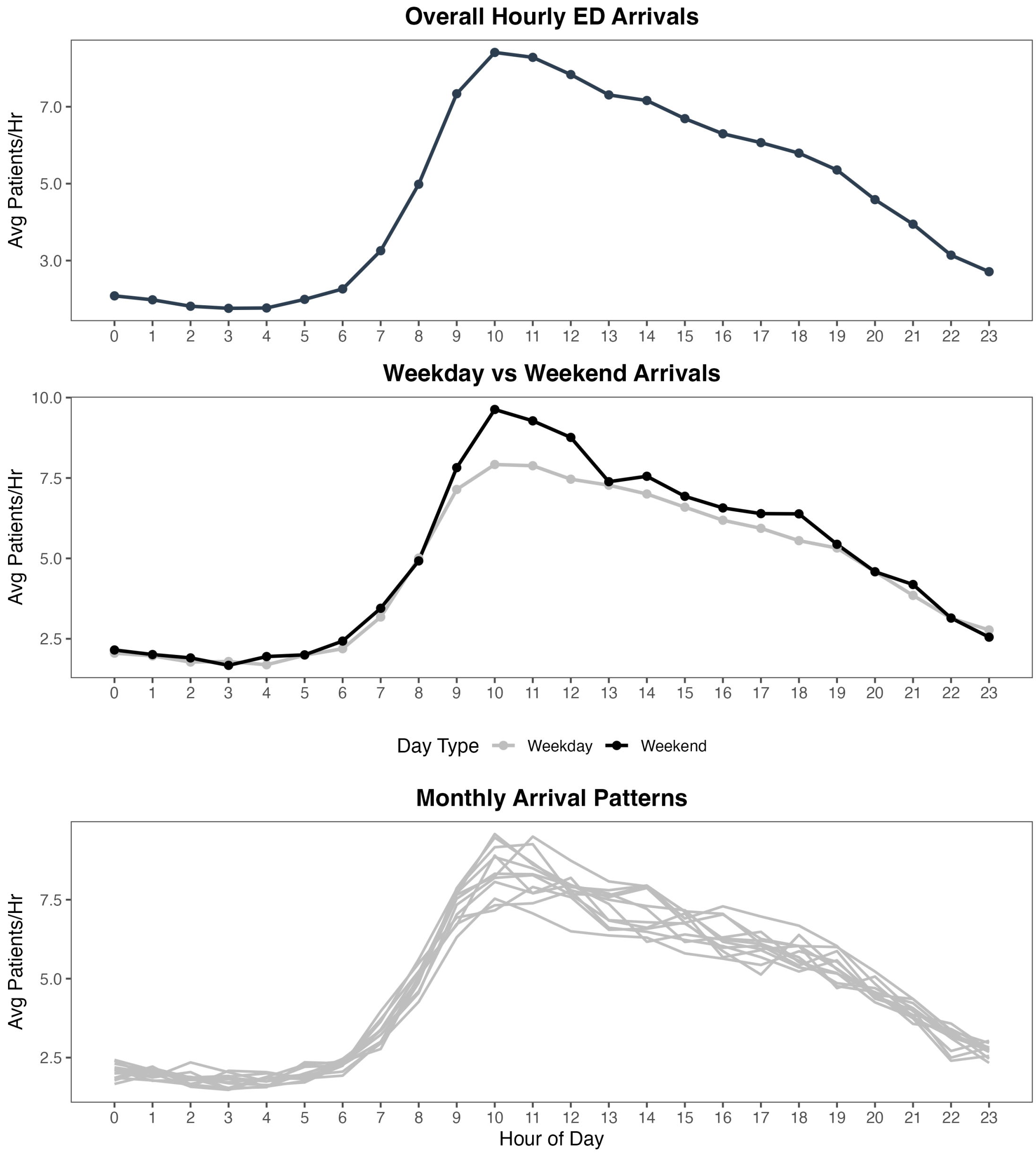
Appendix

EC.1 ED Arrival Patterns

Figure EC.1.1: Arrival Rate by Hour



EC.2 ED Capacity

Table EC.2.1: Internal Guidelines for ED Overcapacity and Saturation Activation

Capacity Level	**Conditions**	**Actions**
Normal Operations	• Waiting room time < 20 minutes • ED census below staffed bed capacity	• No RMA or waiting-room evaluations unless treatment room unnecessary • No waiting-room laboratory testing • Nurse-initiated protocols entered 30 minutes after arrival
Minor Overcapacity	• Waiting room time 21–90 minutes • ≥10 patients waiting • ED census exceeds beds by ≤10 patients • Team Lead discretion	• All normal-operations actions • Waiting-room lab initiation • RMA activation with dedicated staff • Expanded zone placement • Expedited consulting communication
Major Overcapacity	• Waiting room time > 90 minutes • >20 patients waiting • >40 arrivals in 2 hours • ED census exceeds beds by ≥20 patients • Team Lead discretion	• All minor-overcapacity actions • Consider diversion • Second RMA activation if feasible • Hospital-wide throughput escalation • Activation of observation APPs and on-call ED physician

EC.3 Sample Selection

This appendix documents our sample selection process and compares characteristics of included versus excluded encounters. Our analytical sample focuses on encounters where imaging decisions are both consequential and discretionary—the population where our instrumental variables approach can identify meaningful policy-relevant effects.

Figure EC.3.1 displays the sequential exclusion criteria applied to arrive at our analytical sample. Starting with 48,854 adult patient encounters during the study period (October 2018–December 2019), we exclude: (1) encounters with non-full-time providers, (2) rare complaints with fewer than 500 total encounters, (3) complaints where batching occurs in less than 5% of encounters, and (4) encounters without any imaging ordered. These restrictions yield an analytical sample of 11,659 encounters (23.9% of the original sample).

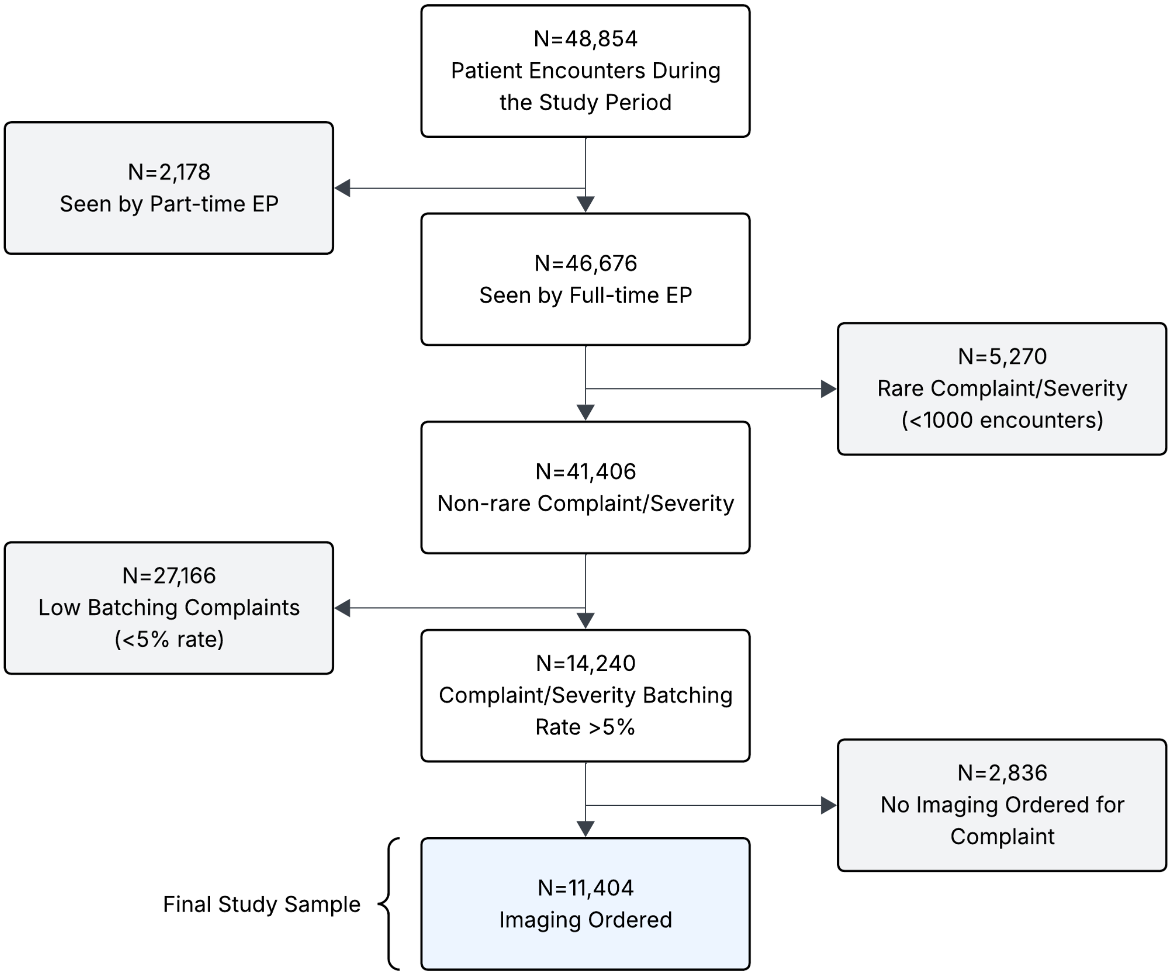


Figure EC.3.1: CONSORT Flow Diagram for Sample Selection

Notes: Sample selection for Mayo Clinic ED encounters (October 2018–December 2019). Rare complaints are those with fewer than 500 total encounters. Low-batching complaints are those where batching occurs in less than 5% of encounters.

Table EC.3.2 compares patient characteristics and outcomes between excluded and included encounters. The analytical sample differs from excluded encounters in clinically expected ways: included patients are older

(62.9 vs. 57.5 years), higher acuity (ESI 2.52 vs. 2.87), and more likely to present with abnormal vital signs. These differences reflect appropriate sample definition rather than selection bias—our sample comprises the population where imaging decisions matter most and where physicians face genuine uncertainty about optimal testing strategy.

Table EC.3.2: Comparison of Excluded vs. Included Encounters

	Excluded Encounters	Included (Analytical Sample)	p-value
Demographics			
Age (years)	57.5 (19.6)	62.9 (18.3)	<0.001
Acuity			
ESI Level (mean)	2.87 (0.64)	2.52 (0.59)	<0.001
Vital Signs			
Tachycardic (%)	18.8	20.1	0.001
Tachypneic (%)	9.1	9.2	0.567
Febrile (%)	1.4	4.6	<0.001
Hypotensive (%)	1.2	2.3	<0.001
Outcomes			
ED LOS (minutes)	238.3 (135.0)	272.0 (133.5)	<0.001
Admission rate (%)	17.0	28.4	<0.001
N	37,195	11,659	
% of total sample	76.1%	23.9%	

Notes: Standard deviations in parentheses for continuous variables. ESI (Emergency Severity Index) ranges from 1 (highest acuity) to 5 (lowest acuity). P-values from two-sample t-tests for continuous variables and proportion tests for binary variables.

These sample restrictions serve two purposes standard in the instrumental variables literature. First, our IV approach requires sufficient variation in batching behavior to generate a strong first stage; including complaints where batching rarely occurs would create weak instrument problems. Second, our research question concerns discretionary decisions—we identify the Local Average Treatment Effect for patients whose testing strategy depends on physician preference rather than clinical necessity. While our results apply to approximately 24% of ED visits, this population consumes 41% of imaging resources and represents the margin where ED management interventions can meaningfully influence practice.

EC.4 Complier, Always-Taker, and Never-Taker Classification

This section describes the method used to calculate the share of compliers, always-takers, and never-takers in the context of batch ordering in the emergency department (ED). Our approach follows Dahl et al. [2014], Dobbie et al. [2018], and Eichmeyer and Zhang [2022].

Overview

Compliers are defined as patients who would not have had their imaging tests ordered in a batch if they had been seen by a low-batch-tendency physician (“sequence”) but would have had their imaging tests batched if they had been seen by a high-batch-tendency physician (“batcher”):

$$\pi_{\text{complier}} = P(B_i = 1|Z_i = \bar{z}) - P(B_i = 1|Z_i = \underline{z}) = P(B_{\bar{z}i} > B_{\underline{z}i})$$

where B_i represents the batch ordering decision for patient i, Z_i represents the batch tendency of patient i’s assigned physician, and $\bar{z}$ and $\underline{z}$ represent the maximum and minimum values of our batch tendency instrument (the highest and lowest batch tendency physicians), respectively.

Always-takers are patients whose imaging tests would be batched regardless of which physician they see. Because of the monotonicity and independence assumptions, the fraction of always-takers is given by the probability of being batched by the most conservative (lowest batch tendency) physician:

$$\pi_{\text{always-taker}} = P(B_i = 1|Z_i = \underline{z}) = P(B_{\bar{z}i} = B_{\underline{z}i} = 1)$$

Finally, never-takers are patients whose imaging tests would never be batched regardless of which physician they see, with the fraction of never-takers given by the probability of not being batched by the most aggressive (highest batch tendency) physician:

$$\pi_{\text{never-taker}} = P(B_i = 0|Z_i = \bar{z}) = P(B_{\bar{z}i} = B_{\underline{z}i} = 0)$$

Number of Compliers

We calculate the shares of patients in each category by examining batch ordering rates for patients assigned to physicians at different points in the batch tendency distribution. Following Dahl et al. [2014], we define the “most aggressive” batch-ordering physicians ($\bar{z}$) as those at the top 1 percentile of batch tendency and the “most conservative” batch-ordering physicians ($\underline{z}$) as those at the bottom 1 percentile.

In the first three columns of Table EC.4.1, we estimate a local linear regression of $Batched_i$ on our residualized measure of physician batch tendency. Under this more flexible analog to our first-stage equation, we find that approximately 20 percent of our sample are compliers, 73 percent are never-takers, and 7 percent are always-takers.

In the last three columns of Table EC.4.1, we estimate our linear specification of the first stage, given by Equation (4). Under this specification, we can recover π_c as $\hat{\alpha}_1(\bar{z} - \underline{z})$, π_a as $\hat{\alpha}_0 + \hat{\alpha}_1\underline{z}$, and π_n as $1 - \hat{\alpha}_0 - \hat{\alpha}_1\bar{z}$, where $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are the estimated first-stage coefficients. Under this linear specification, we find that 21 percent of our sample are compliers, 72 percent are never-takers, and 7 percent are always-takers. We also explore the sensitivity of the estimated share of compliers, always-takers, and never-takers to the exact choice of cutoff for the most aggressive and most conservative physicians. As shown in Table EC.4.1, our results are robust to the particular model specification and cutoff.

Table EC.4.1: Sample Share by Compliance Type

	Local Linear Model			Linear Model		
Batch Tendency Cutoff:	1%	1.5%	2%	1%	1.5%	2%
Compliers	0.20	0.20	0.20	0.21	0.21	0.21
Never-Takers	0.73	0.73	0.73	0.72	0.72	0.72
Always-Takers	0.07	0.07	0.07	0.07	0.07	0.07

Notes: This table presents the estimated share of compliers, never-takers, and always-takers under different model specifications and batch tendency cutoffs. The local linear model estimates a loess regression of batching on batch tendency. The linear model estimates Equation (4) in the main text. Cutoffs define the percentiles used to identify the most aggressive (top percentile) and most conservative (bottom percentile) batch-ordering physicians.

Characteristics of Compliers

We also characterize our population of compliers by observable characteristics, which can be recovered by calculating the fraction of compliers in different subsamples [Abadie and Gardeazabal, 2003, Dahl et al., 2014]. For any binary characteristic X_i, we compute:

$$P(X_i = x|\text{complier}) = \frac{\pi_{c|x} \cdot P(X_i = x)}{\pi_c}$$

where $\pi_{c|x}$ is the complier share estimated within the subsample where $X_i = x$, and π_c is the overall complier share. The ratio $P(X_i = x|\text{complier})/P(X_i = x)$ indicates whether compliers are over-represented (ratio > 1) or under-represented (ratio < 1) among patients with characteristic $X_i = x$.

Table EC.4.2 presents the sample distribution, complier distribution, and relative likelihood for different patient subgroups. Several patterns emerge. Compliers are significantly more likely to be female (ratio = 1.21) and to have normal vital signs at presentation (ratio = 1.13). Compliers are less likely to be tachycardic (ratio = 0.57) or to have any abnormal vital sign (ratio = 0.79). Compliers are also more likely to have laboratory tests ordered (ratio = 1.13) but less likely to present without laboratory orders (ratio = 0.49).

These patterns suggest that compliers—patients at the margin of physician discretion—tend to present with less acute clinical pictures. Patients with abnormal vital signs or without laboratory workups may have more obvious clinical trajectories that dictate a particular imaging strategy regardless of physician preference. In contrast, patients with normal vitals and standard laboratory workups represent the diagnostic "gray zone' ' where physician practice style most influences testing decisions.

Table EC.4.3 presents analogous results by chief complaint category. Compliers are substantially over-represented among patients presenting with Falls, Motor Vehicle Accidents, Assaults, and Trauma (ratio = 1.66) and Neurological Issues (ratio = 1.24). Compliers are under-represented among patients with Abdominal Complaints (ratio = 0.62) and Fevers, Sweats, or Chills (ratio = 0.62). This pattern is consistent with clinical intuition: trauma and neurological presentations often involve diagnostic uncertainty about the extent of injury, creating opportunities for physician discretion in imaging strategy. In contrast, abdominal pain and fever presentations may have more standardized imaging protocols that reduce the influence of individual physician preference.

Together, these complier characteristics help interpret our LATE estimates. Our instrumental variables analysis identifies the effect of batch ordering for the approximately 20 percent of patients whose imaging strategy depends on physician preference rather than clinical necessity. These marginal patients tend to be older, female, with normal vital signs and standard laboratory workups, and are disproportionately likely to present with trauma or neurological complaints. This population represents exactly the "gray zone" where ED operational interventions could influence practice patterns without constraining clinically necessary care.

Table EC.4.2: Characteristics of Marginal Patients (Compliers)

	$P[X=x]$	$P[X=x \mid \text{complier}]$	$\frac{P[X=x \mid \text{complier}]}{P[X=x]}$
Demographics			
Male	0.477	0.407	0.853
Female	0.523	0.633	1.211
White	0.898	0.918	1.022
Non-White	0.102	0.120	1.181
Age			
Age $<$ 50	0.230	0.204	0.887
Age $\geq$ 50	0.770	0.834	1.082
Acuity			
High Acuity (ESI 1–2)	0.486	0.478	0.984
Lower Acuity (ESI 3–5)	0.514	0.560	1.090
Vital Signs			
Tachycardic	0.201	0.115	0.569
Any Abnormal Vital	0.280	0.222	0.793
Normal Vitals	0.720	0.817	1.134
Laboratory Testing			
Labs Ordered	0.780	0.877	1.125
No Labs Ordered	0.220	0.108	0.492

Notes: This table presents the sample distribution, complier distribution, and relative likelihood for different patient subgroups. Column 1 reports the unconditional probability of each characteristic in the full sample. Column 2 reports the probability of each characteristic among compliers, estimated following Abadie and Gardeazabal [2003]. Column 3 reports the ratio of the complier probability to the sample probability, indicating whether compliers are over-represented (ratio $>$ 1) or under-represented (ratio $<$ 1) among patients with each characteristic.

Table EC.4.3: Complier Characteristics by Chief Complaint

Chief Complaint	$P[X = x]$	$P[X = x \mid \text{complier}]$	$\frac{P[X=x\mid\text{complier}]}{P[X=x]}$
Neurological Issue	0.208	0.259	1.241
Falls, MVA, Assaults, Trauma	0.171	0.284	1.656
Extremity Complaints	0.228	0.199	0.870
Dizziness/Lightheadedness/Syncope	0.106	0.094	0.882
Fatigue and Weakness	0.103	0.091	0.887
Abdominal Complaints	0.100	0.062	0.623
Fevers, Sweats, or Chills	0.083	0.052	0.623

Notes: This table presents the sample distribution, complier distribution, and relative likelihood for each chief complaint category. Compliers are patients whose batching decision depends on the assigned physician's batch tendency. See notes to Table EC.4.2 for estimation details.

EC.5 Mediation Analysis: SEM Results

Table EC.5.1: Mediation Analysis of the Effect of Batching on ED Length of Stay (LOS) via Imaging and Admission Decisions

	Estimate	Standard Error	p-value
Number of Imaging Tests (Residualized)			
Batching (b_1)	1.434	0.106	<0.001
Tachycardic	0.077	0.015	<0.001
Tachypneic	0.111	0.020	<0.001
Febrile	0.051	0.028	0.062
Hypotensive	0.111	0.038	0.003
Arrival Age (scaled)	0.002	0.000	<0.001
Admission Decisions (Residualized)			
Batching (b_2)	0.327	0.069	<0.001
Number of Imaging Tests (b_3)	0.162	0.006	<0.001
Tachycardic	0.084	0.010	<0.001
Tachypneic	0.056	0.013	<0.001
Febrile	0.142	0.018	<0.001
Hypotensive	0.210	0.024	<0.001
Arrival Age (scaled)	0.002	0.000	<0.001
Length of Stay (Residualized)			
Number of Imaging Tests (c_1)	0.144	0.006	<0.001
Admission Decisions (c_2)	0.089	0.010	<0.001
Batching Direct Effect (c')	0.031	0.072	0.662
Tachycardic	0.041	0.010	<0.001
Tachypneic	-0.005	0.013	0.721
Febrile	-0.020	0.018	0.283
Hypotensive	-0.025	0.025	0.326
Arrival Age (scaled)	0.002	0.000	<0.001
Indirect and Total Effects			
Indirect via Imaging Tests ($b_1 \times c_1$)	0.207	0.018	<0.001
Indirect via Admission Decisions ($b_2 \times c_2 + b_1 \times b_3 \times c_2$)	0.050	0.008	<0.001
Total Indirect Effect	0.256	0.020	<0.001
Total Effect	0.288	0.073	<0.001

Notes: This table presents the results of a structural equation model (SEM) investigating the relationships between batching, the number of imaging tests, admission decisions, and ED length of stay (LOS). To address potential confounding from fixed effects (e.g., complaint type and time of month), all variables were residualized by regressing them on the fixed effects prior to the SEM. This residualization ensures that the estimates reflect associations net of the fixed effects.

Table EC.5.2: Mediation Analysis of the Effect of Batching on Time to Disposition

	Estimate	Standard Error	p-value
Number of Imaging Tests (Residualized)			
Batching (b_1)	1.434	0.106	<0.001
Tachycardic	0.077	0.015	<0.001
Tachypneic	0.111	0.020	<0.001
Febrile	0.051	0.028	0.062
Hypotensive	0.111	0.038	0.003
Arrival Age (scaled)	0.002	0.000	<0.001
Time to Disposition (Residualized)			
Number of Imaging Tests (c_1)	0.060	0.007	<0.001
Batching Direct Effect (c')	-0.100	0.081	0.216
Tachycardic	0.006	0.011	0.617
Tachypneic	-0.018	0.015	0.236
Febrile	-0.098	0.021	<0.001
Hypotensive	-0.090	0.028	0.002
Arrival Age (scaled)	-0.001	0.000	0.042
Indirect and Total Effects			
Indirect via Imaging Tests ($b_1 \times c_1$)	0.085	0.012	<0.001
Total Effect	-0.015	0.081	0.855

Notes: This table presents results from a structural equation model (SEM) analyzing the relationship between batching, the number of imaging tests, and time to disposition. All variables were residualized to account for fixed effects of day of the week, month, and chief complaint by severity. The coefficients for the number of imaging tests reflect its role as a mediator in the pathway from batching to time to disposition. The direct effect of batching is not significant, while the indirect effect via imaging tests is significant and positive, supporting the hypothesis that increased diagnostic intensity prolongs time to disposition. This analysis is exploratory and intended to provide insight into plausible mechanisms, not causal mediation.

EC.6 Cost-Benefit Analysis of Discretionary Batch Ordering

This appendix details our cost-benefit analysis quantifying the operational and financial burden of discretionary batch ordering. We adopt a societal perspective using Medicare-allowed amounts as proxies for resource costs, following standard health economics practice.

Our analysis focuses on two cost components where we have direct causal evidence: (1) excess imaging from additional tests and (2) excess ED capacity consumed from extended length of stay. Let C_i denote the total incremental cost per batched complier:

$$C_i = C_i^{\text{imaging}} + C_i^{\text{capacity}} = \underbrace{(\Delta\text{X-rays}_i) \cdot p^{\text{xray}}}_{\text{imaging}} + \underbrace{(\Delta\text{Hours}_i) \cdot p^{\text{bed}}}_{\text{capacity}} \tag{1}$$

where $\Delta\text{X-rays}_i$ is the causal increase in X-rays from our UJIVE estimates, p^{xray} is the Medicare-allowed reimbursement per radiograph, ΔHours_i is the increase in bed-hours, and p^{bed} is the cost per ED bed-hour.

For imaging costs, we use Medicare-allowed reimbursement for diagnostic radiography in the ED setting. For a two-view chest radiograph (CPT 71046), Dabus et al. [2025] report the technical component under HOPPS is \$98.53, with an additional \$30–40 for the professional component, yielding $p^{\text{xray}} \approx \$130$ (range \$100–160). Our UJIVE estimates indicate $\Delta\text{X-rays} = 0.959$ ($p < 0.001$), while effects on ultrasound and CT are not statistically significant. To maintain conservative estimates, we include only X-rays:

$$C_i^{\text{imaging}} = 0.959 \times \$130 = \$125 \quad (\$96\text{–}\$153) \tag{2}$$

For capacity costs, we draw on time-driven activity-based costing studies. Schreyer and Martin [2017] report ED personnel costs of \$58.20 per bed-hour (2010–2011 data; approximately \$83 inflation-adjusted), while Canellas et al. [2024] find daily ED boarding costs of \$1,856, implying approximately \$77 per hour. We use $p^{\text{bed}} = \$85$ (range \$75–100). Our UJIVE estimates indicate discretionary batching increases LOS by 65%, translating to $\Delta\text{Hours} = 2.64$ additional hours relative to baseline LOS of 243 minutes:

$$C_i^{\text{capacity}} = 2.64 \times \$85 = \$224 \quad (\$198\text{–}\$264) \tag{3}$$

Combining these components yields total incremental costs of $C_i = \$349$ (\$294–417) per batched complier, with capacity costs accounting for 64% of the total.

To scale these estimates to annual ED operations, we identify the affected population. Our complier analysis (Appendix EC.4) indicates that approximately 20% of patients are compliers—those whose batching status depends on physician assignment rather than clinical necessity. Among batched patients, roughly 51% are compliers whose batching reflects physician discretion. Let N^{ED} denote annual ED volume and ϕ denote the share of imaging-relevant encounters (23.9% in our data). The number of batched compliers is:

$$N^{\text{compliers}} = N^{\text{ED}} \times \phi \times \pi_c \times P(\text{Batched}|\text{Complier}) \approx N^{\text{ED}} \times 0.0167 \tag{4}$$

where $\pi_c = 0.20$ is the complier share and $P(\text{Batched}|\text{Complier}) \approx 0.35$. Table EC.6.1 presents annual cost estimates by ED size.

For our study site, eliminating discretionary batching would free approximately 1,717 bed-hours annually—equivalent to 72 days of ED capacity—enabling approximately 425 additional patient visits (a 1.1% capacity improvement). A 50% reduction in discretionary batching would save approximately \$113,000 annually while freeing 850 bed-hours.

These estimates represent conservative lower bounds. We exclude: (i) non-significant imaging modalities (including ultrasound and CT point estimates would increase costs by 17%), (ii) radiologist interpretation time, (iii) physician cognitive burden from processing simultaneous results, (iv) patient time costs, (v)

Table EC.6.1: Estimated Annual Cost of Discretionary Batch Ordering by ED Size

ED Type	Annual Volume	Batched Compliers	Annual Cost	Range
Small (Rural/Community)	20,000	333	\$116,000	(\$98,000–\$139,000)
Medium (Suburban)	40,000	666	\$232,000	(\$196,000–\$278,000)
Large (Urban)	60,000	999	\$349,000	(\$294,000–\$417,000)
Study Site (Mayo Clinic)	39,083	650	\$227,000	(\$191,000–\$271,000)

Notes: Annual costs calculated as $N^{\text{compliers}} \times C_i$. Range reflects uncertainty in imaging (\$100–160 per X-ray) and bed-hour (\$75–100 per hour) cost parameters.

downstream admission costs (the 40 percentage point increase in admission probability, at \$10,000–15,000 per inpatient stay, would dwarf our direct cost estimates), and (vi) spillover congestion effects on other ED patients. Notably, we find no offsetting benefits: 72-hour return rates are statistically indistinguishable between strategies.

EC.7 Monotonicity

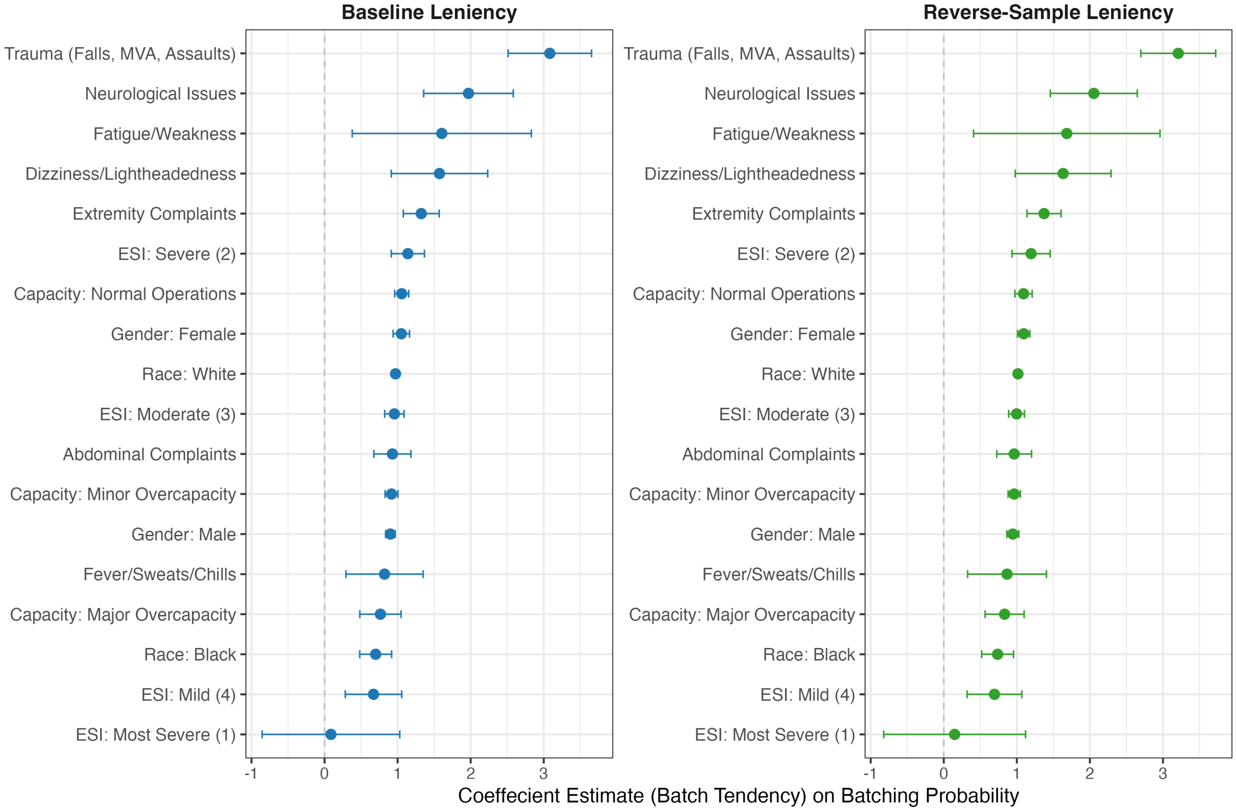


Figure EC.7.1: Testing the Monotonicity Assumption

Notes: The left panel displays the first stage coefficient and 95% CI of batching on the baseline physician batch tendency instrument for the corresponding sub-sample. The right panel constructs a new physician batch tendency instrument using all emergency visits, excluding the corresponding sub-sample ("reverse-sample"), and displays the coefficient and 95% CI of the first stage regression back on that sub-sample. Robust standard errors are clustered at the physician level.

EC.8 Placebo Exercise

We investigate whether the reduced-form effects observed in Section 4.1 are due to differences in batch ordering rates across physicians or due to other provider differences correlated with batching tendency. We examine this by studying reduced-form effects among patients with conditions that rarely require multiple imaging tests, as a falsification check. For example, consider a patient who arrives at the ED with an isolated ankle injury—a condition for which multiple imaging tests are rarely ordered. For such patients, we should expect to see no impact of physician batch tendency if "batchers" and "sequencers" do not systematically differ in other dimensions of care relevant to patient outcomes.

We restrict attention to ED visits for conditions which batching occurs less than 1 percent of the time, as a placebo check. We estimate reduced-form regressions of each main outcome on physician batch tendency for this subsample and the results are presented in Table EC.8.1. The results show no significant association between physician batch tendency and patient outcomes for this subsample, providing evidence that the reduced-form effects observed in the main analysis are not driven by unobserved physician differences correlated with batching tendency.

Table EC.8.1: Placebo Check: Batch Tendency and Patient Outcomes

	Dependent variable		
	Log time to disposition	Log LOS	72hr return with admission
	(1)	(2)	(3)
Batch tendency	3.486	3.580	−0.2158
	(1.724)	(1.750)	(0.1069)
Mean dependent variable	4.57	4.59	0.001
	(0.611)	(0.615)	(0.098)
Time FE	Yes	Yes	Yes
Baseline controls	Yes	Yes	Yes
Adj R^2	0.159	0.160	0.051
Observations	1,244	1,244	1,244

Notes: This table reports the estimated coefficients of a reduced-form regression of our main outcomes on physician batch tendency for patients with conditions for which batching occurs less than 1% of the time. The dependent variables are log time to disposition, log LOS, and 72-hour return with admission. Robust standard errors are clustered at the physician level.

This placebo analysis supports our identification strategy by suggesting that batch tendency is not systematically correlated with other physician practice patterns that might affect patient outcomes. However, we recognize that given the multidimensionality of physician behavior, we also present reduced-form results and maintain focus on imaging-related outcomes, for which such concerns are less pronounced.

EC.9 Sensitivity to Exclusion Restriction Violations

A core identifying assumption in our instrumental variables strategy is that physician batch tendency affects patient outcomes only through its effect on the batching decision (the exclusion restriction). This appendix presents sensitivity analyses following Conley et al. [2012] to assess how our estimates change under plausible violations of this assumption.

Under standard IV assumptions, the structural equation is:

$$Y_i = \beta \cdot \text{Batched}_i + X_i'\lambda + \varepsilon_i \tag{5}$$

which assumes that batch tendency affects outcomes only through its effect on Batched_i. The Conley et al. approach relaxes this assumption by allowing the instrument to have a direct effect on outcomes:

$$Y_i = \beta \cdot \text{Batched}_i + \underbrace{\delta \cdot \text{BatchTendency}_i}_{\text{violation}} + X_i'\lambda + u_i \tag{6}$$

where δ captures any exclusion restriction violation. Such violations could arise through multiple channels: direct effects (e.g., physicians with higher batch tendency may systematically differ in how they manage patients in ways that directly affect LOS) or backdoor paths (e.g., batch tendency may reflect broader physician aggressiveness or thoroughness that independently affects outcomes).

Under this framework, we can no longer point-identify the treatment effect β, but we can trace out how $\hat{\beta}(\delta)$ varies as we consider different magnitudes of the violation parameter δ. For a given value of δ, we compute:

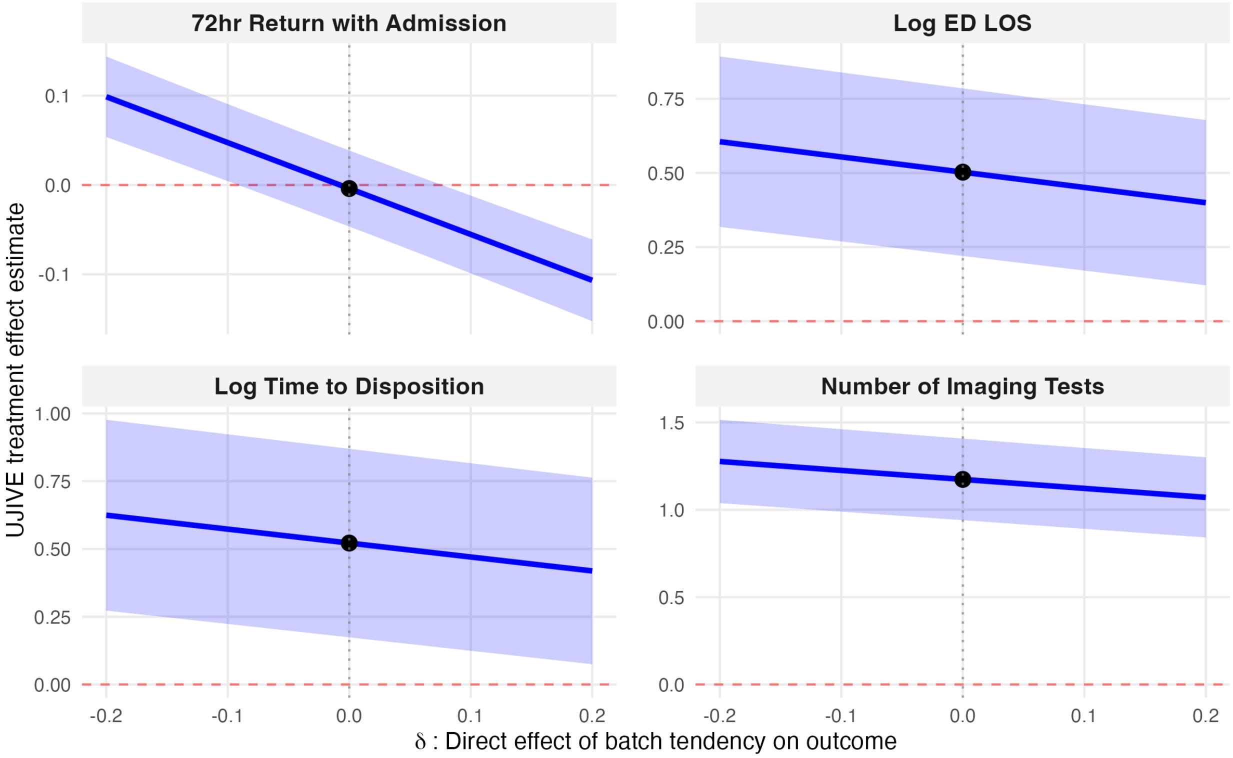
$$\hat{\beta}(\delta) = \hat{\beta}^{IV} - \delta \cdot \frac{\text{Cov}(Z_i, Y_i)}{\text{Cov}(Z_i, \text{Batched}_i)} \tag{7}$$

where $\hat{\beta}^{IV}$ is the standard IV estimate and Z_i denotes the instrument. Confidence intervals are constructed using heteroskedasticity-robust standard errors.

Figure EC.9.1 presents sensitivity analyses for our primary outcomes. We consider $\delta \in [-0.2, 0.2]$, which spans a range where a one percentage point increase in batch tendency directly changes the outcome by up to 0.2 units (in addition to any indirect effect through batching). Across all outcomes, the treatment effect remains statistically significant throughout this range. For log LOS, the 95% confidence interval excludes zero for all values of δ considered. Similar robustness holds for log time to disposition, number of imaging tests, and admission probability.

The stability of our estimates across this range of δ values provides evidence that our conclusions are not sensitive to moderate violations of the exclusion restriction. While we cannot definitively rule out all such violations—no observational study can—these results suggest that any direct effect of batch tendency on outcomes would need to be implausibly large to overturn our main findings.

Figure EC.9.1: Sensitivity of Treatment Effect Estimates to Exclusion Restriction Violations



Notes: This figure displays UJIVE estimates under different assumptions about direct effects (δ) of physician batch tendency on outcomes. The x-axis represents the magnitude of the exclusion restriction violation: $\delta = 0$ corresponds to the standard IV assumption, while $\delta \neq 0$ allows batch tendency to directly affect outcomes. Black points indicate baseline estimates ($\delta = 0$). Shaded areas show 95% confidence intervals using heteroskedasticity-robust standard errors following Conley et al. [2012].

EC.10 Robustness to Nonlinear Outcome Models

Our primary analysis uses linear models for all outcome variables, including binary outcomes (admission, 72-hour return) and count outcomes (number of imaging tests). While nonlinear models such as logit for binary outcomes and Poisson for counts may seem more appropriate given the outcome distributions, the linear 2SLS estimator provides a well-defined Local Average Treatment Effect (LATE) for compliers under the standard monotonicity and independence assumptions. Nonlinear second-stage models would require strong assumptions about the entire joint distribution of errors and generally lack the transparent LATE interpretation that gives our estimates clear policy relevance. The linear probability model, while potentially producing predictions outside $[0, 1]$ for individual observations, yields consistent estimates of the average marginal effect for compliers—exactly the parameter of interest in our quasi-experimental setting.

To verify that functional form assumptions do not drive our results, we estimate nonlinear outcome models while retaining a linear first stage. Specifically, we estimate logit models for binary outcomes and Poisson models for count outcomes using a control-function (two-stage residual inclusion) approach, and compute average marginal effects (AMEs) for comparison with our 2SLS estimates. Table EC.10.1 presents these results.

Table EC.10.1: Robustness of Estimates to Outcome Model Specification

Outcome	Model	2SLS	Nonlinear (AME)
72-hour return with admission	Logit	-0.015 (0.020)	-0.015 (0.023)
Admission	Logit	0.404^{***} (0.088)	0.377^{***} (0.075)
Number of imaging tests	Poisson	1.241^{***} (0.116)	1.197^{***} (0.241)

Notes: This table compares 2SLS estimates (linear second stage) with average marginal effects from nonlinear outcome models (logit for binary outcomes, Poisson for counts). Both approaches use physician batch tendency as the instrument with a linear first stage and include the full control set. Standard errors for nonlinear AMEs are delta-method estimates. $^{***}p < 0.001$.

The nonlinear specifications yield estimates that closely match our primary 2SLS results. For admission, the logit AME (0.377) is within one standard error of the 2SLS estimate (0.404). For number of imaging tests, the Poisson AME (1.197) is nearly identical to the 2SLS estimate (1.241). The 72-hour return estimates remain small and statistically insignificant under both specifications. This consistency across model specifications reinforces confidence that our substantive conclusions are not artifacts of functional form assumptions.

EC.11 Robustness to Alternative Outcome Measures

This appendix presents robustness checks using alternative definitions of our primary outcome measures. We examine (1) treatment time as an alternative measure of ED efficiency, and (2) any 72-hour ED revisit as an alternative measure of care quality.

Our primary time-based outcomes—LOS and time to disposition—may capture variation unrelated to physician decision-making, such as waiting room delays or post-disposition boarding. Treatment time, defined as the period from first physician contact to disposition decision, isolates the clinical processing phase and provides a cleaner measure of how batching affects physician workflow. Table EC.11.1 presents estimates using this alternative specification alongside time to disposition for comparison.

Table EC.11.1: Robustness: Treatment Time and Time to Disposition

	Sequenced	2SLS		UJIVE	
	mean	(2)	(3)	(4)	(5)
Log treatment time	5.180	0.556***	0.628***	0.438**	0.447**
	(0.462)	(0.120)	(0.123)	(0.216)	(0.210)
Log time to disposition	5.237	0.659***	0.651***	0.583***	0.522***
	(0.499)	(0.103)	(0.101)	(0.189)	(0.177)
Necessary controls	—	Yes	Yes	Yes	Yes
Precision controls	—	No	Yes	No	Yes
Observations	11,659	11,659	11,659	11,659	11,659

Notes: Column 1 reports means for sequenced patients (standard deviations in parentheses). Columns 2–3 report 2SLS estimates using physician batch tendency as the instrument. Columns 4–5 report UJIVE estimates using ED provider identifiers as instruments. Precision controls include patient vitals, demographics, complaint-severity fixed effects, physician characteristics, laboratory ordering, hours into shift, and ED capacity. $^{**}p < 0.01$, $^{***}p < 0.001$.

The estimates for treatment time closely track those for time to disposition. Under the full-control specification, the 2SLS estimate for log treatment time (0.628) implies a 87% increase, while UJIVE yields 0.447 (56% increase). The consistency across these measures confirms that batching affects clinical processing time, not merely waiting room or boarding delays.

For care quality, our primary measure is 72-hour ED revisit resulting in hospital admission. We focused on this measure based on guidance from our physician coauthors, who noted that many ED revisits are planned (e.g., wound checks, suture removal) or reflect intentional "watchful waiting' ' strategies rather than quality failures. Returns requiring admission more reliably indicate missed diagnoses or inadequate initial treatment. Nevertheless, we also estimated effects on any 72-hour revisit for comparability with prior literature. Table EC.11.2 presents both measures.

Both quality measures yield small, negative, and statistically insignificant effects across all specifications. The point estimates for any 72-hour return are slightly larger in magnitude than those for returns with admission, but neither approaches statistical significance. These results reinforce our main finding: discretionary batching does not improve short-term quality outcomes, regardless of how returns are measured.

Table EC.11.2: Robustness: Effect of Batching on 72-Hour Return Outcomes

	Sequenced	2SLS		UJIVE	
	mean	(2)	(3)	(4)	(5)
72-hr return with admission	0.012	−0.014	−0.015	−0.008	−0.004
	(0.110)	(0.018)	(0.020)	(0.020)	(0.022)
72-hr return (any reason)	0.030	−0.051	−0.054	−0.044	−0.040
	(0.170)	(0.029)	(0.031)	(0.032)	(0.034)
Necessary controls	—	Yes	Yes	Yes	Yes
Precision controls	—	No	Yes	No	Yes
Observations	11,659	11,659	11,659	11,659	11,659

Notes: Column 1 reports means for sequenced patients (standard deviations in parentheses). Columns 2–3 report 2SLS estimates; Columns 4–5 report UJIVE estimates. Standard errors are heteroskedasticity-robust. For patients admitted during the index visit, the 72-hour window begins at hospital discharge.

EC.12 Subsample Analysis of Discharged Patients

A potential alternative interpretation of our findings is that batching does not represent overuse but rather frontloads diagnostics that would eventually occur during hospitalization. Under this interpretation, the additional ED imaging simply shifts testing from the inpatient setting to the ED without increasing total resource utilization. This appendix presents a subsample analysis to address this concern.

Interpreting our findings purely as frontloading faces a conceptual challenge. Frontloading would imply that batched patients receive ED tests that would have been ordered during a predetermined hospitalization. Yet admission itself is a downstream consequence of the testing strategy—we find that batching increases admission probability by approximately 40 percentage points. This suggests that at least some additional tests are driving new admissions rather than frontloading diagnostics for patients whose admission was already determined. That is, the tests themselves influence disposition decisions, not the reverse.

To empirically investigate this concern, we re-estimated our main specifications restricting the sample to patients who were ultimately discharged from the ED. This subsample is informative because discharged patients, by construction, do not receive subsequent inpatient diagnostic workups. If the additional imaging from batching merely substituted for tests that would have occurred during hospitalization, we would expect attenuated effects among discharged patients. Table EC.12.1 presents these results.

Table EC.12.1: Effect of Batching on Imaging Tests Among Discharged Patients

	2SLS		UJIVE	
	(1)	(2)	(3)	(4)
Number of distinct imaging tests	1.516***	1.447***	1.497***	1.415***
	(0.122)	(0.124)	(0.135)	(0.133)
Necessary controls	Yes	Yes	Yes	Yes
Precision controls	No	Yes	No	Yes
Observations	6,306	6,305	6,306	6,305

Notes: Sample restricted to patients discharged from the ED (66.8% of analytical sample). Columns 1–2 report 2SLS estimates using physician batch tendency as the instrument. Columns 3–4 report UJIVE estimates. $^{***}p < 0.001$.

The estimated effect of batching on imaging tests remains large, positive, and statistically significant among discharged patients, with magnitudes comparable to or exceeding those in the full sample (1.42–1.52 additional tests versus 1.17–1.24 in the full sample). This pattern is inconsistent with a pure frontloading interpretation, as these patients by definition do not receive subsequent inpatient imaging.

This analysis should be interpreted with caution. Discharge status is itself a downstream outcome affected by ED imaging decisions; conditioning on discharge therefore risks post-treatment selection bias and may attenuate or inflate causal estimates. We do not treat this subsample analysis as a primary causal estimate. Rather, it demonstrates that the positive association between batching and imaging intensity is not driven solely by patients who are admitted and might receive further inpatient testing.

We note that multiple causal pathways may operate simultaneously. Some additional tests may reveal genuinely admission-worthy conditions that would have been discovered later. Others may trigger defensive admissions through incidental findings. Without linked ED-inpatient imaging records, we cannot empirically separate these mechanisms. Distinguishing between frontloading and overuse remains an important direction for future research.

EC.13 Standard Error Computation in Leniency Designs

This appendix discusses our choice of heteroskedasticity-robust standard errors without clustering. In instrumental variables designs using decision-maker leniency, the appropriate standard error calculation depends on the assignment mechanism rather than traditional considerations about error correlation [Abadie et al., 2020, 2023, Goldsmith-Pinkham et al., 2025].

The key insight from recent econometric work is that what matters for inference is the correlation structure of the product $\tilde{\ell}_i \varepsilon_i$—the interaction between relative leniency and the outcome residual—not the correlation structure of ε_i alone [Goldsmith-Pinkham et al., 2025]. Under the design-based view of inference, if physician assignment is independent across patients, then relative leniency $\tilde{\ell}_i$ is also independent across patients. Even if outcome residuals ε_i exhibit correlation (e.g., patients treated on the same day may have correlated outcomes due to shared environmental factors), the product $\tilde{\ell}_i \varepsilon_i$ will be uncorrelated across observations because the randomly assigned component $\tilde{\ell}_i$ breaks any dependence structure.

This reasoning parallels the standard approach in randomized controlled trials: we cluster at the level of randomization, not at the level of potential outcome correlation. If treatment is assigned independently to each unit, heteroskedasticity-robust standard errors are appropriate regardless of correlation in potential outcomes [Abadie et al., 2020].

In our setting, the Mayo Clinic ED employs a rotational patient assignment algorithm that assigns each patient independently to physicians based solely on arrival time. The rotation order is predetermined by the ED scheduler, and assignments occur automatically through the electronic health record system without consideration of patient characteristics, physician workload, or prior assignments. This institutional design generates individual-level randomization: each patient's physician assignment is determined by an independent draw from the rotation, conditional on which physicians are on duty.

Because assignment operates at the individual patient level, heteroskedasticity-robust standard errors are the appropriate choice. Clustering would be necessary only if groups of patients were assigned together to the same physician—for example, if all patients arriving during a particular hour were assigned to a single randomly chosen physician. The Mayo Clinic's rotational system does not operate this way; instead, consecutive patients are assigned to different physicians in the rotation sequence, ensuring independence across assignments.

We note that clustering standard errors "just in case" is not innocuous in this setting. When the assignment mechanism implies independent observations, clustering typically produces overly conservative inference by inflating standard errors beyond what the design warrants [Abadie et al., 2023]. Our choice of heteroskedasticity-robust standard errors thus reflects both the institutional reality of Mayo Clinic's assignment process and best practices for inference in leniency designs with individual-level randomization.

References

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132, 2003.

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020.

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.

Maureen M Canellas, Marcella Jewell, Jennifer L Edwards, Danielle Olivier, Adalia H Jun-O'Connell, and Martin A Reznek. Measurement of cost of boarding in the emergency department using time-driven activity-based costing. *Annals of Emergency Medicine*, 84(4):410–419, 2024. doi: 10.1016/j.annemergmed.2024.05.013.

Timothy G. Conley, Christian B. Hansen, and Peter E. Rossi. Plausibly exogenous. *The Review of Economics and Statistics*, 94(1):260–272, 2012. doi: 10.1162/REST_a_00139.

Guilherme Dabus, Joshua A Hirsch, Michael T Booker, and Ezequiel Silva. Practice expense and its impact on radiology reimbursement. *Journal of the American College of Radiology*, 2025. doi: 10.1016/j.jacr.2025.02.047.

Gordon B. Dahl, Andreas Ravndal Kostøl, and Magne Mogstad. Family welfare cultures. *The Quarterly Journal of Economics*, 129(4):1711–1752, November 2014. doi: 10.1093/qje/qju019. URL https://doi.org/10.1093/qje/qju019.

Will Dobbie, Jacob Goldin, and Crystal S. Yang. The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2):201–240, 2018.

Sarah Eichmeyer and Jonathan Zhang. Pathways into opioid dependence: Evidence from practice variation in emergency departments. *American Economic Journal: Applied Economics*, 14(4):271–300, 2022. doi: 10.1257/app.20210048.

Paul Goldsmith-Pinkham, Peter Hull, and Michal Kolesár. Leniency designs: An operator's manual, 2025. URL https://arxiv.org/abs/2511.03572.

Kraftin E Schreyer and Richard Martin. The economics of an admissions holding unit. *Western Journal of Emergency Medicine*, 18(4):553–558, 2017. doi: 10.5811/westjem.2017.4.32740.