Dear Editors,

Thank you for these comments designed to improve our paper, “***Variation in Batch Ordering of Imaging Tests in the Emergency Department and the Impact on Care Delivery***,” which we have addressed below. We greatly appreciate the time and effort put forth by reviewers and editors to improve our paper. If any responses are unclear or you wish additional changes, please let us know.

Kindly,

Jacob Jameson

Editor

1. *Overall, the results of the analysis are modest. One of the ways the paper explores variation in imaging is by evaluating patients based on chief complaint, and find that two of the diagnoses with the most variation are trauma and neurological issues. Both of these chief complaints have a wide-range in patient severity (e.g., is the trauma a gun shot wound, an MVA, a patient falling down the stairs, or a knife injury in the kitchen?). So, without further detail on the chief complaint, the variation may be clinically appropriate. The discussion should focus a bit more on the challenges of this assessment, rather than the implications of this work.*

**There is indeed a wide range in patient severity across chief complaint categories. Appendix Table 3A displays how we mapped each reason for a visit to a broader chief complaint category using a methodology previously used in the literature.**

**In addition to the chief complaint, we control for the ESI score determined at triage. This allows us to differentiate between emergency severity of patients with complaints that fall under the same category, such as trauma or neurological issues.**

1. *In terms of “batching” do we have information on what is batched? Are the batching appropriate? For example, again with trauma, the patient may need a Head CT and a Chest CT as a result of an MVA. But, a chest x-ray and a chest CT ordered at the same time seems less appropriate as a batch (you should wait for the x-ray result).*

**In terms of batching, we do have information on what is batched. From [Page 7; Line 13]:**

**“Each imaging test (e.g., X-ray, CT scan, Ultrasound) is considered a separate, distinct test for our study. Therefore, a batch in our study consists of two or more distinct imaging tests ordered at the start of a patient encounter by the attending physician.”**

**According to our definition of batching, a head CT and chest CT ordered together would not constitute a batch. A chest x-ray and a chest CT ordered simultaneously would meet our batching criteria. We do not make a judgment on what tests ordered are clinically appropriate; instead, we focus more on the operational aspect of test ordering and the variation between physicians who are randomly assigned patients.**

1. *Do you mean “Data Manipulation” or “Data Analysis”?*

**The title of this section has been changed to “Data Transformation” (per Reviewer 1’s suggestion).**

1. *Follow author instructions for tables (hsr.org/authors, Section 2.4.2.6) and figures (Section 2.4.2.7). Use table footnotes and figure legends to define abbreviations and information in parentheses, and explain content so tables and figures are understandable without referring to the main text.*

**We have followed the author instructions for tables and figures.**

1. *Rewrite table and figure titles so they reflect the content not methods.*

**We have rewritten table and figure titles to reflect the content, not the methods.**

1. *Clearly label supplemental material as “supplemental” or “appendix.”*

**We have clearly labeled supplemental materials in the Appendix.**

1. *Figures must be legible when the page is viewed at 100% (hsr.org/authors, Section 2.4.2.7). Consider adjusting the contrast and using black instead of gray to increase readability and comply with accessibility guidelines.*

**We have updated our figures to use darker colors to increase readability.**

Reviewer 1

1. *(Major) The data span 1 year and 2 months in total. Did the authors divide the data into training and test sets? If not, this step is necessary for model development. One limitation exists regarding the construction of the test set given that only 2 calendar months are repeating.*

**We use 10-fold cross-validation to produce predicted probabilities of batching on the test sets. Because our training sets for model development are randomly sampled from the data in this approach, we use training data representative of the entire study period. This approach also allows us to use our full sample of data, rather than a test sample, for all our subsequent analyses.**

**We have added an Appendix Figure 1A, where we also report the ROC/AUC of a version of the model where the training data is the first one year of data and the test data is the last 2 months (the months that repeat in the study period). We show that the test AUC of 10-fold cross-validation is nearly identical to this approach (0.75 vs 0.76). This suggests that a physician’s batch tendency is consistent throughout the study period (i.e., if Physician A has a high batch tendency in the first few months of the data, they will also continue to have a high batch tendency in the last few months of data).**

1. *(Minor) The acronyms “LOS” and “ESI” should be defined the first time the full spellings appear and used for all subsequent references.*

**We have made this change.**

1. *(Minor) Page 6 line 22: Missing citation.*

**The citation has been added.**

1. *(Major) Please provide a citation, description, and evidence for the randomized rotational patient assignment algorithm adopted by the Mayo Clinic ED. The provided citation #12 does not account for this matter.*

**We have enhanced our description of the patient assignment algorithm [Page 6; Line 16]:**

**“In contrast, patients arriving at the Mayo Clinic ED are assigned to physicians via a randomized rotational patient assignment algorithm, which practically removes potential selection bias concerns from our analyses13. A computer algorithm electronically assigned patients to physicians 60s after registration, and assignments are made solely on arrival time, without consideration of patient demographics, chief complaint, ESI, physician-patient load, or acuity of patients recently assigned to the physician.”**

**We have updated the citation to point towards the following, which describes the implementation of this algorithm, provides evidence of its effectiveness, and explains how it works.**

**Traub SJ, Saghafian S, Bartley AC, Buras MR, Stewart CF, Kruse BT. The durability of operational improvements with rotational patient assignment. Am J Emerg Med. 2018;36(8):1367-1371. doi:10.1016/j.ajem.2017.12.045**

1. *(Major) It would be helpful to expand the balance test in Table 1A. Why are there only a subset of chief complaints? Please distinguish all 5 ESI levels. Also, what would the results look like if you run direct linear regressions with chief complaints/ESI as independent variable and physician ID as dependent variable?*

**We have expanded Table 1A to include all 16 chief complaint categories that made it into our analytical sample. We have also expanded to include all 5 ESI levels and other patient characteristics collected at triage (race, age, sex, hypotension, tachycardia, tachypnea, and fever).**

**To clarify our results in Table 1A, as you suggest, we run an intercept-only linear regression model where complaints/ESI/other characteristics are independent variables that serve as a reference to estimate the overall mean of each variable across all observations. We then extend each regression model by including Physician ID as predictors. Each row in the table corresponds to a separate Wald test for a particular independent variable, testing whether the inclusion of Physician ID improved the model fit. The F-statistic reported in each row measures the strength of the test, indicating the extent to which the model fit improves with the inclusion of physician identifiers. The p-value provides the probability that the observed distribution of chief complaints or ESI categories across physicians could arise by chance if there were no actual differences. Additionally, we included Bonferroni-corrected p-values due to the risk of type I errors when conducting this test over so many variables.**

**We have enhanced the notes of Table 1A to provide a more precise description of how we conducted this balance test.**

1. *(Major) The definition of batching is unclear and could be enhanced. Do the authors label batching if such behavior occurs at any time during each patient’s LOS? If 2 or more tests were ordered within a short time window after some initial tests, does this still count as batch ordering? Were some tests ordered before patients were seen by the physician? Are there timestamps in the data regarding the test results time? It is conceivable that some batched orders were clinically needed after an initial assessment and understanding of the patient’s condition were gained.*

**We have enhanced the definition in the text [Page 7; Line 10]:**

**“We operationalize batching as occurring when multiple diagnostic imaging tests are ordered within a 5-minute window, and these are the first tests ordered for a patient.”**

**To meet our primary definition of a batch order, the 2 or more tests ordered must be the first set of tests ordered for a patient. There are timestamps in the data regarding the test results time, and we agree that it is conceivable that some batched orders were clinically needed after an initial imaging test was already ordered and the physician has a better understanding of the patient's condition. This occurs in 192 (0.4%) patient encounters. However, we have added to Appendix Table 2A an additional robustness check that includes considering a broader definition of batching where such a situation would constitute a batch, and the results are nearly identical to our main analysis.**

*(Major) Please provide a sensitivity analysis for additional time window durations such as 20 minutes and 30 minutes. Also, provide a sensitivity analysis for alternative batch sizes such as >=3 and >=4.*

**We have added this sensitivity analysis to Appendix Table 2A. Our results are very robust to these varying definitions, however, because batches >=3 or >=4 are extremely rare in practice we are likely underpowered with this definition of batching. Nonetheless, results are reported in Appendix Table 2A.**

1. *(Major) One of my biggest reservations with the paper is that the regression model contains very crude covariates and is subject to many endogeneity issues and omitted variables. For example, patients’ clinical severity is only captured by the set of variables X\_compalint/severity, which only includes the grouped complaint areas and ESI scores. There can be much more patients’ clinical and severity information (such as vitals, demographics, disposition, whether ambulance was needed, etc.) that determines whether various imaging tests are needed, which is not controlled for in the model.*

**Indeed, we do not initially control for other covariates, such as vital signs, demographics, and disposition. Because of the randomized rotational patient assignment system used, we mimic an RCT where patients are randomly assigned to physicians who are either more or less likely to batch order tests by controlling for physician shift level variation and patient arrival time. However, we have added additional severity covariates to Table 1 and shown that they do not change our estimates on the effect of batch tendency.**

**Additionally, we do not control for disposition because whether the patient is admitted or discharged is not known at the time the physician places their initial test orders and, therefore, is not relevant to the decision to batch and may even be a downstream outcome of that decision if the batching leads to more tests.**

**Hodgson, Nicole R., Souroush Saghafian, Lanyu Mi, Matthew R. Buras, Eric D. Katz, Jesse M. Pines, Leon Sanchez, Scott Silvers, Steven A. Maher, and Stephen J. Traub. “Are Testers Also Admitters? Comparing Emergency Physician Resource Utilization and Admitting Practices.” Journal of Emergency Medicine 36.10 (October 2018): 1865-1869.**

1. *(Minor) Page 8, line 13: Please define “quasi-random assignment assumption.”*

**We have replaced with the following [Page 8; Line 13]: “As stated earlier, these controls are more than required for the patient-to-physician assignment to be deemed as good as random.”**

**This refers to the following in the Details on Data section: “In contrast, patients arriving at the Mayo Clinic ED are assigned to physicians via a randomized rotational patient assignment algorithm, which practically removes potential selection bias concerns from our analyses12. In essence, controlling for patient arrival time and physician shift-level variation, the physician-to-patient matching can be deemed random.”**

1. *(Minor) Page 8 line 44: “By considering batch tendency as a proxy measure for batching itself, we address the problem of endogeneity”. Please elaborate. This also seems somewhat an overstatement regarding “addressing the problem of endogeneity.”*

**We have enhanced the text to better describe how we address the endogeneity problem of studying batching directly [Page 9; Line 5]**

**“The batch tendency variable is constructed to reflect physician j’s underlying tendency to batch at patient i’s encounter and is independent of all patient i’s characteristics. Since the decision to batch order many tests could be related to patient i’s presenting condition, by considering batch tendency as a proxy measure for batching itself, we address the problem of endogeneity.16”**

1. *(Minor) Perhaps change “manipulation” to “transformation”.*

**The title of this section has been changed to “Data Transformation.”**

1. *(Major) Several important questions regarding the three metrics and their derived implications:  
   a. LOS: Does LOS include boarding time? Is disposition an omitted variable here? For example, severe patients are more likely to be admitted to the hospital, have longer LOS partly due to boarding, and require more tests.  
     
   b. 72-hour return: Again, the model does not control for disposition decision. Patients who are admitted to hospital will not have 72-hour return but may need more tests due to severity.  
     
   c. Number of distinct imaging tests ordered: This regression seems expected and does not provide much additional insight, especially due to the potential high correlation between batch ordering and the number of distinct tests ordered. By definition, batch ordering already implies having 2 or more tests ordered. It is unclear how the conclusion can be drawn: “This result indicates that batching may be leading to tests that would not have been otherwise ordered had the physician waited for the results from one test before ordering the next”.*

**(a) The length of stay includes boarding time. Disposition would not be an omitted variable because of the randomized assignment of patients to physicians. What we measure with the batch tendency is the effect of being randomly assigned to a physician who is more likely to batch tests versus a physician who is less likely to batch. In our models that control for patient characteristics, we only control for information that the physician would have at the time of the decision to batch. Because the decision to admit or discharge a patient is made after the initial diagnostic testing, it is not a relevant covariate for our analysis of test batching, and conditioning on it may bias our estimate.**

**(b) We do not control for disposition because this knowledge of patient disposition is not known to the physician at the time of the initial test ordering. However, controlling for disposition, we observe no change in our batch tendency estimate or confidence interval in the regression for 72-hour return with admission -0.002 [95% CI -0.003; -0.001].**

**(c) The result of distinct imaging tests ordered is important. This association shows us that the number of distinct imaging tests ordered for a patient is a function of whether that patient is randomly assigned to Physician A (batcher) or Physician B (sequencer), not their underlying condition. We have enhanced the text to clear up any confusion around the conclusion [Page 11; Line 12]:**

**“Since patients are balanced and randomly assigned to physicians who differ only in their batching probability, this result indicates that batching may be leading to tests that would not have been otherwise ordered had the physician waited for the results from one test before ordering the next.”**

1. *(Major) Please provide some measures of model fit, such as ROC/AUC.*

**We have enhanced our description around the model fit of our batching predicted model in the Statistical Analysis section and reported AUC on the test set. We have also added Appendix Figure 1A, which displays the ROC curve of the model test set.**

1. *(Minor) Page 11 line 54: Typo.*

**Fixed.**