RESEARCH · RESEARCH DESIGN

Observational vs. Experimental Research: Divergent Paths to Knowledge Generation

How observational and experimental research differ in the way they generate knowledge, what each can and cannot establish, the history of conflicts between them (including hormone therapy and advertising), when observational evidence is credible, and how the two approaches work together.

ModQu research article · 2,630 words · 25 references (APA 7th edition)

There are two basic ways to learn how the world works. One is to watch it: record what people, organisations or nature do without interfering, and look for patterns. The other is to intervene: change something deliberately and see what follows. **Observational research** takes the first path and **experimental research** the second. The distinction is one of the most important in research methodology, because it determines what kind of conclusion a study can support. Experiments are often called the gold standard for causal questions, yet some of the most important causal knowledge, such as the link between smoking and lung cancer, came from observation. This article explains how the two paths differ, why they sometimes reach different conclusions, when observational evidence can be trusted, and how the two complement each other.

The defining difference

The difference between the two is who decides which units receive the treatment or exposure. In an **experiment**, the researcher decides, ideally by random assignment. In an **observational study**, the units, or circumstances beyond the researcher's control, decide. People choose whether to smoke, to join a loyalty programme or to take a medication; firms choose whether to adopt a technology; regions adopt policies for their own reasons. Cochran (1965) defined an observational study as one in which the objective is to elucidate cause-and-effect relationships but it is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover. Rosenbaum (2002) adopted a similar definition.

The consequence follows from the logic of causal inference. Rubin (1974) and Holland (1986) framed a causal effect as the difference between the outcome a unit would have under one treatment and the outcome it would have under another. Only one outcome can ever be observed for each unit, so causal inference always involves comparing different units. Random assignment ensures that, on average, the units in each group are alike in every respect, measured or unmeasured, apart from the treatment (Fisher, 1935). When units select their own treatment, they may differ systematically in ways that also affect the outcome. This is **confounding**, and it is the central problem of observational research.

Two paths to causal knowledge

The defining difference is who decides who is exposed

	Observational	Experimental
Who assigns exposure	Units themselves or circumstances	The researcher, ideally at random
Main threat	Confounding by how units were selected	Limited realism, scale or duration
What can be studied	Anything recorded, incl. unassignable exposures	Only what can ethically be assigned
Scale and time span	Often large populations over years	Usually smaller and shorter
Best for	Description, rare harms, hypotheses	Effects of interventions

Based on Cochran (1965); Shadish et al. (2002); Rosenbaum (2002) ModQu.com

Figure 1. Observational and experimental research compared. The defining difference is who decides which units are exposed. Based on Cochran (1965), Shadish et al. (2002) and Rosenbaum (2002); table by ModQu.

What each path does well

Experiments are the strongest design for estimating the effect of an intervention. Random assignment removes confounding by design, and the researcher's control over the treatment means that its timing and content are known precisely. Shadish et al. (2002) describe experiments as having the highest **internal validity** of any design.

Observational research has its own strengths. It can study exposures that cannot be assigned, such as age, income, a disease, a natural disaster or a recession. It can study exposures that would be unethical to assign, such as smoking or neglect. It can study rare outcomes and long time spans, using existing records of millions of people over decades. It often observes behaviour in its natural setting and at full scale, which may make its results more applicable to real conditions. And it is usually cheaper and faster.

Some observational studies have produced causal knowledge as secure as any experiment. Snow (1855) traced cholera deaths in London to a contaminated water supply by comparing households served by different companies. Doll and Hill (1950) compared the smoking histories of lung cancer patients with those of other hospital patients and found a strong association, later confirmed in their long-term study of British doctors. No randomised trial of smoking was ever conducted, nor could it ethically be, yet the causal link is not in doubt.

Types of observational study

Observational research takes several forms, which differ in how units are selected and when they are measured.

- **Cohort studies** identify people by their exposure, such as smokers and non-smokers, and follow them over time to see who develops the outcome. Because exposure is measured before the outcome, the time order is clear.

- **Case-control studies** start from the outcome. They identify people who have a condition (cases) and people who do not (controls) and compare their past exposures. Doll and Hill's (1950) first smoking study used this design. It is efficient for rare outcomes but vulnerable to errors in recalling past exposures.
- **Cross-sectional studies** measure exposure and outcome at the same time, which makes it hard to establish which came first.
- **Ecological studies** compare groups, such as regions or countries, rather than individuals, and risk attributing group-level patterns to individuals.
- **Analyses of routine data**, such as transaction records, health records or website logs, are observational studies of the largest scale, often covering whole populations. Their size reduces random error but does nothing to reduce confounding.

Rothman et al. (2008) describe how the choice among these designs affects which biases are most likely, and how each can be strengthened. Cochran (1965) urged that observational studies be planned with the same care as experiments, with a clear statement of the comparison to be made, the measurements needed to address the main alternative explanations, and the population to which results will apply.

When the two paths disagree

The risk of relying on observational evidence is best illustrated by cases where it was later contradicted by experiments.

Hormone therapy and heart disease

During the 1980s and 1990s, many observational studies found that postmenopausal women who took hormone therapy had lower rates of coronary heart disease than women who did not. Stampfer and Colditz (1991) pooled the evidence and estimated that users had roughly half the risk of non-users. The finding was biologically plausible and widely accepted, and hormone therapy was often recommended partly for heart protection.

The Women's Health Initiative then randomly assigned more than 16,000 women to hormone therapy or placebo. The trial was stopped early in 2002 when it found that hormone therapy slightly increased the risk of coronary heart disease, along with other risks (Writing Group for the Women's Health Initiative Investigators, 2002).

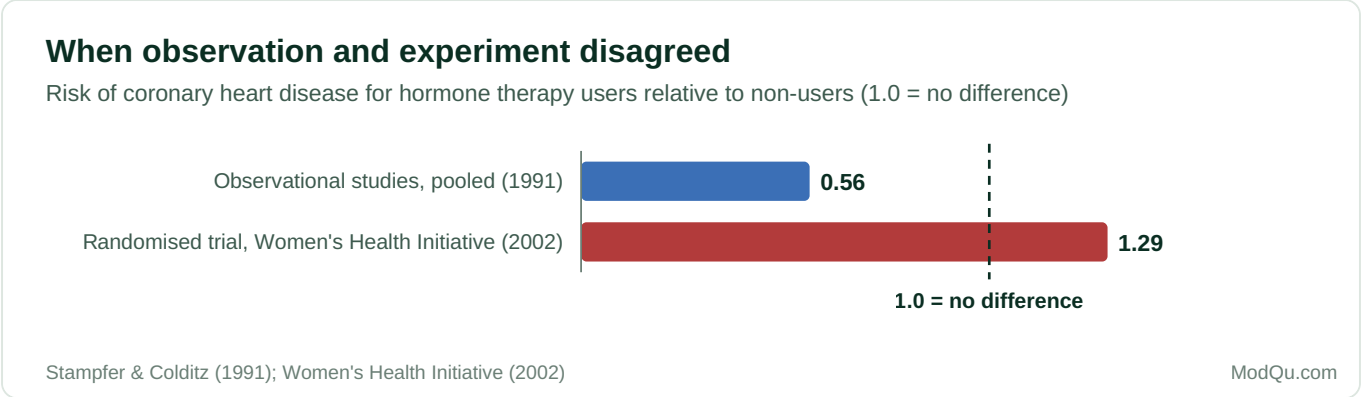


Figure 2. Hormone therapy and coronary heart disease: pooled observational estimate (relative risk) versus the randomised trial estimate (hazard ratio). Values below 1.0 indicate lower risk among users; above 1.0, higher risk. Sources: Stampfer and Colditz (1991); Writing Group for the Women's Health Initiative Investigators (2002).

The explanation was largely confounding: women who chose hormone therapy tended to be healthier, wealthier and more health-conscious than those who did not, and these characteristics, not the therapy, explained much of their lower risk. Later reanalysis showed that part of the discrepancy also came from how the observational studies were analysed. Hernán et al. (2008) reanalysed data from the Nurses' Health Study, one of the main observational studies, designing the analysis to mimic the structure of a trial: comparing women who started therapy with women who did not, from the same starting point. The reanalysed estimates were much closer to the trial's results. Many earlier observational analyses had compared long-term users, who had survived the early period of therapy when risks were highest, with non-users.

Advertising

A similar pattern has appeared in commercial research. Gordon et al. (2019) examined 15 large advertising campaigns on Facebook in which users had been randomly assigned to see or not see the advertisements. They then applied a range of common observational methods to the same data, comparing users who saw the advertisements with those who did not and adjusting for many characteristics. The observational methods often overestimated the effect of the advertising, sometimes by a large factor, because people who see an advertisement differ from those who do not in ways that affect whether they buy. LaLonde (1986) reached a similar conclusion for job training programmes: many observational estimates differed substantially from the experimental benchmark.

How often do they disagree?

The record is mixed. Concato et al. (2000) and Benson and Hartz (2000), comparing observational studies and randomised trials of the same treatments, found that well-designed observational studies usually produced results similar to those of trials, and argued that the assumed hierarchy, with trials always superior, was too simple. Ioannidis (2005), examining highly cited clinical research, found that initial claims were contradicted or found to be exaggerated by later studies more often when the original studies were observational than when they were randomised trials. Vandenbroucke (2004) suggested a useful distinction: observational studies are generally reliable for detecting unintended effects, such

as rare side effects of a drug, because the reasons people take a treatment are usually unrelated to such effects. They are less reliable for intended effects, because the same factors that lead people to choose a treatment often affect the outcomes it is meant to improve.

When observational evidence is credible

Several features make observational evidence more trustworthy.

A strong and consistent association. Hill (1965) proposed a set of considerations for judging whether an association is likely to be causal, including its strength, consistency across studies and populations, specificity, temporal order, a dose-response gradient, plausibility, coherence with other knowledge and analogy. Hill did not intend these as a checklist that proves causation, but they remain a useful framework. The smoking-lung cancer association satisfied almost all of them.

A design that approximates an experiment. Natural experiments, regression discontinuities and other quasi-experimental designs exploit circumstances in which exposure was assigned in a way that is close to random (Angrist & Pischke, 2009). These designs often provide stronger evidence than simple comparisons adjusted for measured differences.

Careful adjustment for measured confounders. Matching and weighting methods, including those based on the **propensity score**, can balance treated and untreated groups on many measured characteristics (Rosenbaum & Rubin, 1983). Their limitation is that they cannot balance characteristics that were not measured.

Target trial emulation. Hernán and Robins (2016) proposed that observational analyses should be designed by first specifying the randomised trial that would ideally answer the question (the "target trial") and then emulating each element of its protocol, including eligibility, treatment strategies, the start of follow-up and outcomes, using observational data. This approach avoids several common errors, such as comparing current users with non-users, that produced misleading results in the hormone therapy studies.

Sensitivity analysis. Rosenbaum (2002) developed methods to ask how strong an unmeasured confounder would have to be to explain away an observed association. If only an implausibly strong confounder could do so, the causal conclusion is more credible.

An explicit causal model. Pearl and Mackenzie (2018) argue that stating the assumed causal relationships among variables, often as a diagram, makes clear which variables must be adjusted for and which must not, and exposes the assumptions on which a causal conclusion depends.

When experiments are not the answer

Experiments have limits of their own. Smith and Pell (2003) made the point with a satirical systematic review that found no randomised controlled trials of parachutes, concluding that people who insist on trial evidence for every intervention might volunteer for one. Some effects are obvious enough, and some interventions harmful enough, that experiments are unnecessary or unethical.

Even where experiments are possible, they may answer a narrower question than the decision requires. Deaton and Cartwright (2018) caution that a trial establishes an average effect in a specific population and setting, and that applying it elsewhere requires understanding why the effect occurred, which often depends on observational evidence and theory. Experiments may be small, short or conducted in unusual conditions, while observational data may cover whole populations over long periods.

Experiments can also be subject to their own biases. Participants who know they are in a study may behave differently, people may drop out unevenly from the treatment and control groups, and those assigned to a treatment may not take it. Shadish et al. (2002) note that such problems can reintroduce some of the selection that randomisation was meant to remove. A poorly conducted experiment is not automatically more reliable than a well-designed observational study, and the quality of the execution matters as much as the label attached to the design.

How the two work together

The most reliable knowledge usually comes from combining the two approaches.

- **Observation generates hypotheses; experiments test them.** Patterns in observational data suggest which interventions might work, and experiments then test the most promising.
- **Experiments calibrate observation.** Where both exist, comparing them shows how much bias observational methods are likely to carry in a particular field, as in the advertising studies.
- **Observation extends experiments.** Once an experiment has established an effect, observational data can show whether it holds in other populations, over longer periods and at scale, and can detect rare harms that trials are too small to reveal.
- **Observation answers questions experiments cannot.** For exposures that cannot be assigned, careful observational designs are the only option, and the history of smoking research shows how much they can achieve.

A worked example

A retailer wants to know whether its free delivery threshold, which offers free delivery on orders above $50, increases spending. Its data show that customers who regularly use free delivery spend 40% more per year than those who do not (illustrative figures).

An observational reading would credit free delivery with the difference. But customers who place large orders would qualify for free delivery anyway, so the comparison is circular. Customers who live far from stores may both value delivery more and spend more online. Loyal, high-spending customers may simply be more aware of the offer.

Several approaches could produce a better answer. An analyst could compare customers just above and just below the threshold to see whether orders bunch just above $50, which would suggest that the offer changes behaviour at the margin. The retailer could look for a natural experiment, such as a period when the threshold changed, and compare spending before and after, using a group of customers unaffected by the change as a comparison. Or it could run an experiment, randomly assigning customers to different thresholds for a period and comparing their total spending. Each

approach answers a slightly different question, and the experiment gives the cleanest answer to whether changing the threshold would change spending. The raw observational comparison, despite being based on millions of transactions, answers none of them.

Implications for organisations

Organisations hold vast amounts of observational data: transactions, website behaviour, customer records and employee data. These data are invaluable for describing what is happening and for generating hypotheses. But when the question is whether an action will change an outcome, such as whether a loyalty programme increases spending or whether a training course improves performance, observational comparisons are vulnerable to exactly the confounding that misled the hormone therapy studies. Customers who join a loyalty programme, employees who take a course and regions that receive a new initiative are rarely comparable to those who do not.

Where an experiment is feasible, such as a randomised test of a new offer, it usually provides the clearest answer. Where it is not, the observational analysis should be designed with the target trial in mind, adjust carefully for differences between groups, test how sensitive the conclusion is to unmeasured factors, and be reported with appropriate caution.

Conclusion

Observational and experimental research are two paths to causal knowledge. Experiments remove confounding by design and give the clearest answers to questions about interventions, within the conditions tested. Observational research can study what cannot be assigned, at scales and over periods that experiments cannot reach, but it must work hard to rule out confounding. Cases such as hormone therapy show how observational evidence can mislead when that work is not done; cases such as smoking show how much it can achieve when it is. The two paths are most powerful when used together, each checking and extending the other.

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, *342*(25), 1878–1886. https://doi.org/10.1056/NEJM200006223422506

Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society: Series A (General)*, *128*(2), 234–266. https://doi.org/10.2307/2344179

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, *342*(25), 1887–1892. https://doi.org/10.1056/NEJM200006223422507

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Doll, R., & Hill, A. B. (1950). Smoking and carcinoma of the lung: Preliminary report. *British Medical Journal*, *2*(4682), 739–748. https://doi.org/10.1136/bmj.2.4682.739

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, *38*(2), 193–225. https://doi.org/10.1287/mksc.2018.1135

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. E., & Robins, J. M. (2008). Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology, 19*(6), 766–779. https://doi.org/10.1097/EDE.0b013e3181875e61

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology, 183*(8), 758–764. https://doi.org/10.1093/aje/kwv254

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300. https://doi.org/10.1177/003591576505800503

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, *294*(2), 218–228. https://doi.org/10.1001/jama.294.2.218

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, *76*(4), 604–620.

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

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-3692-2

Rosenbaum, P. R., & 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

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3rd ed.). Lippincott Williams & Wilkins.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. https://doi.org/10.1037/h0037350

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Smith, G. C. S., & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ*, *327*(7429), 1459–1461. https://doi.org/10.1136/bmj.327.7429.1459

Snow, J. (1855). *On the mode of communication of cholera* (2nd ed.). John Churchill.

Stampfer, M. J., & Colditz, G. A. (1991). Estrogen replacement therapy and coronary heart disease: A quantitative assessment of the epidemiologic evidence. *Preventive Medicine*, *20*(1), 47–63. https://doi.org/10.1016/0091-7435(91)90006-P

Vandenbroucke, J. P. (2004). When are observational studies as credible as randomised trials? *The Lancet*, *363*(9422), 1728–1731. https://doi.org/10.1016/S0140-6736(04)16261-2

Writing Group for the Women's Health Initiative Investigators. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial. *JAMA*, *288*(3), 321–333. https://doi.org/10.1001/jama.288.3.321