Using Delivery-Window Measurements to Estimate LATE with Unobserved Treatment Receipt

Gaurav Sood

Technical note

Abstract

A delivery-window measurement can serve as a proxy for otherwise unobserved treatment receipt. In a randomized experiment with one-sided noncompliance, the fraction flagged by a fixed detection rule identifies compliance only when the rule's sensitivity and false-positive rate are known or identified from suitable validation data. This note applies the established misclassification correction to the Wald estimator, derives its joint delta-method variance, and gives sensitivity bounds when detection sensitivity is only bounded. Separate delivery and outcome windows do not make the classifier independent of the outcome. Simulations illustrate calibration uncertainty, serial dependence, and weak detection. The analysis is an application of existing identification and measurement-error results; it does not establish a new test of monotonicity or an empirically validated method for recovering receipt.

1 The measurement problem

Treatment assignment is often recorded more reliably than treatment receipt. A concentrated delivery window may supply additional evidence: successful delivery might produce a known mechanical change in an activity measure. The question is whether that measurement can recover the compliance fraction needed to interpret an assignment effect.

Let $Z_i \in \{0,1\}$ denote randomized assignment, $D_i(z)$ potential receipt, and $Y_i(d)$ a later outcome. Assume consistency, no interference, assignment independent of potential receipt and outcomes, and exclusion: assignment affects the later outcome only through receipt. Impose one-sided noncompliance, $D_i(0) = 0$ for every unit. Then $C_i = D_i(1)$ indicates a complier, $\pi = P(C_i = 1) > 0$, and

$$\Delta = E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0] = \pi\tau, \qquad \tau = E[Y_i(1) - Y_i(0) \mid C_i = 1]. \tag{1}$$

These are the usual LATE assumptions and estimand (Imbens and Angrist, 1994; Angrist et al., 1996). One-sided noncompliance rules out both always-takers and defiers. Its validity must come from the treatment setting; observing a proxy cannot establish it.

Observe a pre-treatment series, a delivery-window series, and a later outcome for each unit. A fixed rule produces a binary receipt flag R_i among assigned units. Define its sensitivity and false-positive rate in that population as

$$s = P(R_i = 1 \mid C_i = 1, Z_i = 1), \qquad f = P(R_i = 1 \mid C_i = 0, Z_i = 1), \qquad s > f. \tag{2}$$

The flag is evidence about receipt, not an observed compliance type. The measurement relation below does not require R_i to be independent of Y_i conditional on receipt.

2 Calibration and identification

Writing $q = P(R_i = 1 \mid Z_i = 1)$, the law of total probability gives

$$q = \pi s + (1 - \pi) f, \qquad \pi = \frac{q - f}{s - f}, \qquad \tau = \frac{\Delta(s - f)}{q - f}. \tag{3}$$

The compliance correction is the Rogan–Gladen correction for the prevalence of a misclassified binary condition (Rogan and Gladen, 1978). Substituting it into the Wald ratio is an application of that correction, not a new identification result. Without information about s and f, the observed flag rate does not separately identify π.

Calibration can come from an independent validation sample with verified receipt and the same measurement protocol. For the formulas here, validation units with and without receipt are sampled separately and independently. Within each receipt class, their flag probabilities must equal those among assigned experimental units. This transport assumption can fail if delivery strength, baseline variation, or measurement quality differs across samples. Receipt-class sampling proportions need not match the experimental compliance fraction. The detection rule and threshold must be fixed before validation; tuning on validation observations would require additional data or inference accounting for that tuning.

Control units have no receipt under the maintained design, but their flag rate identifies f for assigned nonrecipients only under an additional equality of measurement distributions. Exclusion for the later outcome alone does not imply that equality. Verified receipt among some assigned units is therefore especially useful. No such validation dataset is supplied in this repository.

2.1 A delivery-window rule

For a positive mechanical delivery signature, one possible rule is

$$R_i = 1 \left\{ \frac{\bar{W}_{i,\text{delivery}} - \bar{W}_{i,\text{pre}}}{S_{i,\text{pre}} \sqrt{1/T_{\text{pre}} + 1/T_{\text{delivery}}}} > t_{1-\alpha, T_{\text{pre}}-1} \right\}. \tag{4}$$

If daily errors are independent Gaussian draws with a common within-unit variance, the baseline is stable, and receipt adds a constant positive delivery shift, the statistic has a central t distribution for nonrecipients and a noncentral t distribution for recipients. In this special model $f = \alpha$, and sensitivity follows from the noncentrality parameter. A nominal test level need not equal the actual false-positive rate with trends, serial dependence, censoring, or a misspecified variance. The correction uses actual detection rates, estimated or justified for the population at hand.

A negative delivery-window change does not identify an IV defier. Defiance means that $D_i(1) = 0$ and $D_i(0) = 1$. It concerns receipt, not the sign of an outcome change. Negative changes can diagnose a poor measurement model or a failure of the assumed positive signature, but they cannot distinguish receipt types from these data. Distributional tests of the joint IV restrictions require a different argument (Kitagawa, 2015).

3 Estimation and uncertainty

Let $\hat{a}$ be the difference in outcome means, $\hat{q}$ the treated-arm flag mean, and $\hat{s}$, $\hat{f}$ calibration estimates. Set

$$\hat{\pi} = \frac{\hat{q} - \hat{f}}{\hat{s} - \hat{f}}, \qquad \hat{\tau} = \frac{\hat{a}(\hat{s} - \hat{f})}{\hat{q} - \hat{f}}. \tag{5}$$

Under the maintained assumptions and consistent calibration, these estimates are consistent when π is in the interior of $(0,1)$ and $s-f$ is bounded away from zero. The ratio is not generally unbiased in finite samples. With estimated rates, both experiment and validation sample sizes must grow for this argument.

For $\theta = (a, q, s, f)$ and $g(\theta) = a(s-f)/(q-f)$, the gradient is

$$\nabla g = \left(\frac{s-f}{q-f},\ -\frac{a(s-f)}{(q-f)^2},\ \frac{a}{q-f},\ \frac{a(s-q)}{(q-f)^2} \right)'. \tag{6}$$

The joint delta variance is $\widehat{V}(\hat{\tau}) = \nabla g(\hat{\theta})' \widehat{\Sigma} \nabla g(\hat{\theta})$, where $\widehat{\Sigma}$ estimates the covariance of the estimates, already scaled by their sample sizes.

For independent experimental units with fixed arm sizes n_1, n_0, the experimental block is

$$\widehat{\Sigma}_{a,q} = \begin{pmatrix} S^2_{Y,1}/n_1 + S^2_{Y,0}/n_0 & S_{YR,1}/n_1 \\ S_{YR,1}/n_1 & S^2_{R,1}/n_1 \end{pmatrix}. \tag{7}$$

Here S denotes sample variances or covariance with the usual degrees-of-freedom correction. Independent receipt-class validation samples of sizes m_1, m_0 contribute the block

$$\widehat{\Sigma}_{s,f} = \operatorname{diag}\{S^2_{R,\text{val},1}/m_1, S^2_{R,\text{val},0}/m_0\}.$$

Known rates contribute zero calibration variance. If validation overlaps the experiment or the rates are estimated jointly, the corresponding covariance terms must also be estimated.

The covariance $S_{YR,1}$ cannot be discarded merely because the two measurements use different windows. For example, a perfect flag $R_i = C_i$ and a treated outcome $Y_i = \tau C_i + \varepsilon_i$ with independent mean-zero noise give $\operatorname{Cov}(Y_i, R_i \mid Z_i = 1) = \tau\pi(1-\pi)$. With no noise and zero control outcomes, $\hat{a} = \tau\hat{q}$: the ratio has no variation from the realized compliance composition. The covariance term captures that cancellation. Persistent shocks can create further dependence.

The simulations use normal 95% Wald intervals based on this variance. These are large-sample approximations for sufficiently informative receipt measurements. They can fail when $q-f$ is small or calibration is imprecise. A nonpositive estimated compliance fraction, a fraction above one, or $\hat{s} \le \hat{f}$ is reported as an unavailable estimate rather than clipped or silently dropped. Even an interior estimate does not guarantee reliable Wald inference. Weak-denominator inference would require inversion of a joint test for $a(s-f) - \tau(q-f) = 0$ or another method justified for that regime; it is not implemented here. Cluster-randomized experiments likewise require inference at the assignment-cluster level.

4 Sensitivity when detection is imperfect

Assuming perfect sensitivity when $s < 1$ underestimates compliance, provided f is correctly specified:

$$\widetilde{\pi} = \frac{q-f}{1-f} = \pi \frac{s-f}{1-f} \le \pi. \tag{8}$$

The corresponding population ratio exaggerates the magnitude of a nonzero LATE. Dividing by the uncorrected flag rate has no fixed bias direction because false positives and false negatives can offset each other.

Suppose f is known, $q > f$, and $s \in [s_L, s_U]$. Feasibility requires $s \ge q$. Define $s_* = \max(s_L, q)$ and assume $s_* \le s_U$. Then

$$\pi \in \left[\frac{q-f}{s_U - f}, \frac{q-f}{s_* - f} \right], \qquad \tau \in \left\{ \frac{\Delta(s-f)}{q-f} : s \in [s_*, s_U] \right\}. \tag{9}$$

For positive Δ, the effect interval runs from $\Delta(s_* - f)/(q - f)$ to $\Delta(s_U - f)/(q - f)$; for negative Δ, their order reverses. An empty feasible sensitivity range contradicts the proposed calibration restrictions. These are population identification bounds, not sampling confidence intervals. Uncertainty in Δ, q, or f requires a separate inferential calculation. If f is also bounded, the feasible set must vary both rates rather than holding f at its nominal test level.

Predicting compliance from baseline covariates does not remove the denominator. For a subgroup G defined by a fixed rule using only baseline covariates, randomization and exclusion give

$$E[Y \mid Z = 1, G] - E[Y \mid Z = 0, G] = P(C = 1 \mid G)E[Y(1) - Y(0) \mid C = 1, G]. \tag{10}$$

Thus a predicted-complier subgroup comparison is generally a subgroup assignment effect. It equals a complier effect only with additional justification, such as perfect classification and an appropriate target population. Prediction accuracy does not test independence of compliance type and potential outcomes.

5 Simulation evidence

Each scenario uses 2000 independent replications with 500 units and exactly half randomly assigned. Compliance types are independent Bernoulli draws with probability 0.75 before assignment. Controls receive no treatment. Independent unit baselines follow $N(5, 3^2)$. There are 30 pre-treatment, 6 delivery, and 30 post-treatment observations per unit. Daily Gaussian noise has marginal standard deviation 2. The baseline design has independent daily noise; serial scenarios use stationary AR(1) noise with correlation parameter 0.6 across the entire series. Outcomes are continuous and are not truncated. The untreated mean post-treatment total is 150.

Receipt adds a delivery shift of 5 in the strong-signal scenario, 2 in moderate and serial scenarios, and 0.5 in the weak-signal scenario. It adds 3 per day afterward, except in the null-effect scenario, where that effect is zero. The LATE is the post-treatment total effect. The detector uses the one-sided rule in equation (4) with $\alpha = 0.05$. Known-rate scenarios use the exact independent-Gaussian sensitivity and false-positive rate. Validated scenarios independently generate 500 verified recipients and 500 verified nonrecipients in each replication using the same measurement model and retain calibration uncertainty. These validation samples are simulated, not empirical evidence that receipt can be verified in an application.

Table 1 compares suitable calibration with deliberate calibration errors. Under independent Gaussian noise, moderate-signal sensitivity is 0.705. With independent validation, coverage is 0.960; assuming perfect detection instead gives bias 41.7 and coverage 0.323. For serially correlated outcomes, coverage is 0.952 with validation and 0.949 with the inappropriate independent-noise rates. The incorrectly calibrated serial scenario still has coverage close to the nominal level in this design. Its opposing sensitivity and false-positive errors partly offset at this compliance fraction; that coincidence does not validate those rates. The comparison concerns measurement calibration, not the causal assumptions, which hold by construction.

Weak detection exposes the limits of ordinary ratio inference. Its unavailable fraction is 0.184 and its coverage is 0.814. Bias among available estimates can be unstable when the denominator is close to zero. These results do not justify applying Wald intervals to every interior estimate. The null-effect scenario checks behavior when the assignment effect is zero, while compliance remains positive. All reported results, including unfavorable ones, are generated from the same simulation program.

Table 1: Corrected LATE and 95% Wald intervals across measurement scenarios.

Scenario	True LATE	Bias	Coverage	Width	Unavailable
Strong signal	90	-0.2	0.950	42.6	0.000
Moderate signal	90	0.1	0.949	46.7	0.000
Moderate, validated	90	0.4	0.960	48.0	0.000
Assume perfect detection	90	41.7	0.323	67.8	0.000
Serial noise, validated	90	0.6	0.952	54.8	0.000
Serial noise, iid rates	90	2.9	0.949	49.4	0.000
Weak signal, validated	90	27.8	0.814	320.9	0.184
Null effect, validated	0	0.4	0.947	42.6	0.000

Bias and mean interval width use replications with an available estimate. Coverage uses all 2000 replications, counting unavailable intervals as failures. Unavailable is a fraction of all replications. Monte Carlo standard errors for coverage are at most $1/(2\sqrt{2000})$; row-specific values and all design parameters are in the generated CSV. Strong and moderate signal rows use known rates. "Assume perfect detection" incorrectly sets sensitivity to one. "Serial noise, iid rates" incorrectly applies independent-Gaussian rates to serially correlated measurements. Effects, bias, and width are in post-treatment total-outcome units.

6 Scope

The practical proposal is to evaluate a delivery-window measurement against independently verified receipt, then use calibrated measurement-error correction when the one-sided IV design is credible. Validation must cover both recipients and nonrecipients and the delivery conditions of the intended application. A detector trained or calibrated in an unusually strong delivery setting may not transport to routine delivery.

The correction is established, the delivery signature is an assumption to investigate, and the simulations establish performance only under their specified models. An empirical contribution would require evidence that these measurements recover otherwise-unobserved receipt accurately enough to improve an actual analysis. The repository currently supplies a technical illustration and explicit limitations, not that evidence.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, 83(5):2043–2063.

Rogan, W. J. and Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1):71–76.