Quasi-experimental study designs for evaluating practice, programs and policies: assessing the assumptions

Till Bärnighausen1,2,3, Catherine Oldenburg4, Peter Tugwell5, Christian Bommer6, Cara Ebert6, Mauricio Barreto7,8, Eric Djimeu9, Noah Haber2, Hugh Waddington9, Peter Rockers10, Barbara Sianesi11, Jacob Bor12, Günther Fink2, Jeffrey Valentine13, Jeffrey Tanner14, Tom Stanley15, Eduardo Sierra16, Eric Tchetgen Tchetgen2, Rifat Atun2, Sebastian Vollmer6

1Institute of Public Health, Faculty of Medicine, Heidelberg University, Heidelberg Germany
2Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, USA
3Africa Health Research Institute, Somkhele, South Africa
4Francix I. Proctor Foundation, University of California San Francisco, San Francisco, USA
5Centre for Global Health, University of Ottawa, Canada
6Department of Economics, University of Göttingen, Göttingen, Germany
7Instituto de Saúde Coletiva, Federal University of Bahia, Salvador, Brazil
8National Institute in Science, Technology and Innovation in Health (CITECS), Salvador, Brazil
9International Initiative for Impact Evaluation, London, UK, and Washington DC, USA
10Center for Global Health and Development, Boston University, Boston, USA
11Institute for Fiscal Studies, London, UK
12School of Public Health, Boston University, Boston, USA
13College of Education and Human Development, University of Louisville, Louisville, USA
14Impact Evaluation, World Bank, Washington DC, USA
15Hendrix College, Conway, Arkansas, USA
16Centre for Medicine and Society, Freiburg University, Freiburg, Germany
Corresponding author:
Till Bärnighausen
Institute of Public Health
Faculty of Medicine, University of Heidelberg, Heidelberg
INF 130.3
69120 Heidelberg, Germany

Department of Global Health and Population
Harvard T.H. Chan School of Public Health
655 Huntington Avenue
02115 Boston, MA, USA
Abstract

Quasi-experimental designs are gaining popularity in epidemiology and health systems research – in particular for the evaluation of healthcare practice, programs and policy – because they allow strong causal inferences without randomized controlled experiments. We describe the concepts underlying five important quasi-experimental designs: Instrumental Variables, Regression Discontinuity, Interrupted Time Series, Fixed Effects, and Difference-in-Differences designs. We illustrate each of the designs with an example from health research. We then describe the assumptions required for each of the designs to ensure valid causal inference and discuss the tests available to examine the assumptions.
What’s new

• Causal inference in experiments, quasi-experiments, and non-experiments is today mostly based on the potential outcomes model, which requires the stable unit treatment value assumption (SUTVA) to be met – the value of an outcome for a unit exposed to the treatment of interest is the same irrespective of the mechanism used to assign the treatment and independent of the treatments that other units receive.

• In quasi-experiments the unconfoundedness assumption, which is required for causal inference in non-experiments in addition to SUTVA, is replaced by other – often weaker – assumptions.

• Key assumptions of important quasi-experimental designs include the exogeneity and monotonicity assumptions of Instrumental Variables designs, the discontinuity and monotonicity assumptions of Regression Discontinuity and Interrupted Timer Series designs, and the parallel trends assumption of Difference-in-Differences designs and its generalization (no time-varying confounding) for Fixed Effects designs.

• While the key assumptions can ultimately not be verified, tests exist that can falsify or increase the plausibility of the assumptions.

• Several statistical developments have recently increased our ability to test the assumptions underlying important quasi-experimental study designs.
Introduction

Quasi-experiments offer important opportunities for causal inference in epidemiology, health systems research, and program evaluation. As a group of research designs, they occupy a “sweet spot” between experiments and non-experiments, potentially generating causal evidence of both high internal and high external validity [1]. Like experiments, they can control not only for observed but also for unobserved confounding factors in estimating causal effects – leading to internally valid estimates. Like non-experiments, they typically leave the natural context in which data are generated undisturbed – leading to externally valid estimates [2-4].

Traditionally, quasi-experiments have not been part of the mainstay of epidemiological and health research methods, and they have commonly been excluded in systematic reviews of effectiveness [2-4]. However, a number of recent trends have led to increasing use of quasi-experiments to answer epidemiological and health research questions. These trends are described in more detail in the editorials [5, 6] and the introduction to this themed issue of the Journal of Clinical Epidemiology [7], as well as in a dedicated paper on the value and the uses of quasi-experiments in epidemiology [1].

One important reason for the relative paucity of quasi-experimental studies in epidemiology is that such studies require a detailed understanding and examination of the assumptions that need to be met to ensure valid causal inference. This content is typically not taught in medical schools and schools of public health nor is it covered in epidemiology textbooks. Here, we aim to contribute an accessible yet rigorous introduction to the assumptions underlying quasi-experiments and the tests that can be used to assess whether the assumptions hold. This introduction should be of use to both people intending to become more informed judges of quasi-experimental evidence generated by others and people people planning to use quasi-experimental designs to generate new evidence. The former group includes researchers carrying out systematic reviews that incorporate quasi-experimental evidence [8-10]. For the latter group, this introduction will likely be insufficient to provide sufficient detail and depth of understanding to correctly implement quasi-experimental designs; however, it should be a useful and efficient starting point.

For five important quasi-experimental study designs – Instrumental Variables, Regression Discontinuity, Interrupted Time Series, Fixed Effects, and Difference-in-Differences designs –, we will first describe the design conceptually, and then explain the major assumptions that are required for valid causal inference using the design. We will then discuss important tests to assess whether the assumptions underlying a particular design hold. Finally, for each design we will provide an example of a good quasi-experimental study and briefly discuss how the authors have tested the assumptions.

Instrumental Variables designs

Conceptual introduction

Instrumental Variables designs allows for the estimation of causal effects when there exist one or more exogenous factors (the instruments) with a sufficiently strong association with the exposure or treatment variable of interest (in the following referred to as explanatory variable) [11, 12]. Conceptually, exogeneity means that no factor affecting the outcomes
of interest can have exerted any influence on the instrument. In a regression estimation context, exogeneity means that a variable is uncorrelated with the error term. When the instrument is uncorrelated with the error term, it can be used to estimate the causal effect of treatment without requiring the assumption that there is no bias caused by unobserved factors related to both the explanatory and outcome variables (see Box 1 for an illustrative example).

Instrumental Variables estimation can be used for the estimation of regression models when at least one explanatory variable is endogenous. Endogeneity of an explanatory variable causes bias in the estimates and can arise for various reasons, most importantly: (1) simultaneity (also called reverse causality) – the endogenous explanatory variable is to some extent determined by the outcome variable; (2) confounding – there exists at least one variable which is not part of the statistical model but affects both the endogenous explanatory variable and the outcome variable; (3) measurement error in the outcome variable which is correlated with the endogenous explanatory variable; and 4) measurement error in the endogenous explanatory variable which is correlated with the variable’s observed value. To address endogeneity, the Instrumental Variables design exploits a source of variation in the endogenous explanatory variable that is not related to the outcome other than through the explanatory variable of interest. The instrument identifies this – exogenous – variation in the explanatory variable.

Instrumental Variables estimation is typically carried out in so-called two-stage-least-squares estimation. This estimation approach is called two-stage-least-squares, because it can be conceived as a two-stage process. In the first stage, the endogenous explanatory variable is predicted using ordinary least squares regression with the instruments and all other exogenous covariates of the main regression model as explanatory variables. In the second stage, these predicted values are used in the ordinary least squares estimation of the main regression model instead of the observed values of the endogenous explanatory variable. In practice, the two-stage-least-squares estimator is estimated in one stage, because the two-stage process described above ignores the estimation uncertainty in the first stage and thus leads to underestimates of the standard errors.

Two-stage-least-squares estimation allows us to identify the causal effect of the endogenous explanatory variable by exploiting only the exogenous variation in the endogenous explanatory variable, which is generated by the instruments. Wooldridge [13] and Angrist and Pischke [14] provide a more detailed description of Instrumental Variables estimation. Instrumental Variables estimation is also possible for non-linear models [15-17], which are common in epidemiology and health research.

**Underlying assumptions and interpretation**

Three assumptions or conditions must be met for Instrumental Variables estimation to be able to identify causal effects. First, the instruments must explain sufficient variation in the endogenous explanatory variable. This assumption is referred to as relevance condition. If there is only a weak correlation between the instruments and the endogenous explanatory variable, the instruments are called weak instruments and can lead to inconsistent estimates and amplified bias [12].

Second, the instruments must causally affect the outcome only through the endogenous explanatory variable [18]. This assumption is referred to as the exclusion restriction. The exclusion restriction requires not only that the instruments do not directly affect the
outcome of interest (other than through the explanatory variable of interest), but also that the instruments are not correlated with any other observed covariate or unobserved variable affecting the outcome, i.e. the instruments must be uncorrelated with the error term of the regression model. Violation of the exclusion restriction would result in alternate pathways between the instruments and the outcome, biasing the effect estimate.

Third, all people who are affected by a given instrument are affected by it in the same way [18]. This so-called monotonicity assumption rules out that for a given change in an instrument there are both some individuals whose treatment status changes in one direction and other individuals whose treatment status changes in the opposite direction. A corollary of this assumption is that there cannot be any ‘defiers’ in the population. ‘Defiers’ are individuals whose treatment status would be affected in the ‘wrong’ direction. In the context of our example, the Vietnam Era Draft Lottery (Box 1), ‘defiers’ would be people who would not serve in the military if they received a lottery number leading to a call to serve but who would serve in the military if they did not receive a lottery number leading to a call to serve. The monotonicity assumption must in particular be met if the effect of treatment is heterogeneous across the population and if there is sorting into treatment based on treatment effect [19].

Even if an instrument is relevant and it is assumed that it satisfies the exclusion restriction and the monotonicity assumption, Instrumental Variables estimation cannot identify the average effect of a treatment on the full population. Instrumental Variables estimation can only identify the effect of the treatment on the subpopulation that was induced by the instrument(s) to change their treatment status – i.e., the ‘compliers’. ‘Compliance’ in this context is the exogenous variation predicted by the instruments in the first stage. The Instrumental Variables effect size may not be a valid estimate for the two populations that do not change their treatment status when the instrument changes – the ‘never-takers’ and the ‘always takers’. (The third populations that does not ‘comply’ are defiers, who are assumed not to exist, see above.) In the context of our example (Box 1), the subpopulation of ‘compliers’ are those people who were induced by their lottery number to serve in the military. The results do not generalize to individuals who would always or never serve in the military, regardless of their lottery number. To emphasize that Instrumental Variables estimation generates an effect size that is only valid for the subpopulation of ‘compliers’, the effect estimated by Instrumental Variables approaches is called complier average causal effect (CACE) [20] or local average treatment effect (LATE) [21].

Tests of assumptions and conditions

1. Relevance condition. Relevance can be tested and it is thus a condition rather than an assumption. Relevance can be partly verified by observing a statistically significant correlation of the instruments with the endogenous explanatory variable conditional on all other control variables. However, a significant correlation is not sufficient for relevance. Bound, Jaeger and Baker [22] show that Instrumental Variables estimation even in very large samples leads to inconsistent estimates if the partial correlation between the instruments and the endogenous explanatory variable is not strong enough. This problem is indicated by a small value of the F-statistic in the estimation of the first stage. A small value of the F-statistic is also a good indicator for the finite sample bias of the Instrumental Variables estimator mentioned above. As a rule of thumb, Staiger and Stock propose an F-statistic cut-off of 10 to determine whether an instrument is weak or strong [23]. This suggestion is based on the relative bias of the two-stage-least-squares estimator: at this F-
test value, the probability that the worst-case relative bias of the estimator is approximately 10% or less is approximately 5% [24]. Hahn and Hausman have proposed an alternative test for weak instruments. According to this test, instruments are strong if the null hypothesis that two different two-stage-least-squares estimators – the forward and the reverse estimator – are the same cannot be rejected [25]. Andrews and Stock provide a comprehensive overview of tests for weak instruments and technical corrections to deal with it [26].

2. **Exclusion restriction assumption.** Traditionally, the *exclusion restriction* has been believed not to be testable [21, 27], with the exception of cases with more instruments than endogenous explanatory. In this special case, so-called overidentification tests can be used to falsify sets of instruments [28-30]. The basic idea underlying overidentification tests is that if all of the instruments satisfy the exclusion restriction, using any subset of these instruments should lead to the same estimate of the treatment effect; comparing multiple estimates of the treatment effect based on different subsets of instruments can thus tell us whether all of the instruments satisfy the exclusion restriction. The null hypothesis of overidentification tests is that all of several instruments satisfy the exclusion restriction. If the null hypothesis is rejected, at least one of the instruments fails to satisfy the exclusion restriction; if the null hypothesis is not rejected, it can, however, not be concluded that all of the instruments are valid – they could all be invalid. Thus, overidentification tests are generally falsification tests and can only be used verify that a particular instrument meets the exclusion restriction if one of the other instruments is already known to meet the exclusion restriction based on prior information (e.g., if it was randomly assigned).

Recently, a number of new tests of the exclusion restriction have been developed that do not require more instruments than endogenous explanatory variables [31-37]. These tests derive implications of the exclusion restriction and then test whether estimation using an instrument in a given application generates these implications. As for overidentification tests, these new tests can only falsify the exclusion restriction: If the null hypotheses that the implications are found in the data is rejected, the exclusion restriction is violated; if the null hypothesis is not rejected the exclusion restriction can still be violated.

Because the available tests of the exclusion restriction cannot verify the exclusion restriction, in addition to carrying out these tests researchers should use other strategies to strengthen the credibility of instruments [38, 39], including

- Providing a strong intuitive argument for instrument validity – in many cases, it is an intuition that will have let a research to consider and select an instrument. This intuition will need to be carefully explained to convince others that an instrument is valid.
- Checking the intuition by regressing the outcome variable on the instrument and the exogenous co-variates – if the sign of the coefficient of the instrumental variable in this so-called reduced-form regression is at odds with the intuition, the instrument’s validity is questionable.
- Checking the intuition by regressing the endogenous explanatory variable on the instrument and the exogenous co-variates – again, if the sign of the coefficient of the instrumental variable in regression is at odds with the intuition, the instrument’s validity is questionable.
• Testing the intuition against theory [40] and empirical evidence generated in other populations than the one under study.
• Show that reasonable speculations why an instrument might be invalid are wrong – in examining instruments, critics will raise doubts with plausible speculations why an instrument might violate the exclusion restriction. These speculations can often be empirical tested and if they are found to be untrue, the case for instrument validity is strengthened. This approach has been used in some of the best-known examples of Instrumental Variables estimations in the economic literature [41-43].
• Carefully consider if including other variables in the estimation could strengthen the case for an instrument – instruments will not be valid if they are correlated with variables that determine the outcome of interest but have been omitted from the estimation. To render instruments invalid, it is sufficient that these omitted variables are correlated with the instruments, i.e., these omitted variables do not also have to be confounders of the relationship between the explanatory variable of interest and the outcome. Controlling for variables that are correlated with the instruments and are plausible determinants of the outcome variable can strengthen the case for an instrument.
• Use alternative instruments – if several instruments exist but it is not feasible to include all of the instruments in the same estimation (for formal overidentification tests), the results from applying several instruments separately from one another should be compared. If the parameter estimates using the alternative instruments differ substantially, the case for the validity of the instruments is weakened.

3. **Monotonicity assumption.**

The **monotonicity assumption** tends to be far less discussed and assessed in the Instrumental Variable literature than the **relevance condition** and the **exclusion restriction assumption** [44]. One possible reason for this relative neglect may be that in many cases it is highly likely that the **monotonicity assumption** is satisfied, as in the case of the Vietnam draft lottery (Box 1). However, in some cases monotonicity may be likely to be violated, as in an example described by Imbens and Angrist [21]: If two officials screen applicants to determine admission to a program and the two officials have different admission rates, the identity of the official would be a good instrumental variable because the exclusion restriction is likely met (the official’s identity does not independently affect program effects). In this case, the **monotonicity assumption** would only be met if all applicants admitted by the official with the higher admission rate would also (counterfactually) have been admitted had they been screened by the official with the lower admission rate. In many cases – e.g., there are a number of different criteria that are important for admission – this assumption is unlikely to hold.

Thus, it is important that researchers discuss the plausibility of the monotonicity assumption in their particular Instrumental Variable study. It is important to note that many of the newer tests of the exclusion restriction cited above are both tests of the exclusion restriction and monotonicity [32-34, 37]. As a result, if the tests reject the null hypothesis both the **exclusion restriction** and the **monotonicity**
assumption will be difficult to defend, while failure to reject the null hypothesis does not verify either assumption.

Regression Discontinuity designs

Conceptual introduction

Regression Discontinuity designs have been called the “next best thing to a randomized trial” [45]. Regression Discontinuity designs can be used to obtain causal effect estimates when a treatment is assigned by a threshold rule on a continuous assignment variable (see Box 2 for an illustrative example) – i.e., people with values of the assignment variable on one side of the threshold are eligible for the treatment, while people with values of the variable on the other side of the threshold are not. Threshold decision rules are ubiquitous in clinical practice [46]. For instance, antiretroviral therapy for HIV is prescribed when patients’ CD4 counts fall below a threshold [47]; intensive medical care is provided to newborns weighing less than 1500g, who are classified as “very low birth weight” [48]; and medication against hypertension, diabetes, or hypercholesterinemia is prescribed if certain continuously measured variables exceed predetermined thresholds. But threshold rules are also commonly used in public health to assign health-related interventions to individuals (e.g., to notify students in New York that their body mass index falls outside a healthy weight and that they should review their weight with a healthcare provider [49]) and communities (e.g., to provide a government insurance for tertiary care to poor people in India [50]). Threshold rules simplify complex diagnoses, enabling evidence-based medicine to be practiced at scale. Threshold rules are particularly important for clinical decision-making in low-resource settings, where advanced diagnostic testing is not available or where care is provided by health-care workers who have received less training than doctors or nurses [51].

Threshold rules offer a paradox: at the macro-level, the efficiencies gained (reductions in time and complexity) can improve resource allocation and save lives. On the micro-level, however, there is a decidedly arbitrary nature to such rules: patients just above and below the threshold are essentially identical on all observed and unobserved characteristics – similar to a randomized trial – and yet they receive different diagnoses, different treatments, and may have thus different outcomes. It is this arbitrariness of treatment assignment that makes threshold rules amenable to analysis as a natural experiment.

Underlying assumptions and interpretation

A key assumption for Regression Discontinuity design to generate valid estimates is the so-called continuity assumption -- potential outcomes must be continuous at the threshold [47, 52]. The continuity assumption means that as an individual moves across the threshold from being untreated to being treated the potential untreated outcome (which is not observable above the threshold) is continuous - e.g. no other factors change at threshold that cause a jump in the potential outcome - and similar for the potential treated outcome (which is not observable below the threshold). If the continuity assumption is met, the Regression Discontinuity design is equivalent to a randomized controlled trial within a small neighborhood around the threshold.
The \textit{continuity assumption} is a weak unconfoundedness assumption. It is much more limited and therefore easier to defend than the broad unconfoundedness assumption typically invoked in non-experimental studies that rely on covariate adjustment through regression or matching \cite{53}. It also often viewed as an assumption that is generally weaker than the key assumptions required for other quasi-experimental designs \cite{54}. For the \textit{continuity assumption} to be violated, the33re would need to be another relevant discontinuity at the Regression Discontinuity threshold, i.e., a discontinuity in the relationship between the assignment variable and the outcome of interest or a discontinuity in confounders.

In addition to the continuity assumption, the Regression Discontinuity decision rule and the threshold value must be known \cite{53}. There must be a measured continuous assignment variable (Z), and it must be known whether the treatment is assigned above or below a threshold value of Z. The knowledge about the decision rule includes whether treatment assignment is

- Deterministic – “sharp” Regression Discontinuity, i.e. all patients who are below Z always receive treatment and no patients above Z receive treatment or
- Stochastic – “fuzzy” Regression Discontinuity, such as in cases where clinical judgment factors into treatment decisions and therefore treatment status is not solely determined by the value of the assignment variable (i.e. there are treated and non-treated individuals on both sides of the threshold).

In a sharp Regression Discontinuity design the probability of treatment must jump from zero to one at the threshold, whereas in a fuzzy Regression Discontinuity design the probability of treatment must not be equal for all individuals just below and all individuals just above the threshold. In the fuzzy design, the estimated effect can be interpreted as the intent-to-treat effect of being below (above) the threshold. The treatment effect for the population of ‘compliers’ with the threshold rule (i.e., the CACE or LATE) can be identified using the assignment just below vs. just above the threshold as an Instrumental Variable \cite{47, 52}. In this case, the \textit{monotonicity assumption} (see above) needs to be met in addition to the \textit{continuity assumption}. The \textit{exclusion restriction assumption}, on the other hand, will not be required if the \textit{continuity assumption} is met \cite{52, 54}.

The Regression Discontinuity design can generate strong causal inferences. However, the extent to which these inferences are generalizable may be limited, if treatment effects are heterogeneous across the range of the assignment variable Z. The literature on the generalizability of Regression Discontinuity effect size estimates ‘away from the threshold’ can be summarized in three perspectives. The first perspective is that the effect size estimate is a ‘global’ average treatment effect, which applies across all individuals, analogously to the average treatment effect estimated in a randomized controlled experiment. This perspective is justified under the assumption that the functional form of the potential outcome conditional expectation functions (if treated vs. if not treated) is known across the full range of the assignment variable Z \cite{55}. This assumption is strong and will typically not be testable.

The second perspective is that the Regression Discontinuity effect size estimate applies only ‘locally’ across an arbitrarily small region around the threshold \cite{47, 52}. This perspective requires for justification merely the \textit{continuity assumption}, which is far weaker than the assumption that the functional form of the unobservable potential outcome conditional expectation functions are known. In essence, this perspective is equivalent to conceiving Regression Discontinuity as a randomized controlled experiment including only
individuals in a small region around the threshold. In some cases, this ‘local’ effect size estimate is precisely the effect size of interest. In particular, if the question is whether a threshold used to assign individuals to a treatment (such as a clinical therapy or a health policy) should be changed. Indeed, trials with individuals ‘in small regions’ of an assignment variable are sometimes carried out to test the hypothesis whether changing a clinical threshold so that more people would be eligible to receive a treatment is of significant benefit [56]. Such trials typically differ from the Regression Discontinuity ‘trial’ in that they can only experiment on the one side of the assignment threshold that implies ‘no treatment’ under the current standard of care. However, the results of the clinical controlled ‘threshold’ trial and the Regression Discontinuity ‘trial’ will likely be very similar and serve the same purpose of providing data to inform threshold rules. For this purpose, the second perspective on the generalizability of Regression Discontinuity effect size estimates implies a strength. For the many other purposes where ‘global’ rather than ‘local’ average treatment effects would be more useful the second perspective implies a weakness.

As Lee and Lemieux point out in a third ‘perspective’, however, the weakness implied by the second perspective “may be an overly simplistic and pessimistic assessment” in many instances [54]. Lee and Lemieux show that the Regression Discontinuity effect size estimate is a weighted average treatment effect across all individuals, where the weights are proportional to the ex ante likelihood that an individual’s value of the assignment variable Z will be close to the threshold. The more similar these weights are across individuals, the more similar the Regression Discontinuity effect size estimate will be to the ‘global’ average treatment effect. While only the ex post realization of one value of Z for each individual is observable, rather than the ex ante likelihood, and the similarity of the Regression Discontinuity effect size estimate to the ‘global’ average treatment effect can thus not be quantified, the third perspective emphasizes “that the treatment effect estimated using a RD [Regression Discontinuity] design is averaged over a larger population than one would have anticipated from a purely ‘cut-off’ [or threshold] interpretation” [54], as in the second perspective.

Tests of assumptions and conditions

The assumption of continuity of potential outcomes at the Regression Discontinuity threshold implies that as we approach the threshold from above and from below, the individuals in the two groups become more and more alike in both all observed and all unobserved pretreatment characteristics (in expectation). Analysts can thus partially assess the continuity assumption by comparing the distribution of covariates observed at the point of assignment in those above and below the assignment threshold in a small neighborhood around the threshold. Such tests are analogous to “balance tests” in randomized controlled trials – with the exception that in the Regression Discontinuity case the test is ‘local’ around the threshold rather than ‘global’. Such balance tests can only falsify – if there is significant imbalance in observed pretreatment co-variates, the continuity assumption would be rejected. In contrast, failure to detect significance imbalance in observed co-variates does not rule out significant imbalance in unobserved co-variates and can thus not prove that the continuity assumption is met.

The critique of balance tests in analyzing data from randomized controlled trial has focused on the fact that if it is known with certainty that the randomization process has worked perfectly (based on additional information about the process), imbalance in observed co-variates does not provide any useful information, because one randomization
process can generate any distribution of covariates in treatment and control group [57]. However, in the application to assess Regression Discontinuity designs, balance tests are useful for checking whether there is evidence that the assignment to treatment and control was not as good as random (i.e., whether the continuity assumption is violated) [58]. Balance tests have thus become a standard in the applied regression discontinuity literature [53].

In one common situation – there is measurement error in the assignment variable – the continuity assumption will be automatically met (in expectation) [59], if one of the other assumptions of Regression Discontinuity holds true: individuals cannot precisely manipulate the assignment variable (see below). Measurement error is one important example of randomness in science [59]. In many potential applications of Regression Discontinuity in epidemiology, public health, or program evaluation, the assignment variable will indeed be measured with random error, ensuring randomization to treatment vs. control assignment in a small neighborhood around the threshold. The case for measurement error in the assignment variable can typically be made based on the standard literature on the precision with which particular variables are measured, such as biomarkers (e.g., blood pressure), physical markers (e.g., weight and height), or deprivation scores (e.g., the scores that are used to assign government programs to disadvantaged individuals or communities).

If there is measurement error in the assignment variable, the continuity assumption could still be violated – if relevant people (e.g., a patients or program directors) can and do precisely manipulate the value of the assignment variable. However, this type of manipulation – i.e., precise – will be unlikely in many instances. It is much more likely that some agents will be able to manipulate the value of the assignment variable to some extent. For instance patients may be able to influence an assignment variable, such as blood pressure or cholesterol levels, through their behaviors, but this type of manipulation will not be precise. As long as there is still some random component in the measurement of the assignment variable, the continuity assumption will be met [54].

Whether precise manipulation of the value of the assignment variable has occurred can be formally tested. The idea behind manipulation testing is that, in the absence of manipulation, the density of the assignment variable at the threshold should be continuous. Significant discontinuity in the density of the assignment variable at the threshold is usually interpreted as manipulation and violation of the continuity assumption. This reason for violation should be checked in any application of Regression Discontinuity, but it becomes the key reason for such a violation if the assignment variable is measured with random error. Manipulation test are often viewed as a falsification tests, because absence of a significant discontinuity in the assignment variable could occur despite manipulation, if for some individuals there is a discontinuity in one direction while for others there is an off-setting discontinuity in the other direction. McCrery developed the first and currently most commonly used manipulation test [60], which is based on a local polynomial density estimator [61]; alternative manipulation tests have recently been developed [62, 63].

In addition to balance tests and manipulation tests, the best practice for Regression Discontinuity includes a number of graphs to visually support the plausibility of the design itself – (i) the probability of treatment by the assignment variable and (ii) the outcome of interest by the assignment variable – and the continuity assumption – (iii) covariates by the assignment variable and (iv) the density of the assignment variable [53, 64]. In the
fuzzy Regression Discontinuity design, it is also important to discuss and test the monotonicity assumption [32-34, 37]. Finally, the analyst should carry out sensitivity analyses to test the robustness of Regression Discontinuity findings. These include variations of the main analyses in terms of the

- Size of the neighborhood around the threshold, or the bandwidth. In general, both power and bias in estimating effects in Regression Discontinuity increase with the bandwidth. Formal tests exist to determine the optimal bandwidth given this power-bias trade-off [65], but robustness to alternative bandwidth will increase the confidence in Regression Discontinuity results.
- Covariates controlled for in the analyses.
- Functional form of the relationship between the treatment variable and the outcome of interest. If the assignment variable is continuous, no functional form assumptions are necessary for Regression Discontinuity effect size estimation. In this case, the bandwidth can theoretically be made arbitrarily small, ensuring that the continuity assumption is met in the limit. However, power limits the ability to make the bandwidth arbitrarily small. Moreover, often the assignment variable will be discrete or only reported in coarse intervals. In these cases – limited power, discrete or coarsely reported assignment variable – Regression Discontinuity effect size estimation will require the assumption of a functional form capturing the relationship between treatment and outcome [66]. The confidence in Regression Discontinuity estimation will increase if the results are robust to different (parametric and non-parametric) functional form choices [54].

**Interrupted Time Series designs**

Interrupted Time Series design is a particular case of the Regression Discontinuity design – the assignment variable is calendar time and the threshold is the point in calendar time at which an intervention or policy change was implemented [67] (see Box 3 for an illustrative example). Interrupted Time Series has been described as the most commonly used quasi-experimental design in epidemiology [68]. The differences between Interrupted Time Series and other applications of Regression Discontinuity designs arise because of the particular nature of the assignment variable. For the continuity assumption to be met, it is necessary that no other interventions or confounding covariates than the treatment of interest in an analyses changed at the threshold. For this reason, the continuity assumption may be difficult to defend in the Interrupted Time Series application of Regression Discontinuity: Many factors may plausibly have changed at or around the threshold point in time. For instance, explicit policy changes are often accompanied by implicit changes in practice, and individuals sometimes anticipate treatments (e.g. laws becoming effective at certain dates) and prepone or postpone certain activities in response to this anticipation. In contrast, in many other applications of the Regression Discontinuity design, the assignment variable is very specific to the treatment it assigns (sometimes, it has been specifically constructed for assignment of a particular treatment), and it is thus less likely that other interventions and confounding covariates change at the thresholds used for the assignment of the treatment of interest. As in the case of other Regression discontinuity applications, violations of the continuity assumption can be partially assessed by checking for discontinuities in covariates. However, the number of unobserved covariates that are plausible confounders of the relationship between the treatment and the outcome of interest may be especially large in the case of Interrupted Time Series analyses, limiting the effectiveness of balance tests in increasing the credibility of the design.
Difference-in-Differences designs

Conceptual introduction

Difference-in-Differences designs are alternatively referred to as controlled before and after studies, untreated control group design with independent pretest and posttest samples, and control group design with pretest and posttest [69-72]. They are commonly used to evaluate the effectiveness of practice, programs and policies, including in the healthcare sector [73-79]. Difference-in-differences designs are a particular case of the broader Fixed Effects category of studies (see below) [14]. The logic underlying the design is both simple and intuitive: if a treatment works, we should observe that the outcome of interest improves more rapidly in subjects receiving the treatment than in individuals not receiving the treatment. Conceptually, Difference-in-Differences design can be best thought of as a combination of a before-after comparison and a comparison between treated and untreated individuals (see Box 4 for an illustrative example).

More precisely, the Difference-in-differences estimator compares a first difference – In the treated: the outcome in the treatment period minus the outcome in the period prior to treatment – to a second difference – in the untreated: the outcome in the period when the treated are receiving treatment minus the outcome in the prior period. The treatment effect is then the difference between these two differences [13, 14]. The design therefore requires data from the pre- and post-treatment periods for both the treatment and the control group. This data can be either repeated cross-sectional or longitudinal. In the case of longitudinal data, the design allows adjustment for individual-level fixed effects or other individual-level factors for additional confounding control [14, 73]. Extensions to multiple treatments and time periods are possible [80], and the Difference-in-Differences design also generalizes to cases where it is not time that indicates treatment in a group receiving and another group not receiving the treatment, but another variable.

The advantages of the Difference-in-Differences design are straightforward and become clear in comparison to two designs that are nested within the approach: First, while in the absence of randomization a simple between-group comparison of treated and untreated individuals is likely to suffer from bias due to pre-treatment differences in outcomes, Difference-in-Differences designs allow for a flexible control of any group-level (or individual-level, in the case of longitudinal data) time-invariant confounding factors (same as entity Fixed Effects). Second, the Difference-in-Differences design is also preferable to an uncontrolled before and after comparison of only treated individuals. In the case of uncontrolled before and after studies, it cannot be ruled out that the change between the treatment period and the prior period among the treated is due to time, rather than due to the treatment. The Difference-in-Difference design, in contrast, controls for any time trend in the data that is unrelated to the treatment under study – to the extent that this trend is common to both the treatment and the control group (same as time Fixed Effects).

Underlying assumptions

The key assumption of the Difference-in-Differences design is that in the absence of treatment the average change in the outcome of interest for the treated would have equalled the average change in the outcome for the untreated [81]. This so-called parallel trends assumption is counterfactual because the trend of the outcome in the treated exists and is only observable in the presence of treatment. The parallel trends assumption is
thus ultimately unprovable. The assumption will be violated if individuals select into treatment and control groups based on factors that affect the change in the outcome of interest between the pre- and the post-treatment period. For instance, if eligibility for treatment were based on the values of the outcome in the pre-treatment period, people might alter their behavior before a treatment starts in order to be eligible for the treatment. In such a situation, it is likely that the ability to benefit from treatment differs systematically between treatment and control group and the parallel trends assumption is violated.

It is important to note that non-linear transformations on the outcome - like the ones used to analyze binary outcomes (e.g. logit, probit) - typically violate the parallel trends assumption [82]. In such cases, the “Changes-in-Changes” model developed by Athey and Imbens [83] – a generalization of Difference-in-Differences designs – allows for discrete outcomes using the nonlinear transformations typically used to analyze such variables. In general, in Difference-in-Differences analyses, researchers should make sure to correct for the hierarchical structure of the empirical models. As highlighted in studies by Moulton [84] and Bertrand et al. [85], applying standard statistical models relying on random error distributions, when data are grouped in space or serially correlated in time, can lead to large biases in standard errors and a large number of false positive findings in empirical analysis. All analyses relying on data that are nested within a larger level should explicitly discuss assumed error dependence structures and calculate standard errors that are robust to spatial and temporal dependence [86, 87]

Tests of assumptions

The parallel trends assumption is not directly verifiable. However, several pieces of information and empirical tests can be used to assert the plausibility of this assumption and the credibility of the study overall.

1. The standard test for the parallel trends assumption is to assess whether the assumption holds in the periods preceding the treatment. This test may not always feasible in practice, since it requires data for at least two time periods before the start of the treatment (i.e., the novel practice, program, or policy). If such data are available, difference-in-differences “placebo regressions” should be run on the pre-period to formally test whether treatment assignment identifies differential trends prior to the treatment. Parallel pre-treatment trends are not technically a requirement for the validity of Difference-in-Differences design; observing differential trends prior to treatment launch does however substantially weaken the plausibility that trends absent of the treatment would have been parallel. Loosening the parallel trends assumption, one can control for different trends in intervention and control groups when data from two periods before intervention start are available. In this case, the parallel trends assumption is replaced by the assumption that the different trends in the treatment and control groups observed prior to treatment stayed constant over the treatment period.

2. In addition to examining pre-treatment trends, several types of additional analyses can be used to test the robustness and credibility of Difference-in-Differences findings

- Multiple control groups: If findings remain similar when the main analysis is rerun with a different control group, the credibility of the findings from a Difference-in-
Differences analysis is strengthened. For instance, if the main analysis used as the control group the districts in the same province where the intervention districts are located, rerunning the analysis using districts from other provinces as control group can serve as a robustness check of the findings.

- Multiple outcomes: The main analysis can be re-run with outcomes that – according to theory or prior empirical findings – should not be affected by the treatment under investigation. If significant effects on such alternate or “placebo” outcomes are detected in the analysis, the credibility of significant treatment effects on the original outcomes is diminished.

- Treatment reversals: Just like the introduction of a program or policy can be used to establish causal effects in Difference-in-Differences analysis, so can the termination of the program or policy. The special cases, where the analyst has access to data spanning pre-introduction, post-introduction, and post-termination time periods, allow analysis and comparison of treatment effects from these two perspectives – see, for instance, an analysis of the causal effect of user fees on health services utilization in Malawi [88]. In general, the credibility of findings will be strengthened if introduction and termination effects are opposite in direction and similar in magnitude.

**Fixed Effects designs**

*Conceptual introduction*

Fixed Effects designs are a generalization of Difference-in-Differences designs. Fixed Effects utilize that individuals are measured under different treatment statuses but are nested within a larger level. This nesting level is common to some or all individuals such as belonging to some common entity, e.g. being the same person, living in the same geographic area or family, or being observed at the same point in time. Fixed Effects allow to control for all observed and unobserved factors that are common to all individuals belonging to the same entity. For instance, individual-level fixed effects in longitudinal analysis control for all observed and unobserved individual-level factors that do not vary over time [89]; family fixed effects in longitudinal and cross-sectional analysis control for all factors that are shared by siblings, such as having the same mother [90] (see **Box 5** for an illustrative example); and time fixed effects in longitudinal analysis control for all time-varying observed and unobserved factors that are common to all individuals at the time of observation (e.g. changes in national policy or interest rates). The Fixed Effects methodology can be applied to any situation in which observations with different treatment statuses are nested within a larger level.

*Underlying assumptions*

The primary identifying assumption for Fixed Effects designs is that treatment assignment is as good as random after controlling for the fixed effects and other covariates. In other words, although the fixed effects control for some unobserved factors – and thus substantially strengthen the causal inference in many cases – they can never control for all unobserved factors that could confound the relationship between a treatment and an outcome. For instance, when controlling for individual-level fixed effects in longitudinal analysis, unobserved time-invariant individual factors are controlled for but unobserved time-varying individual factors are not.
Tests of assumptions

Because Fixed Effects designs are generalizations and extensions of Difference-in-Differences designs, the same approaches can theoretically be used to test the underlying assumptions. However, such tests become harder to execute as the complexity of the design increases. For instance, the equivalent to the pre-treatment parallel trends test for Difference-in-Differences designs is the introduction of pre-treatment variables in longitudinal analysis with entity fixed effects – i.e., dummy variables for one time period before treatment, two time periods before treatment, and so on. If these dummy variables are insignificant, the plausibility of the “no time-varying confounding” assumption and the credibility of the findings is strengthened. Similarly, just like in the case of Difference-in-Differences designs the analyses can be re-run with multiple outcomes that – according to theory or prior empirical findings – should not be affected by the treatment of interest. If such “placebo outcome” analyses fail to detect significant effects, the credibility of the main findings in fixed effects analyses can be strengthened.

Discussion

Quasi-experiments have the potential to provide evidence with high internal and high external validity. However, each quasi-experimental design has its own suite of assumptions that must be met for valid causal inference. The major class distinction between quasi-experiments and non-experiments is that the strong unconfoundedness assumption of the latter is replaced by other assumptions (Instrumental Variable, Regression Discontinuity, and Interrupted Time Series designs) or substantially weaker unconfoundedness assumptions (Difference-in-Differences and Fixed Effects Designs).

In this context, it is important to note that the theory underlying this article, and most experiments and quasi-experiments carried out today – the Rubin causal model (also known as counterfactual or potential outcomes model) [91, 92] – does itself depend on an assumption, the stable unit treatment value assumption (SUTVA): the value of an outcome for a unit exposed to the treatment of interest is the same irrespective of the mechanism used to assign the treatment and independent of the treatments that the other units in the population receive [93, 94]. Especially the second part of SUTVA is a strong assumption because it implies that there are no treatment spillover effects between units. However, many treatments effects can spread across units – infectious disease treatments can affect people within transmission networks, behavior changes can spread within social networks, and treating some people within a population (e.g., girls with a voucher to stay in school to prevent “sugar daddy” relationships and HIV transmission) can make other people within the same population more vulnerable (e.g., girls who did not receive voucher may become more likely to engage in “sugar daddy” relationships as a result of the voucher intervention [95]). Researchers designing and analyzing both quasi-experiments and experiments should thus also consider and discuss whether SUTVA may be violated in a particular study and what the likely implications of such violations are [96]. Experiments can be designed allow estimation of both direct and spillover effects [97] and recent literature suggests approaches to estimate causal effects when SUTVA is violated in quasi-experiments [98-101].

The promise of quasi-experiments lies in the fact their assumptions can be weaker and more easily testable than the strong unconfoundedness assumptions of non-experimental studies. This article provides an introductory overview of the different assumptions
underlying different quasi-experimental designs and approaches to test these assumptions. While quasi-experiments promise major knowledge gains about causal effects of health practice, programs and policies, this promise will only be fulfilled if analysts carefully consider the assumptions underlying any particular quasi-experiment, test the assumptions in as far as possible, and report the results of assumption tests in their publications.
Box 1. Instrumental Variables design: illustrative example

A classical example of an instrumental variable is the Vietnam Era Draft Lottery number for military service in the United States. Researchers were interested in studying the effects of military service on civilian outcomes of men after completing military service. One concern was that men who select into military service have different (unobserved) characteristics than men who do not and thus ordinary least squares regression results with military service as an explanatory variable will be biased due to endogeneity. Between 1970 and 1972 men in the United States were quasi-randomly assigned a lottery number based on their birthdates, which determined their probability to be part of the military draft. Hearst, Newman and Hulley [102] used this exogenous variation to estimate the effect of military service on subsequent mortality. Similarly, Angrist [103] used the draft lottery number as an instrument to study the effect of military service on subsequent earnings outcomes. The idea is that the draft lottery explained considerable variation in military service (relevance condition) but that the lottery had no other effect on measured outcomes (exclusion restriction) and that there were no defiers, who would not have served had they received a lottery number that would lead to a call to serve but who would served had they received a lottery that would not lead to a call to serve (monotonicity). Despite the overall strong case for the exclusion restriction in the case of the draft lottery, violations of this assumption cannot be completely ruled out. For instance, because students were exempted from military service it is possible that men who received lottery numbers indicating a high probability of a call to serve stayed in school longer to avoid the draft. In this case, the exclusion restriction would likely be violated, because schooling can affect both mortality and earnings. Angrist, Imbens and Rubin [18] discuss this particular example in some detail, including the extent to which violations of assumptions may have affected the Instrumental Variables estimates.

Box 2. Regression Discontinuity design: illustrative example

A good example of a study using the Regression Discontinuity design to estimate the causal effect of healthcare is a 2010 study by Almond et al. [48]. The researchers used the clinical threshold for the condition "low birth weight" (1500 grams) to estimate the effect, and value, of low-birthweight interventions on infant mortality and healthcare expenditures. They find that low-birthweight interventions cause a one percentage point decrease in infant mortality and induce additional healthcare expenditures of US$4,000, implying the the cost of saving a statistical life of a newborn with birthweight near 1500 grams is around US$550,000. The researchers argue convincingly that [f]rom an empirical perspective, birth weight-based thresholds provide an attractive basis for a regression discontinuity design for several reasons. First, they are unlikely to represent breaks in underlying health risk. A 1985 Institute of Medicine report, for example, notes: "... designation of very low birth weight infants as those weighing 1,500 grams or less reflected convention rather than biologic criteria." Second, it is generally agreed that birth weight cannot be predicted in advance of delivery with the accuracy needed to change (via birth timing) the classification of a newborn from being just above 1500 grams to being just below 1500 grams. Thus, although we empirically investigate our assumption that the position of a newborn just above 1500 grams relative to just below 1500 grams is "as good as random," the medical literature also suggests this assumption is reasonable" [48]. Additionally, they demonstrate visually the continuity in treatment assignment, which showed no bunching at the threshold, and carry out McCrary’s manipulation test [60]. Moreover, to test the continuity assumption the authors carried out a number of balance
tests. Finally, they measured the effect of being below 1500 grams on external causes of death (e.g., accidents), because these causes are unlikely to be affected by low-birthweight interventions. As expected, they find no effect of the interventions on external causes of death, lending further support for the validity of the Regression Discontinuity design in this study.

**Box 3. Interrupted Time Series design: illustrative example**

An example of a study utilizing the Interrupted Time Series design is research by Gorman et al. estimating the effect of the implementation of state laws related to medical cannabis use on cannabis consumption [104]. The researchers used the implementation of medical cannabis use laws in cities in California, Oregon, and Colorado at a specific point in time, and compared outcomes during the time period prior to the implementation of the law to those following implementation. The authors analyzed a nearly ten-year period of cannabis use and found no association between the introduction of the law and use of cannabis. These models accounted for autocorrelation. Although the authors discuss issues with generalizability arising from their specific data sources, there was no discussion of co-occurring time trends that may affect cannabis consumption.

**Box 4. Difference-in-Differences design: illustrative example**

A good example of a study utilizing the Difference-in-Differences design is research by Baranov et al. [105]. The researchers investigated whether the provision of HIV treatment affected mortality risk, mental health, and labor supply of HIV-negative individuals in sub-Saharan Africa. This Difference-in-Differences study compares people living near and far healthcare facilities where HIV treatment was available (as a proxy of HIV treatment exposure status among HIV-negative individuals), before and after treatment became available. The researchers "test the validity of the parallel trends assumption in two ways. First, we use data from 2004 and 2006 to test for pre-intervention trends in work time and available demographic characteristics. We cannot test for differential pre-trends for mental health and subjective mortality risk because these variables are only available in one pre-intervention round. Secondly, we investigate concurrent trends in thirteen measures of economic shocks and support. [105]"

**Box 5. Fixed Effects design: illustrative example**

An example of how a Fixed Effects design can help to identify causal effects of healthcare interventions is a study by Anekwe et al. [90], who investigated the relationship between childhood measles vaccination educational attainment. Given that measles may lead to complications ranging from undernutrition to blindness and brain damage, the authors expected a positive effect of receiving the measles vaccine on educational attainment. To test this hypothesis, they used a longitudinal population-based dataset, including measles vaccination status and educational attainment 4783 South-African children born between 1995 and 2000. They controlled for a number of observed sibling-varying factors, such as age, birth cohort, mother’s age at child’s birth, and birth order. Because many factors that were not observed in this population cohort are plausible confounders of the relationship between measles vaccination and educational attainment – such as attitude toward risk, conscientiousness, and aspirations for children – the researchers additionally controlled for mother fixed effects. Control for mother fixed effects substantially strengthens the causal inference in this study (measles vaccination increases school grade attainment by
an average of about 0.2 grades), because the biases due to all unobserved confounders that are sibling-invariant are removed by the mother fixed effects.
References


89. Bor, J., et al., *In a study of a population cohort in South Africa, HIV patients on antiretrovirals had nearly full recovery of employment.* Health Aff (Millwood), 2012. 31(7): p. 1459-69.


