Augmenting Decision Competence in Healthcare Using AI-based Cognitive Models

Niklas Keller
Simply Rational – The Decision Institute
Berlin, Germany
Dept. of Anesthesiology, Division for Operative Intensive Care Medicine
Charité Universitätsmedizin
Harding Centre for Risk Literacy, University of Potsdam, Potsdam
niklas.keller@simplyrational.de

Mirjam A. Jenny
Science Communication Unit
Robert-Koch-Institute
Berlin, Germany
Harding Centre for Risk Literacy
University of Potsdam, Potsdam
Center for Adaptive Rationality, Max-Planck Institute for Human Development, Berlin
jennym@rki.de

Claudia A. Spies
Dept. of Anesthesiology, Division for Operative Intensive Care Medicine
Charité Universitätsmedizin
Berlin, Germany
Claudia.spies@charite.de

Stefan M. Herzog
Centre for Adaptive Rationality
Max-Planck-Institute for Human Development
Berlin, Germany
herzog@mpib-berlin.mpg.de

***Abstract*—In many critical decisions, such as in medicine, transparency of the underlying decision process is critical. This extends to decision processes that are supported by artificial intelligence. Rather than using a post-hoc explainability approach from explainable AI research (e.g., SHAP or LIME), we develop and test an intrinsically transparent and intuitively interpretable model developed from cognitive science, fast-and-frugal trees, in a comparative analysis with state-of-the-art machine learning models. The resultant decision support can be easily implemented as laminated pocket card, augmenting the decision competence of physicians rather than replacing it.**

Keywords—augmented intelligence, post-operative risk stratification, decision support, fast-and-frugal trees, explainable artificial intelligence

I. Introduction

With a slow but steady shift from implementation to continuous operations, the potential of artificial intelligence (AI) for medical decision-making is being intensively and critically discussed. Especially in medicine, the use of complex and non-transparent AI technologies generates various practical and ethical challenges [1, 2, 3]. An example is the question under what conditions experts find non-transparent algorithmic recommendations acceptable for critical decisions [4]. There are new legal frameworks, such as the General Data Protection Regulation (GDPR) or the European Charter of Patients' Rights which require a minimum of explainability for AI-supported decisions. Scepticism of AI applications in healthcare continues to grow as more organizations try to implement these tools in their decision processes with mixed or sometimes disconcerting results. Zech et al. found, for example, that in 3 out of 5 cases, the performance of neural networks for pneumonia diagnosis deteriorated significantly when applied to different sites (i.e., different hospitals) [5]. In one case, the AI had learned that the word "portable" on the x-ray was key predictor of risk. This is because severely ill patients or those undergoing surgery cannot go to the radiology department and have to be x-rayed using a portable machine. This is, of course, a pure artefact and has nothing to do with underlying medical predictors of pneumonia risk. Critically, with black box models such problems surface only accidentally or when they are particularly egregious. Often, they go unnoticed for extensive periods of time. Secondly, even if such a flaw is discovered, it is not clear how one should go about "debugging" the model. Simply remove the problematic feature from the model? Or maybe a certain piece of data and re-run the analysis? Will the issue go away?

As a response to this and other high-stakes decisions in finance, health and judicial decision making, the field of eXplainable AI (XAI) has gained a lot of attention. In a nutshell, XAI uses further complex methods such as LIME [6] or SHAP [7, 8] to assess the degree to which different predictors contributed to a specific output of a black box algorithm for a particular decision (or generally for its decisions). These post-hoc explanations are thus themselves only models, i.e., approximations, of the underlying black box model. Subsequently, the explanation will deviate from the actual underlying decision process of the black box model. An XAI-method that delivers the "correct" explanation 80% of the time will provide a faulty explanation 20% of the time. As with the underlying black box model, there is no way to tell when this is the case, opening the potential for issues such as infinite regress, i.e., using approximations of black box models with XAI interpretations which may then themselves be subject to being modelled in an attempt to try to understand when deviations occur, these models again being approximations etc..

Underlying all of these efforts is the widespread and fundamental assumption that complex AI models achieve a higher predictive performance than less complex models (see Fig. 1). Under this assumption, a compromise between transparency and accuracy inevitable. This compromise is also seemingly wholly accepted by the XAI community,

978-1-7281-5382-7/20/$31.00 ©2020 IEEE

which employs the driving principle that explainability is a roadblock on the way to the use of more "sophisticated" (i.e., more complex) models and, as we get better at explainability, we can gradually employ more and more complex models in domains in which explainability is critical (although some dissenting voices exist, see [1]).

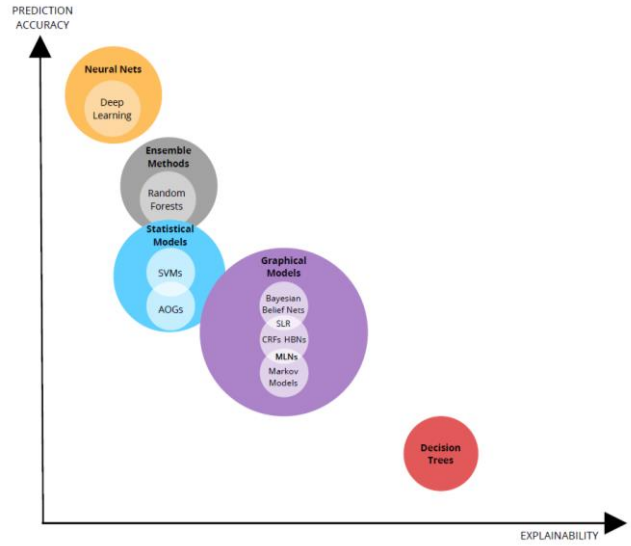

Fig. 1. The classic "effort-accuracy trade-off" assumption as presented by the Defence Advanced Research Projects Agency (DARPA) in their evaluation of the potential of AI. SVM = Support Vector Machine; AOG = And-Or-Graphs; CRF = Conditional Random Field; HBN = Hierarchical Bayesian Network; MLN = Markov Logic Networks; SLR = Simple Linear Regression.

We propose a fundamentally different approach to the issue of transparency, or the lack thereof, in medical (and other high stakes) decision support: to use intrinsically interpretable models instead. Data scientific studies have shown that in many areas of application simple, intuitive models can achieve a similar or even higher predictive performance than complex black box models from machine learning and artificial intelligence (e.g. Random Forests, Support Vector Machines, Artificial Neural Networks; [1, 9, 10]). Specific to medicine, a recent systematic review of clinical prediction models based on 71 studies showed that models from machine learning have no performance advantage over statistical standard procedures such as logistic regression [11], which can be represented as a nomogram (Figure 2c) and thus used as (relatively) simple decision support.

Furthermore, it has been shown in the cognitive sciences that simple, transparent, heuristic models that are intuitively used by humans and can be learned quickly, can achieve similarly high prediction performance as complex models (e.g. [9]). Examples are unit-weight linear models (e.g. "tallying", Fig. 2a) or lexicographic models (e.g. simple decision trees, such "fast-and-frugal trees", Fig. 2b; for an overview, see [12].

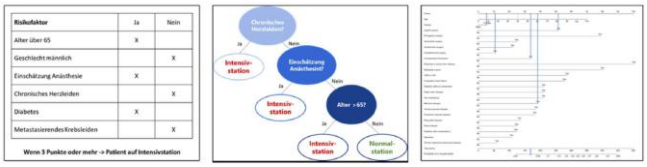

Fig. 2. Visual representation of simple cognitive models and statistical standard procedures. 2a: unit-weight linear model ("tally") as a checklist. 2b: Lexicographic model as decision tree ("Fast-and-frugal Tree"). 2c: Linear regression model as nomogram (adapted from [13]).

The advantages of this approach are manyfold: for one, in an iterative process of problem understanding and model adaptation, better understanding of the model and what it does can itself lead to better model performance. Secondly, integration with human decision making and expertise can further improve the output of the human-machine-system by reducing false positives and false negatives. The expert may be aware of boundary conditions that impair the output or certain features of a model in particular situations. Or they can anticipate technological, structural, and procedural changes and gauge their impact on model performance. This also means that an intrinsically transparent model developed at a different site may be more safely applied out-of-population. Relatedly, such simple tools are far less affected by infrastructural constraints, as they can often be implemented as laminated pocket cards or simply remembered after short use. Lastly, if a model is faulty or if it cannot be applied in a particular situation, this becomes obvious much more immediately than with a black box model or even an XAI solution. In particular, the debugging is much more straightforward for an intrinsically simple and intuitive model.

II. The Problem

In what follows, we used this approach to design a decision support for post-operative risk stratification in intensive care. A large, prospective investigation of post-operative mortality risk in 28 European countries found that 73% of patients that had died within a week of their operation had never been in contact with the intensive care tract [14]. Many of these patients would have benefitted from the higher level of surveillance and care the intensive care unit provides. Currently, post-operative risk stratification is done using risk scores such as the Charlson Comorbidity Score or Index (CCS; a logistic-regression-based risk scoring tool) or the American Society of Anaesthesiologists' Physical Status Score (ASA-PS; a six-point intuitive assessment of a patient's physical status by an expert anaesthesiologist). The goal of this analysis was to construct a simple, intrinsically transparent and interpretable model for the problem of post-operative risk stratification to the ICU or normal ward. Note that this initial analysis concerned itself only with the application of such simple, intuitive models to data. Advantages such as improved performance through more efficient iteration, reduction of false positives and negatives through better understanding of the model *in situ*, or better

care provision through adaptation to and incorporation of patients' wishes and preferences in a process of shared decision making, were not addressed and are not within the scope of this paper.

III. Methods and Data

A. Goal-Directed Task Analysis

A goal-directed task analysis [15] of the decision environment was conducted consisting of a document review, a three-week observational period of the allocation process, and interviews with subject matter experts (N=5) at a large European university hospital. Three decision points were identified at which risk stratification decisions are made: pre-operative establishment of the patient roadmap, updating due to intra-operative events, and a further updating-decision in the post-operative phase for patients administered to the post-anaesthesia care unit (PACU). Our analysis focused on pre-operative roadmap establishment. While both post-operative decision points contain important additional information for deciding on the appropriate level of care for a patient, pre-operative roadmap establishment is central for daily organizational planning and resource allocation. Every time a deviation from the pre-operative roadmap occurs, it requires additional corrective organizational actions to be taken by hospital staff at multiple levels. Achieving high accuracy during initial allocation is therefore critical to smooth organizational functioning.

B. Data Selection and Pre-Processing

Data from 182.886 patients prospectively recruited to the Ko-Moskau study [13] were used as the data pool for analysis. The study protocol was approved by the local ethics committee (EA1/007/13) and pre-registered prior to data collection (ClinicalTrials.gov identifier: NCT01810133). All patients with a digitalized electronic anaesthesia record between January 2006 and December 2011 who were not treated in the intensive care unit were eligible for inclusion. Emergency and ambulatory cases were excluded and only cases with complete data were analysed, leaving a final number of 130.238 case records. The dataset included the following information: ASA-PS, surgical discipline, priority of surgery (elective or urgent), localization of surgery, intraoperative transfusions, age, gender, and in-hospital death. If the patient underwent >1 surgical procedure during their hospital stay, only the data of the first surgical intervention was included in the analysis. Since intraoperative transfusions occur after the decision point that we studied, we ignored this information when building our prediction tools. Outcome criteria was in-hospital death for patients never admitted to the intensive care unit (0.1% mortality baseline). We used RStudio version 1.0.153 for data analysis. For some models (CART, random forest, penalized logistic regression, fast-and-frugal tree induction based on cross-entropy optimization) we dummy-coded all variables leaving us with an additional dummy data set. Pre-processing was done for both data sets. From both sets, we removed all variables that had near-zero or zero variance or were highly correlated with other variables (pair-wise absolute correlation of >.90) or were linearly dependent on other variables [16]. Lastly, as mortality in the dataset was only 0.1%, the data were upsampled to account for this low base rate and resulting class imbalance after the pre-processing.

C. Models

In our model comparisons, we tested classification and regression trees (CART; [17]) using the "one-standard-error" method, support vector machines, three random forests, and three regression models (unpenalized and penalized logistic regression, as well as the Surgical Mortality Score, a regression model developed specifically for post-surgical risk stratification by Kork et al. [13]), using the R-package *caret* [16]. As user-friendly models we tested different algorithms to induce fast-and-frugal trees (FFTs) using the R-package *FFTrees* [18] and not yet published FFT-algorithms. FFTs are binary classifiers that have at least one exit leaf for each node (two for the last). In our implementation, the FFTs were restricted to include maximally 10 nodes. FFTs are a family of very simple and easy-to-use yet powerful decision trees [19]. All of these models' performances were also compared to anaesthesiologists' assessment of the patients' health status (ASA-PS) and the Charlson Comorbidity Score (CCS). Note that both the ASA-PS and the CCS could potentially be included as variables in the models if they were not selected out during the feature selection process. We fitted the models on the years 2006-2010 and predicted mortality in the year 2011, optimizing the area under the receiver operating curve (AUC). We also used a 10-fold cross-validation in the years 2006-2010 (while applying the "one-standard-error" method to the simple tree model CART as suggested in the R-package *caret* and [16]). Subsequently, we used the fixed tuning parameters and fit the remaining parameters again on the same data. Finally, we validated the models using the year 2011 as "holdout" set. The median AUC of all models was used as a robust estimation of the predictability of each criterion. For the SMS model, logistic regression, penalized logistic regression, and support vector machines, the data was centred and scaled [15].

IV. Results

For each model class, we report only the variant that was most predictive in 2011 (CART, unregularized logistic regression, random forest with mtry = 2, and a recursive FFT algorithm using gini index for splits and ordering). The median AUC across all statistical and machine learning models in 2011 was 0.93 (range = 0.89 to 0.98, mean = 0.93). Logistic regression (AUC = 0.98) performed best. The tree resulting from the CART algorithm used only five predictors (CCS, age, whether a patient had cancer, whether the surgery was urgent, and whether the surgeon's field of specialty was general surgery) and had a simple structure with one exit for each of the first four predictors and two for the fifth predictor, i.e., classifying as a fast-and-frugal decision tree (FFT). The CART-FFT had an AUC of 0.89, while the "actual" FFT consisted of only three predictors (# of comorbidities, cardiac surgery, and the anaesthesiologist's assessment of how ill a patient is (ASA-score)) and had an AUC of 0.91.

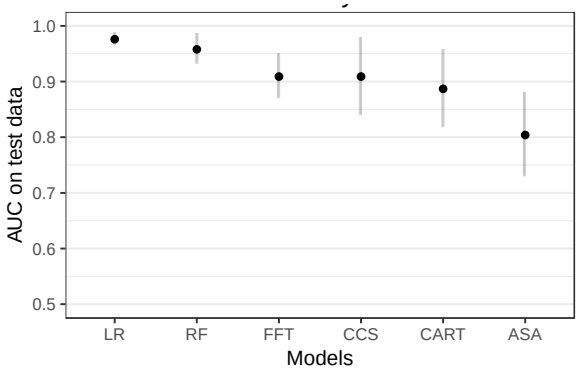

Fig. 3. Results of machine learning analysis for the best performing models in each model class. Error bars are bootstrapped 95% confidence intervals. AUC = Area Under the receiver operating Curve; LR = Logistic Regression; RF = Random Forest; FFT = Fast-and-Frugal Tree; CCS Charlson Comorbidity Score; CART = Classification and Regression Tree; ASA = Anesthesiologist's intuitive assessment.

V. Conclusion

We have used the latest methods in machine learning to assure a high-quality analysis, including an intrinsically interpretable, transparent class of algorithm, so called fast-and-frugal trees (FFT). The resulting FFT asks only three questions and can be easily visualized and understood. Decisions which align with or deviate from the algorithm can be transparently communicated, explained, or defended. The tool can be implemented in absence of any complex information technological infrastructure as a laminated pocket card (see Fig. 4) for physicians to refer to before being simply remembered. While achieving a predictive performance of 7 percentage points below the best-performing model, logistic regression, this only pertains to the pure statistical performance, not the performance in the field. By nature of its intrinsic transparency, which does not rely on post hoc analyses with complex XAI-methods, the physician's decision-making competence is not replaced, but augmented.

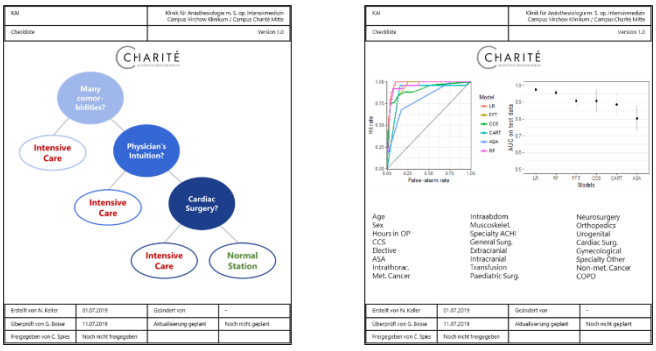

Fig. 4. Exemplary instantiation of the FFT as laminated pocket card. Front contains the decision algorithm, back contains additional information on comparative model performance and predictors included in the analysis to better assess the utility of the tool under certain boundary conditions.

References

[1] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Mach Intell. 2019;1(5):206-215.

[2] Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1-4.

[3] Zweig KA. Ein Algorithmus hat kein Taktgefühl. München: Heyne; 2019.

[4] Burton JW, Stein MK, Jensen TB. A systematic review of algorithm aversion in augmented decision making. J Behav Decis Making. 2019

[5] Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Confounding Variables Can Degrade Generalization Performance of Radiological Deep Learning Models arXiv preprint arXiv:1807.00431.

[6] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).

[7] Shapley, Lloyd S. "A value for n-person games." Contributions to the Theory of Games 2.28 (1953): 307-317.

[8] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017.

[9] Buckmann M, Şimşek, Ö. Decision heuristics for comparison: How good are they? Proceedings NIPS 2016 Workshop on Imperfect Decision Makers, PMLR. 2017;58:1-11

[10] Hand DJ. Classifier technology and the illusion of progress. Stat Sci. 2006;21(1):1-14.

[11] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Calster BV. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22.

[12] Hafenbrädl S, Waeger D, Marewski JN, Gigerenzer G. Applied decision making with fast-and- frugal heuristics. J Appl Res Mem Cogn. 2016;5(2):215-231.

[13] Kork F, Balzer F, Krannich A, Weiss B, Wernecke K-D, Spies C. Association of Comorbidities With Postoperative In-Hospital Mortality. Medicine. 2015;94(8):e576. doi:10.1097/MD.0000000000000576.

[14] Pearse RM, Moreno RP, Bauer P, Pelosi, P, Metnitz, P, Spies, C, et al. Mortality after surgery in Europe: a 7 day cohort study. Lancet. 2012;380(9847):1059-1065.

[15] Klein G, Armstrong AA. Goal-directed task analysis. In Handbook of human factors and ergonomics methods. CRC Press. 2004:373-382.

[16] Kuhn M, Johnson K. Applied Predictive Modeling. Springer; 2013. doi:10.1007/978-1-4614-6849-3.

[17] Breiman L, Friedman J, Stone C., Olshen R. Classification and Regression Trees. New York, NY: Chapman; Hall; 1984.

[18] Phillips ND, Neth H, Woike JK, et al. FFTrees: A toolbox to create, visualize, and evaluate fast-and-frugal decision trees. 2017;:1–25.

[19] Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). "Categorization with limited resources: A family of simple heuristics," Journal of Mathematical Psychology, 52: 352-361.