Experimentation without randomized controls

## 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. It has several major advantages over other approaches to research. 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.

Randomization does not have to be at the individual patient level. You might use cluster randomization where you select a large number of dental practices and randomly assign half of the locations to try a new intervention on all of their patients. All of the patients at the other clinic serve as controls. Cluster randomization requires special care during data analysis and often require matching or other controls. Nevertheless, cluster randomized trials have all the same benefits of individual randomization.

There are times, however, when you should reject individual or cluster randomization in favor of some non-randomized alternatives. The sad truth is that 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. In many settings, things are already messy enough and you don’t need the extra hassle of switching back and forth between two different treatment protocols. This is often true in emergency care settings where the distraction caused by randomization is quite unwelcome.

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.

This paper will not cover observational studies, those where the choice of who gets what is totally out of control of the researcher. Pandis (2024) provides a nice overview of observational studies and contrasts them to randomized studies.

This paper will focus on something in between a randomized study and an observational study. When you have the ability to assign patients to receive a treatment or not, you may choose to 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.

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.”

There is no way to directly test for competing temporal trends. You can, at best, make a qualitative and highly subjective argument that no alternate factor could account for the differences that you saw in your historical controls study.

## The before-and-after design

A close cousin to the historical control design is the before-and-after design also known as a pre-test post-test design. You measure a group of patients at baseline, give all of them the intervention, and then measure them again. There is no control group.

A Korean study of Basic Life Support (BLS) training during Dental School evaluated 98 students who received this training during their third year of dental school (Kim 2020). The researchers gave them a test two years after grauduation to see how much they remembered. These students did poorly, scoring an average of 56 points out of 100. Then they provided a refresher course in BLS and gave the test again. The average score increased to 81 points. The authors concluded that it was “necessary to update BLS training periodically and also implement more effective education methods to maintain BLS knowledge and practical skills.”

In the before-and-after design, you have to make the same untestable assumption assumed earlier. You must assume that untreated patients would not have shown as much improvement as you saw among your treated patients. In this aforementioned study, you might safely conclude that the knowledge level would not have changed much and if anything would probably have declined further over time without the refresher course. This seems quite reasonable in this setting. In other settings, this untestable assumption might fail for the same set of reasons that the historical controls trial failed.

## 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, page 332) 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.

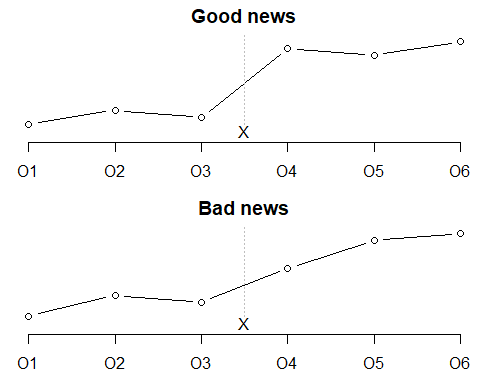


Figure 1. Two hypothetical outcomes from an interrupted time series design.

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.

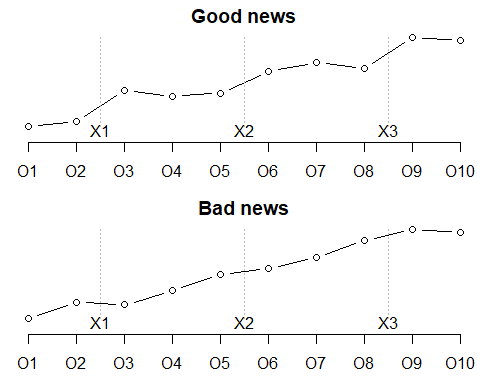


Figure 2. Two hypothetical outcomes for a for an interrupted time series design with phases.

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. Certainly, a flat-jump-flat-jump-flat pattern is very persuasive positive evidence. Equally certainly, a pattern with no jumps and no changes in trend is very persuasive negative evidence. But interpretation becomes difficult when you see a jump at only one of the three intervention points. It could be a negative finding because the one jump might have been caused by something other than the intervention. It could just as easily represent an intervention where only one of the three phases was truly 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 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.

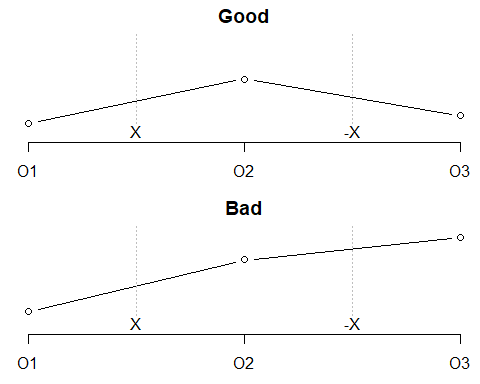


Figure 3. Two hypothetical results from a withdrawal design

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 not be too persuasive. How do you know that this 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.

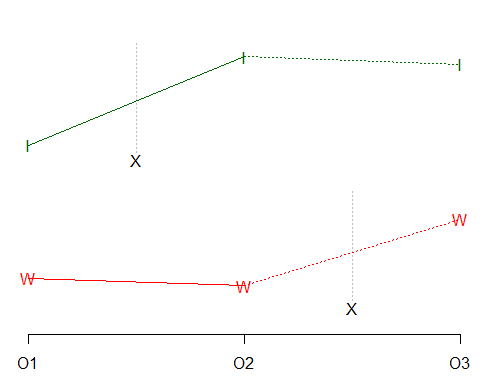
## Waiting list control group

In some settings, researchers cannot or will not withhold the treatment from their patients but they do have the ability to control the timing. These open up an opportunity to randomize times.

The simplest concept is to evaluate every patient at baseline, randomly assign half to receive the intervention immediately and half to receive the intervention at the end of the study. This is a waiting list control design.

If the researchers have sufficient resources, the study does not have to end there. When the waiting list control patients complete their intervention, evaluate everyone a third time. Does the waiting list control group show an improvement late that is comparable to those who received the intervention early? Do the intervention patients show a long term persistence in the effectiveness of their intervention or is there some backsliding?

Figure 4 shows a hypothetical outcome of a waiting list control design. The two groups



# ## Randomizing time

Figure 4.

A study of dental care practices used an Internet driven educational intervention to encourage providers to ask about tobacco use and to advise any smokers so identified to stop smoking (Houston 2008). The researchers identified 190 practices for the study and all of them were evaluated on asking and advising at baseline using patient exit cards. Patients received a card at the end of their visit and asked to fill out the cards at home and return them to the research team. Then half of the clinics were randomly assigned to receive the intervention immediately and half were assigned to receive the intervention at the end of the eight month study. While the intervention had no impact on asking about smoking, it did result in a greater degree of advising among the smokers identified. The intervention clinics improved substantially advising 55% of the time at eight months compared to 44% at baseline. The control group the advising rates were 45% at eight months, only a slight increase over the baseline rate of 42%. This disparity, as measured by a test of statistical interaction, was statistically significant. The authors concluded that an “Internet-delivered intervention designed to promote and support tobacco control in dental practices can be effective.”

This study did not evaluate again sixteen months after the start to see if there was a similar improvement in the waiting list control clinics or to see if the Internet delivery persisted long term in the treatment group.

## Stepped wedge design

Once you have control over timing, you can do even more. Instead of treating a random half of the patients right away and the other half at the end of the study, you can divide things even finer. Start a random group early, another random group late, and the rest at one or more times in between. This is a stepped wedge design. Think of it as a waiting list control on steroids.

When the jump at one treatment time is accompanied by flatness at the patients who were not treated, you have substantial evidence that no external factor is conspiring against you.

Figure 5 shows the hypothetical outcome of a stepped wedge design. The orange line with ones represents the early intervention and shows the pattern of one jump with flatness elsewhere that supports the effectiveness of your intervention. The blue line with twos, representing the middle intervention and the green line with threes, representing the late intervention, also show a pattern supportive of the effectiveness of your intervention. The dotted lines highlight that not every stepped wedge design takes advantage of the information available at times other than the intervention. This is a lost opportunity because flatness at these observations greatly strengthens the credibility of your findings.

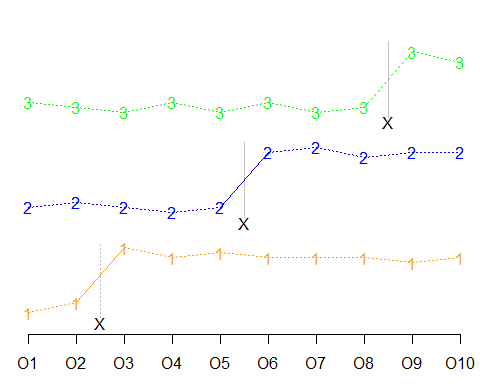


Figure 5. Hypothetical outcome of a stepped wedge design

A study at a large teaching hospital examined the effectiveness of a new care protocol that encouraged nurses to provide better oral health care to patients in their wards (Schafthuizen 2023). The intervention was a seven step Implementation of Change Model for provision of good oral health care. The researchers divided the wards into four groups that were provided this intervention at separate times. They measured adherence to the protocol using a nine item questionnaire given to patients at five observation times bracketing each intervention. The research team also assessed the nurses’ knowledge and attitude towards oral care, but only at the beginning and very end of the study. Knowledge did show a small but statistically significant increase from an average of 69 points (out of 100) to 72 points. Attitudes remained unchanged, and adherence actually declined from 60% at the start of the study to 35% at the end. The researchers did not have a good suggestion for this decline and suggested an evaluation of barriers and facilitators was needed.

There are parallels and differences between the stepped wedge design and the interrupted time series design with multiple phases. The former divides patients into random groups and allocates them at different times. The latter divides the intervention into phases and allocates the phases at different times. See Fok (2105) for a more detailed comparison.

## Regression discontinuity

In all of the previous examples, everybody gets the intervention, sometimes with multiple evaluations, sometimes with the intervention broken into phases, and sometimes with the intervention staggered over time. Another type of design, the regression discontinuity design, examines treatment allocation where a qualifying variable is used to decide controls who gets the intervention. This qualifying variable can be (and often is) associated with the outcome.

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. Only 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 just barely on either side of the dividing line and ignore the patients at the extremes.

There is a good intuition to this. 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 6 through 8 shows a hypothetical setting for a regression discontinuity design. A qualifying variable shows a relationship to the outcome (Figure 6). Patients scoring below a threshold are assigned to the treatment (Figure 7). 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 8).

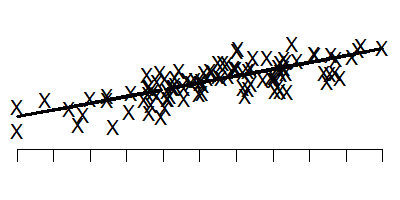


Figure 6. Hypothetical data showing a relationship between a qualifying variable and a health outcome.

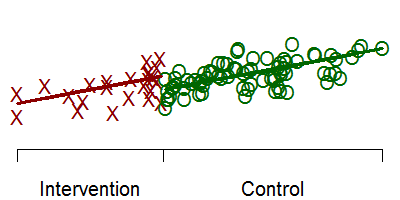


Figure 7. Allocation of patients to the intervention based on the qualifying variable.

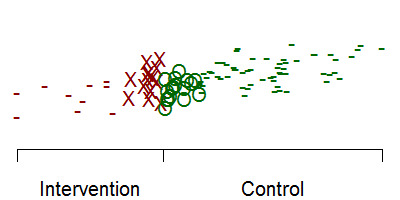


Figure 8. Removal of patients at the extremes.

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).

## Considerations for choosing a non-randomized experiment.

I’m a professional statistician and it may seem to you like a betrayal of everything I’ve learned to try to talk you out of randomizing. I’m sorry if you feel that way, but I do strongly believe that sometimes abandoning randomization is the best course of action. But only sometimes. Here are some settings where you would prefer to do this.

Sometimes an intervention is implemented by someone higher on the food chain than the researcher. A nation might pass a law providing new benefits for dental care or impose restrictions during the pandemic. A court ruling in a liability case leads to major changes in dental practice. A state licensing board changes certification requirements. These settings fit naturally into an interrupted time series design. For a law, court ruling, or certification change that is phased in, even better.

The astute reader might note here that an intervention not under the direct control of the researcher is no longer an experiment, but would be more accurately described as an observational study. The dividing line between quasi-experimental studies and observational studies is quite fuzzy and the dividing line among the various quasi-experimental designs is even fuzzier. One research team developed a checklist of seven questions to help categorize the type of research design, including all the ones described here and more (Reeves 2017). In any case, the categorization of research designs is less important than the careful use of extra information that these designs provide you.

In every one of these settings, though, avoid the temptation of the before-and-after design. If you can’t randomize or prefer not to randomize, you have to work harder to produce persuasive evidence. Adding multiple evaluation times, breaking the intervention into phases or evaluating both the addition and withdrawal of an intervention are all changes that take a lot more time and energy. Even so, the investment is often worth the trouble.

Don’t eliminate randomization, of course, as a method in your research toolbox. A randomized controlled trial is great if it is easy to apply individually and if the the individual patients (and their care providers) are willing to accept it.

## 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.

Campbell, D. T. (1988). Methodology and epistemology for social science: Selected papers. (E. S. Overman, Ed.). University of Chicago Press.

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.

Fok CC, Henry D, Allen J. Research Designs for Intervention Research with Small Samples II: Stepped Wedge and Interrupted Time-Series Designs. Prev Sci. 2015 Oct;16(7):967-77. doi: 10.1007/s11121-015-0569-4. PMID: 26017633; PMCID: PMC4581909.

Houston TK, Richman JS, Coley HL, Ray MN, Allison JJ, Gilbert GH, Gordon JS, Kiefe CI; DPBRN Collaborative Group. Does delayed measurement affect patient reports of provider performance? Implications for performance measurement of medical assistance with tobacco cessation: a Dental PBRN study. BMC Health Serv Res. 2008 May 8;8:100. doi: 10.1186/1472-6963-8-100.

Jones KM, Swearer SM, Friman PC. Relax and try this instead: abbreviated habit reversal for maladaptive self-biting. J Appl Behav Anal. 1997 Winter;30(4):697-9. doi: 10.1901/jaba.1997.30-697.

Kim SY, Shin D, Kim HJ, Karm MH. Changes of knowledge and practical skills before and after retraining for basic life support: Focused on students of Dental School. Int J Med Sci. 2020 Oct 22;17(18):3082-3090. doi: 10.7150/ijms.47343.

Nishi H, Horikoshi S, Yoshida T, Fukushima N, Oshita K, Munenaga S, Edahiro T, Ureshino H, Shigeishi H, Yoshioka Y, Konishi M, Ide N, Ogawa Y, Marukawa R, Shintani T, Ino N, Kajiya M, Kakimoto N, Ohge H, Ichinohe T, and Kawaguchi H. 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.

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.

Pandis N, Tu YK, Fleming PS, Polychronopoulou A. Randomized and nonrandomized studies: complementary or competing? Am J Orthod Dentofacial Orthop. 2014 Nov;146(5):633-40. doi: 10.1016/j.ajodo.2014.08.002. Epub 2014 Oct 28. PMID: 25439214.

Reeves BC, Wells GA, Waddington H. Quasi-experimental study designs series-paper 5: a checklist for classifying studies evaluating the effects on health interventions-a taxonomy without labels. J Clin Epidemiol. 2017 Sep;89:30-42. doi: 10.1016/j.jclinepi.2017.02.016.

Roberts ET, Mellor JM, McInerny MP, Sabik LM. Effects of a Medicaid dental coverage “cliff” on dental care access among low-income Medicare beneficiaries. Health Serv Res. 2023 Jun;58(3):589-598. doi: 10.1111/1475-6773.13981.

Schafthuizen L, Spruit-Bentvelzen L, van Dijk M, van Rosmalen J, Ista E. Implementation of a nursing oral health care protocol in a university teaching hospital: A cluster-randomized stepped-wedge design. Int J Dent Hyg. 2023 Sep 18. doi: 10.1111/idh.12748. Epub ahead of print. PMID: 37722075.

Timm LH, Farrag G, Wolf D, Baxmann M, Schwendicke F. Effect of electronic reminders on patients’ compliance during clear aligner treatment: an interrupted time series study. Sci Rep. 2022 Oct 5;12(1):16652. doi: 10.1038/s41598-022-20820-5.

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