Machine learning is a hammer, but are case-control studies nails?

Travis A. Gerke*

Case-control study designs are commonly used to estimate the effect of exposures on health outcomes, and may be viewed as more efficient variations of cohort studies, particularly for rare diseases. In practice, case-control designs most often proceed by selecting all patients who experienced the health event of interest (cases) and a corresponding subsample of population members who did not experience the event (controls). Dramatic cost savings are possible with this design, since exposures which are expensive to measure (e.g. genomics) need only to be ascertained in a subset of the population at risk.

In the case-control setting, the field of epidemiology has long recognized the odds ratio as a useful estimator of exposure-disease relationships. In addition to approximating the risk ratio for rare diseases, the odds ratio provides a valid estimate of the hazard ratio for many control-selection strategies. As such, logistic regression — which estimates the log-odds ratio to summarize exposure-outcome associations — has remained a ubiquitous tool in the analysis of case-control data.

Increasingly, case-control studies measure a high dimensional set of exposures; for example, data with >1M features are common in -omics research. Machine learning techniques present seemingly ideal tools to build classification or prediction models that distinguish cases from controls, often alongside feature selection. However, few machine learning techniques use the odds ratio as an objective function. This talk will demonstrate both theoretical and simulated explanations for case-control settings in which machine learning tools are appropriate, and settings in which they may mislead.

*Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute