Health Exams in Vietnam Data Analysis Report

Qiao Su

26/11/2019

Executive Summary

Medical care for serious diseases can be very expensive and time and time consuming, especially in the late stage. To allow early detection, patients can get regular check-ups (or "general health examinations," GHEs). However, there are many possible obstacles to getting at risk citizens to go to regular check-ups. Firstly the at-risk patients must be properly identified. Additionally, their own schedules, biases or experiences with checkups may prevent them from wanting to go see the doctor.

To investigate this question, we used a public health dataset from (Vuong 2017). It was collected by public health researchers in Vietnam who wanted to determine what obstacles prevented widespread use of regular check-ups. They surveyed respondants by traveling and conducting in person interviews for about 10–15 minutes. This produced 2,068 valid responses of which half did not know when their last GHE was scheduled. However, the other 1467 surveyed had a GHE before.

The data contains 50 variables concerning many relevant demographics and responses, organized into three main categories. The first category are demographics such as BMI, age, education and sex. The second category quantifies their attitude towards health such as whether they can basic medical equipment, and how much time the respondent spends on sports and physical exercise. The last category quantifies their attitudes relating directly to the GHEs such as their perceived ability of examiner and the percieved attractiveness of information they received in check-ups.

We first used this data to evaluate the perception of the GHEs and found that although GHEs were not percieved badly on average, the participants rated the quality of GHEs higher than the information which was given at the GHE. Secondly, we evaluated the most important variables to determining whether a respondent would obtain a GHE by modelling their time since checkup in terms of the other variables. We found that the most important predictors included whether they had checkups when they obtained diseases, whether they were updated on their own health and that of their family, how frequently they believed that they should have a checkup and their rating of the quality of GHEs. In particular, in our model more frequent incidence of doctor visits increased the probability of getting an unprompted checkup by 5 fold.

Lastly, we checked what proportion of the population might be influenced to have a checkup by building a random forest classifier to distinguish those patients who would obtain a GHE within 12 months from those who would not. The classifier achieved good performance of 76% on the test data and had reasonable diagnostics. We used our classifier to predict what kind of responses would be given by those who answered that they did not know when they would last have a GHE. We found that about 30% of these respondents were similar to those who had answered that they had a GHE within the last 12-24 months. This suggests a significant portion of the population might be amenable to getting a checkup if prompted. Additionally, we studied what were the most important variables in determining their classification.

Overall, our analysis suggests that the information given at checkups may be improved in order to raise public perception of GHEs in Vietnam. Further, those who interact more often with the health system seem more inclined to get regular yearly GHEs.

Introduction

Due to their importance for detecting early incidence of disease, we wanted to determine whether we can determine what factors would explain and predict who would undergo a yearly general health examination (GHE). Based on the previous work of (Vuong 2017) we knew that specific demographics, personal biases or

experiences with checkups may prevent Vietnamese citizens from wanting to undergo GHEs. To investigate this question, we used the public health dataset published by (Vuong 2017). It was collected by public health researchers in Vietnam who wanted to determine what obstacles prevented use of regular check-ups. They surveyed respondants with in person interviews and produced 2,068 responses. Of these cases, 1467 surveyed reported when they had undergone a GHE.

The dataset includes three categories of variables (50 variables total) about the participants. The first category are demographics such as BMI, age, education and sex. The second category quantifies their attitude towards health such as whether they can basic medical equipment and how much time the respondent spends on sports and physical exercise. The last category quantifies their attitudes relating directly to the GHEs such as their perceived ability of examiner and the percieved attractiveness of information they received in check-ups.

We first used this data to evaluate the perception of the GHEs and found that although GHEs were not percieved badly on average, the participants rated the quality of GHEs higher than the information which was given at the GHE. Secondly, we evaluated the most important variables to determining whether a respondent would obtain a GHE by modelling whether they had obtained a checkup within the past year using a logistic generalized linear model. We selected variables by stepwise selection and examined the explanatory effect size. We found that the most important predictors whether they visited the doctor for prompted reasons and their age. This suggests that aside from age which is strongly correlated with health, contact with the medical system is strongly explanatory of willingness to undergo GHEs.

Lastly, we checked what proportion of the population might be influenced to have a checkup by building a random forest classifier to distinguish those patients who would obtain a GHE within 12 months from those who would not. The classifier achieved good performance of 76% on the test data. We used our classifier to predict what kind of responses would be given by those who answered that they did not know when they would last have a GHE. We found that about 30% of these respondents were similar to those who had answered that they had a GHE within the last 12-24 months. This suggests a significant portion of the population might be amenable to getting a checkup if prompted. Additionally, we studied what were the most important variables in determining their classification. Our random forest classifier assigned the highest importance to their health demographics, whether they had an unprompted checkup, and for what reason. These variables were also identified as significantly correlated with the incidence of an unprompted checkup by our logistic model.

Methods

Removal of outliers and missing data

After importing the data, we found there was 90 cases with missing data out of the 2068 participants. The variables with missing data included numeric variables only: height, weight and the ratings of GHEs e.g. percieved timeliness of the checkup. We first discarded height and weight since these two correspond fully to BMI (also see EDA). There seemed to be a possibility that the remainder missing cases were informative missing data, since the participants had answered everything else. Therefore we kept the other observations that had missing data.

We also looked for outliers in the data among the remaining variables. Only the Age and BMI directly inform us about the patient's health so outliers in these variables are very important and were removed (see EDA). Lastly we removed outliers in the date to restrict participants from September to October 2016. This left 1941 responses.

Exploratory Data Analysis

We may hypothesize that the respondent's general health is inversely related to their visits to GHEs. Furthermore their perceptions of GHEs may be approximately correlated with what value they have derived

from the procedures in improving their health. Therefore our hypothesis of the connection between frequency of GHEs and the other covariates is illustrated in Figure 1.

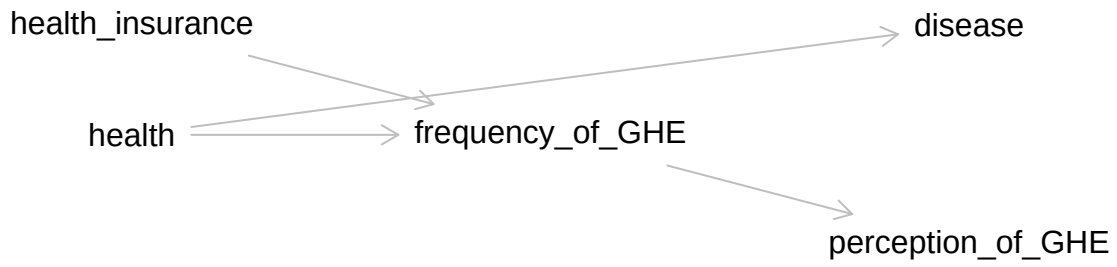

Figure 1: Causal diagram illustrates hypothesized relationships between checkups, disease and cost of public healthcare.

Univariate variable distributions

Data were collected in 2016 from participants from 13 to 83 years of age, on 31 separate dates in the year 2016. There are outlier dates in Figure 2 A which may be typos and fall outside of real dates e.g. '20169828'. Therefore we trimmed the few outlier dates which fall outside of September and October 2016. As seen in Figure 2 B-C, there are also outlier values in the variables age and BMI. We can see that the data are upward skewed by the outliers. We therefore removed data with age greater than 50 and BMI greater than 35.

BMI is an important indication of fitness and calculated as weight (kg) / [height (m)]^2. When we calculated the BMI from the height and weight we found that the quantities corresponded exactly, as seen in Figure 2 D. Therefore since the information is redundant and BMI is more indicative of health, we discarded height and weight variables.

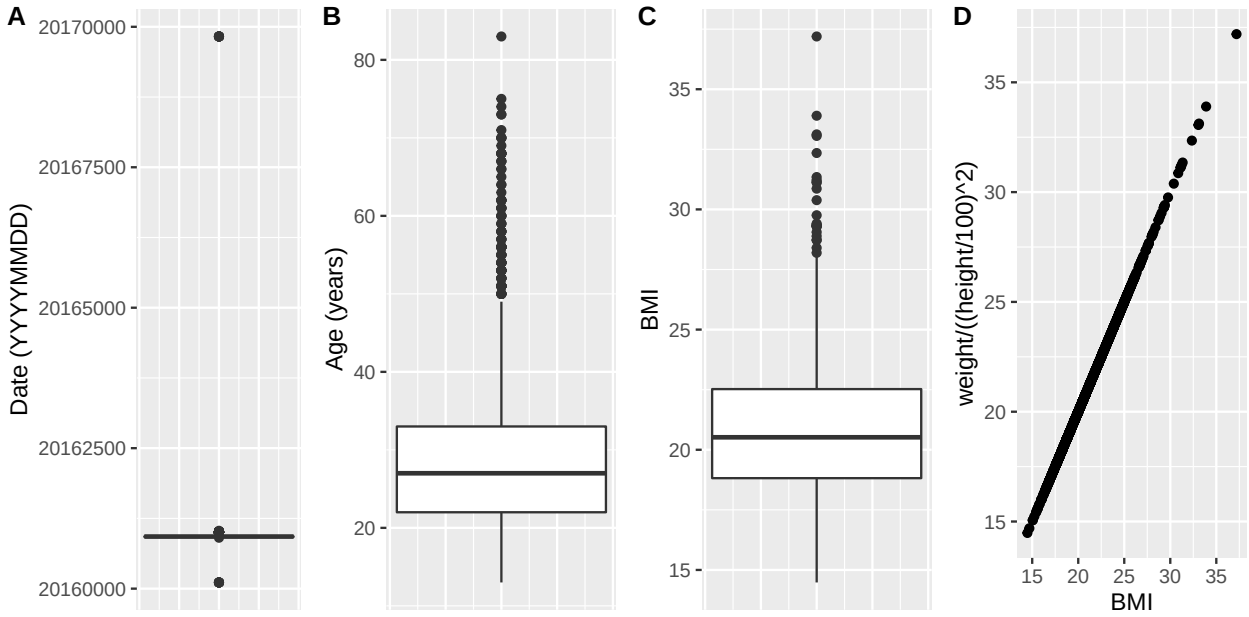

Figure 2: A-C) Distributions of important demographic variables in the raw data show outliers. We trimmed the outliers in the date, age and BMI. D) We can see BMI is corresponding well to BMI calculated from height and weight.

Pairwise distributions

We checked the correlation among the numeric variables. We observed that there doesn't appear to be collinearity among the numeric variables. Furthermore, age and BMI have a reasonable correlation. Additionally, we found that the ratings naturally grouped by correlation into responses concerning the quality of GHEs (Tangibles, percieved quality of tangible equipment and personnel to Empathy, percieved empathy of the staff) and about the type of information they recieve during GHEs (SuffInfo, rating of the sufficiency, to Popular info, rating of the popularity of the information).

Based upon this, we constructed scores to represent the rating of the quality and the information respectively by averaging the corresponding variables, which we could use later on to quickly summarize them. We plotted the distribution of these two score variables in Figure 3. There were 20 missing cases for the quality score and 2 missing cases for the information score.

Current perceptions of checkup quality and the information recieved in checkups

Figure 3 A-B shows that overall, about half of respondents believe that checkups are a waste of time or waste of money. In fact about 25%, or 582 out of our 1941 respondents believed both. However we wanted to know what aspects of the checkups are good versus need improvement.

In order to summarize how respondents felt about the specific parts of checkups, we first looked at the distibutions of the ratings they gave of the checkups. We checked pairwise distributions of the main rating variables (data not shown). As noted, these ratings are also significantly correlated and thus we averaged them to make the quality and information scores which are summarized in Figure 3. The mean quality score is 3.54/5 and the mean information score is 2.82/5.

This suggests that the quality of checkups is viewed more favorably than the information given by checkups. Therefore to improve, it may help to start by changing the sufficiency, attractivenness and impressiveness of the information given in check-ups.

Determining the most important factors that make participants likely to take a general health exam every 12 months

We aimed to determine what factors make a person less likely to get check-up every twelve months by modelling the variable RecPerExam (the time since the respondent got an unprompted checkup). RecPerExam has four levels: less12 = less than 12 months, b1224 = between 12 and 24 months, g24 = over 24 months, unknown = respondent doesn't know. We first trimmed the cases that are 'unknown' since this is not an informative response, leaving 1467 cases. We interpreted responses with the less24 value as 'likely to get a check-up every twelve months', and those with the g24 value as not likely to get a checkup very twelve months. In this setup, we additionally trimmed the data for respondents who got their checkup between 12 and 24 months before the survey since it's unclear which category they wold fall into.

Since the response variable is composed of two discrete events after filtering, we used a logistic generalized linear model (glm). However there is some class imbalance as there are about twice as many cases of less than 12 months than of the other level (data not shown). We also note that another limitation of this model is that we may not have a representative sample for the population which is unwilling to undergo a yearly checkup.

First we picked the terms we should have in the best explanatory model by using stepwise model selection with AIC using the *mass* package. In order to reduce runtime, we used the quality and information scores as summaries of the underlying response ratings. The diagnostic plots for the chosen model shown in Figure 4 are meant for linear models, which is why we see that for example the residuals are not highly normally distributed in the normal QQ plot. However they show that we do not see significant outlier cases. For example, in the plot of the Residuals vs. Leverage, there are no outliers outside the Cook's distance lines. This suggests that the model is reasonable despite the class imbalance in the data.

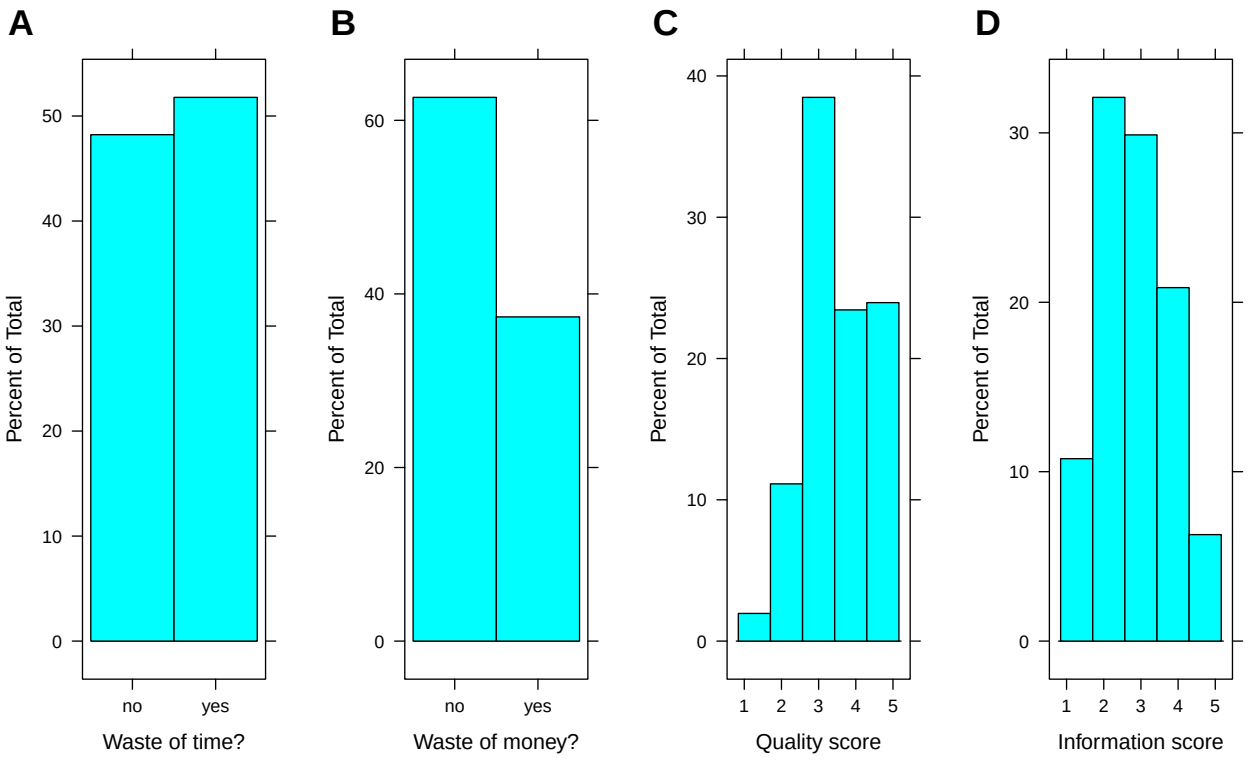

Figure 3: Perception of GHEs. A) About 50% of respondent believe check-ups are a waste of time. B) About 40% of respondent believe check-ups are a waste of money. C-D) The mean quality score is 3.54/5 and the mean information score is 2.82/5.

Starting from these 39 predictors and AIC of 609.5, the model summarized in Table 1 was produced with 21 predictors and a reduced AIC of 539. The baseline value of the response for this model is "g24", whereas the alternate value is "less24". Therefore, these coefficients We evaluated how important each of the terms was by looking at their effect size and the significance level of the effect size. We see that the factor predictors chosen have at least one level ech with effect size with an effect significant at $p < 0.05$. However among these, the variables with the largrest fitted effect size are "RecExam", "ReaExam" and "Age". Compared to the respondent with baseline values for all variables, we can interpret these effects as follows:

- if the time since the respondent last visited a doctor with symptoms of a disease (RecExam) is over 24 months, they are approximately exp(-3.031)= 5% less likely to get a checkup yearly. However if their last visit was within 12 months or less, they are approximately exp(1.716)= 5.5 times more likely than someone whose last visit was between 12 and 24 months.
- if the reason for visit (ReaExam) is voluntary, they are approximately exp(1.598)= 5 times more likely to get a checkup yearly; if the reason is due to someone's request, they are approximately exp(2.060)= 7 times more likely than someone whose last visit was for 'worrying symptoms'.
- if their age (Age) is over 18 and less than 50, they are at least exp(1.574)= 5 times more likely to get a checkup yearly than someone who is 18 or under.

In particular, the effect size of the RecExam variable is largest as well as significant at the $p < 0.001$ level. The coefficients in Table 1 suggest that relative to a respondent with the baseline values of the variables given above, a respondent who makes more frequent contact with the healthcare system is much more likely to get a yearly checkup.

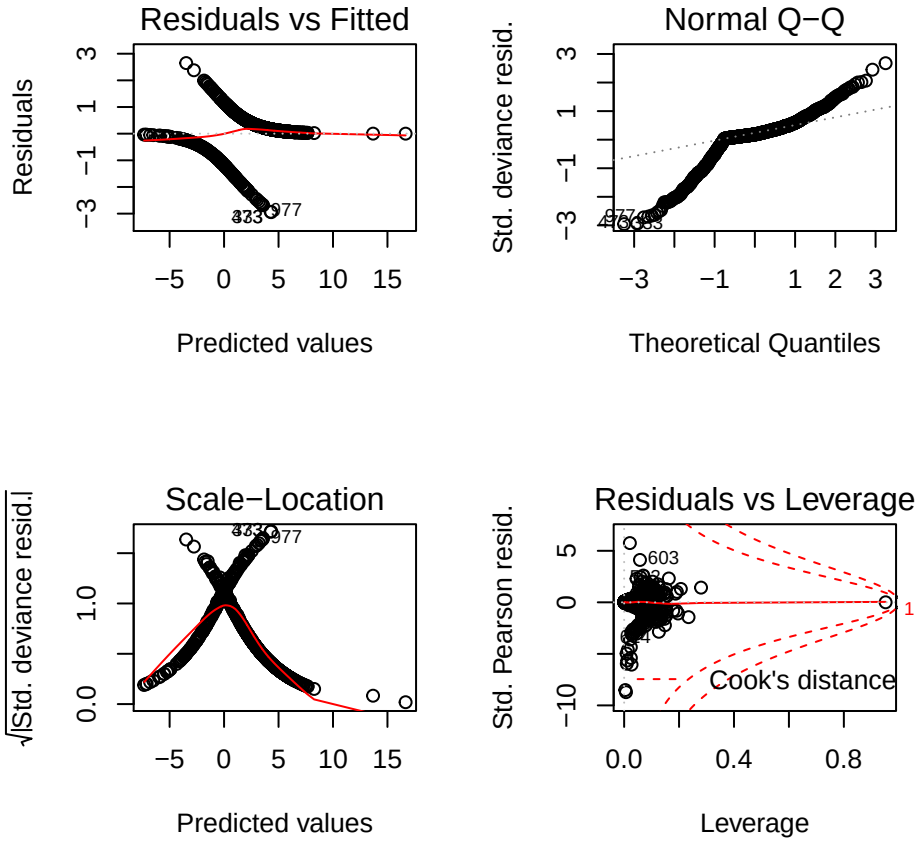

Figure 4: Diagnostic plots for the model chosen by stepwise variable selection.

Table 1: Terms picked through stepwise selection in the logistic glm model of RecPerExam. Stars indicate the level of significance of the fitted effect size.

	Dependent variable:
	RecPerExam
Constant	151,470.300 (102,840.700)
date	−0.008 (0.005)
Sexmale	0.571* (0.293)
BMI	−0.069 (0.049)
HealthInsyes	0.437 (0.291)
RecExamg24	−3.031*** (0.621)
RecExamless12	1.716*** (0.463)
RecExamunknow	−0.498 (0.650)
ReaExamnoti.disease	0.761 (0.703)
ReaExamrequest	2.060*** (0.345)
ReaExamvolunteer	1.598*** (0.310)
Wsttimeyes	−0.750*** (0.289)
Wstmonyes	0.522 (0.339)
DiscDiseaseyes	−0.533* (0.321)
NotImpyes	−0.695*** (0.254)
HthyPriorityyes	0.842*** (0.304)
ComSubsidyyes	0.657** (0.259)
Habityes	0.946*** (0.280)
FlwHealthyes	0.660** (0.278)
ExamToolsyes	−0.785*** (0.269)
UseMonlater	−0.664** (0.295)
UseMonpartly	−0.215 (0.384)
StChoiseclinic	0.833*** (0.304)
StChoiseselfstudy	0.861*** (0.322)
CHPercgood	1.685*** (0.468)
CHPercquite	0.953*** (0.293)
CHPercunknown	1.293** (0.525)
UseITno	0.698** (0.341)
UseITyes	0.175 (0.293)
Age_gr>=50	13.903 (552.484)
Age_gr18-29	1.665** (0.822)
Age_gr30-39	1.574* (0.849)
Age_gr40-49	3.082*** (1.036)
quality_score	−0.265* (0.141)
Observations	863
Log Likelihood	−235.507
Akaike Inf. Crit.	539.015

Predicting which demographic will be most likely to take a general health exam

In order to predict which people would be easiest to convince to take a GHE every 12 months, we again build a model of the time since respondent last visited a doctor for a check-up (RecPerExam). One of the four responses is 'unknown', meaning the respondent doesn't know when they last visited a doctor for a check-up when not prompted by a specific illness. However, we can try to predict for the respondants who answered 'unknown' whether they would fall into one of the other three categries (less12 = less than 12 months, b1224 = between 12 and 24 months, g24 = over 24 months). We first set aside the cases for which RecPerExam is 'unknown'. Working with the remaining 1467 cases, we built a random forest model using the package Ranger to predict RecPerExam.

We chose to use a random forest model due to the large number of predictors available Since here we are interested in predicting RecPerExam, rather than interpretation, we decided to include all of the potential predictor variables (other than id). To avoid missing values as much as possible, we also used our quality and information scores in the place of the underlying ratings. We randomly shuffled the rows and split the data 7:3 into a training and test set.

To check how many trees we should have in the random forest, we first plotted the error versus the number of trees when using default parameters for the random forest in Figure 5. We can see that without tuning, the performance error stabilizes around 100 trees. As well the performance on the cases with less than 12 months are predicted better than the other two cases due to the greater number of cases.

Subsequently we used ranger to tune the following parameters, proceeding with 100 trees.

- mtry: number of variables to randomly sample as candidates at each split. We tried a range from 2 to 46 (the total number of predictors).
- minimum node size: the number of samples in each terminal node. Lower node size means more complexity.
- sample size: number of samples to train upon.

The parameters of the best classifier, which achieved an out of bag root mean squared error of 0.47, were mtry=23, node size of 5, and sample size of 0.6. We verified the performance of this classifier on the test set. It achieved a test set accuracy of 76% as seen in Table 2.

We were also interested in characteristics of the patients who haven't gotten an exam in the last 12 months, but are most like other patients who have. We plotted the variable importance of this classifier in Figure 6 and found considerable overlap with the significant variables found by glm model fitting.

Table 2: Predictions of the tuned random forest model (y-axis) versus true classifications (x-axis). Performance achieved is about 76%.

	b1224	g24	less12
b1224	0.0560748	0.0233645	0.0163551
g24	0.0116822	0.1004673	0.0397196
less12	0.0957944	0.0467290	0.6098131

Next, we used our classifier to predict the status of the 466 unknown cases. The classifier predicts 343 'less12' cases, 101 'g24' cases and 22 'b1224' cases. This suggests that about 101/343 or about 30% of respondents in the population studied could be prompted to get an exam if they had not gotten one in the last 12 months.

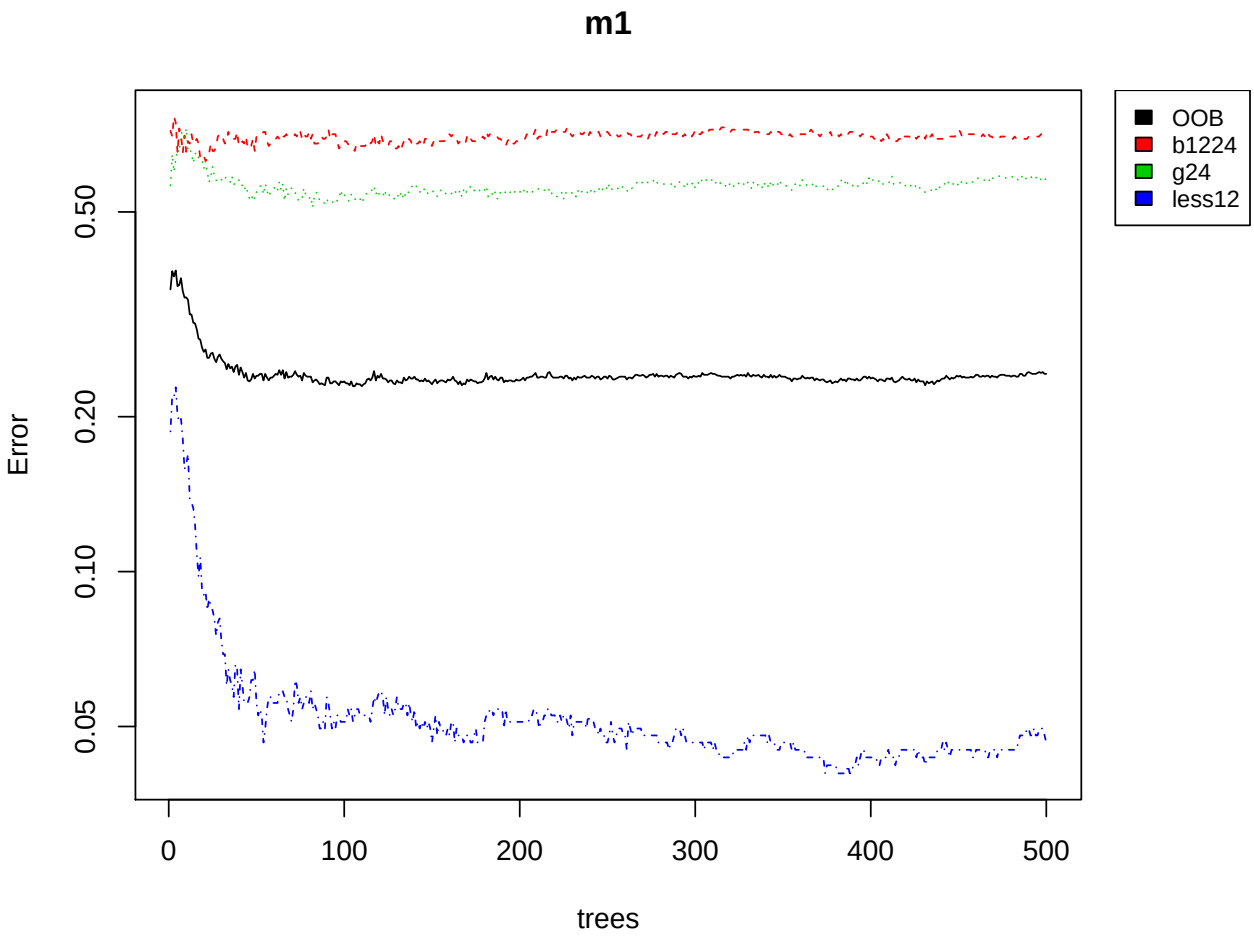

Figure 5: Plot of error of random forest model of RecPerExam versus the number of trees when using default parameters. OOB is the out of bag error. Levels of RecPerExam include less12 = less than 12 months, b1224 = between 12 and 24 months, g24 = over 24 months Performance on the cases with less than 12 months are predicted better than the other two time periods.

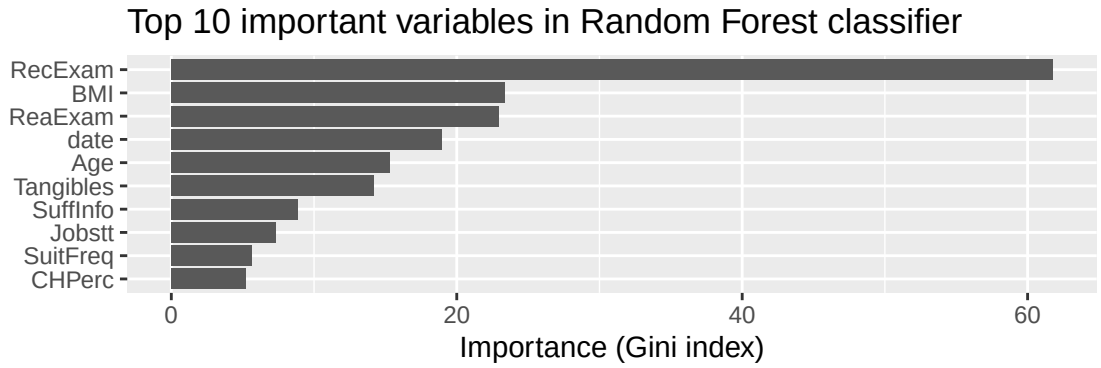

Figure 6: Variable importance of tuned random forest classifier.

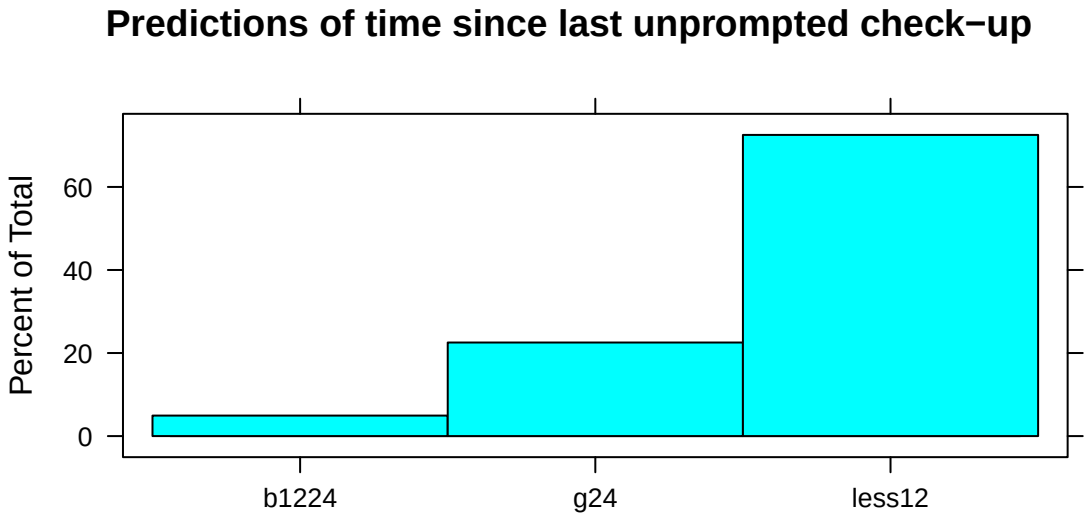

Figure 7: Distribution of predicted RecPerExam for respondents who did not know when their last visit was. RecPerTime represents the time since therespondent last visited a doctor for a check-up, not prompted by a specific illness.

Conclusion

We have examined the relationship between general health exam (GHE) attendance and many covariates of interest including demographic variables, self reported public perceptions, and health awareness of participants for the Vietnam health exams dataset. We first found that in terms of the public perception of GHEs, the quality of care recieved an above-average rating while the information provided in the GHE recieved a slightly below average rating. This suggests that the main area of improvement when trying to encourage the public to take GHEs more frequently is to increase the perception of the information given.

Secondly, we saw that among the most important predictors of whether a respondent is more likely to take a GHE yearly without being prompted is how frequently they get examinations for other reasons. In fact our model suggests those who take a voluntary examination are 5 times more likely than even those who have 'worrying symptoms' to get unprompted checkups yearly. When we examined which variables are most important for the classifier that we built to identify which respondents would take a GHE yearly, we also found that the time of their last checkup for medical reasons was by far the variable with the highest importance. Based upon our classifier, about 30% of the cases which had not reported when their last GHE was had a similar profile

However, there are some caveats to this analysis. We found in our EDA that there is an imbalance between the number of people who self-reported as having a GHE within the past year, and those who gave all other responses. We also used those who reported as having a GHE within more than 2 years as a stand in for those who would be reluctant to undergo a general GHE and those who reported as having a GHE within 12-24 months as a stand in for those who could presumably be persuaded to undergo a general GHE. As well, inherently we did not have a precise measure of the underlying health of the respondents here, so it is difficult to determine their general health status which we hypothesized could be a confounder of how often they will attend a GHE. Lastly we have used GLMs to find important covariates and random forest to classify the cases, however, both methods are supsceptible to class imbalance.

Bibliography

Vuong, Quan Hoang. 2017. "Data Descriptor: Survey data on Vietnamese propensity to attend periodic general health examinations." Nature Publishing Groups. doi:10.1038/sdata.2017.142.