**Question 2:**  The following data provide the survival among cancer patients.  The data provides comorbidities for patients who have or don't have lung cancer.

1. Using SQL, group the diagnoses into commonly occurring strata, report the common odds ratio across strata (10 points).
2. Calculate overlap in the data (5 points)
3. Redo analysis and this time use synthetic controls to improve overlap (10 points)

Provide both the SQL code and your answers to questions 2a through 2d. Use the following [Data►](http://openonlinecourses.com/causalanalysis/ComorbidLungCancer41.zip)

Ans:-

1. And b)

From R

> dataexamq2=read.csv('ComorbidLungCancer41.csv')

> colnames(dataexamq2)

[1] "ï..ID" "Dead" "Survival" "Cancer""I401.9""I496." "I272.4"

[8] "I305.1""I486." "I530.81" "I414.01" "I285.9""I427.31" "I600.00"

[15] "I311." "I491.21" "I276.1""I428.0""I276.51" "I276.8""I599.0"

[22] "I403.90" "IE849.7" "I309.81" "I585.9""I300.00" "I414.00" "I443.9"

[29] "I244.9""I724.2""IV58.61" "I250.00" "I427.89" "I788.20" "I280.9"

[36] "I786.6""I518.89" "I786.59" "I787.91" "IV45.81" "IE849.0" "I070.54"

[43] "I303.90" "I287.5""IV45.82"

> modelexamq1=glm(dataexamq2[,2]~., data=dataexamq2[,5:45], family=binomial)

> summary(modelexamq1)

Call:

glm(formula = dataexamq2[, 2] ~ ., family = binomial, data = dataexamq2[,

5:45])

Deviance Residuals:

Min 1QMedian 3QMax

-3.1046 -0.5532 -0.4778 -0.35712.9484

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.0246930.005479 -369.550 < 2e-16 \*\*\*

I401.9 0.1175440.00695816.892 < 2e-16 \*\*\*

I496. 0.6343350.00843875.180 < 2e-16 \*\*\*

I272.4-0.2793410.007462 -37.434 < 2e-16 \*\*\*

I305.1-0.2054970.008300 -24.759 < 2e-16 \*\*\*

I486. 0.7851180.01108870.809 < 2e-16 \*\*\*

I530.81 -0.1378760.008247 -16.718 < 2e-16 \*\*\*

I414.010.1466840.00958515.304 < 2e-16 \*\*\*

I285.9 0.4700510.00979447.992 < 2e-16 \*\*\*

I427.310.5208820.01024150.862 < 2e-16 \*\*\*

I600.000.1233660.00982712.554 < 2e-16 \*\*\*

I311. -0.0406180.009834-4.130 3.62e-05 \*\*\*

I491.210.5404150.01360239.732 < 2e-16 \*\*\*

I276.1 0.5237840.01264041.437 < 2e-16 \*\*\*

I428.0 0.6624080.01084161.101 < 2e-16 \*\*\*

I276.510.3919540.01334229.377 < 2e-16 \*\*\*

I276.8 0.0689840.012637 5.459 4.80e-08 \*\*\*

I599.0 0.7364750.01081068.130 < 2e-16 \*\*\*

I403.900.2649020.01653916.017 < 2e-16 \*\*\*

IE849.70.0791700.013948 5.676 1.38e-08 \*\*\*

I309.81 -0.5341200.010967 -48.701 < 2e-16 \*\*\*

I585.9 0.3146050.01763417.841 < 2e-16 \*\*\*

I300.00 -0.1680640.014002 -12.003 < 2e-16 \*\*\*

I414.000.3751820.01857520.198 < 2e-16 \*\*\*

I443.9 0.3323450.01522621.828 < 2e-16 \*\*\*

I244.9 0.2692220.01163523.138 < 2e-16 \*\*\*

I724.2-0.2837840.012666 -22.406 < 2e-16 \*\*\*

IV58.61 -0.2045370.014028 -14.580 < 2e-16 \*\*\*

I250.00 -0.2813380.012080 -23.289 < 2e-16 \*\*\*

I427.89 -0.1307480.014193-9.212 < 2e-16 \*\*\*

I788.200.1265610.015434 8.200 2.40e-16 \*\*\*

I280.9 0.2211760.01535114.408 < 2e-16 \*\*\*

I786.6 1.0964360.03175134.532 < 2e-16 \*\*\*

I518.890.4550020.02220020.496 < 2e-16 \*\*\*

I786.59 -0.6773220.015322 -44.206 < 2e-16 \*\*\*

I787.910.1045990.016733 6.251 4.07e-10 \*\*\*

IV45.810.0128000.020486 0.625 0.532

IE849.00.0231990.014446 1.606 0.108

I070.540.4318040.01408630.654 < 2e-16 \*\*\*

I303.90 -0.1203240.013476-8.929 < 2e-16 \*\*\*

I287.5 0.5792060.01552537.308 < 2e-16 \*\*\*

IV45.82 -0.3541470.017506 -20.230 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 760697 on 829798 degrees of freedom

Residual deviance: 660280 on 829757 degrees of freedom

AIC: 660364

Number of Fisher Scoring iterations: 5

/\*\*\*\*\*\* Script for SelectTopNRows command from SSMS \*\*\*\*\*\*/

drop table #Cases

SELECT COUNT(DISTINCT ID) AS nCases

,SUM(IIF([Dead] = 1, 1., 0.)) as a

,SUM(IIF([Dead] = 0,1.,0.)) as b

,[I4019],[I496],[I2724],[I3051],[I486],[I53081],[I41401],[I2859],[I42731],[I60000],[I311],[I49121],[I2761],[I4280],[I27651],[I2768],[I5990],[I40390],[IE8497],[I30981],[I5859],[I30000],[I41400],[I4439],[I2449],[I7242],[IV5861],[I25000],[I42789],[I78820],[I2809],[I7866],[I51889],[I78659],[I78791],[I07054],[I30390],[I2875],[IV4582]

---- except IV4581 and IE849 WHICH IS NOT SIGNIFICANT

INTO #Cases from ComorbidLungCancer41

where Cancer = 1

group by

[I4019],[I496],[I2724],[I3051],[I486],[I53081],[I41401],[I2859],[I42731],[I60000],[I311],[I49121],[I2761],[I4280],[I27651],[I2768],[I5990],[I40390],[IE8497],[I30981],[I5859],[I30000],[I41400],[I4439],[I2449],[I7242],[IV5861],[I25000],[I42789],[I78820],[I2809],[I7866],[I51889],[I78659],[I78791],[I07054],[I30390],[I2875],[IV4582]

--- (10806 rows affected)

--- FOR CONTROLS

drop table #Controls

SELECT COUNT(DISTINCT ID ) AS nControls

,SUM(IIF([Dead] = 1, 1., 0.)) as c

,SUM(IIF([Dead] = 0, 1., 0.)) as d

, [I401 9]

,[I496],[I2724],[I3051],[I486],[I53081],[I41401],[I2859],[I42731],[I60000],[I311],[I49121],[I2761],[I4280],[I27651],[I2768],[I5990],[I40390],[IE8497],[I30981],[I5859],[I30000],[I41400],[I4439],[I2449],[I7242],[IV5861],[I25000],[I42789],[I78820],[I2809],[I7866],[I51889],[I78659],[I78791],[I07054],[I30390],[I2875],[IV4582]

---- except IV4581 and IE849 WHICH ARE NOT SIGNIFICANT

INTO #Controls from ComorbidLungCancer41

where Cancer = 0

group by

[I4019],[I496],[I2724],[I3051],[I486],[I53081],[I41401],[I2859],[I42731],[I60000],[I311],[I49121],[I2761],[I4280],[I27651],[I2768],[I5990],[I40390],[IE8497],[I30981],[I5859],[I30000],[I41400],[I4439],[I2449],[I7242],[IV5861],[I25000],[I42789],[I78820],[I2809],[I7866],[I51889],[I78659],[I78791],[I07054],[I30390],[I2875],[IV4582]

---(186748 rows affected)

-- COMMON ODD RATIO

drop table #COR

SELECT SUM(a\*d/(a+b+c+d))/sum(b\*c/(a+b+c+d)) AS CommonOddsRatio

into #COR

FROM #Cases inner join #Controls on

#Cases. [I401 9] = #Controls. [I401 9] AND

#Cases.[I496 ] =#Controls.[I496 ] AND

#Cases. [I272 4] =#Controls. [I272 4]AND

#Cases. [I305 1] =#Controls. [I305 1]AND

#Cases. [I486 ] =#Controls. [I486 ]AND

#Cases. [I530 81] =#Controls. [I530 81]AND

#Cases. [I414 01] =#Controls. [I414 01]AND

#Cases. [I285 9] =#Controls. [I285 9]AND

#Cases. [I427 31] =#Controls. [I427 31]AND

#Cases. [I600 00] =#Controls. [I600 00]AND

#Cases. [I311 ] =#Controls. [I311 ]AND

#Cases. [I491 21] =#Controls. [I491 21]AND

#Cases. [I276 1] =#Controls. [I276 1]AND

#Cases. [I428 0] =#Controls. [I428 0]AND

#Cases. [I276 51] =#Controls. [I276 51]AND

#Cases. [I276 8] =#Controls. [I276 8]AND

#Cases. [I599 0] =#Controls. [I599 0]AND

#Cases. [I403 90] =#Controls. [I403 90]AND

#Cases. [I309 81] =#Controls. [I309 81]AND

#Cases. [I585 9] =#Controls. [I585 9]AND

#Cases. [I300 00] =#Controls. [I300 00]AND

#Cases. [I414 00] =#Controls. [I414 00]AND

#Cases. [I443 9] =#Controls. [I443 9]AND

#Cases. [I244 9] =#Controls. [I244 9]AND

#Cases. [I724 2] =#Controls. [I724 2]AND

#Cases. [IV58 61] =#Controls. [IV58 61]AND

#Cases. [I250 00] =#Controls. [I250 00]AND

#Cases. [I427 89] =#Controls. [I427 89]AND

#Cases. [I788 20] =#Controls. [I788 20]AND

#Cases. [I280 9] =#Controls. [I280 9]AND

#Cases. [I786 6] =#Controls. [I786 6]AND

#Cases. [I518 89] =#Controls. [I518 89]AND

#Cases. [I786 59] =#Controls. [I786 59]AND

#Cases. [I787 91] =#Controls. [I787 91]AND

#Cases. [I070 54] =#Controls. [I070 54]AND

#Cases. [I303 90] =#Controls. [I303 90]AND

#Cases. [I287 5] =#Controls. [I287 5]AND

#Cases. [IV45 82] =#Controls. [IV45 82]

SELECT \* FROM #COR -- CommonOddsRatio :- 8.947610

-- for calculation of overlap

DROP TABLE #M

select a,b,c,d into #M

from #Cases join #Controls on

#Cases.[I401 9] = #Controls.[I401 9] AND

#Cases.[I496 ] =#Controls.[I496 ] AND

#Cases.[I272 4] = #Controls.[I272 4] AND

#Cases.[I305 1] = #Controls.[I305 1] AND

#Cases.[I486 ] = #Controls.[I486 ] AND

#Cases.[I530 81] = #Controls.[I530 81] AND

#Cases.[I414 01] = #Controls.[I414 01] AND

#Cases.[I285 9] = #Controls.[I285 9] AND

#Cases.[I427 31] = #Controls.[I427 31] AND

#Cases.[I600 00] = #Controls.[I600 00] AND

#Cases.[I311 ] = #Controls.[I311 ] AND

#Cases.[I491 21] = #Controls.[I491 21] AND

#Cases.[I276 1] = #Controls.[I276 1] AND

#Cases.[I428 0] = #Controls.[I428 0] AND

#Cases.[I276 51] = #Controls.[I276 51] AND

#Cases.[I276 8] = #Controls.[I276 8] AND

#Cases.[I599 0] = #Controls.[I599 0] AND

#Cases.[I403 90] = #Controls.[I403 90] AND

#Cases.[IE849 7] = #Controls.[IE849 7] AND

#Cases.[I309 81] = #Controls.[I309 81] AND

#Cases.[I585 9] = #Controls.[I585 9] AND

#Cases.[I300 00] = #Controls.[I300 00] AND

#Cases.[I414 00] = #Controls.[I414 00] AND

#Cases.[I443 9] = #Controls.[I443 9] AND

#Cases.[I244 9] = #Controls.[I244 9] AND

#Cases.[I724 2] = #Controls.[I724 2] AND

#Cases.[IV58 61] = #Controls.[IV58 61] AND

#Cases.[I250 00] = #Controls.[I250 00] AND

#Cases.[I427 89] = #Controls.[I427 89] AND

#Cases.[I788 20] = #Controls.[I788 20] AND

#Cases.[I280 9] = #Controls.[I280 9] AND

#Cases.[I786 6] = #Controls.[I786 6] AND

#Cases.[I518 89] = #Controls.[I518 89] AND

#Cases.[I786 59] = #Controls.[I786 59] AND

#Cases.[I787 91] = #Controls.[I787 91] AND

#Cases.[I070 54] = #Controls.[I070 54] AND

#Cases.[I303 90] = #Controls.[I303 90] AND

#Cases.[I287 5] = #Controls.[I287 5] AND

#Cases.[IV45 82] = #Controls.[IV45 82]

--- (5108 rows affected)

Declare @TotalCases float

set @TotalCases = (select sum(nCases) from #Cases)

select ROUND(sum(a+b) \* 100/@TotalCases,2) as PercentOverlap

from #M /\*PercentOverlap 70.65 \*/

C)

---NOW BY SYNTHETIC CONTROL

SELECT [ID]

,[Dead]

,[Survival]

,[Cancer]

,CAST([I401 9] AS int ) AS I4019

,CAST([I496 ] AS int ) AS I496

,CAST([I272 4] AS int ) AS I2724

,CAST([I305 1] AS int ) AS I3051

,CAST([I486 ] AS int ) AS I486

,CAST([I530 81] AS int ) AS I53081

,CAST([I414 01] AS int ) AS I41401

,CAST([I285 9] AS int ) AS I2859

,CAST([I427 31] AS int ) AS I42731

,CAST([I600 00] AS int ) AS I60000

,CAST([I311 ] AS int ) AS I311

,CAST([I491 21] AS int ) AS I49121

,CAST([I276 1] AS int ) AS I2761

,CAST([I428 0] AS int ) AS I4280

,CAST([I276 51] AS int ) AS I27651

,CAST([I276 8] AS int ) AS I2768

,CAST([I599 0] AS int ) AS I5990

,CAST([I403 90] AS int ) AS I40390

,CAST([IE849 7] AS int ) AS IE8497

,CAST([I309 81]AS int ) AS I30981

,CAST([I585 9] AS int ) AS I5859

,CAST([I300 00]AS int ) AS I30000

,CAST([I414 00] AS int ) AS I41400

,CAST([I443 9] AS int ) AS I4439

,CAST([I244 9] AS int ) AS I2449

,CAST([I724 2] AS int ) AS I7242

,CAST([IV58 61] AS int ) AS IV5861

,CAST([I250 00] AS int ) AS I25000

,CAST([I427 89] AS int ) AS I42789

,CAST([I788 20] AS int ) AS I78820

,CAST([I280 9] AS int ) AS I2809

,CAST([I786 6] AS int ) AS I7866

,CAST([I518 89] AS int ) AS I51889

,CAST([I786 59] AS int ) AS I78659

,CAST([I787 91] AS int ) AS I78791

,CAST([IV45 81] AS int ) AS IV4581

,CAST([IE849 0] AS int ) AS IE8490

,CAST([I070 54] AS int ) AS I07054

,CAST([I303 90] AS int ) AS I30390

,CAST([I287 5] AS int ) AS I2875

,CAST([IV45 82] AS int ) AS IV4582

INTO #DataQ2

FROM [HAP 823 EXAM].[dbo].[ComorbidLungCancer41]

SELECT \* into examq2conrol FROM #DataQ2

where Cancer=0 -- (810174 rows affected)

Exported data examq2control where cancer is 0 to R and found out the coefficient from there for all the variables

> dataexamq2controls=read.csv('examq2controls.csv')

> colnames(dataexamq2controls)

[1] "ID" "Dead" "Survival" "Cancer""I4019" "I496" "I2724"

[8] "I3051" "I486" "I53081""I41401""I2859" "I42731""I60000"

[15] "I311" "I49121""I2761" "I4280" "I27651""I2768" "I5990"

[22] "I40390""IE8497""I30981""I5859" "I30000""I41400""I4439"

[29] "I2449" "I7242" "IV5861""I25000""I42789""I78820""I2809"

[36] "I7866" "I51889""I78659""I78791""IV4581""IE8490""I07054"

[43] "I30390""I2875" "IV4582"

> modelexamq2controls=glm(dataexamq2controls[,2]~., data=dataexamq2controls[,5:45], family=binomial)

> summary(modelexamq2controls)

Call:

glm(formula = dataexamq2controls[, 2] ~ ., family = binomial,

data = dataexamq2controls[, 5:45])

Deviance Residuals:

Min 1QMedian 3QMax

-3.0443 -0.5320 -0.4642 -0.34852.9322

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.1225290.005714 -371.476 < 2e-16 \*\*\*

I4019 0.1244410.00725617.151 < 2e-16 \*\*\*

I4960.5445600.00897760.658 < 2e-16 \*\*\*

I2724 -0.2805130.007771 -36.099 < 2e-16 \*\*\*

I3051 -0.2449720.008754 -27.986 < 2e-16 \*\*\*

I4860.7257160.01166562.212 < 2e-16 \*\*\*

I53081-0.1293740.008564 -15.107 < 2e-16 \*\*\*

I41401 0.1552780.00993715.626 < 2e-16 \*\*\*

I2859 0.4802140.01010247.536 < 2e-16 \*\*\*

I42731 0.5263370.01061449.587 < 2e-16 \*\*\*

I60000 0.1302060.01016712.807 < 2e-16 \*\*\*

I311 -0.0144090.010148-1.420 0.15564

I49121 0.5631410.01415139.794 < 2e-16 \*\*\*

I2761 0.4993850.01315537.962 < 2e-16 \*\*\*

I4280 0.7147830.01109864.404 < 2e-16 \*\*\*

I27651 0.3901010.01375128.370 < 2e-16 \*\*\*

I2768 0.0779290.013033 5.979 2.24e-09 \*\*\*

I5990 0.7826240.01102570.987 < 2e-16 \*\*\*

I40390 0.2807340.01698216.531 < 2e-16 \*\*\*

IE8497 0.0904930.014456 6.260 3.85e-10 \*\*\*

I30981-0.5197610.011363 -45.743 < 2e-16 \*\*\*

I5859 0.3295070.01807218.233 < 2e-16 \*\*\*

I30000-0.1693780.014526 -11.661 < 2e-16 \*\*\*

I41400 0.3791380.01916819.780 < 2e-16 \*\*\*

I4439 0.3259510.01587220.536 < 2e-16 \*\*\*

I2449 0.2906150.01196524.289 < 2e-16 \*\*\*

I7242 -0.2714030.013100 -20.717 < 2e-16 \*\*\*

IV5861-0.1925840.014427 -13.349 < 2e-16 \*\*\*

I25000-0.2685570.012472 -21.532 < 2e-16 \*\*\*

I42789-0.1283690.014649-8.763 < 2e-16 \*\*\*

I78820 0.1422080.015879 8.956 < 2e-16 \*\*\*

I2809 0.2257260.01577414.310 < 2e-16 \*\*\*

I7866 0.9763770.03572227.333 < 2e-16 \*\*\*

I51889 0.4347020.02349418.503 < 2e-16 \*\*\*

I78659-0.6690000.015834 -42.252 < 2e-16 \*\*\*

I78791 0.1119070.017206 6.504 7.83e-11 \*\*\*

IV4581 0.0302700.021099 1.435 0.15139

IE8490 0.0456900.014810 3.085 0.00204 \*\*

I07054 0.4652470.01448532.120 < 2e-16 \*\*\*

I30390-0.0650500.013831-4.703 2.56e-06 \*\*\*

I2875 0.6003640.01586337.847 < 2e-16 \*\*\*

IV4582-0.3616890.018137 -19.942 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 711031 on 810173 degrees of freedom

Residual deviance: 616138 on 810132 degrees of freedom

AIC: 616222

Number of Fisher Scoring iterations: 5

-- Add regression predicted values to the data

Select \*

, exp(-2.122529+I4019\*0.124441+I496\*0.544560+I2724\*-0.280513+I3051\*-0.244972+I486\*0.725716+I53081\*-0.129374+I41401\*0.155278+I2859\*0.480214

+I42731\*0.526337+I60000\*0.130206+I311\*-0.014409+I49121\*0.563141+I2761\*0.499385

+I4280\*0.714783+I27651\*0.390101+I2768\*0.077929

+I5990\*0.782624+I40390\*0.280734+IE8497\*0.090493+I30981\*-0.519761

+I5859\*0.329507+I30000\*-0.169378+I41400\*0.379138+I4439\*0.325951

+I2449\*0.290615+I7242\*-0.271403+IV5861\*-0.192584+I25000\*-0.268557+

I42789\*-0.128369+I78820\*0.142208+I2809\*0.225726+I7866\*0.976377

+I51889\*0.434702+I78659\*-0.669000+I78791\*0.111907+IV4581\*0.030270+IE8490\*0.045690

+I07054\*0.465247+I30390\*-0.065050+I2875\*0.600364+IV4582\*-0.361689)

/(1+exp (-2.122529+I4019\*0.124441+I496\*0.544560+I2724\*-0.280513+I3051\*-0.244972

+I486\*0.725716+I53081\*-0.129374+I41401\*0.155278+

I2859\*0.480214+I42731\*0.526337+I60000\*0.130206+I311\*-0.014409

+I49121\*0.563141+I2761\*0.499385+I4280\*0.714783+I27651\*0.390101+I2768\*0.077929

+I5990\*0.782624+I40390\*0.280734+IE8497\*0.090493+I30981\*-0.519761

+I5859\*0.329507+I30000\*-0.169378+I41400\*0.379138+I4439\*0.325951

+I2449\*0.290615+I7242\*-0.271403+IV5861\*-0.192584+I25000\*-0.268557

+I42789\*-0.128369+I78820\*0.142208+I2809\*0.225726+I7866\*0.976377

+I51889\*0.434702+I78659\*-0.669000+I78791\*0.111907+IV4581\*0.030270+IE8490\*0.045690

+I07054\*0.465247+I30390\*-0.065050+I2875\*0.600364+IV4582\*-0.361689))

AS Predicted -- predicted probability of dead

INTO #Data1 FROM #DataQ2 ---(829799 rows affected)

select \* from #Data1

-- for cancer patient's cases

SELECT COUNT(distinct [ID]) AS nCases

, Sum(IIF([Dead] = 1, 1., 0.)) AS a

, SUM(IIF([Dead] = 0, 1., 0.)) AS b

, Max([Predicted]) AS [Predicted Control] -- If cases do not match then predicted

,[I4019],[I496],[I2724],[I3051],[I486],[I53081],[I41401],[I2859],[I42731],[I60000],[I311],[I49121],[I2761],[I4280],[I27651],[I2768]

,[I5990],[I40390],[IE8497],[I30981],[I5859],[I30000],[I41400],[I4439],[I2449],[I7242]

,[IV5861],[I25000],[I42789],[I78820],[I2809],[I7866]

,[I51889],[I78659],[I78791],[IV4581],[IE8490],[I07054],[I30390],[I2875],[IV4582]

INTO #Case -- Save in temporary file called Case

FROM #Data1

WHERE [Cancer]=1

GROUP BY

[I4019],[I496],[I2724],[I3051],[I486],[I53081],[I41401],[I2859],[I42731],[I60000],[I311],[I49121],[I2761],[I4280],[I27651],[I2768],[I5990],[I40390],[IE8497],[I30981]

,[I5859],[I30000],[I41400],[I4439],[I2449],[I7242],[IV5861],[I25000],[I42789],[I78820]

,[I2809],[I7866],[I51889],[I78659],[I78791],[IV4581],[IE8490],[I07054],[I30390],[I2875]

,[IV4582]

--(11016 rows affected)

select \* from #Case

-- for controls with patients without cancer

SELECT COUNT(distinct ID) AS nControls

, Sum(IIF([Dead] = 1, 1., 0.)) AS c

, SUM(IIF([Dead] = 0, 1., 0.)) AS d

,[I4019],[I496],[I2724],[I3051],[I486],[I53081],[I41401],[I2859],[I42731],[I60000],[I311],[I49121],[I2761],[I4280],[I27651],[I2768]

,[I5990],[I40390],[IE8497],[I30981],[I5859],[I30000],[I41400],[I4439],[I2449],[I7242],[IV5861],[I25000],[I42789],[I78820],[I2809],[I7866]

,[I51889],[I78659],[I78791],[IV4581],[IE8490],[I07054],[I30390],[I2875],[IV4582]

INTO #Control -- Save in temporary file called Control

FROM #Data1

WHERE [Cancer]=0

GROUP BY

[I4019],[I496],[I2724],[I3051],[I486],[I53081],[I41401],[I2859],[I42731],[I60000],[I311],[I49121],[I2761],[I4280],[I27651],[I2768]

,[I5990],[I40390],[IE8497],[I30981],[I5859],[I30000],[I41400],[I4439],[I2449],[I7242],[IV5861],[I25000],[I42789],[I78820],[I2809],[I7866]

,[I51889],[I78659],[I78791],[IV4581],[IE8490],[I07054],[I30390],[I2875],[IV4582]

-- (197198 rows affected)

--- to match cases and controls

SELECT a

,b

,iif(c is null, (a+b)\*[Predicted Control], c) AS c -- Replaces missing dead controls with predicted

,iif(d is null, (a+b)\*(1-[Predicted Control]),d) AS d -- Replaces missing alive controls with 1-predicted

INTO #M1

FROM #Case left join #Control -- left join as cases not matched must have a null control

ON #Case.[I4019]=#Control.[I4019]and

#Case.[I496]=#Control.[I496]and

#Case.[I2724]=#Control.[I2724]and

#Case.[I3051]=#Control.[I3051]and

#Case.[I486]=#Control.[I486]and

#Case.[I53081]=#Control.[I53081]and

#Case.[I41401]=#Control.[I41401]and

#Case.[I2859]=#Control.[I2859]and

#Case.[I42731]=#Control.[I42731]and

#Case.[I60000]=#Control.[I60000]and

#Case.[I311]=#Control.[I311]and

#Case.[I49121]=#Control.[I49121]and

#Case.[I2761]=#Control.[I2761]and

#Case.[I4280]=#Control.[I4280]and

#Case.[I27651]=#Control.[I27651]and

#Case.[I2768]=#Control.[I2768]and

#Case.[I5990]=#Control.[I5990]and

#Case.[I40390]=#Control.[I40390]and

#Case.[IE8497]=#Control.[IE8497]and

#Case.[I30981]=#Control.[I30981]and

#Case.[I5859]=#Control.[I5859]and

#Case.[I30000]=#Control.[I30000]and

#Case.[I41400]=#Control.[I41400]and

#Case.[I4439]=#Control.[I4439]and

#Case.[I2449]=#Control.[I2449]and

#Case.[I7242]=#Control.[I7242]and

#Case.[IV5861]=#Control.[IV5861]and

#Case.[I25000]=#Control.[I25000]and

#Case.[I42789]=#Control.[I42789]and

#Case.[I78820]=#Control.[I78820]and

#Case.[I2809]=#Control.[I2809]and

#Case.[I7866]=#Control.[I7866]and

#Case.[I51889]=#Control.[I51889]and

#Case.[I78659]=#Control.[I78659]and

#Case.[I78791]=#Control.[I78791]and

#Case.[IV4581]=#Control.[IV4581]and

#Case.[IE8490]=#Control.[IE8490]and

#Case.[I07054]=#Control.[I07054]and

#Case.[I30390]=#Control.[I30390]and

#Case.[I2875]=#Control.[I2875]and

#Case.[IV4582]=#Control.[IV4582]

---(11016 rows affected)

-- calculate odd ratio and percent overlap

SELECT \* FROM #M1

SELECT

sum(a\*d/(a+b+c+d))/sum(b\*c/(a+b+c+d)) As [Common Odds Ratio]

, ROUND(100.\*SUM(iif(c is not null, a+b, 0))/sum(a+b),2) as [Percent Overlap]

FROM #M1

/\*Common Odds Ratio Percent Overlap

7.90555363653156 100.000000 \*/