Integration of Individual Participant and Aggregate Data Under Dataset Shift

Ming-Yueh Huang,
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan

Jing Qin,
Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, U.S.A.

and

Chiung-Yu Huang
Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94158, U.S.A.

Abstract

Integrated IPD–AD analysis, which combines individual participant data with aggregate data, is widely used for synthesizing evidence across heterogeneous studies. This talk examines how the form of aggregate data affects integration efficiency, an issue that has received less attention than algorithmic development. Using a constrained maximum likelihood framework, I show that subgroup-specific aggregate summaries substantially improve estimation efficiency, with outcome-stratified summaries consistently outperforming covariate-stratified ones, especially for continuous outcomes. Although outcome-stratified summaries are standard for discrete outcomes, they are rarely reported for continuous endpoints; our results suggest that routinely providing such summaries could meaningfully enhance evidence synthesis. I further extend the framework to accommodate dataset shift and propose a fast, non-iterative estimation procedure, illustrated with applications to income data under covariate shift and housing data under prior probability shift.

Keywords: Constrained maximum likelihood estimation; dataset shift; non-iterative algorithm.