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

the receiver operating characteristic curves (ROC), sensitivity, and specificity measures with Bonferroni adjustment for multiple testing applied were performed to compare the models' performance. ***Results:*** Among 15964 identified patients, a MALE occurred in 26.02% of patients, and death occurred in 18.82% of patients. The most effective predictive model for MALE, as determined by the AUC, was the gradient boosted decision tree (AUC= 0.7539) followed by the LASSO regulated GLM (AUC= 0.749). The most effective predictive model for mortality was the LASSO regularized GLM (AUC=0.7930) followed by the GLM model (AUC=0.7922). The GLM, LASSO regularized GLM model, and gradient boosted decision tree produced similar ROC. ***Conclusions:*** All models showed acceptable discrimination, with an AUC greater than 0.7, when predicting MALE and mortality among patients receiving treatment for lower extremity peripheral artery disease. The machine learning techniques outperformed traditional regression-based techniques and can be leveraged to generate robust predictive models within the clinical space of lower extremity PAD.

Comparative Effectiveness Studies

CE1
A MODELED HEALTH OUTCOMES EVALUATION OF DAROLUTAMIDE PLUS ANDROGEN DEPRIVATION THERAPY FOR HIGH-RISK NON-METASTATIC CASTRATION-RESISTANT PROSTATE CANCER IN CHINA

Ming J,[1] Liu Y,[1] Lu W,[1] Wu Y,[1] Li W,[1] Han R,[2] **Hu S**[3]
[1]*IQVIA, Shanghai, China,* [2]*Bayer Healthcare Company Ltd., Beijing, China,* [3]*School of Public Health, Fudan University, Shanghai, China*

Objectives: Novel antiandrogens have demonstrated clinical benefit for patients with non-metastatic castration resistant prostate cancer (nmCRPC). We modeled the incremental life-years and QALYs of darolutamide+ADT compared to apalutamide+ADT and enzalutamide+ADT, for high-risk nmCRPC, in a Chinese setting. ***Methods:*** A partitioned survival model with three health states (nmCRPC, metastatic CRPC and death) was devised. Data from the ARAMIS, PROSPER and SPARTAN trials were used, in the absence of head-to-head studies. Hazard ratios (HRs) characterizing overall survival (OS) were derived from formal indirect treatment comparisons of the OS HRs between the three trials; the point estimates of the OS HRs characterizing the indirect comparison of darolutamide+ADT vs apalutamide+ADT and enzalutamide+ADT were 0.88 and 0.95, respectively. Parametric survival functions were selected for extrapolation based on the goodness of fit. The baseline characteristics, treatment patterns and utilities for this modeled Chinese cohort were validated with 12 urologists from 11 hospitals. A discount rate of 5% was applied. Univariate and probabilistic sensitivity analyses were performed. ***Results:*** With a 20-year modeling horizon, darolutamide+ADT yielded more life years (5.98LYs) and QALYs (4.61QALYs) than both apalutamide+ADT (5.64LYs, 4.43QALYs) and enzalutamide+ADT (5.82LYs, 4.55QALYs). The incremental QALY gains with darolutamide+ADT vs. apalutamide+ADT (+0.18QALYs) and vs. enzalutamide+ADT (+0.06QALYs) were mainly driven by the improved life years during post-progression survival (+0.92LYs vs. apalutamide+ADT and +0.65LYs vs. enzalutamide+ADT). Results were sensitive to the HRs and utility inputs; however, the probabilities of darolutamide+ADT having incremental QALYs reached 70.4% and 56.8% when compared to apalutamide+ADT and enzalutamide+ADT, respectively. ***Conclusions:*** This modeled evaluation indicates that darolutamide+ADT is likely to be associated with incremental health outcomes over apalutamide+ADT and enzalutamide+ADT, for a high-risk nmCRPC Chinese cohort, over a 20-year time-frame.

CE2
EFFECT OF DIFFERENT VARIANCE ESTIMATION METHODS WITH INVERSE PROBABILITY TREATMENT WEIGHTS (IPTW) ON COMPARATIVE EFFECTIVENESS MEASURE IN MULTIPLE SCLEROSIS

Earla J,[1] Hutton GJ,[2] Thornton JD,[3] Chen H,[4] Johnson ML,[4] Aparasu RR[4]
[1]*Incyte Corporation, Chadds Ford, PA, USA,* [2]*Baylor College of Medicine Medical Center, McNair Campus, Houston, TX, USA,* [3]*University of Houston, The Prescription Drug Misuse Education and Research (PREMIER) Center, Houston, TX, USA,* [4]*University of Houston, Houston, TX, USA*

Objectives: The Inverse Probability Treatment Weighting (IPTW) method provides marginal treatment effects that are more generalizable than other propensity score (PS) methods. The objective of this study was to compare the treatment effects from three different variance estimation methods with stabilized IPTWs on comparative effectiveness between oral fingolimod and injectable Disease Modifying Agents (DMA) users in Multiple Sclerosis (MS). ***Methods:*** This longitudinal retrospective study used adults(≥18 years) with MS diagnosis (ICD-9-CM:340) and a DMA prescription from the IBM MarketScan Commercial Claims and Encounters Database from 2010–2012. Patients were classified into fingolimod or injectable users based on their initial DMA prescription. The composite endpoint (time-to-relapse/DMA switch) was assessed during the one-year follow-up period after DMA initiation. The stabilized IPTW-Cox Proportional Hazards regression model was used to evaluate the composite endpoint with three different variance estimators – (i)Naïve, (ii)Robust sandwich-type, and (iii)Bootstrapping(200 replications). Patients who died/were lost from follow-up due to the lack of insurance coverage were censored. ***Results:*** The new DMA user study cohort consisted of 1,700 MS patients who were initiated with oral(15.82%) or injectable(84.18%) DMAs during 2010-2011. The proportion of patients who had a composite endpoint in fingolimod and injectable DMA users was 16.72% and 27.16%, respectively. The stabilized IPTW-Cox model with naïve and bootstrapping variance estimators revealed that oral fingolimod users were superior to injectable DMAs in reducing the risk of composite endpoint (Naïve estimator: Adjusted Hazards Ratio [aHR]-0.67, 95%CI:0.51-0.87; Bootstrapped estimator: aHR-0.68, 95%CI:0.39-0.97). However, the findings were not significant in the IPTW-Cox model with robust sandwich estimator(aHR-0.67, 95%CI:0.43-1.03). ***Conclusions:*** The analyses revealed that the significance of treatment effect estimates could vary depending on the choice of variance estimation method. Hence, researchers should pay attention to the selection of variance estimation method with small samples in addition to handling of extreme weights while using IPTWs for time-to-event analyses.

CE3
SYSTEMATIC REVIEW AND INDIRECT COMPARISON OF PD-(L)1 INHIBITORS IN COMBINATION WITH PLATINUM-BASED DOUBLET CHEMOTHERAPY (PT-DC) FOR THE FIRST-LINE TREATMENT OF NON-SQUAMOUS, NON-SMALL-CELL LUNG CANCER (NSQNSCLC)

Zhang L,[1] Qian Y,[2] Li J,[2] Cui C,[2] Chen L,[2] Qu S,[3] **Lu S**[4]
[1]*Eli Lilly and Company, Shanghai, 31, China,* [2]*Eli Lilly and Company, Shanghai, China,* [3]*Real World Solutions, IQVIA, Shanghai, China,* [4]*Shanghai Chest Hospital, Shanghai, China*

Objectives: To evaluate relative efficacy and safety of sintilimab compared with other PD-(L)1 inhibitors in combination with platinum-based doublet chemotherapy (PT-DC) for the first-line treatment of nsqNSCLC. ***Methods:*** A systematic literature search was conducted based on published studies in electronic databases up to December 2020. Eligible randomized controlled trials (RCT) that investigated untreated locally advanced or metastatic nsqNSCLC without activating EGFR mutations or ALK translocations patients were analyzed. Outcomes evaluated by adjusted indirect treatment comparison (ITC) using Bucher method mainly included progression-free survival (PFS), objective response rate (ORR), time to response (TTR) and safety profile. Taking the heterogeneity into account, random-effect model will be applied if the p-value of Chi-square test greater than 0.05. ***Results:*** Seven RCTs involving 3,559 patients were included. ITC results suggested that combined PT-DC with sintilimab had a comparable PFS compared with that of combined PT-DC with pembrolizumab (HR=1.00, 95%CI: 0.71 to 1.41), atezolizumab (HR=0.81, 95%CI: 0.59 to 1.10), tislelizumab (HR= 0.75, 95%CI: 0.48 to 1.16), camrelizumab (HR=0.80, 95%CI:0.54 to 1.20) and nivolumab (HR=0.72, 95%CI : 0.51 to 1.02). Little differences were found in ORR between combined PT-DC with sintilimab and combined PT-DC with pembrolizumab (OR=0.66, 95%CI: 0.37 to 1.20), atezolizumab (OR=1.23, 95%CI: 0.70 to 2.14), tislelizumab (OR=1.11, 95%CI: 0.58 to 2.11), camrelizumab (OR=1.05, 95%CI: 0.58 to 1.90) and nivolumab (OR=1.14, 95%CI: 0.64 to 2.00). Incidence of all grades or grades ≥3 adverse events (AE) were comparable between combined PT-DC with sintilimab and other PD-(L)1s. For AE leading to discontinuation, the incidence of sintilimab in combination with PT-DC is significantly lower than that of pembrolizumab in combination with PT-DC (OR=0.27, 95% CI: 0.11 to 0.67). ***Conclusions:*** Combined PT-DC with sintilimab and other PD-(L)1 inhibitors had comparable efficacy and safety profile for the first-line treatment of locally advanced or metastatic nsqNSCLC patients.

CE4
EXPANDING EVIDENCE BASE VS INTRODUCING HETEROGENEITY IN NETWORKS FOR NETWORK META-ANALYSES: A SIMULATION STUDY

Luttenauer H,[1] Le Nouveau P,[2] Gauthier A[3]
[1]*Amaris Consulting, London, UK,* [2]*Amaris Consulting, Levallois-Perret, ON, France,* [3]*Amaris Consulting, Barcelona, Spain*

Objectives: Network meta-analyses (NMAs) are widely used to estimate the relative treatment effect between treatments that have not directly been compared in clinical trials. Challenges arise in precision medicines, where only small networks of evidence are available. This study aimed to explore the trade-off between increasing the evidence base by including additional studies and increasing the level of heterogeneity in the network considered. ***Methods:*** Data were simulated to reflect a small network including four treatments. Four scenarios were simulated with increasing number of studies per direct comparison. Each study was assumed to add an additional level of heterogeneity in the network, that was partly explained by an observable covariate. Scenario 1 represented a small network with a low level of heterogeneity while scenario 4 was a bigger network with a higher heterogeneity level. Standard NMAs were conducted and the mean relative difference (MRD) between the NMA estimates