Peer Review File

Article information: https://dx.doi.org/10.21037/jtd-21-765.

Reviewer A

Comment 1：This article tackles an important topic and provides a summary that will be very useful to clinicians in this field. I think this will be a valuable addition to the literature with some tweaks. Here are the areas I think need to be addressed:

Reply 1：Thank you for your kind words. We provided a point-by-point response below.

Changes in the text：The location of change matches the location in the clean version of the manuscript.

Comment 2：Introduction - discussion of supervised, unsupervised, and reinforcement learning is hard to understand. Would suggest simplifying the language and perhaps including examples. This section also has grammatical errors and incomplete sentences.

Reply 2：We have added examples per type of variable that appeal more to the imagination.

Changes in the text：See page 5 and 6, lines 80-96

(i) supervised ML is concerned with the training of a model towards a known target variable (outcome). By differing the weighted effect of given labeled inputs (e.g., age, sex, cholesterol level, smoking status), it minimizes the prediction error of the desired output (for example, having cardiovascular disease or not). Most applications in medicine apply this principle of machine learning, using either a classification or regression model. (ii) Unsupervised ML is when the algorithm obtains unlabeled data (e.g., large sets of radiological or histological images) and attempts to find patterns. This is a more exploratory method as the algorithm decides what classes and patterns best describe the data. (iii) Reinforcement learning is the technique that is perhaps key to surpassing human capability. This method learns what actions lead to the highest possible reward. This reward is predefined and usually custom-tailored to the problem at hand. In this case, a training set is absent, but it is created by the inputs the model receives through interaction with the environment. An example of such a reward is each time an autonomous vehicle stays within its lane. Through positive and negative reinforcement, the self-driving model learns what the required behavior is and what actions lead to that scenario. For now, the use of ML in medicine is mainly limited to supervised methods.

Comment 3：Introduction - The next section discussing examples of success of ML is too generic. It doesn't give any specifics. I would consider getting rid of this entire "overview" of ML/AI and stick to discussions of perioperative medicine and where this tool could be useful.

Reply 3：This brief overview was intended to inform the reader about the added value of ML in other medical fields, to make clear that it should not only be seen as a new statistical method but that it has proven to be a better method in some cases. However, based on your feedback we agree that this might be confusing.

Changes in the text：As suggested we removed the following paragraph: “Real-time clinical advantages have been illustrated in diagnosing myocardial infarction (6) and their predictive capability is seen across different fields from psychiatry to radiology (8-13), with clear advantages over traditional statistical references in some medical fields (4, 14).”

Comment 4：Prediction of mortality - This section feels too short and the discussion too shallow. I would like to know more about the models discussed and to get into the details more.

Reply 4：We have explained the most contributing variables used in the ML models (if available) to showcase what were important risk factors and whether they were obtained preoperative.

We did not discuss how each model was set up (percentage of training set and validation set, etc.), but we now explicitly mention the type of ML models compared and the respective AUCs.

Changes in the text：See page 10 and 11, lines 156 to 190.

Within the elderly population, six perioperative variables (not further specified by the authors) were found to be strongly correlated with mortality. Based on those variables, a logistic regression (LR) model, Bayesian network (BN), and an artificial neural network (ANN) produced AUCs of respectively 0.854, 0.931, and 0.941, clearly outperforming the EuroSCORE that had an AUC of 0.648 in this population. (18). Overall, the main mortality predictors in RHD were found to be left atrium size, high creatinine, tricuspid procedure, reoperation, and pulmonary hypertension. Using a random forest (RF) model, a new clinical score, the RheSCORE, was built on those predictors. With an area of 0.98, it outperforms the EuroSCORE II, which produces an AUC of 0.857 based on essentially the same predictors. (19).

However, in a mixture of cardiac surgery procedures, the two aforementioned clinical scores perform similarly or slightly less than advanced models (20-22). An ANN yielded comparable predictive properties to the EuroSCORE (AUC 0.80 vs. 0.79), with only a small advantage in the case of valve procedures (AUC 0.76 vs. 0.72, p-value 0.0001) (20). Assembling four ML models (gradient boosting machines (GBM), RF, support vector machines (SVM), and Naïve Bayes (NB)) created a significant but modest benefit with an AUC of 0.795 versus 0.737 for the EuroSCORE II (21). Similarly, modest advantages in accuracy and AUC were seen comparing an advanced ML model (extreme gradient boosting machine (XGBoost)) to the STS clinical score. Interestingly, despite both the STS score and the XGBoost being well-calibrated and having a high area under the curve (respectively 0.808 and 0.795), they identified a large proportion of different patients as being at risk (22). Even one of the first clinical scores, the Parsonnet score, still holds value in predicting in-hospital mortality with a comparable AUC to an advanced LR and ANN model (0.829, 0.852, and 0.873, respectively) (23).

Also, when comparing advanced ML methods, little difference in predicting performance is seen (23-29), with only a slight advantage for nonlinear models (ANN, BN, and multilayer sigmoid perceptron (MLP)) over linear LR models (18, 20, 30). The majority of these studies use a set of preoperative data, including demographic characteristics, medical history, and type of surgery performed. Adding intraoperative hypotension as a dynamic parameter to these preoperative data showed improved AUCs for advanced LR, RF, and XGBoost models.

At the same time, an SVM and ANN did not benefit from this added parameter, outputting AUCs of 0.66 and 0.70, respectively (31).

Comment 5：Line 177 - incomplete sentence

Reply 5：Well noticed, we have modified the sentence as follows:
We did not discuss how each model was set up (percentage of training set and validation set, etc.), but we now explicitly mention the type of ML models compared and the respective AUCs.

Changes in the text：See page 12, line 210 to 211. This is not fully acceptable as they can cause distraction or alarm fatigue.

Comment 6：Automation of IOE - please expand on the mitral valve example. it is unclear what this technology does and why it is useful.

Reply 6：Thank you for the feedback. We have now explained the methods that were employed in the article by Jeganathan et al. and we explain the advantages and disadvantages of current AI software in MV assessment.

Changes in the text：See page 13 and 14, line 239 to 254.

The second AI application in ultrasound automation relates to the analysis of the mitral valve (MV) (50). Patients with a normal biventricular function who underwent an elective CABG surgery were included for ultrasound imaging to evaluate the clinical applicability and accuracy of an AI-based MV analysis software. An experienced echocardiographer captured three end-systolic frames of the MV in each patient. Postoperatively, these frames were analyzed with the AI software. The software automatically traced the valves, and three experienced examiners independently verified the valve tracings. Thus, creating three separate datasets for all frames, as the examiners could administer minor manual adjustments when deemed necessary. Subsequently, the software's six clinically relevant geometric parameters were calculated from the verified MV tracings (annulus anterolateral posteromedial diameter, annulus anteroposterior diameter, annular area, annulus nonplanarity angle, annulus total perimeter, and anterior and posterior leaflet areas). Statistical analyses showed a high precision for the calculated parameters in corresponding end-systolic frames in which only the valve tracings were verified by different examiners. Meaning that the latter did not affect the outcome (50).

Comment 7：Future directions and challenges - I find this section far too long. The beauty in this article is the fact that is goes through the perioperative medicine literature and summaries the ML/AI applications there. I would ask the authors to focus more on that and ensure the discussion of each of these models is thorough enough for the reader to understand why it is important/relevant. This last section seems more opinion based and to me is less useful.
Reply 7：This section focuses more on possibilities that we want to highlight to make clear what could come in the future and what the obstacles or dangers could be. We agree that it is partly speculation, but not completely unfounded. We have removed several paragraphs that were purely hypothetical or did not contribute to clarifying the message. We narrowed it down to one page.
Changes in the text：
Removal of:
"Several important issues will have to be addressed before such implementations can be realized. First of all, it must be established whether actively self-improving models are desirable in hospitals. It could be beneficial to link these models to a continuously […] to unlock their full potential."

Removal of:
"For example, who will be responsible in due course for reviewing self-adapting ML models to verify that they still account for emerging risks following new surgical or anesthetic advancements, the software provider, the provided hospitals, or an external party. The same should be answered for legal consequences. The insight of physicians and nurses will always remain paramount above ML algorithms, but if their actions are based on biased ML data, the liability shifts to a grey area. Can the ML-engineer be kept responsible while the "unlocked" model is continuously learning, is the hospital implementing the model accountable, or is the autonomous physician fully culpable in case of errors?"

Removal of:
"[…] due to impossible judicial protection for all parties involved"

Removal of:

“It may all come down to the conclusion that “unlocked” self-improving models are only appropriate in research settings and not for clinical use. In any case […]”

Reviewer B

Comment 1：The relevance of machine learning in medicine is increasing and the review is therefore timely.

Reply 1：Thank you for this comment.

Comment 2：In the abstract (Background) it is mentioned the focus of the review is on the use of machine learning techniques specifically for anethetic care. We conducted a scoping review to examine the extent and potential limitations of ML implementation in perioperative anesthetic care, specifically in cardiac surgery patients. However in the following section articles relevant to postoperative ICU admission were also considered for the review. Can you provide some explanation linking anesthetic care to ICU admission including the role of ML?

Reply 2：In this review we want to focus on perioperative care. In case of cardiac surgery patients this stretches from the pre-operative assessment of patients (e.g. risk evaluation) to the anesthetized postoperative period which extends to a short post-operative ICU admission. We considered this to be a relevant period, because of the large available data (arterial line measurements, mechanically ventilated time) which lends itself for machine learning. By including this ICU phase we can give a full overview of ML applications covering the entire perioperative phase.