







STUDY PROTOCOL

REVISED Protocol for the perioperative outcome risk assessment with computer learning enhancement (Periop ORACLE) randomized study [version 2; peer review: 2 approved]

Bradley Fritz ¹, Christopher King¹, Yixin Chen², Alex Kronzer¹, Joanna Abraham^{1,3}, Arbi Ben Abdallah¹, Thomas Kannampallil^{1,3}, Thaddeus Budelier ¹, Arianna Montes de Oca¹, Sherry McKinnon¹, Bethany Tellor Pennington¹, Troy Wildes ¹, Michael Avidan ¹

¹Department of Anesthesiology, Washington University School of Medicine, St. Louis, Missouri, 63110, USA

²Department of Computer Science and Engineering, Washington University McKelvey School of Engineering, St. Louis, Missouri, 63130, USA

³Institute for Informatics, Washington University School of Medicine, St. Louis, Missouri, 63110, USA

v2 First published: 14 Jun 2022, 11:653
<https://doi.org/10.12688/f1000research.122286.1>

Latest published: 29 Sep 2022, 11:653
<https://doi.org/10.12688/f1000research.122286.2>





Abstract

Background: More than four million people die each year in the month following surgery, and many more experience complications such as acute kidney injury. Some of these outcomes may be prevented through early identification of at-risk patients and through intraoperative risk mitigation. Telemedicine has revolutionized the way at-risk patients are identified in critical care, but intraoperative telemedicine services are not widely used in anesthesiology. Clinicians in telemedicine settings may assist with risk stratification and brainstorm risk mitigation strategies while clinicians in the operating room are busy performing other patient care tasks. Machine learning tools may help clinicians in telemedicine settings leverage the abundant electronic health data available in the perioperative period. The primary hypothesis for this study is that anesthesiology clinicians can predict postoperative complications more accurately with machine learning assistance than without machine learning assistance.

Methods: This investigation is a sub-study nested within the TECTONICS randomized clinical trial (NCT03923699). As part of TECTONICS, study team members who are anesthesiology clinicians working in a telemedicine setting are currently reviewing ongoing surgical cases and documenting how likely they feel the patient is to experience 30-day in-hospital death or acute kidney injury. For patients who are included in this sub-study, these case reviews will be randomized to be performed with access to a display showing machine learning predictions for the postoperative complications or

Open Peer Review

Approval Status  

	1	2
version 2 (revision) 29 Sep 2022	 view	 view
version 1 14 Jun 2022	  view	

1. **Michael Robert Mathis** , University of Michigan Medical School, Ann Arbor, USA
2. **Pengbin Yin**, Chinese PLA General Hospital, Beijing, China
Ming Chen, Chinese PLA General Hospital, Beijing, China
Yi Li, Chinese PLA General Hospital, Beijing, China

Any reports and responses or comments on the article can be found at the end of the article.

without access to the display. The accuracy of the predictions will be compared across these two groups.

Conclusion: Successful completion of this study will help define the role of machine learning not only for intraoperative telemedicine, but for other risk assessment tasks before, during, and after surgery.

Registration: ORACLE is registered on ClinicalTrials.gov: NCT05042804; registered September 13, 2021.

Keywords

Anesthesiology, Machine Learning, Postoperative Complications, Protocol, Surgery



This article is included in the **Artificial Intelligence and Machine Learning** gateway.



This article is included in the **AI in Medicine and Healthcare** collection.

Corresponding author: Bradley Fritz (bafritz@wustl.edu)

Author roles: **Fritz B:** Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Original Draft Preparation; **King C:** Conceptualization, Methodology, Software, Writing – Review & Editing; **Chen Y:** Conceptualization, Funding Acquisition, Resources, Writing – Review & Editing; **Kronzer A:** Data Curation, Software, Writing – Review & Editing; **Abraham J:** Conceptualization, Funding Acquisition, Writing – Review & Editing; **Ben Abdallah A:** Methodology, Writing – Review & Editing; **Kannampallil T:** Methodology, Writing – Review & Editing; **Budelier T:** Methodology, Writing – Review & Editing; **Montes de Oca A:** Methodology, Writing – Review & Editing; **McKinnon S:** Project Administration, Writing – Review & Editing; **Tellor Pennington B:** Methodology, Writing – Review & Editing; **Wildes T:** Conceptualization, Methodology, Resources, Writing – Review & Editing; **Avidan M:** Conceptualization, Funding Acquisition, Methodology, Resources, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the National Institute of Nursing Research [R01NR017916 to Dr. Avidan]; the Foundation for Anesthesia Education and Research [MRTG08152020 to Dr. Fritz]; the National Center for Advancing Translational Sciences [KL2TR002346 to Dr. King]; the National Science Foundation [1622678 to Dr. Avidan and Dr. Chen] and the Agency for Healthcare Research and Quality [R21HS024581 to Dr. Avidan].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Fritz B *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Fritz B, King C, Chen Y *et al.* **Protocol for the perioperative outcome risk assessment with computer learning enhancement (Periop ORACLE) randomized study [version 2; peer review: 2 approved]** F1000Research 2022, 11:653 <https://doi.org/10.12688/f1000research.122286.2>

First published: 14 Jun 2022, 11:653 <https://doi.org/10.12688/f1000research.122286.1>

REVISED Amendments from Version 1

This revised version was edited in response to the reviewer comments. Additional methods information has been added describing the training and validation of the machine learning predictions. In addition, we have added clarification that patients on preoperative dialysis, undergoing dialysis access procedures, or with baseline creatinine > 4.0 mg/dl will be excluded from acute kidney injury analyses. We have also added some sensitivity analyses and secondary outcomes suggested by the reviewer.

Any further responses from the reviewers can be found at the end of the article

Introduction**Background and rationale**

Each year, more than four million people worldwide die within 30 days after surgery, making perioperative events the third-leading cause of death.¹ One common postoperative complication that is associated with an increased risk of death is acute kidney injury (AKI).^{2–6} In one retrospective study, postoperative AKI occurred in 39% of hospitalized surgical patients who had their creatinine checked after surgery.⁷ In-hospital mortality increased from 0.6% among those without AKI to 8.8% among those who experienced AKI.

Some postoperative deaths and AKI may be prevented through identification and modification of risk factors. Although some risk factors (e.g., age^{8–10}) are not modifiable and other risk factors (e.g., preoperative glycemic control^{8,11,12}) are no longer modifiable once surgery has begun, several important risk factors are under the anesthesiologist's control during surgery. For example, intraoperative hypotension is a well-established risk factor both for postoperative death^{13–15} and for postoperative AKI.^{15–17} Interventions on such risk factors may therefore affect outcomes.^{18,19} Appropriate adjustments to the postoperative care plan may also impact outcomes.¹⁹

Telemedicine has successfully improved mortality and other outcomes in the intensive care unit (ICU), and similar results may be expected in the operating room. In a stepped-wedge trial across units in a single medical center, tele-ICU implementation was associated with an absolute mortality reduction of nearly 2%.²⁰ In a subsequent multi-center pre-post study, tele-ICU was associated with reduced in-hospital mortality and length of stay.²¹ Improved outcomes were associated with enhanced compliance with best clinical practice guidelines.^{20,21} The effect of telemedicine on outcomes may depend on characteristics of the patients in question—in one study, tele-ICU implementation improved mortality only among patients with higher illness severity scores.²²

The effectiveness of a telemedicine intervention depends on how well clinicians working in the telemedicine setting can assess patient risk and identify potential interventions to reduce risk. Accurate intraoperative risk assessment by clinicians is challenging for the following reasons. First, the sheer volume of data available is more than the human brain can comprehend with its limited data processing capacity.^{23,24} Competing clinical demands reduce the cognitive capacity available to process these data.^{25,26} Second, anesthesiology clinicians frequently use the available data to make decisions that are not based on sound logic.²⁷ Clinicians may anchor on the first diagnosis that comes to mind, only seek out data that confirm a previously suspected diagnosis, or interpret ambiguous findings with a false sense of certainty.²⁷ These concerns become increasingly significant as clinicians monitor increasing numbers of patients and have less time to review each patient's information.

Clinicians in telemedicine settings may use machine learning (ML) tools to address their innate cognitive deficiencies. Computers can process larger quantities of data more quickly than a human, model more complex relationships and interactions among inputs, and will not fatigue.^{28–30} The ability of clinicians to integrate the output of ML tools into their overall assessment of the patient depends on the ML output being presented in a manner that makes sense to the clinician and caters to the clinician's information needs.^{31,32} These needs will vary depending on the clinician's background and the context in which the clinician is working.

Preliminary data

This study is possible because the investigators have previously developed ML algorithms for predicting postoperative death and AKI. They used a retrospective cohort of approximately 110,000 patients who underwent surgical procedures at Barnes-Jewish Hospital between 2012 and 2016.³³ This dataset was divided into a training set (for model parameter tuning), a validation set (for hyperparameter tuning), and a test set (for quantifying model performance) in a ratio of approximately 7:1:2. Inputs to the models included patient characteristics, health conditions, preoperative vital signs and laboratory values, and intraoperative vital signs and medications. The resulting model predicted postoperative death with excellent accuracy (area under the receiver operating characteristic curve of 0.91, 95% confidence interval

0.90-0.92).^{34,35} A separate model predicted AKI with good accuracy (area under the receiver operating characteristic curve of 0.82, 95% confidence interval 0.81-0.84).³⁶

In addition, the investigators have recently conducted a series of focus groups and user interviews with anesthesiology clinicians (unpublished data) to learn about their workflows and information needs when working in the intraoperative telemedicine unit at Barnes-Jewish Hospital. Based on these insights, the investigators have designed a display interface that shows ML predictions for postoperative death and AKI in real-time during surgery.

Objective

To determine whether anesthesiology clinicians in a telemedicine setting can predict postoperative death and AKI more accurately with ML assistance than without ML assistance. The hypothesis is that clinician predictions will be more accurate with ML assistance than without ML assistance.

Overall study design

The Perioperative Outcome Risk Assessment with Computer Learning Enhancement (Periop ORACLE) study will be a sub-study nested within the ongoing TECTONICS trial (NCT03923699). TECTONICS is a single-center randomized clinical trial assessing the impact of an anesthesiology control tower (ACT) on postoperative 30-day mortality, delirium, respiratory failure, and acute kidney injury.³⁷ As part of the TECTONICS trial, clinicians in the ACT perform medical record case reviews during the early part of surgery and document how likely they feel each patient is to experience postoperative death and AKI. In Periop ORACLE, these case reviews will be randomized to be performed with or without access to ML predictions.

Methods: Participants, interventions, and outcomes

Study setting

The study will be conducted at Barnes-Jewish Hospital, a 1,252-bed university-affiliated tertiary care facility in St. Louis, MO. About 19,000 inpatient surgeries are performed in the hospital's 58 operating rooms each year.

Eligibility criteria

The participants will include all patients enrolled in the TECTONICS trial during the 12-month sub-study period for whom the ACT clinicians conduct a case review. The inclusion criteria for TECTONICS include (1) surgery in the main operating suite at Barnes-Jewish Hospital, (2) surgery during hours of ACT operation (weekdays 7:00am-4:00pm), and (3) age \geq 18. Exclusion criteria include procedures performed without anesthesiology services.

Recruitment

Participants are currently enrolled in TECTONICS via a waiver of consent. The investigators have obtained a waiver of informed consent to include these patients in Periop ORACLE as well.

Interventions – Machine learning algorithms

The machine learning models used in this study were originally trained and validated on a retrospective cohort of approximately 110,000 adult patients who underwent surgery with general anesthesia at Barnes-Jewish Hospital between 2012 and 2016. Input features included demographic characteristics, comorbid conditions, preoperative vital signs, surgical service, functional capacity as documented during the preoperative assessment, and most recent values of selected laboratory tests. A random forest model was implemented in scikit-learn. In the holdout validation cohort of 21,171 patients, the incidence of postoperative death was 2.2% and the model predicted this outcome with receiver operating characteristic area under curve (AUC) of 0.939 and precision-recall AUC of 0.161. The incidence of postoperative AKI was 6.1% and the model predicted this outcome with receiver operating characteristic AUC of 0.799 and precision-recall AUC of 0.275.

In February 2022, the models were retrained using a newer cohort of 84,455 patients who underwent surgery with general anesthesia at Barnes-Jewish Hospital between 2018 and 2020. This time period was after the hospital had transitioned from its previous electronic health record systems (including MetaVision as its anesthesia information management system) to Epic (Epic, Verona, WI). Input features were the same as the previous models, with the addition of the planned surgical procedure text field. A regularized logistic regression was used to predict the outcome from the words in the planned surgical procedure text field. This was used to initialize a gradient boosted decision tree, which was trained using the remaining features. Hyperparameters were selected by 10-fold cross validation. Models were implemented in XGBoost. In the holdout validation cohort of 16,891 patients, the incidence of postoperative death was 1.9% and the model predicted this outcome with receiver operating characteristic AUC of 0.91 and precision-recall AUC of 0.27. The incidence of postoperative AKI was 13.3% and the model predicted this outcome with receiver operating characteristic AUC of 0.90 and precision-recall AUC of 0.66.

A password-protected web application on a secure server has been created for delivery of machine learning algorithm outputs to clinicians. The design of this web application was informed by input obtained by users (clinicians from the ACT) in focus group sessions. For each patient, the machine learning predictions are presented in the form of predicted probabilities of each outcome (e.g., 4.5% chance of AKI). In addition, a list of features contributing most to the prediction is shown, along with Shapley values. During the focus group meetings, most users said they would find it overwhelming to see confidence intervals around the predicted probabilities. A fact sheet about each model is available on demand, including the receiver operating characteristic curve, precision-recall curve, and calibration curve.

Interventions – Implementation and workflow

Clinicians in the ACT currently conduct case reviews by viewing the patient's records in AlertWatch (AlertWatch, Ann Arbor, MI) and Epic. AlertWatch is an FDA-approved patient monitoring system designed for use in the operating room. Clinicians in the ACT use a customized version of AlertWatch that has been adapted for use in a telemedicine setting.³⁸ Epic is the electronic health record system utilized at Barnes-Jewish Hospital. For patients included in Periop ORACLE, each case review will be randomized in a 1:1 fashion to be completed with or without ML assistance. If the case review is randomized to ML assistance, the clinician will access the machine learning web application (see previous section) during the case review. Study staff will be immediately available in the ACT during all case reviews to assist with accessing the display interface if needed to improve adherence to the protocol. If the case review is not randomized to ML assistance, the clinician will not access this web application. This choice of comparator mimics current practice more closely than using an active comparator. After viewing the patient's data, the clinician will complete a case review from in AlertWatch (as described in the data collection section later in this document).

Outcomes

The co-primary outcomes will be clinician accuracy in predicting postoperative death and clinician accuracy in predicting postoperative AKI. Clinician predictions will be retrieved from the case review forms completed in AlertWatch. Observed death and observed AKI will be retrieved from Epic. Death will be defined as 30-day in-hospital mortality. AKI will be defined as a creatinine increase ≥ 0.3 mg/dl above baseline within 48 hours or an increase to ≥ 1.5 times baseline within seven days, consistent with the kidney disease: improving global outcomes definition.³⁹ If no preoperative creatinine is available, then the upper limit of the laboratory's reference range (1.2 mg/dl) will be used as the baseline.

Secondary outcomes will include AKI stage 2 or greater (creatinine increase to ≥ 2 times baseline within seven days) and AKI stage 3 (creatinine increase to ≥ 3 times baseline within seven days, an increase to ≥ 4.0 mg/dl, or initiation of renal replacement therapy).

Participant timeline

No direct interactions between study staff and the participants are planned, and no anesthetic interventions will be prohibited based on participation in this trial. The initial medical record review and clinician predictions will occur during the participants' surgery. Additional data retrieval to obtain observed death and AKI will occur at least 30 days after surgery (and may occur in bulk 30 days after the final participant's surgery).

Sample size

The sample size calculation is based on the assumption that ML-assisted clinicians would predict each outcome with receiver operating curve area under curve (AUC) similar to the published AUC of the ML algorithms.^{34,36} Incidences of death and AKI were also taken from these previous publications. A simulation population of 100,000 patients was generated. ML-assisted and -unassisted clinician predictions were simulated with beta distributions whose parameters were adjusted to achieve the specified AUC. For each sample size tested, 1,000 random samples were drawn and the difference in AUC between the assisted and unassisted clinician predictions was determined.⁴⁰ Power at each sample size was defined as the fraction of samples for which the two sets of predictions had significantly different AUC at $\alpha = 0.025$. The minimum clinically meaningful difference (MCMD) in AUC was defined as 0.07.

As shown in [Figure 1](#), a sample size of 4,500 will provide 80% power to detect a difference in AUC from 0.91 to 0.84 (the MCMD) for death and 95% power to detect a difference from 0.91 to 0.81. This sample size will give >99% power to detect a difference in AUC from 0.82 to 0.75 (the MCMD) for AKI ([Figure 2](#)).

To allow for 15% missing data, we will enroll 5,300 cases. Currently, about 20 case reviews are performed in the ACT on a typical day. We should therefore be able to complete enrollment over a period of 12 months.

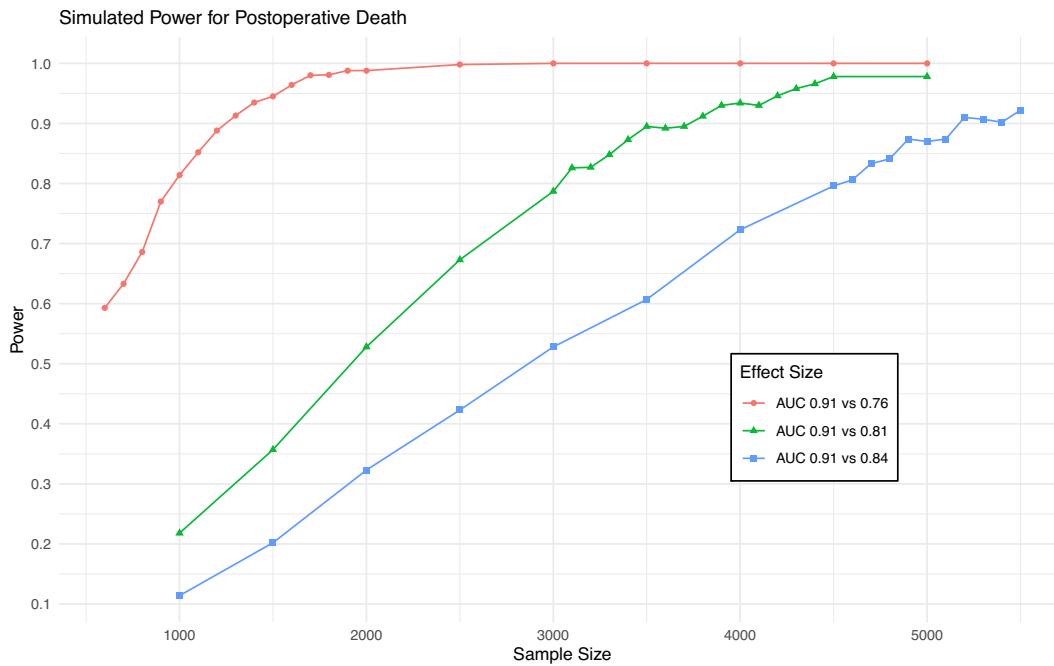


Figure 1. Simulation results from power calculation for death. Power achieved in simulations of various sample sizes and effect sizes for postoperative death. Blue line is minimum clinically meaningful difference.



Figure 2. Simulation results from power calculation for AKI. Power achieved in simulations of various sample sizes and effect sizes for postoperative AKI. Blue line is minimum clinically meaningful difference.

Methods: Assignment of interventions

Allocation

Each participant case review will be randomized in a 1:1 fashion to be completed with or without ML assistance. Randomization will be stratified by intervention/control status of the parent trial because clinicians may be biased to perform case reviews more carefully in the TECTONICS intervention group than in the TECTONICS control group (Figure 3). The allocation sequence will be generated by computer-generated random numbers within the AlertWatch

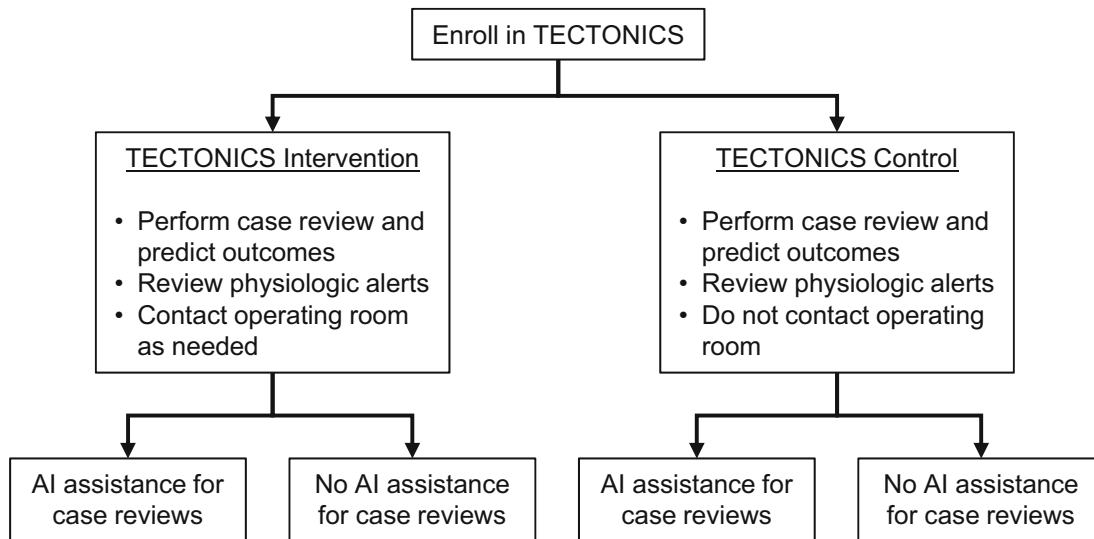


Figure 3. Flow chart showing treatment allocation. Periop ORACLE participants will be randomized to ML assistance or no ML assistance, stratified by intervention or control status of the parent TECTONICS trial.

software. The allocation will be automatically displayed on the case review form when the clinician opens it. The study staff will not have access to the allocation sequence before the case review.

Blinding

By necessity, the clinician completing the case review and the other study staff in the ACT will not be blinded to treatment allocation. Study personnel who retrieve observed postoperative complications from the electronic health record will be blinded.

Methods: Data collection, management, and analysis

Data collection

Clinician predictions will be collected via an existing case review form in AlertWatch that contains two sections (Figure 4). In the first section, the clinician documents how likely the patient is to experience postoperative death within 30 days, postoperative AKI within seven days, and a few other complications. The clinician selects their choice from a five-point ordered categorical scale (very low risk, low risk, average risk, high risk, and very high risk). In the second section (which is part of the parent TECTONICS trial but not used for this sub-study), the clinician selects treatment recommendations (e.g., blood pressure goals, medications to administer or avoid). An additional question asks the clinician to confirm whether they reviewed the ML display interface prior to documenting their predictions in section one. Finally, clinicians are asked whether the ML predictions on the display interface agree with their previous opinion and whether the display interface caused the clinician to change their opinion.

Observed death and observed AKI will be retrieved from Epic via a query of the Clarity database. Observed outcomes will be retrieved for all randomized participants, including those with protocol deviations.

Data management

The clinical applications used in this project (AlertWatch and Epic) can only be accessed over the secure institutional internet network or using virtual private network, and user authentication is required for access. Computations for the ML models will occur on institutional servers that meet the standards of the Health Insurance Portability and Accountability Act. The ML display interface is also hosted on an institutional server, can only be accessed over the secure institutional internet network or using virtual private network, and requires user authentication for access. For data analysis, the amount of protected health information retrieved will be limited to the minimum amount necessary to achieve the aims of the study. All study data will be stored on institutional servers, and access will be limited to study personnel.

Statistical methods

To compare the accuracy of clinician predictions with and without ML assistance, the investigators will construct logistic regressions for death and for AKI. Separate models will be constructed for the ML-assisted and -unassisted groups. The only inputs will be dummy variables encoding the clinician predictions (Table 1). The regression coefficients will be restricted to positive values, thus modeling a monotonic increasing relationship between the clinician predictions and the

Complication Assessment					
Minor	At Risk	High Risk	Major	At Risk	High Risk
Awareness under GA	<input type="checkbox"/>	<input type="checkbox"/>	CHF (acute)	<input type="checkbox"/>	<input type="checkbox"/>
Delirium	<input type="checkbox"/>	<input type="checkbox"/>	Hyperglycemia/Hypoglycemia	<input type="checkbox"/>	<input type="checkbox"/>
Hypothermia	<input type="checkbox"/>	<input type="checkbox"/>	Kidney failure	<input type="checkbox"/>	<input type="checkbox"/>
PONV	<input type="checkbox"/>	<input type="checkbox"/>	Myocardial ischemia	<input type="checkbox"/>	<input type="checkbox"/>
Refractory post-op pain	<input type="checkbox"/>	<input type="checkbox"/>	Respiratory failure	<input type="checkbox"/>	<input type="checkbox"/>
Venous thromboembolism	<input type="checkbox"/>	<input type="checkbox"/>	Stroke	<input type="checkbox"/>	<input type="checkbox"/>
Wound infection	<input type="checkbox"/>	<input type="checkbox"/>	Bleeding	<input type="checkbox"/>	<input type="checkbox"/>
Other (enter)	<input type="checkbox"/>	<input type="checkbox"/>	Other (enter)	<input type="checkbox"/>	<input type="checkbox"/>

Post Operative Predictions (not included in summary)	
DO NOT USE machine learning display when making predictions	
(Mortality within 30 days)	▼
(Mechanical ventilation for > 48 hours or reintubation within 48 hours.)	▼
(Acute kidney injury (creatinine increase of 0.3 mg/dL within 48 hr or increase of 50% within 7 days))	▼
(Did you review the machine learning display when making these predictions?)	▼
((If answered yes to the previous question): How did the machine learning prediction compare with what you thought before looking at it?)	▼

Recommendations	
Heart	PostOp
(MAP monitoring) ▼	(Post-op ICU) ▼
<input type="checkbox"/> Maintain higher MAP due to head up	<input type="checkbox"/> Consider post-op ventilation
<input type="checkbox"/> High BP for spinal cord with pertinent aortic surgery	<input type="checkbox"/> OSA risk orders
(HR monitoring) ▼	<input type="checkbox"/> Consider CPAP/BiPAP (other than OSA)
(Volume monitoring) ▼	Diabetes
(Inotrope administration) ▼	(Glucose monitoring) ▼
(Vasopressor) ▼	<input type="checkbox"/> Initiate insulin infusion if glucose > 180
(Patient has Pacemaker) ▼	<input type="checkbox"/> Infusion or int/long-acting insulin for glucose >180
(Patient has Defibrillator) ▼	<input type="checkbox"/> Always maintain basal insulin for type 1 DM or infusion
Blood	Other
<input type="checkbox"/> Transfuse below Hgb 7.0	<input type="checkbox"/> Hemodynamically significant AS management
<input type="checkbox"/> Consider Hgb during case due to anticipated anemia risk	<input type="checkbox"/> HOCM or HOCM risk management
<input type="checkbox"/> Consider Hgb due to previous anemia	<input type="checkbox"/> Craniotomy ventilation management
<input type="checkbox"/> Consider platelet count during case	<input type="checkbox"/> Proactive warming due to patient hypothermia risk
<input type="checkbox"/> Consider platelet count due to previous thrombocytopenia	<input type="checkbox"/> Assess and document TOF
<input type="checkbox"/> Consider INR and PTT during case	<input type="checkbox"/> Ensure reversal of NMBD due to patient risk
<input type="checkbox"/> Consider INR due to previous coagulopathy	<input type="checkbox"/> Consider ETAC alarm
PONV Prophylaxis	<input type="checkbox"/> Consider EEG/BIS
(At low risk for PONV) ▼	<input type="checkbox"/> Pharmacologic DVT prophylaxis.
(At moderate/high risk for PONV) ▼	<input type="checkbox"/> Redose antibiotics when due
<input type="checkbox"/> Consider Scopolamine Patch	<input type="checkbox"/> Redose antibiotics when due for CPB case
<input type="checkbox"/> Consider 10ml/KG Fluid	<input type="checkbox"/> Consider multimodal pain regimen.
<input type="checkbox"/> Consider using non-opioid analgesics or adjunctive analgesics	<input type="checkbox"/> Fresh gas flow rates
	<input type="checkbox"/> Other (enter)

Figure 4. Case review form in AlertWatch. The first two sections contain fields for documentation of clinician predictions of postoperative complications, and the Periop ORACLE randomization allocation is disclosed in the header of the second section. The treatment recommendations documented in the third section are utilized for the parent TECTONICS trial but not for this sub-study. The amount of time the clinician in the ACT spends completing each case review will be retrieved from the Epic audit log.

Table 1. Dummy variable encoding for clinician predicted risk of postoperative death.

Clinician prediction	Var1	Var2	Var3	Var4
Very low	0	0	0	0
Low	1	0	0	0
Average	1	1	0	0
High	1	1	1	0
Very high	1	1	1	1

true incidence of death and AKI. The AUC of the model constructed from ML-assisted cases will be compared to the AUC of the model constructed from ML-unassisted cases,⁴⁰ using the Holm method to ensure the family-wise error rate remains less than a two-sided $\alpha = 0.05$ across the two co-primary outcomes—accuracy of death prediction and accuracy of AKI prediction. The null hypothesis is that the AUCs will be equal.

The primary analysis will follow an intention-to-treat principle, and all case reviews will be included in the group to which they were randomized. Patients on dialysis preoperatively, undergoing dialysis access procedures, or with baseline creatinine > 4.0 mg/dl will be excluded from the AKI analysis but included in the death analysis. In a secondary per-protocol analysis, case reviews will be grouped according to whether the clinician reported viewing the ML display or not. Exploratory analyses stratified by sex and race will evaluate for biases learned during training.⁴¹ If either clinician predictions or observed outcomes are missing for a given case, then the case will be excluded from the analysis. No interim analyses are planned. To determine whether different levels of clinician engagement in the TECTONICS intervention group versus the TECTONICS control group impact the findings, sensitivity analyses will be conducted stratified by TECTONICS intervention status. To examine the effects of dataset shift and model retraining, the prospective performance of the models will be reported for each month of the study, and sensitivity analyses will be conducted in the subgroups who had surgery before and after the February 2022 retraining event.

To examine human-computer agreement, the proportion of cases for which the clinician reported being surprised by the ML prediction will be determined. Among those cases for which the clinician was surprised, the proportion for which the clinician self-reported agreeing or disagreeing with the ML prediction will be determined.

To examine the potential impact of inaccurate ML predictions, we will conduct the following sensitivity analysis. First, the predicted probabilities output by the ML algorithms will be converted to dichotomous predictions by using the cutoff value that maximizes the Youden index. Second, the cases will be categorized as having correct or incorrect ML dichotomized predictions. Finally, the primary analysis will be repeated in the subgroup with correct ML predictions and in the subgroup with incorrect ML predictions.

To estimate the effect of the ML predictions on case review efficiency, chart review duration (approximated using the time the Epic chart was open, retrieved from the audit log) will be compared between case reviews performed with ML assistance and those performed without ML assistance, using either an unpaired T test or Wilcoxon rank sum test as appropriate.

Ethics and dissemination

Ethical statement

This study has been approved by the Human Research Protection Office at Washington University in St. Louis (approval #202108022) on August 26, 2021. Any protocol amendments will be approved by the study steering committee and communicated with the institutional review board. This study presents patients with minimal risks, other than a small risk for a breach of confidentiality if protected health information were to become unintentionally available to individuals outside the study team. To protect against this risk, all electronic data will be kept in an encrypted, password-protected environment accessible only to the research team (see Data Management section). Because the risks are minimal, no dedicated safety monitoring committee is planned for Periop ORACLE. However, the parent TECTONICS trial does have a safety monitoring committee. The institutional review board has granted a waiver of informed consent to enroll patients in this study.

Study results will be presented at national or international scientific meetings and published in a peer-reviewed publication. Individuals who meet the International Committee of Medical Journal Editors authorship guidelines (<https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html>) will be included as authors on any publications resulting from this work. To comply with data sharing recommendations, de-identified individual participant data underlying the study results will be made available to researchers who provide a methodologically sound proposal for utilizing that data.

Strengths, limitations, and alternative strategies

This project has multiple strengths. First, this protocol has been prepared in accordance with Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) guidelines.^{42,43} Second, the sample size of predictions made by experienced anesthesia clinicians will be large. The pragmatic design of superimposing Periop ORACLE on the established TECTONICS trial where many case reviews are already being conducted on a daily basis makes it feasible to achieve such a large sample size. Third, the ML models to be used have very good AUC, which enhances the likelihood that ML assistance can increase the accuracy of clinician predictions. Fourth, the ML display interface has been created

using a human-centered design framework informed by the clinicians who work in the ACT, which maximizes the chances that clinicians will integrate the ML display interface into their workflows and their decision-making.

This project also has limitations. First, data collection in a telemedicine setting rather than in the operating room may limit the generalizability of the findings to institutions that do not utilize intraoperative telemedicine. However, multiple clinicians have stated during focus group interviews that their workflows for performing case reviews in the ACT closely mimic their workflows for preparing to provide bedside anesthesiology care. Thus, the findings may be relevant in the operating room as well. Second, this is a single-center study, so differences in patient population or practice patterns at other institutions may limit generalizability. However, many large academic medical centers care for patient populations similar to those seen at Barnes-Jewish Hospital. Third, some of the outcomes may be incompletely measured. While in-hospital vital status should always be available, some patients may not have a postoperative creatinine measurement available to distinguish whether AKI has occurred. However, AKI is expected to be extremely uncommon among patients without postoperative labs, minimizing the effect of this potential bias. Fourth, AKI is defined using creatinine only and not urine output, based on anticipated data availability. This may cause some cases of AKI to be missed. However, during ML model training, the incidence of AKI using this definition was compatible with the incidence of AKI reported in other studies. Finally, clinicians may give the ML display interface varying degrees of weight during case reviews randomized to ML assistance. Understanding this variability is of interest to the research team, which is why the case review form will ask the clinician how the ML display impacted their decision-making.

The expected result is that prediction accuracy for death and for AKI to be greater with ML assistance than without ML assistance. If the null hypothesis is rejected for only one of the two co-primary outcomes, this result will still be viewed as evidence that the ML assistance is beneficial. A possible unintended consequence would be if false negative ML predictions lead telemedicine clinicians to pay less attention to some patients who are actually at high risk for death or AKI.²² Even if ACT clinicians monitor these patients less intensely, these patients will still receive standard-of-care monitoring and care by clinicians in the operating rooms. Thus, the reduction in telemedicine monitoring should not result in patient harm. The telemedicine clinicians supplement, but do not replace, standard care.

Conclusion

Intraoperative telemedicine has the potential to improve postoperative outcomes if interventions can be targeted to patients most at-risk for complications, and ML may be able to help clinicians distinguish which patients those are. Periop ORACLE will test the hypothesis that anesthesiology clinicians in a telemedicine setting can predict postoperative complications more accurately with ML assistance than without ML assistance. By nesting this study within the ongoing TECTONICS randomized clinical trial of intraoperative telemedicine, the investigators will efficiently assemble a large dataset of clinician predictions that will be used to achieve the study objective. Successful completion of this study will help define the role of ML not only for intraoperative telemedicine, but for other risk assessment tasks before, during, and after surgery.

Data availability

No data are associated with this article.

Reporting guidelines

This protocol is presented according to the SPIRIT guidelines.

Open Science Framework: "Protocol for the Perioperative Outcome Risk Assessment with Computer Learning Enhancement (Periop ORACLE) Randomized Study". <https://doi.org/10.17605/OSF.IO/GC4ES>.⁴³

Data are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

References

1. Nepogodiev D, Martin J, Biccard B, et al.: **Global burden of postoperative death.** *Lancet* 2019; **393**(10170): 401. [PubMed Abstract](#) | [Publisher Full Text](#)
2. Kheterpal S, Tremper KK, Englesbe MJ, et al.: **Predictors of postoperative acute renal failure after noncardiac surgery in patients with previously normal renal function.** *Anesthesiology*

- 2007; **107**(6): 892–902.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Kheterpal S, Tremper KK, Heung M, *et al.*: **Development and validation of an acute kidney injury risk index for patients undergoing general surgery: results from a national data set.** *Anesthesiology* 2009; **110**(3): 505–515.
[PubMed Abstract](#) | [Publisher Full Text](#)
 4. Engoren M, Habib RH, Arslanian-Engoren C, *et al.*: **The effect of acute kidney injury and discharge creatinine level on mortality following cardiac surgery.** *Crit. Care Med.* 2014; **42**(9): 2069–2074.
[PubMed Abstract](#) | [Publisher Full Text](#)
 5. Rydén L, Ahnve S, Bell M, *et al.*: **Acute kidney injury after coronary artery bypass grafting and long-term risk of myocardial infarction and death.** *Int. J. Cardiol.* 2014; **172**(1): 190–195.
[PubMed Abstract](#) | [Publisher Full Text](#)
 6. Dardashti A, Ederoth P, Algotsson L, *et al.*: **Incidence, dynamics, and prognostic value of acute kidney injury for death after cardiac surgery.** *J. Thorac. Cardiovasc. Surg.* 2014; **147**(2): 800–807.
[PubMed Abstract](#) | [Publisher Full Text](#)
 7. Hobson C, Ozragat-Baslanti T, Kuxhausen A, *et al.*: **Cost and mortality associated with postoperative acute kidney injury.** *Ann. Surg.* 2015; **261**(6): 1207–1214.
[PubMed Abstract](#) | [Publisher Full Text](#)
 8. Palomba H, De Castro I, Neto A, *et al.*: **Acute kidney injury prediction following elective cardiac surgery: AKICS Score.** *Kidney Int.* 2007; **72**(5): 624–631.
[PubMed Abstract](#) | [Publisher Full Text](#)
 9. Abelha FJ, Botelho M, Fernandes V, *et al.*: **Determinants of postoperative acute kidney injury.** *Crit. Care* 2009; **13**(3): R79.
[PubMed Abstract](#) | [Publisher Full Text](#)
 10. van Gestel YR, Lemmens VE, de Hingh IH, *et al.*: **Influence of comorbidity and age on 1-, 2-, and 3-month postoperative mortality rates in gastrointestinal cancer patients.** *Ann. Surg. Oncol.* 2013; **20**(2): 371–380.
[Publisher Full Text](#)
 11. Chrastil J, Anderson MB, Stevens V, *et al.*: **Is hemoglobin A1c or perioperative hyperglycemia predictive of periprosthetic joint infection or death following primary total joint arthroplasty?** *J. Arthroplast.* 2015; **30**(7): 1197–1202.
[Publisher Full Text](#)
 12. Halkos ME, Lattouf OM, Puskas JD, *et al.*: **Elevated preoperative hemoglobin A1c level is associated with reduced long-term survival after coronary artery bypass surgery.** *Ann. Thorac. Surg.* 2008; **86**(5): 1431–1437.
[PubMed Abstract](#) | [Publisher Full Text](#)
 13. Monk TG, Saini V, Weldon BC, *et al.*: **Anesthetic management and one-year mortality after noncardiac surgery.** *Anesth. Analg.* 2005; **100**(1): 4–10.
[PubMed Abstract](#) | [Publisher Full Text](#)
 14. Monk TG, Bronsart MR, Henderson WG, *et al.*: **Association between intraoperative hypotension and hypertension and 30-day postoperative mortality in noncardiac surgery.** *Anesthesiology* 2015; **123**(2): 307–319.
[PubMed Abstract](#) | [Publisher Full Text](#)
 15. Walsh M, Devereaux PJ, Garg AX, *et al.*: **Relationship between intraoperative mean arterial pressure and clinical outcomes after noncardiac surgery: toward an empirical definition of hypotension.** *Anesthesiology* 2013; **119**(3): 507–515.
[PubMed Abstract](#) | [Publisher Full Text](#)
 16. Sun LY, Wijeyesundera DN, Tait GA, *et al.*: **Association of intraoperative hypotension with acute kidney injury after elective noncardiac surgery.** *Anesthesiology* 2015; **123**(3): 515–523.
[PubMed Abstract](#) | [Publisher Full Text](#)
 17. Salmasi V, Maheshwari K, Yang D, *et al.*: **Relationship between intraoperative hypotension, defined by either reduction from baseline or absolute thresholds, and acute kidney and myocardial injury after noncardiac Surgery: A retrospective cohort analysis.** *Anesthesiology* 2017; **126**(1): 47–65.
[PubMed Abstract](#) | [Publisher Full Text](#)
 18. Futier E, Lefrant J-Y, Guinot P-G, *et al.*: **Effect of individualized vs standard blood pressure management strategies on postoperative organ dysfunction among high-risk patients undergoing major surgery: a randomized clinical trial.** *JAMA* 2017; **318**(14): 1346–1357.
[PubMed Abstract](#) | [Publisher Full Text](#)
 19. van den Boom W, Schroeder RA, Manning MW, *et al.*: **Effect of A1C and glucose on postoperative mortality in noncardiac and cardiac surgeries.** *Diabetes Care* 2018; **41**(4): 782–788.
[PubMed Abstract](#) | [Publisher Full Text](#)
 20. Lilly CM, Cody S, Zhao H, *et al.*: **Hospital mortality, length of stay, and preventable complications among critically ill patients before and after tele-ICU reengineering of critical care processes.** *JAMA* 2011; **305**(21): 2175–2183.
[PubMed Abstract](#) | [Publisher Full Text](#)
 21. Lilly CM, McLaughlin JM, Zhao H, *et al.*: **A multicenter study of ICU telemedicine reengineering of adult critical care.** *Chest* 2014; **145**(3): 500–507.
[PubMed Abstract](#) | [Publisher Full Text](#)
 22. Thomas EJ, Lucke JF, Wueste L, *et al.*: **Association of telemedicine for remote monitoring of intensive care patients with mortality, complications, and length of stay.** *JAMA* 2009; **302**(24): 2671–2678.
[Publisher Full Text](#)
 23. McDonald CJ: **Protocol-based computer reminders, the quality of care and the non-perfectability of man.** *N. Engl. J. Med.* 1976; **295**(24): 1351–1355.
[PubMed Abstract](#) | [Publisher Full Text](#)
 24. Buschman TJ, Siegel M, Roy JE, *et al.*: **Neural substrates of cognitive capacity limitations.** *Proc. Natl. Acad. Sci. U. S. A.* 2011; **108**(27): 11252–11255.
[PubMed Abstract](#) | [Publisher Full Text](#)
 25. Gaba DM, Lee T: **Measuring the workload of the anesthesiologist.** *Anesth. Analg.* 1990; **71**(4): 354–361.
[PubMed Abstract](#)
 26. Zenati MA, Leissner KB, Zorca S, *et al.*: **First reported use of team cognitive workload for root cause analysis in cardiac surgery.** *Semin. Thorac. Cardiovasc. Surg.* 2019; **31**(3): 394–396.
[PubMed Abstract](#) | [Publisher Full Text](#)
 27. Stiegler MP, Tung A: **Cognitive processes in anesthesiology decision making.** *Anesthesiology* 2014; **120**(1): 204–217.
[Publisher Full Text](#)
 28. Sinha A, Singh A, Tewari A: **The fatigued anesthesiologist: A threat to patient safety?** *J. Anaesthesiol. Clin. Pharmacol.* 2013; **29**(2): 151–159.
[PubMed Abstract](#) | [Publisher Full Text](#)
 29. Howard SK, Rosekind MR, Katz JD, *et al.*: **Fatigue in Anesthesia: Implications and Strategies for Patient and Provider Safety.** *Anesthesiology* 2002; **97**(5): 1281–1294.
[Publisher Full Text](#)
 30. Gander PH, Merry A, Millar MM, *et al.*: **Hours of Work and Fatigue-Related Error: A Survey of New Zealand Anaesthetists.** *Anaesth. Intensive Care* 2000; **28**(2): 178–183.
[PubMed Abstract](#) | [Publisher Full Text](#)
 31. Lintern G, Motavalli A: **Healthcare information systems: the cognitive challenge.** *BMC Med. Inform. Decis. Mak.* 2018; **18**(1): 3.
[PubMed Abstract](#) | [Publisher Full Text](#)
 32. Johnson CM, Johnson TR, Zhang J: **A user-centered framework for redesigning health care interfaces.** *J. Biomed. Inform.* 2005; **38**(1): 75–87.
[PubMed Abstract](#) | [Publisher Full Text](#)
 33. Fritz BA, Chen Y, Murray-Torres TM, *et al.*: **Using machine learning techniques to develop forecasting algorithms for postoperative complications: protocol for a retrospective study.** *BMJ Open* 2018; **8**(4): e020124.
[PubMed Abstract](#) | [Publisher Full Text](#)
 34. Fritz BA, Cui Z, Zhang M, *et al.*: **Deep-learning model for predicting 30-day postoperative mortality.** *Br. J. Anaesth.* 2019; **123**(5): 688–695.
[PubMed Abstract](#) | [Publisher Full Text](#)
 35. Fritz BA, Abdelhack M, King CR, *et al.*: **Update to 'Deep-learning model for predicting 30-day postoperative mortality' (Br J Anaesth 2019; 123: 688–95).** *Br. J. Anaesth.* 2020; **125**(2): e230–e231.
[PubMed Abstract](#) | [Publisher Full Text](#)
 36. Cui Z, Fritz BA, King CR, *et al.*: **A factored generalized additive model for clinical decision support in the operating room.** *AMIA Annu. Symp. Proc.* 2019; **2019**: 343–352.
 37. King CR, Abraham J, Kannampallil TG, *et al.*: **Protocol for the Effectiveness of an Anesthesiology Control Tower System in Improving Perioperative Quality Metrics and Clinical Outcomes: the TECTONICS randomized, pragmatic trial.** *F1000Res* 2019; **8**: 2032.
[PubMed Abstract](#) | [Publisher Full Text](#)
 38. Murray-Torres T, Casarella A, Bollini M, *et al.*: **Anesthesiology Control Tower—Feasibility Assessment to Support Translation (ACTFAST): Mixed-Methods Study of a Novel Telemedicine-Based Support System for the Operating Room.** *JMIR Hum. Factors* 2019; **6**(2): e12155.
[PubMed Abstract](#) | [Publisher Full Text](#)
 39. Kellum JA, Lameire N: **Diagnosis, evaluation, and management of acute kidney injury: a KDIGO summary (Part 1).** *Crit. Care* 2013; **17**(1): 204.
[PubMed Abstract](#) | [Publisher Full Text](#)
 40. DeLong ER, DeLong DM, Clarke-Pearson DL: **Comparing the Areas under Two or More Correlated Receiver Operating**

Characteristic Curves: A Nonparametric Approach. *Biometrics* 1988; **44**(3): 837–845.
[Publisher Full Text](#)

41. Zou J, Schiebinger L: **AI can be sexist and racist—it's time to make it fair.** *Nature* 2018; **559**(7714): 324–326.
[PubMed Abstract](#) | [Publisher Full Text](#)
42. Chan A-W, Tetzlaff JM, Altman DG, *et al.*: **SPIRIT 2013 statement: defining standard protocol items for clinical trials.** *Ann. Intern.*

Med. 2013; **158**(3): 200–207.

[PubMed Abstract](#) | [Publisher Full Text](#)

43. Fritz B: **Perioperative Outcome Risk Assessment with Computer Learning Enhancement (Periop ORACLE) Randomized Study.** 2022, May 19.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:  

Version 2

Reviewer Report 02 August 2023

<https://doi.org/10.5256/f1000research.138973.r177469>

© 2023 Yin P et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Pengbin Yin

Chinese PLA General Hospital, Beijing, China

Ming Chen

Department of Orthopedics, Chinese PLA General Hospital, Beijing, China

Yi Li

Department of Orthopedics, Chinese PLA General Hospital, Beijing, China

This protocol aimed to clarify the accuracy of machine-learning based prediction models over anesthesiology clinicians in assessing the risk of 30-day in-hospital mortality and AKI in adult surgical patients. This is a well-design protocol and of great clinical significance. However, in order to improve the manuscript's accessibility, readability, and overall quality, a few concerns are raised.

1. Has this prediction model previously been tested for predictive performance in external validation?

2. You mentioned the clinician selects their choice from a five-point ordered categorical scale (very low risk, low risk, average risk, high risk, and very high risk). What are the definitions of the five categories? Were all clinicians involved in the case review of this trial well-trained? Do these evaluation criteria remain consistent during all the practice process?

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Clinical and basic research with focus on extracellular vesicles-related mechanisms discovery, targeted therapy development, biomarker identification, and prognosis improvement for ageing-related musculoskeletal conditions including fractures, osteoporosis, sarcopenia, etc.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 06 October 2022

<https://doi.org/10.5256/f1000research.138973.r151957>

© 2022 Mathis M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Michael Robert Mathis 

Department of Anesthesiology, University of Michigan Medical School, Ann Arbor, MI, USA

I have reviewed the revised manuscript and do not consider any additional revisions necessary. I approve the article for indexing.

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 22 July 2022

<https://doi.org/10.5256/f1000research.134258.r143767>

© 2022 Mathis M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Michael Robert Mathis 

Department of Anesthesiology, University of Michigan Medical School, Ann Arbor, MI, USA

This study is a protocol a sub-study nested within the TECTONICS RCT (assessing the impact of telemedicine approaches to preoperative risk assessment), seeking to understand the impact of machine-learning based prediction models to improve the accuracy of such preoperative risk assessments.

The study follows SPIRIT guidelines for reporting, and is overall well-written and of great clinical significance to anesthesiologists. Strengths of the study include its rigorous design; weaknesses include its single-center nature and perhaps limited exploration of unintended consequences of machine learning algorithm implementation.

Overall I believe this study will be an important scientific contribution to the field of anesthesiology, and look forward to its execution. However as the protocol is currently written, I have several critiques for the authors to address:

- 1) There limitations of using solely a creatinine-based AKI endpoint (as opposed to creatinine + UOP + dialysis), but on balance I recognize the choice to focus on simply creatinine, given it is widely available within the EHR. However, there are other unusual circumstances that the authors might consider, that may lead to misguided conclusions if left unaddressed:
 - a. How are patients with stage 4 or 5 CKD handled? Perhaps they are included, but it is misleading to think that AKI in such patients has anything to do with perioperative care, as most will develop AKI simply due to their native pathophysiology.
 - b. If there is an EGFR-based cut-off for inclusion, how is EGFR computed? Would recommend the 2021 CKD-EPI formula which addresses racial bias.
 - c. How are patients with no preoperative creatinine available handled?
 - d. How are patients undergoing procedures which by definition impair renal function handled? (e.g. renal artery embolization; nephrectomy for RCC, etc.).
 - e. Are dialysis access procedures (e.g. AV fistula placements) included in this study?
- 2) I agree from both a study power / and clinical significance standpoint in using AKI Stage 1 or greater as AKI outcome (and of course mortality is more challenging to power); however you might consider breaking down AKI by stages as a secondary outcome (e.g. Stage 2 or greater; Stage 3 or greater), given their greater clinical significance (although even Stage 1 is important).
- 3) It could be helpful to explore unintended consequences of the machine learning algorithm implementation. For example, workload/efficiency metrics may be useful to track: the authors may

wish to track the total amount of 'active review time' used by clinicians in performing preoperative assessments with and without aid of the machine learning algorithm. If implemented in clinical care, there should be a plan (perhaps covered by the DSMB of the TECTONICS trial) to understand unexpected outcomes (for example, increased risk of stroke or MI, at the expense of decreased risk of AKI).

4) As the AKI and mortality prediction algorithms are imperfect, do the clinicians performing preoperative risk assessments have insight into the performance characteristics of the algorithm (for example, do they know the algorithm's positive predictive value, negative predictive value; perhaps Shapley values, AUROC, AUPRC, etc.), in order to better understanding exactly how much they should be trusting it? It may be interesting to study the impact of the machine-learning based risk assessments, for the cases in which the algorithm was *wrong*... in order to understand the potential impact of automation bias induced by machine learning algorithms on clinical decision-making.

5) One of the major issues confronting ML-based algorithm support in healthcare decision-making is dataset shift – defined as an evolving mismatch between the data upon which the algorithm is trained upon, versus the data the algorithm is using in real-time to make predictions (PMID 34260843). Maintaining algorithm robustness over time remains very challenging, and decreases in algorithm performance can be seen whenever there are changes to practice patterns, EHR documentation patterns, or shifts in the patient population being studied. Although this study is limited by its single-center nature, it is not limited in its ability to assess temporal trends / algorithm de-tuning (and potential effects of algorithm re-tuning, if the authors wish to incorporate this into the study) over the time period studied. I would suggest assessing temporal trends in (i) AKI and mortality prediction algorithm performance, which if decreased may lead to a temporal trend in (ii) outcomes of this study.

6) Whereas the study protocol appropriately follows SPIRIT guidelines, when this study is executed and reported, this study will most likely need to follow DECIDE-AI guidelines (PMID 35585198), although I could be mistaken. The study team may wish to consider that downstream study now, and have the trial protocol written in such a way that the downstream study will be able to best adhere to DECIDE-AI reporting guidelines.

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine learning applied to healthcare; anesthesiology; prediction modeling; large observational database research; acute kidney injury prediction.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 21 Sep 2022

Bradley Fritz

Thank you for your thoughtful review of this protocol. We have revised the protocol to address your concerns, and we feel these changes have made the protocol stronger. Please see our point-by-point response below:

COMMENT 1a. How are patients with stage 4 or 5 CKD handled? Perhaps they are included, but it is misleading to think that AKI in such patients has anything to do with perioperative care, as most will develop AKI simply due to their native pathophysiology.

Patients who are on dialysis preoperatively and patients with baseline creatinine level > 4.0 mg/dl will be excluded from the analyses of AKI, but other patients with CKD will be included. We agree that AKI will be highly prevalent among patients with CKD. Although some (many?) cases of AKI in this population will not be preventable, other cases will be driven at least partly by modifiable risk factors.

All patients, including those on dialysis preoperatively and those with baseline creatinine > 4.0 mg/dl, will be included in the analyses of postoperative death. We have added a sentence to make this clearer:

"Patients on dialysis preoperatively, undergoing dialysis access procedures, or with baseline creatinine > 4.0 mg/dl will be excluded from the AKI analysis but included in the death analysis."
(Methods/Statistical Methods)

COMMENT 1b. If there is an EGFR-based cut-off for inclusion, how is EGFR computed? Would recommend the 2021 CKD-EPI formula which addresses racial bias.

We agree that historically common EGFR equations can introduce racial bias, and we appreciate the reviewer's suggestion for an alternative. As noted above, we will not use an EGFR-based cut-off for inclusion.

COMMENT 1c. How are patients with no preoperative creatinine available handled?

If no preoperative creatinine was available, then the upper limit of the laboratory's reference range (1.2 mg/dl) was used as the baseline. We felt this was a reasonable approximation because patients with abnormal renal function before surgery would most likely have a creatinine value available. We have added a sentence to make this clearer:

"If no preoperative creatinine is available, then the upper limit of the laboratory's reference range (1.2 mg/dl) will be used as the baseline." (Methods/Outcomes)

COMMENT 1d. How are patients undergoing procedures which by definition impair renal function handled? (e.g. renal artery embolization; nephrectomy for RCC, etc.).

These patients will be included in the AKI analyses. Although it may frequently not be possible to prevent AKI in these situations, it may be possible to lessen its severity. Therefore, a clinician would want to be aware of the elevated risk in these patients.

COMMENT 1e. Are dialysis access procedures (e.g. AV fistula placements) included in this study?

Patients undergoing dialysis access procedures will be excluded from the analyses of AKI. However, they will be included in the analyses of postoperative death. We have added a sentence to make this clearer:

"Patients on dialysis preoperatively, undergoing dialysis access procedures, or with baseline creatinine > 4.0 mg/dl will be excluded from the AKI analysis but included in the death analysis." (Methods/Statistical Methods)

COMMENT 2. I agree from both a study power / and clinical significance standpoint in using AKI Stage 1 or greater as AKI outcome (and of course mortality is more challenging to power); however you might consider breaking down AKI by stages as a secondary outcome (e.g. Stage 2 or greater; Stage 3 or greater), given their greater clinical significance (although even Stage 1 is important).

Thank you for this suggestion. We have added these as secondary outcomes:

"Secondary outcomes will include AKI stage 2 or greater (creatinine increase to ≥ 2 times baseline within seven days) and AKI stage 3 (creatinine increase to ≥ 3 times baseline within seven days, an increase to ≥ 4.0 mg/dl, or initiation of renal replacement therapy)." (Methods/Outcomes)

COMMENT 3. It could be helpful to explore unintended consequences of the machine learning algorithm implementation. For example, workload/efficiency metrics may be useful to track: the authors may wish to track the total amount of 'active review time' used by clinicians in performing preoperative assessments with and without aid of the machine learning algorithm. If implemented in clinical care, there should be a plan (perhaps covered by the DSMB of the TECTONICS trial) to understand unexpected outcomes (for example, increased risk of stroke or MI, at the expense of decreased risk of AKI).

We agree it would be insightful to investigate the time spent reviewing each case. If the machine learning algorithms have an impact on active review time, then the impact could

plausibly be in either direction: clinicians may spend less time in the patient's chart for ML-assisted cases because the algorithms have expedited the risk assessment process, or they may spend more time because the algorithms have pointed out concerns the clinicians otherwise would not have noticed. We have added this to the data collection and analysis plans:

"The amount of time the clinician in the ACT spends completing each case review will be retrieved from the Epic audit log." (Methods/Data collection)

"To estimate the effect of the ML predictions on case review efficiency, chart review duration (approximated using the time the Epic chart was open, retrieved from the audit log) will be compared between case reviews performed with ML assistance and those performed without ML assistance, using either an unpaired T test or Wilcoxon rank sum test as appropriate." (Methods/Statistical methods)

Information from the case reviews performed in ORACLE will only be communicated to bedside clinicians in the operating room if the patient is in the intervention arm of the parent TECTONICS trial. (All patients in ORACLE are also enrolled in TECTONICS.) The TECTONICS trial has a data safety monitoring committee that reviews adverse events.

COMMENT 4. As the AKI and mortality prediction algorithms are imperfect, do the clinicians performing preoperative risk assessments have insight into the performance characteristics of the algorithm (for example, do they know the algorithm's positive predictive value, negative predictive value; perhaps Shapley values, AUROC, AUPRC, etc.), in order to better understanding exactly how much they should be trusting it? It may be interesting to study the impact of the machine-learning based risk assessments, for the cases in which the algorithm was *wrong*... in order to understand the potential impact of automation bias induced by machine learning algorithms on clinical decision-making.

We have added a new paragraph to explain what pieces of meta-data are provided along with the predictions:

"For each patient, the machine learning predictions are presented in the form of predicted probabilities of each outcome (e.g., 4.5% chance of AKI). In addition, a list of features contributing most to the prediction is shown, along with Shapley values. During the focus group meetings, most users said they would find it overwhelming to see confidence intervals around the predicted probabilities. A fact sheet about each model is available on demand, including the receiver operating characteristic curve, precision-recall curve, and calibration curve." (Methods/Interventions – Machine Learning Algorithms)

To address the impact of "wrong" predictions, we have added a sensitivity subgroup analysis:

"To examine the potential impact of inaccurate ML predictions, we will conduct the following sensitivity analysis. First, the predicted probabilities output by the ML algorithms will be converted to dichotomous predictions by using the cutoff value that maximizes the Youden index.

Second, the cases will be categorized as having correct or incorrect ML dichotomized predictions. Finally, the primary analysis will be repeated in the subgroup with correct ML predictions and in the subgroup with incorrect ML predictions.” (Methods/Statistical methods)

COMMENT 5. One of the major issues confronting ML-based algorithm support in healthcare decision-making is dataset shift – defined as an evolving mismatch between the data upon which the algorithm is trained upon, versus the data the algorithm is using in real-time to make predictions (PMID 34260843). Maintaining algorithm robustness over time remains very challenging, and decreases in algorithm performance can be seen whenever there are changes to practice patterns, EHR documentation patterns, or shifts in the patient population being studied. Although this study is limited by its single-center nature, it is not limited in its ability to assess temporal trends / algorithm de-tuning (and potential effects of algorithm re-tuning, if the authors wish to incorporate this into the study) over the time period studied. I would suggest assessing temporal trends in (i) AKI and mortality prediction algorithm performance, which if decreased may lead to a temporal trend in (ii) outcomes of this study.

This is an important point that we had considered but failed to mention in the previous version of the protocol. At the time the study was initiated, we were utilizing models that had been trained on a retrospective cohort of patients who had surgery at Barnes-Jewish Hospital between 2012 and 2016. Approximately six months after study initiation, we retrained the models using patients who had surgery between 2018 and 2020. We have added two new paragraphs to the methods section to explain this clearly:

“The machine learning models used in this study were originally trained and validated on a retrospective cohort of approximately 110,000 adult patients who underwent surgery with general anesthesia at Barnes-Jewish Hospital between 2012 and 2016. Input features included demographic characteristics, comorbid conditions, preoperative vital signs, surgical service, functional capacity as documented during the preoperative assessment, and most recent values of selected laboratory tests. A random forest model was implemented in scikit-learn. In the holdout validation cohort of 21,171 patients, the incidence of postoperative death was 2.2% and the model predicted this outcome with receiver operating characteristic area under curve (AUC) of 0.939 and precision-recall AUC of 0.161. The incidence of postoperative AKI was 6.1% and the model predicted this outcome with receiver operating characteristic AUC of 0.799 and precision-recall AUC of 0.275.

In February 2022, the models were retrained using a newer cohort of 84,455 patients who underwent surgery with general anesthesia at Barnes-Jewish Hospital between 2018 and 2020. This time period was after the hospital had transitioned from its previous electronic health record systems (including MetaVision as its anesthesia information management system) to Epic (Epic, Verona, WI). Input features were the same as the previous models, with the addition of the planned surgical procedure text field. A regularized logistic regression was used to predict the outcome from the words in the planned surgical procedure text field. This was used to initialize a gradient boosted decision tree, which was trained using the remaining features. Hyperparameters were selected by 10-fold cross validation. Models were implemented in XGBoost. In the holdout validation cohort of 16,891 patients, the incidence of postoperative death

was 1.9% and the model predicted this outcome with receiver operating characteristic AUC of 0.91 and precision-recall AUC of 0.27. The incidence of postoperative AKI was 13.3% and the model predicted this outcome with receiver operating characteristic AUC of 0.90 and precision-recall AUC of 0.66.” (Methods/Interventions – Machine Learning Algorithms)

To examine the effects of dataset shift on model performance and trends over time, we will report the prospective performance (i.e., AUC) of each model for each month of the study. To examine the effects of model retraining, we will also conduct subgroup analyses repeating the primary analyses in the subgroups who had surgery before and after the retraining event. We have added these steps to the statistical methods:

“To examine the effects of dataset shift and model retraining, the prospective performance of the models will be reported for each month of the study, and sensitivity analyses will be conducted in the subgroups who had surgery before and after the February 2022 retraining event.”
(Methods/Statistical methods)

COMMENT 6. Whereas the study protocol appropriately follows SPIRIT guidelines, when this study is executed and reported, this study will most likely need to follow DECIDE-AI guidelines (PMID 35585198), although I could be mistaken. The study team may wish to consider that downstream study now, and have the trial protocol written in such a way that the downstream study will be able to best adhere to DECIDE-AI reporting guidelines.

Thank you for this suggestion—we agree the study results will be reported according to DECIDE-AI guidelines. To make it easier for the results manuscript to adhere to DECIDE-AI guidelines, we have expanded the methods section of this protocol to add details that are required by the DECIDE-AI guidelines. In particular, we have broken up the Intervention subsection into two new sections describing the AI system (checklist item 4) and describing the implementation (checklist item 5), which are mentioned in our responses to your earlier comments. In addition, we have added statistical methods that will allow us to report on human-computer agreement (checklist item 12):

“To examine human-computer agreement, the proportion of cases for which the clinician reported being surprised by the ML prediction will be determined. Among those cases for which the clinician was surprised, the proportion for which the clinician self-reported agreeing or disagreeing with the ML prediction will be determined.” (Methods/Statistical methods)

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research