Authors' Response to Review of Manuscript MS-HCM-2025-01252

The Impact of Batching Advanced Imaging Tests
in Emergency Departments

I. General Response to the Entire Review Team

We would like to thank the Department Editor (DE), the Associate Editor (AE), and the three referees for the constructive and detailed feedback provided in their reports. We are also grateful for the opportunity to address the reviewers' comments and for the decision to revise.

We have carefully revised the paper to address all comments, and it has significantly benefited from the suggested revisions. In revising the paper, we have also benefited from several conversations with our medical collaborators (and coauthors of this study), who are responsible for running the EDs at two leading US hospitals (our partner hospitals): Mayo Clinic and Massachusetts General Hospital (MGH). In what follows, we first briefly summarize the main changes and then provide individual responses.

1. We have substantially strengthened our identification strategy to address concerns about the exclusion restriction. The review team raised the concern that our instrument (physician batch tendency) might capture general diagnostic intensity rather than imaging-specific ordering behavior. We carefully address this in three ways: (a) we expanded our control set to include laboratory tests ordered, physician characteristics (experience, gender, hours into shift), and ED capacity levels; (b) we implemented the Unbiased Jackknife Instrumental Variables Estimator (UJIVE) following **?**, which is specifically designed for leniency designs with many providers and eliminates the many-instrument bias that arises when 2SLS uses numerous physician indicators; and (c) we conducted various robustness checks, including Conley et al. (2012) sensitivity analyses demonstrating robustness to plausible exclusion restriction violations. Our sensitivity analyses, as well as the close agreement between our 2SLS and the newly added UJIVE estimates—both before and after adding precision controls—provide strong evidence that our approaches isolate batching-specific effects.

2. We have clarified our theoretical framework and the definition of treatment. Following guidance from the AE and all three reviewers, we reframed our comparison throughout the manuscript as "batch ordering versus standard practice" rather than "batching versus sequencing." We added a formal hypothesis development section (Section 2.3) that articulates testable predictions grounded in the cognitive load, information value, and diagnostic momentum literatures. We also clarified that our Local Average Treatment Effect (LATE) estimator identifies effects for "compliers"—the approximately 20% of patients whose testing strategy depends on physician practice style rather than clinical necessity—which is precisely the margin where ED management interventions can influence practice.

3. We have added robustness checks and refined outcome measures. Following Reviewer 3's suggestion, we added treatment time (from physician contact to disposition) as another outcome variable, excluding both waiting room delays and post-disposition boarding. We verified robustness to nonlinear model specifications (logit for binary outcomes, Poisson for counts). We conducted heterogeneity analyses across seven chief complaint categories, finding consistent patterns of increased imaging without efficiency gains or quality improvements across all clinical scenarios. We also provided detailed sample selection documentation, including a CONSORT flow diagram and comparison of included versus excluded encounters.

4. We have updated our estimates to reflect the enhanced specification. Our revised results show that discretionary batching increases ED length of stay by approximately 65–81% (attenuated from the original 130% after adding precision controls), increases imaging tests by 88–105%, and increases admission probability by 40 percentage points, with no effect on 72-hour returns. The attenuation confirms reviewers' intuition that our original specification likely contained some correlation with general diagnostic intensity, while the persistence of large, significant effects across both of the leading US hospitals we have partnered with demonstrates that substantial imaging-specific variation remains.

We hope the review team finds our efforts to extend our results and carefully address the comments satisfactory in this round.

II. Responses to Department Editor (DE) Comments

DE Wrote (Overall assessment):

"The review team is on the border of rejection and high-risk major revision. In view of the potential practical importance of your main finding - batching of tests in ED reduces patient flow - and the large and therefore, if true, a practically very meaningful effect, I would like to encourage you to revise the paper.

This revision will be very challenging. The AE report provides some guidance. My main concern is the effect size on ED LOS. Why is it so large? You will have to provide a more granular story, including more information and empirical evidence on mechanisms (perhaps exploring variation in the batch tendency of physicians or in timing of test returns), and on subsamples where the identification assumptions that you make are most likely to be satisfied."

Response: We are grateful to the Department Editor for the opportunity to revise this manuscript and for the specific guidance on strengthening our empirical approach. We have taken the DE's concerns about effect magnitude and identification seriously, and our revision directly addresses each point raised.

Regarding the large effect size, we have clarified that the large magnitude of the effect is partially due to our interest in (and reporting of) LATE estimates rather than other measures such as ATE. Furthermore, the type of imaging tests we study are the ones known to be the major sources of delays in EDs. Each additional test is indeed known to significantly affect ED LOS. Nevertheless, following the comment, we have taken two approaches. First, our enhanced specification—which adds controls for laboratory ordering, physician characteristics, and ED capacity—reveals that part of the original 130% effect reflected correlation with general diagnostic intensity rather than imaging-specific batching behavior. Our revised estimates of 65–81% reflect this. Second, we now provide a more precise mechanistic explanation: imaging turnaround times at our study site average 60–90 minutes per modality, so the 1.2 additional tests we observe for discretionary batching compound substantially. Because batching increases the total imaging performed without the filtering effect of sequential information revelation, the resulting delays accumulate mechanically, yielding relatively large effect sizes.

Regarding a more granular story and mechanisms, we have added several analyses. Our mediation analysis (Section 4.3) demonstrates that the effect of batching on LOS operates primarily through increased imaging volume and higher admission probability—not through the batching process itself. Our heterogeneity analysis by chief complaint (Figure 4) shows consistent patterns across seven clinical categories, from simple presentations (fevers, extremity complaints) to complex polytrauma cases. Our analysis of determinants (Table 6) reveals that batching decreases as physicians progress through their shifts and during major overcapacity, suggesting that time pressure and resource constraints induce more selective ordering.

Regarding subsamples where identification assumptions are most likely satisfied, our complier analysis (EC.4) characterizes the approximately 20% of patients whose testing strategy depends on physician

assignment—precisely the margin where our assumptions hold by construction. The close agreement between our 2SLS estimates (using a single batch tendency instrument) and the newly added UJIVE estimates (using the complete set of provider indicators with leave-one-out bias correction) provides strong validation. These two approaches rely on different constructions of identifying variation, yet yield similar results, increasing confidence that many-instrument bias is not driving our findings and that the underlying source of variation is robust to alternative estimation strategies. Finally, our replication at our second partner hospital (MGH)—a completely different institutional setting with non-random assignment—shows directionally similar and statistically indistinguishable effects, demonstrating external validity beyond results obtained at our other partner hospital (Mayo Clinic) within its specific operational context.

These revisions provide the granular mechanistic evidence and robust identification that the DE requested, while maintaining transparency about the limitations inherent in any observational study. We hope the DE and the entire review team find our efforts to carefully address the comments we received to their satisfaction.

III. Responses to Associate Editor (AE) Comments

AE Wrote (Overall assessment):

"The study addresses an innovative and interesting research question, and the main finding is indeed surprising. On the surface, simultaneous ordering of diagnostic tests would appear to improve efficiency; this counterintuitive result raises important questions.

However, the referees have raised significant concerns regarding theory and methodology. In its current form, the paper does not sufficiently develop the theoretical foundation necessary to support its empirical strategy or claims of contribution."

Response: We thank the Associate Editor for the thoughtful summary and for highlighting both the novelty of the question and the need for stronger theoretical and methodological grounding. We have undertaken substantial revisions that directly address these concerns.

First, we strengthened the empirical strategy by expanding the control set to differentiate imaging-specific practice styles from broader diagnostic comprehensiveness. Incorporating detailed patient characteristics, laboratory test ordering, physician characteristics, hours into shift, and ED capacity levels allows us to isolate the variation in imaging behavior that is orthogonal to general diagnostic intensity. The revised estimates are more conservative but remain economically and statistically meaningful.

Second, we added the Unbiased Jackknife IV Estimator (UJIVE), following **?**, which is purpose-built for leniency designs with many decision-makers. When 2SLS uses many physician indicators as instruments, it constructs estimated leniency partly from each patient's own treatment status, creating mechanical correlation with the outcome error that biases estimates toward OLS. UJIVE eliminates this bias through leave-one-out estimation, constructing each patient's predicted treatment using only other patients' data. The close agreement between our 2SLS results (using a single batch tendency instrument) and UJIVE estimates (using physician indicators directly) provides important validation: it indicates that many-instrument bias is not driving our findings and that both approaches identify similar underlying variation. Across all outcomes, UJIVE estimates closely match our 2SLS results under the preferred full-control specification:

- 0.447 increase in log treatment time ($\approx 56\%$)

- 0.522 increase in log time to disposition ($\approx 68\%$)

- 0.503 increase in log LOS ($\approx 65\%$)

- 1.174 additional imaging tests ($\approx 88\%$)

Third, we clarified the theoretical framework for discretionary batch ordering. The revised Sections 2.3 and 3.3 explicitly distinguish between clinically mandated imaging and the discretionary margin where physician practice style determines the approach. Our identification strategy targets this

5

margin, by focusing on "compliers" and estimating LATE values, and the complier analysis confirms that physician assignment affects testing strategy for approximately 20% of patients—precisely the subgroup relevant for operational policy.

These revisions tighten the methodological foundation and make the contribution clearer: the paper identifies and quantifies the operational consequences of discretionary batch ordering using quasi-random physician assignment, rich clinical data collected from two leading US hospitals, and modern IV estimators designed for heterogeneous providers. We believe the revised manuscript now provides the theoretical clarity and empirical rigor appropriate for publication in *Management Science*.

AE Wrote (Theory):

"The theoretical foundation requires significant clarification. Key questions remain unaddressed:

- *Is batching truly a discretionary decision by the physician, or is it also driven by, perhaps unobserved clinical factors which affect ED LOS?*
- *Are additional tests medically required at the time of ordering, or are they preemptively ordered to buy time in anticipation of further needs?*
- *Could batching be a consequence of prior diagnostic uncertainty (e.g., pending results, routine requirements for referrals) or contextual factors such as utilization of the unit?"*

Without a clearer articulation of the decision-making process and its drivers, the interpretation of the results remains ambiguous.

Response: We thank the Associate Editor for these important theoretical questions, which have led us to clarify our framework substantially. We address each concern in turn.
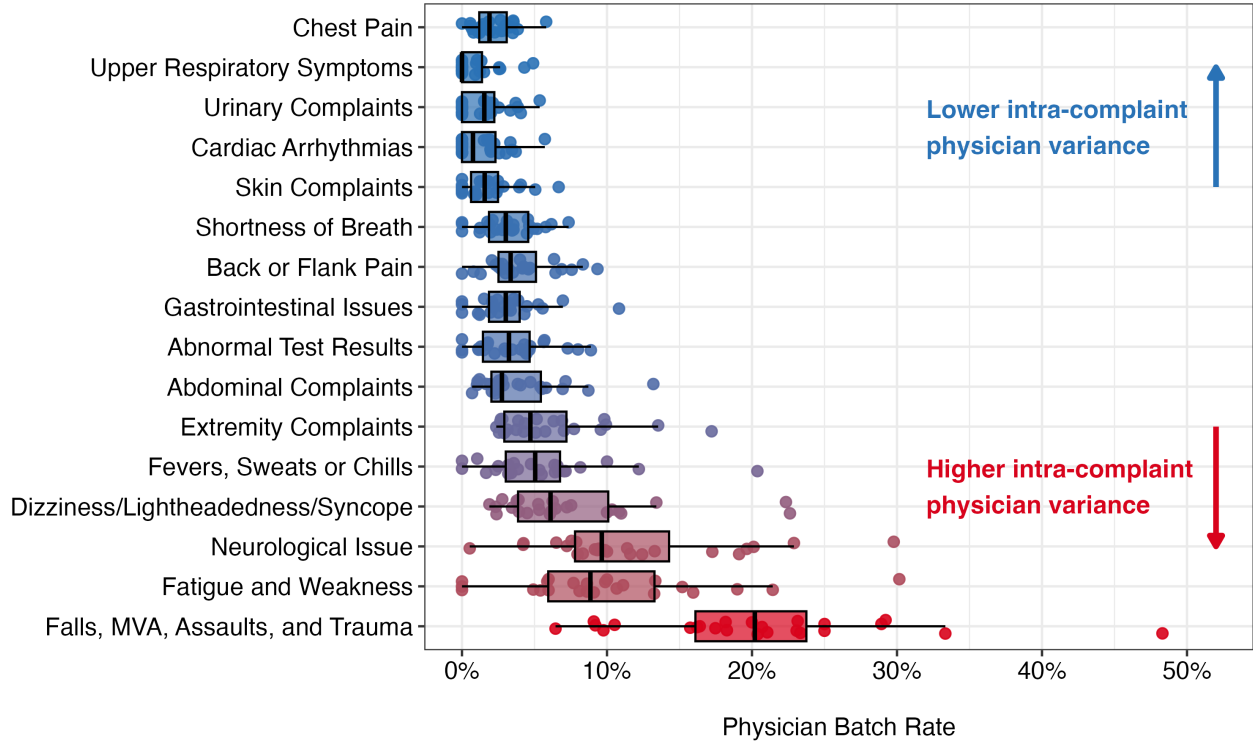
1. Is batching discretionary?

In various discussions, our medical collaborator have emphasized that there are clinical scenarios in which batching is necessary (e.g., polytrauma requiring multiple imaging modalities for comprehensive assessment) and scenarios where it is clearly unnecessary (e.g., uncomplicated urinary tract infections). At the same time, physicians vary substantially in their propensity to batch for clinically similar patients (Figure 1), indicating that practice style plays an important role.

Our empirical strategy isolates the discretionary component of batching—variation driven by practice style, not clinical necessity. First, Mayo Clinic's rotational assignment system provides quasi-random assignment of patients to physicians (Traub et al., 2016a,b, 2018). Assignment is based solely on a rotational basis and is independent of patient characteristics. Figure 2 confirms that while patient characteristics strongly predict whether a given encounter is batched (left panel), they do not predict assignment to high- vs. low-batch-tendency physicians (right panel). This supports the interpretation that practice-style differences, not case mix, generate the variation we exploit.

Second, our Local Average Treatment Effect (LATE) identifies effects for "compliers"—patients whose testing strategy depends on physician assignment rather than clinical protocols. These patients, by

Figure 1: Physician Variation in Batch Ordering Imaging Tests

Notes: This figure highlights the marked differences among Mayo Clinic ED physicians in their propensity to batch order imaging tests. Batch rates are crude rates calculated by dividing the number of patient encounters where the physician batch ordered imaging tests for a complaint by the number of patient encounters they had with that complaint.

definition, represent the discretionary margin where physician preference determines whether their tests are batched. Our complier analysis in EC.4 reveals approximately 20% of patients fall into this category—lacking clear clinical indicators mandating either batching or non-batching.
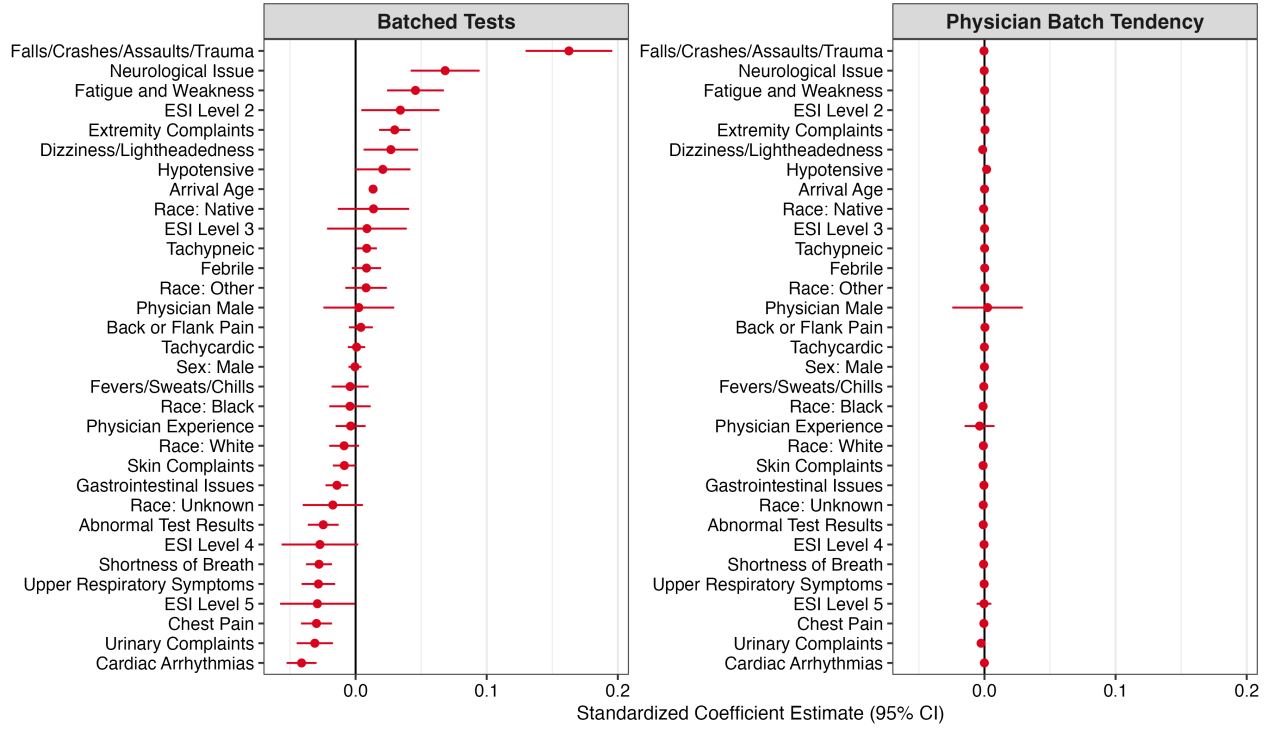
2. Are tests medically necessary or preemptively ordered?

The existence of compliers in our data—identical patients whose batching decision is based on physician assignment not clinical necessity—provides evidence that these tests are not medically necessary at the time of ordering. If all batching were medically necessary, physician-induced variation would be minimal.

These tests may ultimately prove clinically useful later in the encounter, but the margin identified by our design reflects cases where physicians themselves disagree. Some believe that ordering multiple studies up front is efficient; others believe this approach is wasteful and reduces opportunities to update testing based on initial findings. Our goal is to quantify the downstream efficiency and impacts of this discretionary batching margin.

3. Could batching be a consequence of diagnostic uncertainty or contextual factors?

Figure 2: Batch Tendency by Patient Characteristics

Notes: This figure plots a test for quasi-random assignment of patients to physicians in the Mayo Clinic ED. The left panel shows how patient characteristics predict batching decisions. The right panel shows these same characteristics do not predict assignment to physicians with different batch tendencies. Residualization fixed effects include hospital-year-month, hospital-day of week-time of day.

Our focus on early batching (within 5 minutes of encounter initiation) rules out pending results as a driver. At this point, no prior tests have returned results. Late adaptive batching in response to test results occurs in only 1.9% of multi-test encounters, confirming that pending results do not drive the batching decisions we study.

If batching were driven by standardized referral protocols (e.g., all trauma patients require specific imaging for specialist consultation), we would observe minimal physician-induced variation (i.e., all physicians would batch the same patients). Instead, our complier analysis reveals 20% of patients receive different testing strategies based solely on physician assignment. Moreover, Mayo Clinic's random assignment mechanism ensures referral requirements are balanced across physicians, so systematic differences in batching rates reflect physician discretion rather than patient-driven protocols.

The random assignment mechanism directly addresses whether uncertain cases drive batching. If diagnostically uncertain patients systematically received more batching, we would observe that patient characteristics predict assignment to high-batching physicians. Figure 2 (right panel) shows this is not the case—patient characteristics do not predict batch tendency, confirming that physician practice style, not case complexity, drives the variation we exploit.

We directly examine whether contextual factors like ED capacity drive batching decisions. However, as Reviewer 3 correctly noted, our capacity-stratified analysis (Table 5) does not show statistically significant differences across capacity levels. We have removed claims about capacity heterogeneity and instead note that batching occurs across all operational contexts, with consistent operational inefficiencies (increased imaging without time savings) regardless of ED utilization levels.

Manuscript changes:

We have substantially revised Sections 2.3 (Hypothesis Development) and 3.3 (Identification Strategy) to clarify these theoretical foundations, emphasizing:

- The distinction between clinically-necessary and discretionary batching.

- Our LATE estimator targets the discretionary margin: patients whose batching strategy depends on the physician discretion and where ED interventions can improve efficiency.

- The role of enhanced controls in validating that we capture imaging-specific workflow decisions.

- Interpretation of complier effects and what they reveal about physician decision-making.

We appreciate the Associate Editor's guidance, which has enabled us to provide a much clearer theoretical foundation and address concerns about whether we truly isolate discretionary behavior.

AE Wrote (Comment – Empirical concerns):

"The observed effect with an increase of 130% in ED stay as a result of batching is striking and, in fact, lacks credibility. Such a large effect increases the likelihood of confounding and therefore demands a very high degree of confidence in the empirical approach. As pointed out by several reviewers, there are unresolved concerns regarding the physician-patient assignment process. It is unclear whether this assignment is truly exogenous. For example, some physicians might only work certain shifts or treat specific patient types.

All reviewers were critical of the empirical strategy. Nevertheless, I believe the data offer enough potential to address these concerns through a substantial revision. This would require a significant improvement of both the theoretical development and the empirical analysis."

Response: We appreciate the Associate Editor's candid assessment, which precisely identified the central challenges. Our revised analysis directly addresses the concerns raised.

1. Assignment process and exogeneity:

Concerns about endogenous assignment were also raised by Reviewer 1, and we have substantially strengthened Section 3.1 to provide a clearer, more detailed justification of exogeneity. We now document the Mayo Clinic's rotational assignment algorithm in depth, clarifying that the algorithm provides a quasi-random assignment removing endogenous concerns. We discuss prior published work (Traub et al., 2016a,b, 2018) and new operational details provided to us by our physician co-author who is in charge of operations at our study site.

We have added the following detailed description to Section 3.1:

> "The Mayo Clinic ED employs a sophisticated computerized rotational patient assignment algorithm that addresses many empirical challenges in healthcare settings; see, e.g., Traub et al. (2016a,b, 2018). The system automatically assigns patients completely on a rotational basis to physicians 60 seconds after registration through the electronic health record system. At shift start, each physician receives four consecutive patients to establish an initial patient load, after which they enter rotation with other on-duty physicians. The rotation order is predetermined by the ED scheduler and varies across shifts to ensure fairness over time. Critically, these assignments are based solely on arrival time—the algorithm does not consider patient demographics, chief complaint, Emergency Severity Index score, physician workload, or the acuity of recently assigned patients. To maintain system integrity, physicians receive no new patients during their final 120 minutes and are capped at 18 patients per shift.
>
> Our empirical tests show that this rotational mechanism achieves the quasi-randomization necessary for causal inference. Unlike settings where patient-physician matching may be influenced by triage decisions, physician preferences, or informal routing practices, the Mayo Clinic's algorithmic assignment removes discretion from the matching process."

This institutional design rules out the two major concerns that certain physicians might selectively receive specific patient types, and that shift timing might generate systematic differences in the types of cases. The variation used for identification comes from within-shift rotation, and we confirm empirically that patient characteristics predict batching decisions (Figure 2, left panel) but not assignment to high- vs. low-batching physicians (Figure 2, right panel). This pattern is what we would expect under a quasi-randomized assignment mechanism.

Further, our instrument is constructed by residualizing physician-level batching on shift fixed effects, ensuring that the identifying variation is within physician, within shift-type, and orthogonal to systematic differences across time of day, weekday, or ED load. This ensures that the IV captures only practice-style-driven variation in batching propensity, not structural differences across shifts or physician schedules.

2. Effect magnitude and confounding:

Regarding the magnitude of the effect, it should be noted that our estimate is based on LATE, which differs from other measures such as ATE. Nonetheless, we agree with the Associate Editor, that the large effect could be due in part to violations of the exclusion restriction. In response to the review team's concerns about confounding, we substantially expanded our control set to ensure our instrument captured imaging-specific behavior rather than general diagnostic intensity. In our original specification, we controlled only for patient characteristics (vital signs, age, demographics, chief complaint by severity) and temporal factors (shift-level effects). In the revised version, we have substantially expanded our control set to better isolate imaging-specific ordering behavior.

First-Stage Robustness: Batch Tendency Strongly Predicts Batching

	Dependent Variable: *Batched*	
	(1)	(2)
Batch Tendency	1.837***	1.752***
	(0.1379)	(0.1336)
Fixed Effects		
Necessary Controls	Yes	Yes
Precision Controls	No	Yes
Observations	11,679	11,679
R^2	0.02248	0.07359
First-stage F-stat	177.5	171.9

Notes: First-stage regressions of batching on batch.tendency (leave-one-out physician residualized batching propensity). Necessary controls include day-of-week ×time-of-day and month fixed effects. Precision controls include patient vital signs, age, demographics, chief complaint-severity fixed effects, race, gender, physician experience, physician gender, hours into shift, laboratory ordered, and ED capacity controls. *** p<0.001.

Our revised primary specification now includes:

- **Patient characteristics:** Vital signs (tachycardic, tachypneic, febrile, hypotensive), age, demographics, chief complaint by severity

- **Physician characteristics:** Years of experience, gender, hours into shift (capturing potential fatigue effects)

- **Realized laboratory diagnostic intensity:** Laboratory tests performed during the encounter

- **Contextual factors:** ED capacity level (normal, minor overcapacity, major overcapacity),

- **Shift level FE:** Month-year and day-of-week × time-of-day fixed effects

By controlling for laboratory tests ordered for each patient, we test whether imaging effects persist when accounting for the thoroughness of general diagnostics. If batch tendency simply reflected "comprehensive diagnosticians," controlling for realized lab intensity should substantially attenuate the instrument's relevance.

First-stage results validate the concern while demonstrating that we can isolate the batching-specific component. Below, is our revised first-stage regression (Table 2), both without and with the expanded precision controls.

We see that even when controlling for this comprehensive set of precision controls, substantial independent variation remains (F=171.9, well above weak instrument thresholds). This is precisely what we would expect if the instrument captures batching-specific propensity rather than general cautiousness.

To complement our 2SLS specification, we also incorporate the Unbiased Jackknife Instrumental Variables Estimator (UJIVE), following **?**. UJIVE addresses a specific concern when using many physician indicators as instruments: 2SLS constructs estimated leniency partly from each patient's own treatment status, creating mechanical correlation with the outcome error that biases estimates toward OLS. UJIVE eliminates this many-instrument bias through leave-one-out estimation. The agreement between our 2SLS results (using a single batch tendency instrument) and UJIVE estimates (using physician indicators directly) provides important validation that many-instrument bias is not driving our findings.

Our newly added results (Table 4) show close agreement between 2SLS and UJIVE—both with and without precision controls, which increases confidence that both estimators identify batching-specific variation.

The attenuation in the 2SLS estimates after adding precision controls confirms the AE's intuition that our original instrument may have contained some correlation with general diagnostic intensity. At the same time, the close agreement between the 2SLS and UJIVE estimates—both before and after adjustment—indicates that many-instrument bias is not driving our results and that both approaches identify similar underlying variation. This convergence, combined with our expanded controls for laboratory ordering and physician characteristics, strengthens confidence that our revised specification isolates batching-specific effects rather than broader diagnostic style.

Additionally, following an approach outlined by Conley et al. (2012), in the revised version we further test how our 2SLS estimates change if one violates the exclusion restriction. Under standard IV assumptions, the structural equation is

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

which assumes that batch tendency affects outcomes only through its effect on Batched_i.

Conley's approach allows for a small violation of the exclusion restriction by introducing a direct effect of the instrument:

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

where δ captures any exclusion restriction violation, including
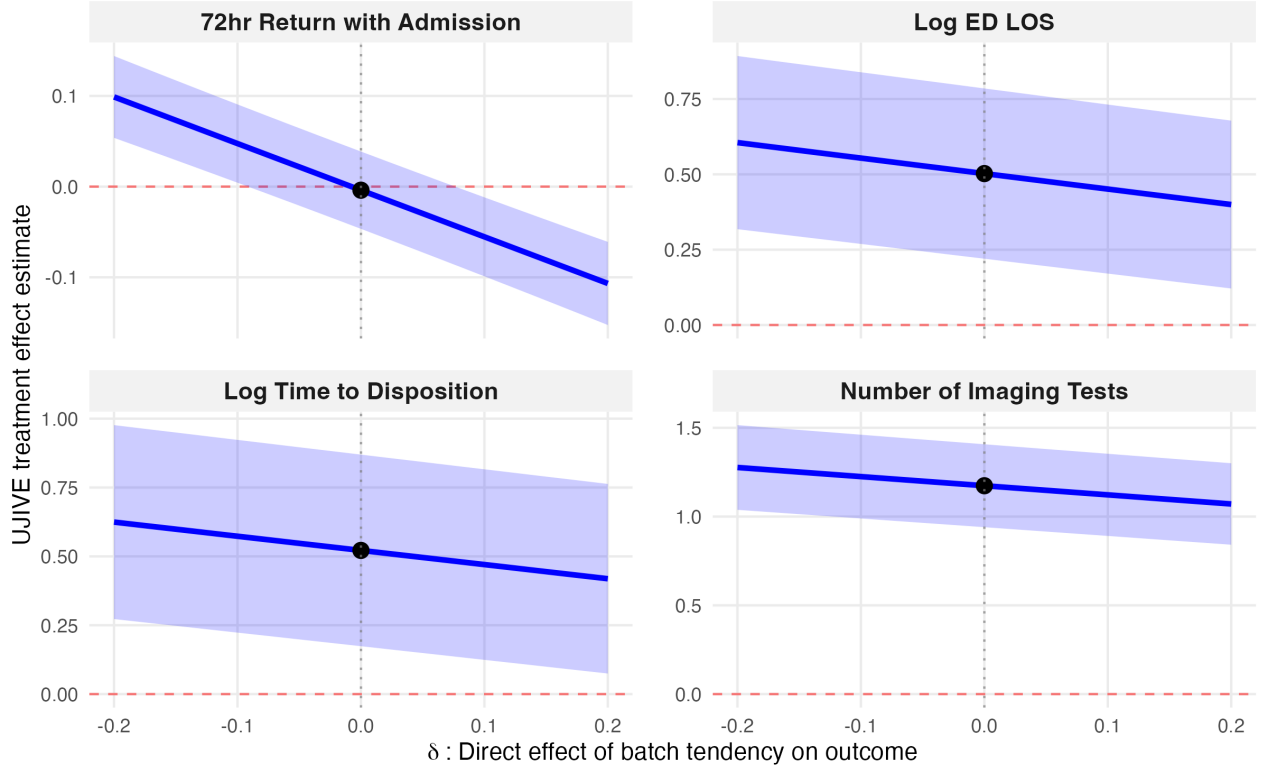
Table 4: Effect of Batching Tests on Patient Outcomes

	Sequenced mean	2SLS		UJIVE	
		(2)	(3)	(4)	(5)
Panel A. Primary Outcomes					
Log time to disposition	5.237	0.659***	0.651***	0.583***	0.522***
	(0.499)	(0.103)	(0.101)	(0.189)	(0.177)
Log LOS	5.490	0.717***	0.597***	0.653***	0.503***
	(0.456)	(0.094)	(0.088)	(0.158)	(0.144)
Number of distinct imaging tests	1.335	1.385***	1.241***	1.316***	1.174***
	(0.572)	(0.118)	(0.116)	(0.126)	(0.119)
72hr return with admission	0.012	-0.0137	-0.0146	-0.0079	-0.0039
	(0.110)	(0.018)	(0.019)	(0.020)	(0.022)
72hr return	0.030	-0.0512	-0.0536	-0.0440	-0.0396
	(0.170)	(0.029)	(0.031)	(0.032)	(0.034)
Panel B. Test Types					
X-ray	0.576	0.943***	0.989***	0.960***	0.959***
	(0.494)	(0.100)	(0.101)	(0.116)	(0.117)
Ultrasound	0.171	0.160**	0.087	0.164*	0.087
	(0.377)	(0.076)	(0.073)	(0.082)	(0.078)
CT without contrast	0.400	0.102	0.062	0.052	0.053
	(0.490)	(0.095)	(0.086)	(0.112)	(0.102)
CT with contrast	0.187	0.180*	0.102	0.140	0.075
	(0.390)	(0.078)	(0.076)	(0.087)	(0.079)
Panel C. Disposition					
Admission	0.279	0.419***	0.404***	0.424***	0.398***
	(0.449)	(0.096)	(0.088)	(0.103)	(0.090)
Necessary controls	—	Yes	Yes	Yes	Yes
Precision controls	—	No	Yes	No	Yes
Observations	11,679	11,679	11,679	11,679	11,679

Notes: Column 1 reports means for sequenced patients (standard deviations in parentheses). Columns 2–3 report 2SLS results using physician batch tendency as the instrument. Columns 4–5 report UJIVE estimates using ED provider identifiers as many weak instruments. All models include day-of-week and month fixed effects. Precision controls described in text. Standard errors are heteroskedasticity-robust.

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

- direct effects (e.g., physicians with higher batch tendency directly prolong LOS),

- backdoor paths (e.g., batch tendency reflects physician aggressiveness, which also affects LOS),

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

Notes: UJIVE estimates under different assumptions about direct effects (δ) of physician batch tendency on outcomes. One unit of δ implies a one percentage point increase in batch tendency directly changes the outcome by δ units. Black points indicate baseline estimates. Shaded areas show 95% confidence intervals using heteroskedasticity-robust standard errors following Conley et al. (2012).

We then estimate $\beta(\delta)$ for a range of plausible values of δ to assess how sensitive our conclusions are to such violations.

The figure below (now included in EC.9) presents sensitivity analyses showing substantial robustness. For all outcomes, the treatment effect remains significant across $\delta \in [-0.2, 0.2]$.

In summary, while we cannot definitively rule out all exclusion restriction violations—no observational study can—the convergence of evidence strongly supports our interpretation:

- Treatment effects persist after adding addition controls for general diagnostic intensity (laboratory ordering), experience, and hours into shift

- UJIVE estimates, which eliminate many-instrument bias, closely match 2SLS results, indicating both approaches identify similar underlying variation

- 2SLS estimates remain robust to potential backdoor paths for plausibly large exclusion restriction violations

14

Finally, as noted earlier, the LATE we identify represents the causal effect for compliers—patients whose imaging strategy depends on physician practice style rather than clinical necessity. This is precisely the margin where ED management interventions can influence practice without constraining clinically necessary care. We have conferred with our physician collaborators and co-authors of this manuscript at both study sites (two leading US hospitals) that this is exactly where the debate among ED physicians lies.

We have revised the manuscript to reflect these more conservative estimates while maintaining transparency about our identification assumptions. The core finding are that discretionary batch ordering increases ED length of stay by approximately 65-81% without improving short-term patient outcomes, suggesting substantial opportunity for operational improvement through policies that preserve diagnostic flexibility. Given that imaging turnaround times at our study site average 60–90 minutes per modality, even modest increases in the number of imaging tests ordered can compound substantially. Because discretionary batching increases the total imaging performed (see our mediation analysis), the resulting delays accumulate mechanically, making effect sizes of this magnitude operationally plausible in this setting.

We are grateful to the AE and reviewers for pushing us to strengthen our empirical strategy, and we believe the revisions substantially improve the credibility and clarity of the findings.

AE Wrote (Comment – Hypothesis development): *"The current framing does not clearly distinguish between "batching" and "sequencing" of diagnostic tests. A more precise hypothesis is needed. See also Ref 2 and Ref 3."*

Response: The Associate Editor is correct, and Reviewers 1 and 2 raised this identical concern. We have reframed our comparison throughout the manuscript as "batch ordering versus standard practice" rather than "batching versus sequencing."

The key clarification is that our counterfactual includes both sequential multi-test encounters and encounters where only a single test is ordered. The single-test patients would have likely received multiple tests if they were assigned to a physician with higher batch tendency; they received only one test because the non-batching practice allowed the physician in charge to remove the need for further tests after observing the result of the first one ("information gain" effect). Our analysis reflects the information problem physicians face: they cannot know ex ante which patients will ultimately need multiple tests. Thus, the relevant comparison is not between two fixed protocols, but between committing upfront to comprehensive imaging versus preserving the option value of information from initial results. Several changes implement this reframing:

Section 2.3 (Hypothesis Development): We now explicitly frame hypotheses as "batch ordering versus standard practice," emphasizing that standard practice retains diagnostic flexibility.

Section 3.3 (Identification): We added language clarifying that the LATE compares batch ordering to standard practice (a composite of single-test and sequential-test encounters). While this yields

a composite counterfactual, it identifies the policy-relevant parameter: the effect of encouraging comprehensive upfront imaging versus allowing information to accumulate before ordering additional tests.

Abstract and Introduction: We now consistently use the "batch ordering versus standard practice" terminology and highlight that the LATE pertains to the discretionary margin—patients for whom physicians genuinely differ in their preferred testing strategy.

We hope these revisions directly address the concern about conceptual precision while preserving an accurate interpretation of what our design identifies.

AE Wrote (Comment – Outcome variable):

"Consider integrating intermediate outcomes / mechanisms, such as the order and type of tests performed. Are all tests necessary in each case? ED length of stay is a problematic outcome due to factors like patient heterogeneity, workload fluctuations, and staffing variability. The reviewer comments offer several suggestions here."

Response: We have substantially enhanced our analysis to address each concern the Associate Editor raises.

Regarding intermediate outcomes and mechanisms, Reviewer 3 raised concerns about reverse causality in Panel B of Table 4, which examines specific test types. We clarified that our IV approach—instrumenting batching with physician tendency calculated from other patients—breaks the endogeneity between patient-specific test needs and batching decisions. The results reveal that discretionary batching nearly universally adds X-rays, while having more minor, non-significant effects on other forms of imaging. This pattern indicates physicians construct comprehensive workups by adding quick, low-cost tests rather than selectively ordering expensive imaging.

Reviewer 2 raised the identical question about whether additional tests represent overuse or appropriate care. While we cannot determine whether specific individual tests are clinically unnecessary, our complier analysis provides definitive evidence of discretionary ordering: 20% of patients receive different testing strategies based solely on physician assignment. For these marginal patients, batching leads to 1.2 additional imaging tests performed without improving quality outcomes as measured by 72 hour returns with admission. This pattern—increased resource utilization without measurable clinical benefits—suggests that at the margin where physicians exercise discretion, additional tests do not provide commensurate value.

The AE correctly notes that ED LOS captures multiple processes beyond physician decision-making. Following Reviewer 3's suggestion, we added treatment time (time from physician assessment to disposition) as an alternative outcome that excludes both waiting room delays and post-disposition boarding. The results (included below and in EC.11) show nearly identical patterns. This consistency demonstrates that batching affects clinical processing time, not just mechanical delays.

Our enhanced specification directly addresses the factors the AE identifies as problematic. For pa-

16

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

	Sequenced mean	2SLS		UJIVE	
		(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	11,679	11,679	11,679	11,679	11,679

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 many weak instruments. All models include day-of-week and month fixed effects. Precision controls include patient vital signs, demographics, complaint-severity fixed effects, physician characteristics, laboratory ordering, hours into shift, and ED capacity. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

tient heterogeneity, we control for vital signs, age, demographics, chief complaint severity, ESI level, and laboratory testing, with Figure 2 confirming these characteristics do not predict assignment to different physician types. For workload fluctuations, we include ED capacity level indicators (normal, minor overcapacity, major overcapacity) and comprehensive temporal fixed effects (day-of-week × time-of-day, month-of-year) to account for systematic demand patterns. For staffing variability, we control for physician experience, gender, and hours into shift to capture systematic differences in practice patterns and potential fatigue effects.

The robustness of our results across these comprehensive controls provides confidence that we capture the effects of genuine physician practice variation rather than confounding from operational factors. Together, we believe these revisions address the AE's concerns about outcome measurement while providing convergent evidence that discretionary batching increases resource utilization without improving efficiency or quality.

AE Wrote (Comment – Relevance):

- *"Focusing on subsets of patients or refining outcomes may help understand and explain the large effect size."*
- *"A counterfactual cost-benefit analysis could strengthen the case for the practical relevance of the findings."*

Response: Thank you for these comments. We have addressed both suggestions to strengthen the practical relevance of our findings.

Regarding patient subsets and refined outcomes, our revisions directly explain the initially reported large effect size. First, our enhanced control set (adding laboratory ordering, physician characteristics, and hours into shift) reveals that the original 130% LOS increase was partially driven by general diagnostic intensity rather than image-specific batching behavior. Our revised estimate presented in Table 4 provide a more accurate and credible magnitude, though the effect remains substantial (65-81% increase in LOS).

Second, following Reviewer 3's suggestion, we conducted a heterogeneity analysis across seven chief complaint categories, which vary in clinical complexity and batching prevalence (Figure 4). This exploratory analysis reveals consistent patterns—increased imaging without time savings or quality improvements—across all complaint types, from simple presentations (fevers, extremity complaints) to complex polytrauma cases. The consistency across subgroups strengthens confidence in our main findings while demonstrating that effects do not depend on specific patient populations.

Third, we refined our time-based outcomes by adding treatment time (Reviewer 3's suggestion), which excludes both waiting room delays and post-disposition boarding to isolate physician processing time. The similar patterns across time to disposition and treatment time confirm that batching affects clinical decision-making processes, not just mechanical delays. We also compared excluded versus included encounters (Reviewer 3's request), demonstrating that our analytical sample appropriately focuses on moderate-to-high acuity patients where imaging decisions are both consequential and discretionary.

Regarding cost-benefit analysis, we developed a cost-benefit analysis quantifying the operational and financial burden of discretionary batching (detailed in our response to Reviewer 2 and EC.6). Under deliberately conservative assumptions, which include include only statistically significant imaging effects and excluding downstream costs from increased admissions, discretionary batching imposes approximately $349 in direct costs per affected patient: $125 from excess imaging and $224 from excess ED capacity utilization. Scaling to annual operations, this amounts to approximately $232,000 per year for a medium-sized ED (40,000 annual visits). These costs arise from the approximately 20% of patients whose batching decisions reflect physician preference rather than clinical necessity, with no offsetting benefits in 72-hour return rates. Even a 50% reduction in discretionary batching would save approximately $113,000 annually while freeing 850 bed-hours of capacity.

These enhancements explaining the effect magnitude through enhanced controls, demonstrating consistency across patient subsets, refining outcome measures, and developing cost-benefit implications—substantially strengthen the practical relevance of our findings for ED management.

AE Wrote (Comment – Additional suggestions):

"The following additional detailed suggestions could help the authors for developing the manuscript:

- *Reassess the definition of the treatment variable (e.g. Ref 3).*
- *Examine the independence of test orders from prior tests and workload conditions (e.g. Ref 2).*

18

- *Include a section discussing sample representativeness.*

- *Expand the empirical strategy to address unobserved physician and non-physician factors. For instance, Ref 1 recommends using physician fixed effects. Also, reconsider model selection in light of the main outcome.*

- *Re-evaluate whether ED length of stay should remain the primary outcome. Alternative metrics such as time from test order to result may offer more insight."*

Response: Thank you. We have addressed each of these suggestions through substantial revisions responding to the three reviewers.

1. Definition of the treatment variable

Reviewers 1 and 2 noted that our original language created ambiguity between "batching" and "sequencing." As noted earlier, we have reframed the treatment as batch ordering versus standard practice, where standard practice includes both single-test encounters and sequential multi-test encounters. Section 3.3 now clarifies that this is a composite counterfactual, but it is precisely the policy-relevant comparison: committing to comprehensive upfront imaging versus preserving diagnostic flexibility when the appropriate testing pathway is uncertain. We also explain that our LATE pertains to patients at the margin of physician discretion, not those with clear clinical indications. This reframing resolves the theoretical concern and aligns with Ref 3's emphasis on clarity in treatment definition.

2. Independence from prior tests and workload

Reviewer 2 raised concerns about whether batching reflects adaptive responses to prior tests or current workload. To address this, we emphasize that our definition of batching is limited to orders placed within the first 5 minutes of the encounter. At this stage, no imaging or laboratory results have returned, and adaptive testing is not yet possible. We also document that "late batching"—i.e., ordering a second image only after the first has returned—accounts for only 1.9% of multi-test encounters, confirming that our treatment definition isolates upfront ordering decisions rather than responses to evolving diagnostic information.

3. Sample representativeness

Reviewer 3 asked for a clearer discussion of representativeness. We now provide a CONSORT-style flow diagram (Figure EC.3.1) detailing all exclusion steps. Table EC.3.2 compares included and excluded encounters on key demographics, vital signs, and complaint categories. As discussed in Section 3.2, the analytic sample appropriately focuses on higher-acuity patients whose presenting complaints plausibly require imaging. This is the clinically meaningful population where discretionary batch ordering has operational consequences. We characterize representativeness in that domain rather than for the ED population as a whole, which we now state explicitly. We additionally characterize compliers in EC.4, confirming that they span a broad range of demographics and complaint types, enhancing external validity.

19

4. Physician fixed effects and model selection

The AE and Reviewer 3 both raised the need for robustness to unobserved provider characteristics. As noted earlier, we now incorporate the Unbiased Jackknife Instrumental Variables Estimator (UJIVE) into our main robustness analysis. UJIVE uses the full set of physician identifiers as instruments and eliminates the many-instrument bias that arises when 2SLS constructs estimated leniency partly from each patient's own treatment status. The close agreement between our 2SLS and UJIVE estimates—despite their very different constructions—indicates that many-instrument bias is not driving our findings. Our expanded controls and Conley sensitivity analysis address the exclusion restriction concerns highlighted by the AE directly. With full controls, the UJIVE estimates are slightly attenuated but remain large and precise (e.g., log LOS: 0.503; time to disposition: 0.522; treatment time: 0.447), reinforcing that the effects we identify are not artifacts of unobserved provider heterogeneity.

We also addressed the concerns about model selection raised by Reviewer 3. EC.10 now includes logit models for binary outcomes and negative binomial models for count outcomes. Across all outcomes, the resulting average marginal effects align closely with those from our linear IV specifications. We retain linear 2SLS in the main text because nonlinear IV models violate core econometric requirements (the "forbidden regression"), a point we clarify in our detailed response to Reviewer 3. The combination of (i) FE-based IV, (ii) UJIVE estimation, and (iii) nonlinear outcome models provides strong evidence that our findings are robust to alternative estimators and modeling choices.

4. Alternative outcomes

We now additionally examine outcomes reflecting more proximate process measures: treatment time (provider contact to decision) and time to disposition (provider contact to disposition decision). These outcomes remove waiting room delays and post-disposition boarding. EC.11.1 reports that the pattern remains consistent across all three measures: batch ordering increases treatment time by 56%, time to disposition by 69%, and total LOS by 65%. Because each metric captures a distinct operational stage of the encounter, we retain all of them in Table 4 and EC.11 to provide a more complete characterization of the effects.

Together, these revisions—clarifying treatment definition, demonstrating independence from prior information, explicitly documenting sample representativeness, adding robustness to unobserved provider factors, and expanding outcome measures—address all five AE suggestions and align the manuscript with the methodological guidance in the referenced literature.

IV. Responses to Referee 1 (R1) Comments

R1 Wrote (Research question & contribution):

"The research question is clear, and it is an important area of study. Emergency departments are high-stress work environments where the stakes are high and patient lives are at stake. Anything we can learn to improve the performance and outcomes of the work that is done there could save lives, and also enhance our understanding of the impact of "work behavior /preferences" on operational outcomes."

Response: We are grateful for the reviewer's recognition of the importance and clarity of our research question. Emergency departments represent uniquely high-stakes environments where operational improvements can have significant implications for patient care. We share the reviewer's view that understanding how physician work behavior affects operational outcomes can bridge individual decision-making and system-level performance. We have strengthened our analysis in response to the methodological concerns raised, and believe the revisions substantially improve the manuscript's rigor and contribution.

R1 Wrote (Major Concern 1):

"The authors apply an interesting method from economics to exploit a setting where patients are assigned to physicians in a random fashion. While this is a nice idea, I have two major concerns about the implementation of the methods in this paper. First, it is unclear to me whether the main explanatory variable is actually testing the main hypothesis presented in the paper, and second, the exclusion restriction assumption (required for an IV to be valid in presenting causal estimates) is not met.

On page 4 of the manuscript (Section 1.2 – Main Findings and Contributions), the authors state "our results show that the marginal batched patient experiences a 130% increase in total ED LOS and an 123% increase in time to disposition compared to patients who have their tests ordered sequentially". It is not clear to me that you are in fact testing the effect of batching compared to sequential test ordering. Here, the main explanatory variable is Batched, which is defined as a binary variable that equals 1 if the physician ordered 2 or more imaging tests of different modalities within a five-minute window at the start of the patient's visit and 0 otherwise. This otherwise, acting as a counterfactual baseline, covers not only cases where the physician is ordering the same number of tests sequentially. It also covers cases where the physician orders fewer tests, or none at all. As such, the Batched variable is actually modeling some element of physician practice style like "degree of cautiousness" or "affinity for comprehensive testing". So, it is unsurprising that patients exposed to high-"batch tendency" physicians also have higher testing volumes and higher LOS – if a physician has a tendency to order more imaging, they may also have a tendency to order more labs or be more comprehensive in other ways that are not observed in the data.

If you are interested in studying the impact of batching imaging orders versus sequentially ordering them, as is currently discussed in the paper, then perhaps consider matching patients who have similar conditions and who we know ultimately have the same number of tests ordered – one patient would have had the orders batched, and the other would have them ordered sequentially. Though, I imagine doing something like this would result in limitations with respect to sample size, similar issues as in the current design where fewer tests are ordered when done sequentially (especially if the patient ends up being admitted and getting these tests when they are on an inpatient unit), among other endogeneity concerns.

Regardless, in its current manifestation, the batched variable is not capturing physician batching compared to sequential testing, and many of the managerial implications and conclusions currently presented in the paper do not logically follow from the presented results."

Response: We thank the reviewer for this thoughtful comment, which has led to substantial improvements in how we frame our research question, specify our models, and interpret our results. The reviewer raises two key concerns: (1) our counterfactual includes heterogeneous cases, and (2) our instrument might capture general physician cautiousness rather than batching behavior specifically. We address each in turn.

1. Clarifying the counterfactual: First, an essential clarification: our sample includes only encounters where at least one imaging test was ordered. The $Batched = 0$ group, therefore, comprises patients who received either (1) a single imaging test, or (2) multiple tests ordered sequentially—but not patients with zero imaging. This distinction is critical because our comparison is between different imaging strategies for patients who require diagnostic imaging, not between testing and no testing.

That said, the reviewer correctly identifies that our counterfactual remains composite—mixing single-test and sequential multi-test encounters. The reviewer suggests matching patients on eventual test count but correctly anticipates the key limitation: this would introduce severe post-treatment bias since batching causally impacts subsequent testing decisions. Physicians cannot know ex-ante which patients will ultimately require multiple tests—they make batching decisions under diagnostic uncertainty. Conditioning on the eventual test count would remove the causal pathway between batching and our outcome measures (e.g., LOS) and would also compare fundamentally different populations after the treatment has already occurred. Furthermore, matching would estimate the Average Treatment Effect on the Treated (ATT)—comparing patients who received batching to observably similar patients who did not. The ATT conflates effects across all patients, including always-takers (where batching is clinically mandated) and never-takers (where single tests suffice). This parameter has limited managerial relevance since ED protocols cannot change testing patterns where clinical necessity dictates the approach.

Our IV approach preserves causal interpretation by identifying effects at the moment of decision-making. The LATE captures effects for "compliers"—patients whose testing strategy depends on physician assignment rather than clinical necessity. These marginal patients represent precisely where ED protocols can influence practice without constraining clinically necessary care. We have conferred with our physician collaborators and co-authors of this manuscript at both study sites (two leading US hospitals) that this is exactly where the debate among ED physicians lies.

In response to this feedback, we have reframed our comparison throughout the manuscript as "batch ordering versus standard practice." This standard practice encompasses both sequential ordering and single-test cases among patients who receive at least one imaging test. The terminology change clarifies that we compare a discretionary practice pattern (early comprehensive imaging) against the standard approach (preserving diagnostic flexibility). We added text in Section 3.3 clarifying:

22

"Our two-stage least squares estimates represent the LATE of batch ordering for 'compliers'—patients whose testing strategy depends on the assigned physician's practice style. This effect compares batch ordering to standard practice, which includes both sequential ordering and single tests. While this involves a composite counterfactual, it provides the policy-relevant parameter: the effect of encouraging comprehensive upfront testing versus allowing diagnostic information to guide testing decisions for patients at the margin of clinical discretion—those whose testing strategy is not dictated by apparent clinical necessity but rather depends on physician practice style and judgment."

This framing reflects the fundamental information structure of emergency medicine. At the time of initial ordering, physicians cannot know which patients will ultimately need multiple tests versus a single test. They must decide whether to commit to comprehensive imaging upfront or preserve diagnostic flexibility (e.g., observe the result the first test and then decide if other tests are needed). We have added a discussion of this information problem in Section 3.2.1:

"We focus on batches that concern the first imaging tests ordered during the patient encounter because this represents the moment of maximum diagnostic uncertainty, when physicians must decide their testing strategy before clinical information unfolds. Physicians cannot know ex-ante which patients will ultimately require multiple tests, leading to instances in which batching is a discretionary choice based on practice style rather than clinical necessity."

We also clarified the policy relevance of this comparison in Section 5.1:

"This comparison reflects the real choice facing ED managers: should protocols encourage comprehensive upfront testing or preserve diagnostic flexibility? Our estimates show that preserving optionality through standard practice—which allows information from initial tests to guide subsequent decisions—reduces testing intensity and improved outcomes such as LOS."

2. Addressing the exclusion restriction concern through expanded controls:

The reviewer notes that if a physician has a tendency to order more imaging, they may also tend to order more labs or be more comprehensive in other ways. This is a valid concern that we take seriously.

In our original specification, we controlled only for patient characteristics (vital signs, age, demographics, chief complaint by severity) and temporal factors (shift-level effects). In the revised version, we have substantially expanded our control set to better isolate imaging-specific ordering behavior.

Our revised primary specification now includes:

Table 2   First-Stage Robustness: Batch Tendency Strongly Predicts Batching

	Dependent Variable: *Batched*	
	(1)	(2)
Batch Tendency	1.837***	1.752***
	(0.1379)	(0.1336)
Fixed Effects		
Necessary Controls	Yes	Yes
Precision Controls	No	Yes
Observations	11,679	11,679
R^2	0.02248	0.07359
Within R^2	0.01939	0.03521
First-stage F-stat	177.5	171.9

Notes: First-stage regressions of batching on batch.tendency (leave-one-out physician residualized batching propensity). Necessary controls include day-of-week ×time-of-day fandmonth fixed effects. Precision controls include patient vital signs, age, demographics, chief complaint-severity fixed effects, race, gender, physician experience, physician gender, hours into shift, laboratory ordered, and ED capacity controls. *** p<0.001.

- **Patient characteristics:** Vital signs (tachycardic, tachypneic, febrile, hypotensive), age, demographics, chief complaint by severity

- **Physician characteristics:** Years of experience, gender, hours into shift (capturing potential fatigue effects)

- **Realized laboratory diagnostic intensity:** Laboratory tests performed during the encounter

- **Contextual factors:** ED capacity level (normal, minor overcapacity, major overcapacity),

- **Shift level FE:** Month-year and day-of-week × time-of-day fixed effects

The addition of laboratory test ordering is critical. If batch tendency captures physicians who are generally more aggressive or comprehensive diagnosticians, they should order both more imaging and more labs. By controlling for whether labs were actually ordered for each patient, we test whether imaging effects persist when accounting for the thoroughness of general diagnostics. If batch tendency simply reflected "comprehensive diagnosticians," controlling for realized lab intensity should substantially attenuate the imaging-specific effects.

First-stage results validate the concern while demonstrating that we can isolate the batching-specific component. Below, is our revised first-stage regression (Table 2), both without and with the expanded precision controls.

24

We see that even when controlling for this comprehensive set of precision controls, substantial independent variation remains (F=171.9, well above weak instrument thresholds). This is precisely what we would expect if the instrument captures batching-specific propensity rather than general cautiousness.

Our revised main results (Table 4) now incorporate the expanded controls and additional methods we employed that are more robust to physician-level unobserved heterogeneity. Specifially, In addition to our standard 2SLS estimates using batch tendency as the instrument, Table 4 includes estimates from the Unbiased Jackknife Instrumental Variables Estimator (UJIVE). We have added this estimator because it is purpose-built for leniency designs with many decision-makers. When 2SLS uses many physician indicators as instruments, it constructs estimated leniency partly from each patient's own treatment status, creating mechanical correlation with the outcome error that biases estimates toward OLS. UJIVE eliminates this many-instrument bias through leave-one-out estimation. The agreement between our 2SLS and UJIVE estimates indicates that many-instrument bias is not driving our findings and that both approaches identify similar underlying variation.

The attenuation in the 2SLS estimates after adding precision controls confirms the reviewer's intuition that our original instrument may have contained some correlation with general diagnostic intensity. At the same time, the close agreement between the 2SLS and UJIVE estimates—both before and after adjustment—provides strong evidence that any remaining correlation between a physician's overall comprehensiveness and patient outcomes is too small to materially bias the results. The close agreement between 2SLS and UJIVE estimates, despite their different constructions of identifying variation, indicates that many-instrument bias is not driving our results. This convergence, combined with our expanded controls for laboratory ordering and physician characteristics, strengthens confidence that our revised specification isolates batching-specific effects.

Additionally, following an approach outlined by Conley et al. (2012), we test how our estimates change in our 2SLS results if one violates the exclusion restriction by allowing batch tendency to have a direct effect on outcomes. Under standard IV assumptions, the structural equation is

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

which assumes that batch tendency affects outcomes only through its effect on Batched_i.

Conley's relaxation allows for a small violation of the exclusion restriction by introducing a direct effect of the instrument:

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

where δ captures any exclusion restriction violation, including
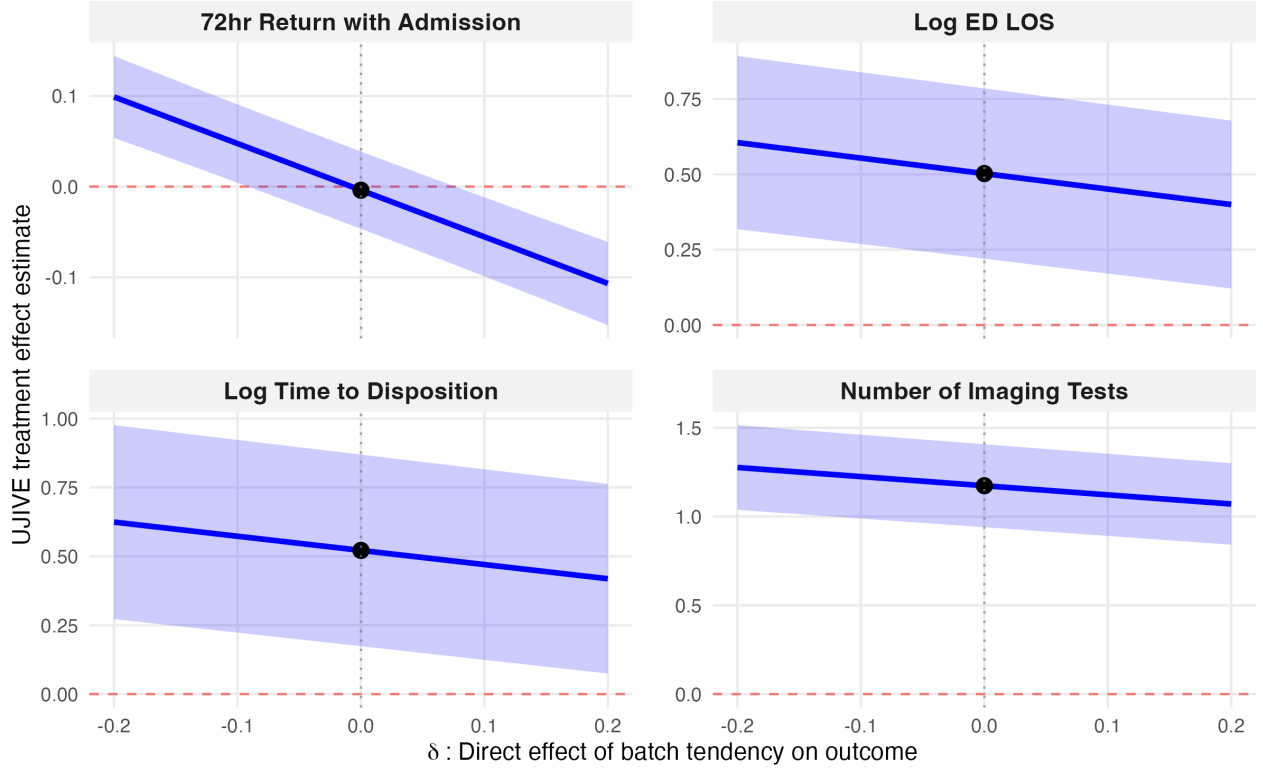
25

Table 4: Effect of Batching Tests on Patient Outcomes

	Sequenced mean	2SLS (2)	(3)	UJIVE (4)	(5)
Panel A. Primary Outcomes					
Log time to disposition	5.237	0.659***	0.651***	0.583***	0.522***
	(0.499)	(0.103)	(0.101)	(0.189)	(0.177)
Log LOS	5.490	0.717***	0.597***	0.653***	0.503***
	(0.456)	(0.094)	(0.088)	(0.158)	(0.144)
Number of distinct imaging tests	1.335	1.385***	1.241***	1.316***	1.174***
	(0.572)	(0.118)	(0.116)	(0.126)	(0.119)
72hr return with admission	0.012	-0.0137	-0.0146	-0.0079	-0.0039
	(0.110)	(0.018)	(0.019)	(0.020)	(0.022)
72hr return	0.030	-0.0512	-0.0536	-0.0440	-0.0396
	(0.170)	(0.029)	(0.031)	(0.032)	(0.034)
Panel B. Test Types					
X-ray	0.576	0.943***	0.989***	0.960***	0.959***
	(0.494)	(0.100)	(0.101)	(0.116)	(0.117)
Ultrasound	0.171	0.160**	0.087	0.164*	0.087
	(0.377)	(0.076)	(0.073)	(0.082)	(0.078)
CT without contrast	0.400	0.102	0.062	0.052	0.053
	(0.490)	(0.095)	(0.086)	(0.112)	(0.102)
CT with contrast	0.187	0.180*	0.102	0.140	0.075
	(0.390)	(0.078)	(0.076)	(0.087)	(0.079)
Panel C. Disposition					
Admission	0.279	0.419***	0.404***	0.424***	0.398***
	(0.449)	(0.096)	(0.088)	(0.103)	(0.090)
Necessary controls	—	Yes	Yes	Yes	Yes
Precision controls	—	No	Yes	No	Yes
Observations	11,679	11,679	11,679	11,679	11,679

Notes: Column 1 reports means for sequenced patients (standard deviations in parentheses). Columns 2–3 report 2SLS results using physician batch tendency as the instrument. Columns 4–5 report UJIVE estimates using ED provider identifiers as many weak instruments. All models include day-of-week and month fixed effects. Precision controls described in text. Standard errors are heteroskedasticity-robust.

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

- direct effects (e.g., physicians with higher batch tendency directly prolong LOS),

- backdoor paths (e.g., batch tendency reflects physician aggressiveness, which also affects LOS),

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

Notes: UJIVE estimates under different assumptions about direct effects (δ) of physician batch tendency on outcomes. One unit of δ implies a one percentage point increase in batch tendency directly changes the outcome by δ units. Black points indicate baseline estimates. Shaded areas show 95% confidence intervals using heteroskedasticity-robust standard errors following Conley et al. (2012).

We then estimate $\beta(\delta)$ for a range of plausible values of δ to assess how sensitive our conclusions are to such violations.

The figure below (now included in EC.9) presents sensitivity analyses showing substantial robustness. For all outcomes, the treatment effect remains significant across $\delta \in [-0.2, 0.2]$.

In summary, while we cannot definitively rule out all exclusion restriction violations—no observational study can—the convergence of evidence strongly supports our interpretation:

- Treatment effects persist after adding addition controls for general diagnostic intensity (laboratory ordering), experience, and hours into shift

- UJIVE estimates, which eliminate many-instrument bias, closely match 2SLS results, indicating both approaches identify similar underlying variation

- 2SLS estimates remain robust to potential backdoor paths for plausibly large exclusion restriction violations

27

The LATE estimator we identify represents the causal effect for compliers—patients whose imaging strategy depends on physician practice style rather than clinical necessity. This is precisely the margin where ED management interventions can influence practice without constraining clinically necessary care.

We have revised the manuscript to reflect these more conservative estimates while maintaining transparency about our identification assumptions. The core finding remains: discretionary batch ordering increases ED length of stay by approximately 65-81% without improving short-term patient outcomes, suggesting substantial opportunity for operational improvement through policies that preserve diagnostic flexibility.

We are extremely grateful to the reviewer for pushing us to strengthen our identification strategy and giving us the opportunity to respond to these concerns. These additional analyses have strengthened the manuscript and our confidence in the robustness of the estimated causal effects.

R1 Wrote (Major Concern 2 – Endogeneity concerns):

"As discussed by the authors on page 16 of the manuscript, ensuring the exclusion restriction is met to establish valid causal estimates using the proposed instrument is challenging, and not directly testable. However, the authors argue that "such violations may likely only have a small impact and may be less concerning than in other healthcare settings". This is a requirement to establish causal claims (which the authors claim to want to do), and as outlined in Major Concern 1 above, "high batching tendency" is likely to be correlated by other elements of a physician's "practice style" that could directly affect the outcomes being tested, outside of only batching imaging. A physician who is more cautious and comprehensive could also order more lab tests, which take time, and thus extend the LOS. A physician who is more cautious and comprehensive could also take longer on the examination, and choose to monitor the patient for a longer period of time before making a disposition decision. Both of these examples could directly impact the outcome without going through imaging batching, thus violating the exclusion restriction. As such, we cannot be certain that the estimates found in this paper are causal estimates on the impact of batching on ED outcomes (even assuming that the main explanatory variable captures batching, which I argue has its own flaw in Major Concern 1)."

Response: We thank the reviewer for reiterating this fundamental concern about exclusion restriction violations. As stated in our response to Major Concern 1, we have taken several steps to address it. For starters, we have substantially expanded our controls to test whether batch tendency captures these correlated behaviors and believe that the manuscript, findings, and implications have been strengthened as a result.

We now control for laboratory tests ordered for each patient. The reviewer's first example—that high-batch physicians might order more labs, which extend LOS—is directly tested by this control. After controlling for lab ordering, imaging effects remain large and significant, while LOS and time to disposition effects show the attenuation the reviewer anticipated (though also remain large and significant). Additionally, we control for physician experience, gender, and hours worked in a shift to capture systematic differences in examination patterns and monitoring decisions.

We acknowledge the reviewer's point that we cannot definitively rule out all exclusion restriction

violations. However, we note two important qualifications to this concern:

First, our revised specification directly tests the reviewer's specific examples and finds that while some correlation exists (as evidenced by attenuation of time effects with lab controls), imaging-specific variation persists. The reviewer's concern that we "argue such violations may likely only have a small impact" referred to our original specification. We have removed this language and instead provide empirical evidence through our expanded controls, which show the existence of correlations and demonstrate that substantial imaging-specific effects persist.

Second, even if exclusion restriction concerns persist, our reduced-form estimates—the direct effect of being assigned to a high-batch-tendency physician—represent causal effects of physician assignment that do not require the exclusion restriction. These estimates inform ED management decisions about physician training, feedback, and protocols regardless of the precise mechanism. When we control for observable pathways (labs, physician characteristics), the reduced-form effects remain substantial for imaging volume, while attenuating for time outcomes, providing managers with actionable evidence about which outcomes are most affected by physician assignment.

Third, as mentioned earlier, we have added formal sensitivity analysis to rigorously test the potential impact the violation of exclusion restriction might have on our main findings (EC.9). Finally, have revised Section 4.8 (Limitations) to acknowledge this concern more directly, removing claims that violations "may likely only have a small impact" and instead transparently presenting our validation evidence through expanded controls, while acknowledging that definitive exclusion restriction tests are not possible in observational settings.

R1 Wrote (Minor Concern – Randomization mechanism clarity):

"In this report I am taking the authors' word regarding the randomization of patient to physician matching as described in the paper. More details on this should be given, and the patient-to physician assignment can be empirically shown – if the patients are assigned to physicians in a round-robin format, and you know when patients arrive at the hospital, and which physicians are working, you can empirically show that assignment is random. This is an important part of the empirical design of the paper, and so future iterations should show this empirically. Related to this, there is also the question of what the the queue configuration of the physicians look like – if it is random, and there are 5 physicians working, which physician gets the first patient? Is it by alphabetical order? More detail to describe this assignment mechanism would be helpful."

Response: We thank the reviewer for requesting additional detail about the patient-physician assignment mechanism, which is crucial to our identification strategy. We have substantially expanded Section 3.1 with detailed description of Mayo Clinic's rotational assignment system, documented in prior publications at this site; see, e.g., Traub et al. (2016a,b, 2018). As we have clarified, the assignment under this system is purely on a rotational basis, thus, providing a quasi-random patient-to-provider assignment mechanism.

The key institutional features address the reviewer's specific questions:

Assignment mechanism: Patients are automatically assigned to physicians 60 seconds after reg-

istration through a computerized algorithm in the electronic health record system. Assignment is based solely on arrival time—the algorithm does not consider patient demographics, chief complaint, Emergency Severity Index score, physician workload, or the acuity of recently assigned patients.

Rotation initialization and ordering: At shift start, each physician receives four consecutive patients to establish an initial patient load, after which they enter rotation with other on-duty physicians. The rotation order is predetermined by the ED scheduler and varies across shifts to ensure fairness over time. Critically, these assignments are based solely on arrival time—the algorithm does not consider patient demographics, chief complaint, Emergency Severity Index score, physician workload, or the acuity of recently assigned patients. To maintain system integrity, physicians receive no new patients during their final 120 minutes and are capped at 18 patients per shift.

System constraints: To maintain integrity, physicians receive no new patients during their final 120 minutes and are capped at 18 patients per shift.
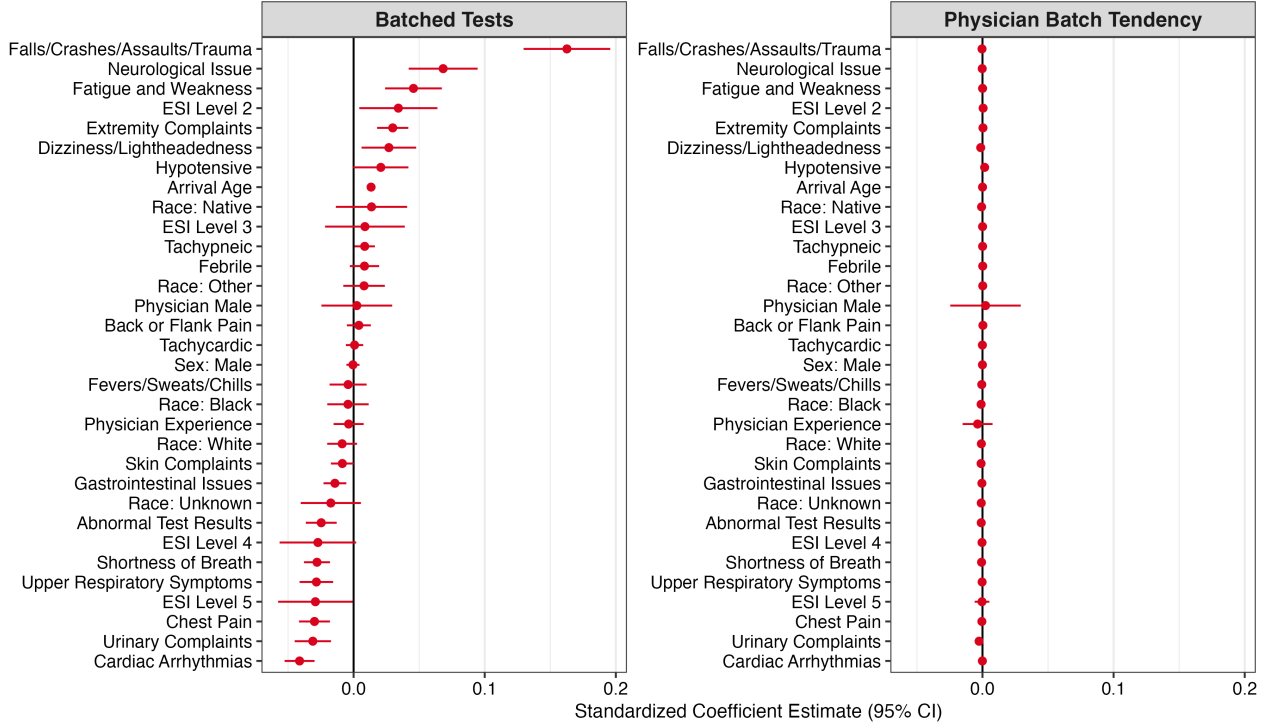
We have added this detailed explanation to Section 3.1:

> "The Mayo Clinic ED employs a sophisticated computerized rotational patient assignment algorithm that addresses many empirical challenges in healthcare settings; see, e.g., Traub et al. (2016a,b, 2018). The system automatically assigns patients completely on a rotational basis to physicians 60 seconds after registration through the electronic health record system. At shift start, each physician receives four consecutive patients to establish an initial patient load, after which they enter rotation with other on-duty physicians. The rotation order is predetermined by the ED scheduler and varies across shifts to ensure fairness over time. Critically, these assignments are based solely on arrival time—the algorithm does not consider patient demographics, chief complaint, Emergency Severity Index score, physician workload, or the acuity of recently assigned patients. To maintain system integrity, physicians receive no new patients during their final 120 minutes and are capped at 18 patients per shift.
>
> Our empirical tests show that this rotational mechanism achieves the quasi-randomization necessary for causal inference. Unlike settings where patient-physician matching may be influenced by triage decisions, physician preferences, or informal routing practices, the Mayo Clinic's algorithmic assignment removes discretion from the matching process."

Empirical verification: Regarding the reviewer's suggestion to empirically demonstrate randomization, Figure 2 of the revised version provides precisely this test. The right panel shows that after conditioning on time fixed effects (which account for the mechanical rotation across shifts), no patient characteristic—including demographics, vital signs, ESI level, or chief complaint—significantly predicts assignment to high versus low batch-tendency physicians. The coefficients are all near zero with confidence intervals crossing zero, confirming that the rotational algorithm effectively randomizes patient assignment conditional on shift timing.

Figure 2: Batch Tendency by Patient Characteristics

Notes: This figure plots a test for quasi-random assignment of patients to physicians in the Mayo Clinic ED. The left panel shows how patient characteristics predict batching decisions. The right panel shows these same characteristics do not predict assignment to physicians with different batch tendencies. Residualization fixed effects include hospital-year-month, hospital-day of week-time of day. Robust standard errors are clustered at the physician level.

We have added explanatory text in Section 3.3 emphasizing this empirical verification:

"Figure 2 provides empirical verification that, while the decision to batch depends on patient characteristics, our measure—batch tendency—is plausibly exogenous. The left panel uses a linear probability model to test whether encounter, patient, ED, and physician characteristics predict the batching decision, controlling for shift-level fixed effects with standard errors clustered at the physician level. As expected, patient characteristics strongly predict batching decisions; for instance, patients with Falls/Assaults/Trauma complaints are 16.2 percentage points more likely to be batched compared to similar patients under similar ED capacity. The right panel assesses whether these same characteristics predict assignment to physicians with different batch tendencies. Importantly, we find that patient characteristics do not significantly predict assignment to high versus low batch-tendency physicians. The coefficients are near zero with confidence intervals crossing zero for all patient characteristics, confirming that, conditional on shift fixed effects (which account for the mechanical rotation), the assignment of patients to physicians with different batching tendencies is effectively random. This validates the rotational as-

31

signment mechanism and establishes batch tendency as an exogenous source of variation for identifying causal effects."

Together, the institutional details and empirical balance tests confirm that Mayo Clinic's rotational assignment mechanism achieves the quasi-randomization necessary for our identification strategy. This distinguishes our study from observational analyses where endogenous patient-physician matching could confound estimates of physician practice effects.

R1 Wrote (Potential impact & writing quality):

"The paper is well written, and studies an important question. I hope the authors find this report helpful in moving this work forward."

Response: We sincerely thank the reviewer for their thorough and constructive feedback. The detailed comments have been invaluable in strengthening our analysis and clarifying our contributions. We believe the extensive revisions addressing the reviewer's concerns about variable definition, empirical strategy, and causal interpretation have substantially improved the manuscript.

V. Responses to Referee 2 (R2) Comments

R2 Wrote (Research question & contribution:):

"The paper's research question is clear. The main contribution is to provide causal evidence that batch ordering of advanced imaging tests in emergency departments (EDs) — commonly assumed to be efficient — increases patient length of stay (LOS), test volume, and admission rates. EDs are under constant pressure to improve throughput, and advanced imaging is a major driver of delays, costs and overcrowding in the EDs. The paper does have the potential to make a significant and novel contribution by causally examining the impact of diagnostic batch ordering in emergency departments. While prior research has explored physician-driven variation in testing and the operational burden of imaging, this study uniquely isolates the effect of batching behavior using a quasi-randomized design."

Response: We appreciate the reviewer's recognition of our contribution to understanding diagnostic test ordering in emergency departments. We agree that challenging conventional assumptions about batch ordering efficiency is important given the operational pressures EDs face. The reviewer's constructive feedback has substantially strengthened our empirical approach and helped us clarify how our quasi-randomized design isolates batching effects from other aspects of physician practice variation.

R2 Wrote (Framing):

"The paper frames the comparison as "batching vs. sequential", but that is not technically correct. In reality, the way the independent variable is defined, the comparison is "early batched imaging" vs. "everything else", which includes a broad mix of clinical pathways. This muddies the interpretation: are the harms of batching due to: The batching itself? Or just differences between patients who need 2+ early tests vs. those who do not? Moreover, the sequencing is not fully explored: the study favors sequential ordering, but the authors do not actually compare different sequencing strategies or evaluate outcomes where delayed imaging causes diagnostic delays."

Response: We thank the reviewer for this important clarification about our treatment comparison. The reviewer is absolutely correct that our empirical strategy compares "early batched imaging" with "everything else" rather than a pure batching versus sequential comparison. The choice of the treatment variable was made to ensure maximum practical relevance after various conversations with our medical collaborators at the two leading US hospitals we have collaborated with (Mayo Clinic and MGH). We appreciate this opportunity to clarify our approach and have made substantial revisions to address this concern.

First, an essential clarification: our sample includes only encounters where at least one imaging test was ordered. The "everything else" comparison, therefore, comprises patients who received either (1) a single imaging test, or (2) multiple tests ordered sequentially, but never patients with zero imaging. This distinction is critical because our comparison is between different imaging strategies for patients who require diagnostic imaging, not testing versus no testing. Furthermore, our LATE estimator captures the effect for "compliers"—patients whose batching decision is determined by the physician's batch tendency rather than clinical necessity. If a patient were assigned to a low-batch tendency

33

physician, they may receive only a single imaging test, because after observing the first test, the physician realizes they have enough information to make a disposition decision. This is what we refer to as the "information gain" in our study. The same patient, if assigned to a high-batch tendency physician, may have received multiple imaging tests because the orders were placed simultaneously.

The reviewer raises a fundamental question: are the harms from batching itself or from differences between patients needing multiple early tests? This potential endogeneity is exactly what our instrumental variables approach addresses. By leveraging quasi-random variation in physician batch tendency, we isolate the causal effect of early batching for patients whose testing strategy depends on physician practice style rather than clinical necessity(i.e., "compliers"). For decisions at the margin of physician discretion, such as the one we study, the batching decision is made under uncertainty about whether additional tests will be needed. Conditioning our analysis on eventual test count would introduce post-treatment bias, as the initial ordering strategy causally affects subsequent testing decisions. Our approach preserves causal interpretation by comparing strategies at the moment of decision-making, which is exactly what our medical collaborators (and co-authors of this paper) were keen to know.

We have made the following changes to address your concerns which we believe have made the manuscript stronger, and we hope you find them to your satisfaction:

- **Reframed our comparison throughout the manuscript**. To remove any confusions, we now consistently describe our comparison as "batch ordering versus standard practice" rather than "batching versus sequencing."

- **Clarified what we identify and why it matters**. We added text in Section 3.3 explaining:

  "Our two-stage least squares estimates represent the LATE of batch ordering for 'compliers'—patients whose testing strategy depends on the assigned physician's practice style. This effect compares batch ordering to standard practice, which includes both sequential ordering and single tests. While this involves a composite counterfactual, it provides the policy-relevant parameter: the effect of encouraging comprehensive upfront testing versus allowing diagnostic information to guide testing decisions for patients at the margin of clinical discretion—those whose testing strategy is not dictated by clear clinical necessity but rather depends on physician practice style and judgment."

- **Added discussion of the information problem**. In Section 3.2.1, we explain why we focus on early batching:

  "We focus on batches that concern the first imaging tests ordered during the patient encounter because this represents the moment of maximum diagnostic uncertainty, when physicians must decide their testing strategy before clinical information unfolds. For patients whose testing is not fully determined by medical necessity, physicians

cannot know ex-ante which patients will ultimately require multiple tests, leading to instances of early batching as discretionary choice based on practice style rather than clinical necessity."

Regarding potential harms from delayed sequential testing, we find no evidence on 72-hour return, suggesting sequential approaches do not cause harmful diagnostic delays.

- **Clarified why "everything else" is the right comparison**. We added the following to Section 5.1:

  "This comparison reflects the real choice facing ED managers: should protocols encourage comprehensive upfront testing or preserve diagnostic flexibility? Our estimates show that preserving optionality through standard practice—which allows information from initial tests to guide subsequent decisions—reduces testing intensity."

The reviewer's observation about mixed clinical pathways in our counterfactual is astute. However, this heterogeneity reflects what makes our estimate policy-relevant. ED managers cannot randomize patients to pure sequential protocols based on eventual diagnostic needs (which are unknown ex-ante). They can only influence whether physicians default to comprehensive early testing or preserve diagnostic optionality when facing uncertainty. Our LATE estimator focuses on "compliers," and hence, provides exactly this parameter.

We believe these revisions substantially improve the manuscript's clarity while maintaining scientific rigor. The comparison we identify—batch ordering versus standard practice for marginal patients—is both what we can credibly estimate given the information structure and what hospital administrators need to know. We are grateful to the reviewer for pushing us to clarify this important distinction, as it has led us to better articulate both the methodological foundations and practical implications of our work.

R2 Wrote (Hypotheses):

"As it currently stands, the study lacks a clearly articulated hypotheses section, which would help clarify the underlying mechanisms the authors expect to observe in the results. This section is typically where one would expect the authors to build a narrative around the anticipated behavioral patterns and how those behaviors are theoretically linked to the operational outcomes under study."

Response: We thank the reviewer for this valuable suggestion to develop formal hypotheses. The reviewer is absolutely correct that explicitly articulating our theoretical predictions strengthens the paper's contribution and clarifies the mechanisms we expect to observe.

Following this guidance, we have added Section 2.3 "Hypothesis Development" that builds from our literature review to develop four formal hypotheses about how batch ordering affects ED operations:

"

Hypothesis Development

Building on the literature reviewed above, we develop formal hypotheses about how batch ordering affects ED operations. While prior work has identified the mechanisms driving batching behavior and its potential consequences, the net effects remain theoretically ambiguous. Our theoretical framework centers on the fundamental tradeoff between the perceived efficiency of parallel processing and the information value of sequential testing.

Information Value and Test Volume

The decision to batch or sequence tests fundamentally involves whether to preserve the option value of information. Sequential testing allows each test result to inform subsequent decisions, potentially eliminating unnecessary tests. When physicians batch tests upfront, they commit to a diagnostic pathway before information unfolds, forfeiting this option value.

Physicians under high workloads face cognitive strain from task switching and may batch tests to defer complex diagnostic reasoning (KC, 2013; Skaugset et al., 2016). However, this cognitive convenience comes at a cost. Without the filtering mechanism of sequential information revelation, physicians must rely solely on their initial assessment. Lam et al. (2020) identify this as a key driver of overtesting—when facing diagnostic uncertainty, physicians order comprehensive test batteries rather than allowing initial results to guide subsequent testing. Given the documented variation in physician testing intensity (Hodgson et al., 2018), with some physicians ordering twice as many tests as their peers, batching likely amplifies these tendencies by removing the natural stopping points that sequential results provide. Therefore:

> **Hypothesis 1.** *Batch ordering will increase the total number of imaging tests performed compared to standard practice.*

Processing Time and Operational Flow

While batching strategies reduce setup times in manufacturing (Fowler and Mönch, 2022), the ED imaging context presents unique operational constraints as noted in our review. Different imaging modalities require separate equipment and cannot be performed simultaneously (Jessome, 2020). This creates a fundamental bottleneck where batched orders must still be executed sequentially, but now with a larger committed workload that cannot be adjusted based on emerging information.

Moreover, the cognitive load literature suggests that processing multiple test results simultaneously increases decision complexity (KC, 2013). When physicians receive multiple

results at once rather than sequentially, they must integrate more information simultaneously, potentially lengthening the diagnostic reasoning process. This "information overload" effect, combined with the additional tests ordered as predicted in H1, suggests that batching may paradoxically increase rather than decrease processing times:

> **Hypothesis 2.** *Batch ordering will increase patient length of stay and time to disposition compared to standard practice, as the operational constraints of imaging and increased test volume outweigh any potential benefits of parallel processing.*

Clinical Decision-Making and Disposition

The medical literature recognizes "diagnostic momentum"—where abnormal findings, even if clinically insignificant, drive further workup and more conservative clinical decisions (Coen et al., 2022; Featherston et al., 2020). When physicians batch order and receive multiple results simultaneously, they encounter more opportunities for incidental findings that may influence disposition decisions (Lumbreras et al., 2010; Berlin, 2011). As our review noted, physicians facing uncertainty and potential legal consequences may opt for more conservative disposition decisions (Rao and Levin, 2012; Lam et al., 2020). The simultaneous arrival of multiple test results, particularly with incidental findings, may trigger defensive medicine behaviors:

> **Hypothesis 3.** *Batch ordering will increase hospital admission rates.*

Contextual Moderators

The literature on physician behavior under capacity constraints consistently shows that resource scarcity forces more selective decision-making (Kuntz et al., 2014; KC and Terwiesch, 2009). When EDs face severe overcrowding, the operational pressures documented in our review intensify. Under these conditions, physicians may reserve batching for cases where it is clinically essential rather than convenient:

> **Hypothesis 4.** *The effects of batch ordering on LOS and test volume will be attenuated under conditions of major ED overcapacity.*

These hypotheses provide testable predictions that we examine using our quasi-experimental design. By leveraging variation in physician batching tendency under random patient assignment, we can identify whether these theoretical mechanisms manifest in actual ED operations."

This new section provides the theoretical foundation that motivates our empirical specifications and helps readers understand which mechanisms we test. We are grateful to the reviewer for this suggestion, which has substantially strengthened the manuscript's theoretical contribution.

R2 Wrote (Shift-of-testing to other settings): *"While the study finds that batch ordering increases imaging in the ED as well as inpatient admissions, it does not evaluate whether this reflects actual overuse or a shift in imaging from the inpatient setting to the ED. In other words, it is unclear whether batching increases unnecessary testing or simply frontloads diagnostics."*

Response: We thank the reviewer for this important distinction between overuse and frontloading of diagnostics. The reviewer is correct that we cannot definitively determine whether the additional imaging represents unnecessary testing or simply shifts testing from inpatient to ED settings. This is an important limitation of our study.

Our medical collaborators have identified disagreements in test ordering between ED physicians and hospital inpatient unit physicians (i.e., after the patient is discharged from the ED and moved to an inpatient unit) as a major issue outside the scope of this study. Our focus is on ED operations, and we do not have access to data outside the ED. However, it is also important to note that the majority of ED patients (66.8% in Mayo Clinic) are discharged after their ED visits. Thus, our findings have important consequences, even though they focus solely on tests ordered by ED physicians.

However, 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 in our data—we find that batching increases admission probability by increasing the number of imaging tests performed. This suggests that at least some of the additional tests are driving new admissions rather than frontloading diagnostics for patients who would have been admitted regardless of their ED imaging results. That is, the tests themselves are creating admissions, not being ordered because admission was already determined.

That said, we acknowledge that multiple causal pathways may occur simultaneously. Some batched patients may receive tests that reveal admission-worthy conditions, which would have been discovered later. Others may be admitted defensively due to incidental findings from comprehensive imaging that are not pursued with sequential testing. Without linked inpatient imaging records, we cannot empirically separate these mechanisms.

To directly address this concern, we re-estimated our main specifications restricting the sample to patients who were ultimately discharged from the ED. This subsample analysis is informative because these patients, by construction, do not receive subsequent inpatient diagnostic workups, eliminating the possibility that ED imaging merely substitutes for tests that would have occurred during hospitalization. As shown in the table below (Table EC.12.1), the estimated effect of batching on the number of distinct imaging tests remains large, positive, and statistically significant across UJIVE and 2SLS specifications, with magnitudes comparable to those in the full sample.

We emphasize that this discharged-only analysis should be interpreted with caution, as discharge status is itself a downstream outcome potentially affected by ED imaging decisions. Conditioning on discharge therefore risks post-treatment selection and may attenuate or otherwise distort causal

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
Adj. R^2	0.0305	0.0952	0.0305	0.0952

Notes: This table reports instrumental-variables estimates of the effect of batching on the number of distinct imaging tests among patients discharged from the emergency department. Columns (1)–(2) report 2SLS estimates using physician batch tendency as the instrument. Columns (3)–(4) report UJIVE estimates using ED provider identifiers as many weak instruments. ***$p < 0.001$.

estimates. For this reason, we do not treat this subsample analysis as a primary causal estimate. Rather, it serves as a robustness check demonstrating that the positive association between batching and imaging intensity is not driven solely by patients who are ultimately admitted and receive further inpatient testing. We report this analysis in EC.12 and reference it here solely to clarify interpretation of our findings in response to the reviewer's concern about frontloading.

We have been careful throughout the manuscript to avoid characterizing the additional tests as "waste" or "overuse," as our LATE identifies effects for marginal patients whose testing decisions vary by physician preference—these tests may have clinical value regardless of timing. However, even if batching represents frontloading rather than overuse, it has operational consequences worth noting. Based on discussions with our physician coauthors at both study sites, ED imaging resources are typically more constrained than inpatient resources; frontloading increases ED congestion and delays care for other ED patients; and inpatient teams often prefer to direct their own diagnostic approach based on evolving clinical information.

We have added a discussion of this limitation in Section 5.3:

"We cannot determine the extent to which the additional imaging from batching represents unnecessary testing or frontloading of diagnostics that would eventually occur in the inpatient setting. However, interpreting our findings purely as frontloading is complicated by the fact that admission itself is affected by batching—we observe a 39 percentage-point increase in admission probability, suggesting that the tests themselves influence disposition decisions. Multiple mechanisms may operate: some additional tests may reveal genuinely admission-worthy conditions, while others may trigger defensive admissions through incidental findings. Distinguishing between these pathways would require linked ED-inpatient imaging data to examine whether patients who are batched receive correspondingly fewer

39

tests after admission. This remains an important direction for future research."

We appreciate the reviewer highlighting this distinction, as it has led us to be more precise about the causal pathways our estimates capture and the important questions that remain for future research.

R2 Wrote (Request for cost-benefit analysis):

"Relatedly, what would help is a cost-benefit analysis – given the findings about overuse, it is surprising the paper does not estimate financial impact or imaging cost burdens. Do the authors have any data they could use to conduct such an analysis?"

Response: We thank the reviewer for this valuable suggestion. We have developed a cost-benefit analysis that quantifies the operational and financial burden of discretionary batch ordering, adopting a societal perspective with Medicare-allowed amounts as proxies for resource costs.

Our analysis focuses on two cost components where we have direct causal evidence: (1) excess imaging costs from additional tests and (2) excess ED capacity consumed from extended length of stay. We deliberately adopt conservative assumptions throughout, excluding several additional cost categories—including physician time, radiologist interpretation burden, patient time costs, downstream costs from increased admissions, and spillover congestion effects—making our estimates lower bounds on the true cost of discretionary batching.

A key input to our analysis is identifying the population affected by discretionary batching. Our complier analysis (EC.4) indicates that approximately 20% of patients in our analytical sample are "compliers"—patients whose batching status depends on physician assignment rather than clinical necessity. Among the 1,595 batched patients in our sample, roughly half (51%) are compliers whose batching reflects physician discretion rather than clinical indication. These 813 "batched compliers" in our 15-month sample (650 annualized) represent the patients who would benefit from interventions targeting discretionary batching.

We estimate imaging costs using Medicare-allowed reimbursement for diagnostic radiography in the ED setting. For a two-view chest radiograph (CPT 71046), the technical component under HOPPS is \$98.53, with an additional \$30–40 for the professional component, yielding total Medicare-allowed reimbursement of approximately \$120–140 per study (Dabus et al., 2025). Our UJIVE estimates (Table 4, Panel B) indicate that discretionary batching significantly increases X-ray utilization (0.96 additional tests, $p < 0.001$), while effects on other modalities (ultrasound, CT) are not statistically significant. To maintain a conservative approach, we include only the statistically significant X-ray effect in our primary analysis. Using a range of \$100–160 per X-ray (mean \$130), we estimate excess imaging costs of **\$125** (\$96–153) per batched complier.

To value the excess ED bed-hours consumed by discretionary batching, we draw on published estimates of ED operational costs using time-driven activity-based costing. Schreyer and Martin (2017) report ED personnel costs of \$58.20 per bed-hour (2010–2011 data; approximately \$83 inflation-adjusted to 2024), while Canellas et al. (2024) find daily ED boarding costs of \$1,856, implying

40

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 batched compliers × $349 per patient. Batched compliers estimated as 23.9% of ED volume (imaging-relevant encounters) × 7.0% (batched complier share of imaging encounters). Imaging costs ($125 per patient) include only statistically significant X-ray effects; including point estimates for other modalities would increase costs by approximately 17%. Capacity costs ($224 per patient) assume $85/bed-hour and 2.64 additional bed-hours per batched complier. Range reflects uncertainty in imaging ($100–160/X-ray) and bed-hour ($75–100/hour) cost parameters.

approximately $77 per hour. Based on these estimates, we use a range of $75–100 per bed-hour (mean $85). Our UJIVE estimates indicate discretionary batching increases LOS by 65% (Table ??), translating to 158 additional minutes (2.64 hours) relative to baseline LOS of 242 minutes for patients receiving standard care. At $75–100 per bed-hour, this yields **$224** ($198–264) in excess capacity costs per batched complier.

Table EC.6.1 summarizes our findings. Combining imaging and capacity costs, discretionary batching imposes an estimated **$349** ($294–417) in direct costs per affected patient. Capacity costs account for 64% of the total, reflecting the substantial LOS increase we document.

Scaling to annual ED operations, discretionary batching costs approximately **$232,000 per year** for a medium-sized ED (40,000 annual visits) and $349,000 for a large urban ED (60,000 visits). For our study site, the annual cost is approximately $227,000. These costs translate to meaningful operational impacts: eliminating discretionary batching would free approximately 1,700 bed-hours annually at our study site—equivalent to 72 additional 24-hour days of ED capacity—enabling approximately 425 additional patient visits (a 1.1% capacity improvement).

Even under conservative assumptions, discretionary batch ordering imposes meaningful costs—approximately $232,000 annually for a medium-sized ED—without corresponding improvements in patient outcomes (Table 4 shows no significant effects on 72-hour returns). These costs arise purely from the approximately 20% of patients whose batching decisions reflect physician preference rather than clinical necessity.

These estimates suggest that relatively modest interventions—such as decision support tools encouraging sequential ordering when clinically appropriate, or providing physicians with feedback on their batching rates relative to peers—could yield meaningful cost savings. A 50% reduction in discretionary batching would save approximately $113,000 annually while freeing 850 bed-hours. Given that compliers represent patients at the margin of physician discretion (not those with clear clinical indications for comprehensive imaging), interventions targeting this group are unlikely to compromise care quality while offering substantial operational benefits.

We have added a detailed cost analysis in EC.4, including the above table and discussion. We believe this addition substantially enhances the practical relevance of our findings for ED administrators considering policies to influence physician batching behavior. We thank the reviewer for prompting us to include this important analysis and have referenced it in the main text (Section 4.7), included here for posterity:

"To quantify the financial burden of discretionary batching—where providers intentionally group multiple diagnostic orders rather than placing them individually—we combine our causal estimates with published cost parameters (see EC.4 for details). Adopting a societal perspective, we use Medicare-allowed reimbursement of $130 per ED radiograph (Dabus et al., 2025) and ED bed-hour costs of $85 based on time-driven activity-based costing studies (Schreyer and Martin, 2017; Canellas et al., 2024).

Under deliberately conservative assumptions, discretionary batching imposes at least $349 in direct costs per batched complier: $125 from excess imaging and $224 from excess capacity utilization. Scaling to annual operations, this amounts to approximately $232,000 per year for a medium-sized ED (40,000 annual visits) and $349,000 for a large urban ED (Table EC.X). These estimates represent conservative lower bounds, include only statistically significant imaging effects (excluding ultrasound and CT point estimates), value bed-hours at operational cost rather than throughput value, and exclude downstream costs from the 40 percentage point increase in admission probability. Inpatient stays cost $10,000–15,000, which would dwarf our direct cost estimates. Notably, we find no offsetting benefits: 72-hour return rates are statistically indistinguishable between strategies. Even under these conservative assumptions, a 50% reduction in discretionary batching would save about $113,000 annually while freeing 850 bed-hours of capacity."

We appreciate the reviewer's suggestion, which has led us to provide a more comprehensive assessment of the operational and financial implications of discretionary batching in EDs.

R2 Wrote (Batch timing and its relation to LOS):

"Batch timing is unclear to me. Specifically, Length of Stay (LOS) is used as a primary outcome. Batching, defined as imaging orders within the first 5 minutes of encounter, is treated as a treatment applied at the beginning of the visit. But in real ED workflows, the decision to batch may itself be a function of how the patient's case has unfolded up to that point. For example, if the diagnosis is taking longer, or prior tests have not resolved the issue, or if the physician senses the patient may be admitted soon, then the physician might batch several tests later in the encounter in order to "wrap things up" and avoid delays — i.e., batching becomes a consequence of extended LOS, not just a cause. The 5-minute window may not capture the delayed batching; e.g., if initial results come back inconclusive or the patient's condition worsens. Thus, the batching variable may be misclassified, and later batches that react to prolonged stays are excluded from the analysis. Even when orders are placed early, test results often take time to return. That delay — especially from CTs or MRIs — inflates LOS. Therefore, the measured LOS may not be a clean posttreatment outcome, but instead partially determined by the batching process itself. This risks simultaneity bias, where cause and effect are entangled

in time. The issue is less severe if the paper really only claims to understand something about initial batching decisions, but I am not sure the authors have made those distinctions clear enough."

Response: We thank the reviewer for this insightful observation about the temporal relationship between batching and LOS. Before addressing the specific timing concerns, we note that in response to feedback from all three reviewers about our original specification, we have substantially expanded our control set to better isolate batching-specific effects. Our revised primary specification now includes controls for physician characteristics (experience, gender, and hours into the shift) and for laboratory tests ordered. This more conservative specification slightly attenuates our time-based estimates. Our revised estimates show discretionary batching increases imaging tests by 92% (1.24 additional tests performed, $p < 0.001$), results in 65-81% longer LOS ($p < 0.001$), and 68-91% longer time to disposition ($p < 0.001$).

Why early batching is the relevant parameter: The reviewer notes that physicians sometimes batch tests later to finish up. This behavior differs from ordering comprehensive imaging at the start. Our instrumental variable (physician batch tendency) measures physicians' early ordering tendencies—distinguishing those who order comprehensively upfront from those who wait for more information.

Empirically, we show that late batching—where a batch follows a single test—happens in only 189 cases (1.91% of multi-test encounters). To check the impact, we re-estimated all models treating these as batched (see Appendix D). The results barely change.

Clarifications made throughout the manuscript: Following the reviewer's observation, we have revised our framing to stress that our parameter of interest is the effect of discretionary batching versus standard practice:

- **Abstract:** Now specifies "discretionary batching decisions made at encounter initiation"

- **Section 1.2:** Clarifies that our LATE identifies "the effect of discretionary batching for patients whose testing strategy depends on physician practice style rather than clinical necessity."

- **Section 3.2.1:** Extensively revised to clarify our focus:

    "We focus on batches that concern the first imaging tests ordered during the patient encounter because this represents the moment of maximum diagnostic uncertainty when physicians must decide their testing strategy before clinical information unfolds. Physicians cannot know ex-ante which patients will ultimately require multiple tests, making early batching a discretionary choice based on practice style rather than clinical necessity."

LOS as outcome and simultaneity bias: The reviewer raises an important concern about whether LOS can serve as a clean post-treatment outcome given that test completion times mechanically

contribute to its measurement. We acknowledge this concern but emphasize that this relationship is not a confounder—it is precisely the causal pathway our study seeks to identify.

The reviewer asks whether this creates "simultaneity bias, where cause and effect are entangled in time." This would be a concern if some third factor simultaneously determined batching and LOS, or if LOS caused batching. However, our design addresses this: we focus on early discretionary batching decisions (the first tests ordered during the encounter), and our instrument (physician batch tendency) is predetermined before the patient encounter begins. The mechanical relationship between test completion and LOS represents the causal pathway through which early batching operates, not simultaneity bias.

Our treatment is the decision to batch, and LOS captures the total consequences of that decision. When physicians choose to batch tests upfront, they set in motion a cascade of operational consequences: patients wait for multiple tests to complete, radiologists must interpret multiple images, physicians must cognitively process all results simultaneously, and clinical decisions must integrate potentially conflicting or incidental findings. These are the mediating mechanisms through which early batching decisions affect patient flow.

To further separate clinical decision-making time from test completion time, we examine "time to disposition"—the duration until physicians make admission/discharge decisions, excluding post-decision boarding. Additionally, per a request from Reviewer 3, we also examine "treatment time" — defined as the time from when the patient is roomed in the ED to when a disposition decision is made. Our estimates show that batching has large and significant effects on both outcomes, demonstrating that early batching affects not only mechanical test completion waiting but also clinical processing efficiency, since these disposition decisions can be made before the imaging study results are read by a radiologist. Even after excluding post-disposition time (when no testing occurs), the early discretionary batching decision yields substantial point estimates of delays in clinical decision-making.

We appreciate the reviewer's careful attention to these temporal dynamics, which has prompted us to more clearly articulate that our study evaluates the full operational consequences of discretionary early batching decisions—precisely the parameter ED managers need to understand when considering protocols to influence physician ordering behavior.

R2 Wrote (Definition of batching): *"The definition of batching is also narrowly defined to be 2+ tests ordered within the 5 first minutes. The study equates batch ordering with guaranteed test completion and additive imaging volume. However, in practice, test results may return asynchronously, and physicians may update their diagnostic plans based on early results—even for batched orders. For example, even if the physicians initially batch ordered the tests, they could cancel some of these as they review the results. The paper would benefit from clarifying how often batched tests were actually completed and whether sequential result review modified downstream test execution. Without this, the causal link between batch ordering and increased imaging intensity may be overstated."*

Response: We thank the reviewer for this important clarification question. Our outcome measures count imaging tests actually performed, not merely ordered. This critical distinction strengthens our

44

causal interpretation: the 1.2 additional tests per marginally batched patient represent completed imaging studies that consumed resources and time, not provisional orders subsequently cancelled.

Based on consultation with our physician co-authors at Mayo Clinic and Massachusetts General Hospital, test cancellations after ordering face substantial operational barriers. Once the radiology department acknowledges an order, cancellation requires physicians to physically call and request the department to "push back" the imaging order. Furthermore, radiology departments often coordinate between modalities (e.g., CT and ultrasound) so patients move directly from one scanner to another. With radiologist read times averaging over 60 minutes, patients typically complete all batched imaging before initial results become available for physician review. While cancellations occasionally occur (e.g., a CT scan revealing appendicitis leading to a cancelled pelvic ultrasound), the operational friction makes these exceptions rather than the standard practice.

Importantly, even if some batched orders were cancelled, this would render our estimates conservative. Cancelled tests still consume operational resources—patients are queued, transported, and prepared for imaging that ultimately does not occur. These inefficiencies from cancelled tests would represent an additional operational burden beyond what our completed test counts capture. Thus, our measure of performed tests likely represents a lower bound on the actual operational impact of batching behavior.

We have clarified the distinction between ordered and performed tests in multiple sections:

Section 3.2.2 (Dependent Variables):

> "Beyond time-based metrics, we examine resource utilization through the number of distinct imaging tests performed during each ED encounter. This count variable helps us understand how batch ordering practices influence the diagnostic workload."

Section 4.2 (Results): Added clarification:

> "Discretionary batching also leads to more intensive diagnostic testing. Specifically, the marginal batched patient receives 1.2 more distinct imaging tests (completed studies with documented results), representing a 91% increase from the mean for standard care patients."

Table 4: Added footnote:

> "Test counts represent performed imaging studies with documented results, not ordered tests. The persistence of increased test volume under batching suggests cancellations do not substantially offset the effect."

Section 5.3 (Limitations): Added:

> "Our data contains imaging tests that were both ordered and performed. We cannot observe tests that were ordered but subsequently cancelled before completion. However,

45

> test cancellation after ordering requires substantial coordination—physicians must call the radiology department to remove patients from imaging queues. This operational friction makes cancellations rare. Moreover, cancelled tests introduce their own inefficiencies: patients experience delays from queuing and preparation, while ED resources are allocated to coordinate cancellations. To the extent that batched orders are more likely to include tests that are ultimately cancelled, our estimates of performed tests would understate the true operational burden of batching behavior."

These clarifications ensure that readers understand our analysis captures actual resource utilization and that potential cancellations would, if anything, strengthen rather than weaken our conclusions about the operational burden of batching.

R2 Wrote (Role of laboratory results):

"While the study restricts its scope to imaging due to its operational constraints, it does not address the interplay and potential dependencies between imaging and other diagnostic inputs—particularly lab results. One can imagine that in practice, lab results often arrive earlier and may prompt physicians to reassess imaging decisions, even when tests are initially batched. Could this potentially overstate the causal impact of batching on downstream outcomes?"

Response: We thank the reviewer for raising this important concern about whether earlier-arriving lab results might affect our estimates of batching's causal impact. This insightful observation prompted us to empirically examine the potential interplay between laboratory and imaging pathways, which has strengthened our analysis.

Empirical test of lab-imaging interplay: In response to similar concerns from Reviewer 1 about whether our instrument captures general diagnostic intensity rather than batching-specific behavior, we substantially expanded our controls to include laboratory tests ordered. This directly tests the mechanism the reviewer describes.

If batch tendency simply reflected physicians who order comprehensive diagnostics across all modalities—with labs potentially modifying imaging pathways—then controlling for whether labs were ordered should substantially attenuate our imaging effects. However, our results show imaging volume effects remain large, highly significant, and unchanged (1.241 additional tests, $p < 0.001$) even after controlling for laboratory utilization.

This pattern demonstrates that while batch tendency may correlate with some general diagnostic intensity, substantial imaging-specific variation persists after accounting for laboratory pathways. In other words, physicians with high batch tendencies order more imaging even when we compare patients who received identical laboratory workups. This directly addresses the reviewer's concern: the lab-imaging interplay does not substantially confound our estimates because we explicitly control for it.

The reviewer correctly notes that lab results often arrive before imaging results (typically 30-60

46

minutes for basic labs versus 90-165 minutes for imaging). Based on consultation with our physician co-author at the Mayo Clinic, normal lab results typically do not lead to imaging cancellation. Cancellation occurs only in rare, extreme cases: when labs provide a definitive diagnosis (e.g., severe anemia fully explaining dyspnea, thereby eliminating the need for a chest CT) or when results indicate instability precluding imaging (e.g., severe hyperkalemia requiring immediate dialysis). More commonly, abnormal lab results might modify the imaging type—for instance, switching from contrast to non-contrast CT for acute kidney injury—but do not eliminate the need for imaging. Importantly, even these modifications require physicians to call the radiology department, creating a significant workflow barrier that discourages treating batch orders as provisional.

Given these operational realities, imaging cancellations based on lab results are rare. Our outcome measures count performed tests, not ordered tests, so any cancellations that do occur would make our estimates conservative—we would be understating the initial commitment to imaging that batching represents. The fact that we observe 1.2 additional performed imaging tests per marginally batched patient demonstrates that batched imaging pathways, once initiated, proceed to completion despite the theoretical possibility of modification based on lab results.

Manuscript changes made:

To reflect this enhanced specification and address concerns about lab-imaging interplay, we have made the following revisions:

- **Enhanced primary specification (Table 4):** Our revised main results now include laboratory test ordering as a control variable. Column 3 shows our primary 2SLS specification controlling for patient characteristics, physician characteristics, hours into shift, and laboratory tests performed. The persistence of large, significant imaging effects (1.241 tests, p<0.001) demonstrates imaging-specific variation.

  These revisions demonstrate that the lab-imaging interplay the reviewer describes does not confound our estimates, as we explicitly control for laboratory pathways in our primary specification.

R2 Wrote (Limited scope of patient conditions):

"The study excludes many complaints where imaging is rarely ordered or batching is infeasible (e.g., dermatologic issues, urinary complaints). The results may not apply to low-acuity patient populations or fast-track EDs. Therefore, the study is more narrowly defined."

Response: We thank the reviewer for highlighting the focused nature of our sample. The reviewer is correct that we exclude complaints where imaging is rarely ordered or batching is uncommon (occurring less than 5% of the time). This exclusion was necessary for both statistical and substantive reasons.

Statistically, our instrumental variables approach requires sufficient variation in batching behavior to

identify effects. Including complaints where batching rarely occurs time would create weak instrument problems and yield unreliable estimates. For these rarely batched complaints, physicians have effectively determined that batching is clinically inappropriate—there is no discretionary decision to be made.

Substantively, the random assignment of patients to physicians at Mayo Clinic allows us to restrict our sample to complaints where batching is a relevant operational decision without introducing selection bias. This focused approach provides clear, actionable insights for specific clinical scenarios—trauma, neurological complaints, abdominal pain—where batching decisions are actively debated among emergency physicians and meaningfully impact ED operations. These complaints represent approximately 25% of the ED volume but account for 41% of imaging resource utilization in our data, underscoring their operational significance.

We agree that this narrows the scope of our study. However, this focus ensures our estimates are both statistically identified and operationally relevant for the clinical contexts where batching represents a genuine debated practice choice. We have added text to the limitations section explicitly acknowledging this scope:

> "Our findings apply to moderate-to-high acuity patients with complaints commonly requiring multiple imaging studies. The effects of batching in low-acuity or fast-track settings, where imaging is less common, remain unexplored."

R2 Wrote (Physician-only ordering & staffing mix):

"At Mayo, only emergency physicians (EPs) order imaging, which is not standard in many EDs (where nurse practitioners, physician assistants, or residents contribute). Does this limit the generalizability to other care settings? How does this difference manifest itself particularly in EDs with significant staffing by parttime or mid-level providers?"

Response: We appreciate the reviewer raising this important point about staffing differences across EDs. The reviewer is correct that Mayo Clinic's physician-only ordering policy differs from many EDs, where nurse practitioners, physician assistants, and residents participate in ordering decisions.

However, our replication at MGH addresses this concern of generalizability directly. MGH employs a mixed staffing model where mid-level providers participate under attending supervision—more representative of typical US emergency departments. As shown in Table 7, the effects at MGH (a 44.3% increase in LOS and 1.8 additional tests) are directionally similar and statistically indistinguishable from those at Mayo, demonstrating that the batching phenomenon persists across different staffing configurations.

Mayo's physician-only setting provides methodological advantages for causal identification by eliminating potential confounding from provider-type variation (e.g., systematic differences in ordering patterns between nurse practitioners and attending physicians) and ensuring all decisions reflect experienced clinical judgment. The random assignment mechanism is also simpler because patients

48

cannot be triaged to different provider types based on acuity or complaint. The consistency of results at MGH—despite its more complex staffing environment—validates that batching effects persist when mid-level providers participate in care delivery.

Our physician co-authors at both sites confirm that, while ordering processes differ, the fundamental trade-offs of batching versus sequential testing remain similar across settings. That said, we acknowledge that EDs with predominantly mid-level staffing may face additional dynamics not fully captured in our analysis, such as supervision requirements or handoff patterns that could interact with batching decisions. We have added text to Section 4.6:

> "While Mayo Clinic's physician-only ordering differs from many EDs, this provides cleaner identification of batching effects. Our replication at MGH—with its mixed staffing model including mid-level providers—demonstrates that these effects persist across different provider configurations, though the specific dynamics in EDs with predominantly mid-level staffing warrant future research."

R2 Wrote (Sensitivity of results – Table 4):

"I would urge the authors to discuss the difference in results between column 4 and 5 of Table 4, specifically with regards to sensitivity to influential controls, since the results before and after adding controls differ significantly."

Response:

We thank the reviewer for this important observation. In response to concerns from all three reviewers about our original specification, we substantially expanded the control set in our primary specification. The revised results of table 4 now includes two 2SLS columns (2 and 3) and two UJIVE columns (4 and 5) that differ only by the inclusion of additional precision controls.

While some coefficients shift in magnitude when precision controls are added, these shifts are moderate, consistent across outcomes, and conceptually aligned with what we would expect when removing non–imaging-related physician heterogeneity. This pattern is exactly what we would expect if the additional controls are successfully removing non–imaging-related physician heterogeneity while leaving intact the imaging-specific component that identifies the causal effect of batching. In other words, the precision controls do not overturn the substantive conclusions; rather, they help us isolate a slightly more conservative and more mechanism-consistent estimate.

Importantly, both 2SLS and UJIVE move in parallel when precision controls are added, and the two estimators remain close in magnitude. The convergence between these methods—despite their very different assumptions and sources of identifying variation—suggests that the core treatment effect is not being driven by omitted physician-level factors.

We have added discussion of this point in Section 4.1 (Results) to help readers interpret the modest but directionally meaningful changes between columns (4) and (5), and to highlight how the stability

Table 4: Effect of Batching Tests on Patient Outcomes

	Sequenced mean	2SLS (2)	2SLS (3)	UJIVE (4)	UJIVE (5)
Panel A. Primary Outcomes					
Log time to disposition	5.237	0.659***	0.651***	0.583***	0.522***
	(0.499)	(0.103)	(0.101)	(0.189)	(0.177)
Log LOS	5.490	0.717***	0.597***	0.653***	0.503***
	(0.456)	(0.094)	(0.088)	(0.158)	(0.144)
Number of distinct imaging tests	1.335	1.385***	1.241***	1.316***	1.174***
	(0.572)	(0.118)	(0.116)	(0.126)	(0.119)
72hr return with admission	0.012	-0.0137	-0.0146	-0.0079	-0.0039
	(0.110)	(0.018)	(0.019)	(0.020)	(0.022)
72hr return	0.030	-0.0512	-0.0536	-0.0440	-0.0396
	(0.170)	(0.029)	(0.031)	(0.032)	(0.034)
Panel B. Test Types					
X-ray	0.576	0.943***	0.989***	0.960***	0.959***
	(0.494)	(0.100)	(0.101)	(0.116)	(0.117)
Ultrasound	0.171	0.160**	0.087	0.164*	0.087
	(0.377)	(0.076)	(0.073)	(0.082)	(0.078)
CT without contrast	0.400	0.102	0.062	0.052	0.053
	(0.490)	(0.095)	(0.086)	(0.112)	(0.102)
CT with contrast	0.187	0.180*	0.102	0.140	0.075
	(0.390)	(0.078)	(0.076)	(0.087)	(0.079)
Panel C. Disposition					
Admission	0.279	0.419***	0.404***	0.424***	0.398***
	(0.449)	(0.096)	(0.088)	(0.103)	(0.090)
Necessary controls	—	Yes	Yes	Yes	Yes
Precision controls	—	No	Yes	No	Yes
Observations	11,679	11,679	11,679	11,679	11,679

Notes: Column 1 reports means for sequenced patients (standard deviations in parentheses). Columns 2–3 report 2SLS results using physician batch tendency as the instrument. Columns 4–5 report UJIVE estimates using ED provider identifiers as many weak instruments. All models include day-of-week and month fixed effects. Precision controls described in text. Standard errors are heteroskedasticity-robust.

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

across specifications supports the robustness of our findings.

R2 Wrote (Table 1 labeling issue):

"In Table 1, the independent variable and dependent variable seem reversed."

Response: We thank the reviewer for catching this labeling error. We have corrected the table to properly display "Batched" as the dependent variable and "Batch Tendency" as the independent variable. The table now clearly shows the first-stage relationship where physician batch tendency predicts the probability of batching for a given patient encounter. We apologize for any confusion this may have caused.

R2 Wrote (Potential impact & writing):

"The insights are policy-relevant, as they directly inform how EDs might design decision support tools, diagnostic protocols, or physician feedback systems to reduce overuse and streamline care. While the methodological innovation (a quasi-experimental IV strategy using random physician assignment) may not be groundbreaking in design, its application to clinical operations and diagnostic behavior is well-executed, adding credibility and potential for translation."

Response: We thank the reviewer for this encouraging assessment of our work's policy relevance and execution. We agree that the key contribution lies not in methodological innovation per se, but in applying rigorous causal inference methods to an important operational decision in emergency medicine that has long been debated. We are grateful that the reviewer sees the potential for our findings to inform practical interventions such as decision support tools and physician feedback systems. In response to this and other reviewers' suggestions, we have strengthened the policy implications section to provide more concrete guidance for ED managers considering interventions to optimize diagnostic test ordering practices.

VI. Responses to Referee 3 (R3) Comments

R3 Wrote (Research question & contribution:):

"Given the growing concern about overdiagnosis in general and its potential role in exacerbating ED over-crowding, I find the research question of this paper highly relevant to emergency medicine management. Due to the individualistic nature of decision-making in EDs, emergency physicians are often perceived as "cowboy doctors." While it may be challenging to implement strict guidelines for test-ordering behaviors in such a dynamic setting, I believe that providing empirical evidence on how ED physicians' decision-making affects system performance can be eye-opening and may encourage more informed behavior where feasible. Thus, this paper has the potential to make a meaningful contribution to the practice and management of emergency medicine. However, I have serious concerns about the empirical strategy employed in this paper, which raise questions about the validity of the findings. I outline my comments in detail below, with the hope that they will help the authors strengthen their work."

Response: We sincerely appreciate the reviewer's recognition of our paper's relevance to emergency medicine management and the important problem of overdiagnosis in contributing to ED overcrowding. We are grateful for your detailed methodological comments, which have substantially strengthened our work. In response to your concerns about the empirical strategy, we have: (1) implemented physician fixed effects as an alternative instrument construction as you suggested, (2) added additional physician level controls to our main specification, (3) added robustness checks using alternative model specifications, (4) provided detailed sample selection documentation, and (5) included heterogeneity analyses by complaint. We believe these revisions, detailed in our responses below, now provide the robust empirical foundation necessary to support our findings. We hope you find the strengthened methodology addresses your concerns.

R3 Wrote (Major concern - Instrumental variable):

"While the authors show that the proposed IV satisfies the relevance condition, I am not convinced that it satisfies the exclusion restriction. Specifically, "batchers" and "sequencers" may differ in their performance in ways that influence the outcome measures—not solely through the batching decision for the focal patient. This means that the IV might be capturing other aspects of physician behavior, which could bias the results.

Although the authors attempt to address this concern through a placebo test in the Appendix, the subsample used for this test appears to differ substantially from the main sample in terms of clinical conditions and outcome measures (comparing Tables 3 and D1). I encourage the authors to explore alternative IVs that better satisfy both the relevance condition and exclusion restriction.

On another point related to the proposed IV, the authors construct the IV using the residual from Equation (1). Although this model accounts for time fixed effects (FE) and observable patient and clinical characteristics, it does not account for unobserved non-physician-specific factors that may also be captured by the residuals. To more accurately estimate the physician-specific effect on batching decisions, I recommend that the authors include physician fixed effects in this model.

Similarly, to control for physician-specific characteristics where possible, I recommend including physician fixed effects in all models throughout the paper."

Response: We appreciate the reviewer's careful attention to whether our instrument satisfies the exclusion restriction. The reviewer's concern—that batchers and sequencers may differ in performance in ways beyond batching decisions—is exactly right to scrutinize. In response, we have substantially strengthened our identification strategy. We address the exclusion restriction concern first, then discuss the alternative instrument construction.

The reviewer is correct that if batch tendency captures other aspects of physician behavior (general diagnostic aggressiveness, thoroughness, experience-based practice patterns), this would bias our results. In our original specification, we controlled only for patient characteristics (vital signs, age, demographics, chief complaint severity) and temporal factors (shift-level effects). In response to concerns from all three reviewers, we have substantially expanded our control set to test whether batch tendency reflects general diagnostic intensity versus imaging-specific timing decisions.

1. Exclusion restriction concerns: Expanded controls

In our original specification, we controlled only for patient characteristics (vital signs, age, demographics, chief complaint by severity) and temporal factors (shift-level effects). In response to concerns from all three reviewers, we have substantially expanded our control set to better isolate imaging-specific ordering behavior.

Our revised primary specification now includes:

- **Patient characteristics:** Vital signs (tachycardic, tachypneic, febrile, hypotensive), age, demographics, chief complaint by severity

- **Physician characteristics:** Years of experience, gender, hours into shift (capturing potential fatigue effects)

- **Realized laboratory diagnostic intensity:** Laboratory tests performed during the encounter

- **Contextual factors:** ED capacity level (normal, minor overcapacity, major overcapacity),

- **Shift level FE:** Month-year and day-of-week $\times$ time-of-day fixed effects

The addition of laboratory test ordering is critical. If batch tendency captures physicians who are generally more aggressive or comprehensive diagnosticians, they should order both more imaging and more labs. By controlling for whether labs were actually ordered for each patient, we test whether imaging effects persist when accounting for the thoroughness of general diagnostics. If batch tendency simply reflected "comprehensive diagnosticians," controlling for realized lab intensity should substantially attenuate the imaging-specific effects.

First-stage results validate the concern while demonstrating that we can isolate the batching-specific component. Below, is our revised first-stage regression (Table 2), both without and with the expanded precision controls:

53

Table 2: First-Stage Results: Batch Tendency and Batching

	Dependent Variable: *Batched*	
	(1)	(2)
Batch Tendency	1.837***	1.752***
	(0.1379)	(0.1336)
Fixed Effects		
Necessary Controls	Yes	Yes
Precision Controls	No	Yes
Observations	11,679	11,679
R^2	0.02248	0.07359
Within R^2	0.01939	0.03521
First-stage F-stat	177.5	171.9

Notes: First-stage regressions of batching on batch.tendency (leave-one-out physician residualized batching propensity). Necessary controls include day-of-week ×time-of-day and month fixed effects. Precision controls include patient vital signs, age, demographics, chief complaint-severity fixed effects, race, gender, physician experience, physician gender, hours into shift, laboratory ordered, and ED capacity controls. *** p<0.001.

We see that even when controlling for this comprehensive set of precision controls, substantial independent variation remains (F=171.9, well above weak instrument thresholds). This is precisely what we would expect if the instrument captures batching-specific propensity rather than general cautiousness.

Our revised main results (Table 4) now incorporate the expanded controls and respond to the reviewers suggestion to employ methods that use physician FE (which we describe in (3)). In addition to our standard 2SLS estimates using batch tendency as the instrument, Table 4 includes estimates from the Unbiased Jackknife Instrumental Variables Estimator (UJIVE), following **?**. UJIVE addresses a specific econometric concern: when 2SLS uses many physician indicators as instruments, it constructs estimated leniency partly from each patient's own treatment status, creating mechanical correlation with the outcome error that biases estimates toward OLS. UJIVE eliminates this many-instrument bias through leave-one-out estimation. We emphasize that UJIVE, like all IV estimators, still requires the exclusion restriction to hold—our expanded controls and Conley sensitivity analysis address exclusion restriction concerns directly. The agreement between 2SLS and UJIVE estimates provides validation that many-instrument bias is not driving our findings

The attenuation in the 2SLS estimates after adding precision controls confirms the reviewer's intuition that our original instrument may have contained some correlation with general diagnostic intensity. At the same time, the close agreement between the 2SLS and UJIVE estimates—both before and after adjustment—provides strong evidence that any remaining correlation between a physician's

Table 4: Effect of Batching Tests on Patient Outcomes

	Sequenced mean	2SLS		UJIVE	
		(2)	(3)	(4)	(5)
Panel A. Primary Outcomes					
Log time to disposition	5.237	0.659***	0.651***	0.583***	0.522***
	(0.499)	(0.103)	(0.101)	(0.189)	(0.177)
Log LOS	5.490	0.717***	0.597***	0.653***	0.503***
	(0.456)	(0.094)	(0.088)	(0.158)	(0.144)
Number of distinct imaging tests	1.335	1.385***	1.241***	1.316***	1.174***
	(0.572)	(0.118)	(0.116)	(0.126)	(0.119)
72hr return with admission	0.012	-0.0137	-0.0146	-0.0079	-0.0039
	(0.110)	(0.018)	(0.019)	(0.020)	(0.022)
72hr return	0.030	-0.0512	-0.0536	-0.0440	-0.0396
	(0.170)	(0.029)	(0.031)	(0.032)	(0.034)
Panel B. Test Types					
X-ray	0.576	0.943***	0.989***	0.960***	0.959***
	(0.494)	(0.100)	(0.101)	(0.116)	(0.117)
Ultrasound	0.171	0.160**	0.087	0.164*	0.087
	(0.377)	(0.076)	(0.073)	(0.082)	(0.078)
CT without contrast	0.400	0.102	0.062	0.052	0.053
	(0.490)	(0.095)	(0.086)	(0.112)	(0.102)
CT with contrast	0.187	0.180*	0.102	0.140	0.075
	(0.390)	(0.078)	(0.076)	(0.087)	(0.079)
Panel C. Disposition					
Admission	0.279	0.419***	0.404***	0.424***	0.398***
	(0.449)	(0.096)	(0.088)	(0.103)	(0.090)
Necessary controls	—	Yes	Yes	Yes	Yes
Precision controls	—	No	Yes	No	Yes
Observations	11,679	11,679	11,679	11,679	11,679

Notes: Column 1 reports means for sequenced patients (standard deviations in parentheses). Columns 2–3 report 2SLS results using physician batch tendency as the instrument. Columns 4–5 report UJIVE estimates using ED provider identifiers as many weak instruments. All models include day-of-week and month fixed effects. Precision controls described in text. Standard errors are heteroskedasticity-robust.

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

overall comprehensiveness and patient outcomes is too small to materially bias the results. The close agreement between 2SLS and UJIVE estimates, despite their different constructions of identifying variation, indicates that many-instrument bias is not driving our results. This convergence, combined with our expanded controls for laboratory ordering and physician characteristics, strengthens

confidence that our revised specification isolates batching-specific effects.

Additionally, following an approach outlined by Conley et al. (2012), we test how our estimates change in our 2SLS results if we relax the exclusion restriction by allowing batch tendency to have a direct effect on outcomes. Under standard IV assumptions, the structural equation is

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

which assumes that batch tendency affects outcomes only through its effect on Batched_i.

Conley's relaxation allows for a small violation of the exclusion restriction by introducing a direct effect of the instrument:

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

where δ captures any exclusion restriction violation, including

- direct effects (e.g., physicians with higher batch tendency directly prolong LOS),

- backdoor paths (e.g., batch tendency reflects physician aggressiveness, which also affects LOS),

We then estimate $\beta(\delta)$ for a range of plausible values of δ to assess how sensitive our conclusions are to such violations.
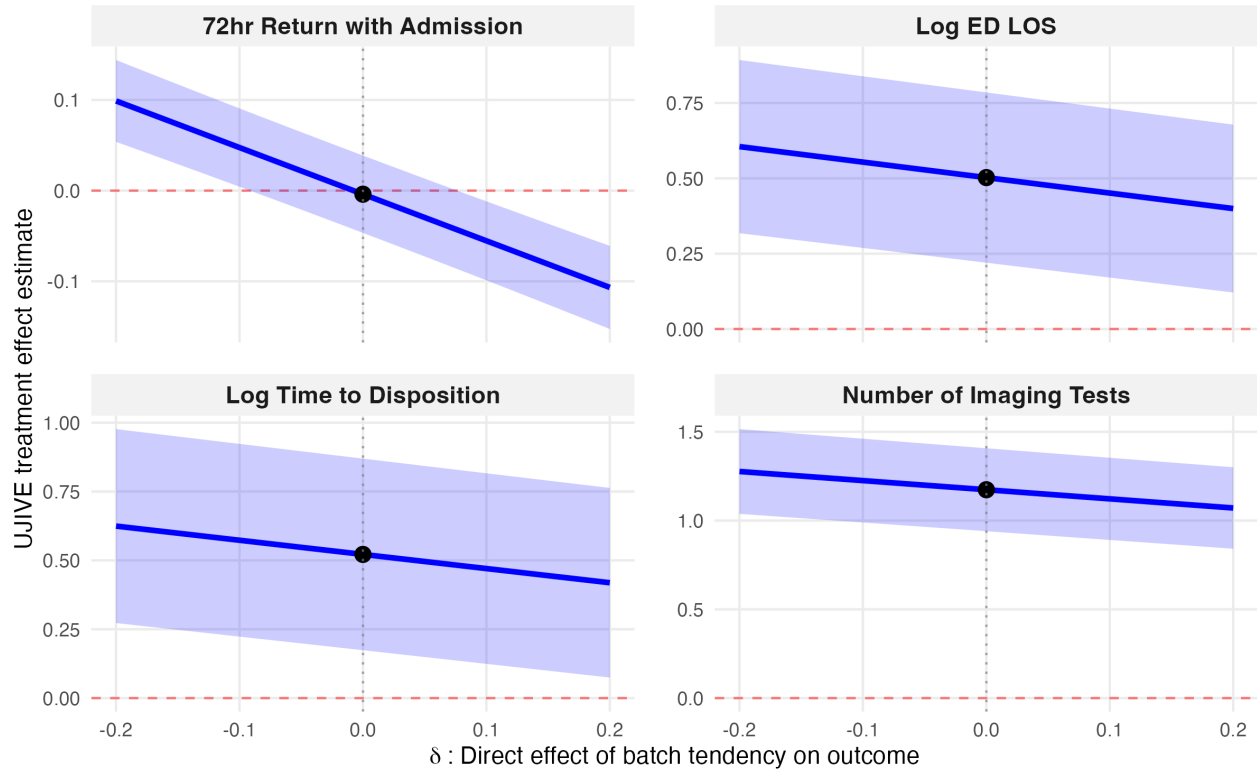
The figure below (now included in EC.9) presents sensitivity analyses showing substantial robustness. For all outcomes, the treatment effect remains significant across $\delta \in [-0.2, 0.2]$.

In summary, while we cannot definitively rule out all exclusion restriction violations—no observational study can—the convergence of evidence strongly supports our interpretation:

- Treatment effects persist after adding additional controls for general diagnostic intensity (laboratory ordering), experience, and hours into shift

- UJIVE estimates, which eliminate many-instrument bias, closely match 2SLS results

- 2SLS estimates remain robust to potential backdoor paths for plausibly large exclusion restriction violations

The LATE we identify represents the causal effect for compliers—patients whose imaging strategy depends on physician practice style rather than clinical necessity. This is precisely the margin where ED management interventions can influence practice without constraining clinically necessary care.

56

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

Notes: UJIVE estimates under different assumptions about direct effects (δ) of physician batch tendency on outcomes. One unit of δ implies a one percentage point increase in batch tendency directly changes the outcome by δ units. Black points indicate baseline estimates. Shaded areas show 95% confidence intervals using heteroskedasticity-robust standard errors following Conley et al. (2012).

We have revised the manuscript to reflect these more conservative estimates while maintaining transparency about our identification assumptions. The core finding remains: discretionary batch ordering increases ED length of stay by approximately 65-81% without improving short-term patient outcomes, suggesting substantial opportunity for operational improvement through policies that preserve diagnostic flexibility.

2. Placebo test clarification:

The reviewer notes our placebo subsample differs from the main sample. This is by design. We examine complaints where batching occurs less than 1% of the time (isolated extremity injuries requiring single X-rays). For these patients, there is no discretionary batching decision—clinical protocols dictate single imaging.

The test asks: if batch tendency captured general physician performance, would we see effects for patients where batching cannot occur? Table D.1 shows null effects (all p>0.05), supporting that batch tendency captures batching-specific behavior. The clinical differences between samples are due

to the test's mechanism—we deliberately select patients where the treatment cannot vary to verify that the instrument does not operate through other channels.

Manuscript changes:

- **Enhanced main specification (Table 4):** All results now reflect our enhanced specification with full controls including physician covariates, contextual factors, and realized laboratory ordering intensity. Column 5 shows our primary 2SLS specification.

- **Enhanced methods discussion (Section 3.3):** We explain our expanded control strategy and why controlling for laboratory ordering addresses concerns about general diagnostic intensity.

- **Expanded limitations (Section 4.8):** We acknowledge that we cannot definitively rule out all exclusion restriction violations. We note that to the extent batch tendency captures general physician intensity, our reduced-form estimates still provide the causal effect of being assigned to a high-batching physician.

3. Alternative instrument construction: Physician fixed effects

The reviewer encourages exploring alternative IVs and suggests including physician FE in Equation (1). However, this would eliminate the very variation we need. If we include physician FE in:

$$\text{Batched}_{it} = \alpha_{\text{time}} + \beta X_{it} + \alpha_j + \varepsilon_{it}, \tag{1}$$

the fixed effects estimator would remove all between-physician variation by construction. The residuals $\hat{\varepsilon}_{it}$ would contain zero physician-specific information. Our leave-out mean instrument:

$$\text{BatchTendency}_{ij} = \frac{1}{N_{-i,j}} \sum_{i' \neq i} \hat{\varepsilon}_{i'j},$$

would converge to zero for all physicians.

Although including physician fixed effects in Equation (1) is infeasible because it would remove exactly the between-physician variation that forms the basis of our instrument, we fully agree with the reviewer's underlying concern and with the spirit of the recommendation. Your comment directly motivated us to adopt an estimator designed precisely for this situation—one that eliminates the component of the instrument correlated with stable, provider-level traits while preserving the identifying variation across physicians. Specifically, we now complement our 2SLS estimates with the *Unbiased Jackknife Instrumental Variables Estimator (UJIVE)* of Kolesár, Chetty, Friedman, Glaeser, and Imbens (2015) and Phillips and Su (2017), which is explicitly built for settings with many potentially weak or noisy instruments and heterogeneous violations of the exclusion restriction.

58

Our primary empirical strategy uses a single continuous instrument—physician batch tendency—but the UJIVE framework requires the full set of physician indicator variables as instruments. Using these indicators allows us to test whether our main 2SLS estimates are affected by the mechanical many-instrument bias documented in the leniency-design literature, while preserving the underlying source of identifying variation.

We proceed in the following way: let Z_i denote the vector of physician indicator variables for encounter i, and let P be the projection matrix onto all controls in the first-stage equation. Define $M = I - P$ as the associated residual-maker matrix. For each observation i, UJIVE re-estimates the physician effects using all observations *except i*:

$$\hat{\pi}_{-i} = \arg \min_{\pi} \sum_{k \neq i} \left(MB_k - Z_k' \pi \right)^2 , \tag{2}$$

where B_k is the observed batching decision. This yields a leave-one-out predicted instrument for encounter i:

$$\hat{\ell}_{i,-i} = Z_i' \hat{\pi}_{-i}. \tag{3}$$

Because $\hat{\pi}_{-i}$ is estimated using all observations except i, the predicted instrument $\hat{\ell}_{i,-i}$ is orthogonal to the structural error by construction. This removes the mechanical correlation that arises when many physician indicators enter the first stage as instruments.

The UJIVE estimator is then obtained in the usual Wald/IV ratio:

$$\hat{\beta}_{UJIVE} = \frac{\sum_i \hat{\ell}_{i,-i} Y_i}{\sum_i \hat{\ell}_{i,-i} Batched_i} = \frac{\widehat{Cov}(\hat{\ell}_{i,-i}, Y_i)}{\widehat{Cov}(\hat{\ell}_{i,-i}, Batched_i)}. \tag{4}$$

By performing leave-one-out estimation of each physician's effect, UJIVE removes the influence of stable, provider-level traits from the instrument, isolating variation driven by the random patient assignment mechanism. While UJIVE does not solve the exclusion restriction; rather, it provides a rigorous robustness check that corrects the potential many-instrument bias of 2SLS. The agreement between UJIVE and 2SLS estimates—despite their very different constructions—indicates that many-instrument bias is not driving our findings and that both approaches identify similar underlying variation.

Our UJIVE estimates closely track our 2SLS estimates across all outcomes, both before and after including the expanded set of precision controls. This parallel movement indicates that any remaining correlation between batch tendency and unobserved physician traits is unlikely to materially bias our estimates. The agreement between these two very different estimators strengthens confidence that our results reflect batching-specific effects rather than broader differences in physician diagnostic style.

We are extremely grateful to the reviewer for pushing us to strengthen our identification strategy and giving us the opportunity to respond to these concerns. These additional analyses have strengthened the manuscript and our confidence in the robustness of the estimated causal effects.

R3 Wrote (Major concern - Model selection):

"The authors use a linear model for all outcome variables, regardless of whether the outcomes are binary, count, or continuous. Are the results robust to more appropriate model specifications that better align with the nature of each outcome variable?"

Response: We thank the reviewer for this methodological question. The reviewer correctly notes that our outcomes include binary (admission, 72-hour return), count (number of tests), and continuous (log time) variables. We use linear models throughout our 2SLS analysis, which is standard practice in the causal inference literature for important econometric reasons we explain below.

While nonlinear models (logit/probit for binary outcomes, Poisson for counts) seem more appropriate for the outcome distributions, using them in instrumental variables settings creates fundamental econometric problems. The key issue is what Hausman (1975, 1978) termed the "forbidden regression."

When using a nonlinear first stage (e.g., probit for our binary batching variable), substituting the fitted values $\hat{d}_i$ into any second stage creates:

$$y_i = \beta_0 + \beta X_i + \gamma \hat{d}_i + [\epsilon_i + \gamma(d_i - \hat{d}_i)]$$

This fails because with a nonlinear first stage, the residuals $(d_i - \hat{d}_i)$ are correlated with $\hat{d}_i$ even asymptotically, unless the first-stage functional form is exactly correct—an untestable assumption. Consistency of 2SLS estimates does not depend on correct specification of the first-stage conditional expectation function, but this robustness is lost with nonlinear first stages (Angrist and Pischke (2009)).

Using nonlinear models in the second stage is even more problematic. The linear 2SLS estimator provides a well-defined Local Average Treatment Effect (LATE) for compliers. Nonlinear second-stage models would require strong assumptions about the entire joint distribution of errors and lack the LATE interpretation. Combining the nonlinear first and second stages compounds these problems.

The linear probability model in 2SLS, while potentially producing predictions outside [0,1] for individual observations, yields consistent estimates of the average marginal effect for compliers—exactly the parameter of policy interest in our quasi-experimental setting.

Nevertheless, to address the reviewer's concern about functional form, we verify that our results are robust to nonlinear outcome models while retaining a linear first stage. Specifically, we estimate logit models for binary outcomes (admission, 72-hour return) and Poisson models for count outcomes (number of imaging tests), using the same linear first stage with batch tendency as the instrument. We then compute average marginal effects (AMEs) from these nonlinear models and compare below

60

EC.10.1: Robustness of Estimates to Outcome Model Specification

Outcome	Type	2SLS	Nonlinear Model (AME)
72-hour return with admission	Binary (logit)	−0.0146	−0.0153
		(0.0195)	(0.0231)
Admission	Binary (logit)	0.404***	0.377***
		(0.088)	(0.075)
Number of imaging tests	Count (Poisson)	1.241***	1.197***
		(0.116)	(0.241)

Notes: Each column reports estimates of the effect of batching. 2SLS uses physician batch tendency as the instrument and includes the full control set. Nonlinear outcome models replace the second-stage linear model with logit for binary outcomes and Poisson for counts Standard errors for nonlinear AMEs are delta-method estimates. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

(now included in EC.10).

Manuscript changes:

We have added a discussion in Section 3.4 (Empirical Specification) explaining our choice of linear models in the IV framework, with reference to the econometric literature. We have also added the included robustness table to EC.10, which shows the OLS robustness check with nonlinear models.

We appreciate the reviewer raising this methodological point, as it allows us to clarify why the linear model is preferable in IV settings while also demonstrating that functional form does not drive our OLS results.

R3 Wrote (Major concern - Sample selection):

"The final sample for the primary data includes less than 25% of all ED encounters. Could the authors provide more details regarding the sample selection process, including the exact number of observations excluded with each criterion?
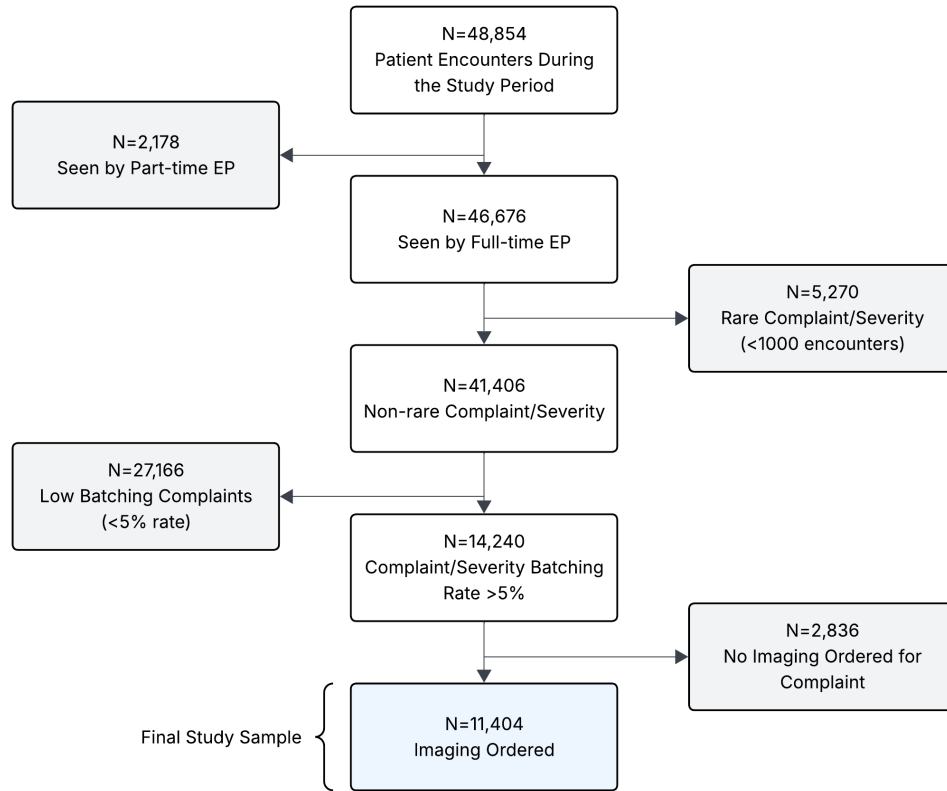
Given such a substantial reduction in sample size, it is crucial for the authors to compare the excluded and included encounters to ensure that the observed effects are not limited to a small, non-representative subsample of ED visits."

Response: We thank the reviewer for requesting greater transparency about sample selection. We address this concern by (1) providing a detailed CONSORT flow diagram showing exact exclusion counts, (2) comparing characteristics of excluded versus included encounters, and (3) explaining why our focused sample strengthens rather than limits our contribution.

1. CONSORT flow diagram:

Figure EC.3.1 (EC.3) shows the complete sample selection process:

Starting with 48,854 encounters of adult patients, we exclude: (1) encounters with non-full time

Notes: Sample selection for Mayo Clinic ED encounters (October 2018 - December 2019). Starting with 48,854 encounters, we apply sequential exclusion criteria to arrive at our analytical sample of 11,404 encounters. Rare complaints have fewer than 1,000 total encounters. Low-batching complaints have batching rates below 5%.

providers, (2) rare complaints (<500 encounters), (3) complaints where batching occurs <5% of the time, and (4) complaints that do not involve any imaging. This yields our analytical sample of 11,679 encounters (23.7%). We made this decisions based on conversations and discussions with our medical collaborators (and co-authors on this paper) who are in charge of operations in two leading US hospitals (Mayo Clinic and MGH).

The table included below (now included in EC.3) compares patient characteristics and outcomes between excluded and included encounters. As can be seen, the analytical sample differs from excluded encounters in clinically meaningful and expected ways. Included patients are older (62.9 vs. 57.5 years), higher acuity (ESI 2.52 vs. 2.87, where lower numbers indicate greater severity), and more likely to present with concerning vital signs (febrile: 4.6% vs. 1.4%, hypotensive: 2.3% vs. 1.2%). Consequently, they experience longer ED stays (272 vs. 238 minutes) and higher admission rates (28.4% vs. 17.0%).

These differences strengthen rather than limit our contribution. The analytical sample comprises precisely the population where imaging decisions matter most: moderate-to-high acuity patients

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	37,450	11,679	
% of total sample	76.2%	23.7%	

Notes: Comparison of patient characteristics and outcomes between excluded encounters (n=37,450) and analytical sample (n=11,651). 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. *** $p<0.001$.

presenting with complaints commonly requiring diagnostic workups. These are the encounters where physicians face genuine uncertainty about optimal testing strategy and where efficiency gains from improved ordering practices would be most impactful. Excluded encounters—primarily low-acuity visits where imaging is rarely needed—represent settings where batching decisions are either clinically inappropriate or operationally irrelevant.

Importantly, these differences reflect an appropriate sample definition rather than selection bias. Mayo Clinic's random assignment mechanism ensures physicians see the full spectrum of acuity and complaints. Our analytical sample focuses on encounters where the batching decision is both consequential and discretionary—precisely the population where our LATE provides actionable policy guidance.

In addition to medical relvance, our sample restrictions serve two purposes, both standard in the econometric literature relative to "judges design."

First, instrumental variable identification requires variation. Our IV approach needs sufficient batching variation to generate a strong first stage. Including complaints where batching rarely occurs (or definitionally could never occur) would create weak instrument problems, yielding imprecise and uninformative estimates of the LATE.

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. This is precisely the population of policy interest: encounters where multiple imaging pathways are clinically plausible and efficiency tradeoffs matter. Our sample includes complaints commonly requiring imaging (Neurological Issues, Abdominal Pain, Chest Pain, Falls/Trauma, Dizziness/Syncope, Extremity Complaints, Constitutional Symptoms) where physicians make consequential ordering decisions.

Importantly, random assignment ensures these exclusions do not introduce selection bias. Physicians receive patient with all complaint types through Mayo Clinic's rotational system. We focus our analysis on complaints where their batching decisions vary meaningfully. We verify balance by: (1) computing batch tendency using physicians' behavior across ALL encounters before exclusions, and (2) confirming patient characteristics remain balanced within our analytical sample (Figure 2).

While our results apply to the 25% of ED visits requiring imaging workups, we would like to highlight again that this represents the clinically and operationally relevant population as indicated by our medical collaborators. Furthermore, this 25% of ED visits consume 41% of imaging resources in our data. Our focused approach identifies the precise effects where physician practice style influences operations—the margin where ED management interventions can meaningfully improve efficiency, and the margin where physicians disagree on optimal testing strategies.

Manuscript changes:

- Added detailed CONSORT diagram (Figue EC.3.1)

- Added comparison of excluded vs. included encounters (Table EC.3.2)

- Revised Section 3.2 to clarify sample selection rationale with appropriate citations

- Added discussion of generalizability in Section 5.3 (Limitations)

We appreciate the reviewer's attention to this issue, which has led us to provide greater transparency about our sample construction and its implications for interpreting our findings.

R3 Wrote (Major concern - Variable selection):

"To estimate the impact of batch ordering on productivity, the authors focus on LOS and time to disposition. However, the most relevant outcome for assessing physician practice is treatment time, defined as the period from the start of assessment to disposition. I recommend that the authors consider this metric, as it excludes both the waiting time before assessment and the boarding time after disposition.

The authors use a 72-hour ED revisit leading to hospital admission as an indicator of care quality. While this is conceptually a valid measure, it is more common in both the operations management and medical literature to use ED revisit—regardless of admission status—as an indicator of adverse outcomes. Are the results robust

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	11,679	11,679	11,679	11,679	11,679

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 many weak instruments. All models include day-of-week and month fixed effects. Precision controls include patient vital signs, demographics, complaint-severity fixed effects, physician characteristics, laboratory ordering, hours into shift, and ED capacity. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

to this alternative measure of quality? From the estimation perspective, this measure should be a better choice given the scarcity of ED revisits leading to hospital admission.

On a related note, how do the authors measure 72-hour ED revisit for patients who were admitted to the hospital during the focal visit?"

Response: We thank the reviewer for these important suggestions about outcome measures. We address each in turn:

1. Treatment time:

We appreciate the reviewer highlighting treatment time as a valuable measure that excludes both waiting room delays and post-disposition boarding. Although our primary time-to-disposition measure already excludes boarding time and our controls mitigate variation in waiting room delays, we fully agree that treatment time provides a clean and complementary measure.

We have therefore added this analysis using treatment time (defined as the time from the first physician contact to the disposition decision). The results, shown in the table below (now in EC.11), demonstrate that our findings are robust to this alternative specification:

The close agreement between the full-control 2SLS and UJIVE estimates further supports the robustness of our findings. Although UJIVE exhibits the expected attenuation from correcting for provider-level heterogeneity, both estimators produce effects of comparable size and identical qualitative implications.

2. 72-hour revisit measures:

The reviewer correctly notes that any 72-hour ED revisit is more commonly used than revisit-with-

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

	Sequenced mean	2SLS		UJIVE	
		(2)	(3)	(4)	(5)
72-hr return with admission	0.012	−0.0137	−0.0146	−0.0079	−0.0039
	(0.110)	(0.0182)	(0.0195)	(0.0203)	(0.0217)
72-hr return (any reason)	0.030	−0.0512	−0.0536	−0.0440	−0.0396
	(0.170)	(0.0287)	(0.0311)	(0.0322)	(0.0344)
Necessary controls	—	Yes	Yes	Yes	Yes
Precision controls	—	No	Yes	No	Yes
Observations	11,679	11,679	11,679	11,679	11,679

Notes: Column 1 reports means for sequenced patients (standard deviations in parentheses). Columns 2–3 report 2SLS estimates using batch tendency as the instrument. Columns 4–5 report UJIVE estimates using ED provider identifiers as many weak instruments. All specifications include day-of-week and month fixed effects. Precision controls include demographic, clinical, physician, and contextual covariates as described in Section 3. Standard errors are heteroskedasticity-robust.

admission. We focused on 72-hour returns requiring admission based on guidance from our physician coauthors (including emergency physicians at both of our partner hospitals), who indicated that this measure better captures true quality failures. Some ED revisits are planned or expected—patients may be instructed to return for wound checks, suture removal, or if symptoms persist after initial treatment. Furthermore, some ED physicians prefer to discharge their patients and ask them return, because they want to check what other symptoms develop before treating them. That is, they intentionally ask for a return, and this should not be measured as an incident of low quality (mis treatment during the first visit). Returns requiring admission, however, are more likely to indicate missed diagnoses or inadequate initial treatment.

Nevertheless, we acknowledge the reviewer's point about statistical power and comparability with prior literature. We therefore re-estimated our models using any 72-hour ED revisit as the outcome.

The full-control 2SLS estimates for treatment time (0.628) and time to disposition (0.651) are nearly identical in magnitude, and UJIVE estimates follow the same pattern (0.447 vs. 0.522), confirming that batching delays patients similarly across both operational time measures.

3. Measurement for admitted patients:

For patients admitted during their index visit, the 72-hour window begins at hospital discharge, not ED departure. Admitted patients cannot revisit the ED while hospitalized. This standard approach ensures fair comparison across all patients regardless of initial disposition. We have clarified this measurement approach in Section 3.2.2.

Manuscript changes:

- Added treatment time analysis (EC.11.1)

- Added any 72-hour return analysis (EC.11.2)

- Clarified 72-hour return measurement for admitted patients (Section 3.2.2)

We are grateful for the reviewer's suggestions, which allow us to demonstrate that our findings are robust to alternative outcome specifications while maintaining our focus on the most clinically meaningful quality measure.

R3 Wrote (Major concern - Heterogeneity analysis):

"The authors correctly discuss the trade-offs between the advantages and disadvantages of batch ordering diagnostic tests compared to sequential test ordering. However, the paper subsequently focuses primarily on the disadvantages of batch ordering. While it is important to quantify the overall net benefit or cost of ordering diagnostic tests in advance, I believe the paper would provide more comprehensive insights for practice if it also identified the conditions under which one strategy outperforms the other. Specifically, are there certain chief complaints for which a batching strategy leads to better outcomes? I would expect this to be the case for more complex chief complaints that generally require a greater number of diagnostic tests.

In addition, is there evidence of heterogeneity in the effect or magnitude of the impact of batch ordering across conditions where batching is more common? Understanding such variation could help tailor diagnostic strategies to different clinical scenarios."
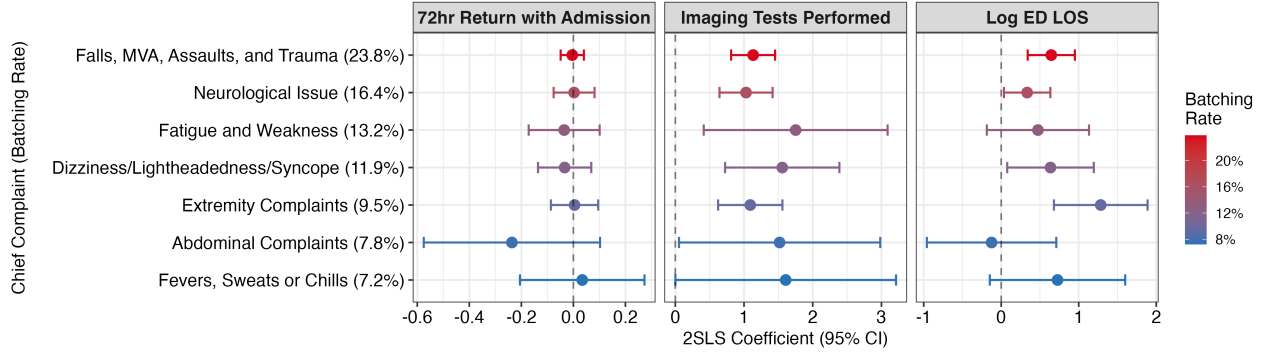
Response: We thank the reviewer for this excellent suggestion to explore heterogeneity in batching effects. We conducted a comprehensive analysis examining effects across our seven chief complaint categories, which vary substantially in both clinical complexity and batching prevalence (ranging from 7% for fevers to 24% for polytrauma).

Figure 4 (included in this document for the reviewer's convenience) presents this heterogeneity analysis, with complaints ordered by batching rate. The color gradient illustrates variation in batching prevalence, from red (high-batching complaints) to blue (low-batching complaints).

We frame this as an exploratory analysis given multiple hypothesis testing concerns across seven categories and three outcomes. Rather than emphasizing statistical significance in subgroups, we focus on whether any clinical scenario shows qualitatively different patterns in direction or magnitude.

The reviewer's intuition that batching might benefit clinically complex complaints is reasonable, but our empirical analysis shows no such pattern. Imaging effects are consistently positive across all seven complaint categories, ranging from 1.03 to 1.76 additional tests, with six of seven showing statistically significant increases despite smaller subgroup sample sizes. Notably, Falls/MVA/Assaults/Trauma—our most complex category with the highest batching rate (23.8%)—shows a 1.13-test increase (SE = 0.16, $p < 0.001$). Similarly large increases are observed for neurological issues (1.03; SE = 0.20) and dizziness/lightheadedness/syncope (1.56; SE = 0.42). Time effects show no evidence of efficiency gains in any category. The complaint-specific effects on log ED LOS closely track the pooled estimate and are non-negative with the exception of abdominal complaints.

67

Figure 4: Heterogeneity in Batch Ordering Effects by Chief Complaint

Notes: Each panel shows 2SLS estimates of batching effects for different chief complaints, ordered by batching rate (shown in parentheses). Point colors indicate batching prevalence. Error bars show 95% confidence intervals. All models include full controls.

Quality effects are uniformly small and indistinguishable from zero. Across all seven complaints, the 72-hour return-with-admission estimates range from -0.24 to +0.06, with all confidence intervals crossing zero ($p > 0.10$). Higher-complexity categories again show no evidence of quality differences (e.g., trauma: -0.004; SE = 0.023; neurological: 0.003; SE = 0.040). These patterns mirror those in our pooled analysis and reinforce that batching does not improve short-term quality outcomes for any clinical scenario in our data.

The consistency across complaint types is striking. Despite meaningful variation in complexity and batching prevalence (7%–24%), the same pattern emerges in every subgroup. This uniformity suggests that recommendations which discourage discretionary batching apply broadly. Even in complex presentations where batching is most common, ordering imaging tests in sequence appears to allow physicians to target diagnostics based on early clinical signals rather than preemptively ordering test bundles that often prove unnecessary.

Manuscript changes:

We have added Section 4.4, which presents this heterogeneity analysis, with Figure 4 showing coefficient plots. Section 5.2 discusses implications for tailoring diagnostic strategies, noting that the lack of heterogeneity simplifies policy recommendations.

We appreciate this suggestion, which strengthens our finding that discretionary batching increases resource utilization without demonstrable benefits across diverse clinical scenarios.

R3 Wrote (Major concern - Results interpretation):

"The argument regarding heterogeneity based on ED capacity status is not valid and requires closer examination. Although the effect magnitudes reported in Table 5 differ across occupancy levels, a closer look at the standard errors indicates that these differences are not statistically significant."

Response: We thank the reviewer for raising this issue. We agree that differences in the estimates

across crowding strata do not imply statistical heterogeneity. We do note a drop in the observed batch rate during major overcapacity relative to normal operations (12.4% vs. 14.8%), suggesting physicians may alter ordering behavior when the ED is severely crowded. However, as pointed out by the reviewer, the estimated effects of batching on all outcomes remain positive and of similar magnitude across capacity levels, with overlapping confidence intervals.

We have updated the manuscript to clarify this and have added a statement emphasizing that the estimated effects of batching are statistically indistinguishable across capacity levels. The stability of the 2SLS estimates across all three operational states reinforces the robustness and generalizability of our main findings.

R3 Wrote (Major concern - Paper organization):

"At several points, I found the paper difficult to follow due to its current organization. Below, I provide a few examples, but I strongly recommend that the authors consider reorganizing the paper to present a more coherent narrative, maintain a logical flow, and avoid abrupt transitions between topics.

- *A dedicated sub-section on empirical challenges and strategy (Section 1.1) in the Introduction seems unnecessary, as it distracts from the main purpose of the paper. I recommend condensing this discussion into a brief paragraph that highlights the key aspects of the empirical strategy, while providing the full details later in Section 3.*
- *From the discussion in Section 3.3, it is not clear that the authors are introducing the IV until the end of this section on page 13. I suggest that the authors begin by clearly presenting the main model (second stage in Equation (4)), explicitly discuss the endogeneity concern, and then introduce the proposed IV as the solution to address this challenge."*

Response: We thank the reviewer for these constructive suggestions about manuscript organization, which have substantially improved clarity and flow.

Regarding Section 1.1, we have condensed the empirical challenges discussion into a single paragraph in the Introduction, which briefly previews our quasi-experimental approach. We have moved all technical details to Section 3, where they can be developed systematically alongside our identification strategy. This revision allows the Introduction to focus on motivating the research question and contribution without prematurely diving into methodological details.

Regarding Section 3.3, we have completely restructured this section to follow the logical progression the reviewer suggests:

- We now begin by presenting the second-stage equation (Equation 4) that defines our estimand
- We then explicitly discuss the endogeneity problem—that batching decisions correlate with unobserved patient complexity
- We introduce our instrumental variable (physician batch tendency) as the solution, explaining how random assignment creates exogenous variation in treatment

69

- We conclude with identification assumptions and validation tests

This reorganization makes the IV strategy immediately clear, eliminating the need for readers to work through preliminary material before understanding the core identification approach. We have also added subheadings throughout Section 3 to improve navigation and added transitional sentences between subsections to maintain narrative flow.

We believe these changes has improved our exposition, and we thank the reviewer for their excellent suggestions.

R3 Wrote (Other concerns):

"In Table 1, please provide summary statistics for all ED performance measures considered in the paper."

Response: We thank the reviewer for this request. We have expanded Table 1 to include all ED performance measures analyzed in the paper. The following variables and their corresponding descriptive statistics have been added to Table 1:

- Time to disposition (mins)

- Treatment Time (mins)

- Number of imaging tests ordered

- 72-hour returns

- 72-hour returns with admission

The updated Table 1 now comprehensively displays all ED performance metrics examined in our analysis, allowing readers to better contextualize the magnitude of our treatment effects.

"Please clarify how you calculate the percentage increase in duration outcomes from the estimates presented in the results tables."

Response: We thank the reviewer for requesting this clarification. Since our outcome variables are log-transformed (ln(ED LOS), ln(time to disposition), ln(treatment time)), the coefficients represent log-point changes. To convert these to percentage changes, we use the standard transformation for log-linear models: $(\exp(\beta) - 1) \times 100\%$.

We have added footnotes to Tables 4 and 5 clarifying the interpretation of coefficients presented.

"Please provide details on how you adjust the IV to account for the non-random assignment in Section 4.6. This clarification is critical because the estimates may be biased if this issue is not properly addressed."

Response: We thank the reviewer for requesting this critical clarification about how we handle the non-random assignment at our second partner hospital (MGH). The reviewer is correct that this methodological detail is essential for interpreting our validation results.

We have expanded Section 4.6 to clarify our approach. The revised text now states:

> "To assess the generalizability of our findings beyond the Mayo Clinic ED, we replicated our analysis using data from the MGH ED, one of the busiest emergency departments in the United States. The MGH dataset comprises 129,489 patient encounters from November 10, 2021, through December 10, 2022. This extensive dataset provides a robust sample to validate the external applicability of our results.

> Unlike the Mayo Clinic ED, where patients are randomly assigned to physicians upon arrival through a rotational system, the MGH ED employs a different patient assignment mechanism. At MGH, patients are triaged into different care areas (e.g., urgent care, fast track, observation) based on acuity and presenting complaints, then assigned to physicians based on availability within those areas rather than through random rotation. To address this non-random assignment and potential selection bias, we adjust our instrumental variable strategy to account for these differences by including additional covariates for care area assignment, acuity level, and presenting complaints in both stages of our 2SLS and instrument construction, thereby accounting for the sorting of patients into different ED zones. While this approach cannot guarantee the same level of causal identification as making use of Mayo Clinic's randomized system, it provides a more robust comparison of the effects of batching on patient outcomes across different ED settings

> After adjusting for institutional differences and using the same exclusion criteria we used with Mayo, we find strong evidence that our key findings generalize to the MGH setting. The 2SLS results in Table 7 suggest that batching leads to a 44.3% increase in length of stay and approximately 1.8 additional imaging tests per patient."

We have also added a footnote explicitly stating: "The MGH estimates should be interpreted as demonstrating external validity rather than providing equally strong causal identification as the Mayo Clinic's results."

"Do the authors observe any instances where the attending physician begins the diagnostic process with a single test, followed by a batch of tests? If so, how do they account for this mixed strategy in their analysis?"

Response: Yes, we observe mixed strategies (single test followed by batch) in 189 encounters, representing 1.91% of multi-test encounters.

We classify these as non-batched ordering based on clinical guidance from our physician coauthors at both sites. Once a physician orders an initial test, they have begun sequential information gathering— even if subsequent tests are ordered simultaneously. True batching requires all tests to be ordered simultaneously without any interim diagnostic process.

71

To verify this classification does not impact our results, we re-estimated all models treating a single test followed by a batch in our definition of batching (Appendix D). The results are identical; the batching rate remains the same, and all coefficients are unchanged, confirming that our classification is appropriate.

We have updated Section 3.2 to clarify:

> "We define "batching" in line with standard emergency medicine practices and focus on batches that include two or more different imaging modalities ordered within a 5-minute window at the start of a patient encounter (Su et al. (2025) Jameson et al. (2024)). We focus on early batching (within 5 minutes) because this represents the moment of maximum diagnostic uncertainty when physicians must decide their testing strategy before clinical information unfolds. Physicians cannot know ex-ante which patients will ultimately require multiple tests, making early batching a discretionary choice based on practice style rather than clinical necessity. Each imaging modality, such as X-ray, contrast CT scan, non-contrast CT, and ultrasound, is considered a separate and distinct test for our study. In particular, we focus on batching instances where the physician orders different imaging tests because such tests cannot be done in a single scanning session (due to differences in equipment and setting). Encounters where a single test precedes subsequent batched tests (1.91% of multi-test cases) are classified as standard care in our primary analysis, as the physician has initiated sequential information gathering before placing additional orders. Sensitivity analyses conducted around this time window, batch size threshold, and the timing of the batch show that our results are robust to variations in these values."

References

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ. Accessed 9 Oct. 2025.

Berlin, L. (2011). The incidentaloma: a medicolegal dilemma. *Radiologic Clinics of North America*, 49(2):245–255.

Canellas, M. M., Jewell, M., Edwards, J. L., Olivier, D., Jun-O'Connell, A. H., and Reznek, M. A. (2024). Measurement of cost of boarding in the emergency department using time-driven activity-based costing. *Annals of Emergency Medicine*, 84(4):410–419.

Coen, M., Sader, J., Junod-Perron, N., Audétat, M.-C., and Nendaz, M. (2022). Clinical reasoning in dire times: Analysis of cognitive biases in clinical cases during the covid-19 pandemic. *Internal and Emergency Medicine*, 17(4):979–988.

Dabus, G., Hirsch, J. A., Booker, M. T., and Silva, E. (2025). Practice expense and its impact on radiology reimbursement. *Journal of the American College of Radiology*.

Featherston, R., Downie, L. E., Vogel, A. P., and Galvin, K. L. (2020). Decision making biases in the allied health professions: A systematic scoping review. *PLoS ONE*, 15(10):e0240716.

Fowler, J. W. and Mönch, L. (2022). A survey of scheduling with parallel batch (p-batch) processing. *European Journal of Operational Research*, 298(1):1–24.

Goldsmith-Pinkham, P., Hull, P., and Kolesár, M. (2025). Leniency designs: An operator's manual.

Hausman, J. A. (1975). An instrumental variable approach to full information estimators for linear and certain nonlinear econometric models. *Econometrica*, 43(4):727–738.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6):1251–1271.

Hodgson, N. R., Saghafian, S., Mi, L., Buras, M. R., Katz, E. D., Pines, J. M., Sanchez, L., Silvers, S., Maher, S. A., and Traub, S. J. (2018). Are testers also admitters? comparing emergency physician resource utilization and admitting practices. *The American Journal of Emergency Medicine*, 36(10):1865–1869.

Jameson, J., Saghafian, S., Huckman, R., and Hodgson, N. (2024). Variation in batch ordering of imaging tests in the emergency department and the impact on care delivery. *Health Services Research*. Epub ahead of print.

Jessome, R. (2020). Improving patient flow in diagnostic imaging: a case report. *Journal of Medical Imaging and Radiation Sciences*, 51(4):678–688. Epub 2020 Sep 17. PMID: 32950432; PMCID: PMC7495148.

KC, D. S. (2013). Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management*, 16(2):168–183.

KC, D. S. and Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498.

Kuntz, L., Mennicken, R., and Scholtes, S. (2014). Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science*, 61(4):754–771.

Lam, J. H., Pickles, K., Stanaway, F. F., and Bell, K. J. L. (2020). Why clinicians overtest: development of a thematic framework. *BMC Health Services Research*, 20(1):1011.

Lumbreras, B., Donat, L., and Hernández-Aguado, I. (2010). Incidental findings in imaging diagnostic tests: a systematic review. *British Journal of Radiology*, 83(988):276–289.

Rao, V. M. and Levin, D. C. (2012). The overuse of diagnostic imaging and the choosing wisely initiative. *Annals of Internal Medicine*, 157(8):574–576.

Schreyer, K. E. and Martin, R. (2017). The economics of an admissions holding unit. *Western Journal of Emergency Medicine*, 18(4):553–558.

Skaugset, L. M., Farrell, S., Carney, M., Wolff, M., Santen, S. A., Perry, M., and Cico, S. J. (2016). Can you multitask? evidence and limitations of task switching and multitasking in emergency medicine. *Annals of Emergency Medicine*, 68(2):189–195.

Su, H., Meng, L., Sangal, R., and Pinker, E. J. (2025). Crisis at the core: Examining the ripple effects of critical incidents on emergency department physician productivity and work style. Available at SSRN: `https://ssrn.com/abstract=5113467` or `http://dx.doi.org/10.2139/ssrn.5113467`.

Traub, S. J., Bartley, A. C., Smith, V. D., Didehban, R., Lipinski, C. A., and Saghafian, S. (2016a). Physician in triage versus rotational patient assignment. *The Journal of Emergency Medicine*, 50(5):784–790.

Traub, S. J., Saghafian, S., Bartley, A. C., Buras, M. R., Stewart, C. F., and Kruse, B. T. (2018). The durability of operational improvements with rotational patient assignment. *American Journal of Emergency Medicine*, 36(8):1367–1371. Epub 2017 Dec 20.

Traub, S. J., Stewart, C. F., Didehban, R., Bartley, A. C., Saghafian, S., Smith, V. D., Silvers, S. M., LeCheminant, R., and Lipinski, C. A. (2016b). Emergency department rotational patient assignment. *Annals of Emergency Medicine*, 67(2):206–215. Epub 2015 Oct 6. PMID: 26452721.