Experimentation without randomization of individual patients

## Introduction

They say that randomized control trials are easy. Hand out drugs or placebos based on a coin flip, wait three years and count the number of dead bodies in each group. Well, maybe it is not that easy, but the methodology is well established.

Randomization of individual patients is considered the gold standard of research. Any time you have sufficient control to select who gets the treatment and who gets the control, you should always consider individual randomization when planning a study.

Randomization prevents covariate imbalance, both among measured and unmeasured covariates. Frail patients and hardy patients are allocated more or less evenly between the treatment and control group. Patients who are terrible about showing up at their monthly evaluations are allocated more or less evenly. Patients who brush their teeth obsessively three or four times a day are allocated more or less evenly.

It sounds good, but there are times when you should reject individual randomization in favor of some non-randomized alternatives. I’m a professional statistician and it may seem to you like a betrayal of everything I’ve learned. I’m sorry if you feel that way, but I do strongly believe that sometimes (but only sometimes) abandoning randomization is the best course of action.

Randomization is not all it’s cracked up to be. Randomization is expensive. You have to set up an elaborate logistical framework to implement randomization. Skip all that stuff and you can get a larger sample size for the same amount of research funding.

Randomization is unpopular. People who provide dental care don’t want to randomly switch back and forth between two different treatment protocols.

Randomization is unnatural. You are used to giving advice about oral health in an assured voice that lets your patients know that your recommendations are based on sound science. When you randomize, you lose this level of assurance and admit that you are leaving their treatment choice up to the flip of a coin.

Now there are some settings where randomization is just flat out impossible. In these settings, you use all the approaches developed for observational designs.

There is a middle ground, however, between randomized designs and observational designs. You have the ability to assign patients to receive a treatment or not, but you control the assignment yourself rather than leaving the assignment to a flip of a coin. If you do this carefully, you can produce evidence that can be just as persuasive as a randomized trial, but you avoid the many limitations associated with randomization.

## The infamous historical control

The simplest way you can avoid randomization is to give everyone who comes to you the new treatment and rummage around in some database for historical records from patients without the new treatment. This is often called a historical control design or a before and after design.

Most researchers sneer at historical control designs. Historical controls have indeed been behind some absolute research disasters.

Nevertheless, you will still see historical control designs in the literature. One study (Nishi 2023) compared the treatment with low-level laser therapy of 21 patients with Oral Mucositis from December 2022 to September 2023 to 96 control patients recruited from another study (Ohbayashi 2021) conducted from August 2009 to December 2019.

Can you compare 2022-2023 patients with those recruited 3 to 14 years earlier? Apparently, according to the authors. And with enough credibility to warrant a call for “development of comprehensive national guidelines, initiation of randomized comparative trials, and the establishment of collaborative research networks among international healthcare institutions.”

You might be a bit more cautious. The problem with historical controls is that they fail to account for temporal trends unrelated to your therapy. If you notice a change in the quality of care provided between 2019 and 2020, is that due to the intervention you’ve been testing or is it due to the many changes in care produced by the COVID pandemic? Did your intervention coincide with the introduction of fluoridation of your local water supply?

It could just be something as simple as your body’s tendency to heal itself over time. It was Voltaire who pointed out “The art of medicine consists of amusing the patient while nature cures the disease.”

## Quasi-experiments to the rescue

You can improve the credibility of studies where the treatment and control groups are separated by time. These are not perfect studies. There are no perfect studies. They can compete, however, on an even playing field with randomized studies. They can provide evidence that is almost as persuasive, as a randomized trial and don’t suffer from the well-known limitations of randomized trials.

Many of these approaches fall into a category of research known as quasi-experimental designs. There are varying definitions for quasi-experimental designs, but a common one is a research design where the researchers have experimental control, meaning having the ability to assign treatments and controls, but decline to use randomization to allocate patients to these groups.

The prefix “quasi” implies inferiority. In fact, a prominent researcher (Campbell 1988) calls them “queasy” experiments. This, however, is an unfair characterization.

In a quasi-experimental design, researchers deliberately decline to randomize because of the known problems with randomization. The researchers recognize abandoning randomization will produce a superior result.

## Interrupted time series

The interrupted time series is a common approach to add rigor without randomization. You need to separate a single “before” evaluation into multiple “before” evaluations and make a similar separation of a single “after” evaluation into multiple “after” evaluations.

Figure 1 shows to possible outcomes for an interrupted time series. Good news for your intervention would be a flat pattern during the multiple evaluations before the intervention, a sudden jump right at the intervention, and a flat pattern for all the evaluations after the intervention. In contrast an upward (or downward) trend at each and every evaluation is very strong negative evidence.

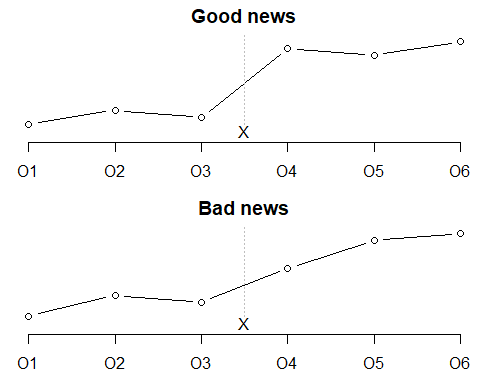


Figure 1. Two hypothetical outcomes from an interrupted time series design.

If you detect a slope or trend prior to the intervention, things get a bit more complicated. Look for a jump at the intervention but also look for a change in slope. Either a flattening of a negative trend or an acceleration of an already positive trend is good news for your intervention.

A study of poor compliance with clear aligner therapy (Timm 2022) monitored patients biweekly in 2019 prior to the introduction of a system of electronic reminders and feedback. Monitoring continued throughout all of 2020. The second half of 2019 was compared to the second half of 2020, to avoid problems with transitioning and with seasonality. The rate of poor compliance was flat during the second half of 2019, hovering around 25%. The data showed a downward trend in the second half of 2020, leveling off at a much better value of 9% in October 2020. The authors concluded that electronic reminders and feedback effectively reduced poor compliance rates.

## Phased intervetions

While a flat-jump-flat pattern for an interrupted time series is good news, you are still left with one concern. Did some other change, occurring at the same time as your intervention, interfere with your experiment? It could work both ways. It could produce a false positive result or it could mask something real.

You can rule out this concern if your intervention is one that can be split and implemented in different phases. If so, make multiple evaluations, implement one phase of the intervention, make additional evaluations, implement a second phase of the intervention, and so forth.

Figure 2. shows two possible results from an interrupted time series with three phases. A flat-jump-flat-jump-flat pattern is good news for your intervention. If the jumps at the intervention points are no bigger than the jumps at other times, that’s bad news for your intervention.

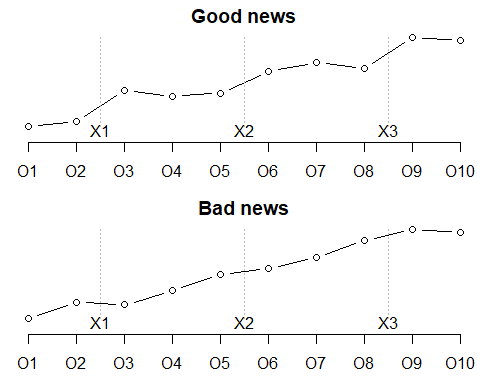


Figure 2. Two hypothetical outcomes for a for an interrupted time series design with phases.

If there is an adverse trend prior to the first intervention, examine whether that trend is reduced after each intervention point. You have to work harder if there are trends, but the work is more tedious than difficult.

It’s not easy to find an example of phased interventions in dental research, but there are good examples in other disciplines. A large teaching hospital implemented three changes to their electronic health record system to reduce inappropriate red blood cell (RBC) and platelet transfusion (Atia 2022). In November 2012, the system displayed a warning was shown when clinicians prescribed RBC for patients who had hemoglobin greater than 100 g/L. In May 2015, a lower threshold triggered the warning and the system displayed the most recent hemoglobin level and date. In May 2016, the threshold for hemoglobin was lowered again. The results were mixed, with statistical significant drops at some but not all of the intervention times. The results also varied by the type of clinic.

This will always be an issue with this type of design. A flat-jump-flat-jump-flat pattern is very persuasive positive evidence. A pattern with no jumps and no changes in trend is very persuasive negative evidence. But when you see a jump at only one of the three intervention points, it could be a negative finding or it could represent an intervention where only one of the three phases was effective.

## Withdrawal design

Can you withdraw a treatment after you apply it? That won’t work for training interventions. You can’t unlearn something. Other interventions, like a remodel in the physical layout of a dental clinic, require too much time and expense to be undone. But when you can withdraw a treatment, you have a very powerful way to demonstrate its effectiveness.

In a withdrawal design (sometimes called the ABA design) you start with a control (A), switch to a treatment (B) and then switch back to your control.

A very simple illustration of this is described in Philip Zimbardo’s book about his infamous prisoner experiments.

He described an experiment (not his infamous prison experiment) that used a withdrawal design. A researcher wanted to show how anonymity increases the tendency to engage in violent and aggressive actions.

He set up an experiment with school children at a Halloween party and found a simple and clever way to measure their aggressive and competitive behavior. This was assessed at the start of the party, before anyone had donned their costumes. Then the children were asked to put on their costumes. The masks gave the a fair degree of anonymity. Aggression was measured again and it rose.

But maybe the kids just got rowdier as time wore on. The researchers had an answer to this objection. They evaluated aggression a third time, after asking the children to remove their costumes. When average aggression levels returned to the level seen at first time point, they could rule out some other temporal trend, such as a change induced by the increasing amount of sugar and candy consumed during the party.

Figure 3 shows hypothetical results from a withdrawal design. A good result is an up-down trend: up when the intervention is added and back down when the intervention is removed, that tends to support your hypothesis. A bad result, one where the upward trend continues in spite of the withdrawal of the treatment, would tend to discredit your hypothesis.

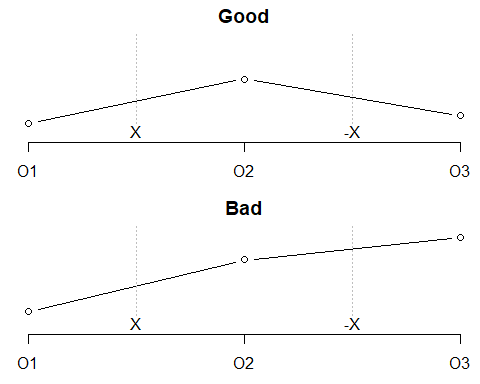


Figure 3. Two hypothetical results from a withdrawal design

Perhaps you can still raise an objection. What if children were like werewolves? They were calm while a full moon was hidden behind clouds, went wild while the full moon emerged from behind the clouds, and then calmed back down when the clouds covered the moon again.

Well, maybe, but most temporal trends are either a sudden jump or a continual upward or downward trend.

You can provide even more credibility with a one extra adaptation. Start with the control, add an intervention, withdraw it, and then add it back again. This is an ABAB design and makes it even harder to cite a “moon behind clouds” alternative explanation. It would be very strange that the clouds would synchronize with the donning and removal of Halloween masks.

A study of self-biting (Jones 1997) illustrates this ABAB design in a single patient. a 15 year old boy had serious behavioral issues including biting his lips hard enough to draw blood. The researchers wanted to test a relaxation therapy and measured lip bleeding before and after implementation of the new therapy. This by itself would be a simple historical control study. How do you know that the problem wasn’t just a problem that goes away over time? The researchers checked this by stopping the relaxation therapy. When the patient reverted to lip biting, the relaxation therapy was re-introduced. Self-biting stopped again, much to the delight of the patient and the researchers.

## Regression discontinuity

Sometimes you might have a qualifying variable, a potential confounding variable that is used to decide controls who gets the intervention.

You might reserve an intervention for patients based on their score on a measure of illness severity. You might offer free or subsidized care only to patients based on their income. You might admit trainees to a special program based on their score on a qualifying exam.

This seems like a terrible setting to conduct research. The worst patients get the intervention? Surely any effect of the intervention will be masked by this lopsided allocation.

The secret to a regression discontinuity design is that you compare patients on either side of the dividing line and ignore the patients at the extremes.

Of course, it makes no sense to compare the very sickest patients with only mildly ill patients, the very rich patients with the desperately poor patients, or the A+ students to the students who flunked. Any difference that you see at the extremes will be strongly influenced by the qualifying measure. But the influence of the qualifying measure is less strong when you restrict yourself to a narrow window on either side of your cut-off.

Figures 4 through 6 shows a hypothetical setting for a regression discontinuity design. A potential confounding variable shows a relationship to the outcome (Figure 4). Patients scoring below a threshold are assigned to the treatment (Figure 5). In this hypothetical dataset, the treatment lifts all patients by a small amount. Then only patients near the threshold are selected for the regression discontinuity study (Figure 6).

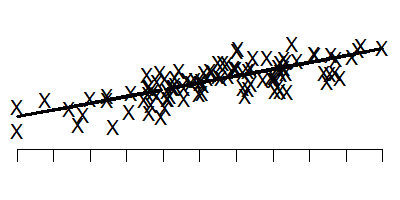


Figure 5. Allocation of patients to the intervention based on the qualifying variable.

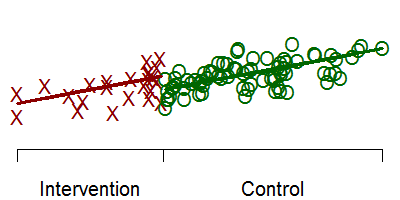


Figure 4. Hypothetical data showing a relationship between a potential confounding variable and a health outcome.

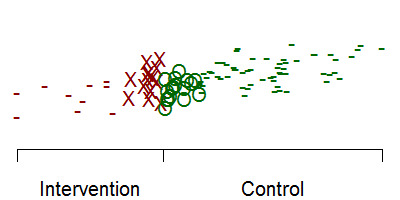


Figure 6. Removal of patients at the extremes.

A regression discontinuity design was used to examine the extent to which Medicaid dental coverage reduced difficulty in accessing dental care (Roberts 2023). The researchers selected patients that were 75 percentage points on either side of the state-specific Medicaid income eligibility threshold. The proportion of patients reporting difficulty accessing dental care was 5% higher in those patients with too much income to qualify for Medicaid coverage.

## How many evaluation points do you need?

For an interrupted time series or some variation of this design, there is no easy answer to how many evaluation points you might need. Certainly you can’t really assess a flat pattern without at least three data points. If you are running a complex model over time including estimation of correlations among the time measurements, a much larger sample (25 or 50) might be needed.

Concern about cyclic patterns, might also require a larger number of evaluation points. For monthly evaluations, a couple of years would be nice so you can factor out effects due to seasons of the year. For daily evaluations, measure for at least a couple of weeks so you can rule out any “Hump Day” or “Thank God It’s Friday” effects.

There is a relationship between the number of evaluation points and the number of patients studied. Which is better, studying a lot of patients with a few evaluation points or studying a few patients with a lot of evaluation points? Often it is six of one, half a dozen of the other as the saying goes. You might find that other aspects of the research, such as logistical constraints, could be better determinants of your choice here.

The interesting issue is when you increase the number of measurement times, you may be able to reduce to a single patient, the smallest sample size possible. This is an example of an “n of 1” design (Davidson 2020). The pattern observed in the single self-biting patient is quite revealing, where changes correspond precisely to the times when the intervention is added (or removed).

## Quote

“Of course, from the quasi-experimental perspective, just as from that of physical science methodology, it is obvious that moving out into the real world increases the number of plausible rival hypotheses. Experiments move to quasi-experiemtns and on into queasy experiments, all too easily.” Donald T. Campbell, in Methodology and Epistemology for Social Science: Selected Papers, page 322.

## Bibliography

Atia J, Evison F, Gallier S, Pettler S, Garrick M, Ball S, Lester W, Morton S, Coleman J, Pankhurst T. Effectiveness of clinical decision support in controlling inappropriate red blood cell and platelet transfusions, speciality specific responses and behavioural change. BMC Med Inform Decis Mak. 2022 Dec 29;22(1):342. doi: 10.1186/s12911-022-02045-8. PMID: 36581868; PMCID: PMC9798655.

Hiromi Nishi, Susumu Horikoshi, Tetsumi Yoshida, Noriyasu Fukushima, Kyoko Oshita, Syuichi Munenaga, Taro Edahiro, Hiroshi Ureshino, Hideo Shigeishi, Yukio Yoshioka, Masaru Konishi, Noriaki Ide, Yuma Ogawa, Rikou Marukawa, Tomoaki Shintani, Natumi Ino, Mikihito Kajiya, Naoya Kakimoto, Hiroki Ohge, Tatsuo Ichinohe, and Hiroyuki Kawaguchi. Efficacy of Low-Level Laser Therapy for Oral Mucositis in Hematologic Patients Undergoing Transplantation: A Single-Arm Prospective Study. J Pers Med. 2023 Nov; 13(11): 1603. doi: 10.3390/jpm13111603. PMCID: PMC10672422. PMID: 38003918.

Ohbayashi Y, Imataki O, Uemura M, Takeuchi A, Aoki S, Tanaka M, Nakai Y, Nakai F, Miyake M. Oral microorganisms and bloodstream infection in allogeneic hematopoietic stem cell transplantation. Clin Oral Investig. 2021 Jul;25(7):4359-4367. doi: 10.1007/s00784-020-03749-9. PMID: 33392808.

Davidson KW, Silverstein M, Cheung K, Paluch RA, Epstein LH. Experimental Designs to Optimize Treatments for Individuals: Personalized N-of-1 Trials. JAMA Pediatr. 2021 Apr 1;175(4):404-409. doi: 10.1001/jamapediatrics.2020.5801. PMID: 33587109; PMCID: PMC8351788.

The Lucifer Effect: Understanding How Good People Turn Evil, by Philip Zimbardo.