**Estimation of Opioid Use Disorder Prevalence Under Unique Data Scenarios: A Simulation Study**

**Authors:**

O’Dea R1, Linas B1, 2, White L2, Barocas JA3, Wang J4,5

1 Boston Medical Center, 2 Boston University School of Public Health, 3University of Colorado School of Medicine-Divisions of General Internal Medicine and Infectious Diseases, 4Massachusetts General Hospital, 5Harvard Medical School

**Background:**

To effectively tackle the overdose crisis, a nuanced understanding of Opioid Use Disorder (OUD) prevalence is crucial, both broadly and within targeted cohorts. Healthcare interactions provide estimates but may overlook those outside the healthcare system, leading to underestimation. Capture-recapture (CRC) analysis is valuable in estimating prevalence by addressing underreporting in surveillance. Conventionally, a stepwise model selection process (MSP) is employed to identify the model that best fits the data. However, the MSP in estimating group-stratified prevalence is less explored, especially with sparse data. This study uses simulations to investigate different MSPs for selecting conventional log-linear CRC models, with a focus on their ability to precisely estimate strata prevalence.

**Methods:**

Using data from the Massachusetts Public Health Data Warehouse, we generated synthetic populations stratified by multiple data sources and demographic groups: age, sex, and race. We simulated observation patterns across multiple data sources, accounting for scenarios with small observed counts. Data were tabulated to reflect varying source and strata combinations. We examined Poisson and Negative Binomial distributions within log-linear models, incorporating interaction terms to adjust for source dependency. Model selection efficacy was benchmarked by comparing stepwise, global AIC-based, and threshold-specific interaction methods against the simulated truth.

**Results:**

**As more demographic information was added, Poisson log-linear models selected through forward-stepwise processes generally outperformed their Negative Binomial counterparts, estimating prevalence within <0.001% discrepancy from true prevalence. In scenarios with data suppression, Poisson models selected through a forward-stepwise approach proved most accurate, within 0.78% discrepancy.**

**Conclusion:**

CRC is a valuable method for estimating the hidden prevalence of OUD, but its effectiveness depends on selecting appropriate models based on available data. Through contrasting different approaches, we highlight the estimation process for strata-specific prevalence and interpret strengths and limitations of common model selection strategies, enhancing the precision of OUD prevalence assessments for uniquely stratified data.