Health Exams in Vietnam Data Analysis Report

Qiao Su

19/11/2019

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.

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.

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

In this report, we are interested to determine whether we can analyze the 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 dataset includes three categories of variables 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 their time since checkup in terms of the other variables using a generalized linear model. 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.

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 generalized linear 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. 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 incidence of disease. Furthermore their perceptions of GHEs may be approximately correlated with what value they have derived from the procedures in improving their health. Lastly the number of GHEs brings up the cost of 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.

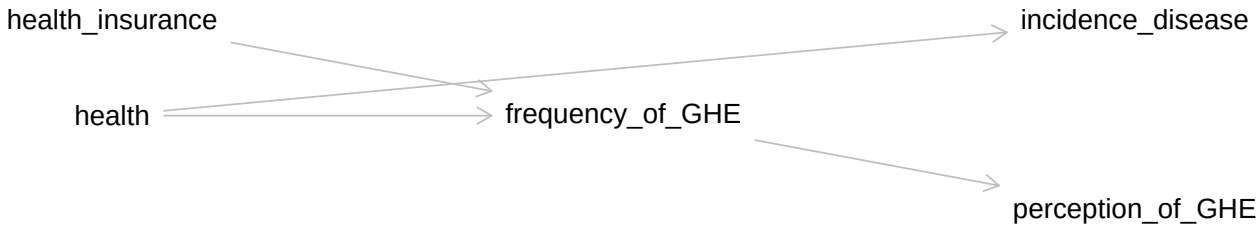


Figure 1: Causal diagram illustrates hypothesized relationships between checkups, disease and cost of public healthcare.

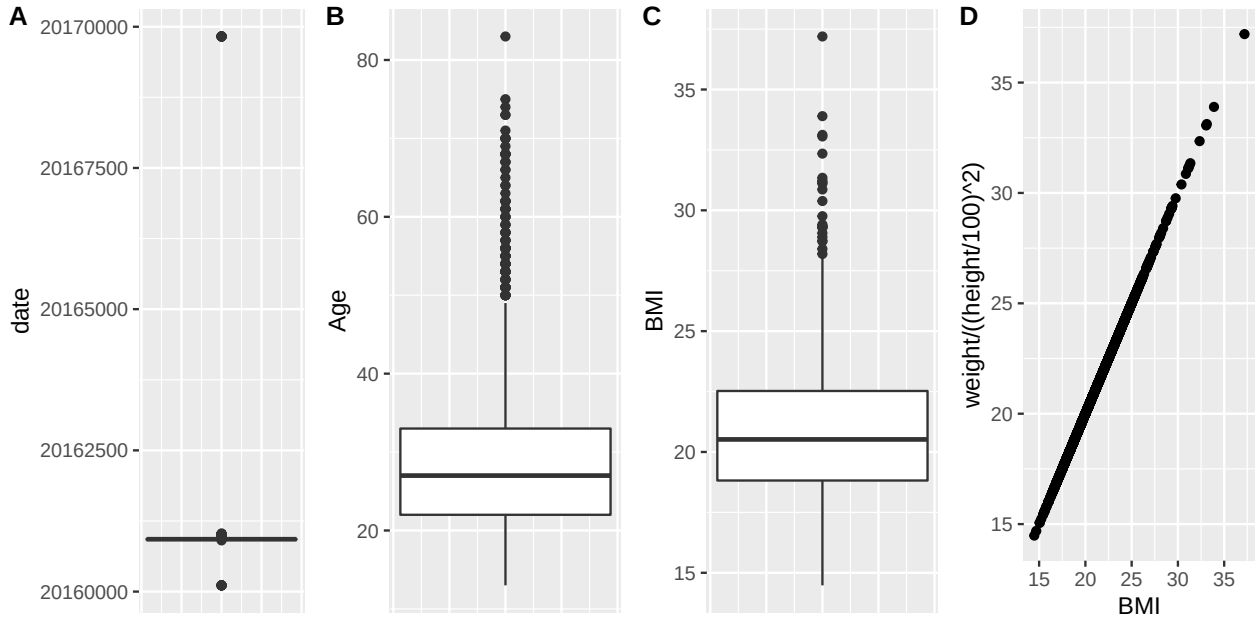


Figure 2: A-C) Distributions of important demographic variables in the raw data show outliers. We trimmed the outliers in the date, age and BMI. D) We can see BMI is corresponding well to BMI calculated from height and weight.

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.

Pairwise distributions

We checked the correlation among the numeric variables of the filtered dataset in Figure 3. 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 4. There were 20 missing cases for the quality score and 2 missing cases for the information score.

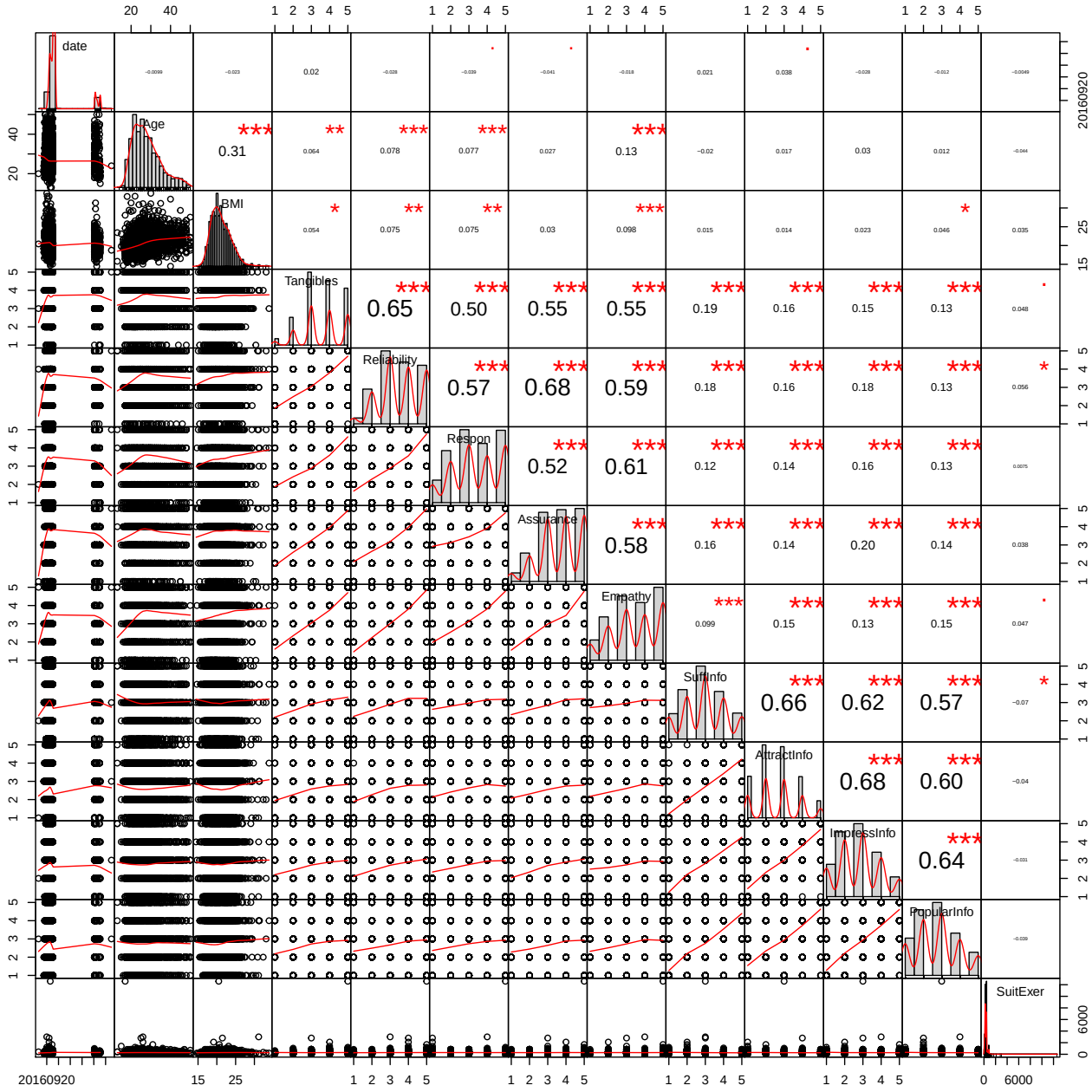


Figure 3: Pairwise correlations of numeric variables including their Pearson correlation coefficient in the top right quadrant. We observed that there doesn't appear to be collinearity among the numeric variables. Furthermore, age and BMI have a reasonable correlation. Additionally, we found that the 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).

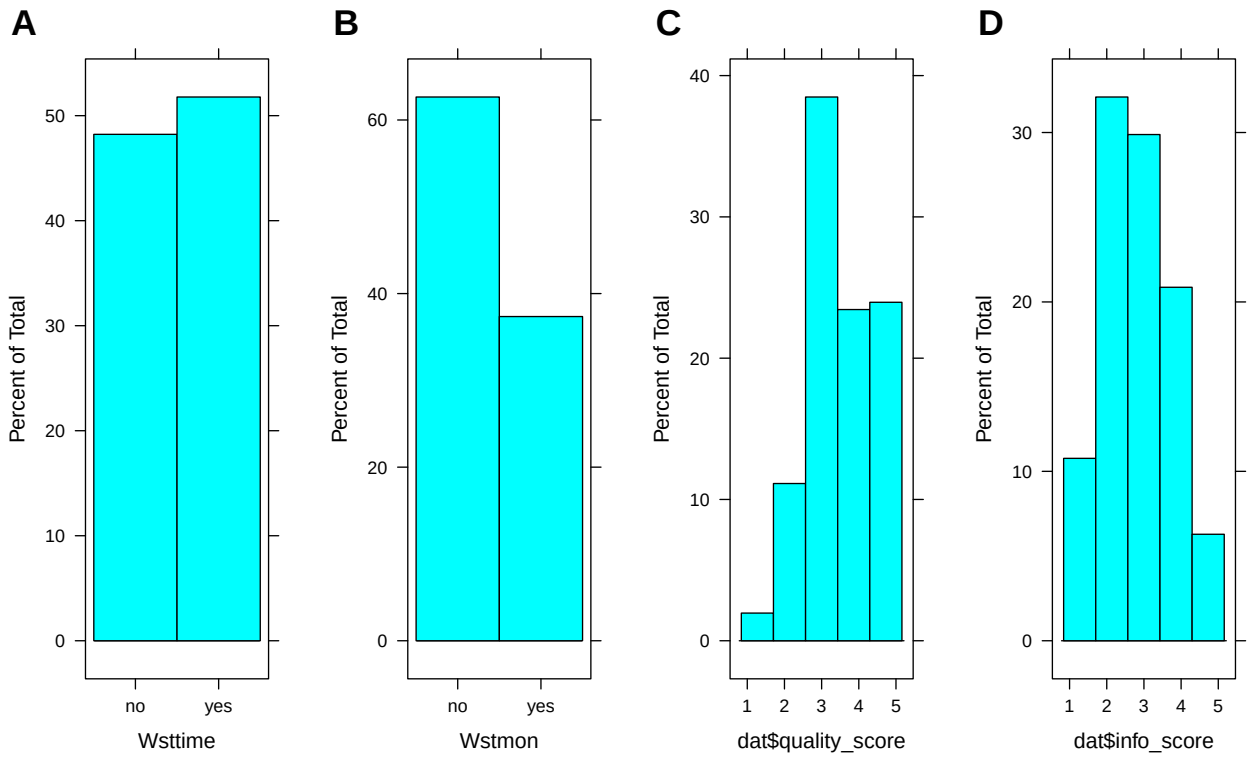


Figure 4: Perception of GHEs. A) About 50% of respondent believe check-ups are a waste of time. B) About 40% of respondent believe check-ups are a waste of money. C-D) The mean quality score is 3.54/5 and the mean information score is 2.82/5.

Introduction

Q1.

Overall, how do people rate the attractiveness, impressiveness, sufficiency, and popularity of information they receive in checkups? Give us some summaries of these variables, as well as variables like assurance, reliability, and empathy that tell us how well our doctors and nurses are doing, so we know how to improve.

Figure 4 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. The distributions across the diagonal in Figure 3 show summaries of the main rating variables. 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 4. 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.

Q2.

What factors make a person less likely to get check-up every twelve months? Find the most important factors that could help us design our advertising, and give us some measure of how

important they are.

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. Since the response variable is composed of discrete events, we used a binomial glm. However there is some class imbalance as there are about twice as many cases of less than 12 months than of the other two levels (data not shown).

First we picked the terms we should have in the best predictive linear model by using stepwise model selection with AIC. In order to reduce runtime, we used the quality and information scores as summaries of the underlying response ratings. Starting from these 39 predictors, the model summarized in Table 1 was produced with only 10 predictors. This suggests that aside from these 10 predictors, the others have a negligible effect size.

We evaluated how important each of the terms shown in Table 1 was by looking at their effect size and the significance level of the effect size. The most important predictors with effect size significant at $p < 0.05$ are the following, with the following interpretation of which effect sizes that are found significant:

- RecExam, all levels: when the participant had a checkup with symptoms of a disease, relative to patients who did between 12 and 24 months.
- ReaExam, request: the reason for their last exam, relative to a patient with worrying symptoms.
- FlwHealth: whether the respondent follows updates on their health measures, relative to one who does not.
- AcqTrmt: whether there is a member of the respondent's family receiving long-term medical treatment, relative to one who does not.
- MedCabinet: whether the respondent keeps a medical cabinet and basic medical equipment, relative to one who does not.
- UseMon: if respondent would use extra money to have a check-up, versus one who would soon.
- SuitFreq: if respondent believes check-ups should be done every 6 months, versus one who believes it should be done every 12 months.
- AfterIT: if respondent needed to have a check-up according to an app, probability that they would arrange one versus probability of someone who responded no.
- quality score: a combined score of their ratings of the quality of GHEs.

Q3.

Can we predict which people would be easiest to convince? That is, some people might be on the edge, and would get an exam with a little extra push; some people are very determined and would not get an exam no matter how hard you try. Using a classifier, can you find the patients who haven't gotten an exam but are most like other patients who have? Be sure to tell us how well your classifier works, so we know whether this is reliable.

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 teh package Ranger to predict RecPerExam. 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.

Table 1: Terms picked through stepwise selection in the binomial glm model of RecPerExam, or the time since the respondent last visited a doctor for an unprompted check-up.

	Dependent variable:
	RecPerExam
Constant	−122,666.100 (84,993.120)
date	0.006 (0.004)
RecExamg24	2.872^{***} (0.510)
RecExamless12	2.438^{***} (0.263)
RecExamunknow	1.942^{***} (0.560)
ReaExamnoti.disease	−0.296 (0.422)
ReaExamrequest	1.169^{***} (0.273)
ReaExamvolunteer	0.220 (0.266)
FlwHealthyes	-0.409^{*} (0.216)
AcqTrmtyes	-0.519^{**} (0.210)
MedCabinetyes	0.556^{**} (0.234)
UseMonlater	-0.590^{**} (0.248)
UseMonpartly	−0.099 (0.303)
SuitFreq18m	−0.207 (0.725)
SuitFreq6m	0.561^{**} (0.221)
SuitFreqg18m	0.011 (0.596)
AfterITno	0.752^{**} (0.348)
AfterITyes	0.064 (0.223)
quality_score	0.310^{***} (0.117)
Observations	1,014
Log Likelihood	−323.569
Akaike Inf. Crit.	685.138

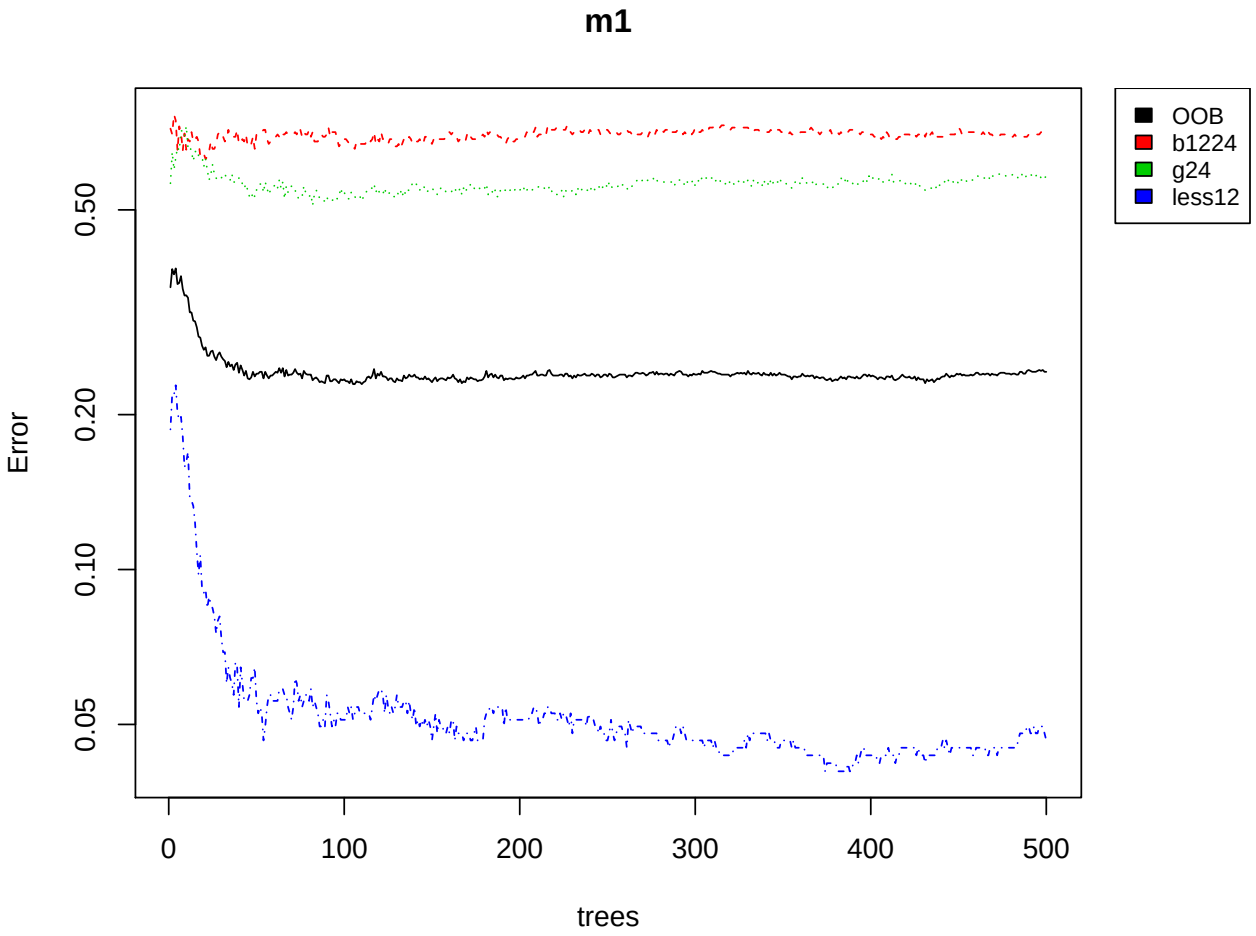


Figure 5: Plot of error of random forest model of RecPerExam versus the number of trees when using default parameters. OOB is the out of bag error. Levels of RecPerExam include less12 = less than 12 months, b1224 = between 12 and 24 months, g24 = over 24 months Performance on the cases with less than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

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.

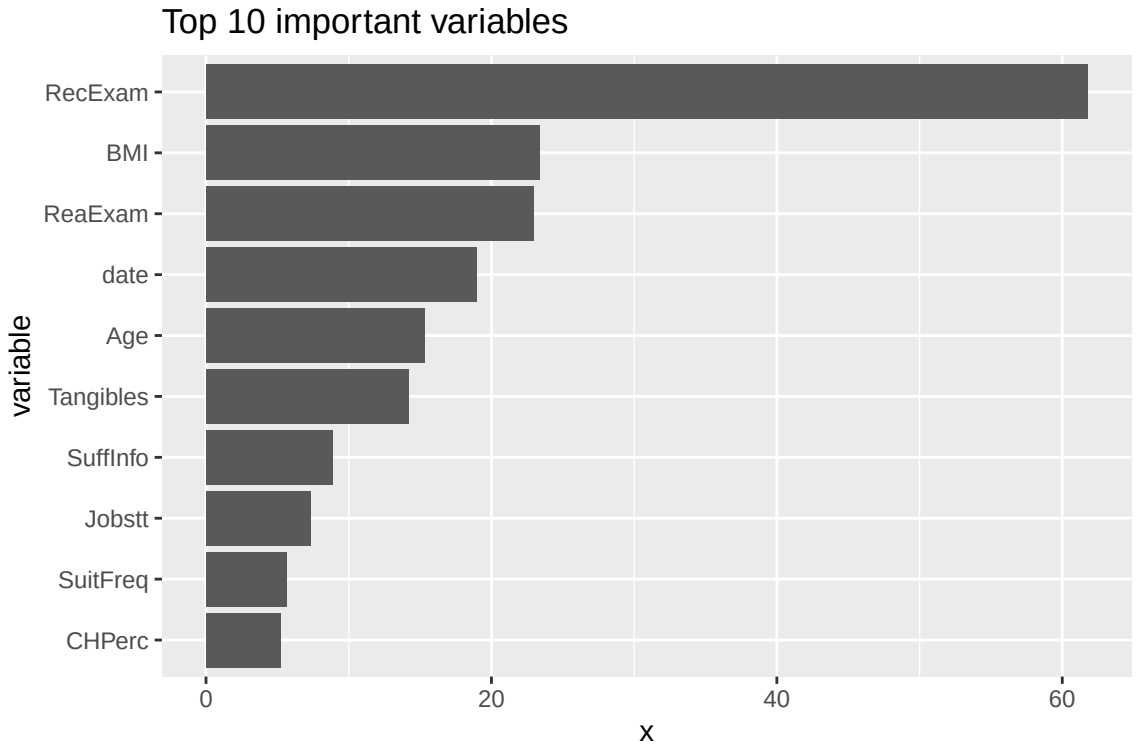


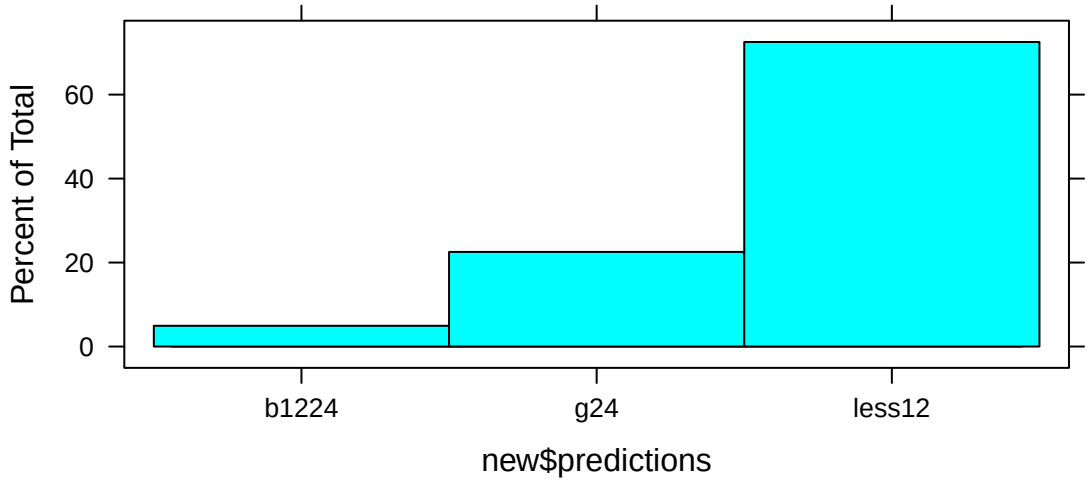
Figure 6: Variable importance of tuned random forest classifier.

[1] 48 [1] NA

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.



Conclusion

Bibliography

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