Stanford MEDICINE | Department of Biomedical Data Science

Data Gluttony: Large Scale Inferential Risk Management for Commonly Used Datasets

Patrick L. Purdon, Ph.D
Postdoctoral scholar in the Department of Cardiothoracic Surgery at Stanford

February 19th, 2026
1:30PM-2:50PM
****R358****

Abstract:

As empirical researchers, we slipped into a new data use framework without even noticing it. Classical statistics gave rise to existing error protection procedures (e.g., p-values, multiple hypothesis testing corrections). But in the traditional setting, novel data sets were generated for each study (e.g., a randomized controlled trial, a survey using a sampling procedure) and then used to evaluate a specified number of hypotheses, and subsequent studies would be conducted leveraging data collected de novo. In the contemporary setting, a new paradigm of data use framework has emerged is better understood as a data reuse framework. Large-scale registries provide vast amounts of data to enable observational studies. In this work, we show that the federated and sequential reuse of data reuse, across many investigators, introduces positive dependence among inferential tasks, leading to cascading inferential errors. Common practice, which we call data gluttony, involves using all available data that meets study criteria, maximizing power at a cost: it shapes the distribution of inferential errors by introducing long tails of error. To address this, we investigate two paths to mitigate this risk.

Reading list:

- Dale, R., Rodu, J., Currie, M. E., & Baiocchi, M. (2025, August 22). *Data gluttony: Epistemic risks, dependent testing and data reuse in large datasets*. arXiv.org. https://arxiv.org/abs/2508.16552