PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Protocol for a Retrospective Study Using Machine Learning Techniques to Develop Forecasting Algorithms for Postoperative Complications: The ACTFAST-2 Study
AUTHORS	Fritz, Bradley; Chen, Yixin; Murray-Torres, Teresa; Gregory, SH; Ben Abdallah, Arbi; Kronzer, Alex; McKinnon, Sherry; Budelier, Thaddeus; Helsten, Daniel; Wildes, Troy; Sharma, Anshuman; Avidan, Michael

VERSION 1 – REVIEW

REVIEWER	Dennis Toddenroth, MD Universität Erlangen-Nürnberg, Germany
REVIEW RETURNED	17-Nov-2017

GENERAL COMMENTS	Thank you for the opportunity to review this interesting manuscript, in which the authors outline their planned development and evaluation of predictive models of postoperative complications, based on electronically recorded preoperative und intraoperative data. The protocol format may not have been originally designed for predictive modeling studies, even though pertinent recommendations such as the TRIPOD statement [1] suggest that their previous publication may be worthwhile. So despite the fact the described study based on observations from between 2012 and 2016 may not really meet the requirement (in the instructions for reviewers) that data collection cannot be 'complete' for protocols, the editors may judge that the present manuscript attains a sufficient degree of traceability to substantiate future results. While the text appears carefully written, and the described methods seem certainly advanced, I would like to propose that the authors consider addressing the following suggestions during a revision: • The authors declare their ultimate goal of eventually supporting intraoperative clinical management, including in their title, but in my view provide little explanation that their retrospective modeling seems to employ data from complete episodes, while potential intraoperative interventions can of course only use information up until each respective point in time. Prediction models based on retrospective summary statistics (such as 'variance, skewness, and kurtosis') that are computed from entire series may not be representative for the predictors that are gradually accrued for future patients during surgery - particularly at the onset of some

relevant crisis halfway into the episode. The intended application of retrospectively developed models for future clinical decision support might thus require some additional explanation, specifically in light of the fact that predictions from complex decision models may sometimes not translate into obvious interventions. Even if accurate models could identify those patients that would experience certain adverse outcomes (e.g., based on a convolutional neural network, or sets of ‘shapelets’ for various attributes), the complexity of determinants could lead to situations where no consequences for intraoperative management ‘in real time’ seem defensible.

- The authors argue that the participants ‘rights and welfare’ will be adequately protected on the grounds that ‘no additional data will be collected’ beyond what has already been recorded. The included patients, however, may also have a legitimate interest in that their confidential clinical data will not become more widely available in any form that might entail personal disadvantages. Some insurer or employer with a hypothetical access to individually identifiable diagnoses, for example, could use this information in a way that may be detrimental to these patients. Since the manuscript refers to the integration with data from outcome registries as well as to a subsequent distribution of patient-level data to other interested researchers, I advise that the authors comment on their planned precautions and procedures to steer clear of theoretical re-identification risks. (Effective anonymization of the described dataset could turn out to be specifically complicated due to its complex structure and its extent.)

- The proposed methods seem diverse and advanced, but simultaneously appear so variegated that they might impair the comprehensibility of future findings. To improve credibility, the authors may want to tone down or replace a few unspecific methodological remarks (such as references to ‘efficient training algorithms’, ‘novel classification algorithms’ or ‘more powerful algorithms’), and instead provide additional descriptive details about down-to-earth practical aspects, such as their intended handling of missing values, or their approach for managing varying observation times (shorter vs. longer surgical interventions) – which might for example involve a targeted exclusion of certain samples or the imputation of values.

- While the ‘data dictionary’ supplements detail a number of clinical attributes that will be used for modeling, I presume that many readers would find a cursory description of typical input records useful. How many vital sign readings approximately accumulate during a representative episode based on its duration and the configured measurement interval, how many lab reading are conventionally available? A data flow diagram that illustrates the overall procedure might also simplify the accessibility of the presentation for many readers.

- The authors state that they will compute different evaluation metrics for model assessment within the training data (‘mean-squared prediction error’) and within another out-of-sample validation (accuracy, precision, and robustness). Since a deteriorating performance between training and validation data is often interpreted to indicate overfitting, I do not understand why inconsistent metrics should be used - please explain. Since classification performance is frequently reported in terms of either pairs of relative frequencies (mostly sensitivity & specificity, or recall

	& precision) or the area under the ROC curve, I also find it counterintuitive to report a combination of accuracy and precision (and its specification as the 'percentage of correctly forecasted events' appears a bit ambiguous in relation to the conventional definition of *recall*). • DETAIL: The authors use the terms generative/parametric as well as discriminative/nonparametric as if these were respectively synonymous. I presume that most clinical researchers would see 'parametric' models as those that build on distributional assumptions that can be specified in the form of a few numbers (parameters), while the 'nonparametric' ones do not. In contrast, 'generative' methods in my view model the joint distribution of inputs and outputs (independent and dependent variables), while 'discriminative' methods model output distributions for given inputs. Please clarify. • DETAIL: If between 40% and 80% of episodes are expected to yield usable datasets (between 50 and 100 of 125 on a given business day?), should we not anticipate that over 4 ¼ years between approximately 32,300 and 64,600 samples accumulate, instead of the quoted 'minimum' estimation between 50,000 and 100,000? A potential ambiguity between 'days' and 'business days' seems to distract from a simple extrapolation that 19,000 yearly episodes should amount to 80,750 episodes over the specified interval (without being a 'conservative' minimum). • DETAIL: The description of the bootstrap-based validation explains that '100 surrogate samples' will be drawn and analyzed, while probably '100 surrogate data sets' were meant. (A single iteration would exhaust the information content of this dataset only insufficiently.) [1] https://www.ncbi.nlm.nih.gov/pubmed/25569120

REVIEWER	Dr C L Gurudatt professor and Head, Department of Anaesthesiology, Critical Care and Pain Medicine JSS Medical College and Hospital, Mysore INDIA
REVIEW RETURNED	04-Dec-2017

GENERAL COMMENTS	Congratulations to the authors for taking up this project as it is going to be beneficial in anticipating complications and preventing them once the algorithm is made and tested successfully. Over 50,000 patients each year sustain a perioperative myocardial infarction (PMI) and hence prevention of a PMI is important to improve overall postoperative outcome. Thus PMI can also be added as a primary objective of the study along with respiratory failure and post operative kidney injury. Regarding the exclusion criteria, patients already in respiratory failure and underwent surgery can be excluded from the study. Regarding the references, some of the references are not as per the Vancouver style. For references with 1 to 6 authors, list all authors. For references with more than 6 authors, list the first 6 authors then add 'et al.'

VERSION 1 – AUTHOR RESPONSE

Editorial Requests:

- Please revise your title to indicate the research question, study design, and setting. This is the preferred format of the journal. You are welcome to include the trial acronym but we suggest clarifying what the acronym stands for in the abstract rather than in the title.

RESPONSE: We have revised the title to address these concerns. The new title reads, "Protocol for a Retrospective Study Using Machine Learning Techniques to Develop Forecasting Algorithms for Postoperative Complications: The ACTFAST-2 Study"

- Please revise the Abstract >> 'Ethics and dissemination' section. Please include information about ethics approval here. It should be similar to what is presented on page 18.

RESPONSE: We have added the following sentence to the Ethics and Dissemination section of the Abstract on page 3: "This study has been approved by the Human Research Protection Office at Washington University in St. Louis."

Reviewers' Comments to Author:

Reviewer: 1
Reviewer Name: Dennis Toddenroth, MD
Institution and Country: Universität Erlangen-Nürnberg, Germany
Competing Interests: None declared

Thank you for the opportunity to review this interesting manuscript, in which the authors outline their planned development and evaluation of predictive models of postoperative complications, based on electronically recorded preoperative und intraoperative data.

The protocol format may not have been originally designed for predictive modeling studies, even though pertinent recommendations such as the TRIPOD statement [1] suggest that their previous publication may be worthwhile. So despite the fact the described study based on observations from between 2012 and 2016 may not really meet the requirement (in the instructions for reviewers) that data collection cannot be 'complete' for protocols, the editors may judge that the present manuscript attains a sufficient degree of traceability to substantiate future results.

While the text appears carefully written, and the described methods seem certainly advanced, I would like to propose that the authors consider addressing the following suggestions during a revision:

• The authors declare their ultimate goal of eventually supporting intraoperative clinical management, including in their title, but in my view provide little explanation that their retrospective modeling seems to employ data from complete episodes, while potential intraoperative interventions can of course only use information up until each respective point in time. Prediction models based on retrospective summary statistics (such as 'variance, skewness, and kurtosis') that are computed from entire series may not be representative for the predictors that are gradually accrued for future patients during surgery - particularly at the onset of some relevant crisis halfway into the episode. The intended application of retrospectively developed models for future clinical decision support might thus require some additional explanation, specifically in light of the fact that predictions from complex decision models may sometimes not translate into obvious interventions. Even if accurate models could identify those patients that would experience certain adverse outcomes (e.g., based on a convolutional neural network, or sets of 'shapelets' for various attributes), the complexity of determinants could lead to situations where no consequences for intraoperative management ‘in real time’ seem defensible.

RESPONSE: We intend that our models will ultimately be used in real time in the operating room, when time series data is only available up until the present time. At any given point in time, we intend to use time series data from the preceding 60 minutes. Thus, an initial prediction can be made 60 minutes after the start of surgery, and this prediction can be updated using the most recent data points as the surgery progresses. When training and validating our models, we will select 60-minute epochs from the historical datasets. We have expanded the Data Analysis section to add this information (page 10, paragraph 1, top of page 11).

We understand the reviewers’ concern regarding whether our predictions can be translated into interventions to prevent adverse postoperative outcomes. However, we believe that successful prediction of adverse postoperative outcomes would be useful. At a minimum, accurate predictions could assist with selecting the most appropriate post-recovery destination for the patient. Prediction of adverse outcomes may help to identify which patients need to be admitted to an intensive care unit rather than a hospital ward, or which patients should be admitted to the hospital rather than discharged home. We have added a sentence to the Implications and Future Directions section on page 17 to address this point.

• The authors argue that the participants ‘rights and welfare’ will be adequately protected on the grounds that ‘no additional data will be collected’ beyond what has already been recorded. The included patients, however, may also have a legitimate interest in that their confidential clinical data will not become more widely available in any form that might entail personal disadvantages. Some insurer or employer with a hypothetical access to individually identifiable diagnoses, for example, could use this information in a way that may be detrimental to these patients. Since the manuscript refers to the integration with data from outcome registries as well as to a subsequent distribution of patient-level data to other interested researchers, I advise that the authors comment on their planned precautions and procedures to steer clear of theoretical re-identification risks. (Effective anonymization of the described dataset could turn out to be specifically complicated due to its complex structure and its extent.)

RESPONSE: We agree that the scope of the dataset compiled for this project increases the consequences of a data breach. There are many methods available to share data with other researchers, and we have carefully balanced our duty to share data for the advancement of scientific knowledge with our duty to protect the privacy of patients. Because so much detail about each patient is included, we have decided not to make the dataset publicly available online. Rather, data will only be shared with researchers who submit a methodologically sound research proposal including a proposal and statistical analysis plan, as described in our Ethics and Dissemination section on page 18. This reduces the risk that health information will fall into the hands of a patient’s employer or insurer.

No patient-identifying fields (including dates or years) will be included in the shared dataset. Age will be provided in years, unless the patient is older than 89 years in which case age will be reported as “>89 years.” Any dates will be presented as “number of days since index surgery.” Time of day will be included, as this is not a HIPAA patient identifier. These details have been added to the Ethics and Dissemination section on page 18-19.

If the same patient undergoes more than one surgery, then all records related to that patient will be linked with a common study identifier number. This number will be distinct from the medical record number and will not be generated from any HIPAA elements. All dates related to that patient will be reported as the number of days since the patient’s first surgery.

• The proposed methods seem diverse and advanced, but simultaneously appear so variegated that they might impair the comprehensibility of future findings. To improve credibility, the authors may want to tone down or replace a few unspecific methodological remarks (such as references to 'efficient training algorithms', 'novel classification algorithms' or 'more powerful algorithms'), and instead provide additional descriptive details about down-to-earth practical aspects, such as their intended handling of missing values, or their approach for managing varying observation times (shorter vs. longer surgical interventions) – which might for example involve a targeted exclusion of certain samples or the imputation of values.

RESPONSE: We have removed several instances of unspecific language from the Data Analysis section. We have removed the sentence "A key area of improvement is feature transformation" from page 10. We have re-structured this paragraph to explicitly focus on the difference between Nadaraya-Watson kernel density estimation and bin-based kernel density estimation.

We have also removed the sentence "We will also develop efficient training algorithms" from page 10.

We have removed the following paragraph in its entirety: "We plan to develop novel classification algorithms that best fit our data. In our preliminary work, we proposed DLR, a novel nonlinear hybrid classification algorithm that integrates kernel density estimation with logistic regression. DLR can achieve nonlinear separability by utilizing a nonlinear feature transformation, but is much more efficient than other nonlinear models since it fits a linear model. It can naturally handle mixed data types. It also offers good interpretability. In this task, we plan to develop more powerful algorithms on top of DLR."

Missing values within the time series datasets will be handled using linear interpolation. We have added this detail to the Data Analysis section, on page 11.

• While the 'data dictionary' supplements detail a number of clinical attributes that will be used for modeling, I presume that many readers would find a cursory description of typical input records useful. How many vital sign readings approximately accumulate during a representative episode based on its duration and the configured measurement interval, how many lab reading are conventionally available? A data flow diagram that illustrates the overall procedure might also simplify the accessibility of the presentation for many readers.

RESPONSE: We have added two sentences to the Data Acquisition section on page 8 to more fully describe how frequently vital signs are measured and how many vital signs might be present in a typical case: "Blood pressure measurements are available at intervals ranging from once per minute to once every five minutes, while other vital signs are captured once per minute. Thus, a three-hour procedure would have about 180 measurements for each vital sign."

We have also added several sentences on page 9 outlining when preoperative lab studies would be available: "In general, a preoperative complete blood count is available if the patient is undergoing major surgery with potential significant blood loss or if other clinical reasons are present. Electrolytes and renal function are available if there is clinical reason to suspect an abnormality (including, but not limited to, patients with hypertension, diabetes mellitus, or chronic kidney disease). Additional tests, such as hepatic function and coagulation studies, are available on smaller sets of patients in whom the tests are clinically indicated."

• The authors state that they will compute different evaluation metrics for model assessment within the training data ('mean-squared prediction error') and within another out-of-sample validation (accuracy, precision, and robustness). Since a deteriorating performance between training and validation data is often interpreted to indicate overfitting, I do not understand why inconsistent metrics should be used - please explain. Since classification performance is frequently reported in terms of either pairs of relative frequencies (mostly sensitivity & specificity, or recall & precision) or the area under the ROC curve, I also find it counterintuitive to report a combination of accuracy and precision (and its specification as the 'percentage of correctly forecasted events' appears a bit ambiguous in relation to the conventional definition of *recall*).

RESPONSE: We have rewritten the Forecasting Algorithm Validation section on page 14. Our new methods incorporate precision and recall, using the conventional definitions. The same metrics will be used in the training and validation datasets. The new text reads as follows:

"For initial model training and validation, the historical database will be divided into a training dataset (60% of the database), a validation dataset (20% of the database), and a testing dataset (20% of the database). Because we expect that our target outcomes will be relatively rare events, overall classification accuracy is not likely to be a useful measure of model performance. Instead, we will use precision (true positives/[true positives + false positives]) and recall (true positives/[true positives + false negatives]). We will optimize model parameters using the training dataset. Then we will pre-specify our desired recall and use the validation dataset to select the decision threshold that leads to the highest precision without sacrificing our desired recall. Then we will apply our model to the testing dataset and report the observed precision and recall.

"Additionally, we propose to perform a validation test of the predictive performance of the developed algorithms prospectively, using patient records that did not belong to the learning database. For this evaluation, we will apply our model to the prospectively-collected data. We will report the observed precision and recall as measures of model performance."

• DETAIL: The authors use the terms generative/parametric as well as discriminative/nonparametric as if these were respectively synonymous. I presume that most clinical researchers would see 'parametric' models as those that build on distributional assumptions that can be specified in the form of a few numbers (parameters), while the 'nonparametric' ones do not. In contrast, 'generative' methods in my view model the joint distribution of inputs and outputs (independent and dependent variables), while 'discriminative' methods model output distributions for given inputs. Please clarify.

RESPONSE: We agree that "nonparametric" is not synonymous with "generative;" nor is "parametric" synonymous with "discriminative." The sentence in question at the top of page 10 should reference generative versus discriminative models. We have removed the references to parametric and nonparametric models.

• DETAIL: If between 40% and 80% of episodes are expected to yield usable datasets (between 50 and 100 of 125 on a given business day?), should we not anticipate that over 4 ¼ years between approximately 32,300 and 64,600 samples accumulate, instead of the quoted 'minimum' estimation between 50,000 and 100,000? A potential ambiguity between 'days' and 'business days' seems to distract from a simple extrapolation that 19,000 yearly episodes should amount to 80,750 episodes over the specified interval (without being a 'conservative' minimum).

RESPONSE: We have amended this paragraph on page 7 to remove some of the distracting numbers. Because we will collect data from a 4.25-year period and our hospital performs 19,000 surgeries per year, we expect 4.25 * 19,000 = 80,750 surgeries. We report this estimate as "80,000-90,000 surgeries."

• DETAIL: The description of the bootstrap-based validation explains that ‘100 surrogate samples’ will be drawn and analyzed, while probably ‘100 surrogate data sets’ were meant. (A single iteration would exhaust the information content of this dataset only insufficiently.)

RESPONSE: The model validation section has been rewritten as outlined above. The sentence in question is no longer included in the protocol.

[1] https://www.ncbi.nlm.nih.gov/pubmed/25569120

Reviewer: 2
Reviewer Name: Dr C L Gurudatt
Institution and Country: Professor and Head, Department of Anaesthesiology, Critical Care and Pain Medicine
JSS Medical College and Hospital, Mysore, INDIA
Competing Interests: None Declared

Congratulations to the authors for taking up this project as it is going to be beneficial in anticipating complications and preventing them once the algorithm is made and tested successfully.
Over 50,000 patients each year sustain a perioperative myocardial infarction (PMI) and hence prevention of a PMI is important to improve overall postoperative outcome. Thus PMI can also be added as a primary objective of the study along with respiratory failure and post operative kidney injury.

RESPONSE: We intend to investigate postoperative myocardial infarction. We have included myocardial infarction as a secondary outcome rather than a primary outcome because we expect to observe a lower incidence of postoperative myocardial infarction compared to postoperative acute kidney injury or postoperative acute respiratory failure.

Regarding the exclusion criteria, patients already in respiratory failure and underwent surgery can be excluded from the study.

RESPONSE: We have added preoperative mechanical ventilation as an exclusion criterion for the analysis of postoperative acute respiratory failure. A sentence describing this has been added to page 9.

Regarding the references, some of the references are not as per the Vancouver style. For references with 1 to 6 authors, list all authors. For references with more than 6 authors, list the first 6 authors then add 'et al.'

RESPONSE: We have amended the reference list to utilize Vancouver style.

VERSION 2 – REVIEW

REVIEWER	Dennis Toddenroth, MD Universität Erlangen-Nürnberg, Germany
REVIEW RETURNED	02-Jan-2018

GENERAL COMMENTS	After reading their responses as well as the ‘marked copy’, I would like to thank the authors for their accommodating revision, which in my view has improved the manuscript. If the review process at this

	point still leaves room for modifications, however, I would like to maintain: • That even though the proposed refinement of predictive models to “60-minute epochs” seems better-suited to underpin future real-time decision aids, deriving epoch-specific datasets from variable-length perioperative episodes may bring up additional issues that could be clarified, especially in the context of model evaluation. Plausible procedures might for instance involve the selection of a random epoch for each perioperative episode, or a replication of episode-specific outcome observations for every available hourly epoch. Hourly data subsets from the same episode, however, might violate conventional assumptions about the statistical independence between units of observation, so if epochs from the same perioperative episode were inadvertently distributed to test sets and validation sets, subtle autocorrelations in episode-specific data in conjunction with overfitting could produce overoptimistic performance estimates even if disjunctive *epochs* were used for training and validation. These considerations could also be exploited in order to better estimate the generalizable model performance in an unbiased fashion, for example by deliberately using epoch data from disjunctive *episodes* for model training and evaluation. The section on model evaluation based on ‘the historical database’ should thus clarify how either episodes (surgeries) or epochs are handled as observational units. • That some graphical overview or data flow diagram might really enhance the general accessibility of the planned investigation for readers who are not versed in predictive modeling – specifically as modeling based on epoch-specific episode subsets may increase the overall complexity of the entire analytical procedure. • That the freehanded widening of an point estimate of 80,750 expected cases (19,000 per annum over 4.25 years) to “approximately 80,000-90,000” cases seems slightly generous, in particular since the authors have tacitly removed their differentiation of observations that are anticipated to “be available for analysis”, and no indications of increasing case numbers over time are provided. • (Detail that I have missed during the initial review round: The “Data Analysis, Part 1” section refers to some “KLR” method, which is likely a spelling error, since this abbreviation seems to be never resolved.)

VERSION 2 – AUTHOR RESPONSE

• That even though the proposed refinement of predictive models to “60-minute epochs” seems better-suited to underpin future real-time decision aids, deriving epoch-specific datasets from variable-length perioperative episodes may bring up additional issues that could be clarified, especially in the context of model evaluation. Plausible procedures might for instance involve the selection of a random epoch for each perioperative episode, or a replication of episode-specific outcome observations for every available hourly epoch. Hourly data subsets from the same episode, however, might violate conventional assumptions about the statistical independence between units of observation, so if epochs from the same perioperative episode were inadvertently distributed to test sets and validation sets, subtle autocorrelations in episode-specific data in conjunction with overfitting could produce overoptimistic performance estimates even if disjunctive *epochs* were used for training and validation. These considerations could also be exploited in order to better estimate the generalizable model performance in an unbiased fashion, for example by deliberately using epoch data from disjunctive *episodes* for model training and evaluation. The section on model evaluation based on ‘the historical database’ should thus clarify how either episodes (surgeries) or epochs are handled as observational units.

We have added additional text on Page 14 to address this concern. The new text reads, “Each training, validation, or testing example will be a 60-minute epoch randomly selected from a single surgery. More than one epoch from the same surgery may be included if the surgery lasted long enough to generate more than one distinct 60-minute epoch. However, all epochs from the same surgery will be included either all in the training dataset, all in the validation dataset, or all in the testing dataset.”

• That some graphical overview or data flow diagram might really enhance the general accessibility of the planned investigation for readers who are not versed in predictive modeling – specifically as modeling based on epoch-specific episode subsets may increase the overall complexity of the entire analytical procedure.

We have added Figure 1 to help clarify the anticipated flow of information during the model-building process.

• That the freehanded widening of an point estimate of 80,750 expected cases (19,000 per annum over 4.25 years) to “approximately 80,000-90,000” cases seems slightly generous, in particular since the authors have tacitly removed their differentiation of observations that are anticipated to “be available for analysis”, and no indications of increasing case numbers over time are provided.

We have revised the statement on Page 7 to read “80,000-81,000 surgeries.”

• (Detail that I have missed during the initial review round: The “Data Analysis, Part 1” section refers to some “KLR” method, which is likely a spelling error, since this abbreviation seems to be never resolved.)

We intended for the abbreviation “KLR” to represent “kernel logistic regression.” Because the final revision of the manuscript includes only one reference to KLR, we have deleted the abbreviation and used the full phrase instead on Page 10.

VERSION 3 – REVIEW

REVIEWER	Dennis Toddenroth, MD Universität Erlangen-Nürnberg, Germany
REVIEW RETURNED	08-Feb-2018

GENERAL COMMENTS	In my view the revision has improved the manuscript, and I am also looking forward to the authors' upcoming publication of their upcoming modeling results.