ANESTHESIOLOGY

Performance of the Hypotension Prediction Index May Be Overestimated Due to Selection Bias

Johannes Enevoldsen, M.D., Simon T. Vistisen, M.Sc., Ph.D.

ANESTHESIOLOGY 2022; 137:283–9

ABSTRACT

The Hypotension Prediction Index is a proprietary prediction model incorporated into a commercially available intraoperative hemodynamic monitoring system. The Hypotension Prediction Index uses multiple features of the arterial blood pressure waveform to predict hypotension. The index publication introducing the Hypotension Prediction Index describes the selection of training and validation data. Although precise details of the Hypotension Prediction Index algorithm are proprietary, the authors describe a selection process whereby a mean arterial pressure (MAP) less than 75 mmHg will always predict hypotension. We hypothesize that the data selection process introduced a systematic bias that resulted in an overestimation of the current MAP value's ability to predict future hypotension. Since current MAP is a predictive variable contributing to Hypotension Prediction Index, this exaggerated predictive performance likely also applies to the corresponding Hypotension Prediction Index value. Other existing validation studies appear similarly problematic, suggesting that additional validation work and, potentially, updates to the Hypotension Prediction Index model may be necessary.

(ANESTHESIOLOGY 2022; 137:283–9)

The Hypotension Prediction Index is a prediction model implemented in the HemoSphere advanced monitoring platform (Edwards Lifesciences, USA). The development and validation of the model have been previously reported.[1] In that report, hypotension was defined as mean arterial pressure (MAP) less than 65 mmHg for at least 1 min. In brief, the Hypotension Prediction Index uses logistic regression with proprietary predictor variables or features derived from the arterial waveform to predict the occurrence of hypotension 5, 10, or 15 min later. The researchers investigated 3,022 individual features and 2,603,125 combined features. After feature selection, the final proprietary model (resulting in Hypotension Prediction Index) included 23 of these features.[1–3] The Hypotension Prediction Index is presented on the monitor as a value between 0 and 100 with higher values indicating a higher likelihood of imminent hypotension.

The external validation of the model shows a compelling area under the receiver operating characteristics curve (AUC) of 0.95 for predicting hypotension 5 min before the event.[1] However, the corresponding receiver operating characteristics curve presented in the index publication's figure 3 (reprinted and adapted here as fig. 1) looks highly improbable. This receiver operating characteristics curve presents the sensitivity (true predictions of hypotension

out of all hypotensive events) and corresponding specificity (true predictions of nonhypotension out of all non-hypotensive events) for all possible thresholds ("alarm limits") of Hypotension Prediction Index. The receiver operating characteristics curve in the index publication shows that a specific (yet unspecified in the publication) threshold is associated with a specificity of approximately 100% and a sensitivity greater than 60%.[1] When the specificity is 100%, there are no false predictions of hypotension. Therefore, the positive predictive value (true predictions of hypotension out of all predictions of hypotension) is also 100%. This means that Hypotension Prediction Index values above this threshold are always associated with future hypotension. We find it difficult to imagine that it is possible to predict with approximately 100% certainty that hypotension will occur while maintaining a reasonable sensitivity. Herein, based on the data selection described in the index publication and a computer simulation illustrating the consequence of that data selection, we present the hypothesis that the index publication and several reported validation studies of the Hypotension Prediction Index may contain a systemic statistical bias influencing its predictive abilities.

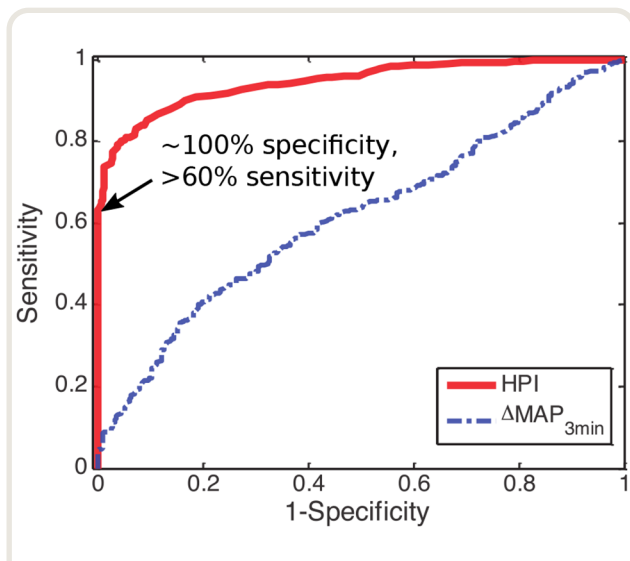<zdoi; DOI:10.1097/ALN.0000000000004320>

Fig. 1. External validation of Hypotension Prediction Index (HPI), 5 min before the event (reprinted and adapted from fig. 3, bottom left from the index publication[1]). MAP, mean arterial pressure.

Data Selection May Be Biased

A probable explanation for the reported high specificity can be found in the methods section of the index publication:[1]

"**Model Feature Selection and Training:** A hypotensive event was calculated by identifying a section of at least 1-min duration such that all data points in the section showed MAP < 65 mmHg. An event, or positive data point, was chosen as the sample recorded 5, 10, or 15 min before the hypotensive event. A nonhypotensive event was calculated by identifying a 30-min continuous section of data points such that the section was at least 20 min apart from any hypotensive event, and all data points in that section showed MAP > 75 mmHg. A nonevent, or negative data point, was the center point of the nonhypotensive event."

Thus, a *non*hypotensive event is defined as a 30-min section where MAP is above 75 mmHg. The sample used to predict a nonhypotensive event is the center point of this section. Therefore, a sample corresponding to a non-hypotensive event will always have a MAP greater than 75 mmHg, while samples corresponding to hypotensive events can have any MAP (this selection of events and samples is illustrated in the model-development paper's supplementary fig. 2, reproduced here as fig. 2). Because of this selection, a sample with MAP less than 75 mmHg will *always* correspond to a future hypotensive event in the training *and* the test sets. This can explain why it was possible for the Hypotension Prediction Index to predict hypotension with

100% specificity: the Hypotension Prediction Index could achieve 100% specificity by simply reflecting the current MAP value—without 22 additional features. Given that the algorithm is proprietary, we do not know which features are included in the Hypotension Prediction Index model, but we do know that MAP was one of the candidates. If the model training was effective, Hypotension Prediction Index should learn that a current MAP less than 75 mmHg is always associated with future hypotension.

As an analogy, one can imagine excluding all subjects younger than 60 yr from the samples corresponding to nonhypotensive events. Then a prediction model applied to the remaining data could simply predict "hypotension" for all subjects younger than 60 yr and be correct every time.

Given this selection problem, receiver operating characteristics curves for MAP and the Hypotension Prediction Index should both show a skew toward high specificity (as shown in fig. 1). Because MAP monitoring is commonly used to titrate blood pressure treatment, we consider that the added clinical value of Hypotension Prediction Index is the difference between the Hypotension Prediction Index's and MAP's ability to predict hypotension. Unfortunately, MAP's ability to predict future hypotension was not reported in the index publication. Instead, the Hypotension Prediction Index was compared to ΔMAP (the change in MAP during, *e.g.,* 3 min), a comparator that is both an unintuitive and poor predictor of hypotension, as discussed in a recent paper.[4]

The likely existence of a significant selection bias in the development and validation of the Hypotension Prediction Index is the *key message* of this discussion.

In the following sections, we first use simulated data to visualize the effect of the biased selection, and demonstrate how the selection can result in skewed receiver operating characteristics curves similar to that in figure 1. Second, we address questions that may naturally arise: Does this bias also affect the numerous validation studies in the literature? What about all the other features used in the Hypotension Prediction Index model? How could the Hypotension Prediction Index be validated appropriately? What is the effect of using Hypotension Prediction Index in clinical trials?

Visualization of the Selection Problem

To visualize how the selection problem can artificially enhance the current MAP's ability to predict hypotension (and thereby likely also the Hypotension Prediction Index's ability), we performed a simple simulation (fig. 3). The simulation is not an attempt to produce realistic data but only serves to visualize how the selection problem can result in a "skewed" receiver operating characteristics curve with very high specificity.

We generated normally distributed data representing the current MAP values available for predicting future hypotension. The MAP values corresponding to hypotensive events have a lower mean than the MAP values corresponding to nonhypotensive events, making current MAP a modest predictor of hypotension (fig. 3A). We then impose a selection
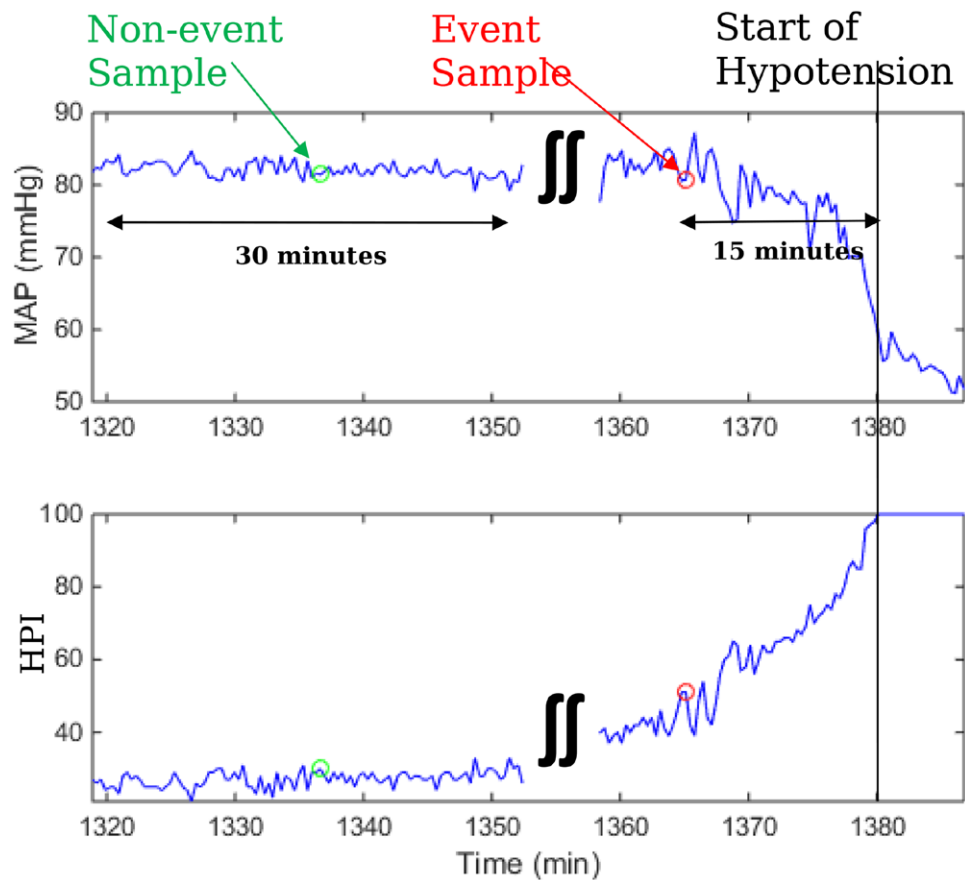
Fig. 2. Illustration of the selection of hypotensive events and nonhypotensive events, and the samples corresponding to each type of event (reprinted from the model development paper's supplementary fig. 2[1]). A nonevent sample is the center of a 30-min section where all data points have a mean arterial pressure (MAP) greater than 75 mmHg. HPI, Hypotension Prediction Index.

that removes current MAP values less than 75 mmHg for nonhypotensive events. The receiver operating characteristics curve corresponding to this biased selection has a characteristic "skew" toward high specificity and a markedly increased AUC (fig. 3B). The simulation computer code and details are presented in Supplemental Digital Content 1 (http://links.lww.com/ALN/C878).

Most Hypotension Prediction Index Validation Studies Seem Biased

We are aware of eight subsequent validation studies of Hypotension Prediction Index.[3,5–11] In these studies, the Hypotension Prediction Index was either downloaded directly from a HemoSphere or EV1000 monitor (Edwards Lifesciences, USA), or calculated *post hoc* from the arterial waveform using the same algorithm. Only one of these studies compares the predictive performance of the Hypotension Prediction Index to that of the concurrent MAP value.

Three studies used the same data selection as the index publication and presented similarly skewed receiver operating characteristics curves with high specificity[5–7] (Wijnberge *et al.*[7] also did the "forward" analysis described below and is counted there as well). Another three showed a "skewed" receiver operating characteristics curve with a very high specificity, but did not specify exactly how nonhypotensive events were selected.[3,8,9] Two used a "forward analysis" starting with a Hypotension Prediction Index alarm and looking in the next 20 min for hypotension.[7,10] Both showed a high predictive performance of Hypotension Prediction Index (*e.g.*, a positive predictive value of 80% and a negative predictive value of 96% at a Hypotension Prediction Index threshold of 85);[7] however, the method for selecting individual prediction–outcome pairs was not explicitly described in the methods sections. Also, the Hypotension Prediction Index's predictive ability was not compared with the concurrent MAP value's predictive ability in these two studies, further complicating the interpretation of these results, as it remains unclear to what extent
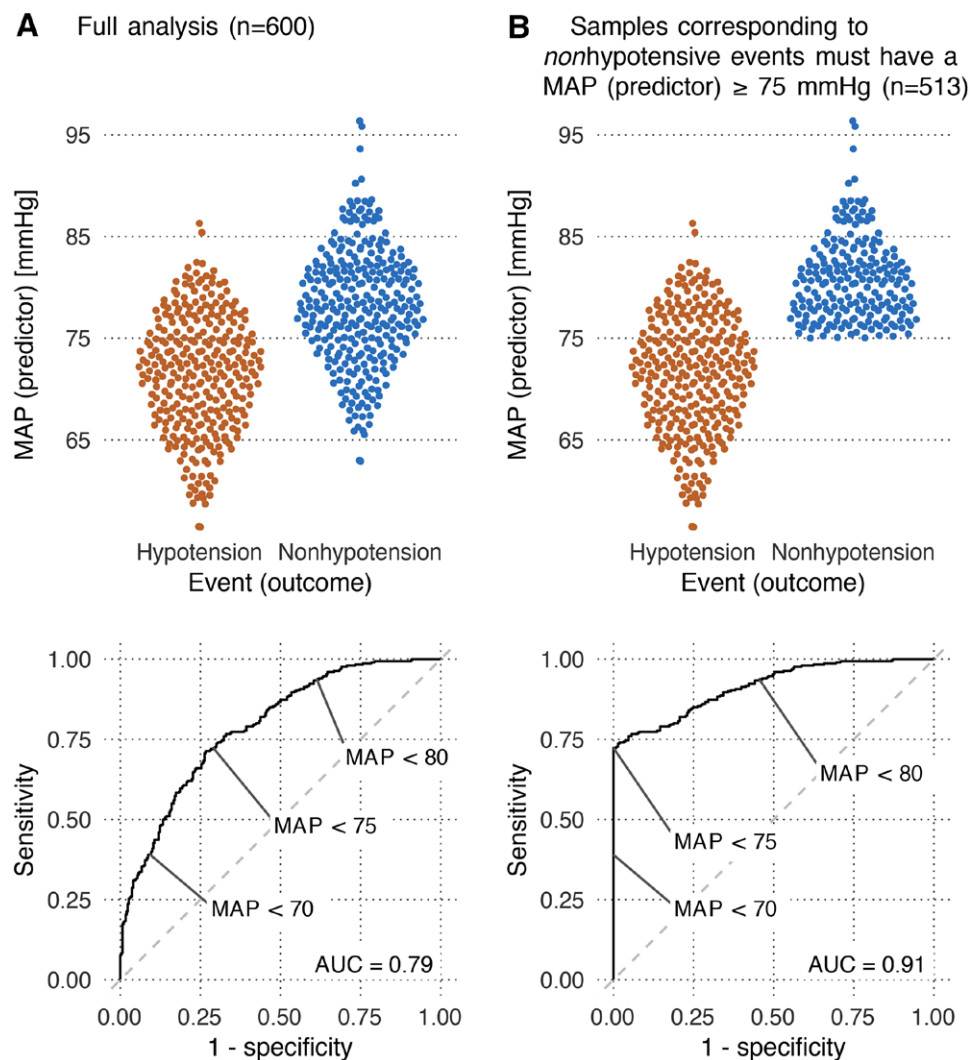
Fig. 3. Simulation of the selection problem. *Columns*, *A* and *B*, illustrate different data selection strategies. *Upper panels* show simulated mean arterial pressure (MAP; mmHg) values for samples corresponding to hypotensive events and nonhypotensive events. *Lower panels* are receiver operating characteristics curves showing MAP's ability to discriminate hypotensive events from nonhypotensive events. The simulation is not an attempt to produce realistic data. It only serves to illustrate how the selection problem can result in a "skewed" receiver operating characteristics curve with very high specificity.

the Hypotension Prediction Index's predictive performance is driven by the concurrent MAP value alone. The study by Ranucci *et al.* selected Hypotension Prediction Index values corresponding to both hypotensive events and non-hypotensive events 5 to 7 min before the event.[11] By using data before the nonhypotensive events, they avoided creating a data set where samples corresponding to nonevents had an artificially high MAP, thereby avoiding the selection bias described above. The authors presented a "symmetric" receiver operating characteristics curve with an AUC of 0.768 for the Hypotension Prediction Index's ability to predict hypotension. While this AUC may be more realistic than what is reported in other studies, it was based on just 77 hypotensive events, and the Hypotension Prediction Index performance was not compared with that of the concurrent MAP value.

Only Davies *et al.*[3] compared Hypotension Prediction Index to MAP. Therein, the Hypotension Prediction Index predicted hypotension markedly better than the concurrent MAP value (AUC 0.926 *vs.* 0.807, respectively), and only the Hypotension Prediction Index receiver operating characteristics curve was skewed toward high specificity (see figure 2 in Davies *et al.*[3]). The methods section in the paper does not provide enough information about the data selection to explain why the Hypotension Prediction Index's receiver operating characteristics curve indicates the

presence of a selection bias while the MAP's receiver operating characteristics curve does not.

Biased Data Make Biased Models

Using multiple features from the arterial waveform to predict hypotension is an admirable idea. However, we speculate that most of the potential added value is lost, because the selection bias forced the Hypotension Prediction Index to learn almost solely from MAP in its development: if MAP is less than 75 mmHg, the patient *will* be classified as hypotensive. When MAP is less than 75 mmHg, other features can only impair this "perfect" prediction. This creates a biased model that overestimates the risk of hypotension. The model will presumably overrepresent MAP and underrepresent other waveform features (at least when MAP is less than 75 mmHg). If our speculation is true, we should expect that the Hypotension Prediction Index is almost a one-to-one transformation of the concurrent MAP (where the Hypotension Prediction Index is high when MAP is low) with only a small impact from other features. The concurrent MAP and Hypotension Prediction Index values in figure 2 and the index publication's figure 5 (not reprinted here) exemplify this one-to-one transformation.

Hypotension Prediction Index Should Be Revalidated

As the Hypotension Prediction Index algorithm may have been subject to selection bias in its development, and since most subsequent validation studies indicate a similar problem, the Hypotension Prediction Index may not predict hypotension as accurately as reported. We suggest that data from previously published Hypotension Prediction Index validation studies be reanalyzed, paying particular attention to ensure an unbiased selection of hypotensive events and nonevents. A more reasonable selection of nonevents could be similar to that by Ranucci *et al.*[11] First, select all 1-min sections with MAP greater than 65 mmHg (optionally, also do an analysis requiring nonevents to have MAP greater than 75 mmHg, corresponding to the intentional "gray zone" implemented in the index publication and most validation studies. Samples before the event should *not* be restricted). Then, exclude nonevents in the first 15 min after an event. For both events and nonevents, the predictor should be Hypotension Prediction Index or MAP 5, 10, or 15 min earlier. This design will allow a reasonable receiver operating characteristics curve analysis for comparing the Hypotension Prediction Index's and MAP's predictive abilities, but it will likely result in an overrepresentation of nonevents, so reporting positive and negative predictive values will not be meaningful (see the section The Case–Control Design Itself Is Problematic). In our view, it is the *difference* between the Hypotension Prediction Index's and the concurrent MAP value's predictive abilities (*e.g.*, AUCs) that represents the *added clinical* value of the Hypotension Prediction Index over simply using MAP to guide blood pressure treatment. It is imperative that any comparison of prediction methods be based on the exact same data selection and outcome labeling.

A recent paper appropriately suggested that a complex and proprietary algorithm like the Hypotension Prediction Index should be compared to a simple model that represents current clinical practice (as opposed to ΔMAP).[4] The authors suggested a linear extrapolation from the current MAP value and the MAP value 1 min earlier ("LepMAP"). The study did not compare LepMAP and Hypotension Prediction Index directly, but when, in a secondary analysis (termed B-analysis), the authors applied a data selection to match that of the index publication, they found an AUC of 0.93 for LepMAP's prediction of hypotension 2 min into the future, with receiver operating characteristics curves that were skewed toward high specificity. In the study's A-analysis, which did not enforce the index publication's selection, they compared the predictive ability of LepMAP to that of the concurrent MAP, and found that they were not statistically different, although the concurrent MAP's AUC had higher point estimates. However, the paper does not address that the B-analysis creates a selection bias.

The Case–Control Design Itself Is Problematic

The data selection in the index publication and most subsequent validation reports (including that by Ranucci *et al.*[11] and what we describe in the section Hypotension Prediction Index Should Be Revalidated) is based on a case–control design: it begins with selection of cases (hypotensive events) and controls (nonhypotensive events). Afterward, predictors (Hypotension Prediction Index or the waveforms used to calculate Hypotension Prediction Index) are selected based on these cases and controls (*e.g.*, 5 min before hypotensive events and the midpoints of nonhypotensive events). Hence, most Hypotension Prediction Index values we observe in a clinical setting will not be represented in the analysis, because they are neither 5 min before a hypotensive event nor the midpoint of a nonhypotensive event. Therefore, even without the described selection bias, this case–control design is known to give results that may not be valid in a clinical setting.[12] This problem may be exacerbated by the exclusion of "gray zone" outcomes (outcome MAP between 65 and 75 mmHg). The index publication argues that while false-positive Hypotension Prediction Index alarms with outcomes in this "gray zone" are not included in the receiver operating characteristics curve analyses, this is not an important limitation, and that these false positives could even be beneficial.[1] Conversely, a review of the Hypotension Prediction Index argues that these false-positive alarms may lead to overtreatment.[13] The case–control design can be useful for model development, but since the proportion of hypotensive events does not represent the true probability of hypotension, it does not

allow calculation of clinically meaningful positive and negative predictive values. Neither the positive predictive value of 12.6% at Hypotension Prediction Index greater than 85, reported by Ranucci *et al.*,[11] nor the 93.2% at Hypotension Prediction Index greater than 39, reported in the index publication,[1] should be interpreted as the probability of imminent hypotension in a continuously monitored patient.

The process of representing clinical data as outcomes and predictors for a risk prediction problem has recently been introduced as the *framing* of the problem.[14] A clinically relevant framing for investigating the Hypotension Prediction Index's ability to predict hypotension could be as follows: When the Hypotension Prediction Index gives an alarm, how often does hypotension actually occur, *e.g.*, 3 to 10 min later (*i.e.*, positive predictive value); and when hypotension occurs, was there an Hypotension Prediction Index alarm 3 to 10 min earlier (*i.e.*, sensitivity)? This way, all false-positive predictions will be counted—not just the ones corresponding to selected "nonevents." We acknowledge that it is not trivial to decide what constitutes a single prediction, and we do not contend that the suggested framing is the single best solution. In the end, it is the model developers' responsibility to make sure that an algorithm is developed and validated to truly address the targeted clinical problem.

Effect of Hypotension Prediction Index in Clinical Trials

Four randomized controlled trials have investigated the Hypotension Prediction Index's preventive effects on hypotension.[2,15–17] Three of these trials, and an additional retrospective study,[18] showed less hypotension in the Hypotension Prediction Index guided group compared to a control group (the trials vary in their outcome definitions and precise management protocols, but the characteristics are beyond the scope of this analysis).[2,16,17] One trial also investigated the effect on postoperative hypotension and found no difference in their primary outcome between the Hypotension Prediction Index–guided group and the standard care group.[19] The fourth and largest study noted no difference between the Hypotension Prediction Index–guided group and the control group in the primary outcome variable: amount of hypotension (time-weighted average MAP less than 65 mmHg).[15] That hypotension may be reduced is important, but these results could be due to increased clinician awareness to hemodynamics (something that might also be achieved with a MAP alarm or treatment threshold of, *e.g.*, 75 mmHg), and do not validate the predictive ability of Hypotension Prediction Index *per se*. In addition, preventing hypotension may come at the expense of overtreatment as indicated in the trial by Tsoumpa *et al.*, where *time spent in hypertension* was increased in the Hypotension Prediction Index–guided group.[17]

Conclusions

A selection bias in the development of the Hypotension Prediction Index may explain a relevant proportion of its reported predictive ability. The same bias seems present in most of the subsequent validation studies. In light of this, we are not yet convinced that using the Hypotension Prediction Index to predict hypotension is meaningfully better than using the concurrent value of MAP. We suggest that data from validation studies be reanalyzed considering the potential for this selection bias, and that the predictive performance of the Hypotension Prediction Index be compared with the predictive performance of the concurrent MAP value.

Correspondence

Address correspondence to Dr. Enevoldsen: Department of Clinical Medicine, Aarhus University, Palle Juul-Jensens Boulevard 99, 8200 Aarhus N, Denmark. enevoldsen@clin.au.dk. This article may be accessed for personal use at no charge through the Journal Web site, www.anesthesiology.org.

Supplemental Digital Content

Computer code for selection bias visualization, http://links.lww.com/ALN/C878

References

1. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, Rinehart J, Cannesson M: Machine-learning algorithm

to predict hypotension based on high-fidelity arterial pressure waveform analysis. ANESTHESIOLOGY 2018; 129:663–74

2. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, Schenk J, Terwindt LE, Hollmann MW, Vlaar AP, Veelo DP: Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: The HYPE randomized clinical trial. JAMA 2020; 323:1052–60

3. 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 2020; 130:352–9

4. Jacquet-Lagrèze M, Larue A, Guilherme E, Schweizer R, Portran P, Ruste M, Gazon M, Aubrun F, Fellahi JL: Prediction of intraoperative hypotension from the linear extrapolation of mean arterial pressure. Eur J Anaesthesiol 2022; 39:574–81

5. Maheshwari K, Buddi S, Jian Z, Settels J, Shimada T, Cohen B, Sessler DI, Hatib F: Performance of the Hypotension Prediction Index with non-invasive arterial pressure waveforms in non-cardiac surgical patients. J Clin Monit Comput 2021; 35:71–8

6. Frassanito L, Giuri PP, Vassalli F, Piersanti A, Longo A, Zanfini BA, Catarci S, Fagotti A, Scambia G, Draisci G: Hypotension Prediction Index with non-invasive continuous arterial pressure waveforms (ClearSight): Clinical performance in gynaecologic oncologic surgery. J Clin Monit Comput 2021 doi:10.1007/s10877-021-00763-4

7. Wijnberge M, van der Ster BJP, Geerts BF, de Beer F, Beurskens C, Emal D, Hollmann MW, Vlaar APJ, Veelo DP: Clinical performance of a machine-learning algorithm to predict intra-operative hypotension with noninvasive arterial pressure waveforms: A cohort study. Eur J Anaesthesiol 2021; 38:609–15

8. Frassanito L, Sonnino C, Piersanti A, Zanfini BA, Catarci S, Giuri PP, Scorzoni M, Gonnella GL, Antonelli M, Draisci G: Performance of the Hypotension Prediction Index with noninvasive arterial pressure waveforms in awake cesarean delivery patients under spinal anesthesia. Anesth Analg 2022; 134:633–43

9. Shin B, Maler SA, Reddy K, Fleming NW: Use of the Hypotension Prediction Index during cardiac surgery. J Cardiothorac Vasc Anesth 2021; 35:1769–75

10. van der Ven WH, Terwindt LE, Risvanoglu N, Ie ELK, Wijnberge M, Veelo DP, Geerts BF, Vlaar APJ, van der Ster BJP: Performance of a machine-learning algorithm to predict hypotension in mechanically ventilated patients with COVID-19 admitted to the intensive care unit: A cohort study. J Clin Monit Comput 2021 doi:10.1007/s10877-021-00778-x

11. Ranucci M, Barile L, Ambrogi F, Pistuddi V; Surgical and Clinical Outcome Research (SCORE) Group: Discrimination and calibration properties of the hypotension probability indicator during cardiac and vascular surgery. Minerva Anestesiol 2019; 85:724–30

12. Yuan W, Beaulieu-Jones BK, Yu KH, Lipnick SL, Palmer N, Loscalzo J, Cai T, Kohane IS: Temporal bias in case-control design: Preventing reliable predictions of the future. Nat Commun 2021; 12:1107

13. van der Ven WH, Veelo DP, Wijnberge M, van der Ster BJP, Vlaar APJ, Geerts BF: One of the first validations of an artificial intelligence algorithm for clinical use: The impact on intraoperative hypotension prediction and clinical decision–making. Surgery 2021; 169:1300–3

14. Lauritsen SM, Thiesson B, Jørgensen MJ, Riis AH, Espelund US, Weile JB, Lange J: The framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. NPJ Digit Med 2021; 4:158

15. Maheshwari K, Shimada T, Yang D, Khanna S, Cywinski JB, Irefin SA, Ayad S, Turan A, Ruetzler K, Qiu Y, Saha P, Mascha EJ, Sessler DI: Hypotension Prediction Index for prevention of hypotension during moderate- to high-risk noncardiac surgery. ANESTHESIOLOGY 2020; 133:1214–22

16. Schneck E, Schulte D, Habig L, Ruhrmann S, Edinger F, Markmann M, Habicher M, Rickert M, Koch C, Sander M: Hypotension Prediction Index based protocolized haemodynamic management reduces the incidence and duration of intraoperative hypotension in primary total hip arthroplasty: A single centre feasibility randomised blinded prospective interventional trial. J Clin Monit Comput 2020; 34:1149–58

17. Tsoumpa M, Kyttari A, Matiatou S, Tzoufi M, Griva P, Pikoulis E, Riga M, Matsota P, Sidiropoulou T: The use of the Hypotension Prediction Index integrated in an algorithm of goal directed hemodynamic treatment during moderate and high-risk surgery. J Clin Med 2021; 10:5884

18. Grundmann CD, Wischermann JM, Fassbender P, Bischoff P, Frey UH: Hemodynamic monitoring with Hypotension Prediction Index *versus* arterial waveform analysis alone and incidence of perioperative hypotension. Acta Anaesthesiol Scand 2021; 65:1404–12

19. Schenk J, Wijnberge M, Maaskant JM, Hollmann MW, Hol L, Immink RV, Vlaar AP, van der Ster BJP, Geerts BF, Veelo DP: Effect of Hypotension Prediction Index-guided intraoperative haemodynamic care on depth and duration of postoperative hypotension: A substudy of the Hypotension Prediction trial. Br J Anaesth 2021; 127:681–8