EENG 436 - Biomedical Informatics

Final Project:

Detecting Adverse

Drug Reaction Signals

Dan Rahm

Nate Wendt

Jhobany Tortolero

December 7, 2012

**Introduction**

Medicine is one of the most necessary fields for the good of the world’s inhabitants. Given that people’s lives are constantly at stake, it should come as no surprise that the medical world constantly strives to improve the effectiveness of drugs and treatments. To do this, data is collected from various sources and compiled into databases within medical information systems. Computers are then utilized to sift through these large quantities of information in order to determine whether a drug is benign to public health or not. This process is a critical part of the medical world since inadequate correlations between drugs and adverse drug effects could lead to the continuation of ineffective drugs and the discontinuation of effective ones.

Among the data collected, are the Adverse Drug Reports (ADRs) compiled by the FDA. These reports include information about patient demographics, drugs taken, reactions observed – all useful in determining the benevolence of the reported drugs when analyzed in large quantities. The purpose of this project is to implement a medical information system that uses a data mining methodology to detect possible signals from spontaneous reports on ADRs. The system was designed with the following specifications in mind.

Specifications:

1. The system shall be a medical information system through which users can check at least three things:
   1. Characteristics of the patient data (i.e., total number of patients, average age, standard deviation of age, percentage of male and female)
   2. The degree of association between a particular drug and an adverse drug reaction (ADR)
   3. A list of top (e.g. 10) pairs with high degree of associations according to your algorithm.
2. It interacts with a database that contains FDA’s spontaneous reports.
3. A web-based interface should be designed in order to help users utilize the system.
4. The system should be thoroughly tested using the FDA’s spontaneous reports.
5. The system’s ability in detecting ADR signals should be analyzed and reported.

**Algorithm**

For the purpose of this project, five different algorithms were given as options. After reviewing their descriptions and uses, the team resolved to make use of the Bayesian Confidence Propagation Neural Network or BCPNN. In their 1998 publication, Bate et. Al, state that this computational approach to data mining has been developed to handle large data sets and find new drug-ADR signals. The authors then go on to describe the inner workings of this data mining methodology.

At first glance, the BCPNN gives the impression of being ruthlessly complicated. Not only can it take into account the four different states for two binary independent variables, but it can also derive the variance associated with the resulting degree of association. After much analysis however, it was decided that the most important piece of information from the 1998 publication is the information component (IC):

where x represents a specific drug and y is a specific reaction; P(x,y) is the coincident probability (the probability that both the specific drug and reaction are contained within the same report); P(x) is the prior probability of the drug (the probability that a specific drug is present within a report); and P(y) is the prior probability of the reaction.

Deriving from Bayes’ law, the IC is the strength of the association between the specific states of two mutually independent variables. For this particular case, the two variables are the drug and the reaction and they have binary properties with regards to either being present or not within any given spontaneous report. There are four possible ICs for any two of these variables; however, for the purposes of drug-ADR association, it is only necessary to look at the instance when both are present. Furthermore, since the system specification does not require calculation of confidence values for the probability associations, the algorithm simplified with respect to the beta distributions. Only the mean value is necessary from the distribution, and thus, a simple calculation of parts divided by whole can replace the complex distribution.

In order to attain P(x) and P(y), running counts must be made of the specified drug and reaction. Additionally, a total count of the total ADRs must be logged within the system along with the count of P(x,y) for each drug-reaction pair. From this information, the IC value can be determined. The computation simplifies to where a higher value of IC corresponds to a higher degree of association between the given drug and reaction while a lower IC represents the opposite. We have normalized the IC values to correspond to a scale that is easier to understand for the frontend user. The association rules are as follows:

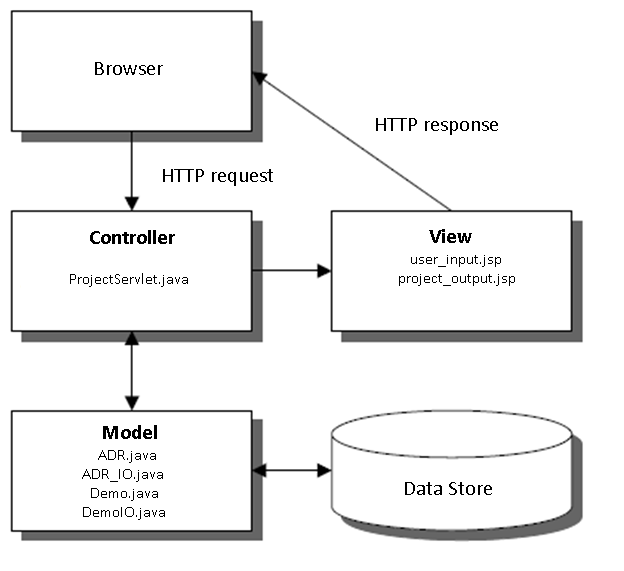
IC < 5 Low association

IC = 9 Medium association

IC > 12 High association

**System Design**

The Model View Controller (MVC) template helped to simplify the layout of the system. Using this design, the team was able to isolate the areas of work into each of the three categories. As a result, the application consists of only a few things: two front-end JSP pages, which make up the View; one servlet, which makes up the Controller; and four java classes, which make up the Model aspect. **Figure 1** (below) shows a visual representation of the suggested project MVC model.



**Figure 1 – Project block diagram.**

**Implementation**

Accessible from any browser, the user\_input.jsp page is the welcome page of the application. This page prompts the user to choose at least one of the three uses of the application as described in the specifications section, and click the submit button. The submit button then calls on the servlet ProjectServlet.java. At this point, the servlet makes sure the user has entered the necessary information to use the system and makes one of two decisions: 1) if the user has not entered the necessary information, the servlet dispatcher sends the user back to the user\_input.jsp file and displays a message according to the type of information needed, or 2) if the user has entered all information adequately, the servlet calls on the files within the Model and sets the dispatcher towards the project\_output.jsp file.

The files in the Model make use of the JDBCWrapper files to sift through and organize the data in the FDA database. For the characteristics of patient data, the classes Demo and DemoIO are used to return all the information mentioned in the Specifications section of the introduction. For the degree of association and the list of pairs with highest association, the system makes use of the ADR and ADR\_IO java classes then returns the requested information from the data store to the servlet, which in turn forwards it to the project\_output.jsp file. This jsp file is displayed in the browser with all the requested information.

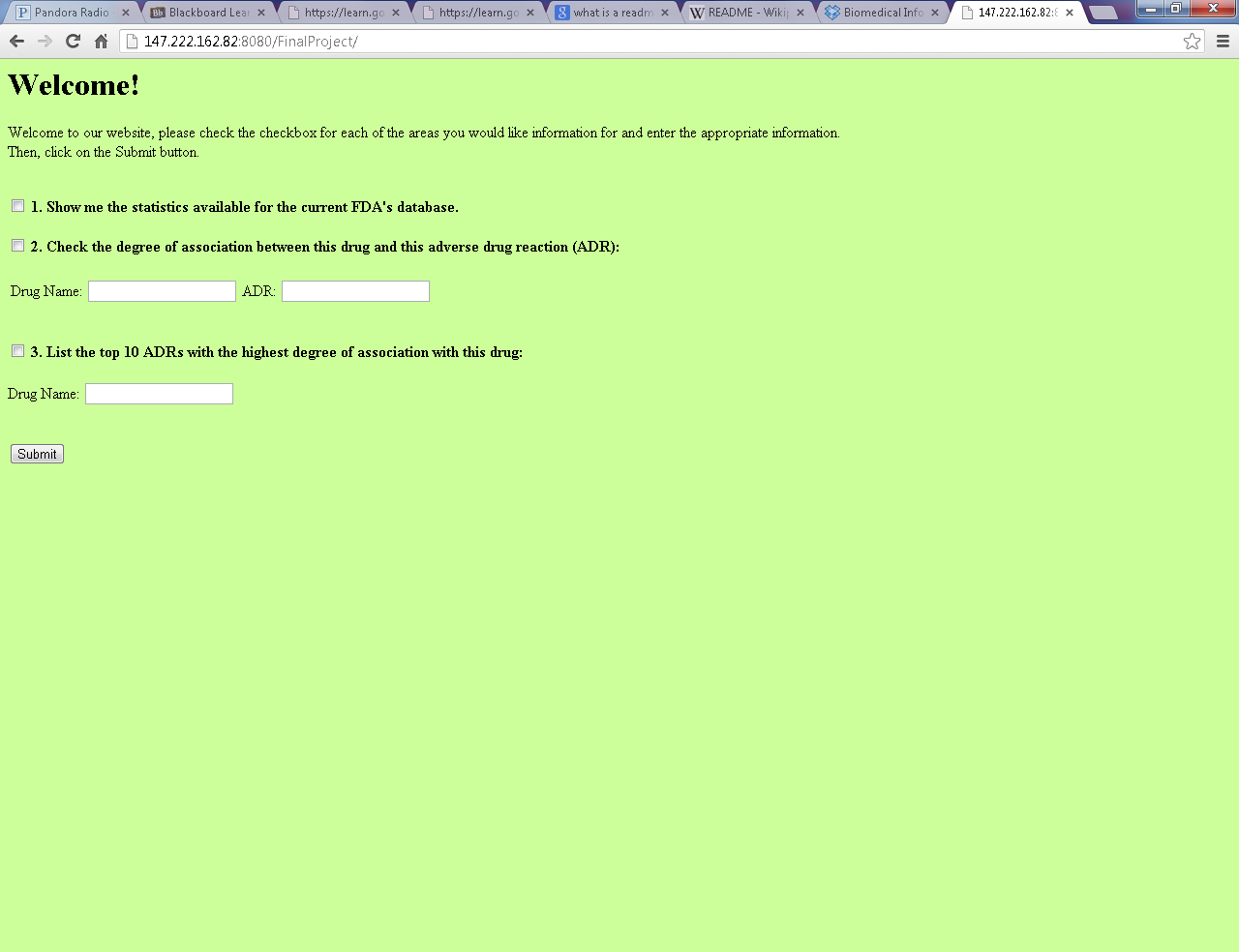
**Testing**

In order to ensure proper functionality of the system, the team tested and debugged the source code at various points throughout the development of the application.

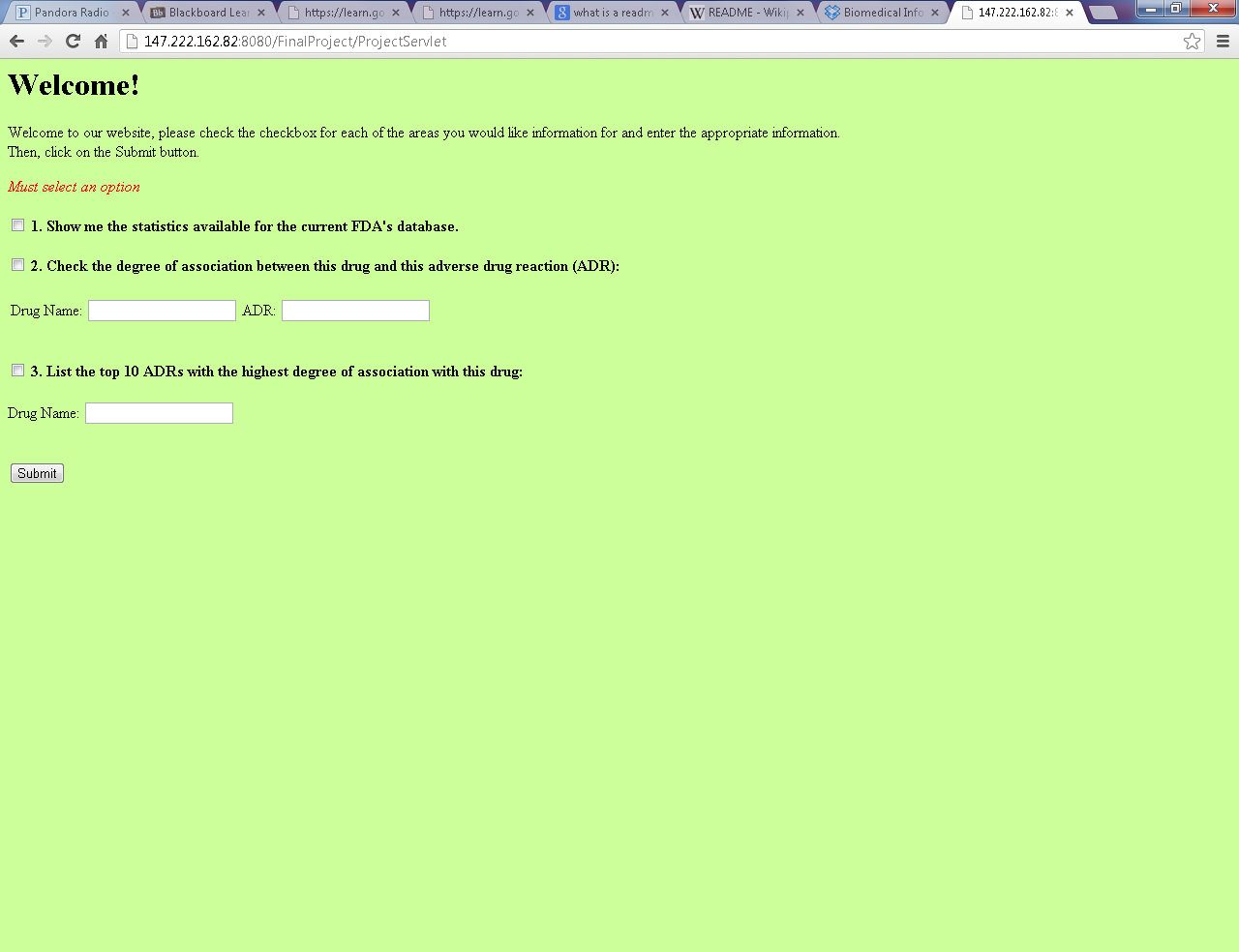
To test the algorithm, it was necessary to run both NetBeans and MySQL Workbench simultaneously and check that each program yielded the same results.

To test the input webpage, the servlet had to be developed and different scenarios were simulated in the browser to make sure the messages were displayed correctly:

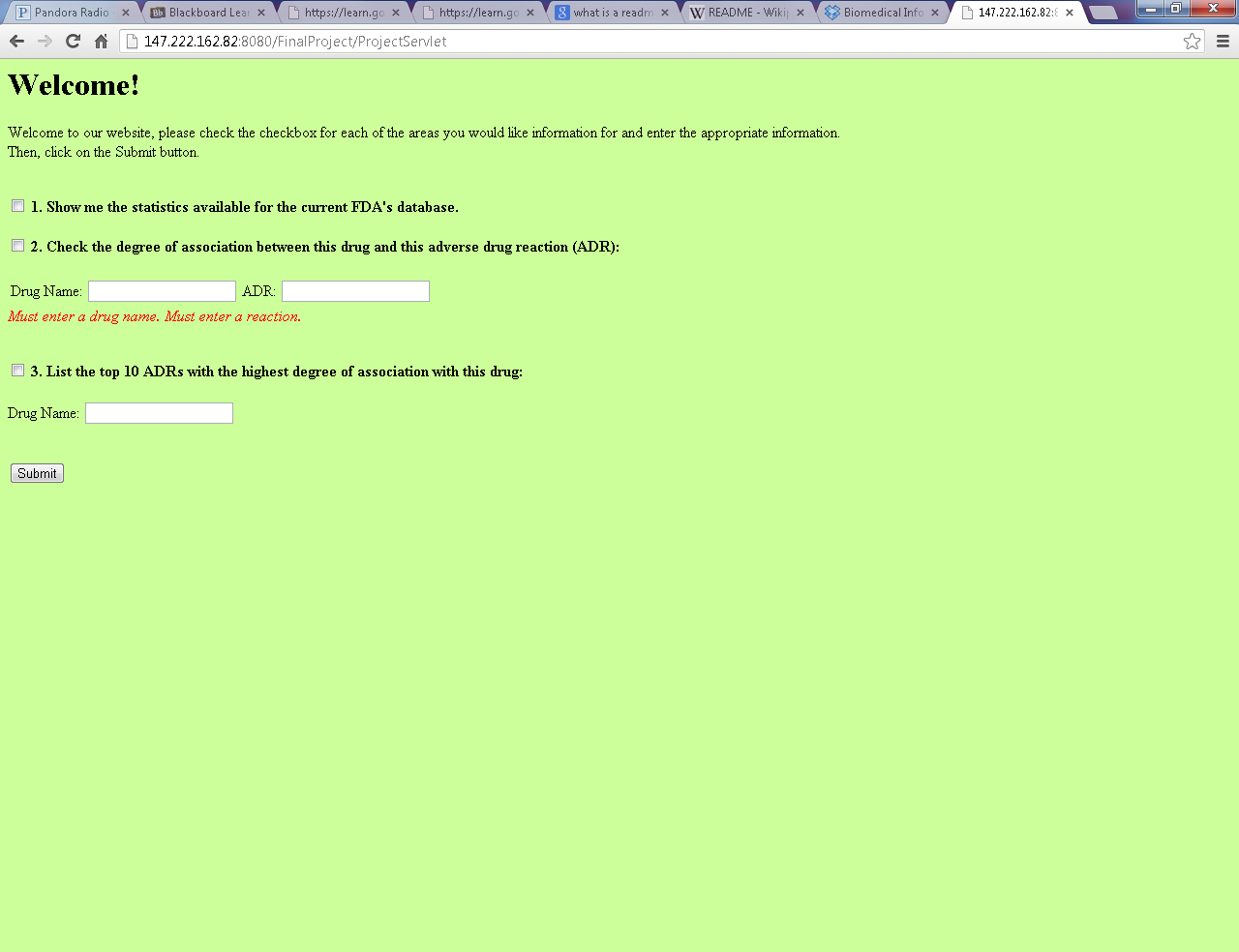
Error Checking Scenarios:

****

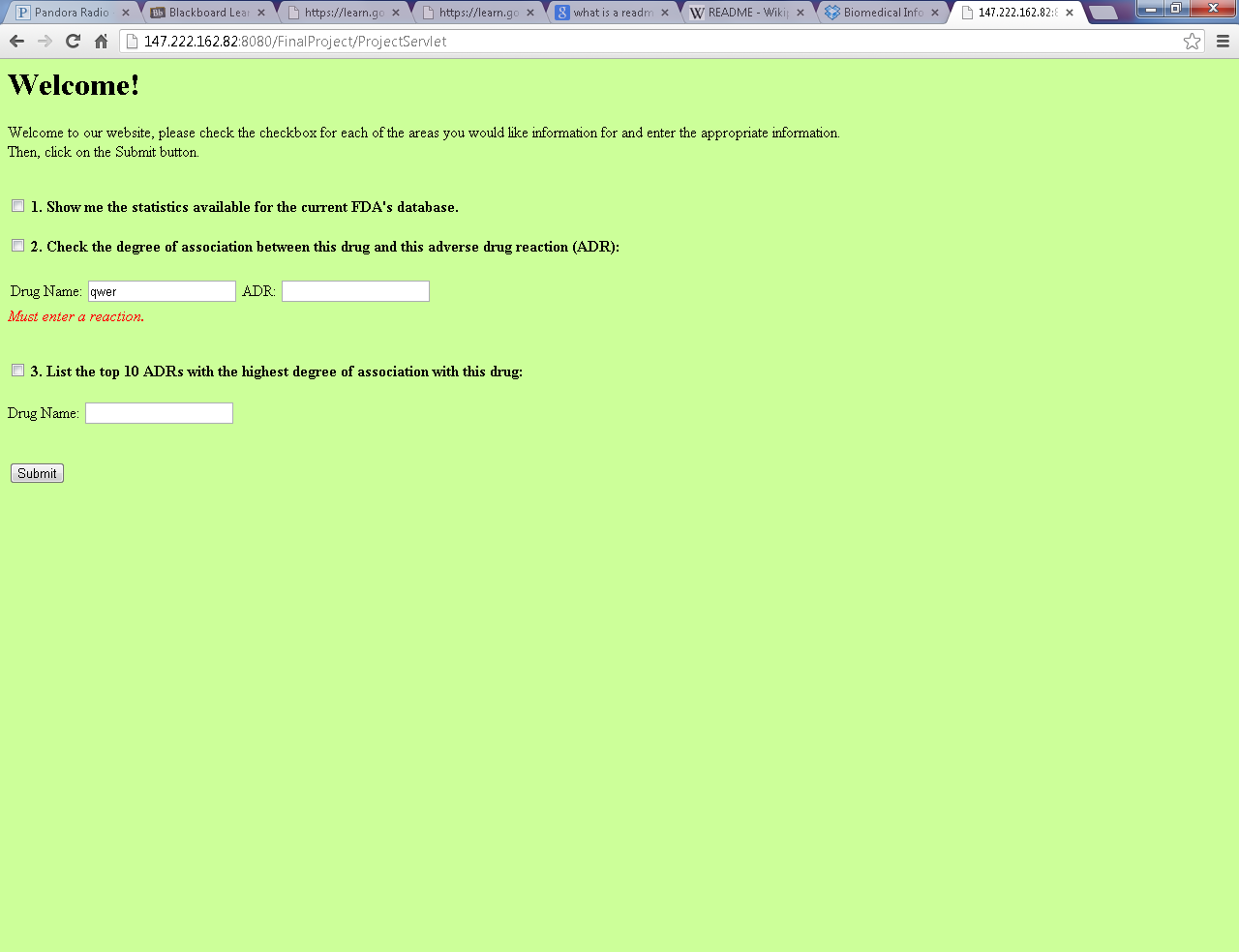
**Figure 2 – Scenario 1**

****

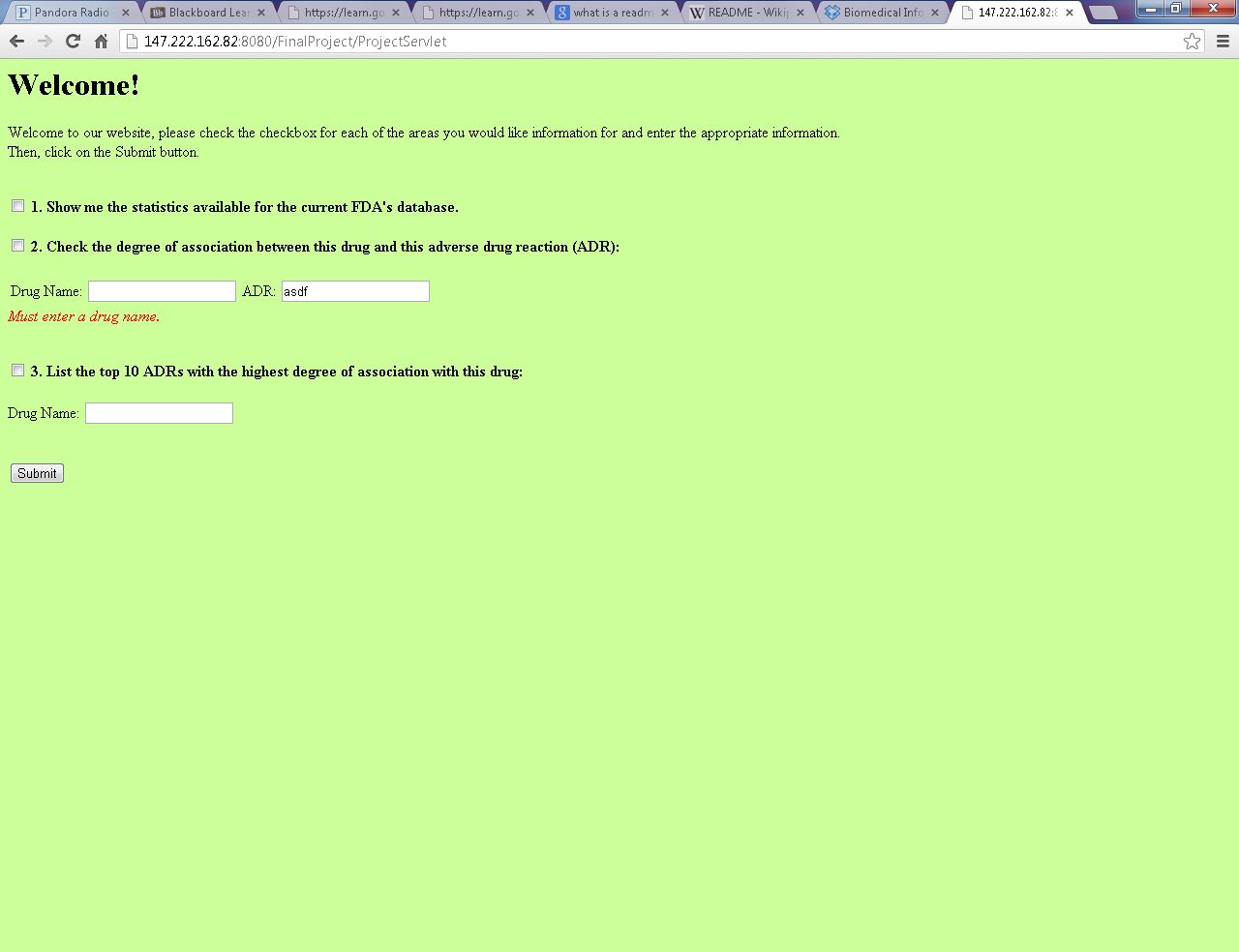
**Figure 3 – Scenario 2**

****

**Figure 4 – Scenario 3**

****

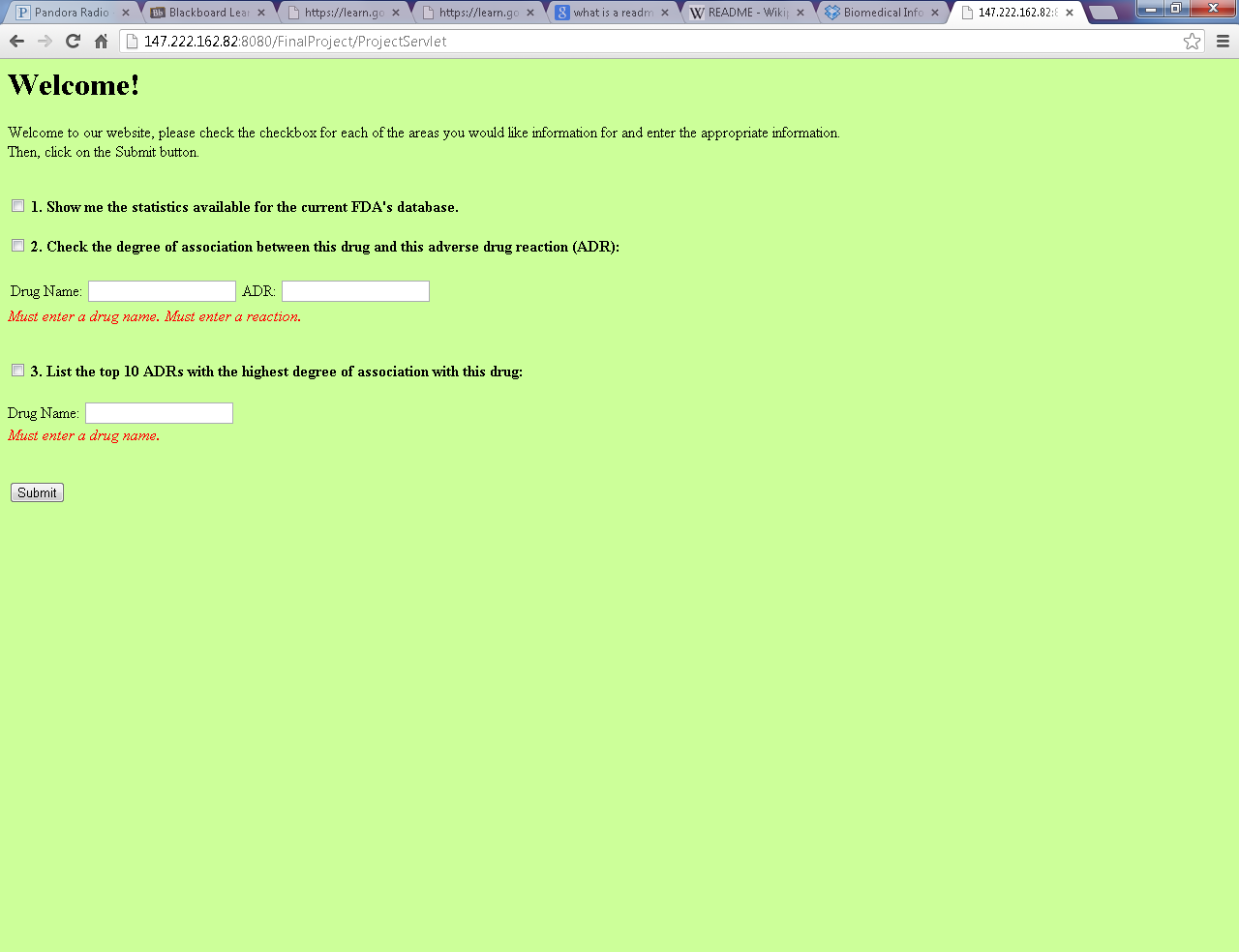
**Figure 5 – Scenario 4**

****

**Figure 6 – Scenario 5**

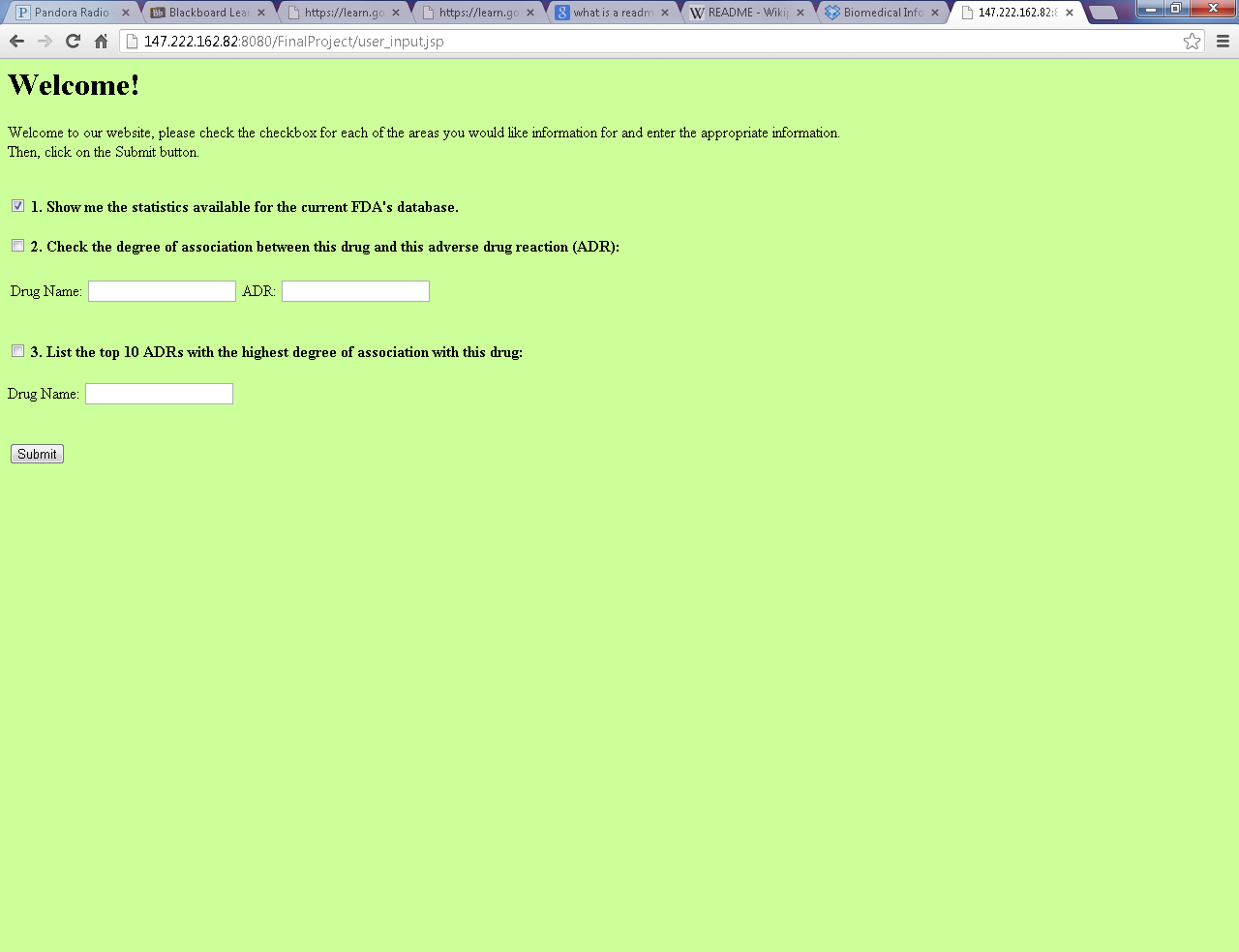
****

**Figure 7 – Scenario 6**

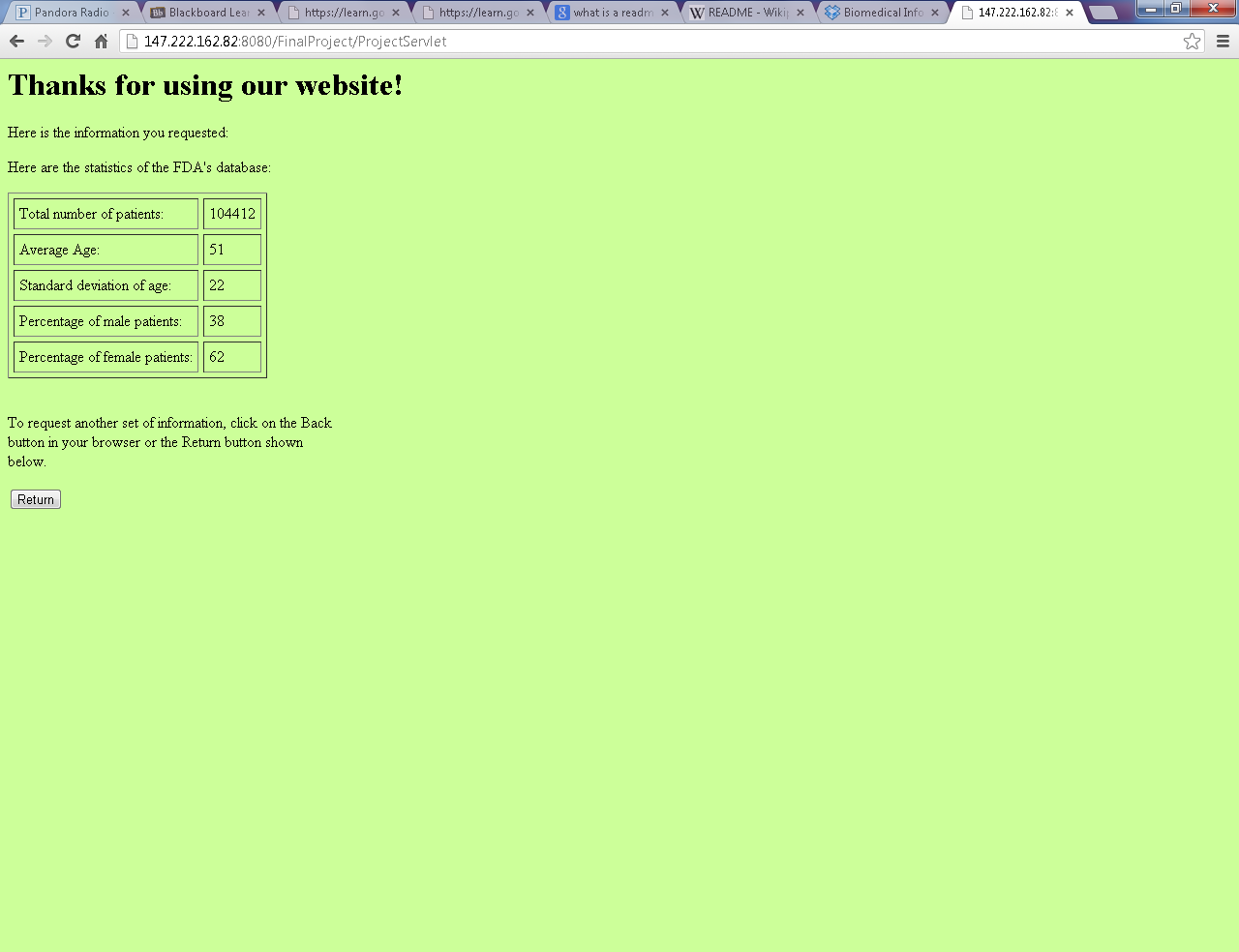
****

**Figure 8 – Scenario 7**

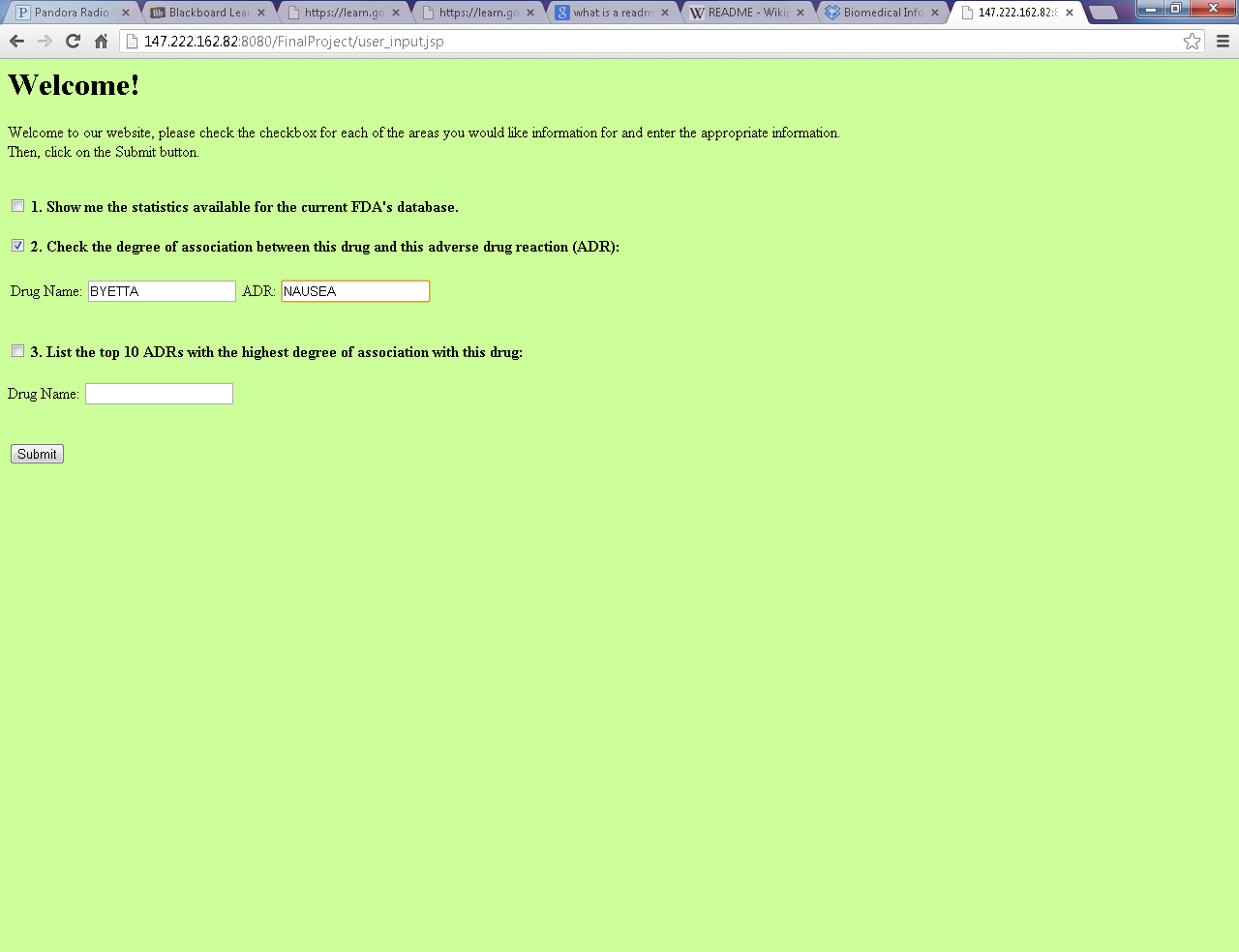
To test the display webpage, many different scenarios were simulated once again.

****

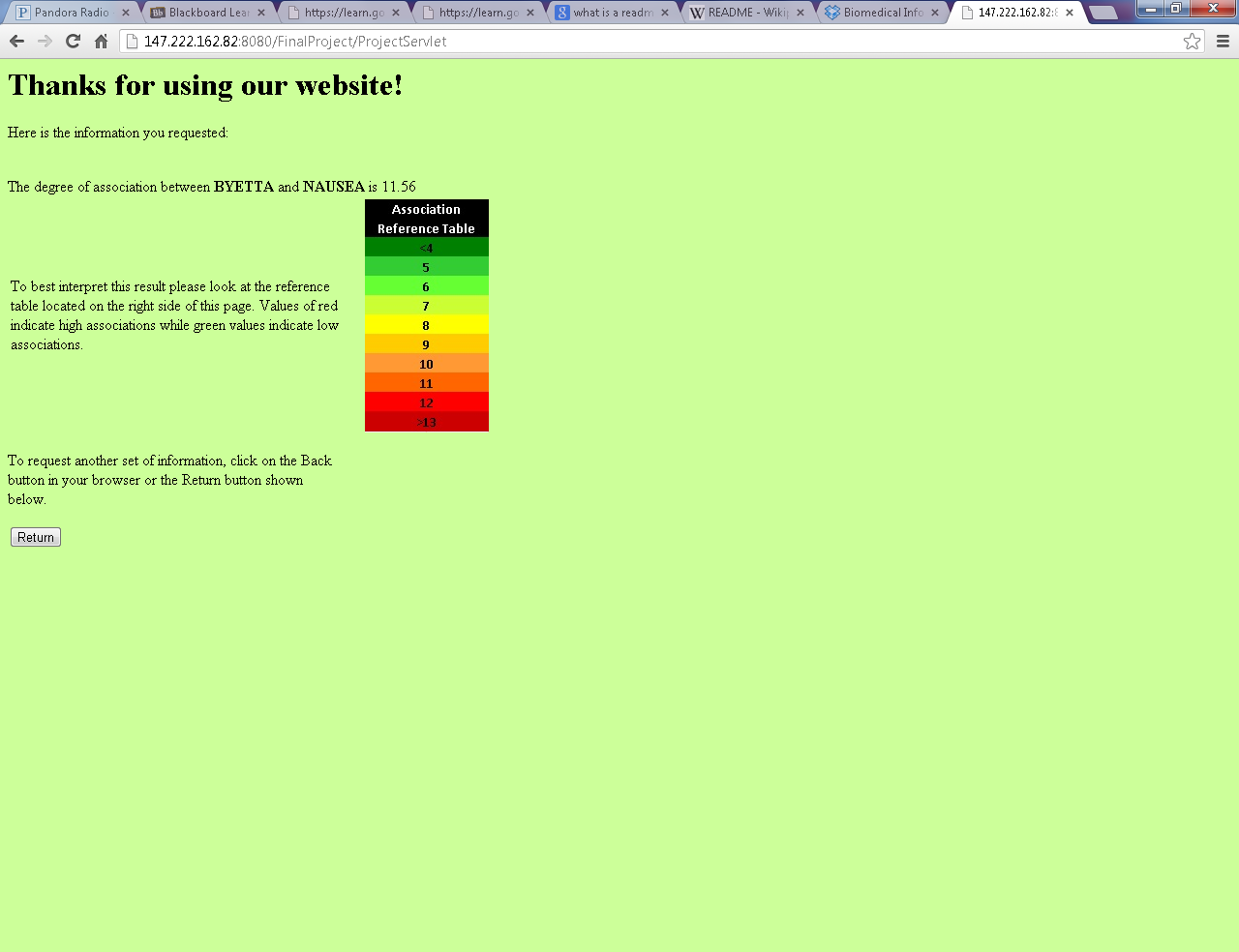
**Figure 9 – Statistics checked.**

****

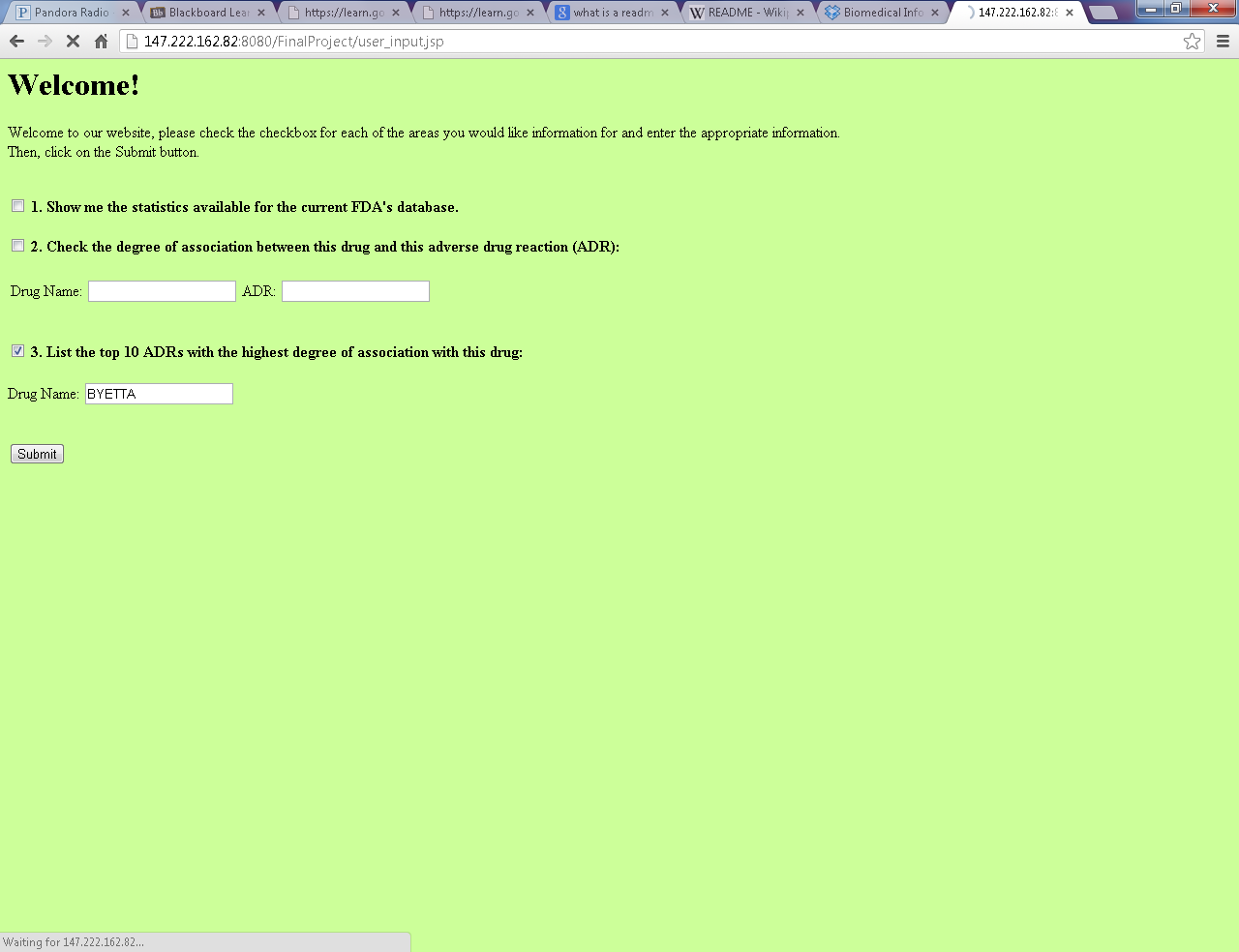
**Figure 10 – Statistics Results**

****

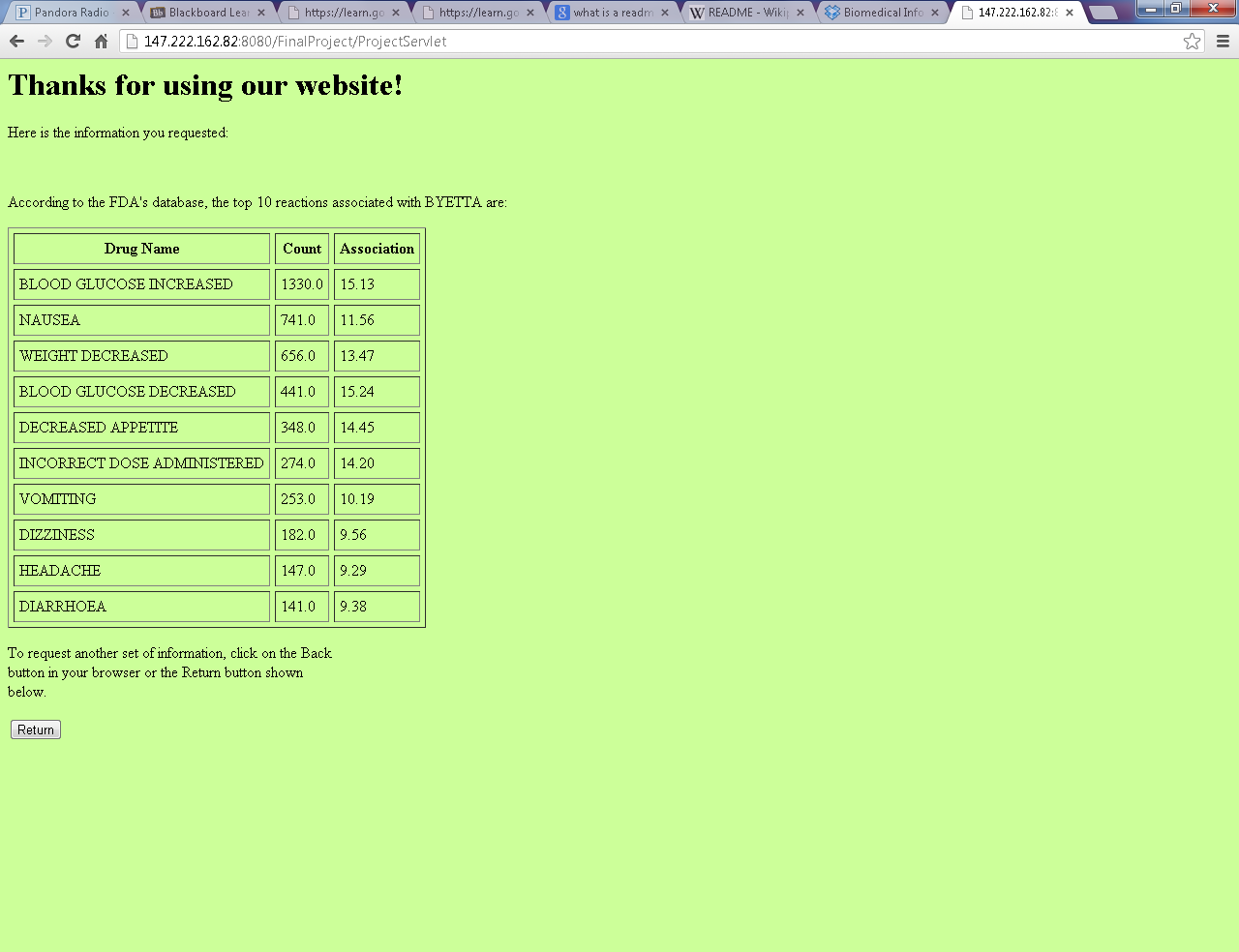
**Figure 11 – Checking association between BYETTA and NAUSEA.**

****

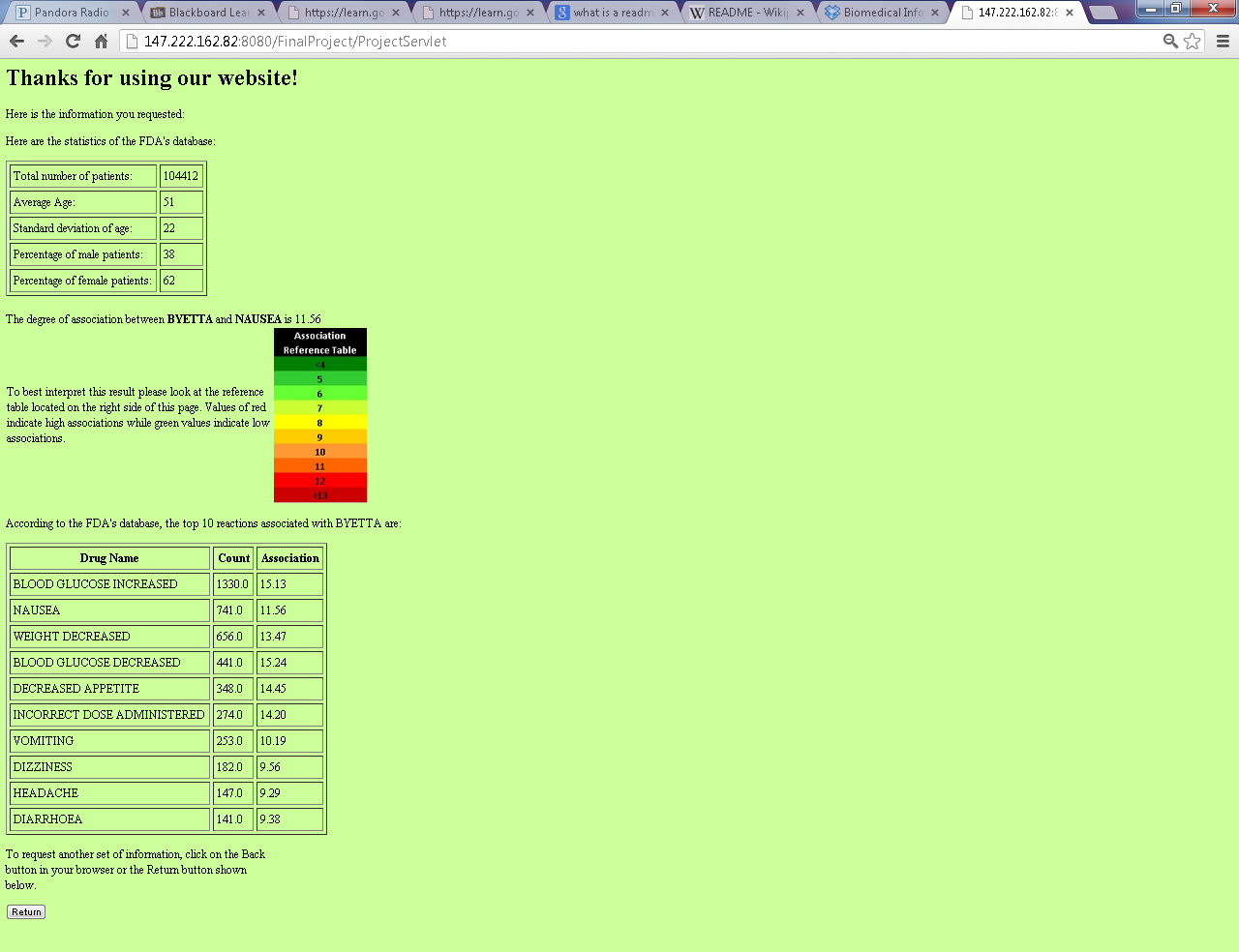
**Figure 12 – Results of association between BYETTA and NAUSEA.**

****

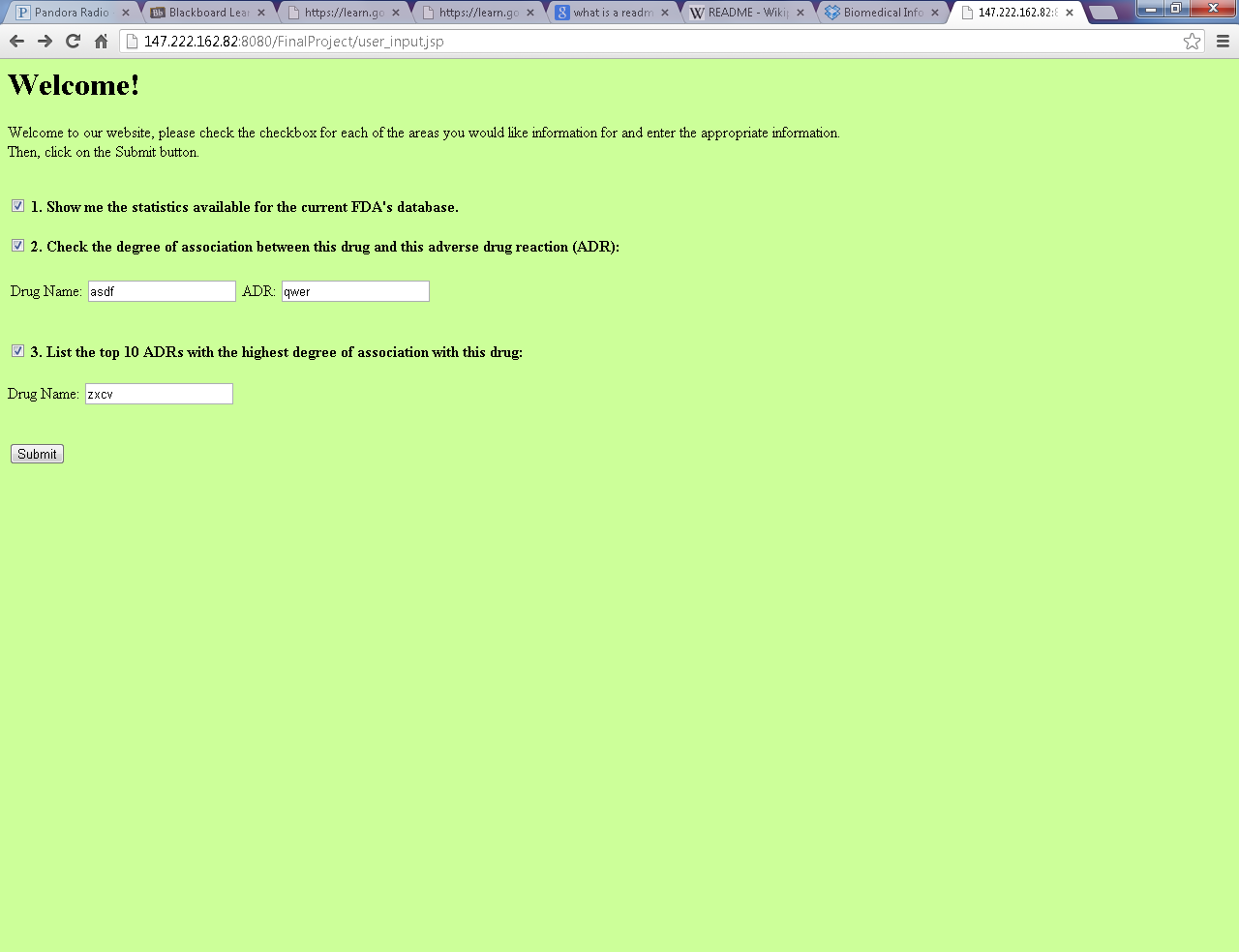
**Figure 13 – Checking the top 10 reactions associated with drug BYETTA.**

****

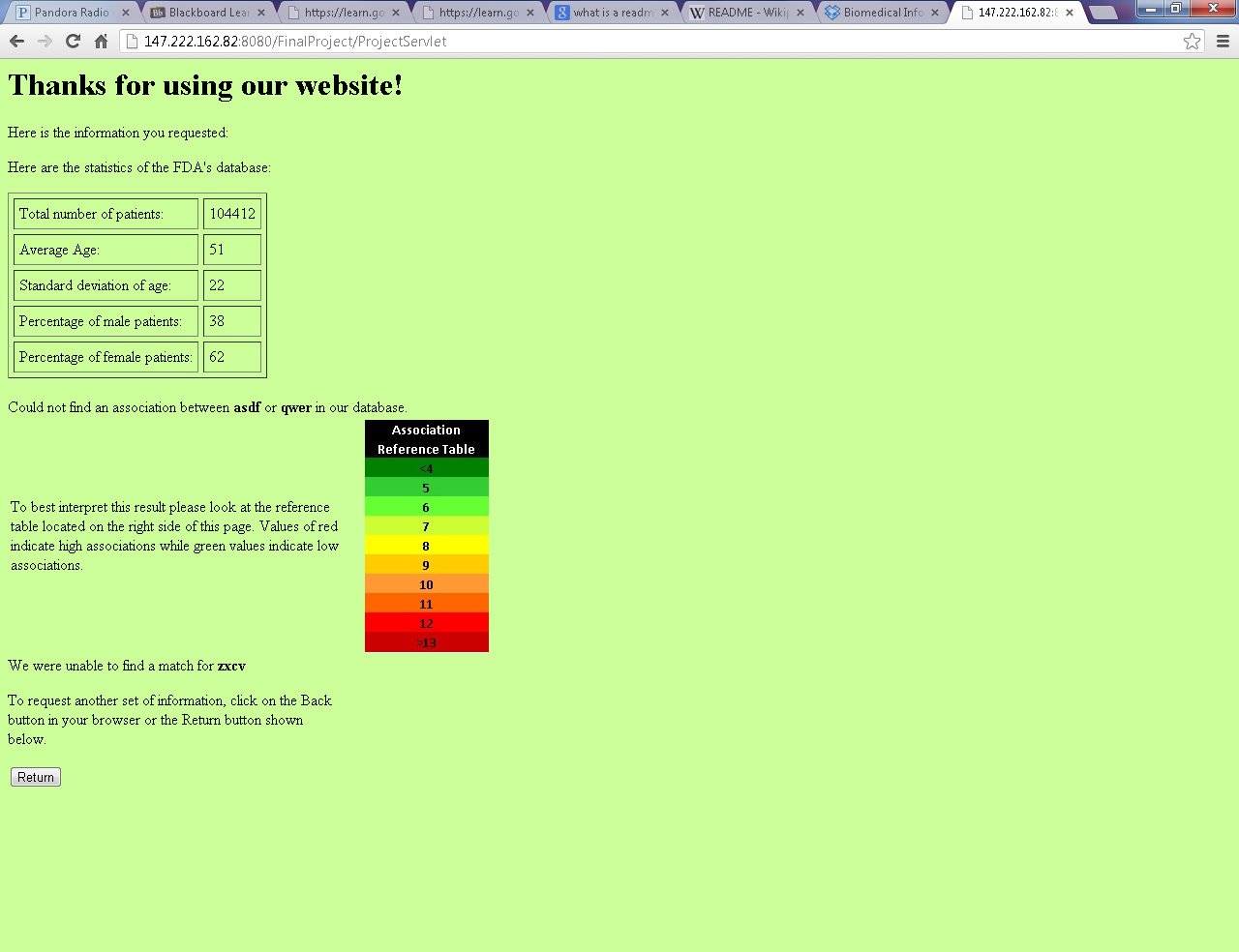
**Figure 14 – Results of top 10 reactions associated with drug BYETTA.**



**Figure 15 – All three features working simultaneously.**



**Figure 16 – Entering variables not available in the database.**



**Figure 17 – Output of variables not found.**

**Discussion and Conclusion**

A major issue in this project was relating our database schema to the schema assumed in the source journal article. The article’s database system assumed certain drugs were flagged as having a high degree of relation with certain reactions, and thus, assumptions could be made about the data mining. Furthermore, their data table was a combination of the reac08q1 and drug08q1 tables. Therefore, it was necessary to create a new table that served as a cache relating drugs and reactions with their ISR numbers. A significant hurdle in performing this SQL query was that it took longer than 3 hours to execute. Ultimately some table indexing had to be performed in order to speed up the execution of the query to a more realistic time frame. Finally the needed cache table was achieved and the top 10 association was performed much more effectively.

From testing, the system behaves according to the rules of association defined in the model journal article. Reaction and drug associations with a high number of P(x,y) (probability of both drug and reaction occurring) values do not always guarantee high IC associations. If a drug or reaction occurs numerous times in the database, the IC expects a high degree of association due to its frequency. Therefore, the highest IC values can be obtained from a high number of P(x,y) values while the drug and reaction occur rarely in the database.

Although the implemented algorithm is effective and efficient for this project’s scope, a larger scale operation might stretch its limitations. In the third part, the user is allowed to select a drug name from which to issue a top 10 reaction calculation. If the top 10 associations in the database were to be found without one variable known, the algorithm would scale in complexity significantly. The journal article worked with information already given to them that indicated certain pairs that would be likely to occur together. The database configuration for this project does not have this kind of knowledge pre-processing and one would need to consider all of the possible combinations of drugs and reactions in the system and compute their associated IC’s. This would introduce new challenges in data mining that the journal article did not address.

Overall, the system works well for large scale database querying and determining associations between drugs and reactions. However, it does not scale effectively for scanning a large data set in which no prior knowledge is known about associations and in which all combinations of pairs must be considered.

**Contributions**

The completion of this project depended highly on the collaboration of the team members. Dan Rahm and Nate Wendt developed the controller and model while Jhobany Tortolero handled the View aspect of the project. In the end, the team members had to collaborate with each other in order to fine-tune the final product – each one editing aspects in different parts of the program. Furthermore, the formal report was written as a team effort by all the team members.

**References**

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lanser, A., & De Freitas, R. M. (1998). A bayesian neural network method for adverse drug reaction signal generation. *PHARMACOEPIDEMIOLOGY AND PRESCRIPTION*, (54), 315-321.