# Simulating Simpson's Paradox

Maulino, J[1], Siño, MF[1], Horton, R[2]

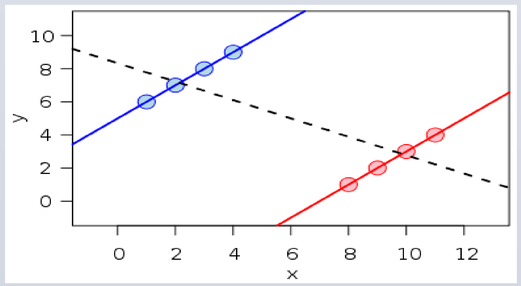
[1]University of San Francisco, [2]Microsoft

## ABSTRACT

Simpson's paradox is a statistical paradox which occurs when a trend in data disappears or reverses when additional factors are included in the analysis. Our project focuses on simulating this paradox in a large-scale simulated healthcare dataset. The dataset we worked with was created by reverse engineering admission histories and engineering analytical features from simulated data sources on an existing dataset about hospital readmission rates for diabetic patients. The simulated data sources include grocery purchases and wearable vital sign sensors and were used to engineer features such as percent calories from carbohydrates and minutes spent walking. In order to simulate Simpson's paradox, we extended this dataset by adding a genetic trait to act as a statistical confounder.

## BACKGROUND




Simple example of Simpson's Paradox:
Considered on their own, the red and blue lines display a positive correlation. However, when considered together, a negative correlation (the dotted line) appears.

## SIMULATING THE DATA

### Description of the Data

Response Variables

- readmitted - categorical variable, places patients in one of three groups: {NO, >30, <30}
- readmit30 - logical variable, TRUE if patients are readmitted within the first 30 days

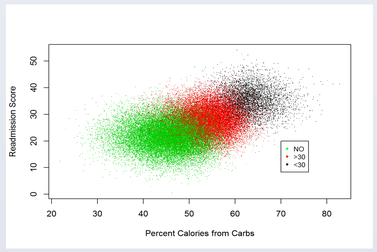
Predictors

- bmi: Body mass index
- pct_calories_from_carbs: percentage of calories from carbohydrates consumed
- sd_glucose: standard deviation of glucose level
- minutes_walking: minutes spent walking

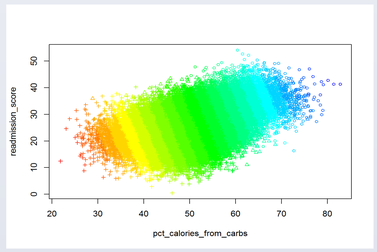
### Adding the Confounding variables

We selected the gene IRS1 as the confounding variable. We selected a subgroup from the data set to assign the rare IRS1 allele (RR). We looked at the different plots to see which group would be good for assigning the allele

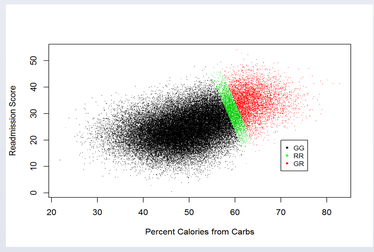
After adding the IRS1 column to the data set, we checked the model fit to check the coefficient of the IRS1 RR and percentage of calories from carbohydrates consumed interaction to see if it is negative.




This is scatterplot of the data set with the different colors representing the readmission of the patients




We separated the data into 25 stripes/groups and we aimed to choose a group that was around has 60-65% from carbohydrates consumed




We chose group 16 as the group with the IRS1 RR allele. The groups that consumed higher percentage of calories from carbohydrates consumed were assigned GR and the rest were assigned GG.

## ANALYZING THE SIMULATED DATA

After storing the new data set with the IRS1 allele, we proceeded to doing an exploratory analysis of the data as if we have no prior knowledge of the given data set.

```
glm(formula = readmit30 ~ bmi + pct_calories_from_carbs + minutes_walking +
    sd_glucose, family = binomial(link = "logit"), data = patients)
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -47.881200 | 1.084286 | -44.16 | <2e-16 | *** |
| bmi | 0.532937 | 0.014733 | 36.17 | <2e-16 | *** |
| pct_calories_from_carbs | 0.523030 | 0.013651 | 38.32 | <2e-16 | *** |
| minutes_walking | -0.053886 | 0.001572 | -34.28 | <2e-16 | *** |
| sd_glucose | 1.056010 | 0.044112 | 23.94 | <2e-16 | *** |

---

```
glm(formula = readmit30 ~ bmi + pct_calories_from_carbs + minutes_walking +
    sd_glucose + irs1, family = binomial(link = "logit"), data = patients)
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -31.561223 | 1.221225 | -25.84 | <2e-16 | *** |
| bmi | 0.534114 | 0.015392 | 34.70 | <2e-16 | *** |
| pct_calories_from_carbs | 0.219612 | 0.018621 | 11.79 | <2e-16 | *** |
| minutes_walking | -0.055099 | 0.001678 | -32.84 | <2e-16 | *** |
| sd_glucose | 1.072624 | 0.046475 | 23.08 | <2e-16 | *** |
| irs1GR | 3.442293 | 0.189642 | 18.15 | <2e-16 | *** |
| irs1RR | 1.625508 | 0.138784 | 11.71 | <2e-16 | *** |

Comparing the models with and without the IRS1 gene we can see that coefficient of pct_calories_from_carbs dropped after adding gene IRS1 to the model. Perhaps certain values of IRS1 are causing this so we fitted a model that includes the interaction of the IRS1 gene to the other variables.

## ANALYSIS (cont.)

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -33.787875 | 1.887187 | -17.904 | < 2e-16 | *** |
| irs1GR | 0.734116 | 2.958161 | 0.248 | 0.80401 | |
| irs1RR | 37.960662 | 4.458518 | 8.514 | < 2e-16 | *** |
| pct_calories_from_carbs | 0.262145 | 0.027672 | 9.473 | < 2e-16 | *** |
| sd_glucose | 1.264137 | 0.091085 | 13.879 | < 2e-16 | *** |
| minutes_walking | -0.059267 | 0.003153 | -18.795 | < 2e-16 | *** |
| bmi | 0.517459 | 0.024865 | 20.811 | < 2e-16 | *** |
| irs1GR:pct_calories_from_carbs | 0.021996 | 0.041917 | 0.525 | 0.59975 | |
| irs1RR:pct_calories_from_carbs | -0.605379 | 0.072458 | -8.355 | < 2e-16 | *** |
| irs1GR:sd_glucose | -0.388612 | 0.114109 | -3.406 | 0.00066 | *** |
| irs1RR:sd_glucose | -0.044864 | 0.135251 | -0.332 | 0.74011 | |
| irs1GR:minutes_walking | 0.008708 | 0.004035 | 2.158 | 0.03090 | * |
| irs1RR:minutes_walking | 0.001083 | 0.004693 | 0.231 | 0.81755 | |
| irs1GR:bmi | 0.052482 | 0.037249 | 1.409 | 0.15885 | |
| irs1RR:bmi | -0.015863 | 0.038736 | -0.410 | 0.68217 | |

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | 4.172787 | 4.039115 | 1.033 | 0.302 | |
| bmi | 0.501596 | 0.029698 | 16.890 | < 2e-16 | *** |
| pct_calories_from_carbs | -0.343234 | 0.066960 | -5.126 | 2.96e-07 | *** |
| minutes_walking | -0.058185 | 0.003475 | -16.742 | < 2e-16 | *** |
| sd_glucose | 1.219273 | 0.099973 | 12.196 | < 2e-16 | *** |

---

The first image shows the summary of the coefficients of the model with the interactions and we can see that the IRS1 RR allele "flipped" the coefficient of the percentage of calories from carbohydrates consumed. After getting the subset of the data with the IRS1 RR allele, we the coefficient of the percentage of c calories from carbohydrates consumed is still negative indicating that the patients possessing this allele has a lower likelihood of being readmitted despite having more calories consumed than other patients.

## CONCLUSION

Some unusual results surfaced when taking into account a patient's IRS1 genotype.

Specifically, patients with the IRS1 RR genotype appear to be less likely to be readmitted even with a higher percentage of calories from carbohydrates in their diet.

We recommend that more research be conducted on this gene in diabetic patients in regard to how it affects the way their body metabolizes calories from carbohydrates.

## RECOMMENDATION

**How can this project be further improved?**

- Adding more confounders
- Simulating more data sources from which more features can be engineered (ie. nutrition database, more sophisticated vital sensor data, etc.)