Journal of Clinical Monitoring and Computing (2019) 33:949–951
https://doi.org/10.1007/s10877-019-00343-7

EDITORIAL

Predicting vital sign deterioration with artificial intelligence or machine learning

Simon T. Vistisen[1,2] · Alistair E. W. Johnson[3] · Thomas W. L. Scheeren[4]

Received: 17 June 2019 / Accepted: 24 June 2019 / Published online: 28 June 2019

Cardiorespiratory instability of patients during their hospital stay is a frequently occurring, undesired complication that often requires prompt treatment to prevent the downstream consequences of reduced oxygen delivery to tissues. This is one of the primary reasons why such patients have their vital signs, such as heart rate (HR), mean arterial pressure (MAP), respiratory rate, and peripheral oxygen saturation (SpO_2) more or less continuously monitored, at least in high care units such as the operating room, the intensive care unit, and the post anesthesia care unit. In this way, real-time detection of e.g. tachycardia, hypotension or hypoxia has made it possible to promptly *react* to deterioration, i.e. treating it shortly after it has occurred.

Still, most (if not all) emergency doctors, intensivists and anesthesiologist have experienced cases where earlier warning of impending cardiorespiratory deterioration could have saved precious time and perhaps even saved the life of a patient. A natural development of this type of monitoring is therefore *predictive analytics*: moving from real-time detection and *reaction* to deterioration to *proactive* treatment of impending deterioration likely holds great potential to improve outcomes for these patients. As an example, the hypotension prediction index has recently been introduced [1] and validated [2], a unitless number ranging from 0 to 100 derived from arterial waveform analysis with the help of machine learning techniques. Higher numbers indicate a higher probability of a hypotensive event from occurring within the following minutes, bearing the potential to prevent such events from occurring by proactive treatment.

In this issue of JCMC, Yoon et al. [3] describe a method of *predicting tachycardia as a surrogate for instability in the intensive care unit*. The study data was derived from the MIMIC-II Clinical and Waveform databases, which hold detailed clinical and monitoring information for each patient. The authors extracted data from 3-h periods preceding *tachycardia episodes* (n = 787) and *control periods* not preceding tachycardia (n = 705). Then, they used a multiparametric approach (multiple characteristics/features), where time series from each of the vital signs *HR, MAP, and* SaO_2 had features derived. Features were either statistical (means, standard deviations, regression coefficients, autocorrelations and entropy) or frequency based (amplitude measures of the spectra). The total of 42 vital sign features were fed into two *classifiers; one linear (logistic regression) and one non-linear (random forest).*

In order to evaluate the generalization performance of their approach, the authors undertook a standard approach of building models on a subset of data (the training set), and evaluating on a distinct subset never before seen by the classifier (the test set). The authors showed that it was possible to predict tachycardia with their algorithm with a compelling accuracy of 0.81 and area under the ROC curve of 0.87, using the random forest classifier.

With this accompanying editorial, we would like to highlight some aspects of these types of *predictive analytics* studies, where vital sign deteriorations are predicted, because the study of Yoon et al. [3] serves as a good example of these aspects. Throughout the remaining content of this editorial, Fig. 1 serves to illustrate the data workflow, which is typical

✉ Thomas W. L. Scheeren
t.w.l.scheeren@umcg.nl

Simon T. Vistisen
vistisen@clin.au.dk

1 Institute of Clinical Medicine, Aarhus University, Palle Juul-Jensens Boulevard 99, C319-128, 8200 Aarhus N, Denmark

2 Department of Anaesthesiology & Intensive Care, Aarhus University Hospital, Aarhus, Denmark

3 Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, E25-505, 77 Massachusetts Ave, Cambridge, MA 02139, USA

4 Department of Anaesthesiology, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9700RB Groningen, The Netherlands

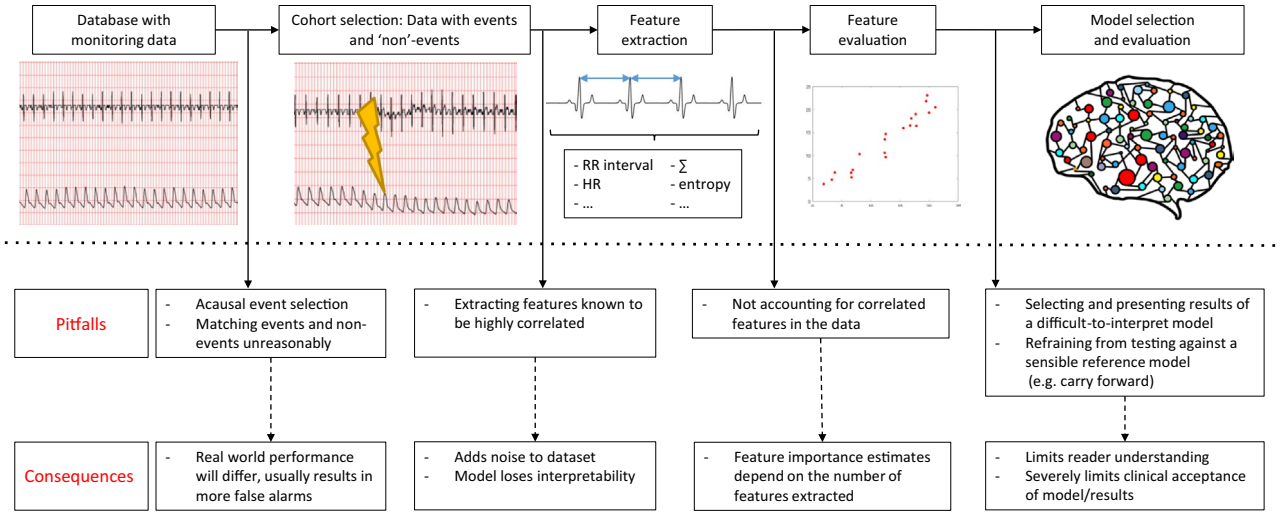

Fig. 1 Illustration of a typical analysis flow from raw physiologic data to a model to predict vital signs deterioration. Through this analysis flow, researchers make important choices as to (1) how (or if) to create a data subset for the event to predict, (2) which features to extract, (3) how to evaluate and/or select a subset of the extracted features – whether being a strict feature evaluation or an evaluation/selection that is embedded in the prediction model development itself, and (4) which prediction model to build. Each of these choices are associated with possible pitfalls, all of which should be sought remediated by wise decisions of the researchers in order to produce a prediction model that will convince readers of its value

for these studies and what pitfalls there may be for each step with the data.

1 Acausal data extraction

An ongoing challenge of classifying decompensation is the design of the cohort. If the existence of an event is used in the *creation* of the training dataset, then the model is fundamentally acausal. All evaluation measures are predicated on knowing who gets tachycardia, and knowing when they get it. We would expect that a clinical implementation of a model derived in this method would perform dramatically differently, as the set of patients presented to the model in practice will markedly differ, see Fig. 1. This change of patient "cohort" would likely result in an increase in the false positive rate of the algorithm, i.e. more false alarms than the current data identifies are expected. However, minimization of artefacts and false alarms to avoid alarm fatigue are crucial for a new algorithm to be implemented in clinical routine.

2 Matched patients

Another subtle but challenging aspect of event detection is deciding upon a suitable control population. While it is clear what time an event occurred for a positive case, it is decidedly more difficult to determine at what time to extract data for a control case—when was the non-event? Yoon et al. [3] have adopted a sensible strategy in this regard. In the absence of other knowledge, randomly sampling from the patient time series will reduce the impact of any unmeasured confounding on the data extraction, and should make the model broadly applicable to patient time series at any time. Other authors have explored matching controls based upon demographics or severity of illness, which usually increases the difficulty of the prediction task [4].

3 Features and model evaluation

3.1 Feature selection

An important aspect of the prediction model is which features are selected to enter the model. Features that are highly correlated with each other may simply just add noise to the prediction model, depending on the chosen model. Models need to be able to handle correlated features, see Fig. 1. In the present study by Yoon et al. [3.], there are several features that are likely highly correlated such as *mean heart rate*, *last 5 min mean heart rate (mean_5min_HR)*, *last 10 min mean heart rate (mean_10min_HR)*, and the summarized amplitudes of the *fast Fourier transform of heart rate.* The authors chose two models, which both incorporate some kind of feature selection (regularized lasso logistic regression and random forest regression). The use of regularization is good as that model selects the feature that is most correlated to the outcome first. The feature selection procedure for the random forest model is more difficult to interpret

since each tree only sees a subset of features, but we do not expect a feature selection problem. Yet, in the results of the study, the above mentioned (presumably) highly correlated features are all ranking very high when classifying with the random forest model (e.g. the four variables are all in top-5 when predicting tachycardia 10 min in advance). Since the list is a summarized rank from the study's cross-validations, it is somewhat unclear which of these correlated features would be most useful in a final model to implement into a monitor for tachycardia prediction. One approach that might have accommodated a possible issue with expectedly correlated features could be to use e.g. mean_5min_HR *and* (mean_5min_HR—mean_10min_HR). In this way, the features would likely be less correlated and still interpretable, which was a goal of the study.

3.2 Reference model

In most studies, the *current* and *historic* values of the vital sign to be predicted are also used for the prediction itself, sometimes solely. The present study defined tachycardia as a HR > 130/min—a threshold decision appropriately advised by data, which the authors should be commended for. Clinical intuition dictates that a HR of e.g. 125/min is obviously associated with a higher probability of impending tachycardia compared with a lower HR (of e.g. 80/min). Therefore, in this situation, the model should as a minimum be compared with a very simple reference model including the "current" value and possibly also its historic values, see Fig. 1—perhaps implemented as a trend value (e.g. a 1st derivative over a couple of minutes). Such an approach is often referred to as *sample-and-hold* or *carry forward*. This is what any bedside clinician would intuitively do anyway—and has the mental capacity to do. In addition, all modern monitors have a tachycardia alarm, which essentially does this work already. Comparing with a simple carry forward model for prediction (and hopefully performing better) will increase the chance that clinicians would eventually endorse the new algorithm. Performing better would mean that more events are detected/predicted and/or the frequency of false alarms is reduced. Otherwise it is unclear whether a new algorithm is simply *crossing the river to get water* – i.e. adding complexity when it is not needed. In the present study, Yoon et al. [3] have unfortunately not reported in their paper, how well HR (combined with its trend) alone could predict tachycardia. The reported high ranking of (essentially) heart rate features, makes this reference model reporting even more important.

4 Conclusion

In conclusion, innovative ways of improving the monitoring of our patients' vital signs in the perioperative period are highly appreciated. This includes validation, integration and analysis of all vital signs (big data), and the development of predictive algorithms based on machine learning. Automated continuous minimally and non-invasive monitoring combined with machine learning-based algorithms will enable subtle changes in vital signs to be recognized early and thus allows earlier treatment or even prevention of hemodynamic catastrophic events, most probably improving patient safety and outcome. However, not every innovation subsumed under the hot topics like artificial intelligence, big data and machine learning should automatically be embraced without critical evaluation. Otherwise, we risk ending up with just another version of Hans Christian Andersen's story of *The Emperor's New Clothes* where eventually clinicians refrain from endorsing and adopting these exciting new methods.

Compliance with ethical standards

Conflict of interest TWLS received research funding and honoraria from Edwards Lifesciences and Masimo Inc. (Irvine, CA, USA) for consulting and lecturing and from Pulsion Medical Systems SE for lecturing in the past. TWLS is associate editor of the Journal of Clinical Monitoring and Computing but had no role in the handling of this paper. The other authors declare no conflict.

References

1. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. Anesthesiology. 2018;129(4):663–74.
2. Davies SJ, Vistisen ST, Jian Z, Hatib F, Scheeren TWL. Ability of an arterial waveform analysis-derived hypotension prediction index to predict future hypotensive events in surgical patients. Anesth Analg. 2019. https://doi.org/10.1213/ANE.0000000000004121.
3. Yoon JH, Mu L, Chen L, Dubrawski A, Hravnak M, Pinsky MR, et al. Predicting tachycardia as a surrogate for instability in the intensive care unit. J Clin Monit Comput. 2019. https://doi.org/10.1007/s10877-019-00277-0.
4. Futoma J, Hariharan S, Sendak M, Brajer N, Clement M, Bedoya A, et al. An Improved multi-output gaussian process RNN with real-time validation for early sepsis detection. Proc Machine Learn Healthcare. 2017;68:arXiv:1708.05894.