"Third factors" are variables that can affect the estimated relationship between outcome and exposure by being ignored. In the parlance of regression methodology, they are "omitted variables". In health analytics, they are "confounders", in which all of the relationship between outcome and exposure is due to their individual relationship with this third variable, or are "modifiers", in which the strength of the relationship between outcome and exposure depends on this third variable.

Using the dataset "Confounder Example.csv", let us demonstrate how these confounding effects can manifest themselves and how they can be addressed at the analytical stage.

The dataset contains 1,000 patients and their diagnosis for cardiovascular disease (CVD) (1 = yes; 0 = no). In addition, the dataset contains the patient characteristics BMI (1 = obese; 0 = not obese) and AGE_GT50 (1 = yes; 0 = no).

```
proc import datafile='/…/Confounding Example.csv' out=heart replace; run;

proc format;
      value Hd_DiagFmt          1="Positive"
                                0="Negative";
      value AgeFmt              1="Age >=50"
                                0="Age < 50";
      value BMIFmt              1="Obese"
                                0="Not Obese"; run;
```

1. **Using PROC FREQ**
   a. **Determine the OR of a patient with an obese BMI being diagnosed with CVD.**

```
proc sort data=heart; by descending cvd descending bmi; run;
proc freq data=heart order=data;
      format bmi bmifmt.  cvd hd_diagfmt.;
      tables bmi*cvd / chisq relrisk nocol norow nopercent; run;
```

The FREQ Procedure

Frequency

Table of BMI by CVD			
	CVD		
BMI	Positive	Negative	Total
Obese	46	254	300
Not Obese	60	640	700
Total	106	894	1000

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	1.9318	1.2811	2.9128
Relative Risk (Column 1)	1.7889	1.2487	2.5628
Relative Risk (Column 2)	0.9260	0.8780	0.9767

We see that the odds of a patient with obese BMI being diagnosed with CVD are 1.9x higher (as compared to not obese BMI). Moreover, the 95% CI does not include 1.0 so we conclude the estimated relationship is statistically significant.

   b. **Determine the OR of a patient with an obese BMI being diagnosed with CVD by AGE strata.**

```
proc freq data=heart order=data;
      format bmi bmifmt.  cvd hd_diagfmt.;
      tables age_gt50*bmi*cvd / chisq relrisk nocol norow nopercent;
run;
```

AGE_GT50 = 0

Frequency	Table 1 of BMI by CVD		
	Controlling for AGE_GT50=0		
	CVD		
BMI	Positive	Negative	Total
Obese	10	90	100
Not Obese	35	465	500
Total	45	555	600

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	1.4762	0.7056	3.0882
Relative Risk (Column 1)	1.4286	0.7316	2.7895
Relative Risk (Column 2)	0.9677	0.9027	1.0375

AGE_GT50 = 1

Frequency	Table 2 of BMI by CVD		
	Controlling for AGE_GT50=1		
	CVD		
BMI	Positive	Negative	Total
Obese	36	164	200
Not Obese	25	175	200
Total	61	339	400

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	1.5366	0.8839	2.6711
Relative Risk (Column 1)	1.4400	0.8990	2.3066
Relative Risk (Column 2)	0.9371	0.8621	1.0187

c. Based on your findings, does AGE seem to matter? Explain.

We see that the OR between BMI and CVD differs substantially when we account for AGE:

```
If >=50, OR for obese BMI being diagnosed with CVD = 1.537
If <50, OR for obese BMI being diagnosed with CVD = 1.476

These values differ from the overall OR of 1.932 => AGE may be a
confounding variable.
```

2. Test for whether AGE is a confounding variable.

a. First, does AGE meet these 2 criteria for being a confounder? Support statistically where possible.

i. Is AGE related to CVD?

```
proc sort data=heart; by bmi; run;
proc corr data=heart; var age_gt50 cvd; by BMI; run;
```

Pearson Correlation Coefficients, N = 700 Prob > |r| under H0: Rho=0		
	AGE_GT50	CVD
AGE_GT50	1.00000	0.08876 0.0188
CVD	0.08876 0.0188	1.00000

Appears to be a positive relationship between AGE and CVD but the strength of the correlation is somewhat weak.

ii. Is AGE related to BMI?

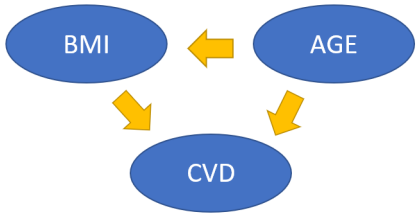
```
proc corr data=heart; var age_gt50 bmi; run;
```

Pearson Correlation Coefficients, N = 1000 Prob > |r| under H0: Rho=0		
	AGE_GT50	BMI
AGE_GT50	1.00000	0.35635 <.0001
BMI	0.35635 <.0001	1.00000

Appears to be a positive relationship between AGE and BMI with a moderately strong correlation.

Note: Another test is whether AGE is an intermediate step in the casual relationship between BMI and CVD?

Theoretically, it does not appear that path from BMI to CVD is through AGE (i.e., BMI causes AGE which cause CVD). Rather, the evidence suggests that (i) both AGE and BMI affect CVD; and (ii) AGE affects BMI. Hence, AGE appears to be a confounding variable to the relationship between BMI and CVD:



b. Second, using logistic regression, test for whether AGE is a confounder.

```
/* run logistic model using full model and capture OR on BMI */

proc logistic data=heart descending;
      class cvd (ref='0') bmi (ref='0') age_gt50 (ref='0') / param=ref;
      model cvd = bmi age_gt50 / rsq; run;

      /* Adjusted OR is 1.515 */

/* run logistic model removing AGE and capture OR on BMI */

proc logistic data=heart descending;
      class cvd (ref='0') bmi (ref='0') / param=ref;
      model cvd = bmi / rsq; run;

      /* Unadjusted OR is 1.932 */

/* calculate the % change

      (1.932 - 1.515) / 1.932 = .2158 or 21.58%
      (1.932 - 1.515) / 1.515 = .2752 or 27.52%

      Since both > 10%, we conclude that AGE is a confounding variable.

*/
```

3. **Calculate the "adjusted OR" for BMI and CVD that accounts for AGE.**
   a. **First, use the contingency table approach.**

We have the data from the frequency tables in #1 above to populate the following formatted tables:

Age < 50

	CVD	No CVD	Total
Obese	10	90	100
Not Obese	35	465	500
Total	45	555	600

The unadjusted OR of an obese subject age < 50 having CVD is 1.476

Age >= 50

	CVD	No CVD	Total
Obese	36	164	200
Not Obese	25	175	200
Total	61	339	400

The unadjusted OR of an obese subject age < 50 having CVD is 1.537

The adjusted OR is effectively a weighted average of the OR from the two AGE strata. The Cochran-Mantel-Haenzel estimate is such an average:

$$\widehat{OR}_{cmh} = \frac{\sum \frac{a_i d_i}{n_i}}{\sum \frac{b_i c_i}{n_i}} = \frac{\frac{10(465)}{600} + \frac{36(175)}{400}}{\frac{90(35)}{600} + \frac{164(25)}{400}} = \frac{7.75+15.75}{5.25+10.25} = 1.52$$

   b. **Second, use PROC FREQ with the CMH option.**

```
proc freq data=work.heart order=data;
      format bmi bmifmt.  cvd hd_diagfmt.;
      tables age_gt50*bmi*cvd / cmh;
run;
```

Common Odds Ratio and Relative Risks				
Statistic	Method	Value	95% Confidence Limits	
Odds Ratio	Mantel-Haenszel	1.5161	0.9739	2.3602
	Logit	1.5146	0.9730	2.3577
Relative Risk (Column 1)	Mantel-Haenszel	1.4364	0.9770	2.1117
	Logit	1.4362	0.9770	2.1111
Relative Risk (Column 2)	Mantel-Haenszel	0.9515	0.9007	1.0052
	Logit	0.9551	0.9054	1.0075

Note that the CMH option also produces the Cochran-Mantel-Haenszel "general association" test statistic which considers the presence of a control variable in determining whether there is an association between CVD and BMI. As shown in the output below, we see that the test statistic is not significant at the 5% level of significance. So there appears to be no association between BMI and CVD after controlling for AGE.

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	3.4138	0.0647
2	Row Mean Scores Differ	1	3.4138	0.0647
3	General Association	1	3.4138	0.0647

c. Third, use logistic regression.

```
proc logistic data=heart descending;
      class cvd (ref='0') bmi (ref='0') age_gt50 (ref='0') / param=ref;
      model cvd = bmi age_gt50 / rsq; run;
```

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
BMI 1 vs 0	1.515	0.974	2.355
AGE_GT50 1 vs 0	1.924	1.243	2.980

So, we see that logistic regression yields the same adjusted OR (1.52) as the CMH/contingency table approach.

Note that the estimated model coefficients (β) are consistent with the CMH test performed in part (c). The β on BMI is not statistically different from 0 at the 5% significance level (95% confidence level). So, once we control for AGE, the apparent relationship between BMI and CVD disappears in this sample.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.5923	0.1629	253.2787	<.0001
BMI	1	1	0.4151	0.2252	3.3980	0.0653
AGE_GT50	1	1	0.6547	0.2232	8.6066	0.0033

4. Is AGE an "effect modifier"? Explain with the help of relevant SAS output.

Above we found that AGE is a confounder of the relationship between BMI and CVD. A related question is whether the relationship between BMI and CVD differs across the strata of AGE. If the strata OR are homogeneous (i.e., are roughly the same), then there likely is no interaction between AGE and BMI and AGE is not an effect modifier.

The Breslow-Day test reveals that we cannot reject the H_0 of homogeneity. The CMH option for PROC FREQ also produces this test statistic. As shown, the p-value is quite large so we cannot reject the H_0.

Breslow-Day Test for Homogeneity of Odds Ratios	
Chi-Square	0.0073
DF	1
Pr > ChiSq	0.9320

We can also use PROC LOGISTIC to generate the deviance chi squared statistic which indicates we cannot reject the H0 that the coefficients on the interaction terms in a saturated model are jointly = 0.

```
proc logistic data=work.heart descending;
        class cvd (ref='0') bmi (ref='0')
        age_gt50 (ref='0') / param=ref;
        model cvd = bmi age_gt50 /
aggregate scale=none; run;
```

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	0.0073	1	0.0073	0.9320
Pearson	0.0073	1	0.0073	0.9321

Number of unique profiles: 4

And we can estimate a fully saturated model as well which shows the interaction term as statistically insignificant:

```
proc logistic data=work.heart descending;
        class cvd (ref='0') bmi (ref='0') age_gt50 (ref='0') / param=ref;
        model cvd = bmi age_gt50 bmi*age_gt50/ rsq;
run;
```

Analysis of Maximum Likelihood Estimates							
Parameter			DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept			1	-2.5867	0.1753	217.7908	<.0001
BMI	1		1	0.3895	0.3766	1.0694	0.3011
AGE_GT50	1		1	0.6408	0.2765	5.3717	0.0205
BMI*AGE_GT50	1	1	1	0.0401	0.4706	0.0073	0.9321

So, we conclude that AGE is a confounder but not an effect modifier.