searches between July and August 2025. The search string was conducted in the electronic databases Web of Science - Core Collection and SCOPUS (Search by TITLE-ABS-KEY) using institutional access from the University of Oxford and the Bodleian Libraries (see **Supplementary Table 9**). No additional filters had been placed for these searches. We conducted our search using the protocol and created a compiled library using EndNote 20 (https://endnote.com/). Our search yielded 12,682 results.

1.2. *Grey literature searches*

Grey literature was searched with two different methods and implemented within our systematic review to reduce the risk of publication bias (Haddaway et al. 2020). First, we reused our search string in ProQuest Dissertations & Thesis Global for any unpublished dissertations and theses. Second, we searched through the; (i) IUCN Library System; (ii) the World Bank; (iii) the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES); (iv) the UN Environment Programme World Conservation Monitoring Centre (UNEP-WCMC); (v) World Wildlife Fund for Nature (WWF); (vi) Wildlife Conservation Society (WCS); (vii) Conservation International (CI); (viii) Birdlife International; (ix) The Nature Conservancy (TNC) (x) Flora and Fauna International (FFI); (xi) the Zoological Society of London (ZSL); and (xii) the British Ecological Society's Applied Ecological Resources using simplified search strings to capture any additional impact evaluations that might be associated with PAs (**Supplementary Table 10**). We decided to stop our search for grey literature at this stage as outside of unpublished thesis and dissertation, the IUCN Library System and the World Bank, very limited results were found. Our grey literature search resulted in 1,725 documents, none of which were duplicates. Merging our grey and scientific literature resulted in a dataset of 14,407 results, and after deduplication through the Systematic Review Accelerator's Deduplicator (Forbes et al. 2024), our dataset yielded 11,641 unique results (see **Figure 2** in the main manuscript). We further filtered our dataset to only

Supplementary Table 9. Boolean search terms used in the Web of Science - Core Collection, SCOPUS and in the ProQuest Dissertations Thesis Global Database, adapted from Langhammer et al. (2024)

Index	Date range	Covered?
Social Science Index (SSCI)	1970 - Present	Yes
Science Citation Index Expanded (SCI-EXPANDED)		Yes
Emerging Sources Citation Index (ESCI)		Yes
Conference Proceedings Citation Index - Science (CPCI-S)		Yes
Arts & Humanities Citation Index (A&HCI)		Yes
Conference Proceedings Citation Index - Social Science & Humanities (CPCI-SSH)		Yes
Book Citation Index - Social Science & Humanities (BKCI-SSH)		Yes
Book Citation Index - Science (BKCI-S)		Yes

include studies post-2006 documents to align with Ferraro and Pattanayak (2006)'s publication year, resulting in our final dataset having 10,280 documents.

1.3. *Search comprehensiveness and benchmarking*

To ensure our search protocol was comprehensive and would capture all relevant studies, we had compiled a benchmark dataset of 20 peer-reviewed protected area impact evaluations that meet our inclusion criteria and assessed whether our searches captured the benchmark studies. When we compared our dataset (n = 10,280) with this benchmark, we achieved a hit rate of 100% (n = 20). To further enhance the comprehensiveness of our dataset, we conducted citation chasing (Haddaway et al. 2022). Citation chasing is a popular approach used within evidence synthesis that makes use of the connections a peer-reviewed article generates, through both a document's references (*Backward citation chasing*) and new documents that have cited it (*Forward citation chasing*). There are no strict guidelines on how to conduct citation chasing. For this study, we selected 20 peer-reviewed papers from a range of well-known papers throughout the years. In total, 812 documents were identified through backward citation chasing and 923 through forward citation chasing. From these documents, 32 fit the inclusion criteria and were added to our dataset of studies. Both the benchmarking process and citation chasing are transparently documented within Caruana et al. (2026).

2. SCREENING PROCESS

Articles were screened in two stages: (i) Title & Abstract and (ii) Full-text. Each article was screened based on our inclusion and exclusion criteria below throughout each stage of the screening process. If there were any doubts about whether an article was relevant, we included it to the next screening stage. All title and abstract screening was done through "ASReview" (van de Schoot et al. 2021). ASReview (https://asreview.nl/) is a free and open-sourced software that uses machine learning tools to reduce the labour involved in the screening phase of meta-research (Callaghan et al. 2024; Quan et al. 2024). Each time the researcher labels a study as relevant or non-relevant the algorithm updates its understanding of what studies are relevant and eventually only shows potentially relevant studies to the researcher (van de Schoot et al. 2021). We stopped screening following a four-fold stopping heuristic (Boetje and van de Schoot 2024); (i) All our benchmarks had been identified within our relevant documents; (ii) We had screened 2,073 documents, which is seven times a crude estimate of the total relevant documents calculated by randomly sampling and labelling 1% of our dataset (n = 103) and multiplying this by the full dataset (see Caruana et al. (2026) for full calculations). The minimum recommended is twice this crude estimate; (iii) Our screened documents represent 20% of the entire dataset (minimum being 10% recommended); and lastly (iv) We identified no further relevant documents in the last 278 consecutively screened documents (with the recommended being at least 50 consecutive records). **Supplementary Table 11** provides a detailed description of our inclusion and exclusion criteria at the Title & Abstract and Full text stage. A decision tree (Foo et al. 2021) illustration of **Supplementary Table 11** is available in **Supplementary Figure 6**. Since our study is focused on the causal assumptions made within a study's design (further details in **Section 3.3**), formal critical appraisal was not conducted. However, note that the causal assumptions are partially captured by Criterion 1 of the Collaboration for Environmental Evidence Critical Appraisal Tool (Konno et al. 2021).

Supplementary Table 10. Grey literature search protocol

Organisation	Search Configuration	Keywords
IUCN Library System	Not applicable	"Ecological evaluation", "Economic evaluation", "Evaluation", "Evaluation and monitoring", "Evaluation reports", "Evaluation techniques", "Monitoring and evaluation", "Programme evaluation", "Project evaluation", "Resources evaluation", "Site evaluation"
World Bank	Not applicable	"Biodiversity conservation evaluate"
World Wildlife Fund for Nature (WWF)	Not applicable	"Evaluation"
Wildlife Conservation Society (WCS)	Export results to excel, filter by publications: WCS Manually checking	"Evaluation" "impact"
Conservation International (CI)	Publication titles	"Evaluation", "Impact evaluation", "Impact"
Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES)	Not applicable	"Evaluation"
UN Environment Programme World Conservation Monitoring Centre (UNEP- WCMC)		"Evaluation", "Impact evaluation"
BirdLife International	Search results type: Report	"Evaluation"
The Nature Conservancy (TNC)	Not applicable	"Evaluation"
Flora and Fauna International (FFI)	Downloaded only "Publication & Reports" documents	"Evaluation"
Zoological Society of London (ZSL)	Not applicable	"Evaluation"
British Ecological Society's Applied Ecological Resources (AER)	Non peer-reviewed documents only	"Evaluation"

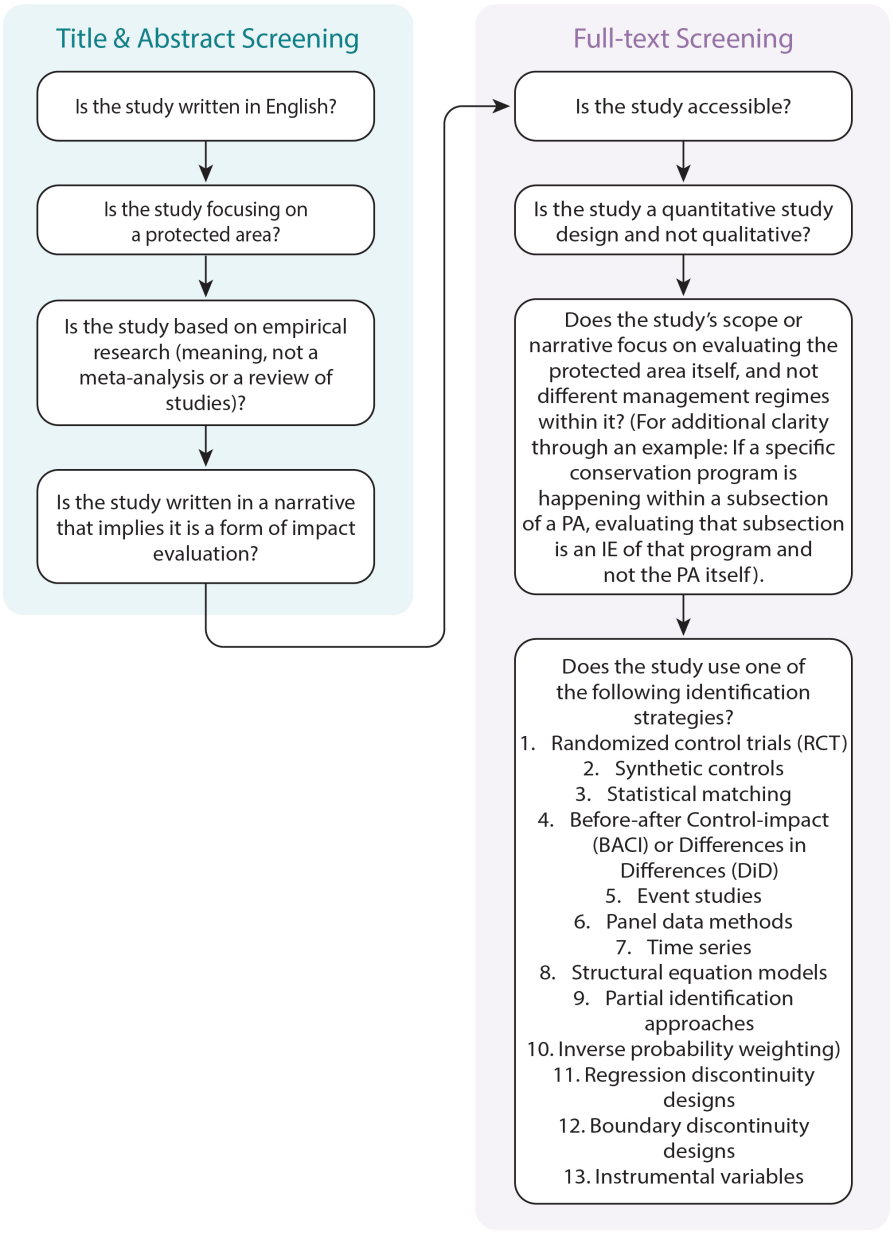
We did not exclude studies on this basis and chose to retain all studies regardless of how well they acknowledged or adhered to these assumptions.

Supplementary Table 11. Inclusion and exclusion criteria for article screening at the Title and Abstract, and Full-text level.

Question Element	Screening at Title & Abstract		Screening at Full Text	
	Inclusion	**Exclusion**	**Inclusion**	**Exclusion**
S - Type of Study Design	Studies will be included if they refer to evaluation or assessment (through keywords such as but not limited to: (i) *evaluating*; (ii) *evidence*; (iii) *impact*; (iv) *efficacy*; (v) *effective*; (vi) *comparison*; (vii) *assessments*; (viii) *causality* of a protected area. If studies do not refer to evaluation but hint one has been carried out (for example, *"A protected area reduced X"*) they will proceed to the next stage of the review.	All other studies are excluded.	Studies must be a quantitative impact evaluation of an entire protected area. Additionally, they must have used at least one of the following study designs: (i) *Randomized control trials* (RCT); (ii) *Synthetic controls* (iii) *Statistical matching*; (iv) *Before-after Control-impact (BACI)* or *Difference in Differences (DiD)*; (v) *Event studies*; (vi) *Panel data methods*; (vii) *Time series*; (viii) *Structural equation models*; (ix) *Partial identification approaches*; (x) *Inverse probability weighting*); (xi) *Regression discontinuity designs*; (xii) *Boundary discontinuity designs*; (xiii) *Instrumental variables*	Studies that use qualitative approaches (e.g., *Process tracing*) are excluded from our review as they operate on the Potential outcomes framework or the Structural causal model framework. Studies which assess the effectiveness of *different interventions* within the boundaries of a protected area only (e.g., different management regimes) are excluded. Our review focuses on studies that assess the effectiveness of protected area designation as a policy decision. Studies evaluating other interventions within a protected area, shift the focus toward management decisions, rather than the policy decision itself. Studies using simple study designs such as *Control-impact (CI)*; *Before-after (BA)*; and *After only* are excluded from full-data extraction in this review but were kept for partial data extraction.

Continued on next page

Question Element	Screening at Title & Abstract		Screening at Full Text	
	Inclusion	**Exclusion**	**Inclusion**	**Exclusion**
D – Types of Data	Not applicable at the title-screening level.	Articles mentioning they are evidence-synthesis such as systematic reviews and systematic maps in the titles are excluded.	Studies must include empirical analysis. Forums, essays and methodological studies can be included if worked examples using empirical data are provided.	Articles based on evidence-synthesis such as systematic reviews, systematic maps, meta-analyses, opinion articles, commentaries, and forums are excluded from this systematic review's data analysis to prevent double-counting of papers.
M – Types of Methods	Our study aims to review all protected area impact evaluations that have used the study designs outlined in S – Types of Study Design, it is beyond our scope to categories different methods within the categories of study design.			
O – Types of Outcome	We include any outcome variables including ecological (e.g., habitat extent condition, species population trends, species community composition, biodiversity indices and the presence or absence of rare species), socio-economic (e.g., economic wellbeing, welfare and multidimensional poverty) and human pressures (e.g., Land-use change & human footprint).			
Other	Studies must be in English	Any other language is excluded.	Studies must be in English	Any other language is excluded.



Supplementary Figure 6. Decision tree illustration for Title & Abstract and full-text screening

3. DATA CODING STRATEGY

Full data extraction were conducted on the 308 studies using causal identification strategies. Extracted variables were organised into six categories: (i) metadata; (ii) study design specifications; (iii) causal reasoning; (iv) causal assumptions; (v) open science practises; and (vi) research integrity measures. Each category is described in the subsections below, and the full data extraction form is presented in **Supplementary Table 12**. All extraction was conducted via Covidence (https://www.covidence.org/). Given the methodological scope of this review, a traditional evidence table is not applicable.

Supplementary Table 12. Data extraction form

Category	Data extraction Question	Type of data recorded
Metadata	1. Document Title	1. Text
	2. First Author	2. Text
	3. Year of Publication	3. Continuous
	4. DOI Link	4. Link
	5. Country/s of Study	5. Text
	6. Scale of study (Global, Multinational, National, Regional, Individual)	6. Categorical (Nominal)
	7. Protected area name/s & WDPA ID/s	7. Text
	8. Protected area realm (Terrestrial, Marine or Freshwater)	8. Categorical (Nominal)
Study design specification	9. Identification strategy/s used	9. Categorical (Nominal)
	10. Study design/s used	10. Categorical (Nominal)
	11. Study's years covered	11. Text
	12. Main outcome variable/s assessed	12. Text
	13. Outcome variable classification (Ecological, Socioeconomic, Human Pressure	14. Multi-label categorical
Causal reasoning	14. Was a Theory of Change depicted?	14. Binary
	15. Was a Direct Acyclic Graph depicted?	15. Binary
	16. Was a Causal Estimand (e.g., ATT, ATE, LATE) specified?	16. Binary
	17. If Yes, what type?	17. Text
Causal assumptions	18. Exchangeability assumptions Tier	18. Categorical (Nominal)
	19. Exchangeability assumption components	19. Multi-label categorical
	20. Positivity assumptions Tier	20. Categorical (Nominal)
	21. Positivity assumption components	21. Multi-label categorical
	22. Consistency assumption Tier	22. Categorical (Nominal)
	23. Consistency assumption components	23. Multi-label categorical
	24. Consistency sub-classification of treatment heterogeneity	24. Categorical (Nominal)
	25. No-Interference assumption Tier	25. Categorical (Nominal)
	26. No-Interference assumption components	26. Multi-label categorical
Open science practises	27. Was the study pre-registered?	27. Binary
	28. Is the code publicly available?	28. Categorical (Nominal)
	29. Is the archived data publicly available?	29. Categorical (Nominal)
Research integrity measures	30. Did the study state they underwent an ethics review (if human participants are involved)?	30. Categorical (Nominal)
	31. Is there any conflicts of interest explicitly disclosed in the document	31. Binary

3.1. *Metadata*

We extracted a comprehensive set of metadata for each study, covering both document-level details and study context. Document-level information included the title, first author, publication year, journal names, and Digital Object Identifier (DOI). For study context, we recorded the countries covered, the spatial scale of the study, the realm, and the names of any identifiable protected areas, which were cross-referenced with the World Database on Protected Areas (UNEP-WCMC and IUCN 2026).

3.2. *Study design specification*

For each included study, we extracted detailed information on its specific study design. We first recorded the years covered by the study and the study design as described by the authors, inferring a broad category where none was specified. We then recorded the main outcome variables of interest analysed within each study (e.g., deforestation), alongside the respective dataset used. Based on this information, we assigned a broad label indicating whether a study ecological outcomes, socioeconomic outcomes, human pressures or some combination of the three.

3.3. *Causal assumptions*

As discussed within the main manuscript, our main objective was to assess whether studies acknowledge and discuss the core causal assumptions. To do so, we developed a classification rubric (Tiers 1-4) based on three components: (1) Statement of the assumption, (2) Justification, and (3) Interrogation. Tier 1 is the aspirational benchmark and represents the highest level of comprehensiveness and explicitness and must cover all components, while the lower tiers reflect fewer components and thus lower levels of comprehensiveness and explicitness. **Supplementary Table 13** below provides a detailed explanation of each of these components. The rubric was piloted independently by the authors on 6 papers, with strong consensus reached across all components for each study.

Supplementary Table 13. Causal assumption reporting rubric

Classification	**Exchangeability**	**Positivity**	**Consistency**	**No-Interference**
Tier 1	3 components must be mentioned	2 components must be mentioned	3 components must be mentioned, and interrogation must include sub-group analysis	3 components must be mentioned
Tier 2	2 components must be mentioned	1 component must be mentioned	2 components must be mentioned	2 components must be mentioned
Tier 3	1 component must be mentioned	Implicit within the study design[1]	1 component must be mentioned	1 component must be mentioned
Tier 4	No mention	No mention	No mention	No mention
Components[2]	**1. Statement**: The authors must either; (a) mention the assumption explicitly or (b) describe the assumption (either through text or mathematical notation that describes the assumption in relation to the Potential Outcomes framework[3]). This component is standard across all the assumptions. For more information on the concept behind each assumption, see **Supporting Information 5**.			

Continued on next page

Classification	Exchangeability	Positivity	Consistency	No-Interference
	2. Justification: The authors must provide a rationale for how their study design addresses or explains this assumption. For *exchangeability*, the authors must explain the treatment-assignment mechanism (for protected area designation) or, at minimum, justifying why the selected covariates are believed to be affecting treatment and outcome variables (i.e. confounding). Merely citing other literature without additional explanation does not meet this criterion.	**2. Justification**: The authors must provide a rationale for how their study design addresses or explains this assumption. For *positivity*, an example of the authors addressing the assumption could be through intentionally excluding a select group of treated units within their study design	**2. Justification**: The authors must provide a rationale for how their study design addresses or explains this assumption. For *consistency*, the authors must explain why heterogeneity within the treatment (protected area designation) is not an issue within their study design or how it has been addressed. The authors must, at the very least, attempt to discuss the implications of what this assumption being violated might do to their inferences.	**2. Justification**: The authors must provide a rationale for how their study design addresses or explains this assumption. For *no-interference*, the authors must explain or at least acknowledge the implications of spillover effects occurring within their study. The creation of buffer zones around protected areas are considered are considered as an acknowledgement of this assumption and thus would fulfil this component.

Continued on next page

Classification	Exchangeability	Positivity	Consistency	No-Interference
	3. Interrogation: The authors must either; (a) assess the implications of the exchangeability assumption being potentially violated (e.g., through sensitivity testing for hidden confounding); or (b) assess whether there is any evidence of its violation (e.g., through placebo tests). Other approaches such as the use of confounder-robust estimators also satisfy this component.	**3. Interrogation**: The positivity assumption is an inherently untestable assumption in causal inference, and thus this component is not considered.	**3. Interrogation**: The authors must incorporate heterogeneity within the treatment in their study design. For this component, there are two additional distinctions to be made. First, studies which attempted to group based on IUCN protected area categories are capped at Tier 2. Tier 1 is reserved for studies which have attempted to incorporate sub-group analysis other than IUCN strictness classification. In simpler terms, a national park can have multiple protected area designations within its boundaries. However, even within those designations, management interventions can differ which can influence any inferences made.	**3. Interrogation**: The authors must assess whether spillover effects are occurring within their study design. The authors can use approaches such as interference-robust estimators, or conduct additional analyses within the buffer areas to test for evidence of violation.

[1] Certain study designs, such as statistical matching, implicitly have the assumption embedded (The matching process itself, down-weighs units with low probability of receiving treatment). Certain design choices such as setting calipers, whilst primarily done to improve the exchangeability assumption, also inherently improves positivity, as the tolerances can be seen as a form of intentionally excluding treated units. For this rubric, we only consider calipers to satisfy justification if they also refer to the positivity assumption, else they are designated to Tier 3.

[2] We give the authors the benefit of the doubt. For example, if the authors describe the rationale of statistical matching but then use covariates which might be mediators rather than confounders, they would not be penalized within this rubric as we are ultimately searching for explicitness and comprehensiveness, not validity.

[3] Structural Causal Model (SCM) Framework was rarely used within our identified studies. The select few that did use this framework were not penalized within our grading rubric and were assessed with reference to Correia et al. (2026)

3.4. *Causal reasoning, open science practices and research integrity measures*

Finally, we assessed the reporting of causal reasoning, open science practices, and research integrity measures within the field. Specifically, we recorded whether studies: (i) presented an explicit theory of change or directed acyclic graph; (ii) specified a causal estimand; (iii) were formally pre-registered or referenced a pre-analysis plan; (iv) made their code publicly available; (v) archived the data used in their analyses for replication purposes; (vi) declared any potential conflicts of interest; and (vii) had undergone an ethics review.

4. LIMITATIONS

This review is subject to some limitations. First, we focus solely on English-language studies as we lack the resources to conduct our review in multiple languages (Amano et al. 2021; Chowdhury et al. 2022). Second, although pre-registration of the systematic review protocol was originally intended, this was not feasible given the constraints of

the corresponding author's funding timeline. Finally, as data extraction was conducted solely by the corresponding author, the possibility of some human error cannot be entirely excluded.

REFERENCES

Amano, T., Berdejo-Espinola, V., Christie, A. P., Willott, K., Akasaka, M., Báldi, A., Berthinussen, A., Bertolino, S., Bladon, A. J., Chen, M., Choi, C.-Y., Kharrat, M. B. D., Oliveira, L. G. d., Farhat, P., Golivets, M., Aranzamendi, N. H., Jantke, K., Kajzer-Bonk, J., Aytekin, M. K., Khorozyan, I., Kito, K., Konno, K., Lin, D.-L., Littlewood, N., Liu, Y., Liu, Y., Loretto, M.-C., Marconi, V., Martin, P. A., Morgan, W. H., Narváez-Gómez, J. P., Negret, P. J., Nourani, E., Quintero, J. M. O., Ockendon, N., Oh, R. R. Y., Petrovan, S. O., Piovezan-Borges, A. C., Pollet, I. L., Ramos, D. L., Segovia, A. L. R., Rivera-Villanueva, A. N., Rocha, R., Rouyer, M.-M., Sainsbury, K. A., Schuster, R., Schwab, D., Şekercioğlu, H., Seo, H.-M., Shackelford, G., Shinoda, Y., Smith, R. K., Tao, S.-d., Tsai, M.-s., Tyler, E. H. M., Vajna, F., Valdebenito, J. O., Vozykova, S., Waryszak, P., Zamora-Gutierrez, V., Zenni, R. D., Zhou, W., and Sutherland, W. J. (2021). Tapping into non-English-language science for the conservation of global biodiversity. *PLOS Biology*, 19(10):e3001296. https://doi.org/10.1371/journal.pbio.3001296.

Boetje, J. and van de Schoot, R. (2024). The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. *Systematic Reviews*, 13(1):81. https://doi.org/10.1186/s13643-024-02502-7.

Callaghan, M., Müller-Hansen, F., Bond, M., Hamel, C., Devane, D., Kusa, W., O'Mara-Eves, A., Spijker, R., Stevenson, M., Stansfield, C., Thomas, J., and Minx, J. C. (2024). Computer-assisted screening in systematic evidence synthesis requires robust and well-evaluated stopping criteria. *Systematic Reviews*, 13(1):284. https://doi.org/10.1186/s13643-024-02699-7.

Caruana, A., Bull, J. W., Ferraro, P. J., Wauchope, H. S., Christie, A. P., and Jones, J. P. G. (2026). Still money for nothing? Two decades of empirical evaluation of conservation investments V1.0 [Dataset]. https://doi.org/10.5281/zenodo.20439985.

Chowdhury, S., Gonzalez, K., Aytekin, M. K., Baek, S.-Y., Bełcik, M., Bertolino, S., Duijns, S., Han, Y., Jantke, K., Katayose, R., Lin, M.-M., Nourani, E., Ramos, D. L., Rouyer, M.-M., Sidemo-Holm, W., Vozykova, S., Zamora-Gutierrez, V., and Amano, T. (2022). Growth of non-English-language literature on biodiversity conservation. *Conservation Biology*, 36(4):e13883. https://doi.org/10.1111/cobi.13883.

Correia, H. E., Dee, L. E., Byrnes, J. E. K., Fieberg, J. R., Fortin, M.-J., Glymour, C., Runge, J., Shipley, B., Shpitser, I., Siegel, K. J., Sugihara, G., von Holle, B., and Ferraro, P. J. (2026). Best practices for moving from correlation to causation in ecological research. *Nature Communications*, 17(1):1981. https://doi.org/10.1038/s41467-026-69878-z.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale university press.

Ferraro, P. J. and Hanauer, M. M. (2014). Advances in Measuring the Environmental and Social Impacts of Environmental Programs. *Annual Review of Environment and Resources*, 39(Volume 39, 2014):495–517. https://doi.org/10.1146/annurev-environ-101813-013230.

Ferraro, P. J. and Pattanayak, S. K. (2006). Money for Nothing? A Call for Empirical Evaluation of Biodiversity Conservation Investments. *PLOS Biology*, 4(4):e105. https://doi.org/10.1371/journal.pbio.0040105.

Foo, Y. Z., O'Dea, R. E., Koricheva, J., Nakagawa, S., and Lagisz, M. (2021). A practical guide to question formation, systematic searching and study screening for literature reviews in ecology and evolution. *Methods in Ecology and Evolution*, 12(9):1705–1720. https://doi.org/10.1111/2041-210X.13654.

Forbes, C., Greenwood, H., Carter, M., and Clark, J. (2024). Automation of duplicate record detection for systematic reviews: Deduplicator. *Systematic Reviews*, 13(1):206. https://doi.org/10.1186/s13643-024-02619-9.

Haddaway, N. R., Bethel, A., Dicks, L. V., Koricheva, J., Macura, B., Petrokofsky, G., Pullin, A. S., Savilaakso, S., and Stewart, G. B. (2020). Eight problems with literature reviews and how to fix them. *Nature Ecology & Evolution*, 4(12):1582–1589. https://doi.org/10.1038/s41559-020-01295-x.

Haddaway, N. R., Grainger, M. J., and Gray, C. T. (2022). Citationchaser: A tool for transparent and efficient forward and backward citation chasing in systematic searching. *Research Synthesis Methods*, 13(4):533–545. https://doi.org/10.1002/jrsm.1563.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Konno, K., Livoreil, B., and Pullin, A. S. (2021). Collaboration for Environmental Evidence Critical Appraisal Tool version 0.3. https://environmentalevidence.org/cee-critical-appraisal-tool/.

Langhammer, P. F., Bull, J. W., Bicknell, J. E., Oakley, J. L., Brown, M. H., Bruford, M. W., Butchart, S. H. M., Carr, J. A., Church, D., Cooney, R., Cutajar, S., Foden, W., Foster, M. N., Gascon, C., Geldmann, J., Genovesi, P., Hoffmann, M., Howard-McCombe, J., Lewis, T., Macfarlane, N. B. W., Melvin, Z. E., Merizalde, R. S., Morehouse, M. G., Pagad, S., Polidoro, B., Sechrest, W., Segelbacher, G., Smith, K. G., Steadman, J., Strongin, K., Williams, J., Woodley, S., and Brooks, T. M. (2024). The positive impact of conservation action. *Science*, 384(6694):453–458. https://doi.org/10.1126/science.adj6598.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, Cambridge, 2 edition.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Quan, Y., Tytko, T., and Hui, B. (2024). Utilizing ASReview in screening primary studies for meta-research in SLA: A step-by-step tutorial. *Research Methods in Applied Linguistics*, 3(1):100101. https://doi.org/10.1016/j.rmal.2024.100101.

UNEP-WCMC and IUCN (2026). Protected Planet: The World Database on Protected Areas (WDPA) and World Database on Other Effective Area-based Conservation Measures (WD-OECM). https://www.protectedplanet.net/en/about.

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., and Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133. https://doi.org/10.1038/s42256-020-00287-7.

SUPPLEMENTARY INFORMATION 5 – ADDITIONAL RESULTS

Supplementary Table 14. The top 10 most studied countries

Ranking	Country	No. of PA Impact evaluations[1]
1	China	56
2	Brazil	31
3	United States of America	17
4	Indonesia	15
5	Costa Rica	12
6	Spain	12
7	Madagascar	11
8	Peru	10
9	Australia	10
10	Australia	10

[1]Global-scale studies are excluded from this count.

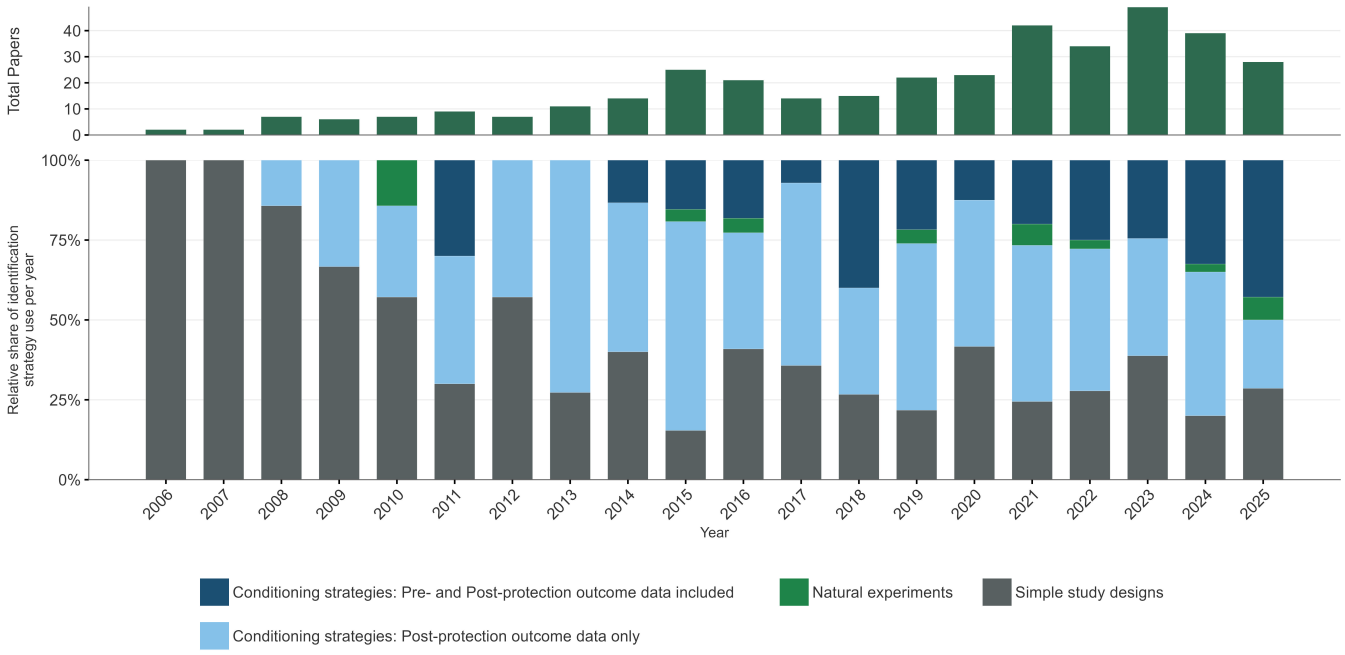
Supplementary Table 15. The top 10 most studied protected areas[1]

Ranking	Protected area name	Country	WDPA ID	No. of PA Impact evaluations
1	Wolong Nature Reserve	China	3015	8
2	Sichuan Giant Panda Sanctuaries - Wolong, Mt Siguniang and Jiajin Mountains	China	902902	7
3	Xishuangbanna	China	61403	5
4	Mount Wuyi	China	198295	5
5	Baishuijiang	China	220258	5
6	Sian Ka'an	Mexico	1850	4
7	Virunga National Park	Democratic Republic of the Congo	2017	4
8	Wuyishan	China	12469	4
9	Masoala	Madagascar	303695	4
10	Calakmul	Mexico	306780	4

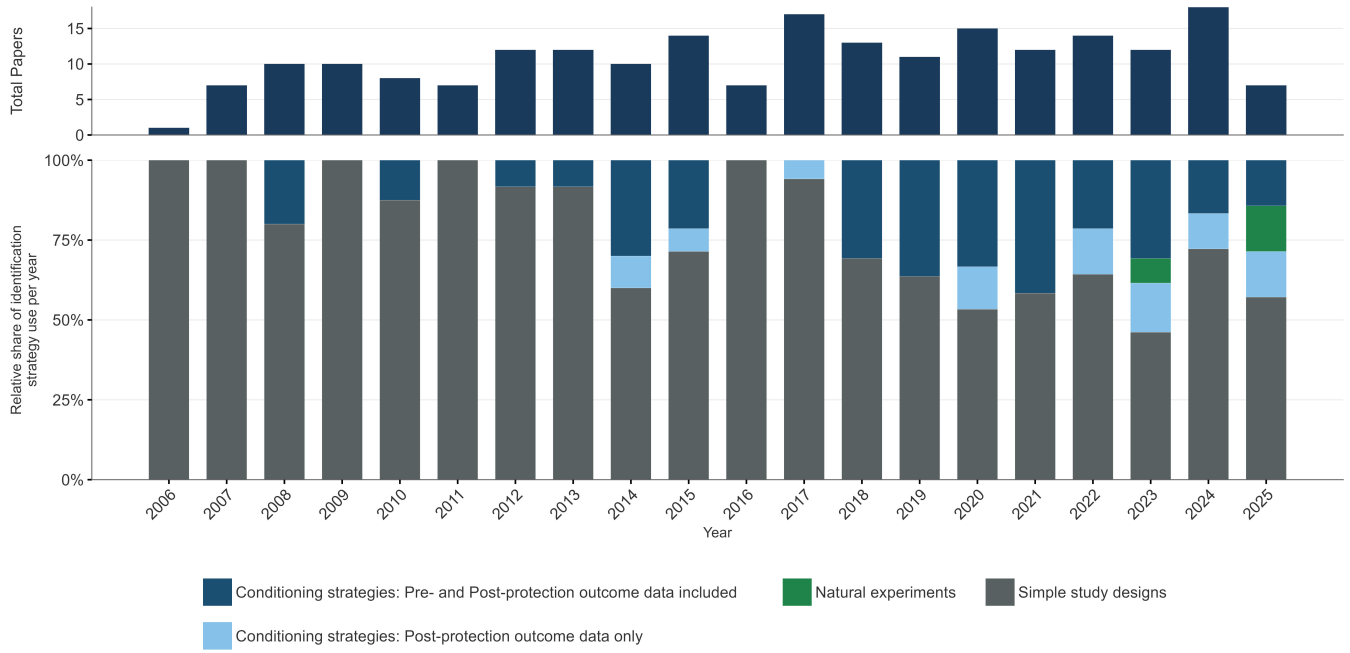
[1]The protected areas listed above are only those that were identifiable within the World Database on Protected Areas.

Supplementary Table 16. Distribution of study design scale

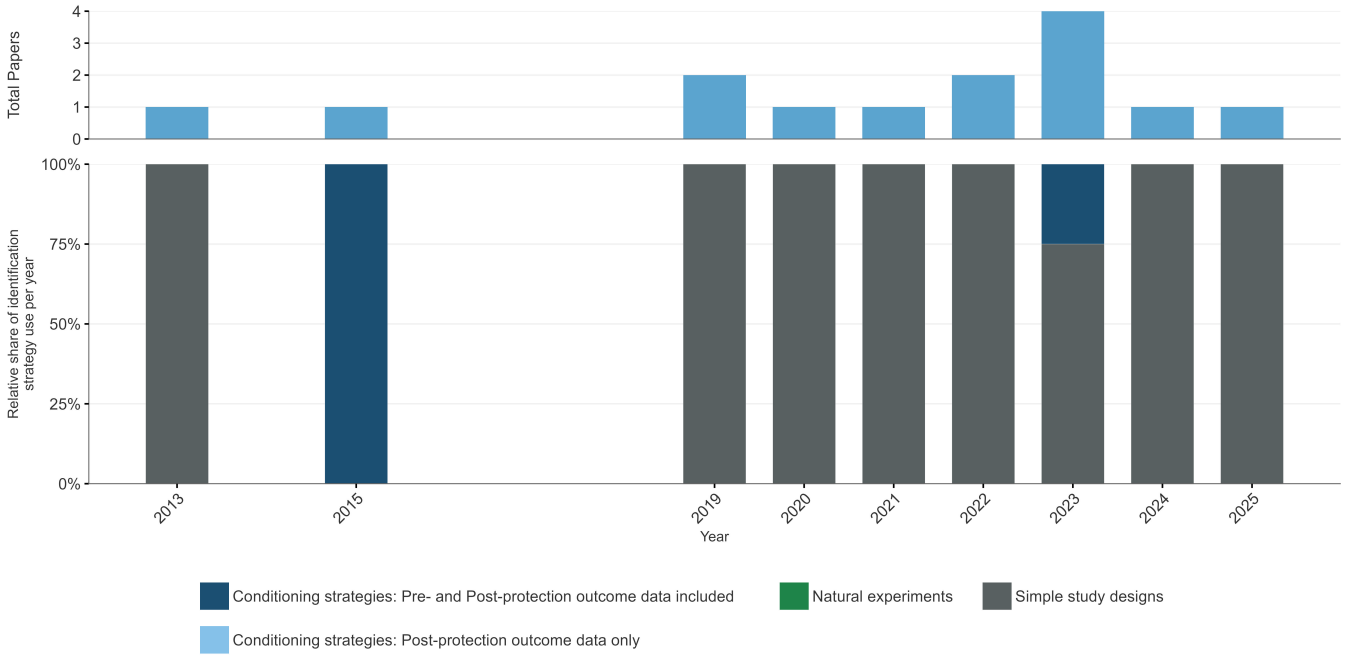
Scale	Description	No. of PA Impact evaluations	Percentage
Global	The study covers the entire globe.	31	10.1%
Multinational	The study focuses on multiple protected areas spanning over multiple countries, however not on a global scale. For example, studies focusing on either Europe or Africa fit this category. Studies focusing on distinct natural features such as the Caucasus Mountain ranges also fit in this category.	31	10.1%
National	The study focuses on protected areas across the entire country.	92	29.9%
Regional	The study focuses on multiple protected areas within a specific region of a country.	115	37.3%
Individual	The study is focused on one individual protected area.	39	12.7%



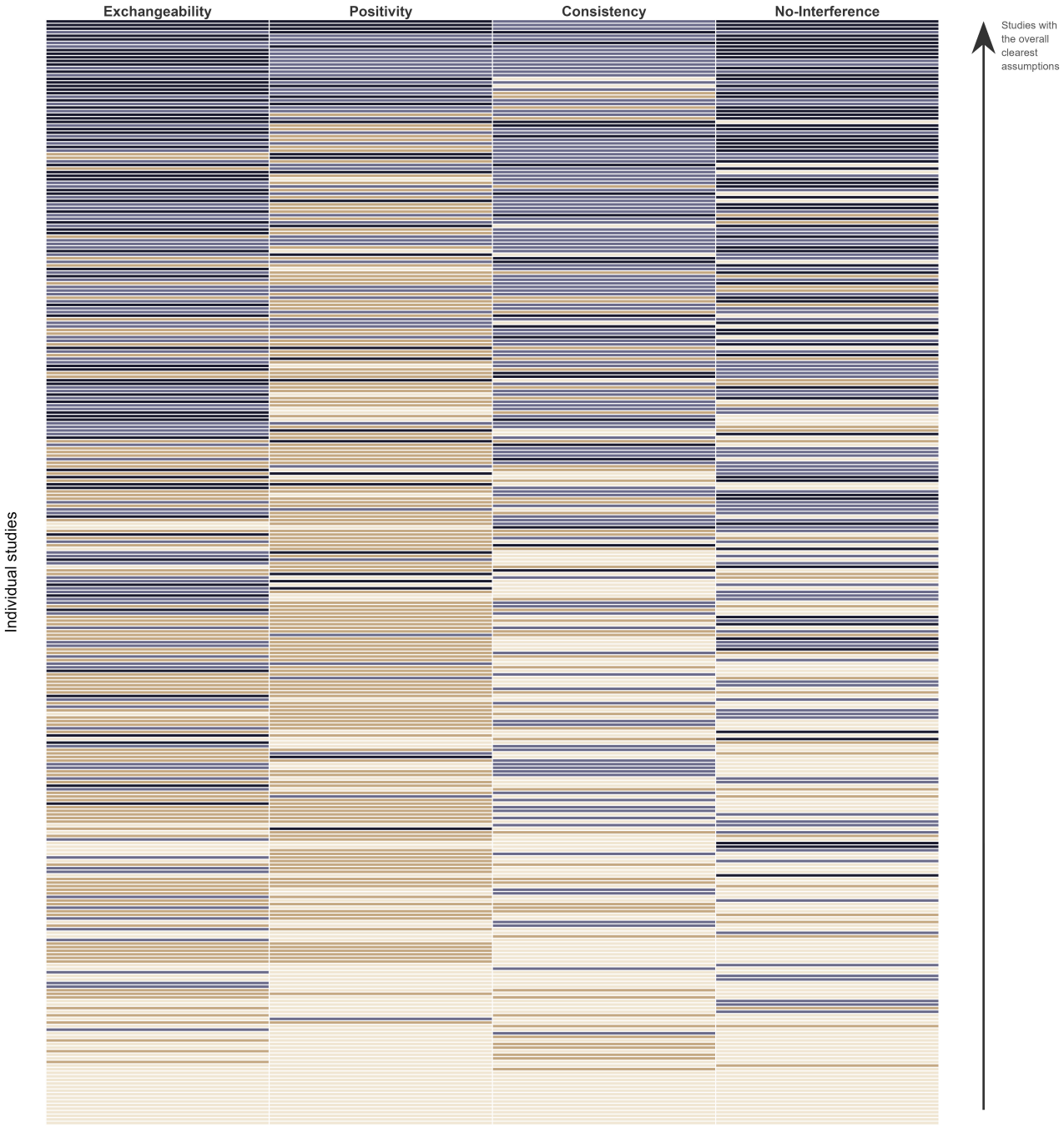
Supplementary Figure 7. The temporal trends of study design usage over time from 2006 to 2026, categorized based on the identification strategy used for terrestrial protected areas. In total, 128 studies (33.6%) used simple study designs, while 253 (66.4%) used a causal identification strategy. The majority used conditioning strategies, either with pre- and post-protection outcome data (33.2%) or post-protection outcome data only (67.2%), while natural experiments were rare (4.3%). Note that some studies used more than one causal identification strategy



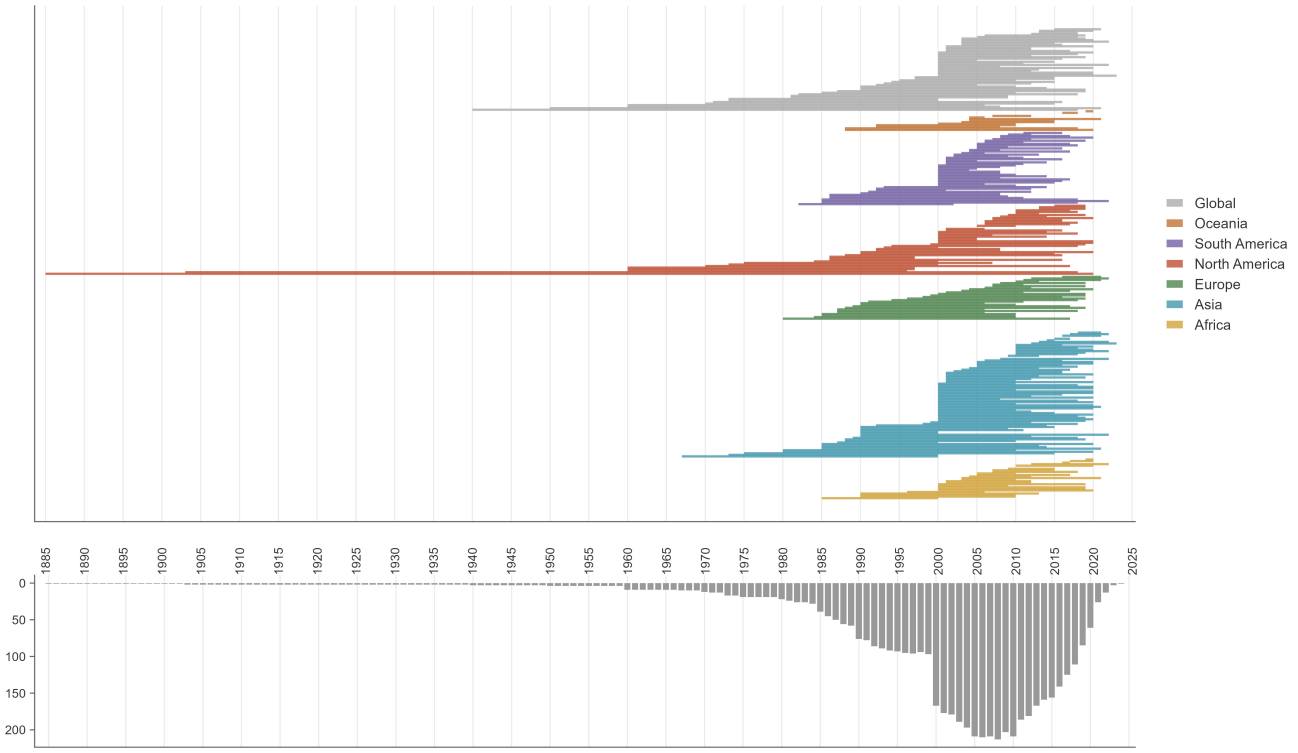
Supplementary Figure 8. The temporal trends of study design usage over time from 2006 to 2026, categorized based on the identification strategy used for marine protected areas. In total, 165 studies (76.7%) used simple study designs, while 53 (24.3%) used a causal identification strategy. The majority used conditioning strategies, either with pre- and post-protection outcome data (75.5%) or post-protection outcome data only (22.6%), while natural experiments were rare (3.8%). Note that some studies used more than one causal identification strategy.



Supplementary Figure 9. The temporal trends of study design usage over time from 2006 to 2026, categorized based on the identification strategy used for freshwater protected areas. In total, 12 studies (85.7%) used simple study designs, while 2 (14.3%) used a causal identification strategy. Both studies used conditioning strategies with pre- and post-protection outcome data. No natural experiments were documented.



Supplementary Figure 10. A study-level disaggregation of these tier classifications presented in **Figure 6** of the **Main Document**. Studies with the highest-level of clarity are at the top and those with the least amount of clarity at the bottom.



Supplementary Figure 11. The temporal coverage of all the studies taken to full-data extraction within our review, with each row representing an individual study categorized based on which continent the study was focusing on.

Supplementary Table 17. The top 5 outcome variables assessed within terrestrial studies per subgroup.

Broad Group	Sub-Group	Outcome variable	Count
Ecological outcomes	Biodiversity metrics	Species abundance & biomass	15
		Species diversity	7
		Biodiversity intactness index	1
		Elephant poaching/deaths	1
		IUCN red list threat to mammals	1
	Deforestation	Deforestation	74
		Forest/tree cover loss	34
	Ecosystem functioning & services	Carbon stocks	9
		Net primary productivity (NPP)	7
		Carbon emissions	4
		Water yield/retention	3
		Soil erosion and retain sediment	2
	Other habitat and landscape structure	Fire	8
		Habitat quality	8

Continued on next page

Broad Group	Sub-Group	Outcome variable	Count
		Normalized difference vegetation index (NDVI)	7
		Forest fragmentation	3
		Forest regrowth	2
Socioeconomic outcomes	Governance, Employment & conflict	Labour market/force	3
		Employment	2
		Violent acts	1
		Comprehensiveness development index	1
		Conflict	1
	Health & Nutrition	Acute respatory infection	1
		Depression	1
		Diarrhoea	1
		Disability	1
		Fertility	1
	Income, poverty and inequality	Poverty measures	21
		Gross Domestic Product (GDP)	3
		GINI Coefficient	3
		Tourism metrics	3
		Household income	2
	Livelihoods & wellbeing	Socioeconomic metrics	2
		Household development index	1
		Household welfare indicators	1
		Households' risk perception	1
		Human development index	1
			1
Human Pressures	Human footprint	Impervious surfaces	4
		Human activity	2
		Human footprint	2
		Anthropogenic pressure	1
		Human demographics	1
	Land use change	Land use/cover change	32
		Avoided natural-cover loss	1
		Cropland expansion	1
		Production-living-ecological space (PLES)	1
	Resource extraction	Coca crops and gold mining	1
		Fuel source	1
		Grain production	1
		Hectares of utilized agricultural area (UAA)	1
		Households' natural resource exploitation	1

Supplementary Table 18. The top 5 outcome variables assessed within marine studies per subgroup.

Broad Group	Sub-Group	Outcome variable	Count
Ecological outcomes	Biodiversity metrics	Species abundance & biomass	34
		Catch per unit effort (CPUE)	10
		Species diversity	7
		Trophic levels	1
	Ecosystem functioning & services	Carbon emissions	1
	Other habitat and landscape structure	Coral cover	2
		Algal cover	1
		Habitat cover	1
		Mangrove cover	1
		Perceived coral reef status	1
Socioeconomic outcomes	Governance, Employment & conflict	Employment	1
		Marine Protected Area Management Status	1
		Sub-industry shares	1
		Village infrastructure index	1
	Health & Nutrition	Diet diversity	1
		Food security	1
		Meat consumption	1
		Mortality rates	1
		Satisfaction with food from the sea	1
	Income, poverty and inequality	Income	3
		Annual net revenue	1
		DHS Wealth Index	1
		Economic vulnerability	1
		Gross Domestic Product (GDP)	1
	Livelihoods & wellbeing	Diversification (income activities)	1
		Non-dependence on fisheries for income	1
		Perceived effects of the sustainability	1
		Perceptions about benefits from management of customary fishing grounds (iqoliqoli)	1
			1
Human Pressures	Resource extraction	Fishing production/practices	2
		Number of coastal fishing boats	1
		Diversification (food production)	1
		Expected catch rate	1
		Recreational fishing license sales	