All Actuaries Summit 2026
25 – 27 May 2026, Melbourne

From Data Taker to Data Generator: Introduction to Causal Inference and Randomised Control Trial Playbook

Prepared by Laura Zhao and Fei Huang

Presented to the Actuaries Institute
2026 All-Actuaries Summit
25-27 May 2026

Actuaries Institute
ABN 69 000 423 656
Level 34, Australia Square, 264 George Street, Sydney NSW 2000
P +61 (0) 2 9239 6100 | actuaries.asn.au

Abstract

As actuaries, we rely on observational data to build models, understand historical patterns and tell stories. Those datasets show us *what* has happened, but rarely tell us *why* things happen. The *why* matters because it pinpoints to action and consequence, it enables decision makers to understand causal effects of changes rather than being blinded by correlations, and it provides robust policy evaluations.

Understanding the *why*, or the cause and effect is challenging, there are techniques and methods to illuminate the hidden causal relationship, randomised controlled trial (RCT) is widely considered as the gold standard for uncovering causality, as it directly addresses the problem of endogeneity by ensuring treatment assignment is exogenous, while other causal methods such as differences-in-differences, regression discontinuity, or instrumental variables arguably suffer from the endogeneity problem to various degrees.

RCT, as a method of field experiment, generates meaningful data, sharper insights and business guidance beyond what observational data can provide. It adds to the actuary's toolkit and redefines our understanding of *why* things happen. This paper will introduce causal inference, the endogeneity problem, an RCT playbook with practical reflections, and a case study, which hopefully will inspire more actuaries to explore and embrace the spirit of experimentation.

Keywords: causal inference; randomised controlled trial; field experiment.

1. Introduction

Actuaries have historically focused on prediction problems, from liability valuations to capital and risk modelling, quantifying uncertainties in long-term financial promises, supporting regulatory supervision and solvency oversight, and translating analytical results to business recommendations. However, predictive accuracy alone is often not enough for decision making. It does not isolate underlying causes of observations due to confounding factors, nor does it guarantee an optimal recommendation due to the absence of causal insights into actions. While forecasts and statistical association-based modelling play an important role in informing decisions, much of an organisation's value is ultimately realised through actions.

To illustrate, identifying the customer segment of the lowest retention rate may provide useful descriptive insights, but it does not necessarily imply this segment will respond effectively to targeted retention campaigns (Ascarza, 2018). From a business perspective, a more value-adding exercise is to target the segment most responsive to churn management interventions, enabling more efficient resource allocations and maximising the impacts of retention strategies. The shift from prediction optimisation to decision optimisation expands analytical toolkit including causal inference techniques, in order to understand causal consequences of business actions and guide effective decisions.

This paper introduces some fundamental concepts of causal inference and RCT in section 2, and presents a playbook for RCT, covering pre-experiment, implementation and post-experiment stages, incorporating practical reflections and a case study in section 3, provides a brief overview of alternative causal inference methods in section 4, and concludes with key takeaways for actuaries in section 5.

2. Causal Inference

Causal inference is a field that leverages both theory and domain knowledge to estimate the impact of events and choices on potential outcomes of interest. There is a key distinction between correlation and causation, as a causal effect is not only about two features being related but also a change in one feature that will directly result in a change in the other feature. Correlations in observational (non-experimental) data rarely reveal causal relationships. Identifying causal effects typically requires exogenous variation. But in real-world settings, individuals make decisions optimally based on available information, rendering choices endogenous. As a result, observed correlations between choices and outcomes generally do not reflect causal effects (Cunningham, 2021).

2.1 Endogeneity problem

In an observational setting, naive comparisons of outcomes between treated and untreated groups are typically biased due to confounding, selection, time trends, and behavioural responses. These biases arise from endogeneity, where explanatory variables are correlated with unobserved determinants of the outcome. As a result, estimated relationships cannot be interpreted as causal. Below are some examples.

1. *Omitted variables.* Unobserved factors that affect both the treatment and the outcome can bias estimated relationships. For example, higher sales may coincide with increased marketing spend, but failing to control for seasonality may lead to overstating the causal effect of marketing.

2. *Treatment selection.* When treatment assignment is not random, individuals receiving the treatment may systematically differ from those who do not. For example, younger customers

may be more likely to receive discounts, leading to biased estimates of price elasticity if age is not properly accounted for.

3. *Simultaneity* (*reverse causality).* Treatment and outcome may be jointly determined, making it unclear which causes which. For example, a popular product attracts more influencer promotion, while influencer promotion simultaneously increases product popularity, leading to upward-biased estimates.

4. *Measurement error.* Errors in measuring variables can distort estimated relationships, particularly when measurement error is systematic rather than random. For example, home insurance sum insured might be underestimated for older buildings due to lack of understanding of higher repair costs associated, sum insured measurement error in this case is correlated with the age of the building.

5. *Sample selection.* The observed sample may not be representative of the target population, leading to biased inference. For example, if data disproportionately represent a specific customer segment with different risk profiles, estimated effects may not generalise to the broader population.

Ignoring endogeneity can lead to biased and misleading estimates of causal effects, ultimately resulting in suboptimal or incorrect decision-making. Addressing endogeneity is therefore central to any credible causal analysis. Randomised controlled trials address these issues by breaking the link between treatment assignment and unobserved factors, thereby eliminating endogeneity by design.

2.2 Mathematical Framework of RCT

Randomised controlled trial (RCT) is widely regarded as the gold standard for causal inference because random assignment ensures that treatment is exogenous. In other words, both observed and unobserved factors affecting outcomes are on average, balanced between treatment and control groups (Gerber and Green 2012).

Causal inference is about understanding *counterfactual* comparisons and comparing potential outcomes. If we define Y as an outcome variable, X as a treatment variable (binary), the causal question is “does X cause Y?”. Mathematically, a causal effect τ of the treatment can be expressed as follows:

$$\tau_i = Y_i(1) - Y_i(0),$$

where i refers to the subject i, $Y_i(1)$ represents the potential outcome of subject i being treated, $Y_i(0)$ represents the potential outcome of subject i if untreated.

If the subject i is treated, then $Y_i(1)$ is the observed outcome, and $Y_i(0)$ is the counterfactual (unobservable) outcome. Only one of these two potential outcomes is observed for each subject i, giving rise to the fundamental problem of causal inference.

To address this, causal effects are typically defined at the population level. Under RCT setting, customers are randomly allocated into control and treatment groups. RCT estimates causal effects on average, which is called the *average treatment effect* (ATE):

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

When subjects are randomly assigned to treatment group, a comparison of average outcomes in treatment and control groups is an unbiased estimator of the ATE, because all subjects have the

same probability of receiving the treatment, the subjects that are randomly selected for the treatment is a random subset of all subjects, therefore, the expected value of $Y_i(1)$ potential outcome of treated subjects is the same as the expected value of $Y_i(1)$ potential outcome of all subjects:

$$E[Y_i(1) \mid D_i = 1] = E[Y_i(1)],$$

where D_i indicates the treatment status of subject i.

Since the assignment selection is random, the untreated subjects group (control group) is also a random subset of all subjects, the expected value of $Y_i(1)$, the potential outcome of untreated subjects is the same as the expected value of $Y_i(1)$ potential outcome of all subjects:

$$E[Y_i(1) \mid D_i = 0] = E[Y_i(1)].$$

Combining the above equations, under random assignment, the expected potential outcome of treatment and control groups is the same:

$$E[Y_i(1) \mid D_i = 1] = E[Y_i(1) \mid D_i = 0].$$

Similarly, subjects who do not receive the treatment have the same expected potential outcome as the treatment group that would have if untreated:

$$E[Y_i(0) \mid D_i = 0] = E[Y_i(0) \mid D_i = 1] = E[Y_i(0)].$$

Above means the average treatment effect can be derived from the observed outcomes, which is the outcome difference between the treatment group and the control group:

$$ATE = E[Y_i(1)] – E[Y_i(0)] = E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0].$$

Key assumptions for this identification to hold true include:

1. Random assignment in which all subjects are allocated by the same random procedure;
2. Excludability that potential outcomes only respond to treatment and not some other features of the experiment such as the outcome measurement procedure;
3. Non-interference that the treatment status of i has no effect on the outcome of j (where i =/= j), i.e. no spillovers between subjects.

Potential outcome Y_i can be represented in a regression form of d_i (1 if treated, 0 if untreated):

$$Y_i = \alpha + \beta_1 d_i + \epsilon_i,$$

where α represents the expected outcome for untreated subjects when $d_i = 0$, β_1 is the average treatment effect (ATE), the term ϵ_i denotes the residual.

The above form can be extended to include covariates, although it is not required for unbiased estimation of ATE under randomisation, but it can improve precision when they are measured prior to treatment and are predictive of the outcome. X_i. Below is a pre-treatment covariate that is predictive of the outcome, β_2 captures the effect of the covariate on the outcome:

$$Y_i = \alpha + \beta_1 d_i + \beta_2 X_i + \epsilon_i.$$

When treatment effects are expected to vary across subgroups, for example, different age groups may respond to treatment differently, interaction terms between the treatment indicator

and selected pre-treatment covariates may also be included to assess heterogeneous treatment effects:

$$Y_i = \alpha + \beta_1 d_i + \beta_2 X_i + \beta_3 d_i X_i + \epsilon_i,$$

where β_3 captures how the treatment effect varies with X_i.

Ordinary least squares (OLS), which is a common method for estimating the relationship between variables in a linear regression model, will be used in estimating the ATE as the main goal of an RCT. Since one of the key assumptions of OLS is homoscedasticity, which is unlikely to be satisfied in practice, as the potential outcomes are likely to have unexplained non-constant variability by individual observations, the residuals will be computed using the Huber-White sandwich estimator, which is commonly available in statistical packages.

3. RCT Playbook

This playbook provides a comprehensive overview of key components in an RCT journey, encompassing important checkpoints before, during and after an experiment. It highlights both methodological steps and practical reflections that support design, execution and evaluation of RCT aligned with rigorous standards and meaningful business outcomes.

3.1 Check RCT feasibility

Before investing efforts in RCT design, check the following criteria if RCT is well-suited to answer your causal question:

- *Assignable*: subjects can be randomly allocated into control and treatment groups.
- *Traceable*: assignment and exposure to treatment can be captured deterministically.
- *Measurable*: outcomes can be captured and recorded.
- *Containable*: non-spillover or interference between subjects or handled via clustering.
- *Ethical and compliant*: no violation of duty of care, regulatory constraints or unfair treatment or non-treatment rules.

Practical reflections

> Feasibility assessment acts as a gating step. It determines if RCT is the appropriate method or an alternative causal inference method should be considered. In the modern business environment, most operational and technical feasibility criteria are readily met, but ethical and compliance considerations often remain as the primary source of contention. *Early engagements and alignments with key stakeholders are critical to ensure a smooth project initiation*.

3.2 Experiment design

A well-designed RCT is the foundation of credible causal inference, as it sets the scene for decision context and aligns the experiment with objectives. This section provides *key elements* of RCT to include in a proposal document. It is recommended to engage impacted stakeholders and seek inputs at this ideation stage, which is not only for early alignments towards common goals, but also to minimise implementation risk.

Objectives

Objectives can be split into primary and secondary, with a clear priority of hypothesis testing.

Practical reflections

> Often multiple objectives are involved in experiment setting, define primary and secondary goals is to ensure clear decision making, manage trade-offs and maintain analytical focus.

Hypothesis

Statements capturing the core of causal inference, it may take the form of "action A will causally change an outcome relative to a baseline/control condition".

Practical reflections

> Defining a hypothesis is a critical step of design. It establishes the link between stakeholder interests and decision criteria, and sets the scope for data analysis.

Treatment

Actions that apply to the treatment group.

Practical reflections

> Treatment condition can be single- or multi-factor based. A two-factor RCT is more informative than a single-factor RCT, as both main effects and interaction effects can be assessed, and achieve higher efficiency per sample from richer information gained. At the same time, a larger sample size might be required to detect interaction effects, and it will add implementation complexity. If the second factor offers limited marginal gains, it may not justify the increased complexity. A factorial design is recommended when there is a credible interaction hypothesis on multiple treatments.

Control

A baseline condition as a benchmark comparison, and it should be consistent with the treatment group condition, except for the treatment/action.

Practical reflections

> Control condition should also be carefully thought through. It is not simply "no treatment"; it is a counterfactual condition constructed to isolate the causal effect of the treatment while holding all other aspects constant, i.e. *ceteris paribus principle*. For example, if the objective is to test a promotion message treatment while not holding delivery timing or user flow consistent, then the treatment effect is a mixture of messaging, timing and user flow effects. In other words, the control group should be exposed to identical traffic allocation, information presentation and measurement except for the test feature for the treatment.

Treatment versus Control

Define the exact treatment and control condition, specify every difference in conditions as well as common conditions.

Expected effect

Outcomes, including key performance indicators (KPI) and timeline.

Practical reflections

> Considering multiple KPIs for analysis is preferred as business decisions are multi-dimensional rather than depending on one single metric, and certain KPIs are often noisy or underpowered, a broader set of metrics is more useful to detect signals and support decision making.

Type of study

Within-subjects vs between-subjects design refers to whether the same subject is exposed to multiple treatment conditions. In a within-subjects design, each subject experiences more than one treatment condition over time, allowing comparisons within the same individual. In a between-subjects design, each subject is exposed to only one treatment condition, and comparisons are made across different groups of subjects. For example, in an insurance context, a between-subjects design would assign different customers to different pricing or communication strategies, while a within-subjects design would expose the same customer to multiple strategies over time.

Practical reflections

> In most field experiments, a between-subjects design is more common, as it avoids carryover effects and is simpler to implement operationally.

Targeted population

Clearly define inclusions and/or exclusions criteria of subjects that are exposed to the experiment, to be applied in IT implementation and used to filter data for analysis.

Randomisation procedure

Common choices include the followings or a combination of the followings:

- Simple randomisation, each subject is assigned to either control or treatment with a fixed probability like a coin flip 50/50, used in large sample size context.
- Block randomisation, used in small to medium sample size context to ensure size balance between control and treatment groups.
- Stratified randomisation, used when certain covariates are known to strongly predict potential outcomes. Randomisation is performed within strata defined by covariates to ensure mix balance between control and treatment groups.
- Cluster randomisation, used when impractical to administer individual subject level randomisation, and to randomise entire groups of individuals.
- Factorial randomisation, used when there are multiple treatments simultaneously.

Sample size

Estimated as a function of significance level, power, baseline value of interested outcome, minimum detectable effect, and attrition rate[1], which is also used to derive experiment duration considering traffic constraints.

Practical reflections

[1]Attrition rate is defined as the percentage of subjects fails to be treated due to drops off or data loss, e.g. policy cancellation prior to expiry

Sample size estimation at the scoping stage provides a guide to experiment duration, but traffic and/or attrition rate may deviate from expectations. Don't wait until the expected end date of the experiment to check the sample size. A frequent monitoring process should be set up upfront if the required sample size is achieved.

Datasets and variables

List datasets and variables available for analysis.

Analysis plan

Include tools and statistical methods to compare results, and any additional analysis, such as heterogeneity or subgroup analysis.

Practical reflections

In a mature experimentation environment, a data analysis plan should be set up upfront rather than improvised after results arrive, which prevents confirmation bias ex post or "cherry picking" a specific segment that might have a significant result. Pre-built analysis pipelines also allow near real-time interpretations once results arrive.

Timeline

Time period of experiment, including start and end dates.

Practical reflections

Sometimes, due to business feasibility, an experiment may be required to cap at a certain duration, resulting in a smaller than originally required sample size, which should lead to changes in design to reduce the required sample size. The followings are potential changes to experiment design with trade-offs considered:

- If there are multiple treatments, consider reducing the number of variants to increase exposure to each group and focus on primary objective of testing.
- Consider increasing minimal detectable effect (MDE), with trade-off of increasing Type II error (false negatives), a larger MDE bias decisions towards big bets and discourages incremental optimisation culture. Any increases should be justified by business reality, which is interpreted as the smallest true effect that would change decision.
- Consider changing to a different funnel stage of testing during customer journey, for example, if the test is run at get quote stage versus final pay stage, exposure will be larger.

Ethical and compliance

Consider regulatory and legal frameworks, fairness and ethical principles, and auditability of this experiment, which will be used in stakeholder communications.

Risks and limitations

Identify potential risks and limitations of this experiment design, prepare for no significant results and decision flow, for example, a follow-up experiment.

3.3 Pre-implementation Setup

Core activities at this stage include building operational monitoring and data analysis pipelines and seeking approval from relevant stakeholders. RCT approval requires alignments across execution owners, decision authority and risk governance. It is recommended to map out stakeholder groups beforehand and prepare answers to address potential concerns.

Below is an illustrative RACI-style stakeholder mapping, which varies depending on the scope of the experiment and organisation context:

	Actuarial & Data Science	Executive & Finance	Risk, Compliance & Legal	IT	Product & Operation	Marketing & Sales
Experimental design	R	C	C	C	C	C
Approval	R	A	I	I	I	I
Execution & Monitoring	R	I	I	R	I	I
Analysis & Business Interpretation	R	C	I	I	I	I
Decision & Rollout	R	A	C	I	I	I

R: Responsible; A: Accountable; C: Consulted; I: Informed.

3.4 Execution and Monitoring

The operational monitoring report should be regularly checked to ensure the experiment is running as expected. Unlike common actuarial reports for valuation, pricing, or capital, which are produced on a yearly, quarterly, or monthly basis, experiment monitoring reports can be run in real-time or daily to support early detection of operational issues.

3.5 Data Analysis

Results analysis should be pre-defined, including metric definitions, data quality checks, hypothesis testing from standard statistical tests to regression analysis of average treatment effect (ATE) and heterogeneous treatment effect (HTE). Most importantly, analysis should be tied to the decision framework; otherwise it risks leading to analysis paralysis.

Standard statistical tests are the first step. For each outcome variable of interest, a standard hypothesis test is conducted. If the p-value falls below 0.05, the null hypothesis is rejected, indicating that the result is statistically significant at the 5% level.

Three statistical significance testing options are commonly used depending on the potential outcome: the z-test is used for comparing proportions between two independent groups e.g. a binary outcome variable. The chi-square test is used for a categorical association test that can also be used for binary outcome testing, or can be used for larger contingency tables with more than two groups. The t-test is used for comparing means of continuous variables e.g. average claim size that can be of any value.

Regression analysis is the second step of hypothesis testing, and provides the significance level of the average treatment effect, after allowing for covariate controls. In an ideal RCT, randomisation ensures baseline balance across control and treatment groups. Controlling for covariates is not required for an unbiased estimation of the treatment effect, but including

covariates can increase the precision of the causal estimate, especially in large samples. By precision, we mean how narrow the confidence interval is around an estimate. A more precise estimate has a smaller standard error, and higher confidence in where the true causal effect lies. Not all covariates that possibly can be found should be controlled. One should only control the covariates that are collected before the treatment, and can predict the potential outcome in order to explain the variance in Y and reduce the standard error of the causal estimate.

For identifying the average treatment effect under randomisation, non-linear modelling is not required: simple comparisons or linear models already yield unbiased estimates. The case for more flexible specifications arises when the goal shifts to estimating heterogeneous treatment effects, where the treatment effect varies across individuals or subgroups rather than being uniform. Detecting this variation has both research and practical value, since it can reveal which segments respond most strongly to an intervention. In practice, heterogeneity is tested by introducing interaction terms between the treatment indicator and relevant covariates, with the significance and magnitude of these interactions indicating whether and how the effect differs across subgroups.

The choice of model depends on the nature of the outcome variable. When the outcome is continuous, ordinary least squares (OLS) is commonly used and provides a straightforward interpretation of average effects. When the outcome is binary, such as whether a claim occurs, models such as logit and probit are more appropriate because they ensure predicted probabilities lie between 0 and 1.

The logit model assumes that the probability of the outcome follows a logistic function of the covariates. The model result is intuitive and interpretable when expressed as an odds ratio, i.e. taking the exponential of the coefficient. For example, if the coefficient of the treatment flag is 1.2, then the odds ratio is calculated as e^1.2 = 3.32, which means the causal effect of treatment will increase the odds of the outcome rate by 3.32 times of the control group.

The probit model is commonly used when the outcome variable is binary, such as whether a claim occurs. It is derived from a latent variable framework in which an unobserved continuous propensity underlies the binary outcome, with the error term assumed to follow a standard normal distribution. This assumption implies that the probability of the outcome is given by the CDF of the standard normal applied to a linear combination of covariates. The analogous model with a logistic error distribution yields the logit model, with the two typically producing similar results in practice.

Unlike linear models, the coefficients in a probit model do not directly represent changes in probability. Instead, they affect the probability of the outcome through the normal CDF. For a binary treatment indicator, the change in predicted probability associated with treatment can be expressed as:

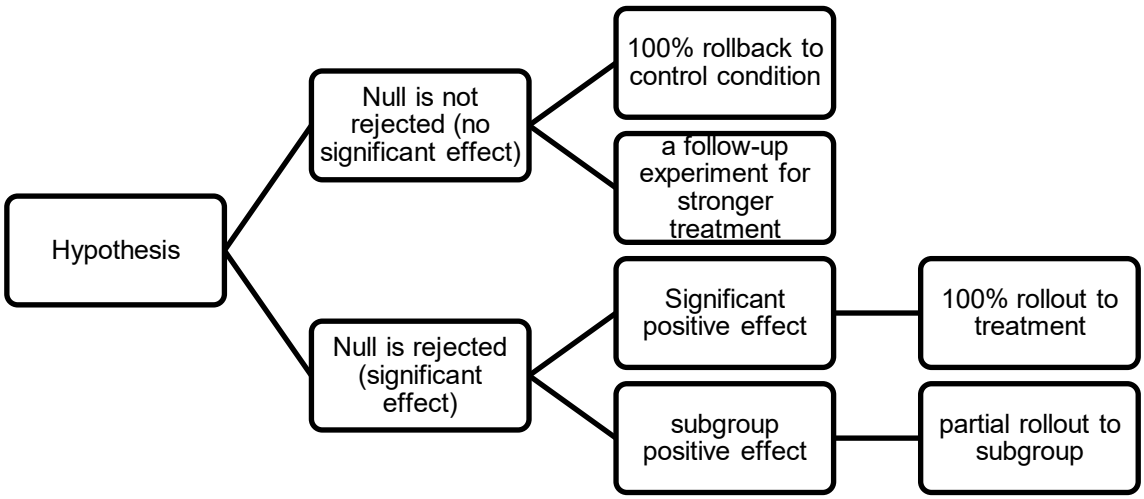
$$\Delta P = \Phi(\beta_0 + \beta_1) - \Phi(\beta_0)$$

For example, when the intercept is -2 and the coefficient of the treatment flag is 0.5, then the change in probability is 4.4%, meaning the treatment increased the probability by 4.4%.

When there is no strong theoretical preference for logit versus probit, fitting both models and assessing consistency in directional signs, marginal effects and significance level can provide a robustness check of results.

3.6 Decisions and Rollout

A decision tree framework, like below, can be applied once results arrive:



In the book *The Voltage Effect*, behavioural economist John A. List explores in depth "the scaling problem", where an idea works effectively in a small and controlled setting but loses its effectiveness when scaling into general public. Before rolling out more broadly, understand the potential changes in conditions, spillover effects, diminishing marginal gains and additional implementation costs at scale helps to prevent overestimation of effects and enable more efficient scaling.

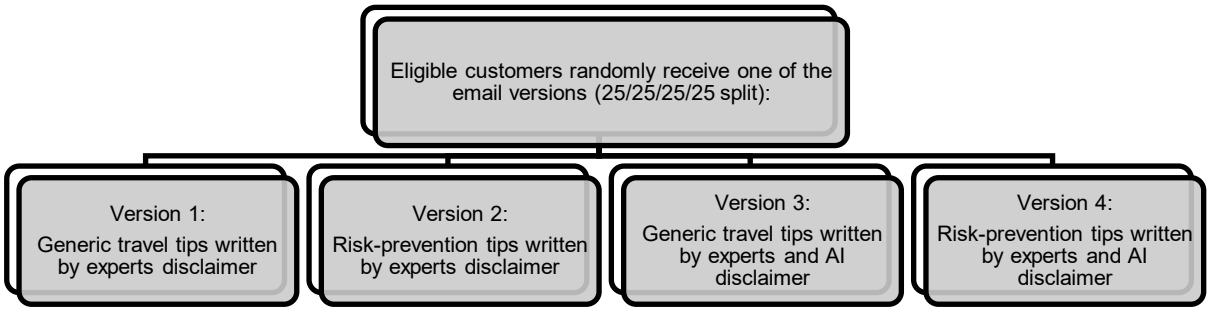
3.7 A Case Study

The Australian travel insurance industry has experienced spikes in gastro claims (e.g. "Bali Belly") from Southeast Asia destinations (News.com.au, 2025), the causes of which are multifaceted. An experiment is designed under this context and an email campaign is launched for travellers' diarrhoea risk awareness to see if informational nudges can improve risk prevention.

A 2x2 factorial randomised controlled trial was designed and implemented. The first treatment consists of a targeted risk-prevention email providing actionable tips on reducing gastro-related risks when traveling to Southeast Asia, testing whether those tips can cause changes in claims behaviours as the primary objective. The second treatment is an AI-generated content disclaimer appended to the top of email, informing recipients that the message was produced by combined insights from experts and AI to test if it may change content engagements as the secondary objective.

Policyholders are randomly assigned to one of four groups and receive the corresponding email on the departure date of their trip. The factorial design allows the estimation of the main and interaction effects of risk-prevention tips and an AI disclaimer on customer engagement and

claims outcomes. The following diagram illustrates email version allocations:



Hypothesis testing and analysis span from claims outcomes to customer engagements, including subgroup analyses such as heterogeneous treatment effects by age. A mix of results has been identified to contribute to business strategies around underwriting and marketing content.

This case study demonstrates how experiments can align cross-functional stakeholders around shared business objectives, and the role actuaries can play in this space.

4. Other Causal Methods

Although a randomised controlled trial is the gold standard for causal inference, it is not always practical due to its resource intensity, nor is it always ethical to implement real-world experiments for treatment, such as funding. Other quasi-experimental designs, including difference-in-differences (DiD), regression discontinuity (RD), and instrumental variables (IV), are also common econometric methods for causal inference.

4.1 Difference-in-Differences

To illustrate the difference-in-differences method, of which causal effect is estimated by calculating changes in potential outcomes over time between control and treatment groups. For example, if a treatment is only implemented for brand A but not to brand B between Jan 2026 to Feb 2026, the causal effect is estimated as follows:

Y: potential outcome	Jan 2026	Feb 2026	Difference
Brand A (Treatment group)	Y1	Y2	Y2-Y1
Brand B (Control group)	Y3	Y4	Y4-Y3
Difference	Y1-Y3	Y2-Y4	**Causal effect: (Y2-Y1) – (Y4-Y3)**

However, for the above causal effect to hold true, it must satisfy the *parallel trends assumption*, a strong assumption that may be unrealistic. Parallel trends assumption in this example means the two brands have similar customers with similar behaviour over time so that the difference arising is purely due to the treatment, but in reality, many external factors could influence mix of customers between brands, e.g. if over the period of treatment, brand B runs a targeted marketing campaign that results in customer profile mix changes, then causal effect suffers from endogeneity problem with differential trends bias.

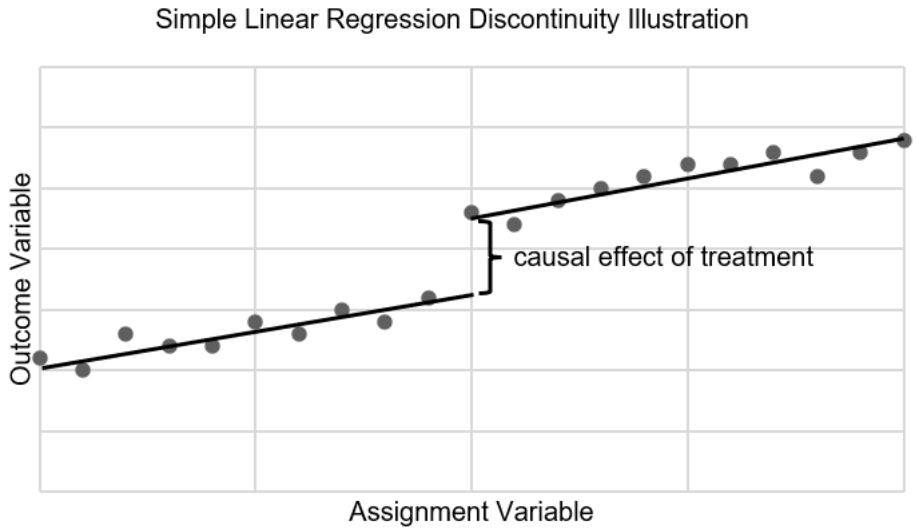
4.2 Regression Discontinuity

The core idea of RD is that any jump or discontinuity in the outcome variable at the cutoff is interpreted as a causal effect of treatment, and the cutoff is defined as the value of an assignment variable where treatment assignment changes. Treatment is implemented at the cutoff point, either through sharp RD design, where the probability of treatment jumps from 0 to 1 when X is above the cutoff, or through fuzzy RD design, allowing for a smaller jump in the probability of assignment to the treatment at the cutoff point, which only needs to satisfy the following:

$$\lim_{\varepsilon \downarrow 0} \Pr(D = 1 | X = c + \varepsilon) \neq \lim_{\varepsilon \uparrow 0} \Pr(D = 1 | X = c + \varepsilon)$$

where D is the treatment variable, 0 meaning untreated and 1 meaning treated, and X is the assignment variable (Lee and Lemieux 2010).

Here is a hypothetical example in a travel insurance context: imagine a treatment only applies to trip duration (assignment variable) above 30 days, and assuming prior to the treatment, medical claims rate (outcome variable) is continuous (in this case linear) for trip duration 29 days to 30 days, then the causal effect of the treatment is claimed to be any significant changes in medical claims rate between trip duration 29 days and 30 days policies, which visually can be represented as below:



Key assumptions of RDD include no manipulations around the cutoff, continuity of potential outcomes and causal effect is only valid for individuals near the cutoff, which is defined as local treatment effect.

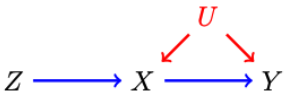
4.3 Instrumental Variables

Instrumental variables (IV) design is the least intuitive method among those quasi-experimental design, and good instruments are weird (Cunningham, 2021). Mathematically, in a potential outcome equation, a valid instrument Z must satisfy two key conditions that Z is not correlated with the error term but is correlated with X:

$$\mathrm{Y_i} = \alpha + \beta * X_i + \epsilon_\mathrm{i} \ ,$$

$$Cov(Z_i, \epsilon_\mathrm{i}) = 0 \ \ and \ \ Cov(Z_i, X_i) \neq 0.$$

Imbens (2014) summarised the relationships as follows, the unobserved component U has a direct effect on both the treatment X and the outcome Y, the instrument Z is not related to the unobserved U, and is only related to the outcome Y through the treatment X:



Four key assumptions of IV are:

1. The instrument is as good as randomly assigned and is independent of all unobserved confounders,
2. the exclusion restriction that the instrument only impacts the outcome through the treatment,
3. the monotonicity of the treatment that it impacts all individuals in the same direction, and
4. the relevance criteria that the instrument is correlated with the treatment.

It is often challenging to find an instrument that satisfies all of the above, and good instruments come from detailed institutional knowledge, rigorous analysis and measurement tailored to a particular context (Angrist and Krueger 2001).

Below summarises methods that have been described in this paper:

Causal Method	**Typical use case**	**Key Strength**	**Key assumptions**	**Limitations**
RCT	Test direct effect of treatment	Minimal confounding, simple causal interpretation	Random assignment; Excludability; Non-interference	Resource intensive; Subject to practical or ethical constraints
DiD	Quasi experiment with time series data	Easy to implement with panel data	Parallel trends	Not allow for time-varying unobserved confounders
RD	A cutoff or threshold based policy evaluation	Credible causal estimates near threshold	No manipulations around the cutoff; Continuity of potential outcomes	Causal effect is only valid for individuals near the cutoff
IV	Observational studies with endogenous treatment	Useful when randomisation is unrealistic	Instrument relevance; Exclusion restriction; Independence from unobserved confounders	Find a valid instrument is challenging; IV estimates the local average treatment effect (LATE), which represents the causal effect for individuals whose treatment status

				is affected by the instrument.

5. Key take-aways for actuaries

Understand the business context

An impactful experiment is grounded on a comprehensive understanding of business dynamics, an appreciation of stakeholders' interests, and alignment with business strategies. It is not only critical for initiating experiments and obtaining approvals, but also for translating results into value added decisions.

Cross discipline collaborations

Moving from prediction problems to decision problems requires even more collaborations across multiple business functions, because business decisions rarely affect a single function; most often, they have cross-functional implications. Partnership with multiple stakeholders diversifies perspectives and reduces the risk of biased decision-making, which strengthens the overall quality of analysis and decisions.

Beyond a single-factor design

Maximising the value of experiments often involves multiple treatments to generate richer insights, especially when justified by clear business decision needs and a favourable cost-complexity trade-off. A 2x2 factorial design is a practical choice and allows for assessing the causal impacts of combined treatments, whose effects can be quite different from simply adding the effects of single treatments. This approach empowers more informed and high-impact decision making.

Treatment effect heterogeneity

Experimental data analysis does not stop at the average treatment effect, heterogeneous treatment effects look into which segments benefit more or less, and inform more targeted treatments and resource allocations that optimise business efficiency, or intentionally non-targeted treatments to uphold fairness and ethical accountability in practice.

Set right expectations

Often, experimental results show either no significant effects or even adverse effects, which could be due to many internal or external factors, e.g. treatment is not strong enough to achieve the expected lift, or customer behavioural dynamics shift. Those results are not failures. They are informative and can be particularly valuable because they update beliefs/priors before scaling to a broader customer base. By mapping these outcomes to decision flows, organisations can proactively adapt strategies and identify opportunities for improvements. In this sense, every experiment contributes valuable insights and reinforces a culture of learning and evidence-based decision-making.

References

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research, 55*(1), 80–98. https://doi.org/10.1509/jmr.16.0163

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *The Journal of Economic Perspectives, 15*(4), 69–85. https://doi.org/10.1257/jep.15.4.69

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press. https://mixtape.scunning.com

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. W.W. Norton & Company.

Imbens, G. (2014). Instrumental Variables: An Econometrician's Perspective. *NBER Working Paper Series*, 19983. https://doi.org/10.3386/w19983

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281

List, J. A. (2022). *The voltage effect: How to make good ideas great and great ideas scale.* Currency.

News.com.au. (2025, March 3). 'Dropping like flies': illness rampant in Bali. https://www.news.com.au/travel/travel-updates/health-safety/dropping-like-flies-why-tourists-are-reconsidering-their-bali-holidays/news-story/334068f899717b0cffb32d5a64287ab1