**LITERATURE SURVEY**

## 1) Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

**AUTHORS: R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad**

# In machine learning often a tradeoff must be made between accuracy and intelligibility. More accurate models such as boosted trees, random forests, and neural nets usually are not intelligible, but more intelligible models such as logistic regression, naive-Bayes, and single decision trees often have significantly worse accuracy. This tradeoff sometimes limits the accuracy of models that can be applied in mission-critical applications such as healthcare where being able to understand, validate, edit, and trust a learned model is important. We present two case studies where high-performance generalized additive models with pairwise interactions (GA2Ms) are applied to real healthcare problems yielding intelligible models with state-of-the-art accuracy. In the pneumonia risk prediction case study, the intelligible model uncovers surprising patterns in the data that previously had prevented complex learned models from being fielded in this domain, but because it is intelligible and modular allows these patterns to be recognized and removed. In the 30- day hospital readmission case study, we show that the same methods scale to large datasets containing hundreds of thousands of patients and thousands of attributes while remaining intelligible and providing accuracy comparable to the best (unintelligible) machine learning method.

# 2) Evaluating Reinforcement Learning Algorithms in Observational Health Settings

**AUTHORS:**  **O. Gottesman, F. Johansson, J. Meier, J Dent, D. Lee, S. Srinivasan, L. Zhang, Y. Ding, D. Wihl, X. Peng, J. Yao, I. Lage, C. Mosch, L. H. Lehman, M. Komorowski, A. Faisal, L. A. Celi, D. Sontag, and F. Doshi-Velez**

Much attention has been devoted recently to the development of machine learning algorithms with the goal of improving treatment policies in healthcare. Reinforcement learning (RL) is a sub-field within machine learning that is concerned with learning how to make sequences of decisions so as to optimize long-term effects. Already, RL algorithms have been proposed to identify decision-making strategies for mechanical ventilation, sepsis management and treatment of schizophrenia. However, before implementing treatment policies learned by black-box algorithms in high-stakes clinical decision problems, special care must be taken in the evaluation of these policies. In this document, our goal is to expose some of the subtleties associated with evaluating RL algorithms in healthcare. We aim to provide a conceptual starting point for clinical and computational researchers to ask the right questions when designing and evaluating algorithms for new ways of treating patients. In the following, we describe how choices about how to summarize a history, variance of statistical estimators, and confounders in more ad-hoc measures can result in unreliable, even misleading estimates of the quality of a treatment policy. We also provide suggestions for mitigating these effects---for while there is much promise for mining observational health data to uncover better treatment policies, evaluation must be performed thoughtfully.

# 3) Incidence, risk factors and severity of retinopathy of prematurity in Turkey (TR-ROP study): a prospective, multicentre study in 69 neonatal intensive care units

**AUTHORS** **: Ahmet Yagmur Bas, Nihal Demirel, Esin Koc, Dilek Ulubas Isik, İbrahim Murat Hirfanoglu, Turan Tunc, TR-ROP Study Group**

To evaluate the prevalence, risk factors and treatment of retinopathy of prematurity (ROP) in Turkey and to establish screening criteria for this condition. A prospective cohort study (TR-ROP) was performed between 1 April 2016 and 30 April 2017 in 69 neonatal intensive care units (NICUs). Infants with a birth weight (BW)≤1500 g or gestational age (GA)≤32 weeks and those with a BW>1500 g or GA>32 weeks with an unstable clinical course were included in the study. Predictors for the development of ROP were determined by logistic regression analyses. The TR-ROP study included 6115 infants: 4964 (81%) with a GA≤32 weeks and 1151 (19%) with a GA>32 weeks. Overall, 27% had any stage of ROP and 6.7% had severe ROP. A lower BW, smaller GA, total days on oxygen, late-onset sepsis, frequency of red blood cell transfusions and relative weight gain were identified as independent risk factors for severe ROP in infants with a BW≤1500 g. Of all infants, 414 needed treatment and 395 (95.4%) of the treated infants had a BW≤1500 g. Sixty-six (16%) of the treated infants did not fulfil the Early Treatment for Retinopathy of Prematurity requirements for treatment. Screening of infants with a GA≤34 weeks or a BW<1700 g appears to be appropriate in Turkey. Monitoring standards of neonatal care and conducting quality improvement projects across the country are recommended to improve neonatal outcomes in Turkish NICUs.

# 4) Accurate Intelligible Models with Pairwise Interactions

**AUTHORS** : **Y. Lou, R. Caruana, J. Gehrke, and G. Hooker**

Standard generalized additive models (GAMs) usually model the dependent variable as a sum of univariate models. Although previous studies have shown that standard GAMs can be interpreted by users, their accuracy is significantly less than more complex models that permit interactions. In this paper, we suggest adding selected terms of interacting pairs of features to standard GAMs. The resulting models, which we call GA2M-models, for Generalized Additive Models plus Interactions, consist of univariate terms and a small number of pairwise interaction terms. Since these models only include one- and two-dimensional components, the components of GA2M-models can be visualized and interpreted by users. To explore the huge (quadratic) number of pairs of features, we develop a novel, computationally efficient method called FAST for ranking all possible pairs of features as candidates for inclusion into the model. In a large-scale empirical study, we show the effectiveness of FAST in ranking candidate pairs of features. In addition, we show the surprising result that GA2M-models have almost the same performance as the best full-complexity models on a number of real datasets. Thus this paper postulates that for many problems, GA2M-models can yield models that are both intelligible and accurate.

**5)** **Interpretable Machine Learning in Healthcare**

**AUTHORS**: **M. Aurangzeb Ahmad, C. Eckert, A. Teredesai, and G. McKelveye**

This tutorial extensively covers the definitions, nuances, challenges, and requirements for the design of interpretable and explainable machine learning models and systems in healthcare. We discuss many uses in which interpretable machine learning models are needed in healthcare and how they should be deployed. Additionally, we explore the landscape of recent advances to address the challenges model interpretability in healthcare and also describe how one would go about choosing the right interpretable machine learnig algorithm for a given problem in healthcare.