STEVENS
INSTITUTE OF TECHNOLOGY
1870

Ph.D. Dissertation Defense

Candidate:	James Pleuss
Degree:	Doctor of Philosophy
School/Department.:	Interdisciplinary / Data Science
Date:	Tuesday, April 15, 2025
Time:	3:00 – 4:00 pm
Location:	Gateway North 303
Title:	Data-Driven Approaches to Nutritional Epidemiology and Dietary Assessment
Chairperson:	Dr. Samantha Kleinberg, Computer Science, School of Engineering and Science
Committee Members:	Dr. Ping Wang, Computer Science, School of Engineering and Science
	Dr. Hyewon Oh, Marketing, School of Business
	Dr. Onur Asan, Human Systems Interaction, School of Systems and Enterprises
	Dr. Andrea Deierlein, School of Global Public Health, New York University

Abstract

The links between nutrition and health are well documented, connecting foods and nutrients to long and short-term health outcomes. Lacking, however, are data-driven approaches that take advantage of advancements in machine learning, a particularly useful endeavor given the complex and multi-faceted nature of nutrition. In this thesis we look at how data-driven approaches to nutritional epidemiology and dietary assessment can offer a new perspective from which to view our diets, particularly during pregnancy, an under-researched stage of life.

First, we focus on ultra-processed foods (UPFs), which are ubiquitous and growing in the global diet, and associated with many negative health factors. We leverage the natural hierarchy of foods and a hierarchical ranking methodology to test whether specific UPFs or categories of UPFs might improve predictions of cardiometabolic health risks. In doing so, we learn which UPFs are most important for predicting cardiometabolic risk and advocate for further research of these UPFs which might enable causal discoveries.

Next, we introduce a new dataset from the Temporal Research in Eating, Nutrition, and Diet during Pregnancy (TREND-P) study which collected up to two rounds of 14-day dietary records with timestamped eating occasions from 150 pregnant individuals during their second trimester with associated biometric and health data. The study reveals significant within- and between-person variation across the micro- and macro-nutrients most critical during pregnancy. Our next paper employed the detailed timing of this dataset to develop a new chrononutrition scoring system, pairing morning and night intake scores with machine learning methods to identify associations between the time at which foods are consumed and maternal impaired glucose tolerance.

We then step back to consider the validity of current methodological approaches to dietary assessment. Existing calculations to determine the required number of days of dietary assessment for nutrition studies focus on within-person variation across cohorts but fail to consider how frequently individuals meet the requirements. We capitalize on the length of the TREND-P study and the nonparametric nature of bootstrapping to provide researchers with an understanding of how their research design choices impact the accuracy of recorded intakes.

Finally, nutrition research strives to establish causal relationships associated with dietary intakes, but these relationships are often limited in their generalizability to other populations. Here we introduce a new algorithm for causal discovery that goes beyond causal relationship identification. It incorporates static characteristics (age, education, comorbidities, etc.) to determine whether they moderate the strength of the causal relationship. We validate our algorithm against benchmarks and apply it to physical activity and nutrition data.

This thesis covers a wide range of data science initiatives (feature selection, bootstrapping, causal inference), applying them to advancing areas of nutrition research (UPFs, chrononutrition, diet/physical activity relationships), while providing a new dataset (TREND-P) that is suited for complex machine learning tasks. Our hopes is that this research will spark further data science advances in nutritional epidemiology.