Exploratory Data Analysis

1. Sale Behavior

The usable features in the medicare dataset include, drug name, total drug cost, and total drug amount. These features have to be made useful in order to build an effective classification model. In a preliminary investigation shown in Fig. 1, we plotted the average total cost and amount (size of the bubble) of all drugs prescribed by fraud and typical providers (blue). We can see that the average cost and amount are similar. However, when we narrow down to just narcotics*, we can see a distinctly different sale between fraud and typical providers (red). Fraud providers sell significantly higher amount of narcotics and total to the higher average cost. The average narcotics sold by typical provider is $3,900, while it is over $41,000 for fraud providers. Therefore, we will look more into the statistics of this anomalies.

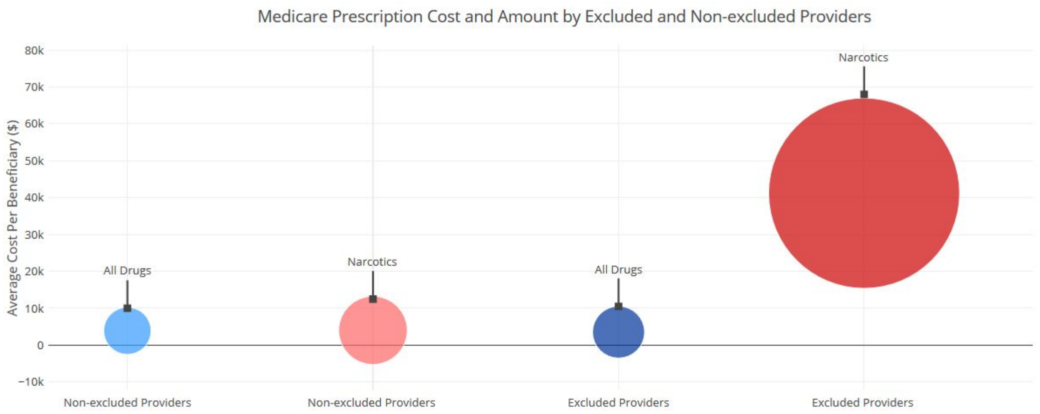


Fig.1 Average total medicare cost per beneficiary by excluded (fraud) and non-excluded (typical) providers.

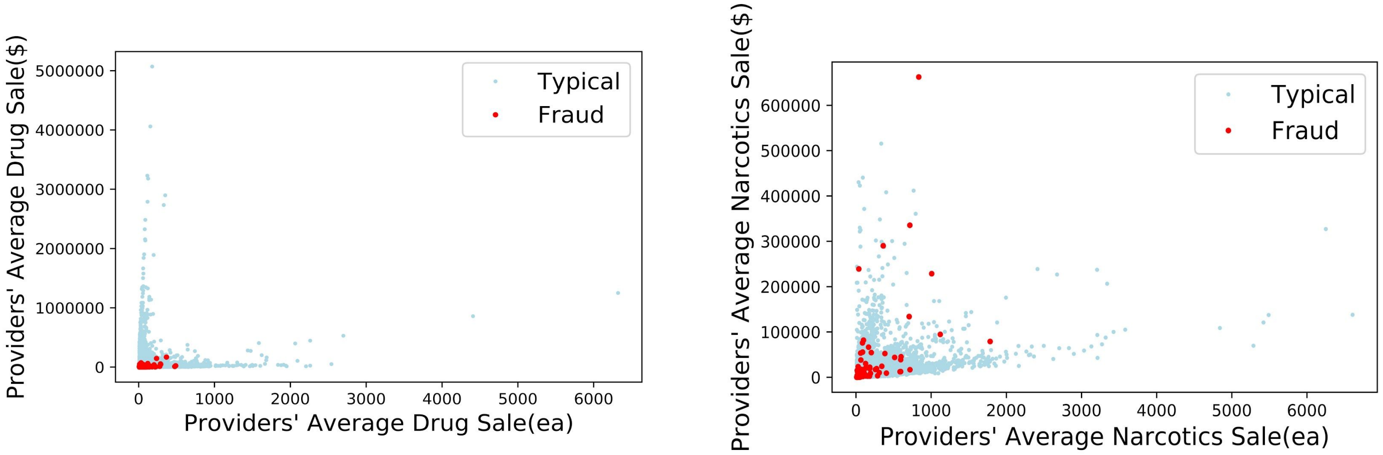


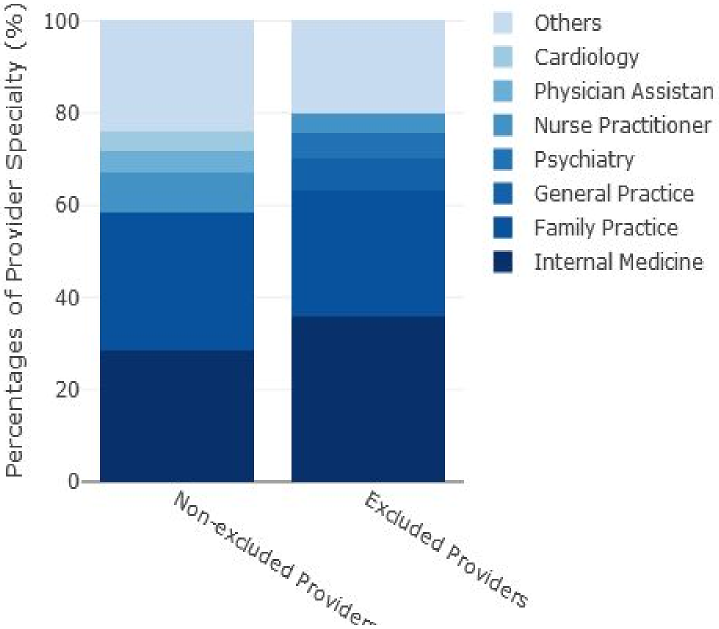
Fig.2 Scatter plot of average amount of all drugs (left) and narcotics (right) VS the average total cost by typical (light blue) and fraud providers (red).

*narcotics includes ['OXYCONTIN', 'OXYMORPHONE HCL ER', 'MORPHINE SULFATE ER', 'OXYCODONE HCL', 'OXYCODONE-ACETAMINOPHEN', 'FENTORA', 'SUBSYS', 'HYDROCODONE-ACETAMINOPHEN', 'SUBOXONE', 'OPANA ER', 'HYDROCODONE-ACETAMINOPHEN'], which are among the top 25 of the most sold drugs by fraud providers

The scatter plot of the drug/narcotics amount and cost sale from typical and fraud providers in Fig. 2 showed that there is a huge overlaps and maybe inseparable. However, we can observe that in narcotics only plot (Fig. 2 right), the trend is more slightly more scattered. To determine that population mean of the cost from fraud and typical providers are in fact significantly different, we applied one-way ANOVA test.

The results showed that the null hypothesis that fraud providers sell all drugs with the similar total cost as typical provider has a p-value of 0.6987, therefore, we do NOT reject the hypothesis. On the other hand, the null hypothesis that fraud providers sell narcotics with the same total cost as typical provider has a p-value of 0.0000, therefore, we REJECT the hypothesis. This implies that the drug type and sale of narcotics by a provider could be useful as a feature for classification purposes.

2. Specialty of Fraud Providers



The second objective is to determine if there is any correlation between the specialty of providers and whether that provider is a fraud one. For this objective, we employed Spearman’s correlation because it is suitable for comparison between two categorical classes and it does not assume the datasets are normally distributed. The correlation between providers' specialty and whether the provider is fraud is 0.0023 with a p-value of 0.0003. The low p-value indicates that the null hypothesis that specialty and whether the providers are fraud are uncorrelated, is rejected. However, the correlation value was virtually 0, which means that the correlation would not be necessarily useful, and hence we would not leverage this feature for the classification model.