1-) Introduction to the problem.

In the health economics literature, moral hazard and supplier-induced demand are well-known issues that threaten the stability of health insurance schemes. According to moral hazard, insured consumers tend to be less cost-conscious and, as a result, overspend on health care. According to supplier-induced demand, health care providers can stimulate their patients' moral hazard behavior in order to increase their income or protecting themselves against malpractice lawsuits. Because of these two issues, health insurance companies need to monitor and evaluate the medical practice of doctors serving their clients. However, other factors such as specialty, locality and even time may also explain the variability of medical practice. The present study analyzes a sample of the database of a Colombian health insurance company. The sample contains the number of patients, visits and ambulatory surgeries performed for each doctor working with the company during the years between 2008 and 2012. The sample also informs the doctor's specialty and locality (more specifically, the Colombian department). There is also information regarding the cost of a visit and a surgery, but unfortunately that information is not disaggregated per doctor: the sample contains the mean cost of a visit and a surgery per specialty.

The main purpose of the present study is to understand the role of specialty, locality and perhaps time in explaining the variability of the medical practice in terms of number of patients, visits and ambulatory surgeries. If the role of specialty, locality and time is weak, then other factors such as moral hazard and supplier-induced demand may be relevant and worth further investigation.

2-) The data.

As explained, the data comes from a sample of the database of a Colombian health insurance company. Originally, the data was divided into seven csv files that can be found in the original_data directory (please visit https://github.com/tmivanus/Springboard/tree/master/Capstone_01):

- *datos_ambulatorios_2008*, *datos_ambulatorios_2009*, *datos_ambulatorios_2010*, *datos_ambulatorios_2011* and *datos_ambulatorios_2012* have the number of patients, visits and ambulatory surgeries performed per doctor in the respective year as well as the doctor's specialty.
- *datos_departamentos* informs the doctor's locality (Colombian department) for 2008-2012.
- *costos_promedios_ambulatorios* informs the mean cost of a visit and an ambulatory surgery per specialty for 2008-2012 in United States Dollar (USD).

The first task was to remove the names of the doctors from the first five csv files, substituting them for identification codes in order to preserve data confidentiality. The 00_erase_doc_name_datos_ambulatorios_ano code file performs that task and creates five new csv files to replace those with confidentiality issues: *datos_ambulatorios_2008_new*, *datos_ambulatorios_2009_new*, *datos_ambulatorios_2010_new*, *datos_ambulatorios_2011_new* and *datos_ambulatorios_2012_new*. For obvious reason, the old five csv files are not made available in the original_data directory.

After that, the 01_data_outpatient_years code file organize the newly created five csv files into python pandas dataframes, which are then saved in the data directory as the following csv files: *data_outpatient_2008*, *data_outpatient_2009*, *data_outpatient_2010*, *data_outpatient_2011* and *data_outpatient_2012*. The reader may wish to see step-by-step what tasks are being performed by opening the 00_optional_data_outpatient_year_details code file and choosing a year in the second cell.

The 02_mean_outpatient_costs_usd code file deals with the *costos_promedios_ambulatorios* csv file, organizing it into a python pandas dataframe and saving it in the data directory as the *mean_outpatient_costs_usd* csv file. That code file also translates the specialty of each doctor from Spanish to English using Google Translate.

Finally, the 03_data_departments code file organizes the *datos_departamentos* csv file into a python pandas dataframe, saving it in the data directory as the *data_departments* csv file.

All the recently created csv files in the data directory are put together to form a panel data by the 04_panel code file, which concatenates all the *data_outpatient* files from 2008 to 2012 and then merges it with *mean_outpatient_costs_usd* and *data_departments* files. The resulting panel data is saved in the data directory as the *data_panel* csv file. It contains 8294 rows and 12 columns of information that is going to be analyzed in the present study. It is worth noting that 10 rows were discarded because of errors involving the doctor's specialty name or specialty identification code.

3-) First analyses.

05_first_analyses code file contain the first analyses made using the resulting panel data. Because many medical specialties and departments (locations) don't have enough data, I initially suggested concentrating on 8 specialties and 9 departments. All of them have data from 2008 to 2012. Table 1 shows the 8 specialties and 9 departments:

Table 1

Medical specialties	Departments
- 137, 'general surgery' - 143, 'plastic surgery' - 200, 'dermatology' - 341, 'gynecology and obstetrics' - 480, 'ophthalmology' - 514, 'orthopedics and traumatology' - 521, 'otorhinolaryngology' - 750, 'urology'	- 5, 'antioquia' - 8, 'atlantico' - 11, 'bogota' - 13, 'bolivar' - 17, 'caldas' - 54, 'n. de santander' - 66, 'risaralda' - 68, 'santander' - 76, 'valle del cauca'

This subpanel from the *data_panel* csv file contains 6444 entries. The variables n_patients and n_visits are not the same but are very correlated, so I suggest to use only n_visits. I want to know how much specialty, department and year can explain the variability of the medical practice in terms of n_visits and n_surgeries. If specialty, department and year don't explain much, then there is a good chance that other factors like moral hazard and supplier-induced demand play an important role.

A potential problem is that some specialties in some departments in some years still don't have enough data or don't have data at all. For example, there is no data for general surgery in Santander in 2011 or for plastic surgery in Bolivar in 2010.

4-) Correlations.

06_correlations code file contain analyses involving correlations and statistical tests using bootstrap techniques. Because the potential problem aforementioned, I decided to investigate a little more the number of observations (doctors) per specialty/department/year. I wanted to make sure I have more than 30 observations for each specialty in each department in each year. As a result, I ended up with a smaller subpanel with 4 specialties, 2 departments and 5 years. That smaller subpanel obeys the following condition: there are more than 30 observations (doctors) for each specialty in each department in each year. Table 2 shows the 4 specialties and 2 departments.

Table 2

Medical specialties	Departments
- 200, 'dermatology' - 341, 'gynecology and obstetrics' - 480, 'ophthalmology' - 521, 'otorhinolaryngology'	- 5, 'antioquia' - 76, 'valle del cauca'

This new subpanel from the *data_panel csv* file contains 2552 entries. In addition, I have created the variable surgeries2visits to represent the ratio between number of surgeries and number of visits. I have also created dummy variables for each specialty and department. Table 3 summarizes the information in the subpanel:

Table 3

```
Int64Index: 2552 entries, 0 to 2551
Data columns (total 19 columns):
doc_code            2552 non-null int64
spec_code           2552 non-null int64
spec_es             2552 non-null object
spec_en             2552 non-null object
year                2552 non-null int64
n_visits            2552 non-null int64
n_surgeries         2552 non-null int64
n_patients          2552 non-null int64
c_visit             2552 non-null float64
c_surgery           2552 non-null float64
dep_code            2552 non-null int64
department          2552 non-null object
surgeries2visits    2552 non-null float64
spec_der            2552 non-null uint8
spec_gyn            2552 non-null uint8
spec_oph            2552 non-null uint8
spec_oto            2552 non-null uint8
dep_ant             2552 non-null uint8
dep_val             2552 non-null uint8
```

Once I have decided the size of the subpanel, I started to analyze how specialty, department and year may affect number of surgeries, number of visits and surgeries-to-visits ratio. I first used “violin” graphs (please see cell 5 in the code file). The horizontal width of each “violin” figure shows how the data is scattered (i.e., roughly speaking, the size of the uncertainty) whereas the vertical height informs the data density in each point. Specialty seems to be main factor affecting surgeries, visits and surgeries-to-visits ratio.

Next I presented heatmaps illustrating correlations between variables (see cell 6). The numbers are linear correlation coefficients that goes from 1 (perfect positive linear correlation) to -1 (perfect negative linear correlation). As expected, there seems to be a positive linear correlation between visits and surgeries. The first heatmap uses the codes for specialty and department. This is not good because the codes provide an arbitrary scale (why one specialty has a code number higher than other?). That arbitrary scale may affect the correlation with other variables, so we don't know if we are seeing the true correlations or some random correlations due to the arbitrary scale. The main problem is specialty because it has 4 different possibilities (department only has 2 possibilities, so the use of an arbitrary scale is not so problematic and can be easily substituted for 0 and 1). The second heatmap introduces the cost of surgery (c_surgery) and the cost of visit (c_visit). Unfortunately, those variables are not disaggregated: they are average costs for each specialty in each year. That's why I haven't used them much. However, because they are related to specialty, they could provide an alternative non-arbitrary scale for specialty. Not surprisingly, c-surgery and c-visit show correlations with spec_code and year (remember: they are average costs for each specialty in each year). Despite not being disaggregated, those variables show interesting correlations with number of surgery and number of visits. The third heatmap introduces dummy variables for each specialty and department. Here we can see even better the expected correlation between them and c-surgery or c-visit. The use of dummies for specialty and department plus the inclusion of c-surgery and c-visit is an interesting approach, although we may run into a multicolinearity issue. Of course, we need to be careful not to introduce all dummies in a regression to avoid perfect correlation among them (as seen between the dummies for departments in the heatmap) and stress the aggregated nature of c_surgery and c_visit. I would also use dummies for years so “year” = 0 has meaning.

After analyzing those heatmaps, I started to make a series of bootstrap statistical tests to see how specialty, department and year may be affecting number of surgeries, number of visits and surgeries-to-visits ratio. First I tested the following H0 hypothesis: difference in means for surgeries/visits/surgeries-to-visits ratio in different categories is by chance (they actually come from the same distribution). Then I focused on the means themselves by testing the following H0 hypothesis: there is no difference in means for surgeries/visits/surgeries-to-visits ratio in different categories. In particular, for this last test, I used

95% confidence intervals. If intervals overlap, H0 is not rejected. If intervals do no overlap, H0 is rejected. Based on those tests, I was able to reach the following conclusions:

- Specialty definitely changes the mean of surgeries, not necessarily the mean of visits. It definitely changes the mean of surgeries-to-visits ratio.
- Department definitely changes the mean of surgeries and the mean of visits. It doesn't necessarily change the mean of surgeries-to-visits ratio (the department with more visits also has more surgeries).
- Year definitely changes the mean of surgeries and the mean of visits in the long run, if you compare 2008 to 2012 for example. As time pass, both measurements tend to increase. However, because of that, year does not change the the mean of surgeries-to-visits ratio.

I also wanted to check how specialty, department and year affect the relationship between number of surgeries and number of visits. So I implemented a series of regressions between number of visits (as the x-variable) and number of surgeries (as the y-variable) for different categories (see cell 15). I used 95% confidence intervals to test the slope and the intercept of each regression. As a result, I was able to reach the following conclusions:

- Dermatology and otorhinolaryngology may have the same slope, but not the same intercept. Gynecology & obstetrics and ophthalmology may have the same intercept, but not the same slope.
- Departments may have the same slope and intercept. There is more uncertainty regarding Valle del Cauca than Antioquia.
- Year may have the same slope and intercept, although slope may be increasing with year. Uncertainty seems to increase with year.

5-) Final results.

07_supervised_classification_and_regression code file contains the final results. Please see the tables, graphs and explanations in the file.