# METHODS

OUTCOME ASCERTAINMENT

The response we used here is salary. Since the distributions of variable salary of two years are approximately the same as shown in graph1, and since we need to meet the requirement that the distribution of response is normal, average of two years after log transformation were used as the final response here. Log transformation was decided by R using boxcox function(graph2).

COVARIATE ASSESSMENT

Information on department, gender, primarily emphasis, certified status, publication rate, number of years since obtaining medical doctor degree, rank was collected. Department was classified as Biochemistry/ Molecular Biology, Physiology, Genetics, Pediatrics, Medicine and Surgery. Gender was classified as female and male. Primarily emphasis was categorized as primarily clinical emphasis and primarily research emphasis, certified status was classified as board certified and not certified. Rank was categorized as assistant professor, associate professor and full professor.

STATISTICAL ANALYSIS

All statistical analysis was based on the “lawsuit” dataset. Linear regression model was applied here to address the main question. Two models were selected here to discuss whether the data support the claim of gender discrimination in setting salaries. Covariates selected in both models are “department”, “gender”, “primarily emphasis”, “certified status”, “experience” and “rank”. “Publication rate” was eliminated from this model because of high collinearity with variable “primarily emphasis” as shown in Table1. VIF > 10 was the criteria for deciding multicollinearity in the model. Correlation table (Table2) was used to show correlation between two variables. Covariate such as “department” and “primarily emphasis” were highly correlated with “publication rate”, so this variable was taken out of the model. After eliminating variable publication rate, no VIF was larger than 10 shown by Table 1.

Then, confounders were investigated. Reason for testing confounding before interaction is that testing interaction first would cover the influence of confounding by R outcome. Adding potential confounders one by one is to find out which covariate could be a confounder. If adding one confounder into the model with “gender” resulted in scale of coefficients of “gender” change over 10%, we consider this covariate confounder (Table4). “Rank” is a special one here. The scale of coefficients of “gender” changed over 9% by “rank”. Though lower than 10%, many articles supported the fact that “rank” is a confounder of gender in setting salaries. Therefore, we kept “rank” in the model.

Next, interaction between gender and other variables were tested by adding to the model one at a time. Putting all the interactions between “gender” and other variables into the model once is not reasonable because coefficient of one interaction will be influenced by other interactions, therefore cover the “true” coefficient of one specific interaction. Outcome of interaction was shown in table5. It was shown that “experience” and “rank” interacted with gender in setting salaries. An interesting finding about “rank” after stratification shows that coefficient of gender\*rank(full professor) is not zero but the coefficient of gender\*rank(associate) equals to zero. According to estimate value, we could say that compared to assistant professor and associate professor, the log salary difference between two genders who were both full professors will decrease around 0.105. So, salary gap between gender is smaller among full professors than assistant professors and associate professors. (may be in the result part???)

ANOVA test was applied here to choose whether one interaction in better than two interactions for the model. The result was shown in table6. Based on Fstatistics and P.value, we conclude that one interaction is better than two interactions. Then we compare which interaction fits the model better. According to the result of R by summary function (Table7), there is not much difference between these two models based on adjusted R square and Fstatistics. Also, based on the articles, we found these two interactions make sense conceptually. So, we chose two models in the end.

All statistical analyses were conducted using R language version 1.2.1335. Two-sided p values <0.05 was considered statistically significant.

Result

# Data Description

Among the 261 participants in this study, 40.6% (n = 106) were female. As shown in Table 8, Graph 3, participants who were female were more likely to be in the department of medicine , primarily clinical emphasis, board certified and assistant professor. Participants who were male were more likely to be in the department of medicine, primarily clinical emphasis, board certified and full professor. The mean publication rate was 4.6(sd = 1.9) for male and 5.4(sd = 1.9) for female. The mean number of years since obtaining MD was 12.1(sd = 6.7)for male and 7.5(sd = 4.2) for female. The mean salary in 1994 was 177338.8(sd = 85930.5) for male and 118871.3(sd = 56168.0) for female. Salary after increment in 1995 was 194914.1(sd = 94902.7) for male and 130876.9(sd = 88778.4) for female.

# Data Distribution

In Graph 3, distributions of publication rate among male and female are alike. The publication rate of male concentrated around 3 pieces per year and 7.4 pieces per year. The publication rate of female concentrated around 3.7 pieces per year and 7.6 pieces per year. Female published faster than male on average. Plot of distribution of number of years since obtaining MD showed that female, on average, had less experiences than male. The distributions of two genders were all right-skewed, meaning there were some extreme high values of experiences. Distribution plot of salary in 1994 indicated that female earned less than male on average. Some extreme high values of salary biased the distribution into right-skewed. Distribution plot of salary in 1995 after increment illustrated that female earned less than male on average. Some extreme high values of salary biased the distribution into right-skewed. Overall, salary distribution in 1994 and 1995 looks alike.