Paper 3897-2019

The Benefits of Keeping it Casual: Investigating Stress Mediated Health Outcomes with PROC CAUSALMED

Daniel Muzyka[1,2], Matthew Lypka[2]

Grand Valley State University[1], Spectrum Health Office of Research[2]

ABSTRACT

In epidemiological science, researchers aim to associate exposures to specific health outcomes. In psychological sciences, researchers want to investigate how an intervention improves quality of life. In the social sciences, policy makers want to quantify how effective work release programs are in reducing re-incarceration rates. When investigating these relationships, it is tempting to conclude that an independent *X* causes a dependent *Y*. However, though significant associations may exist, the mechanisms behind those relationships are not always apparent.

Often, additional variables influence the impact of your independent *X* variable. To accurately model these additional variables in inferential statistics, you must classify them based on their hypothesized role in the relationship. In some situations, it may be hypothesized that *X* acts on, and by proxy, exerts effect on *Y* through a mediating variable, M. To properly attribute the significance and influence of mediating variables when preforming analyses, researchers may use Causal Mediation Analysis (CMA).

A procedure known as CAUSALMED, introduced in SAS ® 9.4, allows for the use of CMA when looking at how *X* indirectly affects *Y* via M. In this paper, we highlight how PROC CAUSALMED, in conjunction with various other regression and pathway modeling procedures, was used to conduct CMA on data from publicly available data sets to investigate how stress may act as a mediating variable in health outcomes.

INTRODUCTION

This paper will serve as an introduction to Causal Mediation Analysis using a real world example to illustrate how to conduct a preliminary search for a mediating relationship, and how to interpret the results.

The goal of Causal Mediation Analysis (CMA) is to model the pathways through which predictor (independent) variables relate to outcomes (dependent variables), including intermediary variables. To illustrate this idea, we can glean examples from real world studies. An example that may be commonly understood by most is the idea that stress may not only relate to increased caloric intake, but also a shift in food preference from lower fat to higher fat foods (Zellner et al., 2006). In these studies, individuals were randomized to either receive stressful stimuli or to be part of the control group. The causal diagram in Figure 1 below depicts the causal associations investigated in these randomized experiments as *X (predictor)* relating to *Y (outcome)*.

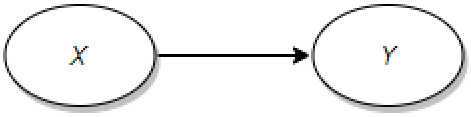


Figure 1. Basic Cause and Effect Relationship Diagram

The randomization present in these experiments allows us to assume that any propensity towards dietary behavior or other confounding variables are equally represented across groups. This grants the researchers the ability to conclude that any change in dietary behavior is a result of the treatment, allowing them to estimate a causal treatment effect. While the resulting associations of these studies are intriguing, they are not fully explanatory. Although researchers are within their means to postulate that a cause and effect relationship exists, their conclusion would not be explicit in how that effect was manifested. The aforementioned studies lack what would be considered a mechanism of action. A mechanism of action can be described as a process (typically a biological one) in which a substance produces an effect (Spratto, Woods, 2009). Other studies have utilized the concept of a mechanism of action through the hypothesis that the hormone cortisol may play a role in the causal relationship due to it links to both physiological stress and hunger. (Epel et al., 2001).

Stringing these hypotheses together, we could hypothesize two concurrent pathways leading from a stressful situation to increased caloric intake. The first is the natural direct pathway, stressful stimuli relating to increased caloric intake. The second is referred to as the indirect pathway in CMA. This pathway goes as follows: The presence of a stressful stimulus leads to an increased level of cortisol production, which in turn leads to an increased caloric intake. The two concurrent pathways are shown in Figure 2. In CMA, the main goal is to assess and interpret these as competing pathways with the aim to determine which pathway accounts for most of the effect.

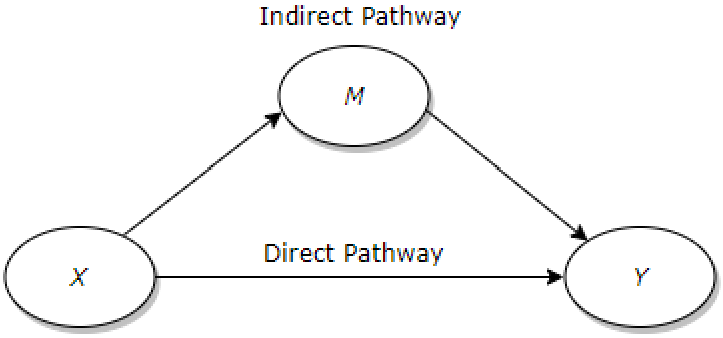


Figure 2. Direct and Indirect Pathway Diagram

Learning which pathway accounts for more of the effect may have practical decision making implications. Therefore, when looking at the diagram above in Figure 2, it is useful to imagine the two pathways as competing so we can decompose their effects. Decomposition is a method of interpretation in CMA where the first objective is to estimate how much of the outcome can be derived as the result of the indirect pathway. After estimating how much of the effect can be attributed to the indirect pathway, the residual difference can be reasonably attributed to the direct pathway. If it is determined that most of the effect is being transmitted through the indirect pathway’s mechanism of action, the conclusion is that the relationship is the result of full mediation, i.e. through indirect effects. Much more commonly, the indirect pathway will have only an ancillary effect on the outcome, and in these cases, the relationship would be considered the result of partial mediation. The difference observed in the influence of each pathway may inform researchers if they are within their means to intervene in this process to generate desirable outcomes.

To conduct CMA, we will be using PROC CAUSALMED, a new procedure introduced in SAS/STAT package 14.3.PROC CAUSALMED was created as an alternative to PROC CAUSALTRT for estimating direct and indirect effects of a treatment variable on an outcome. Along with the including the mediating variable, the CAUSALMED Procedure can also include covariates as well as interactions between the treatment and mediator variable (SAS Institute, 2017). Covariates are a useful addition to PROC CAUSALMED because including these allows for some control of confounding between *X* (Treatment), *Y* (Outcome), and *M* (Mediator). Referencing a paper published by Yiu-Fai Yung, et al. (2018) from the SAS Institute, PROC CAUSALMED can incorporate variables of the following type into the four roles in the modeling process:

- outcome variable *Y* : binary, continuous, or count
- treatment variable *X* : binary or continuous
- mediator variable *M* : binary or continuous
- covariates *C*: categorical or continuous

As described in the stress eating example, relationships may be subject to indirect effects via mediating variables. In order to further explore this idea of mediation, this paper will utilize public data from the STRIDE project to assess if stress mediates other health outcomes. This data set, created by researchers from a consortium of universities and funded by the National Institutes of Health/National Institute of Mental Health, was conceived with the purpose of assessing the link between stress and mental health in a community largely composed of minority identities relating to race, gender, and sexual orientation (Meyer et al., 2018). The observational study followed up with participants after one year. This data set was selected for demonstration of causal mediation for its attention to the multiple types of stressors, as well as having pre and post measurements. The results of the analysis in the paper are exploratory and are intended to serve an illustrative function for PROC CAUSALMED.

PREPARING THE DATASET

SELECTING THE VARIABLES OF INTEREST

The dataset used throughout this paper to model stress as a mediating variable with PROC CAUSALMED was taken from the "Project STRIDE: Stress, Identity, and Mental Health" dataset (Meyer et al., 2018).To begin, the subject pool was limited to participants who had the variables of interest measured at time one and time two (N=370). To narrow our scope, we grouped the available variables of interest into three categories based on our preliminary understanding of the possible relationships: potential predictors (*X*), measures of stress (*M*), and possible outcomes (*Y*). The next step was to use prior knowledge to narrow our list based on some assumptions and the previously stated capabilities of SAS PROC CAUSALMED.

Adhering to the capabilities of the procedure, some variables like our measures of stress which were ordinal responses (true, somewhat true, and false) had to be recoded into binary variables (at least somewhat true, false). Another variable that was changed was ordinal household income. As a categorical variable, household income had a range of $0 to over $1,000,000 in 34 intervals of increasing size. Household income was recoded to represent ordinal categories of equal intervals ($25,000). Due to this recoding, some participants who had income over $100,000 had to be removed because they did not fall into one of the redefined interval bins (N=8 removed). For reference, the data step used for all cleaning and recoding purposes of the variables used in the final model can be found as Appendix Item 1.

After confirming our list of variables was compatible, the list was further refined through assumption checking. The first implicit assumption considered was temporal precedence, meaning we had to be certain that our pathways occurred in a reasonable chronological progression in one direction. Due to the observational nature of the data, temporal precedence cannot be guaranteed, but steps were taken to justify the approach taken. This was accomplished in a time relative sense between predicting and outcomes by using predictors and mediators as measured from time one, and the outcome measures from time two. The other facet of temporal precedence we considered was limiting the possibility of feedback between our *X* and *M*. Feedback in these circumstances would be our *X* affects the level of our *M*, and then our level of *M* affects our level of *X*.

Looking to Valeri and VanderWeele (2013), we can see four other assumptions that are required in estimation of causal mediation effects:

- No unmeasured confounding of the treatment-outcome relationship
- No unmeasured confounding of mediator-outcome relationship
- No unmeasured confounding of the treatment-mediator relationship
- No mediator-outcome confounder that is affected by the treatment

The first two assumptions are described by Valeri and VanderWheele as necessary for identifying a controlled direct effect, while all four assumptions are necessary when attempting to identify the natural direct or natural indirect effects. The assumptions of no confounder existing for the treatment-outcome and treatment-mediator outcomes can be reasonably assumed with randomization of the treatment, but it is importance to recognize that even RCT cannot guarantee the non-existence of a confounder between the mediator and outcome. When selecting variables, care was taken to select a comprehensive list of variables that may be considered as confounders.

Once a tentative list of variables was assembled, an exploratory process of automated and hand curated model selection began. The first step was to regress our outcome variables such as depression, hypertension, and sleep issues on our predictors such as education, relationship status, and sexual orientation. After refining our list of predictors to those that reliably predict some outcome variable, mediating variables were analyzed. The mediator variables in question were our best attempt to find a proxy for stress, chronic strains. These variables included but were not limited to general chronic strain, chronic financial strain, chronic parental strain, and chronic relationship strain. After looking at which mediators were reliably predicted by our list of *X*, we settled on a list of seven variables to further investigate.

The list of variables remaining in our consideration is as follows:

- **HI_CESD_2**, outcome, binary high vs low, Center for Epidemiologic Studies Depression (CES-D) scores
- **Edu_hsd,** predictor, binary level of education, (> high school diploma or ≤ high school diploma)
- **Fin_strain**, mediator, binary, presence of financial strain [At least somewhat financially strained, Not financially strained]
- **Age**, predictor, covariate, continuous variable, age
- **Gender**, predictor, covariate, binary, Male or Female
- **Ethnic**, predictor, covariate, categorical, ethnicity
- **Hi_ord2**, predictor, covariate, ordinal, household income grouped by us into $25k intervals up to $100k

INSPECTING SIMPLE RELATIONSHIPS

Knowing which variables are of interest, basic numerical summaries and frequencies were examined to understand the data. Basic univariate statistics were generated with PROC MEANS and PROC FREQ as follows:

```
proc means data=stride maxdec=2;
      class edu_hsd;
      var age;
run;

proc freq data=stride;
      table edu_hsd*(ethnic fin_strain gender hi_cesd_2 hi_ord2)
      /nocol nocum nopercent;
run;
```

Looking at the PROC MEANS output below in Table 1, it is shown that that our only continuous variable, age, seems to be relatively similar across groups with a mean of 32.69 ± 8.97 in the greater than high school diploma group and 32.35 ± 9.59 in the high school diploma or less group.

Table 1. Numeric Summary of Age

Analysis Variable : AGE

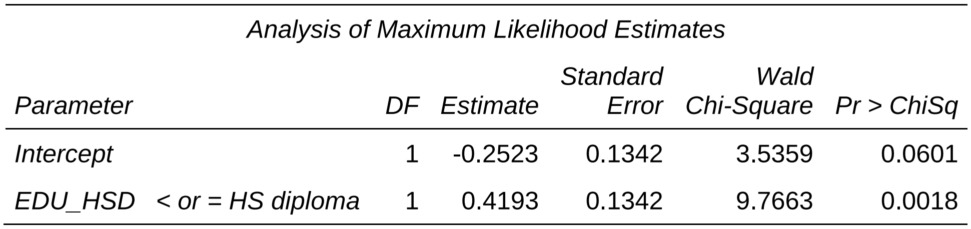
EDU_HSD	*N Obs*	*N*	*Mean*	*Std Dev*	*Minimum*	*Maximum*
> HS education	278	278	32.69	8.97	18.00	58.00
< or = HS diploma	72	72	32.35	9.59	18.00	54.00

Referencing the output from Appendix Item 2 through Appendix Item 6, it is possible to see how our categorical variables differ by the category being considered as the treatment variable, continuation of education after high school. For all of the categorical variables besides gender, we see uneven representations across our education levels when looking at row percentages. Since these variables are considered to play a role in our outcome of interest, we would like to somehow control for these. If we had access to a larger sample size, a method like propensity matching may have been appropriate. Since we are working with a limited sample, we can use the COVAR statement in PROC CAUSALMED to include these as covariates in our final model.

With exploratory analysis completed, it is useful to assess how reliably *X* predicts *M*. We accomplished this using PROC LOGISTIC with a ROC option:

```
proc logistic data=stride plots=roc;
   class hi_cesd_2 edu_hsd;
   model hi_cesd_2 = edu_hsd;
run;
```

Table 2. Parameter Estimates of *X* Predicting *Y*



Analysis of Maximum Likelihood Estimates

				Standard	*Wald*	
Parameter		*DF*	*Estimate*	*Error*	*Chi-Square*	*Pr > ChiSq*
Intercept		1	-0.2523	0.1342	3.5359	0.0601
EDU_HSD	*< or = HS diploma*	1	0.4193	0.1342	9.7663	0.0018

In Table 2 above, we can see that while continuing education beyond a high school diploma had strong evidence of statistical association with CESD depression score (p = 0.0018), the resulting ROC plot seen below in Figure 3 illustrates that this education level alone is not a reliable predictor. The area under the curve of the ROC and c-statistic tell us that the model is only expected to be accurate 57.06% of the time in a population like this one. This low reliability, which is only marginally better than a coin toss, would be considered a red flag for CMA. But for the purpose of illustrating the functions of PROC CAUSALMED in this paper, we will continue assessing the model.

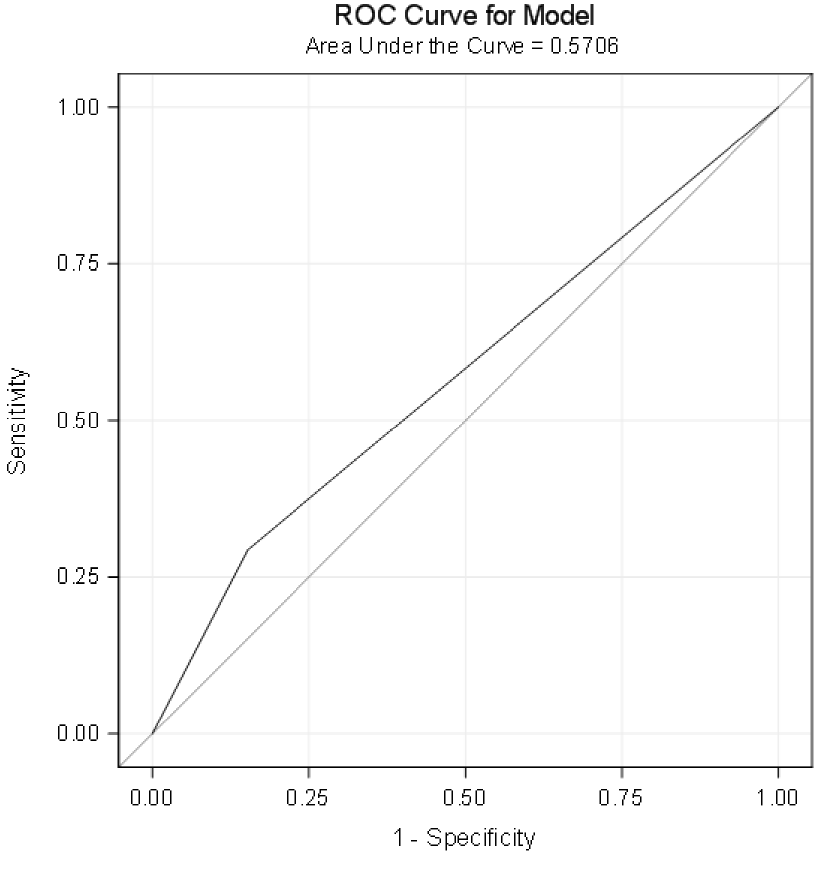


Figure 3. ROC for *X* Predicting *Y*

The next step in preliminary model assessment was to check how reliably our *M*, chronic financial strain, is predicted by our *X,* high school education level. This was once again accomplished using PROC LOGISTIC with ROC plot options. The results, which can be found as below in Table 3 indicated that while high school completion had strong evidence of sharing an association with financial strain (p=0.008), our ROC in Figure 4 illustrated that the model was accurate only 56.07% of the time in predicting chronic financial strain by education level.

Table 3. Parameter Estimates for *X* Predicting *M*

Analysis of Maximum Likelihood Estimates

Parameter		*DF*	*Estimate*	*Standard Error*	*Wald Chi-Square*	*Pr > ChiSq*
Intercept		1	0.7689	0.1516	25.7393	<.0001
EDU_HSD	*< or = HS diploma*	1	0.4052	0.1516	7.1492	0.0075

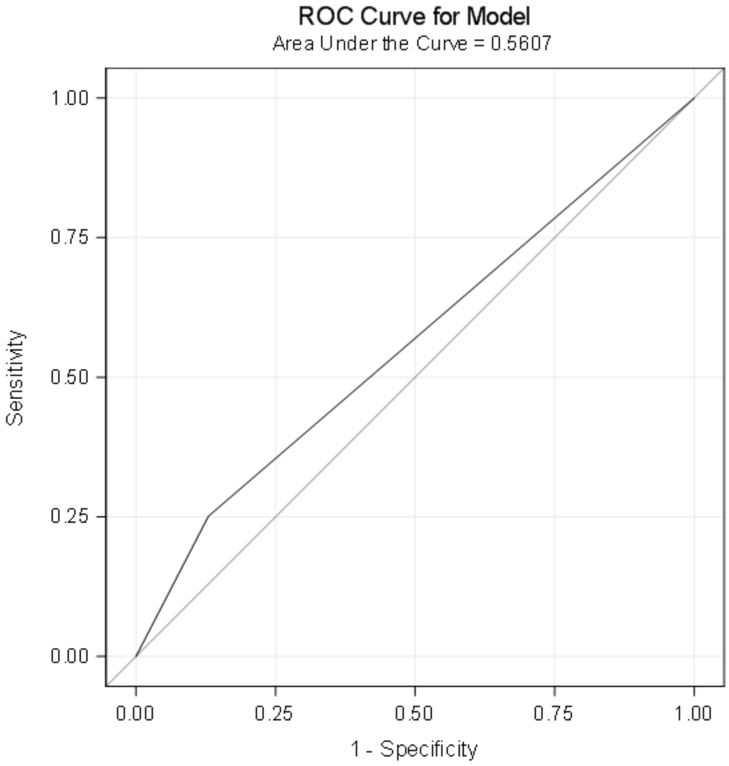


Figure 4. ROC for *X* Predicting *M*

USING PROC CAUSALMED

A BASIC MODEL

The model and output that follow are an attempt to analyze our data using PROC CAUSALMED in its most basic form. This model does not take into consideration the covariates we know to exist. The analysis using this basic model serves two purposes in the context of this paper. First, this example provides us with the best chance for interpretation of a model from the real-world data set analyzed in this paper. This serves the purpose of illustrating comprehensible output generated by the CAUSALMED Procedure. Secondly, this is an example of the spurious results that can be obtained in observational studies when proper precautions are not taken. Without randomization to control for cofounding variables, we cannot reasonably assume they are balanced among groups. The concern of unmeasured or uncontrolled covariates along with the low predictive association between our *X&Y* and *X&M* have raised concerns for our model, but we will interpret for illustrations sake at an ɑ=0.1.

We can use PROC CAUSALMED to analyze our data without including known covariates with the following code:

```
 proc causalmed data=stride alpha=.1;

      class edu_hsd hi_cesd_2 fin_strain gender ethnic hi_ord2;
      model hi_cesd_2=edu_hsd | fin_strain;
      mediator fin_strain=edu_hsd;
 run;
```

Table 4. Model Information

Model Information	
Data Set	WORK.STRIDE
Outcome Variable	HI_CESD_2
Treatment Variable	EDU_HSD
Mediator Variable	fin_strain
Outcome Distribution	Binomial
Outcome Link Function	Logit
Mediator Distribution	Binomial
Mediator Link Function	Logit

Table 5. Observations Used

Number of Observations Read	350
Number of Observations Used	350

Table 6. Class Level Information

Class Level Information		
Class	*Levels*	*Values*
EDU_HSD	2	< or = HS diploma > HS education
HI_CESD_2	2	High CES-D score Low CES-D score
fin_strain	2	At least somewhat financially strained Not financially strained

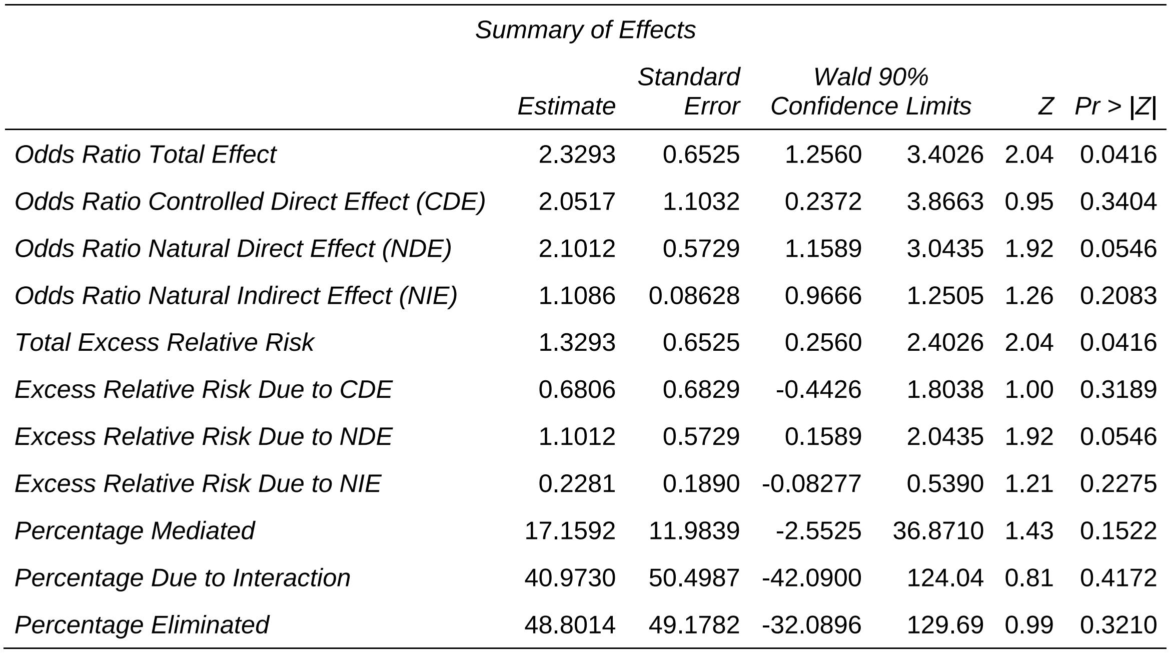
The output shown above is the first output generated by PROC CAUSALMED. Table 4 shows us that SAS recognized that both our mediator and outcome a binary responses and assumed a binomial distribution with a logit link for both pathways. You could explicitly specify these distributions and links using the DIST= and LINK= keywords respectively. We can see in Table 5 that only 350 observations were included in this basic model. The 20 records that were not included were intentionally removed with list wise deletion due to missing values on variables that will be used in our full model. This list wise deletion was done now so that when we compare the models we are certain the same observations are included. Finally, in Table 6 we can see the class levels of our *X*, *Y*, and *M* respectively.

Since all of our variables are categorical, SAS also provided the response profiles for all of the variables included. These outputs are located at the end of this document as Appendix Item 8 through Appendix Item 10. These outputs also tell us which level is being modeled as the outcome. For this analysis, we modeled the probability of a high CES-D as the outcome, and being at least somewhat financially strained as the mediator.

Table 7 below shows the main estimation results of PROC CAUSALMED when a binary outcome is used. The results of the odds ratio total effect shown below in Table 7 demonstrate evidence of an overall association between *X* and *Y* with a Z-score of 2.04 (p=.042). Looking at our response modeled, this output says that the odds of having a high self-reported depression score are higher for those who have not graduated high school or continued to higher education. The output tells us the odds of having a high self-reported CES-D score are estimated to be 2.33 times greater for people who did not finish or continue education after high school. The 90% Wald Confidence Limits tell us that we can conclude with 90% confidence from what we observed in this data that the odds of this relationship are between 1.26 and 3.40 times greater.

The natural direct effect was found to have moderate statistical support when evaluated at a threshold of ɑ=0.1 (p=.055) while the natural indirect effect was not found to have strong statistical support (p=.208).This informs the conclusion that a majority if not all of the overall effect we observed can be attributed to the natural direct pathway. Again, remember this model is ill-informed by not considering covariates.

Table 7. Summary of Effects for Basic Model



Summary of Effects

		Standard	Wald 90%			
	Estimate	Error	Confidence Limits		Z	Pr > |Z|
Odds Ratio Total Effect	2.3293	0.6525	1.2560	3.4026	2.04	0.0416
Odds Ratio Controlled Direct Effect (CDE)	2.0517	1.1032	0.2372	3.8663	0.95	0.3404
Odds Ratio Natural Direct Effect (NDE)	2.1012	0.5729	1.1589	3.0435	1.92	0.0546
Odds Ratio Natural Indirect Effect (NIE)	1.1086	0.08628	0.9666	1.2505	1.26	0.2083
Total Excess Relative Risk	1.3293	0.6525	0.2560	2.4026	2.04	0.0416
Excess Relative Risk Due to CDE	0.6806	0.6829	-0.4426	1.8038	1.00	0.3189
Excess Relative Risk Due to NDE	1.1012	0.5729	0.1589	2.0435	1.92	0.0546
Excess Relative Risk Due to NIE	0.2281	0.1890	-0.08277	0.5390	1.21	0.2275
Percentage Mediated	17.1592	11.9839	-2.5525	36.8710	1.43	0.1522
Percentage Due to Interaction	40.9730	50.4987	-42.0900	124.04	0.81	0.4172
Percentage Eliminated	48.8014	49.1782	-32.0896	129.69	0.99	0.3210

Table 8 below shows the Percentage Decomposition table, which can be requested using the PALL option, to help quantify the percentage of effect being transmitted through each pathway. The natural direct effect is definitely seen as carrying the most weight with an estimated 82.84% of the effect being transmitted via this path. The indirect effect path estimated as representing 17.16% of the transmitted effect.

Table 8. Percentage Decompositions for Basic Model

Percentage Decompositions of Total Excess Relative Risk

			Standard	Wald 90%			
Decomposition	*Excess Relative Risk*	*Percent*	*Error*	*Confidence Limits*		*Z*	*Pr > |Z|*
NDE+NIE	*Natural Direct*	82.84	11.98	63.13	102.55	6.91	<.0001
	Natural Indirect	17.16	11.98	-2.55	36.87	1.43	0.1522
CDE+PE	*Controlled Direct*	51.20	49.18	-29.69	132.09	1.04	0.2978
	Portion Eliminated	48.80	49.18	-32.09	129.69	0.99	0.3210
TDE+PIE	*Total Direct*	92.17	5.42	83.26	101.09	17.00	<.0001
	Pure Indirect	7.83	5.42	-1.09	16.74	1.44	0.1487
NDE+PIE+IMD	*Natural Direct*	82.84	11.98	63.13	102.55	6.91	<.0001
	Pure Indirect	7.83	5.42	-1.09	16.74	1.44	0.1487
	Mediated Interaction	9.33	11.80	-10.08	28.74	0.79	0.4292
CDE+PIE+PAI	*Controlled Direct*	51.20	49.18	-29.69	132.09	1.04	0.2978
	Pure Indirect	7.83	5.42	-1.09	16.74	1.44	0.1487
	Portion Due to Interaction	40.97	50.50	-42.09	124.04	0.81	0.4172
Four-Way	*Controlled Direct*	51.20	49.18	-29.69	132.09	1.04	0.2978
	Reference Interaction	31.64	39.08	-32.63	95.92	0.81	0.4181
	Mediated Interaction	9.33	11.80	-10.08	28.74	0.79	0.4292
	Pure Indirect	7.83	5.42	-1.09	16.74	1.44	0.1487

Note: NDE=CDE+IRF, NIE=PIE+IMD, PAI=IRF+IMD, PE=PAI+PIE, TDE=CDE+PAI.

A MODEL INVOLVING COVARIATES

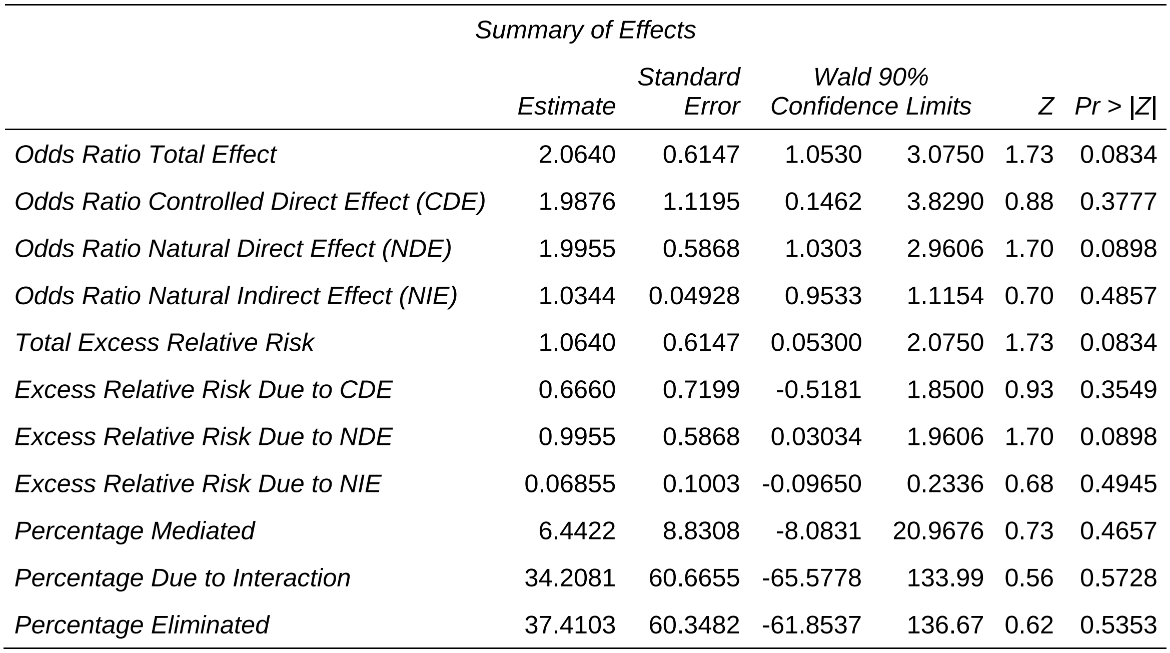
To more accurately model the real world data and better satisfy the underlying assumptions of CMA, we also ran a model including the covariates we had specified previously. These covariates are gender, age, ethnicity, and ordinal household income. Covariates can be included by using a COVAR statement as seen in the code below.

```
proc causalmed data=stride alpha=.1;

      class edu_hsd hi_cesd_2 fin_strain gender ethnic hi_ord2;
      model hi_cesd_2=edu_hsd | fin_strain;
      mediator fin_strain=edu_hsd;
      covar gender age ethnic hi_ord2;
run;
```

The results of this model vary slightly from the initial model, but not in any way that changed our conclusions at ɑ=0.1. Breaking down the changes, we can see that we still observed moderate support for an overall odds ratio total effect (p=0.083) and moderate evidence for natural direct effect (p=0.090). We can note that all of the odds ratio estimates in Table 9 have moved closer to one compared to those in Table 7, representing less statistical support for these effects contributing to the outcome. This is because when we included covariates that the underlying regression models deemed important and some of the effect observed was therefore attributed to these covariates. This illustrates the concern of unmeasured confounding variables in observational studies.

Table 9. Summary of Effects for Model Including Covariates



Summary of Effects

	Estimate	*Standard Error*	*Wald 90% Confidence Limits*		*Z*	*Pr > |Z|*
Odds Ratio Total Effect	2.0640	0.6147	1.0530	3.0750	1.73	0.0834
Odds Ratio Controlled Direct Effect (CDE)	1.9876	1.1195	0.1462	3.8290	0.88	0.3777
Odds Ratio Natural Direct Effect (NDE)	1.9955	0.5868	1.0303	2.9606	1.70	0.0898
Odds Ratio Natural Indirect Effect (NIE)	1.0344	0.04928	0.9533	1.1154	0.70	0.4857
Total Excess Relative Risk	1.0640	0.6147	0.05300	2.0750	1.73	0.0834
Excess Relative Risk Due to CDE	0.6660	0.7199	-0.5181	1.8500	0.93	0.3549
Excess Relative Risk Due to NDE	0.9955	0.5868	0.03034	1.9606	1.70	0.0898
Excess Relative Risk Due to NIE	0.06855	0.1003	-0.09650	0.2336	0.68	0.4945
Percentage Mediated	6.4422	8.8308	-8.0831	20.9676	0.73	0.4657
Percentage Due to Interaction	34.2081	60.6655	-65.5778	133.99	0.56	0.5728
Percentage Eliminated	37.4103	60.3482	-61.8537	136.67	0.62	0.5353

A CONDITIONAL MODEL

Another type of model we can apply using PROC CAUSALMED is a conditional model. By default, PROC CAUSALMED models the effect using the "average" covariate levels (Yiu-Fai Yung et al., 2018). This helps to create a model that applies to the population in general, but researchers may be interested in how mediation varies for different sub-populations. Using the EVALUATE statement allows you to specify the levels of the covariates for which you are interested in examining the mediation relationship. This conditional modeling can be useful if there are pre-specified sub-populations that may be considered at risk or have different characteristics that imply the relationship may be different.

In the code below you can see four ESTIMATE statements were added. These statements include a label for the estimate as well as the specification of covariate levels. Notice that the categorical variables are referenced using their formatted levels. The first two statements represent creating models for differing ages, either 18 or 45. The second two statements create models for a white male and another for an African-American female.

```
proc causalmed data=stride alpha=.1;
      class edu_hsd hi_cesd_2 fin_strain gender ethnic hi_ord2;
      model hi_cesd_2=edu_hsd | fin_strain;
      mediator fin_strain=edu_hsd;
      covar gender age ethnic hi_ord2;
      evaluate 'age=18' age=18;
      evaluate 'age=45' age=45;
      evaluate 'White Male' ethnic='White' gender ='Male';
      evaluate 'African-American Female' ethnic='Black/African-American'
      gender ='Female';
run;
```

Looking below at the output from our first two evaluate statements in Table 10 and Table 11 we can see some differing estimates. Most notably, the overall odds ratio total effect found support below the α=0.1 level when controlling for age = 45 (0.088), while for age = 18 the support for an overall association was weaker (p=0.101). This indicates that the over association may not be as pronounced for younger ages. We can see that for both age groups the natural direct effects have diminished support compared to the values observed in our overall model Table 9.

Table 10. Summary of Effects for Model Age=18

Summary of Effects: age=18

		Standard	*Wald 90%*			
	Estimate	*Error*	*Confidence*	*Limits*	*Z*	*Pr > |Z|*
Odds Ratio Total Effect	2.0785	0.6570	0.9978	3.1591	1.64	0.1007
Odds Ratio Controlled Direct Effect (CDE)	1.9876	1.1195	0.1462	3.8290	0.88	0.3777
Odds Ratio Natural Direct Effect (NDE)	1.9931	0.6516	0.9213	3.0649	1.52	0.1275
Odds Ratio Natural Indirect Effect (NIE)	1.0428	0.06677	0.9330	1.1527	0.64	0.5213
Total Excess Relative Risk	1.0785	0.6570	-0.00216	2.1591	1.64	0.1007
Excess Relative Risk Due to CDE	0.7626	0.8339	-0.6091	2.1342	0.91	0.3605
Excess Relative Risk Due to NDE	0.9931	0.6516	-0.07873	2.0649	1.52	0.1275
Excess Relative Risk Due to NIE	0.08535	0.1286	-0.1261	0.2968	0.66	0.5067
Percentage Mediated	7.9140	12.2202	-12.1864	28.0143	0.65	0.5172
Percentage Due to Interaction	25.3581	51.1013	-58.6961	109.41	0.50	0.6197
Percentage Eliminated	29.2918	51.8529	-55.9987	114.58	0.56	0.5721

Table 11. Summary of Effects for Model Age=45

Summary of Effects: age=45

		Standard	*Wald 90%*			
	Estimate	*Error*	*Confidence*	*Limits*	*Z*	*Pr > |Z|*
Odds Ratio Total Effect	2.0780	0.6329	1.0370	3.1191	1.70	0.0885
Odds Ratio Controlled Direct Effect (CDE)	1.9876	1.1195	0.1462	3.8290	0.88	0.3777
Odds Ratio Natural Direct Effect (NDE)	1.9938	0.6189	0.9758	3.0118	1.61	0.1083
Odds Ratio Natural Indirect Effect (NIE)	1.0422	0.06410	0.9368	1.1477	0.66	0.5099
Total Excess Relative Risk	1.0780	0.6329	0.03700	2.1191	1.70	0.0885
Excess Relative Risk Due to CDE	0.7335	0.7991	-0.5810	2.0479	0.92	0.3587
Excess Relative Risk Due to NDE	0.9938	0.6189	-0.02416	2.0118	1.61	0.1083
Excess Relative Risk Due to NIE	0.08423	0.1254	-0.1220	0.2905	0.67	0.5018
Percentage Mediated	7.8132	11.5753	-11.2264	26.8528	0.67	0.4997
Percentage Due to Interaction	28.0789	54.4453	-61.4756	117.63	0.52	0.6060
Percentage Eliminated	31.9625	54.8358	-58.2343	122.16	0.58	0.5600

Looking at the output from our second two EVALUATE statements, the results below show some more differing estimates. We can see both models still have evidence of an odds ratio total effect at a threshold of ɑ=0.1, but we can see difference in the natural direct effect when using this threshold. For African-American females, there is slightly more evidence for the direct effect (p=0.090) compared to white males (p=0.116).

Table 12. Summary of Effects for Model ‘White Male’

Summary of Effects: White Male

		Standard	*Wald 90%*			
	Estimate	*Error*	*Confidence*	*Limits*	*Z*	*Pr > |Z|*
Odds Ratio Total Effect	2.0786	0.6426	1.0216	3.1356	1.68	0.0933
Odds Ratio Controlled Direct Effect (CDE)	1.9876	1.1195	0.1462	3.8290	0.88	0.3777
Odds Ratio Natural Direct Effect (NDE)	1.9935	0.6325	0.9531	3.0339	1.57	0.1163
Odds Ratio Natural Indirect Effect (NIE)	1.0427	0.06561	0.9348	1.1506	0.65	0.5151
Total Excess Relative Risk	1.0786	0.6426	0.02162	2.1356	1.68	0.0933
Excess Relative Risk Due to CDE	0.7467	0.8147	-0.5933	2.0867	0.92	0.3594
Excess Relative Risk Due to NDE	0.9935	0.6325	-0.04690	2.0339	1.57	0.1163
Excess Relative Risk Due to NIE	0.08514	0.1274	-0.1245	0.2947	0.67	0.5040
Percentage Mediated	7.8931	11.9125	-11.7013	27.4874	0.66	0.5076
Percentage Due to Interaction	26.8505	52.9518	-60.2474	113.95	0.51	0.6121
Percentage Eliminated	30.7738	53.5071	-57.2375	118.79	0.58	0.5652

Table 13. Summary of Effects for Model ‘African-American Female’

Summary of Effects: African-American Female

		Standard	*Wald 90%*			
	Estimate	*Error*	*Confidence*	*Limits*	*Z*	*Pr > |Z|*
Odds Ratio Total Effect	2.0638	0.6147	1.0527	3.0749	1.73	0.0835
Odds Ratio Controlled Direct Effect (CDE)	1.9876	1.1195	0.1462	3.8290	0.88	0.3777
Odds Ratio Natural Direct Effect (NDE)	1.9955	0.5868	1.0303	2.9607	1.70	0.0898
Odds Ratio Natural Indirect Effect (NIE)	1.0342	0.04896	0.9537	1.1147	0.70	0.4846
Total Excess Relative Risk	1.0638	0.6147	0.05269	2.0749	1.73	0.0835
Excess Relative Risk Due to CDE	0.6653	0.7193	-0.5178	1.8484	0.92	0.3550
Excess Relative Risk Due to NDE	0.9955	0.5868	0.03032	1.9607	1.70	0.0898
Excess Relative Risk Due to NIE	0.06828	0.09974	-0.09577	0.2323	0.68	0.4936
Percentage Mediated	6.4186	8.7749	-8.0147	20.8520	0.73	0.4645
Percentage Due to Interaction	34.2715	60.7467	-65.6479	134.19	0.56	0.5726
Percentage Eliminated	37.4620	60.4225	-61.9242	136.85	0.62	0.5353

DISCUSSION OF THE MODELS

Looking at overall model fit, we also saw low prediction from *X* to *Y* and *X* to *M*. Even still, if we look to the model without covariates (Table 7), just for interpretations sake, it somewhat informs possible policy decisions to reduce depression. The results show that the natural direct path is the only path with even moderate evidence. This could say that the reducing stress (intervening on the indirect path) may be less effective for reducing depression in individuals, but for the portion of the population who would not continue education past high school, there is a large effect to capitalize on by encouraging further education. Increasing efforts to assist people to not only finish high school but continue after would be beneficial for this sub-population.

In full model with covariates, we added the third and fourth ESTIMATE statements to model the covariates levels of white male and African-American female (Table 12 and Table 13). We observed a small difference in the evidence of a total effect between these, but the more pronounced difference was on the evidence for the direct effect. For white males, the natural direct effect (p=0.116) was less supported by the data than the direct effect for African-American females (p=0.090). This lends some evidence to the conclusion that there may be other factors in play for these demographics. It seems that the direct effect is buffered for white males, amplified for African-American females, or both. This could illustrate disparities that result in white males faring better with less education than their African-American female counterparts, at least in regards to depression. While these conclusions are nothing more than new hypotheses drawn from exploratory work on observational data, they can inform research decisions for the future.

Looking at some of the limitations in this paper, there was an aim to use the new CAUSALMED Procedure to model stress as a mediator on health outcomes. Searching to find available data resulted in us building this model on a relatively small data set with a rather specific population. Working with observational data also presented issues for assumption checking, but in the end, we have a model that can be assessed, even if we conclude that a causal mediation model is not the most appropriate fit in this instance.

We sought to model stress as a mediator; we did the best we could by finding a set of proxy variables, chronic strains. This is a limitation for our desired interpretations, but addressing the construct of stress through observational data was expected to be troublesome. We also cannot not expect our mediating variable to be precise with the ordinal categories it had, along with our compulsory recoding to a binary variable. Also, this is a self-reported measure so test-retest validity could be low especially if we think about how unexpected financial burdens can be a factor, even if the strain is considered in a 'chronic' sense. Since multiple chronic strains were available, we probably could have considered the interaction of all the chronic strains. Instead of doing this, there was a category for 'general chronic strain' which we examined, but this was not as reliably predicted from our list of predictors.

When selecting our predictor from those significantly associated with our outcomes, we chose high school education because we thought for most people this was temporally 'locked' in place. In reality, this is not the case especially since the criteria includes graduated but did not go further. A person could go back to finish high school or start college at any time. The assumption that high school was temporally locked would also create an inconsistency that people have had varying amounts of time since the 'treatment'. This is why the conditional model involving age could be more practical. We observed that there was less evidence for the odds ratio total effect in the 18 year old population vs 45 year old. We can hypothesize this is because the effects of not continuing your education have not manifested at the age of 18.

CONCLUSION

Overall, the CAUSALMED Procedure was easy to use and has practical applications, but researchers must be cognizant of the underlying assumptions. Aside from this project, we may find that PROC CAUSALMED will have increased practicality when it comes to the data housed in Electronic Health Records. Physicians and researchers should be aware of potential mediating relationships in both clinical and research settings when looking to evaluate the effectiveness of interventions.

REFERENCES

Epel, E., Lapidus, R., McEwen, B., & Brownell, K. (2001). Stress may add bite to appetite in women: a laboratory study of stress-induced cortisol and eating behavior. Psychoneuroendocrinology, 26(1), 37-49.

Meyer, Ilan H., Dohrenwend, Bruce Philip, Schwartz, Sharon, Hunter, Joyce, and Kertzner, Robert M. Project STRIDE: Stress, Identity, and Mental Health, New York City, 2004-2005. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2018-11-28. https://doi.org/10.3886/ICPSR35525.v2

Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychological methods, 18(2), 137.

Yiu-Fai Yung, Michael Lamm, and Wei Zhang, SAS Institute Inc. (2018). “Causal Mediation Analysis with the CAUSALMED Procedure.” In Proceedings of the SAS Global Forum 2018 Conference. Cary, NC: SAS Institute Inc. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1991-2018.pdf

SAS Institute Inc. 2017. SAS/STAT® 14.3 User’s Guide. Cary, NC: SAS Institute Inc. “SAS/STAT® 14.3 User’s Guide The CAUSALMED Procedure”

Zellner, D.A., Saito, S., & Gonzalez, J. (2007). The effect of stress on men's food selection. Appetite, 49, 696-699.

Zellner, D. A., Loaiza, S., Gonzalez, Z., Pita, J., Morales, J., Pecora, D., & Wolf, A. (2006). Food selection changes under stress. Physiology & behavior, 87(4), 789-793.

ACKNOWLEDGMENTS

Jessica Parker – Spectrum Health

Bioinformatics Core Team – Spectrum Health

RECOMMENDED READING

Hong, G. (2015). Causality in a Social World: Moderation, Mediation, and Spill-Over. New York: John Wiley & Sons.

Keele, L. (2015). Causal Mediation Analysis. American Journal of Evaluation, 36(4), 500-513. http://doi:10.1177/1098214015594689

Lamm, M., and Yung, Y.-F. (2017). “Estimating Causal Effects from Observational Data with the CAUSALTRT Procedure.” In Proceedings of the SAS Global Forum 2017

Conference. Cary, NC: SAS Institute Inc. http://support.sas.com/resources/papers/proceedings17/SAS0374-2017.pdf

Pearl, J. (2014). Interpretation and identification of causal mediation. Psychological Methods, 19(4), 459-481. http://dx.doi.org/10.1037/a0036434

VanderWeele, T. J., & Vansteelandt, S. (2014). Mediation Analysis with Multiple Mediators. Epidemiologic Methods, 2(1), 95–115. http://doi.org/10.1515/em-2012-0010

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Daniel F. Muzyka
Grand Valley State University
Spectrum Health Office of Research
muzykad@mail.gvsu.edu

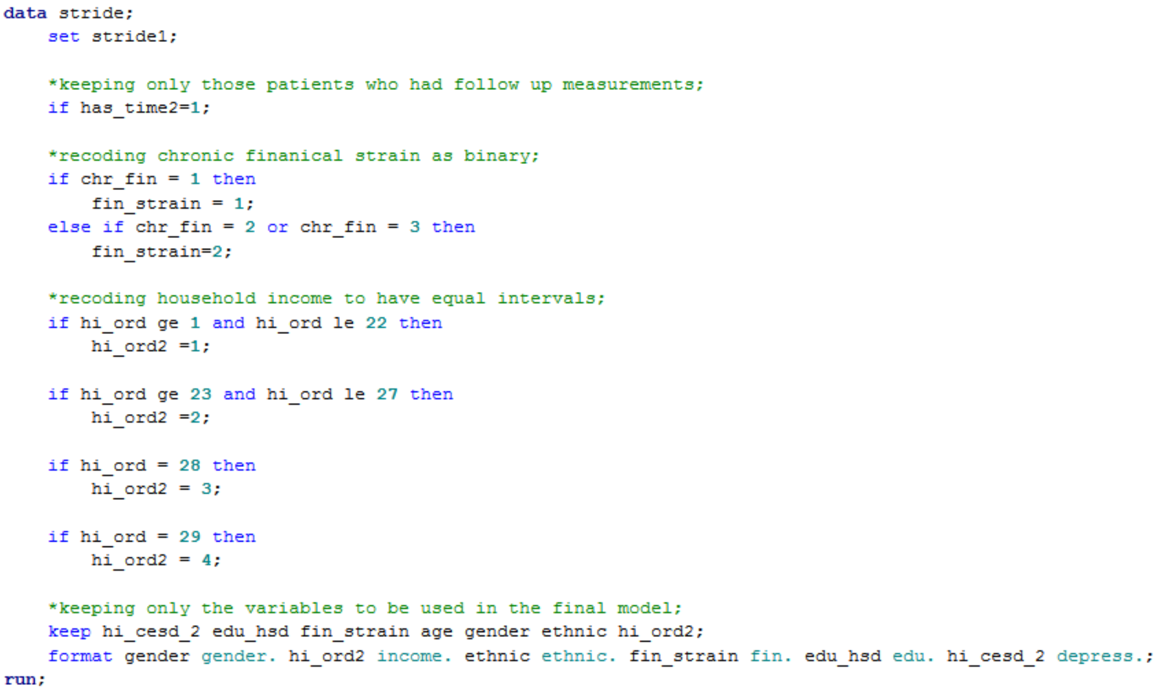
Matthew M. Lypka
Spectrum Health Office of Research
Matthew.Lypka@spectrumhealth.org

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

APPENDIX

Appendix Item 1. Code Used in Data Preparation



```
data stride;
    set stride1;

    *keeping only those patients who had follow up measurements;
    if has_time2=1;

    *recoding chronic finanical strain as binary;
    if chr_fin = 1 then
        fin_strain = 1;
    else if chr_fin = 2 or chr_fin = 3 then
        fin_strain=2;

    *recoding household income to have equal intervals;
    if hi_ord ge 1 and hi_ord le 22 then
        hi_ord2 =1;

    if hi_ord ge 23 and hi_ord le 27 then
        hi_ord2 =2;

    if hi_ord = 28 then
        hi_ord2 = 3;

    if hi_ord = 29 then
        hi_ord2 = 4;

    *keeping only the variables to be used in the final model;
    keep hi_cesd_2 edu_hsd fin_strain age gender ethnic hi_ord2;
    format gender gender. hi_ord2 income. ethnic ethnic. fin_strain fin. edu_hsd edu. hi_cesd_2 depress.;
run;
```

Appendix Item 2. Frequencies of Ethnicity by high school education

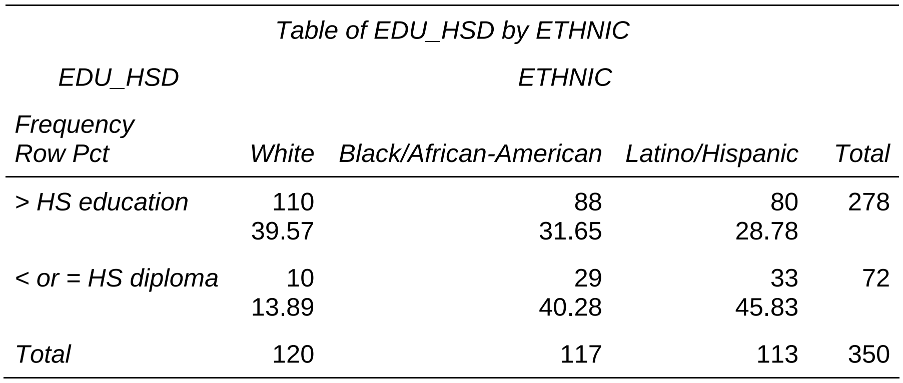


Table of EDU_HSD by ETHNIC

EDU_HSD	*ETHNIC*			
Frequency *Row Pct*	*White*	*Black/African-American*	*Latino/Hispanic*	*Total*
> HS education	110 39.57	88 31.65	80 28.78	278
< or = HS diploma	10 13.89	29 40.28	33 45.83	72
Total	120	117	113	350

Appendix Item 3. Frequencies of Financial Strain by high school education

Table of EDU_HSD by fin_strain

EDU_HSD Frequency Row Pct	fin_strain Not financially strained	At least somewhat financially strained	Total
> HS education	114 41.01	164 58.99	278
< or = HS diploma	17 23.61	55 76.39	72
Total	131	219	350

Appendix Item 4. Frequencies of Gender by high school education

Table of EDU_HSD by GENDER

EDU_HSD Frequency Row Pct	GENDER Male	Female	Total
> HS education	140 50.36	138 49.64	278
< or = HS diploma	38 52.78	34 47.22	72
Total	178	172	350

Appendix Item 5. Frequencies of Depression by high school education

Table of EDU_HSD by HI_CESD_2

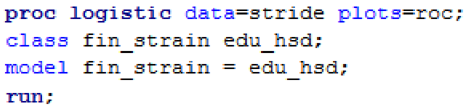
EDU_HSD Frequency Row Pct	HI_CESD_2 Low CES-D score	High CES-D score	Total
> HS education	184 66.19	94 33.81	278
< or = HS diploma	33 45.83	39 54.17	72
Total	217	133	350

Appendix Item 6. Frequencies of income levels by high school education

Table of EDU_HSD by hi_ord2

EDU_HSD		*hi_ord2*			
Frequency *Row Pct*	*<$25k*	*$25k to < $50k*	*$50k to < $75k*	*$75k to < $100k*	*Total*
> HS education	86 30.94	155 55.76	21 7.55	16 5.76	278
< or = HS diploma	48 66.67	19 26.39	3 4.17	2 2.78	72
Total	134	174	24	18	350

Appendix Item 7. Code for regressing financial strain on high school education



```
proc logistic data=stride plots=roc;
class fin_strain edu_hsd;
model fin_strain = edu_hsd;
run;
```

Appendix Item 8. Response profile for Depression

Response Profile

Ordered Value	*HI_CESD_2*	*Total Frequency*
1	High CES-D score	133
2	Low CES-D score	217

Outcome probability modeled is HI_CESD_2='High CES-D score'.

Appendix Item 9. Response profile for financial strain mediator

Mediator Profile

Ordered Value	*fin_strain*	*Total Frequency*
1	At least somewhat financially strained	219
2	Not financially strained	131

Mediator probability modeled is fin_strain='At least somewhat financially strained'.

Appendix Item 10. Profile for the treatment variable high school education

Treatment Profile

Ordered Value	*EDU_HSD*	*Total Frequency*
1	< or = HS diploma	72
2	> HS education	278