Fast Detection of Risk Signals in Post-marketing Drug Surveillance Using Costs in Claims Data: A Machine Learning Approach

Yihan Guan[1], Yiliang Jin[2]

December 9, 2010

Abstract

We investigate a machine learning approach to fast detection of risk signals in post-marketing drug surveillance using costs in health care insurance claims data. We show that by employing a locally weighted linear regression model to predict post-drug cost distribution of a population taking a well-known risky pain killer (Vioxx), the safety signal can be discovered four months earlier compared to a recent study using the same datasets. This project demonstrates the potential value of machine learning algorithms in improving real-time post-marketing drug surveillance.

1.Introduction

Traditional post-marketing drug surveillance systems using health care insurance claim data monitor procedure codes and diagnoses codes to detect adverse drug events (ADEs) [1,2,4,5]. One recent study [3] designed an active post-marketing drug surveillance system from a new angle─monitoring costs in health care insurance claims data, comparing post-treatment cost profiles of populations using alternative drugs based on risk-adjusted group-sequential analysis, and detecting increased spending related to drug side effects. They showed that signals of excess risks can be detected earlier by tracking costs than by tracking procedure/diagnoses codes. However, in that study, the expected cost distribution of the population under surveillance was constructed empirically rather than employing any learning method. Furthermore, the effectiveness of this empirical method has not been evaluated. Error in predicting post-treatment cost distributions could possibly delay the detection of a true risk signal. To address this concern, the present project explores a machine learning approach to find the best prediction model of the post-treatment cost of the population taking the drug under surveillance, which can potentially facilitate fast signal detection. The primary objective of this project is to find a supervised learning algorithm that works most effectively in predicting post-drug cost distribution. We further implement risk-adjusted group sequential analysis to evaluate whether the learning algorithm improves the timeliness of detecting a known risk signal.

2.Methods

2.1 Dataset Description and Features

In this project, two pre-processed health care insurance claims dataset are used. The first dataset corresponds to a population taking a top-selling and well-known safe pain killer, Naproxen. This dataset has cost information of 13,628 individuals. The second dataset corresponds to a population taking another top-selling pain killer,Vioxx. Vioxx was withdrawn from the market in September 2004 for its increased risk of heart attacks and strokes. This dataset has cost information of 6,014 individuals. Notice that these two datasets are mutually exclusive, i.e., members who took both drugs are excluded from both datasets. The detailed population selection rule can be found in [3]. An example of one line of record in our datasets is shown in Table 1.

Member ID	Drug-start date	Gender	Age	Pre-drug monthly cost[3] ($)	Post-drug monthly cost ($)[4]
XXXXX	3/23/2004	Male	41	118	1506

Table 1. An example of a record in the datasets

Features of each member are gender, age, and pre-drug monthly cost. To categorize a member's post-drug cost as either high or low, we adopt the following decision boundary developed in [3]:

$$\text{High post-drug cost} = \mathbf{1}\{\text{post-drug cost} > \text{Max}(800, 2*\text{pre-drug cost})\}$$

[1] Department of Management Science &Engineering. Email: yihan@stanford.edu

[2] Department of Computer Science. Email: yljin@stanford.edu

[3] Pre-drug cost is defined to be a member's average monthly cost during the six months before his/her first prescription of the drug.

[4] Post-drug cost is defined to be a member's average monthly cost during the six months after his/her first prescription of the drug.

2.2 Supervised Learning Algorithms to Predict Post-drug Cost Distribution

2.2.1 Classification

One way to obtain the post-drug cost distribution is to first apply a classification algorithm to predict the binary post-drug cost (1=high, 0=low) for each member in the test set, where gender, age, and pre-drug cost are used as input features. As a next step, members with post-drug cost labels are stratified into risk groups according to their gender, age and pre-drug cost. Then we obtain the distribution of high post-drug cost members in each risk group, which will be used in group sequential hypothesis testing. Figure 1 illustrates the procedures in using classification algorithms.

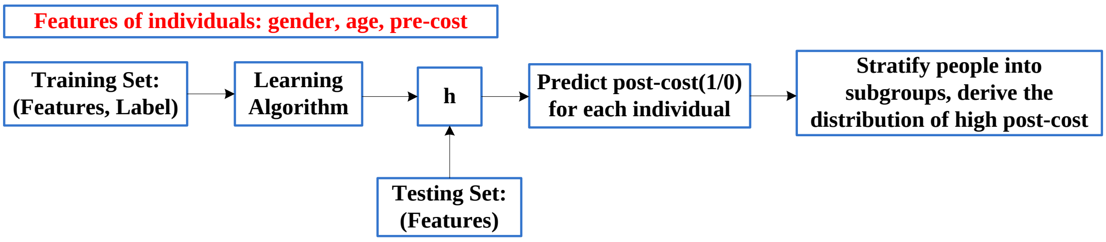
Features of individuals: gender, age, pre-cost
Training Set: (Features, Label)
Learning Algorithm
h
Predict post-cost(1/0) for each individual
Stratify people into subgroups, derive the distribution of high post-cost
Testing Set: (Features)

Figure 1. Procedures in using classification algorithms

Specifically, features are defined as the following:

- Gender = {0,1} (0=Male,1= Female);
- Age = {1,2,3,…,10} (1=1-10 year-old, 2=11-20 year-old, 3=21-30 year-old,…,10=91 year-old or above);
- Pre-drug cost is discritized into 8 buckets: [0,50], (50,100], (100,150], (150,200], (200,400], (400,600], (600,800], and (800,1000][5], so that the number of members in each bucket is roughly balanced.

2.2.2 Linear/Non-linear Regression Algorithms

An alternative way to obtain the post-drug cost distribution is to use a regression algorithm. Specifically, we first stratify members into risk groups according to their individual features and obtain an ID of each risk group as the group feature. Next, a regression algorithm is applied to predict the probability of each group having high post-drug cost. Figure 2 illustrates the procedures of this method.

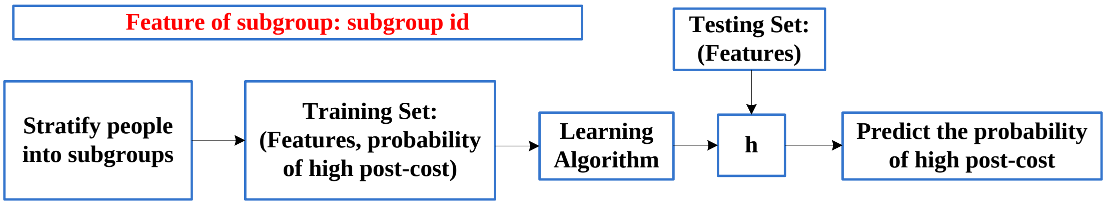
Feature of subgroup: subgroup id
Testing Set: (Features)
Stratify people into subgroups
Training Set: (Features, probability of high post-cost)
Learning Algorithm
h
Predict the probability of high post-cost

Figure 2. Procedures in using regression algorithms

3. Learning Algorithm Selection

To find a suitable supervised learning algorithm to accurately predict the post-drug cost distribution, we experiment several algorithms on the first dataset. Hold-out cross validation (70% of the dataset is used as a training set, and 30% of the dataset is used as a test set) is used to compare the performances of different algorithms.

3.1 Classification Algorithms

3.1.1 Logistic regression and Gaussian discriminant analysis(GDA)

For logistic regression and Gaussian discriminant analysis (linear boundary case), the input features are defined as a vector

[5] Excluded from the datasets are members with pre-drug costs over $1000. The rationale for this exclusion is that members with high pre-drug cost have considerable pre-treatment conditions, and thus it could be argued that the effects under study are due to confounding factors rather than the treatment drug. In addition, these members show high variance in their health service utilization, which can translate into unstable statistical estimates [3].

$X = (x_1\ x_2\ x_3)^T$, where x_1, x_2, x_3 are gender, age, and pre-drug cost, respectively.

3.1.2 Naive Bayes and support vector machine (SVM) with linear kernel

To apply Naive Bayes with Laplace smoothing and SVM algorithm, the input feature vector $X \in R^{20}$ can be written as the following, based on the definitions of features in Section 2.2.1:

$$X = \left(\overbrace{\underbrace{\boxed{1}\boxed{0}}_{2}}^{\text{gender}} \overbrace{\underbrace{\boxed{0}\boxed{1}\boxed{\;\cdots\;}\boxed{0}\boxed{0}}_{10}}^{\text{age}} \overbrace{\underbrace{\boxed{0}\boxed{0}\boxed{\;\cdots\;}\boxed{1}}_{8}}^{\text{pre-cost}} \right)^T$$

3.1.3 Hold-out Cross Validation

In this section, we use hold-out cross validation to compare the performance of the above classification algorithms in predicting post-drug cost distribution. The training set consists of 70% of the dataset (randomly selected), and test set consists of the rest 30% of the data set. The results of hold-out cross validation is shown in Table 2.

Learning Algorithm	Logistic regression	GDA	Naive Bayes	SVM
Hold-out Cross validation error	11.44%	11.44%	11.58%	11.62%

Table 2. Cross validation result of four classification algorithms

The learning curves of logistic regression and GDA are plotted in Figure 3. We observe that these two prediction models both have high bias, which is likely to be caused by the fact that there are too few features. However, due to the limitation of the current dataset, more features are not accessible to us.

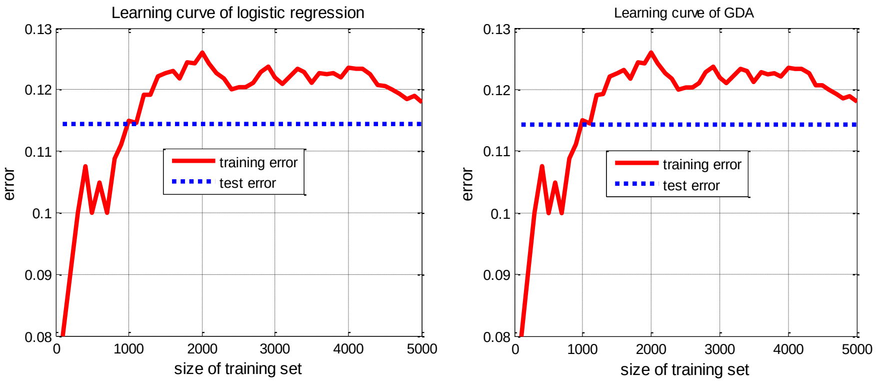

Figure 3. Learning curves of logistic regression and GDA

3.2 Linear/Non-linear Regression

To tackle the problem of overly high bias, we consider applying linear and nonlinear regression algorithms. Since the ultimate goal of this project is to detect the risk signal in a series of group sequential hypothesis testing which only requires to know the predicted distribution of high post-drug cost in risk groups, linear/non-linear regression can be utilized to predict the probability of each risk group having high post-drug cost. Namely, we focus on each risk group, instead of predicting each individual's post-drug cost as in Section 3.1.

3.2.1 Regression Models

As described in Section 2.2.2, the input feature is the ID of risk groups. Unweighted regression and weighted regression are experimented, respectively.

- *Unweighted regression*. Linear, quadratic, cubic and 4th order polynomial models are considered to fit the training set.
- *Weighted regression.* Locally weighted linear regression is used to fit the training set, where the local weight is defined as:

$$w^{(i)} = \exp\left(-\frac{(x - x^{(i)})^2}{2\tau^2}\right),$$

where τ is set to be 0.8.

3.2.2 Reducing variance: choosing an appropriate number of risk groups

As described in Section 2.2.1, the entire population can be stratified into 160 (=2*8*10) risk groups. The left panel in Figure 4 depicts the probability of having high post-drug cost of each risk group given by unweighted regression models, and the right panel plots the fit given by weighted linear regression model. Evidently, high variance appears in both the training data and the test data under all regression models that are considered. This comes from the fact that many risk groups only contain few people (less than 5), which makes the estimates not reliable.

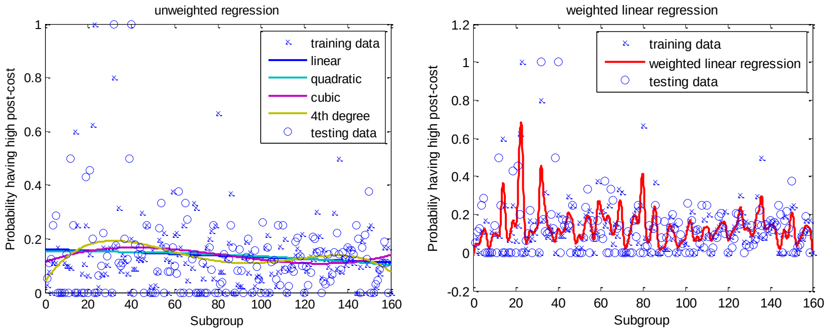

Figure 4. Probability of having high post-drug cost in 160 risk groups

Therefore, to ensure each group has a good sample size, we combine all age groups, namely we only use gender and pre-drug cost to define risk groups and get 16 risk groups (eight female and eight male groups). The left panel of Figure 5 demonstrates the fit of the eight female groups using unweighted regression models, and the right panel shows the fit of the same groups using the weighted linear regression model. It can be seen that the weighted linear regression and the 4th order polynomial regression model fit the training set well. On the other hand, the weighted linear regression fits the testing set best.

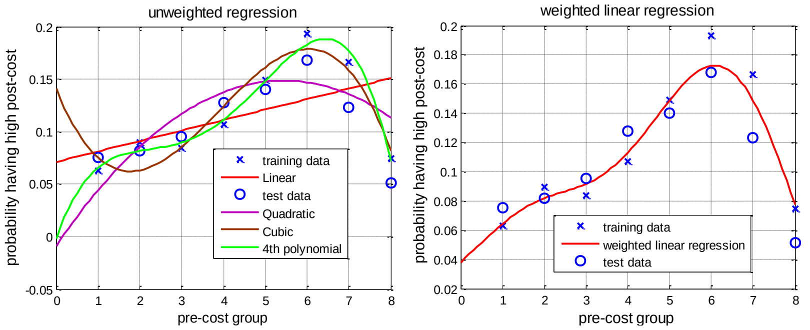

Figure 5. Probability of having high post-drug cost in eight female risk groups

3.2.3 Choosing the best prediction model: hold-out cross validation

To choose the one that performs the best, we compare five regression models via hold-out cross validation. Table 3 implies that the locally weighted linear regression has the smallest variance both for female and male groups. Hence, we choose this model to construct the expected post-drug cost distribution of Vioxx group[6] and implement group sequential analysis to see how early the sequential tests can raise the signal of excess spending in Vioxx group.

			Male	Female
Variance	Locally weighted linear regression		3.482×10^{-3}	8.704×10^{-4}
	Unweighted regression	Linear	4.055×10^{-3}	6.189×10^{-3}
		Quadratic	4.004×10^{-3}	3.050×10^{-3}
		Cubic	3.760×10^{-3}	1.581×10^{-3}
		4th order polynomial	3.488×10^{-3}	1.970×10^{-3}

Table 3. Cross validation result of regression models

[6] The null hypothesis of the sequential tests is that Vioxx group has the same post-drug cost distribution as Naproxen group. Hence Naproxen group serves as a baseline, whose post-drug cost distribution is regarded as the expected post-drug cost distribution of Vioxx group.

4. Group Sequential Analysis

Detailed descriptions of the group sequential analysis can be found in [3]. Briefly, 37 monthly hypothesis tests are conducted on accumulating Vioxx data from July, 1st, 2001 to July 1st, 2004. At each month (t), the p-value of a Chi-square goodness of fit test p-value(t) is compared to a significance level α(t) which is given by *alpha spending function approach* [6] so that the overall significance level of these 37 tests are controlled at 0.05. A signal is detected at t once p-value(t) falls below α(t). Using the locally weighted linear regression model to calculate the expected post-drug cost distribution of Vioxx group, we detect the risk signal in the 26th month since the testing starts. In the previous work where an empirical method was used to calculate the expected post-drug cost distribution of Vioxx group [3], the signal was found in the 30th month. Figure 6 compares the group sequential testing results of the present work (blue dashed curve) to those of the previous work (green dashed curve). Consequently, the locally weighted linear regression model speeds up the signal detection by 4 months, which is a remarkable improvement in the context of post-marketing drug surveillance.

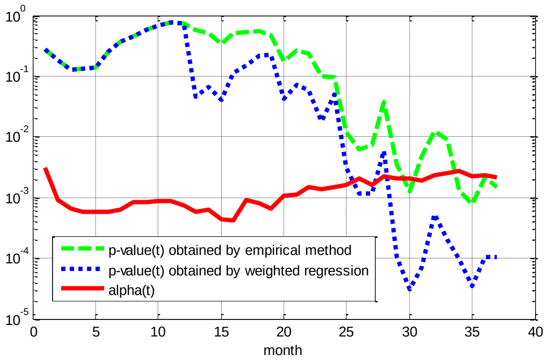

Figure 6. Group sequential testing results

5. Conclusion

This project demonstrates the potential value of machine learning algorithms in improving real-time post-marketing drug surveillance. We show that by employing a locally weighted linear regression model to predict post-drug cost distribution of the population taking a risky drug, the safety signal can be detected considerably faster compared to a recent study using the same datasets. Rapid signals detected by our method can trigger timely investigation for underlying reasons of excess spending. If the excess spending is indeed caused by adverse drug events, our method can potentially save lives and reduce health care costs.

References

1. J. A. Berlin, S. C. Glasser, and S. S. Ellenberg. Adverse event detection in drug development: recommendations and obligations beyond phase 3. *Am J Public Health*, 98(8):1366-1371, Aug 2008.
2. M. V. Bjarnadottir. Data-driven approach to health care: applications using claims data. Ph.D. thesis, Massachusetts Institute of Technology, 2008.
3. M.V. Bjarnadottir, Y. Guan. Follow the money: monitoring cost in health care claims data for real-time post marketing drug surveillance. Working paper, 2010.
4. J. S. Brown, M. Kulldorff, K. A. Chan, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiol Drug Saf*, 16(12):1275-1284, 2007.
5. J. S. Brown, M. Kulldorff, K. R. Petronis, et al. Early adverse drug event signal detection within population-based health networks using sequential methods: key methodologic considerations. *Pharmacoepidemiol Drug Saf*, 18(3):226-234, 2009.
6. D. L. Demets and K. K. G. Lan. Interim analysis: The alpha spending function approach. *Stat Med*, 13(13-14):1341-1352, 1994.