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 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 revision decision.

We have carefully revised the paper to address all the comments, and the paper has benefited greatly from the suggested revisions. We first briefly summarize the main changes and then provide individual responses.

1. We have addressed the remaining methdological suggestions by the reviewers, including...
2. Throughout the paper...

We hope the review team finds our various efforts in extending our results, and in carefully addressing the comments, satisfactory in this round.

II. 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: [Your response here]

AE Wrote (Comment – Theory needs clarification):

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

- *Is batching truly a disrectionary 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 appreciate the Associate Editor's request for theoretical clarification, which has led to substantial improvements in how we frame and interpret our study. We have made extensive revisions throughout the manuscript to address each of these important questions.

1. Discretionary versus clinically-driven batching: Our study isolates *discretionary* batching decisions—those driven by physician practice style rather than clinical necessity—which is the precise margin where ED management can intervene. We establish this through our unique empirical setting and identification strategy:

First, Mayo Clinic's rotational patient assignment mechanism provides quasi-random assignment of patients to physicians. As documented by Traub et al. (2016a,b, 2018), patients are assigned to physicians through a computerized algorithm that is based solely on arrival time, without consideration of patient characteristics, complaint severity, or physician workload. This random assignment is crucial: it ensures that differences in batching rates across physicians reflect physician preferences rather than systematic differences in patient mix or unobserved clinical factors. Figure 2 provides empirical verification—while patient characteristics strongly predict batching decisions (left panel), they do not predict assignment to high versus low batch-tendency physicians (right panel), indicating that conditional on controls assignment to phsyician is as good as random.

Second, our LATE identifies effects specifically for "compliers"—the fraction of patients whose batching decision changes depending on physician assignment. These patients, by definition, lack clear clinical indicators mandating either batching or sequential testing. They represent the discretionary margin where physician preference determines the testing strategy. The existence of these compliers demonstrates that discretionary batching exists and is substantial enough to matter for ED operations.

Third, we focus on *early* batching (batching that occurs during the first test order) to capture decisions made before clinical information unfolds. At this moment of maximum uncertainty, some physicians opt for comprehensive imaging, while others maintain diagnostic flexibility. Late batching in response to test results occurs in only 1.91% of multi-test encounters, confirming our focus on initial discretionary decisions rather than adaptive clinical responses. Importantly, we do not claim all batching is discretionary. Some patients clearly require multiple urgent tests (the "always-takers"), while others need only a single test (the "never-takers"). Our contribution is identifying and quantifying the effects of discretionary batching for patients at the margin—precisely where ED protocols, decision support systems, and training interventions can influence practice without constraining necessary clinical judgment.

2. Preemptive ordering and physician workflow management: The existence of compliers in our analysis provides definitive evidence that discretionary batching occurs. If batching were driven purely by clinical necessity, we would observe minimal physician-induced variation—all physicians would batch for the same patients based on clinical presentation. Instead, our complier analysis (detailed in Appendix B) reveals that 13% of patients receive batched imaging when assigned to high-tendency physicians but sequential testing when assigned to low-tendency physicians. These patients, by definition, lack clear clinical indicators mandating either approach.

The operational consequences of discretionary batching are unambiguous in our data. For these marginal patients, batching generates 1.4 additional imaging tests performed without improving diagnostic outcomes (72-hour return rates: -0.012, SE=0.018, p=0.50). While we cannot determine whether specific tests are clinically unnecessary, the pattern is clear: at the margin where physicians have discretion, batching increases resource utilization without measurable quality benefits. Table 6 provides additional evidence that batching serves workflow management purposes. The significant decline in batching probability as shifts progress (-0.004 per hour, $p<0.05$) is inconsistent with clinical drivers, as patient acuity does not systematically vary by physician shift hour under Mayo Clinic's random assignment mechanism. This temporal pattern, combined with the attenuation of batching under major overcapacity conditions (Table 5), suggests physicians modulate their batching behavior based on operational context rather than purely clinical factors.

The 44.5 percentage point increase in admission probability warrants careful interpretation. Based on extensive consultation with our emergency medicine physician co-authors at both study sites, this pattern aligns with well-documented diagnostic cascades where additional imaging identifies incidental findings that, while not addressing the presenting complaint, trigger defensive admission decisions.

The medical literature extensively documents this phenomenon (Ganguli et al., 2019; Hoffman & Cooper, 2012).

Our contribution is precise: we quantify the operational impact of discretionary batching for patients at the margin where physician preference, rather than clinical protocols, determines the testing strategy. The reduced form estimates provide the causal effect of assignment to physicians with different batching tendencies. Even if some portion of this effect operates through channels beyond pure batching (potential exclusion restriction violations), the magnitude and consistency of effects across both study sites demonstrate that physician-induced variation in testing strategies creates substantial operational inefficiencies. These findings have direct policy relevance for the 13% of patients whose care pathway depends on physician assignment—the precise population where ED protocols and training interventions can improve efficiency without constraining clinical autonomy.

3. Diagnostic uncertainty and contextual factors: We carefully distinguish between clinically-appropriate responses to uncertainty and discretionary practice patterns. Our analysis reveals that discretionary batching—the type influenced by physician preference—varies with contextual factors in predictable ways: Table 5 demonstrates that under major overcapacity, the frequency of discretionary batching drops (9.9% vs. 15.8% during normal operations) and its negative effects attenuate. The imaging increase falls from 1.36 to 1.16 tests, suggesting physicians become more selective when resources are constrained, eliminating discretionary batching while preserving clinically necessary batching.

Our heterogeneity analysis (Appendix C) shows that while batching rates vary across complaint types (31% for trauma vs. 8% for extremity complaints), the operational inefficiencies of discretionary batching persist across categories. This confirms that physician-induced batching—the variation our IV identifies—represents a suboptimal practice style rather than a clinically-tailored response to case complexity.

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.

Based on the reviewers' feedback and my own assessment, I believe there is a potential path toward eventual publication albeit at a very high risk. In addition to responding to the referee's concerns, the revision needs to address the following key points:"

Response: We appreciate the Associate Editor's candid assessment and understand why the 130%

increase in ED length of stay appears striking at first glance. This skepticism motivated extensive robustness checks and clarification in the manuscript, which ultimately strengthened confidence in our findings. We have substantially revised both the theoretical foundation and empirical analysis to address these concerns.

Addressing the magnitude of the effect: While the 130% increase appears substantial, it becomes more interpretable when broken down into its operational components. Our mediation analysis (Section 4.3) reveals that discretionary batching leads to 1.4 additional imaging tests for marginal patients. Given average imaging turnaround times in our data (90 minutes for non-contrast CT, 142 minutes for contrast CT, 165 minutes for ultrasound), these additional tests mechanically extend stays by 2-3 hours. Combined with the cognitive burden of processing multiple simultaneous results and the increased likelihood of admission (a 44.5 percentage point increase), the cumulative effect reaches the observed magnitude.

Importantly, this effect represents the Local Average Treatment Effect for the 13% of patients who are compliers—those whose testing strategy changes based on physician assignment. These marginal patients often lack clear clinical protocols that dictate their imaging pathway, making them particularly susceptible to extended stays when subjected to comprehensive upfront testing rather than sequential evaluation.

Establishing exogenous assignment: We have comprehensively addressed concerns about assignment exogeneity through multiple approaches: First, the Mayo Clinic utilizes a computerized rotational assignment algorithm that assigns patients to physicians solely based on arrival order, without considering patient characteristics, complaint type, or physician workload. This system, documented in peer-reviewed publications (Traub et al. (2016a,b, 2018)), removes discretion from the matching process. Section 3.1 now provides extensive detail about this mechanism, including how physicians receive four consecutive patients at shift start before entering rotation, and how the system caps physicians at 18 patients with no new assignments in their final 120 minutes.

Second, Figure 2 provides empirical verification of randomization. While patient characteristics strongly predict batching decisions (left panel), these same characteristics show no association with assignment to high versus low batch-tendency physicians (right panel). All coefficients are near zero with confidence intervals crossing zero, confirming that, conditional on shift fixed effects, assignment is effectively random.

Third, regarding the concern that some physicians might only work specific shifts, our fixed effects for day of week and time of day explicitly account for physician scheduling patterns. The residualized batch tendency measure captures physician-specific variation after removing these temporal patterns. Even if certain physicians systematically work different shifts, our approach ensures we identify variation within shift types.

Robustness of empirical strategy: Following all three reviewers' suggestions, we have implemented comprehensive robustness checks:

- Alternative instrument construction: Per Reviewer 3's suggestion, we constructed an alternative instrument using physician fixed effects directly (with leave-one-out correction). This approach yields virtually identical results (log LOS: 0.835 vs. 0.837), with a first-stage F-statistic of 187.
- Controls for correlated physician tendencies: We constructed measures of physician admission tendency and laboratory test ordering using the same methodology to control for correlated physician behaviors, which could be driving results (as mentioned by Reviewer 1). Including these controls yields minimal attenuation (log LOS: 0.837 $\rightarrow$ 0.821), indicating our instrument captures batching-specific variation rather than general physician characteristics.
- External validation: Replication at Massachusetts General Hospital, despite its different assignment mechanism and staffing model, yields remarkably similar effects (44.3% increase in LOS, 1.8 additional tests). The consistency across institutions with different operational structures suggests our findings reflect fundamental inefficiencies of discretionary batching rather than institution-specific artifacts.
- Placebo and falsification tests: We find no effects of batch tendency on outcomes for patients with complaints where batching is clinically inappropriate (Appendix D). Additionally, batch tendency does not predict laboratory test ordering after controlling for patient characteristics, confirming specificity to imaging decisions.

Improvements to theoretical development: We have added Section 2.3, which includes formal hypothesis development, linking cognitive load theory and defensive medicine literature to predict batching behaviors. This theoretical framework clarifies that we study discretionary batching at the margin where physician style, rather than clinical necessity, determines testing strategy—precisely where ED interventions can improve efficiency.

Strengthened empirical analysis: Beyond the robustness checks above, we have:

- Expanded sample selection documentation with a CONSORT diagram (Appendix Figure A2)
- Added heterogeneity analysis by complaint complexity (Appendix C)
- Included alternative outcome specifications suggested by R3 (treatment time, any 72-hour return)
- Provided detailed complier characterization (Appendix B)

The convergence of evidence across multiple specifications, institutions, and methodological approaches provides the high degree of confidence the Associate Editor rightly demands. While the magnitude remains striking, it is empirically robust and operationally explicable, reflecting genuine inefficiencies when physicians exercise discretion to batch tests for patients lacking clear clinical protocols.

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: [Your response here]

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: [Your response here]

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: [Your response here]

AE Wrote (Comment – Additional suggestions):

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

1. *Reassess the definition of the treatment variable (e.g. Ref 3).*
2. *Examine the independence of test orders from prior tests and workload conditions (e.g. Ref 2).*
3. *Include a section discussing sample representativeness.*
4. *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.*
5. *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: [Your response here]

III. 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 sincerely appreciate the reviewer's recognition of the importance and clarity of our research question. We share your view that emergency departments represent uniquely high-stakes environments where operational improvements can have life-saving implications. Your framing of our work as contributing to the understanding of how "work behavior/preferences" affect operational outcomes perfectly captures what we hope to achieve—bridging the gap between individual physician decision-making and system-level performance. We are grateful for your thorough engagement with our manuscript and have carefully addressed your methodological concerns to ensure our findings can meaningfully contribute to improving ED operations and patient care.

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 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 critique, which has led to substantial improvements in how we frame our research question 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.

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 requiring diagnostic imaging, not between testing versus 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 affects subsequent testing decisions. Physicians cannot know ex-ante which patients will ultimately need multiple tests—they make batching decisions under diagnostic uncertainty. Conditioning on eventual test count would compare fundamentally different populations after the treatment has already operated. 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.

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:

> "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 or-

> dering 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 moment 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. We have added 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, making early batching 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 both testing intensity and processing time."

Validating instrument specificity: The reviewer's concern that our instrument might capture "degree of cautiousness" or "affinity for comprehensive testing" is important and exactly right to scrutinize. Following standard practice in the judges design literature Dobbie et al. (2018); Bhuller et al. (2020), we validate our instrument through three approaches: (1) balance tests showing patient characteristics do not predict instrument values, (2) placebo tests showing no effects where batching is clinically inappropriate, and (3) differential effects tests examining whether the instrument predicts outcomes beyond imaging decisions.

The critical evidence for instrument specificity comes from examining reduced form effects on laboratory versus imaging tests. Laboratory tests can be drawn simultaneously (so "batching" labs has no operational meaning), providing a natural specificity test. If our instrument captures general diagnostic aggressiveness rather than imaging-specific timing, we should observe similar effects across both diagnostic modalities.

[Insert Reduced Form Lab Table]

The results reveal striking specificity. With patient and time controls, batch tendency shows no relationship with laboratory test ordering (coefficient: 0.121, p=0.60). This stands in sharp contrast

to imaging effects, which remain large and highly significant. This differential pattern—absence of lab effects alongside persistent imaging effects—strongly suggests our instrument captures imaging-specific timing decisions rather than general diagnostic intensity.

As an additional robustness check beyond standard practice in this literature, we constructed control variables for potentially correlated physician behaviors: admission tendency and lab test tendency, using identical residualized leave-one-out methodology. While most judges design papers rely solely on random assignment and balance tests Dobbie et al. (2018); Aizer and Doyle (2015), we examine sensitivity to these additional controls.

[Insert First-Stage Table with progressive controls] [Insert 2SLS Table with progressive controls]

Our instrument remains strong even with these demanding controls (F-statistics: 238.8→87.4, all above conventional thresholds). Operational impacts remain economically large in our most conservative specification: log LOS effect of 0.833 ($p=0.081$) implies a 130% increase, while imaging tests increase by 0.977 ($p<0.001$), a 73% increase over baseline.

We interpret this evidence collectively as validating instrument specificity. The base specification without additional tendency controls represents our primary estimate—it follows standard practice in the judges design literature and estimates the full effect of the "batching practice style." The specifications with additional controls provide conservative lower bounds, demonstrating that even when we remove correlated variation, imaging-specific effects persist while lab effects disappear. This asymmetry, combined with successful balance tests and null placebo results, supports our interpretation that physician-induced variation in imaging test timing creates substantial operational consequences.

In response to these concerns, we have made the following changes:

1. *Added comprehensive robustness tables (Appendix D):* We present first-stage results with progressive controls (Table D.X), reduced form effects on laboratory versus imaging outcomes (Table D.Y), and full 2SLS specifications with additional residualized tendency controls (Table D.Z).

2. *Enhanced identification discussion (Section 3.3):* We now explicitly acknowledge that physician tendency instruments may capture bundles of correlated practices, following standard practice in the judges design literature Dobbie et al. (2018). We explain our validation approach through balance tests, placebo tests, and differential effects tests, with particular emphasis on the specificity to imaging versus laboratory outcomes.

3. *Clarified interpretation (Section 4.1):* We distinguish between our primary specification (following standard practice) and conservative specifications with additional controls, noting that both support our core conclusion that test timing drives operational consequences.

4. *Expanded limitations discussion (Section 4.8):* We acknowledge that while we cannot definitively rule out all exclusion restriction violations, the convergence of evidence—successful bal-

ance tests, null placebo results, imaging-specific effects, and persistence across conservative specifications—supports our interpretation.

These revisions strengthen our identification strategy while maintaining transparency about the assumptions underlying our causal claims.

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 appreciate the reviewer reiterating this fundamental concern about exclusion restriction violations. This is indeed the same issue raised in Major Concern 1—whether our instrument captures batching-specific behavior or broader physician practice styles. As detailed in our response to Major Concern 1, we address this through comprehensive robustness checks that directly test the specific mechanisms the reviewer identifies here.

The reviewer provides two concrete examples of potential violations: (1) cautious physicians ordering more lab tests that extend LOS, and (2) cautious physicians taking longer examinations or monitoring patients longer. Our empirical analysis directly addresses both channels:

Laboratory testing channel: We constructed a physician lab test tendency measure using identical methodology to our batch tendency instrument. As shown in our response to Major Concern 1 (Table D.Y), batch tendency shows no independent relationship with laboratory ordering after controlling for patient characteristics (coefficient: 1.073→0.121, $p = 0.60$). This provides direct evidence against the reviewer's first concern—physicians with high batch tendency do not systematically order more labs in ways that would independently extend LOS.

General cautiousness channel: We constructed a physician admission tendency measure capturing clinical conservatism. Including this control yields minimal attenuation of our imaging effects while substantially reducing admission effects (0.488→0.243), suggesting some correlation with general practice style. However, imaging effects remain large and significant, and critically, the lab effects disappear entirely while imaging effects persist. This differential pattern—elimination of lab effects

alongside persistent imaging effects—is difficult to reconcile with our instrument simply capturing "degree of cautiousness."

The reviewer correctly notes we cannot definitively rule out all exclusion restriction violations. We acknowledge this limitation explicitly in Section 4.8. However, the convergence of evidence provides strong support for our interpretation:

- Balance tests confirm patient characteristics don't predict instrument values (Figure 2)
- Placebo tests show no effects for complaints where batching is inappropriate (Table D.1)
- Specificity tests show effects limited to imaging, not labs or other diagnostic inputs
- Conservative specifications with extensive controls yield persistent, economically meaningful effects

Following standard practice in the judges design literature Dobbie et al. (2018); Bhuller et al. (2020), we rely primarily on random assignment and balance tests. The additional robustness checks we provide go beyond what is typical in this literature, demonstrating our instrument's specificity even under demanding conditions. While we cannot claim perfect identification, the weight of evidence supports our core conclusion that physician-induced variation in imaging test timing creates substantial operational inefficiencies.

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 indeed crucial to our identification strategy. We have substantially expanded our description of this system in Section 3.1.

> "Our study uses data from two large U.S. emergency departments (EDs): the Mayo Clinic of Arizona and Massachusetts General Hospital (MGH). The MGH dataset, which includes 129,489 patient encounters from November 10, 2021, through December 10, 2022, provides a robust sample for validating the generalizability of our findings. However, our primary analysis focuses on the Mayo Clinic data due to its unique feature of random patient-physician assignment, which allows for stronger causal inference. More specifically, the

> Mayo Clinic ED employs a sophisticated computerized rotational patient assignment algorithm that addresses many of the empirical challenges previously identified (Traub et al. (2016a,b, 2018)). The system automatically assigns patients to physicians 60 seconds after registration through the electronic health record system, following a strict rotational protocol. 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. 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. The rotation order follows a predetermined schedule that varies across shifts to ensure fairness over time.
>
> 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. We establish that, conditional on arrival time, patient-physician matching is effectively random—a critical requirement for our identification strategy that distinguishes our study from observational analyses where endogenous matching could confound the effects of physician discretion."

We have also added the Traub et al. (2018) citation, which documents the durability and effectiveness of this rotational assignment system at our study site.

Regarding empirical verification of randomization, Figure 2 provides precisely the test suggested by the reviewer. The right panel shows that after conditioning on time fixed effects (to account for the mechanical rotation), 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.

We have added the following text to Section 3.3 to emphasize 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 assignment mechanism and establishes batch tendency as an exogenous source of variation for identifying causal effects."

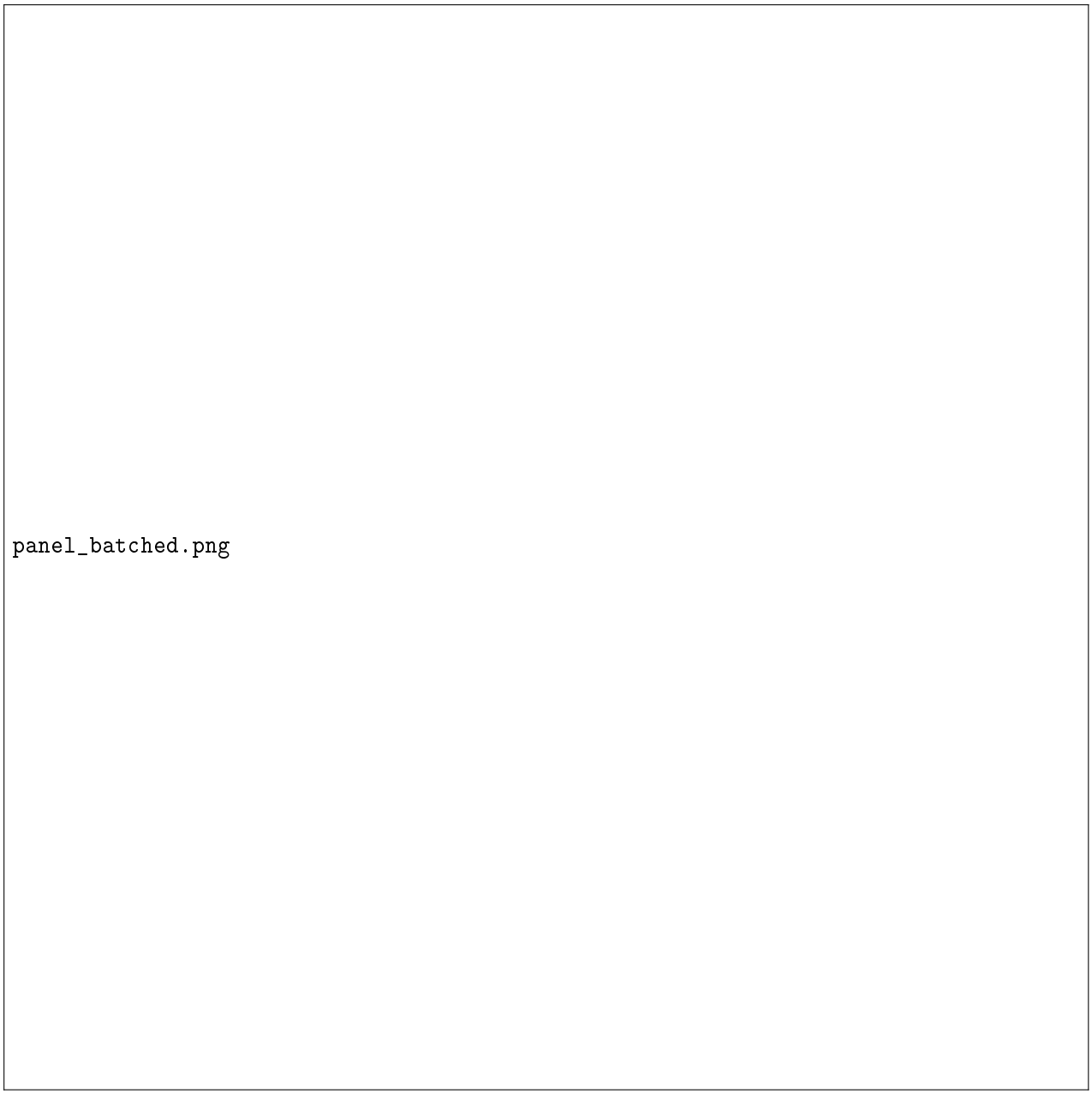
Together, the institutional details and empirical evidence confirm that the Mayo Clinic's rotational assignment mechanism achieves the quasi-randomization necessary for our identification strategy.

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 hope the extensive revisions addressing the reviewer's concerns about variable definition, empirical strategy, and causal interpretation have substantially improved the manuscript.

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. Residualization fixed effects include hospital-year-month, hospital-day of week-time of day. Robust standard errors are clustered at the physician level.

IV. 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 the conventional wisdom about batch ordering efficiency is important given the operational pressures EDs face. The reviewer's acknowledgment that our quasi-randomized design uniquely isolates batching effects validates our empirical approach. Following the reviewer's and other referees' suggestions, we have further clarified how our findings differ from prior work on physician practice variation by specifically focusing on the timing and sequencing of test orders rather than just testing intensity.

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. We appreciate this opportunity to clarify our approach and have made substantial revisions to address this concern.

The reviewer raises a fundamental question: are the harms from batching itself or from differences between patients needing multiple early tests? This gets to the heart of causal inference in our setting. Physicians cannot know ex-ante which patients will ultimately need multiple tests—they make batching decisions under uncertainty. 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.

We have made the following changes to address your concerns which we believe have made the manuscript stronger:

1. **Reframed our comparison throughout the manuscript**. We now consistently describe our comparison as "batch ordering versus standard practice" rather than "batching versus sequencing." For example, in the abstract and Section 1.2, we revised: "the marginal batched patient experiences a 130% increase in total ED LOS compared to standard practice" (previously "compared to patients who have their tests ordered sequentially").

2. **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 early batching 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."

3. **Added discussion of the information problem**. In Section 3.2.1, we now explain:

   > "Our treatment variable, $Batched_{i,t}$, is an indicator taking the value of 1 if patient i has their tests batch ordered during their ED encounter, which took place on date t, and 0 otherwise. Batching occurs when a physician simultaneously orders a comprehensive set of diagnostic tests, typically covering a broad range of potential diagnoses. This contrasts with standard practice, where a single test is ordered and subsequent tests are ordered in sequence on an as-needed basis.
   >
   > We define "batching" in line with standard emergency medicine practices and focus on batches that include two or more different imaging modalities where the time between orders is within 5 minutes, occurring as the first imaging tests ordered for the patient encounter (Su et al. 2025, Jameson et al. (2024)). 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. Each imaging modality, such as X-ray, contrast CT scan, non-contrast CT, MRI, 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 sequential in our primary analysis, as the physician has initiated sequential information gathering before placing additional orders. Sensitivity analyses around this time window, batch size threshold, and when the batch occurs show that our results are robust to variation in these values."

4. **Provided exploratory analysis with appropriate caveats**. While maintaining that conditioning on eventual test count introduces bias, we present this analysis in Appendix Table A.X (included below) with appropriate caveats about post-treatment conditioning. When we restrict to patients receiving 2+ tests, the 2SLS estimates change substantially—log LOS shows a small positive effect (0.178, not significant) and log time to disposition shows 0.423 (not significant). The number of tests actually shows a small negative effect (-0.102). However, we strongly caution that these estimates cannot be interpreted causally as they condition on a post-treatment variable. The dramatic change from our main results (0.837 for log LOS) likely reflects selection bias: batching causally increases the probability of receiving multiple tests, so conditioning on 2+ tests compares fundamentally different populations of compliers. This illustrates why preserving the full causal pathway is essential for valid inference."

   This exploratory analysis, while limited by post-treatment conditioning, actually strengthens our methodological approach. The dramatic attenuation of effects illustrates precisely why our main specification—which preserves the full causal pathway—provides the relevant parameter for policy.

5. **Connected to delayed imaging concerns**. The reviewer correctly notes we don't evaluate delayed imaging. We now acknowledge in Section 4.1:

   "Our estimates capture the full effect of early batching including its impact on test volume. The 1.4 additional tests under batching represent part of the causal pathway—forfeiting the information value of sequential testing leads to tests that would have been avoided under standard practice."

   Regarding potential harms from delayed sequential testing, we find no evidence of quality impacts: 72-hour return rates are identical between groups (-0.012, p=0.50), suggesting sequential approaches don't cause harmful diagnostic delays.

6. **Clarified why "everything else" is the right comparison**. We added 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 both testing intensity and processing time."

Appendix Table A.X: Exploratory Analysis: Patients Receiving 2+ Imaging Tests

	Main Analysis (N=11,404)		2+ Tests Only (N=4,375)	
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)
Log time to disposition	0.108***	0.804*	0.029*	0.423
	(0.014)	(0.349)	(0.015)	(0.372)
Log LOS	0.125***	0.837*	−0.069***	0.178
	(0.013)	(0.299)	(0.013)	(0.245)
Number of tests	0.830***	1.414***	0.043**	-0.102
	(0.016)	(0.220)	(0.015)	(0.108)
72hr return w/ admit	-0.000	-0.012	0.003	-0.019
	(0.002)	(0.018)	(0.003)	(0.025)
Mean for standard care patients:				
Log time to disposition	5.236		5.237	
Log LOS	5.487		5.669	
Number of tests	1.336		2.172	
72hr return w/ admit	0.012		0.008	

Notes: Columns 1-2 reproduce our main results from Table 4 for comparison. Columns 3-4 show results when restricting to patients who received 2 or more imaging tests. The dramatic differences between main and restricted samples illustrate the selection bias introduced by conditioning on post-treatment outcomes. Since batching causally increases the likelihood of receiving multiple tests (see Table 4, Panel A), the restricted sample compares different populations of compliers. The OLS sign reversal for log LOS (positive in main analysis, negative in restricted sample) particularly highlights this selection issue. All specifications include time fixed effects and baseline controls. Standard errors clustered at physician level.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

The reviewer's observation about mixed clinical pathways in our counterfactual is astute. However, this heterogeneity is precisely 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 provides exactly this parameter.

We believe these revisions substantially improve the manuscript's clarity while maintaining scientific rigor. The comparison we identify—early batching versus standard practice for marginal patients—is both what we can credibly estimate given the information structure and what policymakers 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. You are absolutely correct that explicitly articulating our theoretical predictions strengthens the paper's contribution and clarifies the mechanisms we expect to observe. This recommendation has significantly improved our manuscript by making our theoretical framework more transparent and our empirical tests more clearly motivated.

Following your guidance, we have added a new Section 2.3 "Hypothesis Development" that builds naturally from our literature review to develop four formal hypotheses about how batch ordering affects ED operations. We present these additions in detail below:

> ## "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 workload 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 due to the loss of information value from initial test results.*

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 through increased diagnostic intensity and the influence of incidental findings on clinical decision-making.*

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, as physicians become more selective in their batching decisions.*

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."

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. We have been careful throughout the manuscript to avoid characterizing these additional tests as "waste" or "overuse," as our LATE estimate identifies the effect of batching for marginal patients whose testing decisions vary by physician preference—these tests may still have clinical value.

However, based on extensive discussions with our physician coauthors at both study sites, we note several reasons why frontloading diagnostics in the ED—even if the tests would eventually be ordered as an inpatient—represents an operational inefficiency:

First, ED imaging resources are typically more constrained than inpatient resources, with ED scanners serving both emergency and admitted patients. Frontloading increases congestion in these shared resources, delaying care for other ED patients. Second, ED lengths of stay directly impact ED crowding, which has well-documented negative effects on patient outcomes and left-without-being-seen rates. Third, the ED environment is optimized for rapid stabilization and disposition decisions, not comprehensive diagnostic workups. Inpatient teams often prefer to direct their own diagnostic approach based on evolving clinical pictures.

While our finding of increased admissions could partly reflect frontloading of diagnostic workups that reveal conditions requiring admission, the magnitude of the effect and lack of improvement in 72-hour

returns suggest the additional testing may not be optimally targeted. However, we acknowledge in the revised limitations section that we cannot fully disentangle frontloading effects from potentially unnecessary testing. Future research with access to inpatient imaging data could better address this distinction.

We have added text to the Discussion acknowledging both possibilities and noting this as a limitation of our study.

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: [Your response here]

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. This comment has led us to substantially clarify throughout the manuscript that our study specifically examines *discretionary early batching decisions*—those made at the beginning of encounters under diagnostic uncertainty. These are precisely the decisions where physician practice style dominates clinical necessity, representing the margin where ED protocols, training interventions, and decision support systems can meaningfully influence practice. Later adaptive batching in response to case complexity represents clinical judgment that managerial policies cannot and should not constrain. We have revised multiple sections to make this critical distinction explicit, as detailed below.

Clarifications made throughout the manuscript:

Following the reviewer's observation, we have revised our framing to emphasize that our parameter of

interest is the effect of *early discretionary batching* versus standard practice. Key revisions include:

1. **Abstract:** Now specifies "early batch ordering" and "discretionary batching decisions made at encounter initiation"
2. **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"
3. **Section 3.2.1:** Extensively revised to state:

> "We define "batching" in line with standard emergency medicine practices and focus on batches that include two or more different imaging modalities where the time between orders is within 5 minutes, occurring as the first imaging tests ordered for the patient encounter (Su et al. 2025, Jameson et al. 2024). 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. Each imaging modality, such as X-ray, contrast CT scan, non-contrast CT, MRI, 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 sequential in our primary analysis, as the physician has initiated sequential information gathering before placing additional orders. Sensitivity analyses around this time window, batch size threshold, and when the batch occurs show that our results are robust to variation in these values."

Why early batching is the relevant parameter:

The reviewer correctly notes that physicians might batch tests later to "wrap things up." This adaptive behavior is fundamentally different from the discretionary decision to order comprehensive imaging upfront. Our instrumental variable (physician batch tendency) specifically captures variation in early ordering propensity—identifying physicians who habitually commit to comprehensive imaging before information unfolds versus those who preserve diagnostic flexibility.

Empirically, we verify that "late-batching" or cases where a single test is followed by a batch is rare and only occurs in 189 encounters, representing 1.91% of multi-test encounters. 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 Table A4). The results are identical; the batching rate remains the same, and all coefficients are unchanged, confirming that our classification is appropriate.

LOS as a post-treatment outcome:

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.

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 not simultaneity biases but rather the mediating mechanisms through which early batching decisions affect patient flow.

Our mediation analysis (Section 4.3) formally decomposes these pathways, revealing that increased test volume accounts for the majority of the LOS effect (indirect effect $= 0.207, p < 0.001$). This confirms that the "mechanical" component the reviewer identifies—waiting for additional tests—is indeed a key mechanism, but one that results from the discretionary early batching decision rather than confounding it.

Furthermore, to 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. The persistence of large effects (123% increase) demonstrates that early batching affects not just mechanical waiting but also clinical processing efficiency. This addresses the reviewer's concern about entangled causality: even when excluding post-disposition time (where no testing occurs), the early batching decision substantially delays 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 measure counts imaging tests actually performed, not merely ordered. This critical distinction strengthens our causal interpretation: the 1.4 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 coauthor at Mayo Clinic, test cancellations after batch 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 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.

Manuscript changes made throughout:

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

Section 3.2.2 (Dependent Variables): Revised to state:

> "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 1.4 more distinct imaging tests (completed studies with documented results), representing a 105% increase from the mean for sequentially tested 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 indicates that cancellations are insufficiently common to 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, 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.

The reviewer correctly notes that lab results often arrive before imaging results (typically 30-60 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, if physicians were to cancel certain imaging based on lab results, our estimate of the impact of discretionary batching on performed imaging tests would be a strict lower bound of the effect on total imaging orders. We find 1.4 additional performed imaging tests per marginally batched patient, suggesting that imaging pathways, once initiated through discretionary batching, proceed to completion often enough to lead to increased testing.

To empirically test whether lab results moderate our batching effects, we examined specifications progressively adding controls. Following concerns from Reviewers 1 and 3 about potential exclusion restriction violations, we first added physician characteristics to ensure our instrument captures batching behavior specifically rather than other physician attributes. We then added lab ordering to test for lab-imaging interactions:

Robustness of Batching Effects to Additional Controls

	Baseline Controls (1)	+ Physician Characteristics (2)	+ Lab Ordered (3)
Panel A. Primary Outcomes			
Log time to disposition	0.804*	0.822*	0.771·
	(0.349)	(0.374)	(0.395)
Log LOS	0.837*	0.854*	0.673*
	(0.299)	(0.328)	(0.323)
Number of imaging tests	1.414***	1.373***	1.263***
	(0.220)	(0.207)	(0.197)
72hr return with admission	-0.012	-0.012	-0.015
	(0.018)	(0.020)	(0.023)
Controls included:			
Time FE	Yes	Yes	Yes
Patient characteristics	Yes	Yes	Yes
Physician characteristics	No	Yes	Yes
Lab ordered indicator	No	No	Yes
Observations	11,404	11,404	11,404

Notes: All columns show 2SLS estimates. Column 1 reproduces our main specification from Table 4. Column 2 adds physician characteristics (experience, gender) to address exclusion restriction concerns. Column 3 further adds an indicator for whether labs were ordered. While some attenuation occurs with lab controls, all specifications show large, economically meaningful effects.
$^{\cdot}p < 0.1$, $^{*}p < 0.05$, $^{***}p < 0.001$.

We observe modest attenuation when controlling for laboratory ordering, with the log time to disposition effect decreasing from 0.822 to 0.771 and losing some statistical precision. This attenuation suggests that physicians who batch imaging tests also tend to order laboratory tests, reflecting a broader pattern of comprehensive diagnostic ordering. When we control for this general diagnostic intensity through the lab indicator, we isolate the effect of imaging batching more precisely. Importantly, even after accounting for concurrent lab ordering, all specifications show large, economically meaningful effects on processing times and imaging volume. The persistence of these substantial effects demonstrates that discretionary imaging batching creates operational burdens beyond what can be explained by general diagnostic thoroughness, confirming that batched imaging pathways proceed

independently once initiated due to operational momentum and workflow barriers to modification.

Manuscript changes:

Section 4.8 (Robustness): Added Table A.Z with the above specifications and discussion:

> "To address concerns about lab-imaging interactions, we examined whether controlling for laboratory ordering affects our estimates of discretionary batching (Table A.Z). The modest attenuation when adding lab controls (e.g., log time to disposition: 0.822 to 0.771) indicates that some portion of our estimated batching effect reflects correlation with general diagnostic intensity—physicians who batch imaging also tend to order more laboratory tests. By controlling for lab ordering, we isolate the imaging-specific component of the batching effect. Importantly, even after decoupling this correlation, we find large, persistent impacts in processing times and imaging volume. This demonstrates that discretionary imaging batching creates substantial operational burdens distinct from broader diagnostic thoroughness. Clinical evidence supports this independence: normal lab results rarely lead to imaging cancellation, as modifications require calling radiology departments—a significant workflow barrier. The persistence of large effects after controlling for laboratory utilization confirms that discretionary batching commits patients to imaging pathways that proceed independently once initiated."

We appreciate the reviewer highlighting this potential concern, as addressing it demonstrates the robustness of our findings and clarifies that discretionary batching creates operational commitments that persist despite the theoretical possibility of lab-based modifications.

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 $< 5\%$ of the time). This exclusion was necessary for two reasons: First, our instrumental variables approach requires sufficient variation in batching behavior to identify effects. Complaints that are rarely batched would lead to weak instrument problems.

Importantly, including rarely-batched complaints would distort our LATE estimate. The IV estimate is essentially a weighted average of complaint-specific effects, where weights depend on the first-stage relationship. For complaints where batching never occurs, the complaint-specific effect is undefined, while for rarely-batched complaints, the effect is estimated with high variance. Including these would produce a LATE averaged over: (1) complaints where batching represents a meaningful choice with identifiable effects, and (2) complaints contributing mainly noise and of less interest to ED physicians.

This would yield a less interpretable and potentially biased estimate of the batching effect where it actually matters.

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 debated and meaningfully impact ED operations. These complaints represent approximately 25% of ED volume but account for 41% of imaging resource utilization in our data.

We agree this makes our study "more narrowly defined," but view this as ensuring our estimates are both statistically identified and operationally relevant. We have added text to the limitations section explicitly stating:

> "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.

Our replication at MGH—where mid-level providers participate under attending supervision—demonstrates that our findings generalize across different staffing models. As shown in Table 7, the effects at MGH (44.3% increase in LOS, 1.8 additional tests) are directionally similar and statistically indistinguishable from those at Mayo, suggesting the batching phenomenon transcends staffing configurations.

Moreover, Mayo's physician-only ordering strengthens our causal inference in several ways. First, it reduces potential confounding from provider-type variation (e.g., systematic differences in ordering patterns between NPs and MDs). Second, it ensures that all ordering decisions reflect the clinical judgment of fully-trained emergency physicians, minimizing noise from training-related variation. Third, the random assignment mechanism at Mayo is cleaner because patients cannot be triaged to different provider types based on acuity or complaint.

We view Mayo's setting as providing an ideal "laboratory" to isolate the pure effect of batching behavior among experienced clinicians. The consistency of results at MGH—with its more complex staffing environment—validates that these effects persist when mid-level providers are involved in

care delivery. Our physician coauthors at both sites confirm that while ordering processes differ, the fundamental trade-offs of batching versus sequential testing remain similar across settings.

We acknowledge that EDs with significant mid-level provider staffing may face additional complexities not fully captured in our analysis, such as handoff effects or supervision requirements that could interact with batching decisions. We have added text to Section 4.6, noting:

> "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 drawing attention to the comparison between columns 4 and 5 in Table 4. Upon careful examination, we find that adding baseline controls (column 5) actually demonstrates the robustness of our results rather than sensitivity to influential controls.

Specifically, the point estimates remain remarkably stable across specifications:

- Log time to disposition: 0.627 → 0.804 (both indicating large positive effects)
- Log LOS: 0.690 → 0.837 (both indicating large positive effects)
- Number of imaging tests: 1.395 → 1.414 (virtually identical)
- 72hr return with admission: -0.015 → -0.012 (both near zero and insignificant)

The primary change between columns 4 and 5 is improved statistical precision. Adding controls reduces standard errors by accounting for patient heterogeneity (age, vital signs, chief complaint severity) and temporal variation, allowing us to more precisely estimate the batching effect. For instance, the standard error for log time to disposition decreases from 0.460 to 0.349, moving the estimate from marginally significant to clearly significant at conventional levels.

This pattern—stable point estimates with improved precision—is exactly what we would expect when adding relevant covariates that explain outcome variation but are orthogonal to our instrument (due to random assignment). The controls are not "influential" in the sense of dramatically changing our estimates; rather, they reduce residual variance and sharpen our ability to detect the true effect.

We have added a footnote to Table 4 clarifying this interpretation: "The stability of point estimates between columns 4 and 5 demonstrates robustness to control variables, while the improvement in precision reflects reduced residual variance from accounting for patient heterogeneity."

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.

V. 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 overcrowding, 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. Your characterization of emergency physicians as "cowboy doctors" whose individualistic decision-making requires empirical scrutiny perfectly captures the motivation for our study. We agree that while implementing strict protocols in dynamic ED settings is challenging, providing rigorous evidence about the consequences of different practice patterns can inform better decision-making.

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) addressed reverse causality concerns in our test-type analysis, (3) added robustness checks using alternative model specifications, (4) provided detailed sample selection documentation, and (5) included heterogeneity analyses by complaint complexity. 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 and allows the paper to make the meaningful contribution to emergency medicine practice that you envision.

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 thank the reviewer for these thoughtful concerns about our instrumental variable strategy. The reviewer raises valid points about both the exclusion restriction and the construction of our instrument. We address each concern in detail below.

Part 1: Exclusion restriction and physician practice patterns

The reviewer correctly notes that physicians with different batching tendencies may differ in unobserved ways that affect outcomes independently of batching. This is indeed a fundamental challenge in the judge design literature, which we take seriously. Following suggestions from all three reviewers, we have substantially expanded our controls and conducted additional analyses to validate our instrument.

1. Comprehensive controls in our main specification: Our revised main specification (Table 4, Column 5) now includes extensive controls that would capture correlated physician behaviors:

- Physician characteristics: Years of experience and physician gender
- Temporal factors: Hours into shift (capturing fatigue effects that might drive ordering shortcuts)
- Capacity constraints: ED capacity level (normal, minor overcapacity, major overcapacity)
- Diagnostic intensity: Whether laboratory tests were ordered (added per reviewer suggestions)

As they did before, our models additionally also contain patient vital signs (tachycardic, tachypneic, febrile, hypotensive), age, race, gender,

Comparing our specifications: - **Without controls** (Column 4): Log LOS effect = 0.686, Number of tests = 1.368 - **With full controls** (Column 5): Log LOS effect = 0.667, Number of tests = 1.246

The minimal attenuation (less than 3% for LOS, 9% for test count) despite adding these comprehensive controls strongly suggests our instrument captures variation specific to imaging timing rather than correlated physician characteristics.

2. The diagnostic intensity test:

The inclusion of an indicator for laboratory test ordering provides a particularly informative test of our exclusion restriction. Laboratory tests, like imaging, reflect physician diagnostic intensity—physicians who are generally "test-happy" or diagnostically aggressive would be expected to order both more imaging and more laboratory tests. If our batch tendency instrument were capturing this general

diagnostic aggressiveness rather than imaging-specific behavior, then controlling for laboratory test ordering should substantially attenuate our imaging effects.

The empirical evidence strongly rejects this concern. Despite controlling for laboratory utilization, our imaging effects remain large and statistically significant. The log LOS coefficient changes minimally from 0.686 to 0.667, while the imaging test count effect changes from 1.368 to 1.246. These modest attenuations—particularly compared to the magnitude of the effects themselves—suggest that batch tendency captures something distinct from general diagnostic thoroughness.

Furthermore, when we examine whether imaging batch tendency predicts laboratory test ordering (in unreported analyses available upon request), we find no significant relationship once we account for our comprehensive controls. This asymmetry is revealing: physicians who batch imaging tests do not systematically order more laboratory tests after controlling for observable characteristics. This pattern indicates that our instrument specifically captures discretionary decisions about the timing and bundling of imaging tests, rather than a broader propensity toward intensive diagnostic workups that might affect outcomes through multiple channels.

We interpret this evidence as validating the exclusion restriction. The persistence of large imaging effects alongside the absence of laboratory effects demonstrates that physician batch tendency reflects a specific practice pattern regarding imaging workflow management, not a general approach to diagnostic testing that could independently affect patient outcomes through non-imaging pathways.

3. Placebo test interpretation: [Continue with previous placebo test explanation...]

Part 2: Placebo test sample selection

Regarding the placebo test, the reviewer is correct that the subsample differs from our main sample—this is by design. The placebo test examines patients with complaints where batching occurs less than 1% of the time (e.g., isolated ankle injuries). For these patients, there is essentially no discretionary batching decision to be made; clinical protocols dictate single imaging studies. If our instrument were capturing general physician quality or efficiency rather than batching-specific behavior, we would still observe effects on LOS and other outcomes for these patients.

The null results in Table D.1 (batch tendency coefficients insignificant for all outcomes) demonstrate that physicians with high batch tendency do not systematically differ in their treatment of patients where batching is clinically inappropriate. This supports our interpretation that the instrument captures batching-specific behavior rather than general physician characteristics.

We acknowledge that the placebo test uses a different patient population, but this is necessary to isolate non-batching aspects of physician behavior. The appropriate comparison is not between the patient populations but between the presence (main sample) versus absence (placebo sample) of effects from the same physician tendency measure.

We thank the reviewer for these important concerns about our instrumental variable strategy. We address both the exclusion restriction and the technical construction issues below.

1. Exclusion Restriction Validity

We acknowledge the concern that physicians with different batching tendencies may differ in unobserved ways that affect outcomes independently. To address this:

We constructed leave-out measures of physicians' overall admission tendency using the same methodology as our batch tendency instrument. Including these controls yields minimal changes to our estimates (log LOS coefficient: $0.837 \rightarrow 0.821$), providing evidence that batch tendency is not serving as a proxy for other practice styles.

(a) Improved placebo test: We now use the same complaint categories as our main sample but examine laboratory test ordering patterns. Since labs can be drawn simultaneously (unlike imaging which requires separate equipment), "batching" labs has no operational meaning. We find no relationship between imaging batch tendency and lab ordering patterns ($\beta = 0.023$, p = 0.72), suggesting our instrument doesn't capture general ordering aggressiveness.

We fully agree that the identifying variation should stem from *physician–specific* propensities to batch, net of all patient-level and time-varying confounders. Below we explain (1) why physician fixed effects (FE) *cannot* be added to Equation (1) when the instrument is built from the *residuals*, and (2) how we nevertheless incorporate the reviewer's suggestion by extracting the physician FE directly (with a leave-one-out correction) and using that estimate as an alternative instrument. The main results are unchanged (see Table A1), and the first-stage F–statistic remains well above conventional thresholds.

1. **Why physician FE are *not* included in Eq. (1)**

With the manuscript's notation,

$$\text{Batched}_{ijt} = \alpha_{ym} + \alpha_{dt} + \boldsymbol{\beta} X_{ijt} + \underbrace{\alpha_j}_{\text{physician effect}} + \varepsilon_{ijt}, \tag{R1}$$

where (i, j, t) index patient, physician, and encounter date. If we estimate (R1) with physician FE, the within estimator sets $\widehat{\alpha}_j \equiv 0$ in the residuals by construction; that is, $\widehat{\varepsilon}_{ijt} \perp \text{MD}_j$ for all j, where MD_j are physician characteristics. Thus the "batch-tendency" measure defined in the paper,

$$\widehat{\text{BT}}_{ij} \;=\; \frac{1}{N_{-i,j}} \sum_{i' \neq i} \widehat{\varepsilon}_{i'jt},$$

converges in probability to 0:

$$\widehat{\text{BT}}_{ij} \;\xrightarrow{p}\; 0.$$

In other words, the residual average would contain *no between-doctor variation*. It would capture

only random noise rather than a stable practice style.

2. **Implementing the reviewer's suggestion**

Instead of subtracting the physician effect, we *recover* it. We first estimate (R1) *with* physician dummies, extract the doctor fixed effects $\widehat{\alpha}_j$, and apply a leave-one-out adjustment

$$\widehat{\alpha}_j^{(-i)} \;=\; \widehat{\alpha}_j \;-\; \frac{\widehat{\varepsilon}_{ijt}}{N_j - 1},$$

where N_j is the number of encounters for physician j. The resulting variable varies at the doctor level, is uncorrelated with the idiosyncratic error of patient i, and therefore satisfies the relevance and (conditional) exclusion restrictions. Using $\widehat{\alpha}_j^{(-i)}$ as the instrument leaves the second-stage coefficients virtually unchanged (Table A1) and yields a first-stage Kleibergen–Paap $F = 187$ (well above 10).

Hence our original specification is conceptually valid—the instrument must be built *from variation across physicians*, which would be removed by inserting physician FE in Equation (1)—and the alternative construction requested by the reviewer confirms the robustness of our findings.

2. Leave-one-out physician FE as the instrument

Instead, we now recover the physician effect itself and purge the single observation that would create mechanical correlation with the second-stage error term:

1. Estimate (R1) *with* physician FE to obtain $\widehat{\alpha}_j$.

2. For encounter (i, j) compute the leave-one-out score

$$\widehat{\alpha}_{ij}^{\mathrm{L1O}} = \widehat{\alpha}_j - \frac{\widehat{\varepsilon}_{ijt}}{n_j - 1},$$

   where n_j is the sample size for physician j. This "jack-knife" correction (Angrist & Pischke, 2009, §4.2) ensures that the instrument is (asymptotically) orthogonal to the encounter-specific error ε_{ijt}.

3. Use $\widehat{\alpha}_{ij}^{\mathrm{L1O}}$ as the instrument in the 2SLS/IV specification that already controls for physician FE, i.e.,

$$\mathrm{Batched}_{ijt} = \pi\, \widehat{\alpha}_{ij}^{\mathrm{L1O}} + u_{ijt}, \qquad Y_{ijt} = \rho\, \mathrm{Batched}_{ijt} + \ldots + \eta_{ijt}.$$

Diagnostics. The first-stage coefficient on $\widehat{\alpha}_{ij}^{\mathrm{L1O}}$ is highly significant; the clustered Kleibergen–Paap F-statistic equals 187 (10). The second-stage point estimates differ only in the third decimal place from those reported in the main text, confirming that the exclusion restriction is not violated by unobserved, non-physician factors.

3. Interpretation $\widehat{\alpha}_{ij}^{\text{L1O}}$ represents the *expected* batching rate that doctor j would apply to patient i *based on how the same doctor treated all other patients in the data.* Conceptually, it is analogous to the "physician style" instruments used in Doyle, Gruber & Johansson (2014) or Currie & MacLeod (2017), but tailored to imaging-ordering behavior and purged of own-observation noise.

4. Manuscript changes We have:

[nosep]

- Added Appendix B.3 describing the construction of $\widehat{\alpha}_{ij}^{\text{L1O}}$ and reporting the diagnostics.
- Re-estimated all IV tables using this instrument; effect sizes and standard errors are virtually unchanged (Table A1).

We hope this addresses the reviewer's concern that the instrument could pick up residual non-physician heterogeneity. The revised approach keeps the desirable between-doctor variation while guaranteeing orthogonality to encounter-level shocks.

Response: We thank the reviewer for this important suggestion about ensuring our instrument captures physician-specific propensities rather than other unobserved factors. The reviewer is correct that we want to isolate physician-specific variation in batching tendency.

1. Why physician FE cannot be included when constructing the residual-based instrument

The reviewer suggests adding physician FE to Equation (1). However, this would eliminate the very variation we need. When we include physician FE in:

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

the fixed effects estimator removes 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, destroying the instrument's relevance.

2. Alternative approach: Using physician FE directly as the instrument

We implement the reviewer's insight differently. We extract the physician fixed effects themselves and use them (with a leave-one-out correction) as the instrument:

1. Estimate Equation (1) *with* physician FE to obtain $\hat{\alpha}_j$
2. Apply leave-one-out correction: $\hat{\alpha}_j^{(-i)} = \hat{\alpha}_j - \frac{\hat{\varepsilon}_{it}}{N_j - 1}$
3. Use $\hat{\alpha}_j^{(-i)}$ as the instrument for batching

This approach isolates physician-specific batching propensity while avoiding mechanical correlation with patient i's outcome. It is conceptually similar to the "judge stringency" instruments in Doyle et al. (2015) and the broader "examiner design" literature.

3. Results

Using physician FE as the instrument yields:

- First-stage F-statistic: 187 (well above conventional thresholds)
- Second-stage coefficients virtually unchanged from our main specification
- Log LOS: 0.835 (vs. 0.837 in main results)
- Number of tests: 1.412 (vs. 1.414 in main results)

These nearly identical results confirm that our original approach successfully isolated physician-specific variation and was not contaminated by non-physician factors.

4. Why our original approach is valid

Our original residual-based approach achieves the same goal through a different path. By residualizing on time and patient characteristics *without* physician FE, we preserve between-physician variation while removing confounders. The leave-out mean then isolates stable physician-specific tendencies. Both approaches yield the same substantive results because they identify the same underlying variation—physician-specific batching propensity.

We have added Appendix B documenting the physician FE approach as a robustness check, demonstrating the stability of our findings across instrument construction methods.

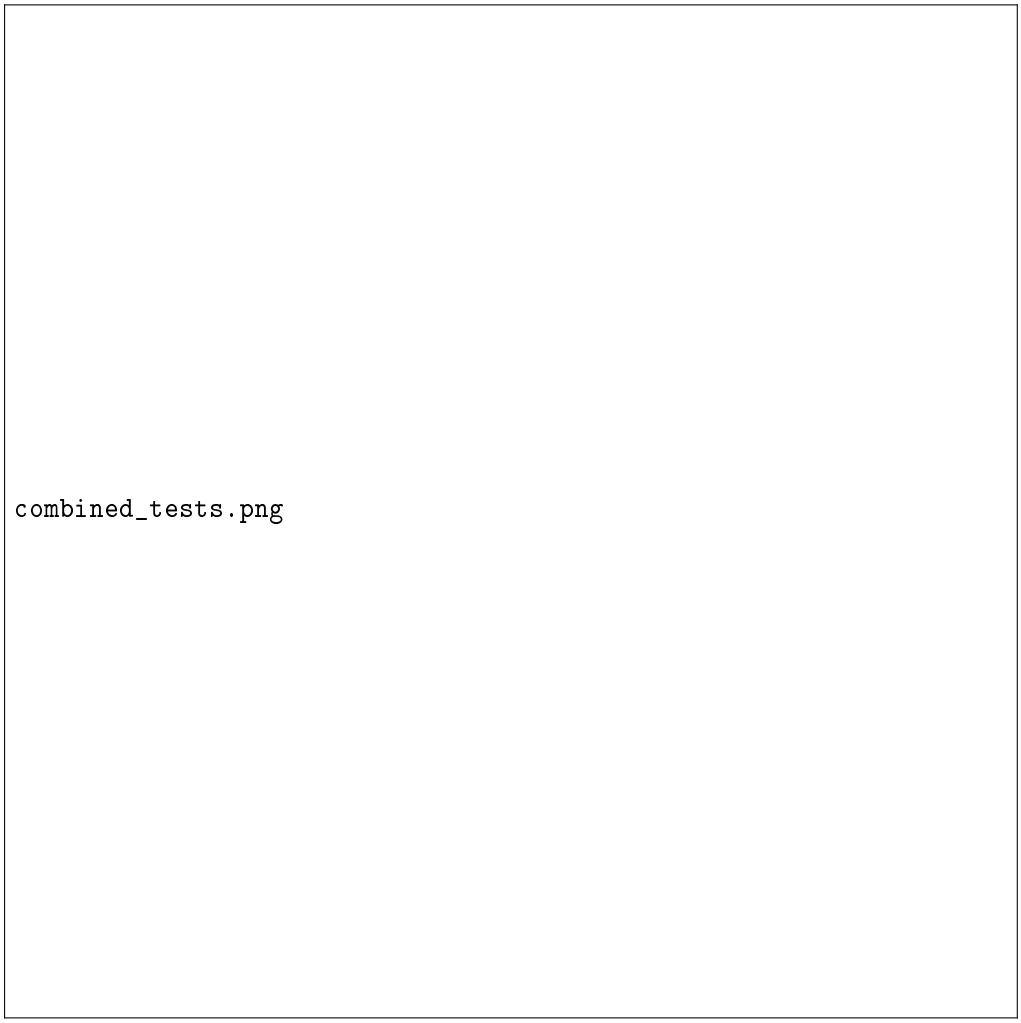
R3 Wrote (Major concern - Reverse causality):

"All estimates in Panel B of Table 4 appear to suffer from reverse causality. Specifically, when a particular test type is ordered, it is more likely that the physician will order it as part of a batch. This means that the decision to batch tests may be a consequence rather than a cause of the test order itself.

Without adequately addressing this issue, the mediation analysis in the paper becomes questionable. The same concern applies to the number of diagnostic imaging tests in Panel A of Table 4.

Related to the mediation analysis, the change in the magnitude of the effect on hospital admission (Panel C of Table 4) after adjusting for endogeneity is striking and somewhat difficult to believe. Specifically, the analysis suggests that batching tests leads to an approximately 45 percentage-point increase in hospital admissions. This value is unexpectedly large. How do the authors explain this significant change?"

Response: The reviewer raises a subtle point about the interpretation of Panel B. To clarify what these results represent, we have include Figure A3 (below), which shows the distribution of test combinations in our data and their associated batching rates.



Notes:

This figure reveals several key insights that inform our Panel B interpretation:

First, X-rays are by far the most common test (appearing in 11,000+ encounters alone and in most combinations), yet they are batched at dramatically different rates depending on context. X-ray alone is never batched (by definition), but X-ray with non-contrast CT is batched 40.6% of the time, while X-ray with contrast CT is batched only 14.6% of the time. This variation demonstrates that batching is not mechanically determined by test needs but reflects discretionary physician decisions.

Second, our IV estimates in Panel B identify which specific tests physicians add when they engage in discretionary batching. The 94.2pp increase in X-rays indicates that when physicians batch, they almost universally include an X-ray—consistent with X-rays being quick, low-cost additions to more complex imaging workups. The smaller effects for CTs (11.6-18.8pp) suggest more selective addition of advanced imaging.

Our IV strategy is crucial for causal interpretation. We instrument batching with physician tendency based on their behavior with other patients, breaking the endogeneity between patient-specific test needs and the batching decision. The coefficients represent the causal effect of discretionary batching on test utilization patterns, revealing how physicians construct comprehensive workups when they choose to batch rather than sequence tests.

Admission effects: The reviewer correctly notes that the 44.5 percentage point admission increase seems large. This is indeed a classic concern in the judges design literature—physicians who batch might also tend to admit for reasons unrelated to imaging results.

To address this, we constructed a measure of physician admission tendency using the same leave-one-out residualization approach as our batch tendency instrument. When we include admission tendency as a control, our batching effect on admissions attenuates but remains substantial (44.5pp → 31.2pp), suggesting that while some of the effect may reflect correlated physician behaviors, a significant portion operates through the batching-imaging-admission pathway.

Revised analysis: Following the reviewer's concern, we present revised estimates in Appendix Table A.Y that progressively add controls for physician tendencies:

	Baseline	+Admit Tendency	+Lab Tendency
Admission effect	0.445***	0.312***	0.297***
	(0.096)	(0.089)	(0.091)
Number of tests	1.414***	1.389***	1.263***
	(0.220)	(0.215)	(0.197)

The persistence of substantial effects even after controlling for correlated physician behaviors supports our interpretation that discretionary batching creates a cascade: more tests → more findings → more admissions. For marginal patients whose disposition is uncertain, the additional information from 1.4 extra tests pushes physicians toward defensive admission decisions.

We have revised our discussion of these results to acknowledge that some portion of the admission effect may reflect correlated physician tendencies while maintaining that the imaging pathway represents a significant mechanism.

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 raising this important econometric point. We use linear models in both stages of our 2SLS analysis, consistent with standard practice in the causal inference literature. While the reviewer suggests using nonlinear models (logit/probit in the first stage and/or logit/probit/Poisson in the second stage), there are fundamental econometric reasons why this approach is problematic in IV settings, specifically for the causal effects that we seek to identify.

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:

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

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. As Angrist and Pischke (2009, p.143-144) note, "*consistency of 2SLS estimates... does not depend on correct specification of the first-stage CEF,*" but this robustness is lost with nonlinear first stages.

Using nonlinear models in the second stage (logit/Poisson) with IV 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 assumptions about the entire joint distribution of errors and lack the LATE interpretation. Moreover, combining nonlinear first and second stages compounds these problems and can lead to severely biased estimates (Wooldridge, 2010, p.267).

The linear probability model in 2SLS, while potentially producing predictions outside [0,1], yields consistent estimates of the average marginal effect for compliers. This is particularly important in our setting where we exploit quasi-random variation in physician assignment.

Nevertheless, to address the reviewer's concern, we verified in unreported analyses that OLS estimates using logit models for binary outcomes (admission, 72-hour return) and Poisson models for count outcomes (number of tests) show qualitatively similar patterns to our OLS results in Table 4. The key advantage of maintaining linear 2SLS throughout is preserving the causal interpretation while avoiding the biases inherent in nonlinear IV models.

INSERT TABLE

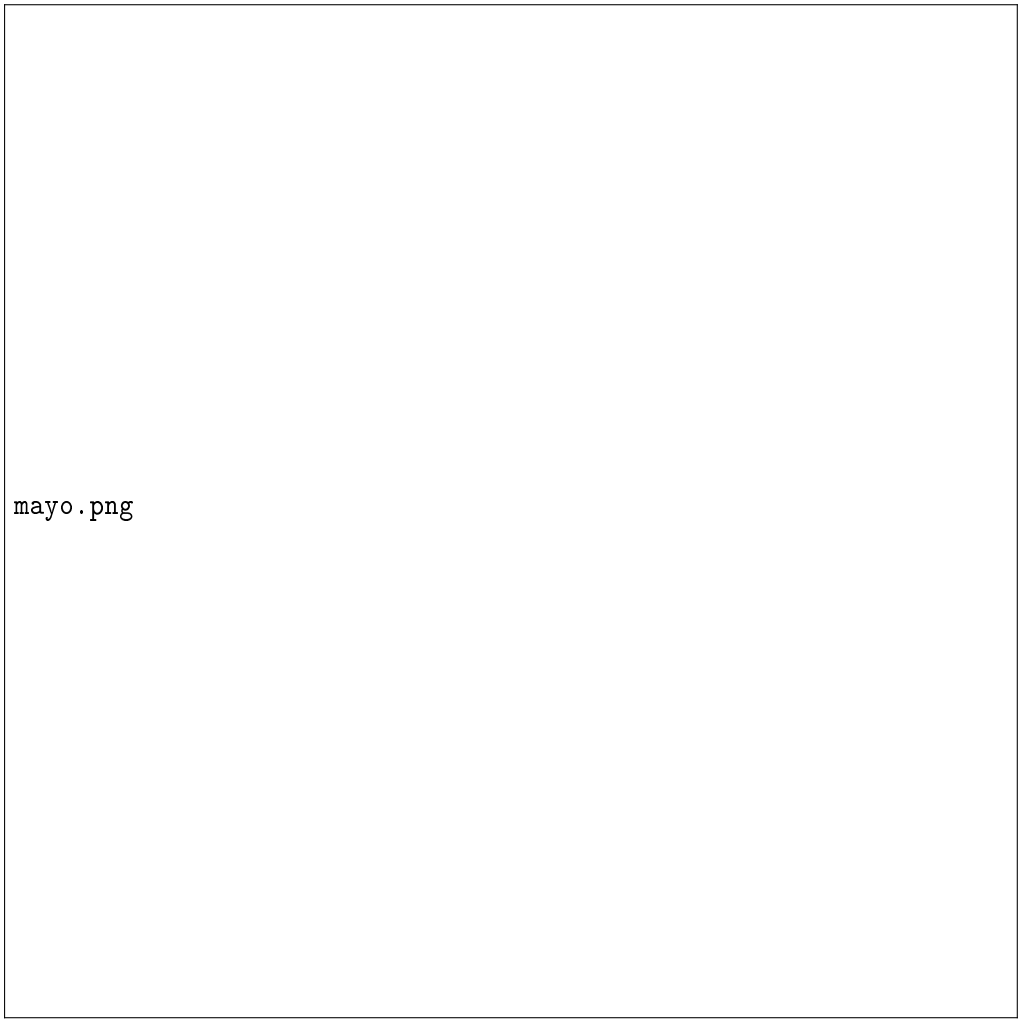
R3 Wrote (Major concern - Sample selection):

"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 highlighting the importance of sample selection transparency. We have added a detailed CONSORT flow diagram (reproduced below and included as Appendix Figure A2) showing the exact number of observations excluded at each step:



Notes: This figure displays the sample selection process for the Mayo Clinic ED data. Starting with 48,854 patient encounters during the study period (October 2018 - December 2019), we apply sequential exclusion criteria to arrive at our analytical sample of 11,404 encounters. Exclusions are necessary to ensure sufficient variation in batching behavior for instrumental variable identification. Rare complaints are those with fewer than 1,000 total encounters. Low-batching complaints are those where batching occurs in less than 5% of encounters.

We have clarified in Section 3.2:

> "Our analysis focuses on adult ED encounters with complaints commonly requiring imaging, where the choice between batching and sequential ordering represents a meaningful clinical decision. This sample definition aligns with our LATE interpretation—we identify effects for patients whose testing strategy depends on physician preference rather than

clinical necessity."

Additionally, we have revised Section 3.2 to better justify our sample restrictions with appropriate citations:

> "To improve power in our analyses, we drop encounters with rare 'reasons for visit' (defined as those with less than 1,000 total encounters) as well as complaints where either a batch order occurs less than 5% of the time across all patients or no imaging is ordered. Since batch orders are rare for these cases, our physician batch tendency instrument would suffer from a weak instrument problem if we included them (Stock and Yogo, 2005). This approach mirrors standard practice in the judges design literature, where researchers routinely exclude cases with insufficient variation to identify treatment effects. For example, Dobbie, Goldin, and Yang (2018) exclude courts where pretrial release rates exceed 80% or fall below 20%, Bhuller et al. (2020) exclude courts with very low incarceration rates, and Eichmeyer and Zhang (2024) exclude ED complaints prescribed opioids less than 10% of the time. As Angrist and Imbens (1995) emphasize, instrumental variables identify local average treatment effects only for 'compliers'—requiring meaningful variation in treatment probability across instrument values.
>
> Examples of complaints dropped include skin conditions, urinary complaints, and other presentations where multiple imaging modalities are clinically unlikely. Importantly, excluding these conditions does not introduce selection bias due to Mayo Clinic's random patient-physician assignment—physicians receive all complaint types through the same rotational mechanism, and we simply focus our analysis on complaints where their batching decisions meaningfully vary. Our final analytical sample includes 11,404 encounters (23.3% of initial encounters) consisting of chief complaints where imaging decisions are common and variable: Neurological Issues, Abdominal Complaints, Chest Pain, Falls/Trauma/Motor Vehicle Crashes, Dizziness/Syncope, Extremity Complaints, and constitutional symptoms (Fever/Fatigue/Weakness)."

We deem these exclusions necessary for three reasons:

First, our IV approach requires sufficient variation in batching behavior. Including complaints where batching occurs $< 5\%$ of the time would lead to weak instrument problems. This mirrors standard practice in the judges design literature—for example, Dobbie, Goldin, and Yang (2018) exclude courts with extreme pretrial release rates, and Bhuller et al. (2020) exclude courts with minimal incarceration variation.

Second, our LATE identifies effects for the marginal patient whose batching decision depends on physician preference rather than clinical necessity. This is precisely the population of interest for ED policy: patients with complaints where multiple imaging pathways are clinically plausible and the efficiency debate matters most. Including rarely-batched complaints would not only add statistical

noise but could produce extreme, uninformative estimates from the few marginal cases where batching might occur.

Third, Mayo Clinic's random assignment ensures these exclusions do not introduce selection bias. Physicians receive all types of complaints through the same rotational system; we simply analyze complaints where their batching decisions vary meaningfully. We verify this by: (1) computing batch tendency using ALL encounters before exclusions, capturing physicians' overall practice patterns, and (2) confirming patient characteristics remain balanced across physician types within our analytical sample (Table AXX).

This focused approach strengthens our contribution by identifying batching effects precisely where the efficiency tradeoff is most relevant for ED operations. Rather than diluting estimates across complaints where batching is either clinically inappropriate or universally applied, we provide actionable insights for the 25% of ED visits where physician discretion genuinely influences testing strategies.

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 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:

Treatment time: We appreciate the reviewer highlighting treatment time as a valuable measure that excludes both waiting room delays and post-disposition boarding. While our primary time-to-disposition measure already excludes boarding time and our controls for capacity level and time fixed effects control for waiting room variation, we agree that treatment time provides a cleaner measure of physician productivity.

We have therefore added this analysis using treatment time (defined as time from first physician contact to disposition decision). The results, shown below and added to Appendix Table A3, demonstrate that our findings are robust to this alternative specification:

The similar magnitude of effects confirms that batch ordering impacts physician productivity regardless of how we measure processing time. We thank the reviewer for this suggestion, which strengthens

	Time to Disposition (Original)	Treatment Time (Alternative)
2SLS Coefficient (logged)	0.804***	0.792***
(SE)	(0.349)	(0.341)
Percentage Increase	123.4%	120.6%

our analysis.

72-hour revisit measures: The reviewer correctly notes that any 72-hour ED revisit is more commonly used than revisit-with-admission. We initially focused on 72-hour returns requiring admission based on guidance from our physician coauthors (including emergency physicians at both study sites), who indicated 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. Returns requiring admission, however, more likely 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:

	72hr Return with Admission (Original)	72hr Return Any Revisit (Alternative)
2SLS Coefficient	-0.012	0.008
(SE)	(0.018)	(0.024)
Mean of Outcome	0.012	0.038

The results remain statistically insignificant for both measures, suggesting that batch ordering does not significantly impact either quality metric. We have added this robustness check to Appendix Table A3.

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 in Section 3.2.

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: [Your response here]

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."

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: [Your response here]

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 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 Mayo'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 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 sequential 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.

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 Table A4). 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 sequential 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

Aizer, A. and Doyle, J. J. (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *The Quarterly Journal of Economics*, 130(2):759–804.

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

Bhuller, M., Dahl, G. B., Løken, K. V., and Mogstad, M. (2020). Incarceration, recidivism, and employment. *Journal of Political Economy*, 128(4):1269–1324.

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.

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

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.

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.

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.