ScienceDirect

Contents lists available at **sciencedirect.com**
Journal homepage: **www.elsevier.com/locate/jval**

Artificial Intelligence Method Studies

AI1
QUANTIFYING BIAS IN ML-EXTRACTED VARIABLES FOR INFERENCE IN CLINICAL ONCOLOGY

Lee J,[1] Estevez M,[2] Segal BD,[2] Sondhi A,[2] Cohen AB,[2] Cherng ST[2]
[1]*Flatiron Health, Baltimore, MD, USA,* [2]*Flatiron Health, New York, NY, USA*

Objectives: Machine learning (ML) approaches can extract clinically relevant information from electronic health records (EHRs) to be used for research purposes, such as comparative effectiveness analyses. This study assessed the effects of misclassification error in ML-extracted clinical variables when used in statistical analyses. ***Methods:*** We selected a cohort of 2,948 patients with advanced NSCLC treated with one of two common second line monotherapies from the nationwide Flatiron Health EHR-derived de-identified database. Focusing on smoking and PD-L1 status information extracted from free-text EHR notes, we analyzed the performance of an ML approach against manual abstraction (reference). We fit a Cox proportional hazards model to estimate overall survival (OS) hazard ratios (HRs) between treatments in cohorts reweighted by propensity scores based on a set of confounders (gender, histology, advanced diagnosis age, first-line treatment class, stage, smoking status, and PD-L1 status). We performed sensitivity analyses by corrupting abstracted labels at varying error rates. ***Results:*** Using manually abstracted PD-L1 and smoking status to estimate propensity scores, the HR (95% CI) of treatment A vs B was 0.797 (0.686, 0.911). Using ML-extracted PD-L1 and ML-extracted smoking status, the HR increased slightly, 0.839 (0.721, 0.968). Using ML-extracted PD-L1 and manually abstracted smoking status the HR was 0.848 (0.725, 0.971), and using ML-extracted smoking status and manually abstracted PD-L1 the HR was 0.790 (0.692, 0.896). In a sensitivity analysis, errors introduced into smoking status did not affect HR estimates, though errors in PD-L1 did. ***Conclusions:*** The impact of using ML-extracted instead of manually-abstracted variables is potentially greater for strong confounding variables (i.e., PD-L1 as opposed to smoking). This argues for using downstream analyses as a way to validate ML-extracted variables, as impact on analytical results cannot be inferred by standard ML performance metrics alone.

AI2
THE IMPACT OF INCLUDING RACE AND ETHNICITY IN RISK PREDICTION MODELS ON RACIAL BIAS

Khor S,[1] Hahn EE,[2] Haupt EC,[2] Shankaran V,[3] Clark S,[1] Rodriguez P,[1] Chen Y,[1] Bansal A[1]
[1]*University of Washington, Seattle, WA, USA,* [2]*Southern California Permanente Medical Group, Pasadena, CA, USA,* [3]*Fred Hutch, Seattle, WA, USA*

Objectives: Risk prediction algorithms can support clinical decision-making but there is a lack of consensus on when and how sociodemographic factors, especially the social construct of race/ethnicity, should be included in these algorithms. Our objective is to assess the impact of including race as a predictor in a risk prediction algorithm on racial biases in model performance. ***Methods:*** We used data from a large integrated health care system to develop a recurrence risk prediction model for adults with colorectal cancer who underwent resection. We fitted three Cox proportional hazard models using clinical and demographic variables: one excluded race/ethnicity as a predictor ("race-blind"), one included race/ethnicity ("race-sensitive"), and one with interactions between predictors and race/ethnicity. We compared racial biases in model performances between these models, measured by discrimination (area under the receiver operator curve (AUC)) and sensitivity at a fixed specificity (80%). ***Results:*** Among 4515 patients (mean age 65; 48% female), 53% were non-Hispanic White (NHW), 22% Hispanic, 13% Black/African American (AA), 12% Asian/Pacific Islander. 5-year cumulative incidence of recurrence for Stage 3 patients differed across racial groups: 25%, 33%, 35%, and 32% among NHW, Hispanic, AA, and Asian, respectively. In the "race-blind" model, AUCs varied across groups (0.74, 0.58, 0.66, and 0.61 among NHW, Hispanic, AA, and Asian, respectively). Adding race improved the AUCs only very slightly. Adding interaction terms resulted in AUCs of 0.75, 0.57, 0.63, and 0.64. Sensitivities also varied across groups (50%, 42%, 36%, and 33% among NHW, Hispanic, AA, Asian, in both the "race-blind" and "race-sensitive" models). Including interaction terms increased sensitivities only for some groups. ***Conclusions:*** Risk prediction models had worse performances in minority racial subgroups compared to NHW, even with the explicit inclusion of race as a predictor or interaction terms. Risk model developers and users need to identify algorithmic disparities and understand their potential implications.

AI3
DISEASE-FREE SURVIVAL (DFS) AS A SURROGATE ENDPOINT FOR OVERALL SURVIVAL (OS) IN ADULTS WITH RESECTABLE ESOPHAGEAL OR GASTROESOPHAGEAL JUNCTION CANCER: A CORRELATION META-ANALYSIS

Leung L,[1] Kurt M,[2] Singh P,[2] Kim I,[2] Donnellan G,[1] Kanters S[1]
[1]*Evidinno Outcomes Research Inc., Vancouver, BC, Canada,* [2]*Bristol Myers Squibb, Lawrenceville, NJ, USA*

Objectives: To evaluate the appropriateness of using DFS as a surrogate for OS in trials investigating adults with resectable esophageal or gastroesophageal junction cancer (EC/GEJ) receiving therapies in the (neo)adjuvant and perioperative settings. ***Methods:*** A systematic literature review was conducted to identify EC/GEJ trials that reported OS and DFS, or progression-free survival (PFS) compatible with the definition of DFS. The primary analysis was restricted to studies meeting the proportional hazards (PH) assumption which was assessed statistically after digitization of the reported KM curves. Secondary analysis was restricted to the adjuvant setting only. Sensitivity analyses consisted of removing studies with outlier characteristics and other potential biases, and estimating differences between restricted mean survival times to assess the impact of including studies with non-proportional hazards. The surrogacy relationship was assessed using bivariate random-effects meta-analysis and weighted linear regression. Surrogate threshold effect (STE), the minimum DFS/PFS benefit that would translate into OS benefit, was also estimated. The robustness of the models was tested via leave-one-out cross-validation (LOOCV). ***Results:*** The primary analysis included 26 trials after removing 7 trials not meeting the PH assumption. The estimated correlation coefficient was 0.83 (95% CI: 0.70-0.90). The resulting surrogacy equation was $\log(HR_{OS}) = 0 + 0.80 \times \log(HR_{DFS/PFS})$ with a corresponding STE of 0.82. LOOCV verified the stability of the results with 95% alignment between observed and estimated HR_{OS}. Correlation estimates across secondary analyses and sensitivity scenarios were similar to those from the primary analysis. ***Conclusions:*** The results and their validation point out a correlation between DFS and OS in EC/GEJ. The estimated surrogacy equation and the corresponding STE can enable $HR_{DFS/PFS}$ as a surrogate predictor of HR_{OS} in EC/GEJ trials. These findings can be useful when evaluating the long-term efficacy of treatments where OS may not be immediately available.

AI4
COMPARING MACHINE-LEARNING METHODS FOR THE PREDICTION OF MAJOR ADVERSE LIMB EVENTS AND MORTALITY AFTER A PERCUTANEOUS INTERVENTION

Gressler L,[1] Marinac-Dabic D,[2] Dosreis S,[3] Goodney P,[4] Mullins CD,[3] Shaya FT[5]
[1]*University of Maryland, Baltimore, MD, USA,* [2]*U.S. Food and Drug Administration, Silver Spring, MD, USA,* [3]*University of Maryland School of Pharmacy, Baltimore, MD, USA,* [4]*Dartmouth-Hitchcock Medical Center, Lebanon, NH, USA,* [5]*University of Maryland Baltimore, Baltimore, MD, USA*

Objectives: The objective was to formulate, test, and compare the performance of regression-based and machine learning models in the prediction major adverse limb events (MALE) and mortality among patients receiving treatment for lower extremity peripheral artery disease (PAD). ***Methods:*** Patients undergoing atherectomy, stent, and combination stent atherectomy for lower extremity PAD were identified in the Vascular Quality Initiative registry. Thirty-nine variables summarizing demographic, medical history, pre-operative, indication-specific, and procedure-specific characteristics were utilized to predict MALE and mortality events. For both events, we compared the performance of four different prediction models: a generalized linear model (GLM), a Least Absolute Shrinkage and Selection Operator (LASSO) regularized GLM, a gradient boosted decision tree, and random forest model. The area under the curve (AUC) evaluated the effectiveness of each prediction model. For validation purposes, 5-fold cross-validation was repeated three times. Pairwise comparisons of