Causal Inference for Personal Health Data:
A Case Study at Wellesley College

Jyontika Kapoor '24
Data Science Honors Thesis under the supervision of Professor Cassandra Pattanayak

Motivation and Research Questions

With the rise of personal health devices like smartwatches, understanding how to infer causality within the same person is important. This poster explores how to extend existing causal inference methods to scenarios where people are monitored over time.

Research Questions:

1) Can the Rubin Causal Model be extended to handle **within-person interference** in longitudinal studies involving multiple people?
2) Can generalized propensity scores for a continuous treatment variable be effectively used in this new methodological framework?
3) Does a causal relationship exist between step count and stress levels among Wellesley College students in the Fall 2022 microbiome study?

Background and Relevant Literature

Rubin Causal Model is a framework to infer whether X **causes** Y. This method is based upon the **no-interference** assumption, otherwise known as the Stable Unit Treatment Value Assumption (SUTVA), which says that the treatment of one unit is independent from the outcome of another unit [3]. However, this assumption is often **violated** in observational studies [4].

The Stable Unit Treatment on Neighborhood Value Assumption (**SUTNVA**) extends SUTVA to account for interference [1]. SUTNVA considers the influence of both a unit's treatment and the treatments of their neighbors within a defined neighborhood. The methodology and example in this thesis represent a specific instance of SUTNVA, addressing within-person interference by treating one neighborhood as one person.

The Stable Unit Treatment on Self Value Assumption (SUTSVA)

SUTSVA holds when:

- The Forward-in-Time assumption holds (see Figure 1).
- For each unit within the network, and under all possible treatment allocations, a change in treatments for units outside a given unit's neighborhood does not impact the outcome of that unit, provided the treatments of the unit and its neighbors remain unchanged.

For instance, a treatment change in person *i* at any time point would not affect the potential outcomes of person *j* at any time point, as long as $i \neq j$.

Person	Week	W	W′	$Y(\mathbf{W})$	$Y(\mathbf{W}')$
Person 1	Week 1	1	1	2	2
Person 1	Week 2	1	1	1	1
Person 1	Week 3	0	0	3	3
Person 2	Week 1	0	0	3	3
Person 2	Week 2	1	0	2	3
Person 2	Week 3	1	1	2	5

Table 1. The Stable Unit Treatment on Self Value Assumption. People are independent and spillover effects can occur between time points as time moves forward.

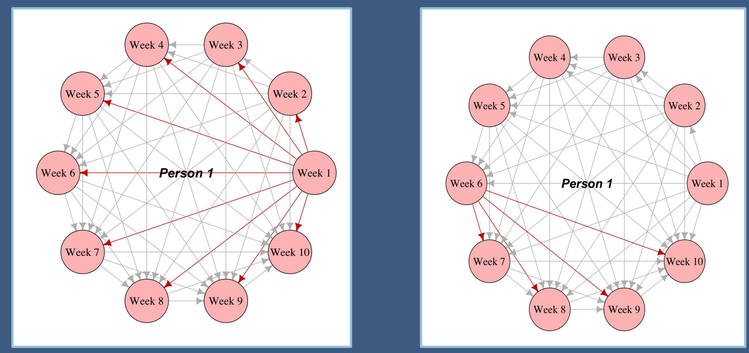

Figure 1. Forward in Time assumption says that when examining a person at week *t*, the treatments assigned in subsequent weeks (e.g., week *t+1* or after) do not have an influence on week *t* itself.

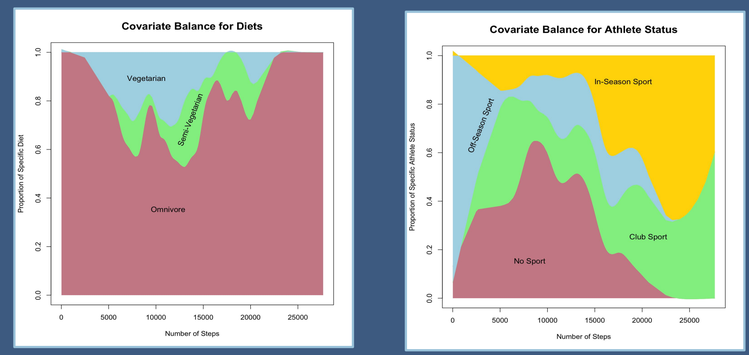

Figure 2. Stacked splines in the left graphic depict initial covariate balance for different dietary habits. Stacked splines on the right graphic show initial covariate balance for athlete status. A balanced distribution would show horizontal splines with equal proportion of each category at each level of treatment.

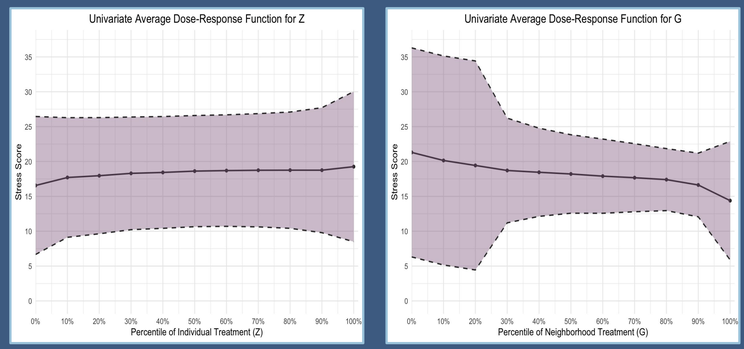

Figure 3. As the percentile of *Z* increases, there is a slight increase in the stress score. As the percentile of *G* increases, there is a slight decrease in the stress score. This suggests that higher step counts in the long-term cause slightly lower stress scores. Conversely, higher step counts in the short term cause slightly higher stress scores.

Estimation Procedure

Note that the below estimation procedure is adapted from Forastiere [2].

I. Estimate the Individual Treatment Model (*Z*).
Use a mixed effects linear regression to model the individual treatment (*Z*) as a function of **X**, including random effects for people.

II. Estimate the Neighborhood Treatment Model (*G*).
Use a mixed effects linear regression to model the weighted average of past treatment (*G*) as a function of **X** and *Z*, including random effects for people.

III. Estimate Propensity Scores.
Use parameters obtained from steps 1 and 2 to compute individual propensity scores ($\hat{\Phi}_{pt}$) and neighborhood propensity scores ($\hat{\Lambda}_{pt}$) for each unit.

IV. Estimate Outcome Model.
Use a mixed effects linear regression model to regress the outcome on *Z*, *G*, $\hat{\Phi}_{pt}$, and $\hat{\Lambda}_{pt}$.
Then, at every 10th percentile of *Z* and *G*, Use *Z* model to estimate $\hat{\Phi}_{pt}$ for each unit and use *G* model to calculate $\hat{\Lambda}_{pt}$ for each unit.

V. Impute Missing Potential Outcomes.
Use the computed values for $\hat{\Phi}_{pt}$ and $\hat{\Lambda}_{pt}$ to impute potential outcome at every 10th percentile of *Z* and *G* using the outcome model in step 4, resulting in an (Nx11x11) grid of imputed potential outcomes.

VI. Estimate the Average Dose-Response Function.
Compute the average of the above potential outcome grid across all N units, resulting in an 11x11 grid of the averaged imputed potential outcomes.

VII. Estimate the Jackknife Variances.
Generate *N* datasets, each with *N-1* units by systematically excluding one unit each time. For each dataset, repeat steps 3-6 in the estimation process and store the resulting 11x11 grid. Then, compute the jackknife variances for each cell in these grids, yielding a final 11x11 grid of jackknife variances.

Data and Application

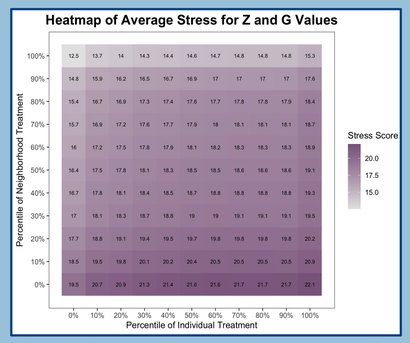

Figure 4. Shows an increase in stress score when *Z* increases and a decrease in stress score when *G* increases.

The longitudinal study application was conducted at Wellesley College in Fall 2022. Participants, aged 18+ with a vagina, were recruited for a 10-week study. For inclusion in this thesis analysis, participants had to complete at least one DASS survey during the study period (n=44). Each participant represents multiple data points (N=212). This study examines the effect of step count (measured by Fitbit) on stress (measured by Depression Anxiety Stress Scales).

Note that the analysis of steps and stress serve as examples for forthcoming microbiome data recorded in the study.

Limitations and Future Work

Assessing covariate balance with a continuous treatment variable poses challenges. Recent literature introduces new methods to address these complexities. However, this thesis does not adopt these methods, thus covariate balance is not evaluated after fitting the outcome model. Future research will explore and extend these methods further.

References. [1] Laura Forastiere, Edoardo M. Airoldi, and Fabrizia Mealli. Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 115(531):901–918, 2020. [2] Laura Forastiere, Davide Del Prete, and Valerio Leone Sciabolazza. Causal inference on networks under continuous treatment interference. *Social Networks*, 76:88–111, 2024. [3] D. B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980. [4] Michael E Sobel. What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.