Assessing Cholangiocarcinoma Risk Using TOP5 Features in Real-World Data

Ryan Goodson[1], Inuk Zandvakili[2], Feliz van Haag[3], Carolin Victoria Schneider[3]

[1]University of Cincinnati College of Medicine, Cincinnati, Ohio; [2] Division of Gastroenterology and Hepatology, Department of Internal Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio; [3]Department of Internal Medicine, Gastroenterology, Metabolic Diseases and Intensive Care, University Hospital RWTH Aachen, Aachen, Germany

Introduction: Cholangiocarcinoma, a rare but deadly cancer of the bile ducts, has a low incidence in the US, 1.26 cases per 100,000 annually. Often diagnosed at an advanced stage, the median survival of 6 to 7 months may improve with early detection by screening. Our collaborators, van Haag, *et al*. applied machine learning to UK Biobank data to design a risk assessment score. This "TOP5" model utilizes the 5 most associated factors of cholangiocarcinoma (age, elevated GGT, complications of cirrhosis, cholangitis, and bile duct obstruction) to identify at-risk individuals and improve prognosis. We aimed to utilize TriNetX, an ultra-large scale clinical database, to externally verify this screening tool and evaluate its predictive value.

Methods: TriNetX contains electronic health records from over 155 million de-identified patients. Six cohorts were defined using GGT values and ICD codes for TOP5 criteria (age incorporated into later analysis): (1) healthy individuals without risk factors, (2) "at risk" with ≥1 general risk factor excluding TOP5, (3) elevated GGT only, (4) any TOP5 feature excluding GGT, (5) elevated GGT plus ≥1 additional TOP5 feature, named "GGT+1" and (6) all TOP5 features. Analyses were stratified by sex. Patients with pancreatic cancer were excluded because of diagnostic overlap with cholangiocarcinoma. Propensity score matching was utilized to balance demographics, BMI, and diabetes. Cholangiocarcinoma (ICD: C22.1, C24) must have occurred ≥1 year after index to exclude non-detected cancer at enrollment.

Results: Among males, cholangiocarcinoma rates were 0.006% in healthy patients, 0.09% in "at risk" patients, 0.17% in elevated GGT only patients, 0.22% in one TOP5 excluding GGT, 0.39% in "GGT+1", and 1.32% in all TOP5. Compared with the healthy cohort, hazard ratios were 32.15 for GGT only (95% CI: 20.82-49.63), 41.70 for one TOP5 excluding GGT (31.00-56.09), and 78.67 for the "GGT+1" cohort (32.52-190.30). Among females, cholangiocarcinoma rates were 0.004% in healthy patients, 0.05% in "at risk" patients, 0.11% in elevated GGT only patients, 0.14% in one TOP5 excluding GGT, 0.27% in "GGT+1", and 1.05% in all TOP5. Compared with the healthy cohort, hazard ratios were 34.37 for GGT only (95% CI: 20.21-58.47), 39.86 for one TOP5 excluding GGT (27.00-58.84), and 82.54 for the "GGT+1" cohort (26.41-257.96).

Conclusion: In the application of the TOP5 model derivative, cholangiocarcinoma rates increased with the accumulation of risk features. Analyses revealed strong associations, evidenced by high hazard ratios compared to the healthy cohort, suggesting the predictive utility of the TOP5 model as a screening tool. This study, limited by reliance on ICD-based data that may be unreliable because of incorrect charting, also highlights that the absolute risk of cholangiocarcinoma remains low even in high-risk groups, limiting immediate clinical utility. Future work will include incorporating age into local data analyses and applying the model to prospective cohorts to further validate predictive performance.

Acknowledgements: This study was supported in part by NIH grant T35 DK060444.