Predicting prolonged apnea during nurse-administered procedural sedation using machine learning

Aaron Conway PhD1,2,3,✉, Carla Jungquist PhD4, Kristina Chang MScN1, Navpreet Kamboj BScN2, Joanna Sutherland MBBS5,6, Sebastian Mafeld7, and Matteo Parotto MD, PhD8,9

1 Peter Munk Cardiac Centre, University Health Network, Toronto, Canada  
2 Lawrence S. Bloomberg Faculty of Nursing, University of Toronto, Toronto, Canada  
3 School of Nursing, Queensland University of Technology (QUT), Brisbane, Australia  
4 The University of Buffalo, Buffalo, New York, USA  
5 Rural Clinical School, University of New South Wales, Coffs Harbour, NSW, Australia  
6 Department of Anaesthesia, Coffs Harbour Health Campus, Coffs Harbour, NSW, Australia  
7 Interventional Radiology, JDMI, Toronto General Hospital  
8 Department of Anesthesia and Pain Management, Toronto General Hospital, UHN, Toronto, Canada  
9 Department of Anesthesia and Interdepartmental Division of Critical Care Medicine, University of Toronto, Toronto, Canada

✉ Correspondence: [Aaron Conway PhD <[aaron.conway@utoronto.ca](mailto:aaron.conway@utoronto.ca)> | +1 416 946 7112 | 585 University Ave | Toronto, ON M5G 2N2](mailto:aaron.conway@utoronto.ca)

# Abstract

## Introduction

Capnography is commonly used for nurse-administered procedural sedation. This study aimed to determine the accuracy of machine learning models for predicting, at the 15-second time point, if apnea will persist for 30 seconds or more. This information would be useful for determining whether operationalizing these predictions in practice as alarm triggers would be beneficial.

## Methods

A secondary analysis of an observational study was conducted. We selected several candidate models to evaluate, including a random forest model, generalized linear model (logistic regression), lasso regression, ridge regression and the XGBoost model. Out of sample accuracy of the models was calculated using 10-fold cross-validation. The net benefit decision analytic measure was used to assist with deciding whether using the models in practice would lead to better outcomes on average than the default capnography alarm management strategies currently in place. The default strategies are: 1) the aggressive approach, which involves triggering an alarm after brief periods of apnea (typically 15 seconds); and 2) the conservative approach, which involves triggering an alarm for only prolonged periods of apnea (typically 30 seconds).

## Results

A total of 384 apneas longer than 15 seconds were observed in 61 of the 102 patients who participated in the observational study. Nearly half of the apneas (n=180) were prolonged. The random forest had the best discrimination (AUROC 0.66) and calibration. Net benefit associated with the random forest model exceeded the aggressive alarm management strategy but was lower than the conservative alarm management strategy.

## Conclusion

Decision curve analysis indicated that using a random forest model would lead to a better outcome for capnography alarm management in comparison to an aggressive strategy where alarms are triggered after 15 seconds of apnea. The model would not be superior to the conservative strategy, where alarms are only triggered after 30 seconds.

## Introduction

A capnography waveform displays the level of expired carbon dioxide (CO2) over time to show changes in concentrations throughout the respiratory cycle. Capnography waveform abnormalities assist in the detection and diagnosis of specific conditions, such as partial airway obstruction and apnea. For this reason, implementing capnography into practice for respiratory monitoring is considered a high priority to improve patient safety by leading authorities, including peak professional organizations for anesthesia in Canada, the United States, and Europe.[1–3] It is commonly used for nurse-administered procedural sedation, [4–6] including in the interventional radiology setting.[7–11]

A potential consequence to consider in regard to the implementation of capnography for patient monitoring in inpatient wards is alarm fatigue. Alarm fatigue has been linked to patient deaths resulting from clinicians becoming desensitized to alarms leading to delayed responses to clinical deterioration.[12] Deciphering which capnography waveform abnormalities deserve intervention (and therefore alarms to signal the event to clinicians) from those that do not is an essential step towards successful implementation of this technology into practice. For example, triggering alarms after short periods of apnea leads to frequent interruptions and potentially increases the risk of alarm fatigue. Conversely, only intervening once an apneic period reaches a longer threshold negates the potential benefits of capnography on patient safety by improving ventilation. In practice, two alternative strategies for capnography alarm management are typically used. The ‘aggressive’ strategy is for alarms to be triggered after short periods of apnea (e.g. 15 seconds). The ‘conservative’ approach triggers the alarm only once the patient has been apneic for a prolonged period (e.g. 30 seconds).

It is possible that capnography monitor alarm management may be improved upon by using machine learning to create a ‘smart alarm’ that can alert clinicians for apneic events that are predicted to be prolonged. Such an approach aligns with a call from The Society for Critical Care Medicine Alarm and Alert Fatigue Task Force, that machine learning techniques should be used to advance the quality of alerts that clinicians receive and to individualize alert delivery based on clinician response characteristics, such as alert frequency and severity.[13]

For the aggressive strategy, only triggering an alarm for apneas predicted to be prolonged would reduce the total alarm burden and potentially reduce the risk of alarm fatigue. The downside of applying a machine learning approaching to the aggressive apnea treatment strategy would be that some prolonged apneas may not receive early intervention if the model incorrectly predicted that the apnea will not last for >30 seconds (i.e. the false negatives). For the conservative strategy, triggering an alarm at the 15-second timepoint for apneas predicted to be prolonged could reduce the total time of apnea, as treatment could be initiated earlier. The downside of applying a machine learning approaching to the conservative apnea treatment strategy would be that there is the potential that it could increase the total alarm burden if the ratio of false positives (predicting the apnea will last for >30 seconds, but it does not) to true positives (correctly predicting at the 15-second time point that the apnea will last for >30 seconds) is high. This study aimed to determine the accuracy of machine learning models for predicting at the 15-second timepoint if a period of apnea will persist for 30 seconds or more with the goal of determining whether operationalizing these predictions in practice as alarm triggers would be beneficial.

## Methods

A secondary analysis of a prospective observational study was undertaken. Results for the primary aim of the observational study are reported elsewhere.[14] All participants provided written informed consent and the study was approved by human research ethics committees (UCH HREC 1614; SVHAC HREC 16/26; QUT 1600000641). The study was undertaken in accordance with the Australian National Statement on Ethical Conduct in Human Research[15] and it was registered prospectively (ACTRN12616001132437).

### Prediction Goal

The prediction goal was to classify apneic events at the 15-second timepoint as being either short (i.e. terminated prior to 30 seconds) or prolonged (persisted for 30 seconds or longer). The prediction algorithm was compared against typical default alarm settings for capnography monitors. These are the aggressive strategy, which triggers an alarm after 15 seconds of apnea, and the conservative strategy, which triggers an alarm after 30 seconds of apnea.

### Participants

Participants in the study were consecutive adult patients scheduled to undergo an elective procedure in the cardiac catheterization laboratory with moderate sedation. Patients with severe cognitive impairment (due to inability to provide informed consent) or inability to understand and speak English (if an interpreter was unavailable) were excluded.

### Sedation and monitoring

The sedation regimen used for patients included in this study comprised bolus doses of intravenous midazolam and fentanyl. Sedation was administered by nurses who were trained in advanced life support. Routine clinical monitoring included cardiac rhythm and pulse oximetry monitoring as well as non-invasive blood pressure measurements every 5-10 minutes. The Respironics LoFlo sidestream CO2 sensor was used for capnography monitoring. A CO2 sampling cannula was inserted into the sideport of an oxygen face-mask or was integrated as a separate line for nasal cannulas. The capnography waveform was displayed on the main physiological monitoring screen. A default ‘No breaths detected’ alert was triggered for apnea, but no other audible or visual alarms were set for the capnography monitor. No restrictions or specific instructions were provided to clinicians regarding the detection of capnography waveform abnormalities as part of the research protocol because the study used an observational design.

### Data collection

Demographic data and clinical characteristics were collected from medical records and self-report questionnaires prior to procedures. Intra-procedural data were collected in real-time by the researcher who was present in the procedure room. Direct observation of the participant was required to record the timing of sedation administrations and any interventions applied by sedation providers.

### Predictor variables

Several raw demographic (age, sex) and clinical (ASA score, diagnosis of sleep apnea, body mass index, dose and type of sedation and analgesia administered) variables were used as predictors. Features related to sedation dosing used as predictors in the model were the total dose of sedation and number of doses of sedation administered, time since first sedation and the time since the previous dose of sedation. Other features were extracted from the capnography waveform to use as predictors, such as the previous respiratory state (defined as either normal breathing or abnormal), duration of the previous apneic event, time since the previous apneic event, and the total number of apneic events. A total of 18 predictor variables were used.

### Statistical analysis

Analyses were performed using R[16]. Data as well as details about how to access the code and a reproducible computing environment required to verify the results is available from this webpage: <https://github.com/awconway/capnopred>.

#### Modelling

We selected several candidate models to evaluate, including a random forest model, generalized linear model (logistic regression), lasso regression, ridge regression, and the XGBoost model. Out of sample accuracy of the models was calculated using 10-fold cross-validation. Many participants in the study contributed multiple apneic events to the dataset used for modeling. To take this dependency into account, we ensured that apneic events from individual participants were not included in both the training and testing partitions of the 10-fold cross-validation process. Pre-processing steps included normalizing numeric predictors and using an interaction term for the duration of the previous respiratory state and the total number of apneic events.

#### Decision curve analysis

We used the net benefit decision analytic measure to assist with deciding whether using the models in practice would lead to better outcomes on average than the default capnography alarm management strategies currently in place. The default strategies are: 1) the aggressive approach, which involves triggering an alarm after brief periods of apnea (typically 15 seconds); and 2) the conservative approach, which involves triggering an alarm for only prolonged periods of apnea (typically 30 seconds). The net benefit is calculated using the formula:

Where is the net benefit, is the total number of true positives (apneic event at 15 seconds which was predicted to be prolonged and did persist for >30 seconds), is the total number of false positives (apneic events at 15 seconds that was predicted to be prolonged but did not persist for >30 seconds), is the sample size (number of predictions made), and is the probability threshold.

This formula essentially transforms the total number of true positives and false positives into a standardized scale, weighted by the relative harm of a false-positive result.[17] For example, a net benefit of 0.07 means the net benefit of using the model would be seven true positives from every 100 predictions from the model. This net benefit can result from any combination of true positives and false positives.[18] A probability threshold of 0.5 indicates that avoiding a false positive is as important to a clinician as identifying a true positive. Preferences for probability thresholds below 0.5 are weighted such that identifying a true positive is more valuable than avoiding a false positive. Preferences for probability thresholds above 0.5 are weighted such that avoiding a false positive is more valuable than identifying a true positive. For example, for a of 0.75, the ‘value’ of a false positive is three true positives (0.7/0.25 = 3). In other words, to create a net benefit from using the model at this probability threshold, there must be more than three true positives for every false positive prediction made from the model. Conversely, for a of 0.25, the ‘value’ of a false positive is weighted far lower, at only a third of a true positive (0.25/0.75 = 0.33). This means that a net benefit would be achieved if there was more than one true positive for every three false positives. Decision curves can be interpreted such that the strategy with the highest net benefit at each probability threshold has the highest clinical value.[18]

We created a decision curve to plot net benefits across a range of probability thresholds for the aggressive alarm management strategy (alarm triggered at 15 seconds of apnea) as well as for the conservative alarm management strategy (alarm triggered at 30 seconds of apnea). The decision curve takes into account the full range of reasonable clinician preferences for the point at which an alarm should be triggered to signal that the patient is apneic. We tested thresholds in the range from 0.3 to 0.5 for the aggressive alarm management strategy. In a practical sense, this means that we decided that all clinicians who usually use the aggressive alarm strategy would not ever accept a probability for prolonged apnea lower than 0.3 as a useful alarm trigger because there would be little difference between this strategy and just setting the alarm for all apneas. We also decided that all clinicians would always consider that an alarm is indicated for the aggressive alarm management strategy if the probability of prolonged apnea was higher than 0.5. A range of values was used because these probability thresholds can be interpreted as value preferences that individual clinicians may reasonably choose in the context of clinical practice. For example, a clinician who is more ‘risk-averse’ may elect for a more conservative probability threshold (closer to 0.3). Individual participant characteristics will also influence clinicians’ decisions about probability thresholds. A clinician may elect to intervene when the probability for prolonged apnea is 0.3 for an elderly patient with multiple comorbidities but not for a young patient who may be more likely to be able to tolerate longer periods of apnea. For the conservative alarm management strategy comparison, we chose to plot the range of probability thresholds from 0.7 to 0.8. Higher values were chosen because the number of false positives would be an important consideration for clinicians already using a conservative alarm management approach.

## Results

A total of 384 apneic events of at least 15 seconds duration from 61 of the 102 patients who participated in this observational study were included in the present analysis. A summary of participant characteristics is presented in Table . Nearly half of the (n=180) of the apneic events were prolonged, indicating a balanced dataset for the prediction problem at hand.

### Discrimination

A plot of the area under the receiver operating characteristics curves for the models using predictions from 10-fold cross-validation is presented in Figure .

The random forest had the best discriminatory power of the models, with a mean AUROC score of 0.66 (SE 0.03). A threshold performance plot, which summarises the discriminatory power for the models, including values for sensitivity, specificity, positive predictive value, and negative predictive value across all probability thresholds, is presented in Figure .

### Calibration

The random forest model had the best calibration. It approximated observed risk at moderate (0.5) to high (0.8) thresholds (Figure ), although the risk was overestimated at very low thresholds and slightly underestimated between 0.4 to 0.5. Other models severely overestimated risk at low probability thresholds and underestimated risk at high probability thresholds.

### Decision curve analysis

As the random forest performed the best in terms of discrimination and calibration, we chose this model to evaluate using decision curve analysis. The net benefit associated with the random forest model exceeded the aggressive alarm management strategy across all probability thresholds from the range of 0.3 to 0.5 (Figure ). The interpretation is that the best clinical outcome would be achieved for clinicians who are willing to initiate an intervention to treat apnea at the 15-second mark if the probability of it being prolonged was more than 40% by using the random forest model. The net benefit associated with the random forest model was lower than the conservative alarm management strategy across all probability thresholds from the range of 0.7 to 0.8 (Figure ).

## Discussion

Results from prior research indicated that using information about the history of previous respiratory states may be a promising solution in regards to predicting durations of apneic periods. It was found in a study of capnography waveform abnormalities during nurse-administered sedation that the hazard for apnea increased two-fold (HR 2.14, 95% CI 1.75 - 2.62) when a patient was in a state of hypoventilation (defined as >10% reduction in ETCO2 from baseline).[19] The risk of apnea also increased with each additional sedation dose (HR 2.86; 95% CI 2.15 - 3.81).[19] Results from an earlier study in a different population also supported the observations that the onset of apneic periods during sedation are associated with the history of previous respiratory states. Krauss and colleagues[20] used survival analysis to model the time to first apneic events in a sample of 312 patients undergoing procedural sedation with propofol or ketamine in the emergency department. It was identified that having an abnormal ETCO2 measurement 30, 60 and 90 seconds prior to an apneic event increased risk for apnea (hazard ratios with 95% CI for the different lag durations of 2.45 (1.63 - 3.69), 1.88 (1.21 - 2.92) and 2.06 (1.36 - 3.11), respectively).

We leveraged these preliminary findings about the associations between apneic events and the history of previous respiratory states in the present study by building a predictive model using a machine learning approach. A random forest model had the best discriminative ability and calibration from those that we tested. However, it should be noted that the accuracy of this random forest model was still quite low (AUROC = 0.66). Additional research should be undertaken with larger sample sizes to validate our initially promising findings.

Many prediction modeling studies focused on predicting clinical outcomes have yielded similarly low AUROC scores. For example, a recent study of the predictive ability of vital sign parameters for detection of clinical deterioration in subacute care patients yielded an AUROC score of 0.57.[21] When presented with a model with low AUC scores, decision curve analysis can help elicit whether the model is “good enough” to use in practice. Based on our results, nurses currently using the conservative alarm management strategy (alarm only at 30 seconds of apnea) who are willing to ‘value’ a false positive about 2-3 times more than a true positive would not derive an overall net benefit from instead using the random forest model as a trigger for apnea alarms. This is because using the random forest model would produce a worse outcome, in terms of the balance between true positives and false positives, for detecting if an apneic event will be prolonged, than a default strategy of waiting to trigger the alarm until the 30-second threshold.

Conversely, nurses currently using the aggressive capnography alarm management strategy (alarm triggered at 15 seconds of apnea) who are willing to ‘value’ a false positive about two-thirds the importance of a true positive would derive an overall net benefit from instead using the random forest model as a trigger for apnea alarms. Using the random forest model as an additional input for an alarm trigger would reduce the total alarm burden and could be considered as an option for implementation into practice. To operationalize these predictions into capnography monitors, partnerships with industry would be required.[22] Integrating predictive models into alarm management strategies for respiratory monitoring devices is also indicated in other contexts. For example, a recent study found that opioid-induced respiratory depression during recovery from anesthesia can be accurately predicted using a machine learning approach.[23] In addition, user-centered design considerations, such as how the predictions should be communicated to nurses responsible for decision-making, would be important avenues for further research prior to implementation.[24]

The finding that the model was able to produce an overall net benefit against the aggressive capnography alarm management strategy but not the conservative approach is a noteworthy finding. Further research with larger sample sizes is needed to increase the predictive power of models aimed at predicting the duration of apnea. Such research is warranted because triggering an alarm after 30 seconds of apnea that would have self-resolved without clinical intervention only 5 seconds later is just as clinically inconsequential as triggering an alert after 5 seconds of apnea that would similarly have resolved after a short period of time. In both of these circumstances, there would not have been enough time for the clinician’s intervention to take effect. Yet, presumably in an attempt to reduce alarm burden, many capnography monitoring devices alarm settings default to triggering after apnea has been observed for 30 seconds. An ideal alternative to the conservative alarm strategy would be for clinicians to receive capnography monitor alarms as *early as possible* in the course of an apneic event, *but only* if the event will be prolonged for enough time for a clinical intervention to be implemented and exert its intended effect. Results from the present study did not achieve this goal. Yet, previous research indicates that finding such a solution may be worthwhile. An analysis of over 500 thousand patients found that respiratory compromise in interventional radiology procedures performed with moderate sedation contributed to worse clinical outcomes and higher costs.[25]

### Limitations

Although the number of apneas included in the models was relatively high, these were contributed by a small number of patients. We used cross-validation to minimize the possibility of overfitting. It should also be noted that the data used in this analysis were produced from an observational study conducted at two hospitals that used a convenience sampling approach. Therefore the possibility of selection bias should be considered. The context in which the study was conducted should also be considered in regard to external validity. Participants in the study were patients undergoing procedures in a cardiac catheterization laboratory where small, bolus doses of midazolam and fentanyl were used for sedation. Other procedural sedation contexts may use different doses of sedation and the types of medications. It would not be expected that the results of this study could be generalized to those other contexts.

### Conclusion

We evaluated several candidate models to determine their accuracy in predicting if an apnea at 15 seconds would be prolonged for more than 30 seconds. The random forest model provided the best discrimination and calibration. Decision curve analysis indicated that using the random forest model would lead to a better outcome for capnography alarm management in comparison to an aggressive strategy where alarms are triggered after 15 seconds of apnea. The model would not be superior to the conservative strategy, where alarms are only triggered after 30 seconds.

## References

1. Apfelbaum J, Gross J, Connis R et al. Practice guidelines for moderate procedural sedation and analgesia 2018: A report by the American Society of Anesthesiologists Task Force on Moderate Procedural Sedation and Analgesia, the American Association of Oral and Maxillofacial Surgeons, American College of Radiology, American Dental Association, American Society of Dentist Anesthesiologists, and Society of Interventional Radiology. *Anesthesiology* 2018; **128**: 437–79.

2. Hinkelbein J, Lamperti M, Akeson J et al. European Society of Anaesthesiology and European Board of Anaesthesiology guidelines for procedural sedation and analgesia in adults. *European Journal of Anaesthesiology* 2018; **35**: 6–24.

3. Dobson G, Chong MA, Chow L et al. Procedural sedation: A position paper of the Canadian Anesthesiologists’ Society. *Canadian Journal of Anesthesia* 2018; **65**: 1372–84.

4. Conway A, Rolley J, Page K, Fulbrook P. Trends in nurse-administered procedural sedation and analgesia across cardiac catheterisation laboratories in australia and new zealand: Results of an electronic survey. *Australian Critical Care* 2014; **27**: 4–10.

5. Conway A, Rolley J, Page K, Fulbrook P. Clinical practice guidelines for nurse-administered procedural sedation and analgesia in the cardiac catheterization laboratory: A modified delphi study. *Journal of Advanced Nursing* 2013; **70**: 1040–53.

6. Conway A, Rolley J, Page K, Fulbrook P. Issues and challenges associated with nurse-administered procedural sedation and analgesia in the cardiac catheterisation laboratory: A qualitative study. *Journal of Clinical Nursing* 2013; **23**: 374–84.

7. Pella L, Lambert C, McArthur B et al. Systematic review to develop the clinical practice guideline for the use of capnography during procedural sedation in radiology and imaging settings: A report of the association for radiologic & imaging nursing capnography task force. *Journal of Radiology Nursing* 2018; **37**: 163–72.

8. Brast S, Bland E, Jones-Hooker C, Long M, Green K. Capnography for the radiology and imaging nurse: A primer. *Journal of Radiology Nursing* 2016; **35**: 173–90.

9. Long M, Green K, Bland E et al. Capnography monitoring during procedural sedation in radiology and imaging settings: An integrative review. *Journal of Radiology Nursing* 2016; **35**: 191–7.

10. Mafeld S, Musing E, Conway A, Kennedy S, Oreopoulos G, Rajan D. Avoiding and managing error in interventional radiology practice: Tips and tools. *Canadian Association of Radiologists Journal* 2020; **71**: 528–35.

11. Werthman JA, Dietrich MS, Maxwell CA, Jordan LM, Minnick AF. The administration of moderate sedation and the role of the radiology nurse. *Journal of Radiology Nursing* 2020; **39**: 92–8.

12. Chopra V, McMahon LF. Redesigning hospital alarms for patient safety: Alarmed and potentially dangerous. *Jama* 2014; **311**: 1199–200.

13. Winters BD, Cvach MM, Bonafide CP et al. Society for critical care medicine alarm and alert fatigue task force. Technological distractions (part 2): A summary of approaches to manage clinical alarms with intent to reduce alarm fatigue. *Crit Care Med* 2018; **46**: 130–7.

14. Conway A, Collins P, Chang K, Mafeld S, Sutherland J, Fingleton J. Sequence analysis of capnography waveform abnormalities during nurse-administered procedural sedation and analgesia in the cardiac catheterization laboratory. *Scientific Reports* 2019; **9**: 1–9.

15. Health N, (Australia) MRC. *Australian Code for the Responsible Conduct of Research: Revision of the Joint NHMRC/AVCC Statement and Guidelines on Research Practice*. National Health and Medical Research Council, 2007.

16. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021.

17. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making* 2006; **26**: 565–74.

18. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352**.

19. Conway A, Collins P, Chang K et al. Pre-apneic capnography waveform abnormalities during procedural sedation and analgesia. *Journal of Clinical Monitoring and Computing* 2019: 1–8.

20. Krauss BS, Andolfatto G, Krauss BA, Mieloszyk RJ, Monuteaux MC. Characteristics of and predictors for apnea and clinical interventions during procedural sedation. *Annals of Emergency Medicine* 2016; **68**: 564–73.

21. Considine J, Street M, Hutchinson AM et al. Vital sign abnormalities as predictors of clinical deterioration in subacute care patients: A prospective case-time-control study. *International Journal of Nursing Studies* 2020; **108**: 103612.

22. Glasgow MES, Colbert A, Viator J, Cavanagh S. The nurse-engineer: A new role to improve nurse technology interface and patient care device innovations. *Journal of Nursing Scholarship* 2018; **50**: 601–11.

23. Jungquist CR, Chandola V, Spulecki C et al. Identifying patients experiencing opioid-induced respiratory depression during recovery from anesthesia: The application of electronic monitoring devices. *Worldviews on Evidence-Based Nursing* 2019; **16**: 186–94.

24. Risling TL, Risling DE. Advancing nursing participation in user-centred design. *Journal of Research in Nursing* 2020; **25**: 226–38.

25. Urman RD, Moucharite M, Flynn C, Nuryyeva E, Ray Jr CE. Impact of respiratory compromise in inpatient interventional radiology procedures with moderate sedation in the united states. *Radiology* 2019; **292**: 702–10.

## Figure legend

Fig. 1 Area under receiver operating characteristics curve

Fig. 2 Threshold performance plot for all models evaluated

Fig. 3 Calibration plot for all models evaluated

Fig. 4 Decision curve analysis plots. Panel A is the comparison for the aggressive alarm management strategy and panel B is the comparison for the conservative alarm management strategy.

Table : Participant characteristics

| Characteristic | N = 611 |
| --- | --- |
| Age | 76 (68, 80) |
| Body mass index | 26.4 (24.6, 29.8) |
| Sex |  |
| Female | 22 (36%) |
| Male | 39 (64%) |
| Obstructive sleep apnea |  |
| No OSA | 46 (75%) |
| OSA | 15 (25%) |
| American Society of Anesthesiology Physical Classification Status |  |
| I or II | 37 (61%) |
| III or IV | 24 (39%) |
| 1Median (IQR); n (%) | |

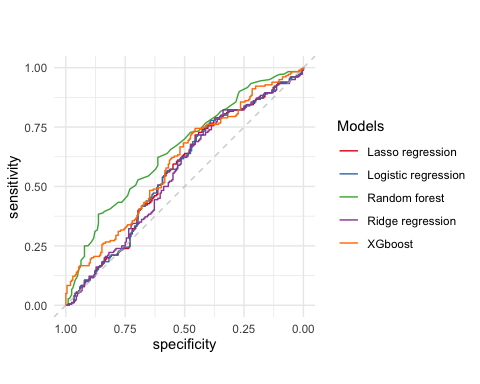


Figure : Area under receiver operating characteristics curve

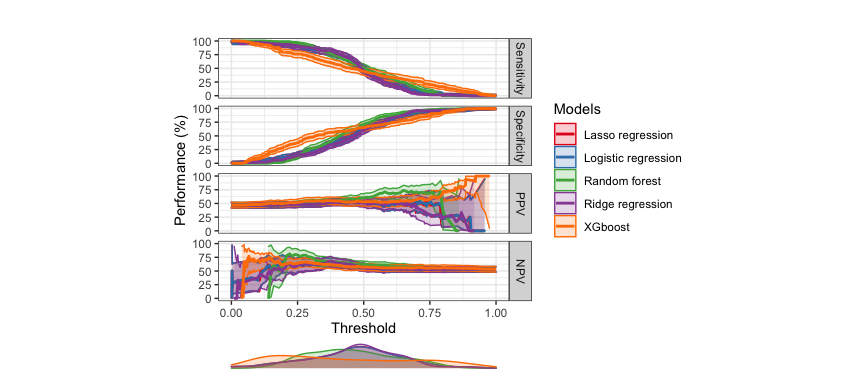


Figure : Threshold performance plot for all models evaluated

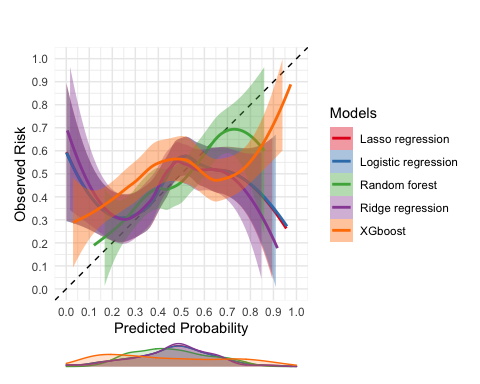


Figure : Calibration plot for all models evaluated

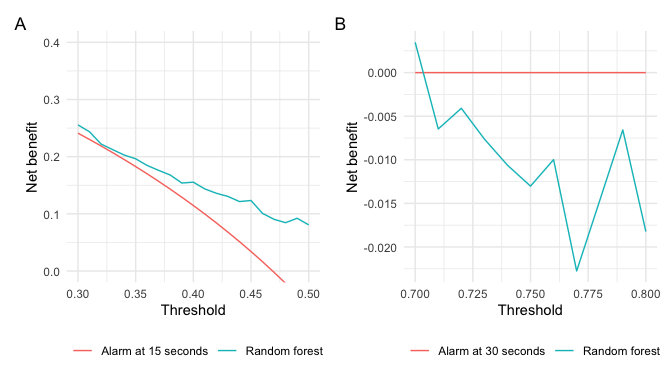


Figure : Decision curve analysis plots. Panel A is the comparison for the aggressive alarm management strategy and panel B is the comparison for the conservative alarm management strategy