

Causal Evaluation and Redesign of Anonymized Data

Supervisors : Maryline Laurent (TSP), Chi Tran (TSP), Louis Philippe Sondeck (Clever Identity)

1 Background and Motivation

From prediction to causal questions

Data are often shared so that other researchers can build predictive models or evaluate interventions. These two goals answer different questions. A predictive question asks, for example, “Which individuals are likely to find a job?” A causal question asks “Would participating in a training programme change an individual’s chance of finding a job?”

An *intervention*, also called a *treatment*, is an action whose effect we want to study. It can be a training programme, a medical treatment, a scholarship, or a public policy. The *outcome* is the result that may be affected by this intervention, such as employment, recovery, graduation, or income.

Prediction alone cannot answer a causal question. A model may accurately predict who will find a job by using age, education, and previous experience. However, this does not show whether a training programme *causes* an improvement. People who choose to participate may already differ from those who do not participate.

What is a causal effect?

Let $T_i = 1$ mean that individual i receives an intervention and let $T_i = 0$ mean that the individual does not. We imagine two possible outcomes for the same individual:

- $Y_i(1)$: the outcome if the individual receives the intervention;
- $Y_i(0)$: the outcome if the same individual does not receive it.

These are called *potential outcomes*. The individual causal effect is

$$Y_i(1) - Y_i(0).$$

The difficulty is that we can observe only one of these outcomes. A person either participates in the programme or does not; we cannot observe both possibilities for the same person at the same time. This is the fundamental problem of causal inference.

A common target is therefore the *Average Treatment Effect* (ATE):

$$\tau_{\text{ATE}} = \mathbb{E}[Y(1) - Y(0)].$$

The ATE is the average difference between the outcome that would occur if the population received the intervention and the outcome that would occur if the same population did not receive it.

For example, let $Y = 1$ mean that a person finds a job within six months. Suppose the probability of employment would be 70% if everyone received training and 50% if the same population did not receive training. Then

$$\tau_{\text{ATE}} = 0.70 - 0.50 = 0.20.$$

The programme increases the probability of employment by an average of 20 *percentage points*.

The effect may differ across groups. The *Conditional Average Treatment Effect* (CATE) for people with characteristics $X = x$ is

$$\tau_{\text{CATE}}(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x].$$

For example, training may help people with little work experience more than experienced workers. The main project will focus on ATE; CATE will be an optional extension.

Why can a simple comparison be misleading?

In an observational dataset, treatment is not assigned randomly. A *confounder* is a variable that influences both treatment participation and the outcome. For example, age may affect both the decision to join a training programme and the probability of finding a job. Comparing all treated individuals with all untreated individuals without accounting for age may therefore confuse the effect of age with the effect of training.

Consider the following hypothetical data. Age is assumed to be the only confounder, and individuals of the same age group are assumed to be sufficiently comparable.

Age group	Training: employed/total	No training: employed/total
20–29	18/20 = 90%	64/80 = 80%
50–59	32/80 = 40%	6/20 = 30%

Within each age group, trained individuals have an employment rate that is 10 percentage points higher. Under the assumptions of this simple example, the estimated ATE after adjusting for age is therefore +10 percentage points.

If the two age groups are merged, the analyst observes only

$$\Pr_n(Y = 1 \mid T = 1) = \frac{50}{100} = 0.50, \quad \Pr_n(Y = 1 \mid T = 0) = \frac{70}{100} = 0.70.$$

The unadjusted difference is now –20 percentage points. This does not mean that training is harmful. The treated group contains many more older people, who have a lower employment rate in this example. The negative difference is an *association* created by the unequal composition of the groups, not a valid estimate of the causal effect.

Why does anonymization create a causal problem?

Before releasing personal data, an organization may apply *anonymization*. Attributes such as age, education, postcode, and occupation are often *quasi-identifiers* (QIs): they may not directly identify a person, but their combination can make a person recognizable. Anonymization may therefore replace exact values with broader categories, for example replacing age 27 with the interval 20–39, or replacing a postcode with a region.

The same variables may also be necessary for causal inference. If anonymization merges the two age groups in the previous example into the interval 20–59, the released table no longer contains the information needed to adjust for age. The treatment and outcome values have not changed, but the estimated causal effect can change sign because an important confounder has been hidden.

The same problem appears in other domains. For example, a hospital may use observational data to estimate whether a new treatment improves recovery. Older patients may be more likely to receive the treatment and, independently, less likely to recover. Age is then both a QI and a confounder. If exact ages are replaced with broad intervals, the analyst may no longer be able to separate the effect of age from the effect of the treatment. The estimated ATE may consequently be biased, just as in the employment example.

Anonymization may also hide differences in treatment effects across subgroups. For example, a scholarship may benefit low-income students more than high-income students, but this difference can no longer be estimated if family income is suppressed. Preserving such subgroup-specific effects, represented by CATE, is an optional extension; the main project focuses on ATE.

These examples illustrate three possible consequences of anonymization: incomplete adjustment for confounding, misleading comparability between treated and untreated groups, and loss of heterogeneous effects. A released dataset may still support accurate prediction while no longer supporting a valid causal conclusion. Consequently, predictive accuracy and conventional information-loss measures are not sufficient measures of causal utility.

The central research question is:

Does an existing anonymized dataset retain enough information to estimate the causal effect of a specified intervention reliably? If it does not, can an alternative anonymization be constructed that satisfies the same privacy requirements while preserving causal information better?

The project therefore has two stages. It first evaluates the causal utility of a dataset that has already been anonymized. Only when this release loses too much information for a reliable causal conclusion does the project search for a better anonymization of the original data. In the second stage, privacy requirements remain mandatory constraints, while preservation of causal information guides the choice of the new release. Predictive performance and conventional information loss are secondary evaluation criteria.

2 Problem Definition

Input to the problem

The main input is an existing anonymized table D^a that researchers intend to use for causal analysis. When available, the corresponding original table D and the anonymization rules used to create D^a are also provided. Access to D is required only for direct comparison and for constructing an alternative release; an analyst who receives only D^a can still perform a more limited sensitivity or partial-identification analysis.

Each row of the data contains:

- **pre-treatment characteristics** X , such as age, education, work experience, or initial health status;
- **treatment** T , with $T = 1$ when the intervention is received and $T = 0$ otherwise;
- **outcome** Y , such as employment or recovery; and
- **a sensitive attribute** S that must be protected.

The provider also specifies the QIs, the sensitive attribute, the required values of k and l , and the causal question. If an alternative release may be constructed, the provider additionally specifies the allowed transformations. Some characteristics may play two roles. For example, age can be a QI because it contributes to re-identification and a confounder because it influences both treatment and outcome.

The initial project will consider a restricted setting in which anonymization replaces QI values with intervals or broader categories. It will not delete rows or change treatment, outcome, or sensitive values. These restrictions keep the first version understandable and make it possible to isolate the effect of QI generalization.

Stage 1: evaluating the existing anonymized table

The first task is to determine whether D^a is suitable for the stated causal question. The method checks whether the release satisfies the required privacy criteria and whether the generalized variables still support meaningful comparisons between treated and untreated individuals. In particular, it examines whether important confounders have been generalized so broadly that treated and untreated individuals who look similar in D^a may have been substantially different in D .

When the original table is available, the method compares the ATE estimate, its direction, and its uncertainty before and after anonymization. This comparison measures the stability of the analysis; it does not prove that the estimate from the original table is the true causal effect. On simulated or semi-synthetic data, where the true ATE is known, causal error can be measured directly. On real data, the evaluation must also use sensitivity analysis and, when appropriate, a range

$$[\tau_{\min}, \tau_{\max}]$$

of effects compatible with the released information and stated assumptions.

The existing release is considered adequate only if it satisfies privacy and meets clearly defined causal-utility criteria, such as preserving the sign of the effect, keeping the estimated change or error below a threshold, and avoiding an uninformatively wide range of possible effects.

Stage 2: proposing an alternative anonymization

If D^a does not meet the causal-utility criteria and the original table is available, the method generates alternative anonymizations. For example, age could be generalized to 10-year intervals, 20-year intervals, or one broad interval. Each choice protects privacy differently and removes a different amount of confounding information.

Rows with the same released QI values form an *equivalence class*. Every candidate release must satisfy:

- **k -anonymity:** each equivalence class contains at least k individuals; and
- **distinct l -diversity:** each equivalence class contains at least l different sensitive values.

Candidates that do not satisfy both conditions are discarded. These criteria address specific disclosure risks and do not protect against every possible attacker. Among the remaining candidates, a release is preferable when it better preserves the direction, magnitude, and uncertainty of the intervention effect.

Possible approaches for constructing an alternative release

The project does not prescribe one solution technique for Stage 2. The student should investigate and justify a method for constructing an alternative release that satisfies the privacy requirements and improves causal utility. Possible approaches include mathematical optimization, clustering guided by confounders or propensity scores, greedy or heuristic search, causal-aware generalization hierarchies, and multi-objective search. Other well-motivated approaches are also permitted.

One possible formulation is to search for a transformation with small causal loss among the allowed privacy-compliant transformations. However, an exact optimization model is only one option and is not a mandatory project requirement. Regardless of the technique chosen, the proposed method must define how candidate transformations are created, how privacy is enforced, and how improvement in causal utility is measured.

Defining causal utility is itself an important research component. A simple starting measure is the change in the estimated ATE between the original and anonymized tables. However,

matching these two estimates does not prove causal validity because the original estimate may also be biased. Stronger alternatives include an estimated upper bound on causal error or the width of the compatible ATE range. Predictive loss and conventional information loss may be used as secondary criteria, but they must not replace the causal evaluation.

Required output

The method should return:

1. a privacy-compliance report for the existing anonymized table D^a ;
2. an ATE estimate or a justified range of possible ATE values from D^a ;
3. causal-preservation measurements and a clear conclusion about whether D^a is adequate for the stated causal analysis;
4. a diagnosis of which generalized confounders are responsible for important information loss;
5. if D^a is inadequate and redesign is possible, an alternative table D^* satisfying the specified k -anonymity and l -diversity requirements, together with its generalization rules and evidence of improved causal utility; and
6. a clear warning when causal reliability cannot be established or no allowed privacy-compliant alternative provides sufficient causal utility.

The problem can therefore be summarized as follows:

Given an anonymized dataset and a causal question, evaluate whether the released information supports reliable estimation of the ATE. If it does not, and the original data are available, find a privacy-compliant alternative that preserves causal information better.

3 Objectives and Requirements

The main objective is to develop a framework that first evaluates whether an existing anonymized dataset remains suitable for causal-effect estimation and, when necessary, proposes a better privacy-compliant anonymization. The project will proceed through the following tasks.

1. **Formalize the causal target.** Specify treatment, outcome, pre-treatment covariates, QIs, sensitive attributes, and the ATE to be studied. Clearly state the causal assumptions.
2. **Evaluate the existing release.** Verify its privacy properties and assess whether generalized confounders still permit meaningful comparisons between treated and untreated individuals.
3. **Measure causal preservation.** Compare ATE estimates when an original reference is available, examine effect direction and uncertainty, and use true causal error on simulated or semi-synthetic data. Develop sensitivity or bounding procedures for real data.
4. **Diagnose causal information loss.** Identify which generalized or suppressed confounders make the causal conclusion unstable, biased, or insufficiently informative.
5. **Propose an alternative anonymization method.** If the existing release is inadequate, investigate a method for constructing a privacy-compliant alternative with improved causal utility. The solution technique is left open.

6. **Evaluate the complete framework.** Measure ATE bias and absolute error, coverage and width of uncertainty regions, whether the sign is correctly determined, and whether the proposed alternative improves on the initial anonymized table.

The initial mandatory scope is deliberately limited to binary treatment, ATE, a small number of discrete or interval-coarsened confounders, and generalization without row deletion. CATE, multiple treatments, continuous treatments, and outcome-dependent anonymization may be studied as extensions.

Release-selection baselines will include methods guided only by conventional information loss, methods guided only by predictive utility, and a privacy method that does not explicitly consider causal effects. Causal-analysis baselines will include naive adjustment using released categories, midpoint replacement for numerical intervals when meaningful, and estimation from the original table as a reference. The contribution will be judged by causal preservation under the same privacy requirements, not by the number of QI patterns repaired.

4 Experimental Datasets

Simulated data

The primary experiments will use simulated data because the data-generating process can provide both potential outcomes, and therefore the true ATE is known. The simulations will vary:

- the strength of confounding;
- the degree of overlap between treated and untreated groups;
- the width of generalized intervals;
- the values of k and l ;
- treatment-effect heterogeneity;
- sensitive-attribute imbalance; and
- sample size.

For each simulated dataset, one or more anonymized tables will first be treated as existing releases and evaluated by the proposed framework. Some releases will deliberately preserve important confounders, while others will generalize them strongly. Because the true ATE is known, the experiment can test whether the framework correctly accepts adequate releases, detects problematic ones, and, when needed, constructs an alternative with lower causal error than the initial release and the baselines.

Semi-synthetic data

Public tabular datasets, such as Adult, may provide realistic covariates. Treatment assignment and potential outcomes will then be generated by a known simulation model. This creates a semi-synthetic benchmark with realistic feature distributions and a known causal effect. The candidate anonymizations will be compared under identical privacy settings. The simulated treatment effect must not be presented as a real causal effect contained in the original public dataset.

Real observational data

A real dataset may be included only if it has a meaningful treatment, outcome, and defensible set of pre-treatment adjustment variables. Because the true causal effect is normally unknown, these experiments can compare the releases selected by different anonymization objectives and evaluate estimate stability and sensitivity, but they cannot establish causal accuracy by themselves.

Expected outcomes

The expected results are: (i) a framework for evaluating whether an existing anonymized table supports reliable causal analysis; (ii) a diagnostic procedure for identifying causal information lost through anonymization; (iii) an ATE estimation or bounding procedure for the evaluated release; (iv) a student-proposed method for constructing or selecting an alternative release satisfying k -anonymity and distinct l -diversity when the existing release is inadequate; and (v) experiments showing whether the proposed alternative better preserves the direction and magnitude of an intervention effect.

The central contribution is causal evaluation of anonymized data, followed by causal-aware redesign only when necessary. The redesign technique is an open research choice to be developed and justified by the student, while prediction and conventional information loss provide complementary comparisons.