**Question 1:** Balance the data by propensity to seek care from Dr. Smith.  This involves first predicting probability of a patient type utilizing services of Dr. Smith; then weighting the data inversely proportional to the probability of using Dr. Smith.  Note that patients cared for by Dr. Smith and by his peer group will have a different set of weights.  The net results of weights is that patients cared for by Dr. Smith and his peer will have the same rate of various diseases.

\*\*answer does not include (a) and (b) in original question\*\*

**Overview of steps in propensity scoring**

1. Create regression model that predicts the propensity of patients that participate in the treatment.
2. Use these propensity scores to balance the data.
3. Once data is balanced, calculate the impact of treatment.

For more information: \*https://www.youtube.com/watch?v=HHZy-BR3T60&feature=youtu.be

**Coding Steps:**

1. **Import data. Name your data frame in this step. Mine is called ‘SMITH’**

> SMITH <- read.csv('EXP\_LOS.csv')

1. **Create a regression model.** ‘Cared.for.by.Dr.Smith’ is your dependent variable. Include all diagnoses as independent variables. My model name is regres\_Smith. The significant predictors are indicated by the Asterix(\*) next the coefficients below.
	1. Findings: Significant predictors = *Heart.Failure* and *Metastetic.Cancer*.
		1. If patients suffer from Heart Failure, they are 52% less likely to be treated by Dr. Smith.
		2. If patients suffer from Metastetic Cancer, they are 52% less likely to be treated by Dr. Smith.

> regres\_Smith = glm(Cared.for.by.Dr.Smith ~ Hypertension + Anemia + Diabetes + HIV + Stomach.Cancer + Lung.Cancer + Myocardial.Infarction + Heart.Failure + Metastetic.Cancer, data = SMITH, family = 'binomial')

> summary(regres\_Smith)

Call:

glm(formula = Cared.for.by.Dr.Smith ~ Hypertension + Anemia +

 Diabetes + HIV + Stomach.Cancer + Lung.Cancer + Myocardial.Infarction +

 Heart.Failure + Metastetic.Cancer, family = "binomial", data = logregres)

Deviance Residuals:

 Min 1Q Median 3Q Max

-1.580 -1.294 0.834 1.031 1.300

Coefficients:

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

(Intercept) 0.8782745 0.0763391 11.505 <2e-16 \*\*\*

Hypertension 0.0041271 0.0510593 0.081 0.936

Anemia -0.0092170 0.0510635 -0.181 0.857

Diabetes -0.0008922 0.0510630 -0.017 0.986

HIV 0.0132470 0.0510956 0.259 0.795

Stomach.Cancer 0.0132470 0.0510956 0.259 0.795

Lung.Cancer -0.0143422 0.0514615 -0.279 0.780

Myocardial.Infarction -0.0847723 0.0515814 -1.643 0.100

Heart.Failure -0.5259409 0.0525712 -10.004 <2e-16 \*\*\*

Metastetic.Cancer -0.5259409 0.0525712 -10.004 <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: 8950.0 on 6656 degrees of freedom

Residual deviance: 8695.9 on 6647 degrees of freedom

AIC: 8715.9

Number of Fisher Scoring iterations: 4

1. **Balance covariates using inverse propensity weighting.**
	1. Predicted propensity scores (regres\_Smith) are attached to the data frame (SMITH), predicted response from the regression is also attached to the data frame

 SMITH$psvalue <- predict(regres\_Smith, type ="response")

* 1. weights are calculated. Ifelse assigns weights for patients who were treated by Dr. Smith, and those who were not treated by Dr. Smith

SMITH$weight<-ifelse(SMITH$Cared.for.by.Dr.Smith==1,1/SMITH$psvalue,1/(1-SMITH$psvalue))

* 1. weighted regression. Notice that the dependent variable is now LOS and not Cared.for.by.Dr.Smith.

According to the regression output, patients who were cared for by Dr. Smith have no significance on length of stay.

final\_mod<-lm(formula =SMITH$LOS ~. - psvalue - weight, data = SMITH, weights=weight)

summary(final\_mod)

Call:

lm(formula = SMITH$LOS ~ . - psvalue - weight, data = SMITH,

weights = weight)

Weighted Residuals:

Min 1Q Median 3Q Max

-5.6829 -2.2188 0.0244 2.0819 5.9991

Coefficients:

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

(Intercept) 3.838095 0.070220 54.658 < 2e-16 \*\*\*

Hypertension 0.082109 0.044757 1.835 0.06662 .

Anemia -0.066588 0.044760 -1.488 0.13689

Diabetes 0.122423 0.044759 2.735 0.00625 \*\*

HIV -0.003799 0.044782 -0.085 0.93240

Stomach.Cancer 0.018399 0.044782 0.411 0.68119

Lung.Cancer -0.029782 0.045001 -0.662 0.50812

Myocardial.Infarction 0.033271 0.045121 0.737 0.46092

Heart.Failure 0.015603 0.046483 0.336 0.73713

Metastetic.Cancer 0.109730 0.046483 2.361 0.01827 \*

Cared.for.by.Dr.Smith -0.062532 0.044759 -1.397 0.16243

---

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

Residual standard error: 2.58 on 6646 degrees of freedom

Multiple R-squared: 0.003384, Adjusted R-squared: 0.001885

F-statistic: 2.257 on 10 and 6646 DF, p-value: 0.01256

Code Page

SMITH <- read.csv('EXP\_LOS.csv')

regres\_Smith = glm(Cared.for.by.Dr.Smith ~ Hypertension + Anemia + Diabetes + HIV + Stomach.Cancer + Lung.Cancer + Myocardial.Infarction + Heart.Failure + Metastetic.Cancer, data = SMITH, family = 'binomial')

summary(regres\_Smith)

SMITH$psvalue <- predict(regres\_Smith, type ="response")

SMITH$weight<-ifelse(SMITH$Cared.for.by.Dr.Smith==1,1/SMITH$psvalue,1/(1-SMITH$psvalue))

final\_mod<-lm(formula =SMITH$LOS ~. - psvalue - weight, data = SMITH, weights=weight)

summary(final\_mod)