Statistica Sinica **36** (2026), 1-30
doi:https://doi.org/10.5705/ss.202024.0100

THE POPULATION AND PERSONALIZED AREAS UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE

Haben Michael* and Lu Tian

University of Massachusetts and Stanford University

Abstract: We consider two generalizations of the area under the receiver operating characteristic curve ("AUC"), a popular measure of discrimination, to accommodate clustered data. We describe situations in which the two cluster AUCs diverge and other situations in which they coincide. Differences are described using concrete models and visualizations, while quantitative results are used to relate the two generalizations. Procedures for joint estimation and inference are also presented, along with a simulation study. We apply the results to data collected on urban policing behavior.

Key words and phrases: AUC, clustered data, confounding, Simpson's paradox.

1. Introduction

The AUC is a widely used measure of how well a scalar predictor discriminates between two outcomes. As a population parameter, the AUC is the probability that the value of a randomly sampled predictor from one of the outcome classes is less than an independently sampled predictor from the other outcome class. There are several ways to generalize the AUC to accommodate clustered data. What we refer to as the "population AUC" appears to be the most commonly studied. The population AUC evaluates the predictor's typical effect on an entire population, as further discussed below.

While the population AUC is an important part of understanding the usefulness of a predictor, the medical field has lately focused on personalizing treatment. For example, in 2018 the National Academy of Medicine concluded: "The individuality of the patient should be at the core of every treatment decision. One-size-fits-all approaches to treating medical conditions are inadequate; instead, treatments should be tailored to individuals based on heterogeneity of clinical characteristics and their personal preferences."

We examine a "personalized AUC" in conjunction with the population AUC. These two evaluations may give different accounts of the usefulness of a marker. In the extreme case, the phenomenon known as Simpson's paradox may occur: The personalized AUC may be nearly uninformative while the population

*Corresponding author. E-mail: hmichael@math.umass.edu