|  |
| --- |
| Technical Note  Issued: |
| Correlation Analytics of Chinese National Cardiovascular Data Repository |
|  |
| QiZhong Lin, Cyrus CHAN; Choo Chiap Chiau;  Philips Research China |
| Philips Restricted  © Koninklijke Philips N.V. |

|  |  |  |
| --- | --- | --- |
| QiZhong Lin |  | QiZhong.LIN@philips.com |
| Cyrus CHAN |  | Cyrus.chan@philips.com |
| Choo Chiap Chiau |  | Choo.chiap.chiau@philips.com |

Authors’ address

© KONINKLIJKE PHILIPS NV   
All rights reserved. Reproduction or dissemination in whole or in part is prohibited without the prior written consent of the copyright holder .

|  |  |
| --- | --- |
| Title: | Correlation Analytics of Chinese National Cardiovascular Data Repository |
|  |  |
| Author(s): | QiZhong Lin; Cyrus CHAN; C.C. Chiau |
|  |  |
| Reviewer(s): | Daniel Elgort, Li, Zuofeng. |
|  |  |
| Technical Note: |  |
|  |  |
| Additional Numbers: |  |
|  |  |
| Subcategory: |  |
|  |  |
| Project: | Healthcare Big Data Analytics (2011-0147) |
|  |  |
| Customer: |  |

|  |  |
| --- | --- |
| Keywords: | Correlation Analytics, multi-variable association, Data Visualzation, Statistical Analytics, Cardiology |
|  |  |
| Abstract: | Currently Cardiovascular Information System (CVIS) is only being used for reporting, scheduling and management purposes. There is no deep analysis tool to turn rich data to information even though there are a number of third party business analytics tools available for generic usage. The proposed correlation analytics method is to automatically generate multi-variable associated/correlated information for the data in the clinical data repository.  The proposed correlation analytics method transforms the data from Cardiovascular Data Repository (CDR), into useful information, and then visualizes the information in an interative and convenient way in web browsers. |
|  |  |

|  |  |
| --- | --- |
| Conclusions: | The method provides quantitative analytics and visualization results of statistical significance for generating research hypotheses. The embodiment for Correlation Analytics is to be implemented in the Cardiovascular Data Repository (CDR) – a cardiology solution developed by BU-HIT China. |
|  |  |

Management Summary

One-page summary for Management.

A variety of data source from Cathlab, Philips Xper, Philips Xcelera and others like HIS/EMR, RIS and LIS are collected and stored in CDR. There is implicit knowledge underlying this big data, but it is not easy or even impossible for human to digest the raw data directly. So how to transform this big data into knowledge automatically shows more and more importance in improving cardiac patient care, research & registry outcomes.

There are a number of third party business analytics tools available for users to conduct data analysis,however, these tools are developed for generic usage. This project data analysis tries to analyze the CNCDR data in depth to provide knowledge/information to the physician in a friendly way.

The correlation analytic module includes three parts: data preprocessing, correlation calculation and visualization. The user case has been refined with coorperation with clinical doctor and BIU, user interface is designed with coorperation with data visualization team, UI design, Philips Netherland, and implementation is coorperated with Pins Philips Netherland and R&D team from BIU.

In physician’s view, current correlation analytic module is meaningful, and then it has been accepted and transferred to BIU, BIU tries to integrate correlation analytic module into CNCDR platform, meanwhile correlation analytic module will be integrated into DHP. Our next step will focus on patter discovery.

Contents

[Correlation Analytics of Chinese National Cardiovascular Data Repository i](#_Toc407094015)

[1. Introduction 7](#_Toc407094016)

[1.1. Background 7](#_Toc407094017)

[1.2. Motivations 7](#_Toc407094018)

[1.3. Solution approach 7](#_Toc407094019)

[2. Methodology 9](#_Toc407094020)

[2.1. Data Preprocessor 9](#_Toc407094021)

[2.1.1. join table between devices and lesions 10](#_Toc407094022)

[2.2. Data Correlation 11](#_Toc407094023)

[2.3. Data Visualization 11](#_Toc407094024)

[2.3.1. Visualization in graph connection (in Matlab Environement) 11](#_Toc407094025)

[2.3.2. Data Visualization with intuitive GUI(final product) 14](#_Toc407094026)

[3. Implementation 15](#_Toc407094027)

[3.1. System Structure 15](#_Toc407094028)

[3.2. Data interface 15](#_Toc407094029)

[3.3. Component Diagram 16](#_Toc407094030)

[3.4. Class diagram 18](#_Toc407094031)

[3.5. sequence diagram in client view 19](#_Toc407094032)

[4. Results 20](#_Toc407094033)

[4.1. view category correlation and view top 5 outcome measures 20](#_Toc407094034)

[4.2. view Attribute correlation 21](#_Toc407094035)

[4.3. view bar chart 2D 21](#_Toc407094036)

[4.4. view bar chart 3D 22](#_Toc407094037)

[4.5. view a list of patients 22](#_Toc407094038)

[5. Discussion and Conclusions 24](#_Toc407094039)

[6. Acknowledgements 24](#_Toc407094040)

[References 25](#_Toc407094041)

# Introduction

In this section, we first introduce the background with a focus area in cardiology. Then we elaborate the motivations of our module from the perspective of data analytics. We then present the solution approach in the rest of this technical note.

## Background

Cardiovascular Information System (CVIS) has been around for many years in the healthcare industry. The main function of current CVIS is to store and assess patients’ records. Some sophisticated CVIS will also integrate with other HIT system in the hospital such as Electronic Health Record (EHR), Laboratory Information System (LIS), and etc. As such, physician will have an over view of patients’ health record.

However, usually the CVIS is being used for reporting, scheduling and management purposes. There is no method currently in the CVIS that allows users/physician to turn the information stored into knowledge.

One of the main reasons that currently the information in the CVIS is not being analysed and transformed into new knowledge for users is because there is a lot of information (over 200 data attributes per patients) recorded in the system and on top of that there is a lot of dependency among the data attributes.

There are a number of third party business analytics tools available for users to conduct data analysis. However, these tools are developed for generic usage. Therefore, as a user of CVIS, one needs to first extract data from the CVIS system or/and other HIT systems within the hospital, then pre-process them to a correct format before users can start using the third party analytics tool. This will require a large amount of time and resources every time when physicians want to conduct data analysis using patients’ records.

## Motivations

Increasingly more data are collected and stored in the cloud due to the advancing in IT, there is implicitly knowledge inside, how to turn rich data into knowledge and summary for human to easily digest the big data is definitely important. The proposed method is to process and analyze big data and then visualize the result in an easy and simple way to understand. The final goal of analytics solution is to improve cardiac patient care, research & registry outcomes.

## Solution approach

As mentioned before, a variety of data source from Cathlab, Philips Xper, Philips Xcelera and others like HIS/EMR, RIS and LIS are collected and stored in CDR, there is implicit knowledge inside this big data. Technique to transform the rich data into knowledge and summary for human is very important. In this report, we propose the analytics solution as following diagram.

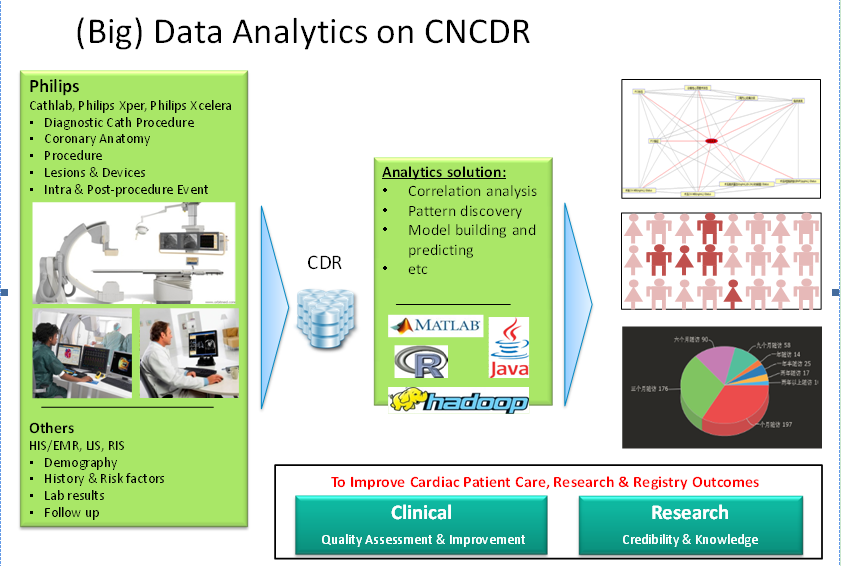


Figure 1 The proposed solution of data analytics on CNCDR

In our data analytics plan, the data analytics modules are being researched and developed step by step. First the correlation analytic is used to search the strength of pair-wise or multi-variable relationship with statistics significance. Based on result from correlation analytics, the module pattern discovery will be developed and used to further reveal the implicit knowledge in terms of predictive power from related multiple dependent attributes. Finally a predictive model will be built to predict unkown cases. This technical note focuses on the correlation analytic module.

In this report, we focus on the proposed method of correlation analystic, which automatically pre-processes the entire data base stored in the CDR, then performs correlation test analysis on all the data attributes, and finially visualizes the correlation results and are integrated into the CDR platform.

# Methodology

## Data Preprocessor

The workflow diagram of data preprocessor is shown in Figure 2 below:

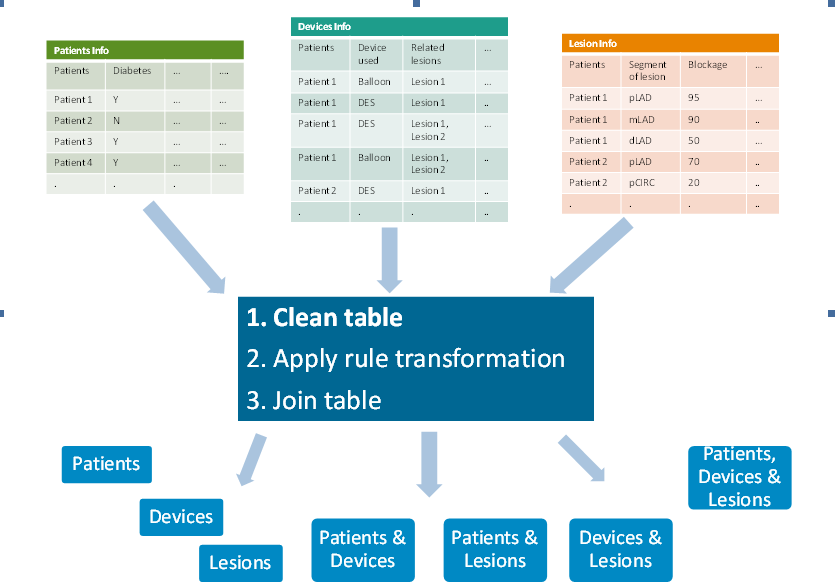


Figure 2 The workflow diagram of data preprocessor

First, PCI registry data including patients information, devices information and lesions information acquired from Intellispace CDR comes from various tables. Because the raw data is noisy and incomplete, the data should be normalized with unified labels. For example, there are a lot of variants representing missing data like “”, “?”, “N/A” and “<undefined>” etc. Meanwhile, if the number of missing data in one row or column is over the threshold, this row or column had better be removed. The above process is denoted as table cleansing.

Second, the data types of attributes include numeric and categoricalones. There are two types of rules. Single rule is to transform numeric type into category type. For example the value range of the attribute “Proximal LAD Vascular Stenosis Percent” between 0 and 70 is normal, the other value is abnormal. Compose rule is to compose multiple attributes to generate a new attribute. For example, the attribute “chronic total occlusion (CTO)” is Yes and the attribute “Post-operative TIMI Flow” is equal to 3, then the new generated attribute “CTO Outcome” is successful; if “chronic total occlusion” is Yes and “Post-operative TIMI Flow” is not equal to 3, then “CTO Outcome” is Failed; if “chronic total occlusion” is No, then “CTO Outcome” is No CTO.

Third, joining table is usually required for data preprocessing. For example, the correlation of pair-wise or multiple attributes is not only from one table, but also from between tables. In PCI registry data, the relationship of patient, device and lesion is complex. For example, one patient has multiple devices and lesions, multiple devices are used to one lesion, multiple lesions maybe using one device, so how to join tables should be handled carefully. Because the relationship between patient and device or patient and lesion is one to many, common inner join is sufficient. The following sub-section focuses on how to join tables between devices and lesions where the relationship is more complex.

### join table between devices and lesions

As shown in tables, both the registry ID of devices and lesions are based on the column called patient id, meanwhile, the device is used for lesion based on the column called relevant lesion, apparently the relationship between devices and lesion is many to many. So the key point of how to join devices and lesions is to generate the unified ID for devices and lesions.

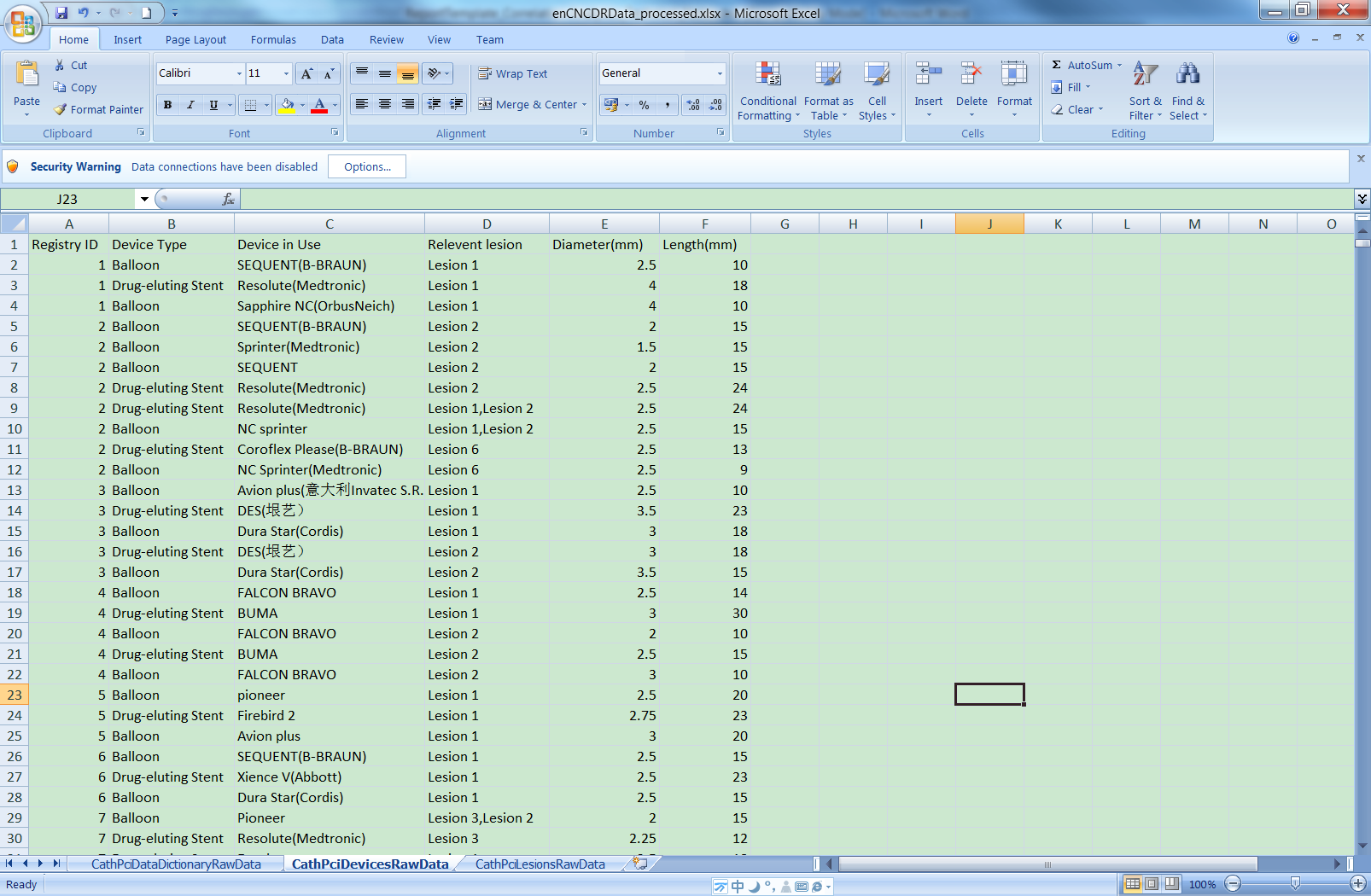
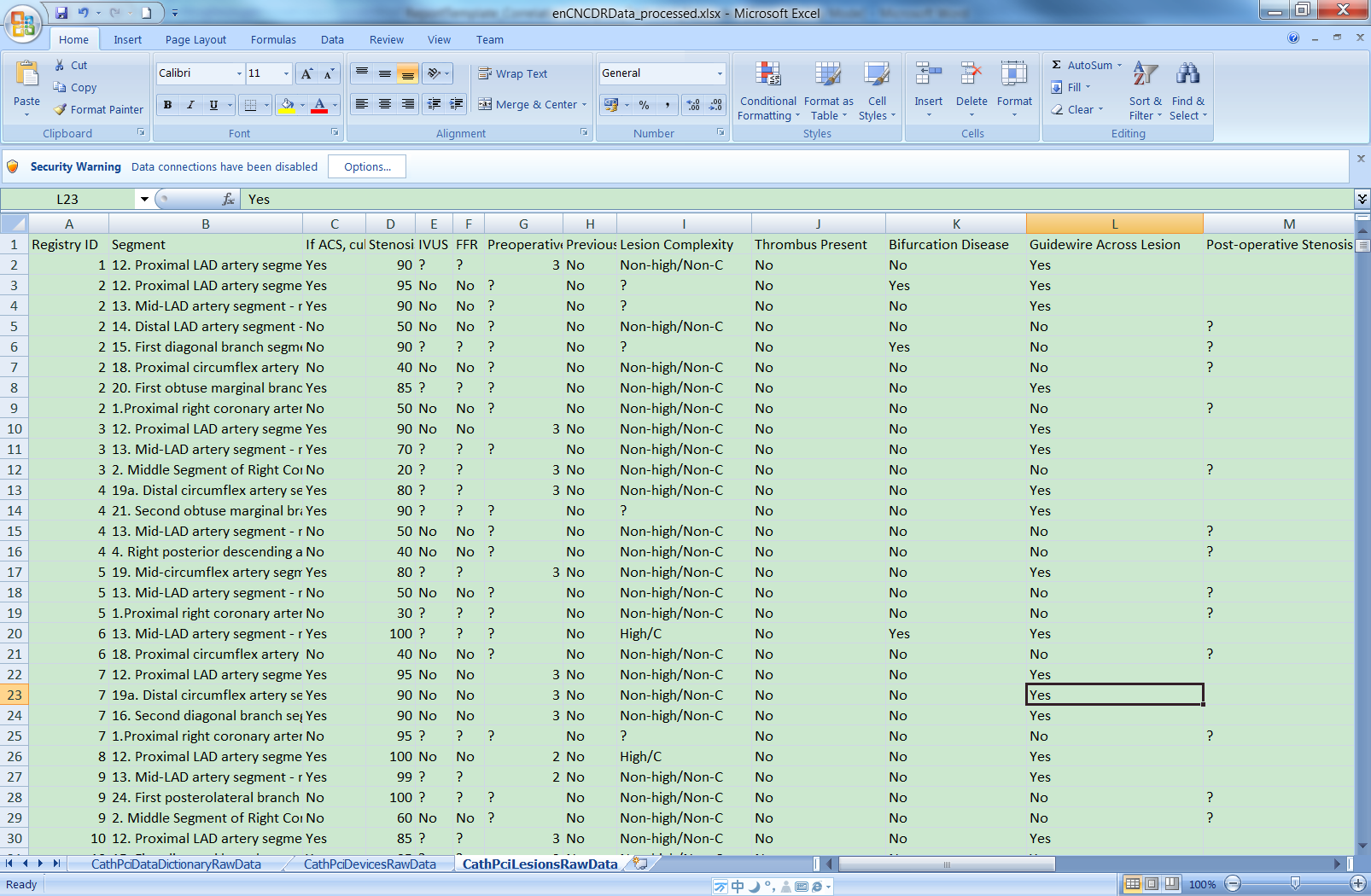
In the device table, the actual unique ID of devices-lesions could be generated by combining Registry ID and Relevent lesion. For example, if Registry ID of the devices is 2 and Relevent lesion is “lesion 1, lesion 2”, two IDs of devices-lesions could be generated as tuples {2, 1} and {2, 1}.

In the lesion table, the ID of devices-lesions could be generated by Registry ID and Registry ID number. For example, the Registry ID 2 has 7 lesions, then seven IDs of devices-lesions could ge nerated as tuples {2, 1}, {2, 2}, {2, 3}, {2, 4}, {2, 5}, {2, 6} and {2, 7}.

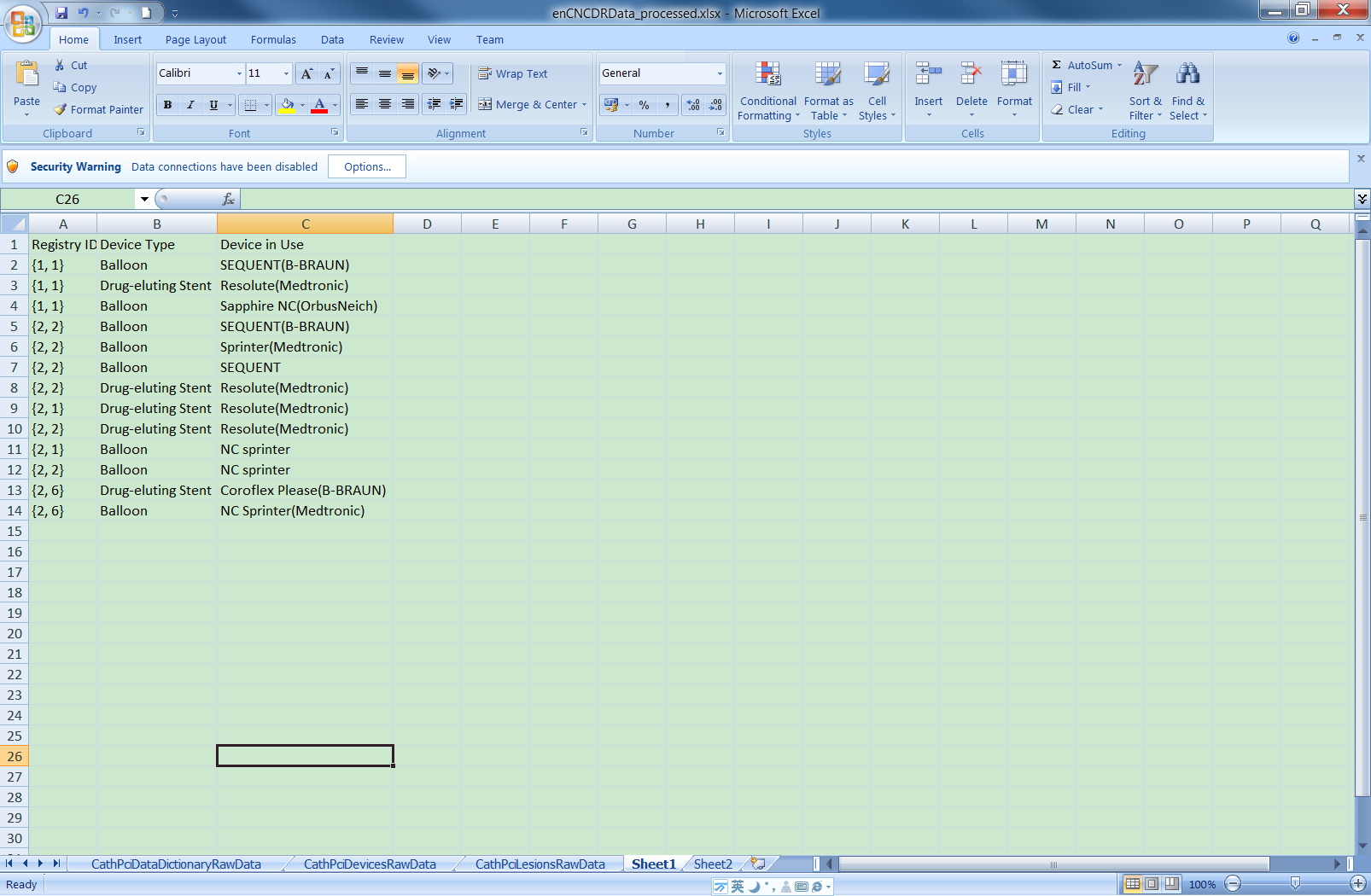
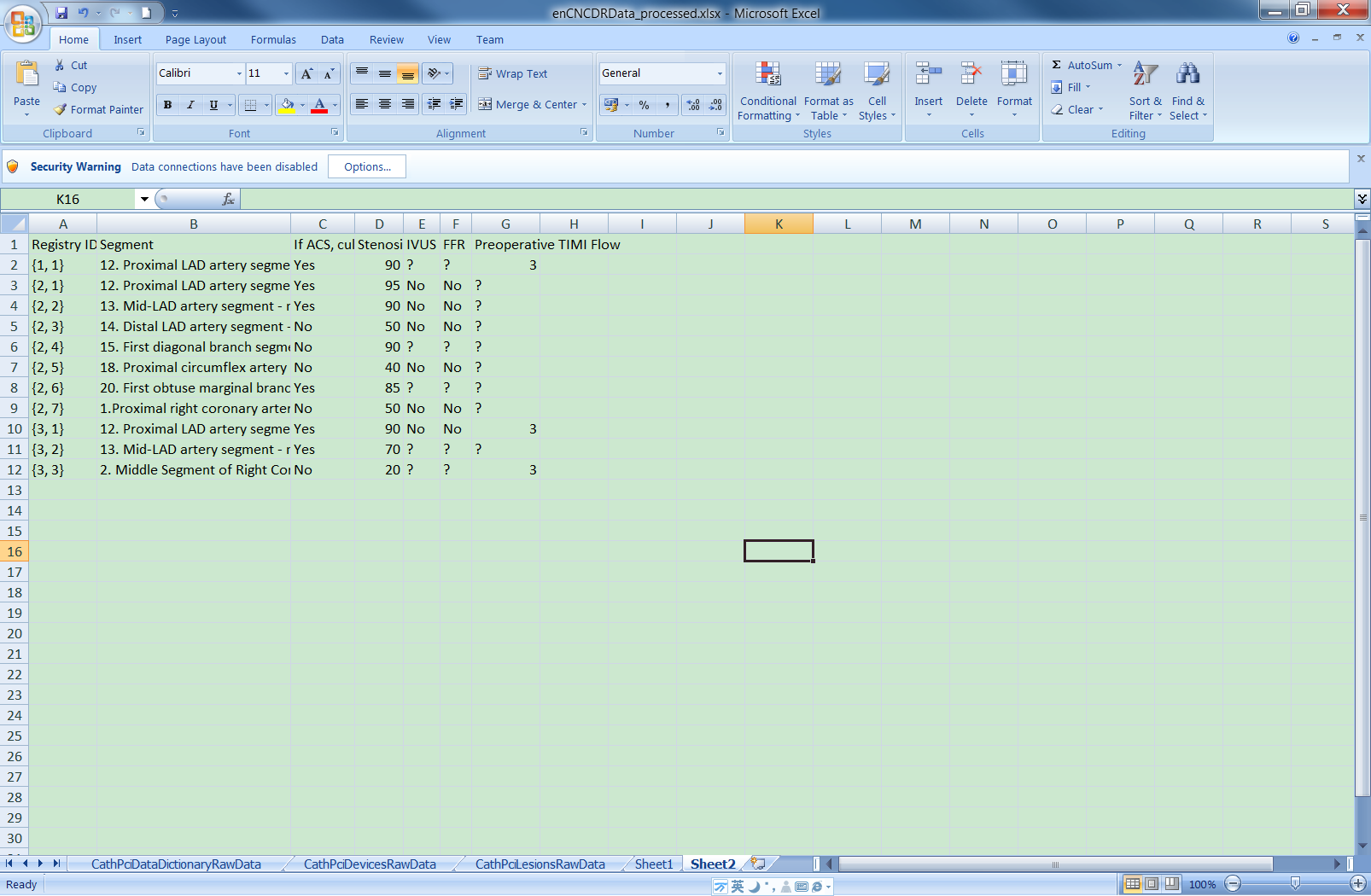
After both device and lesion are standardized with unified ID, then the relationship between the standardized device and lesion is one to many, then common inner join can be applied to join them.

The devices table and lesions table are joined to output devices-lesions tables, as shown in Figure 3 below.

Device Lesion

Standardized device standardized lesion

Device-lesion

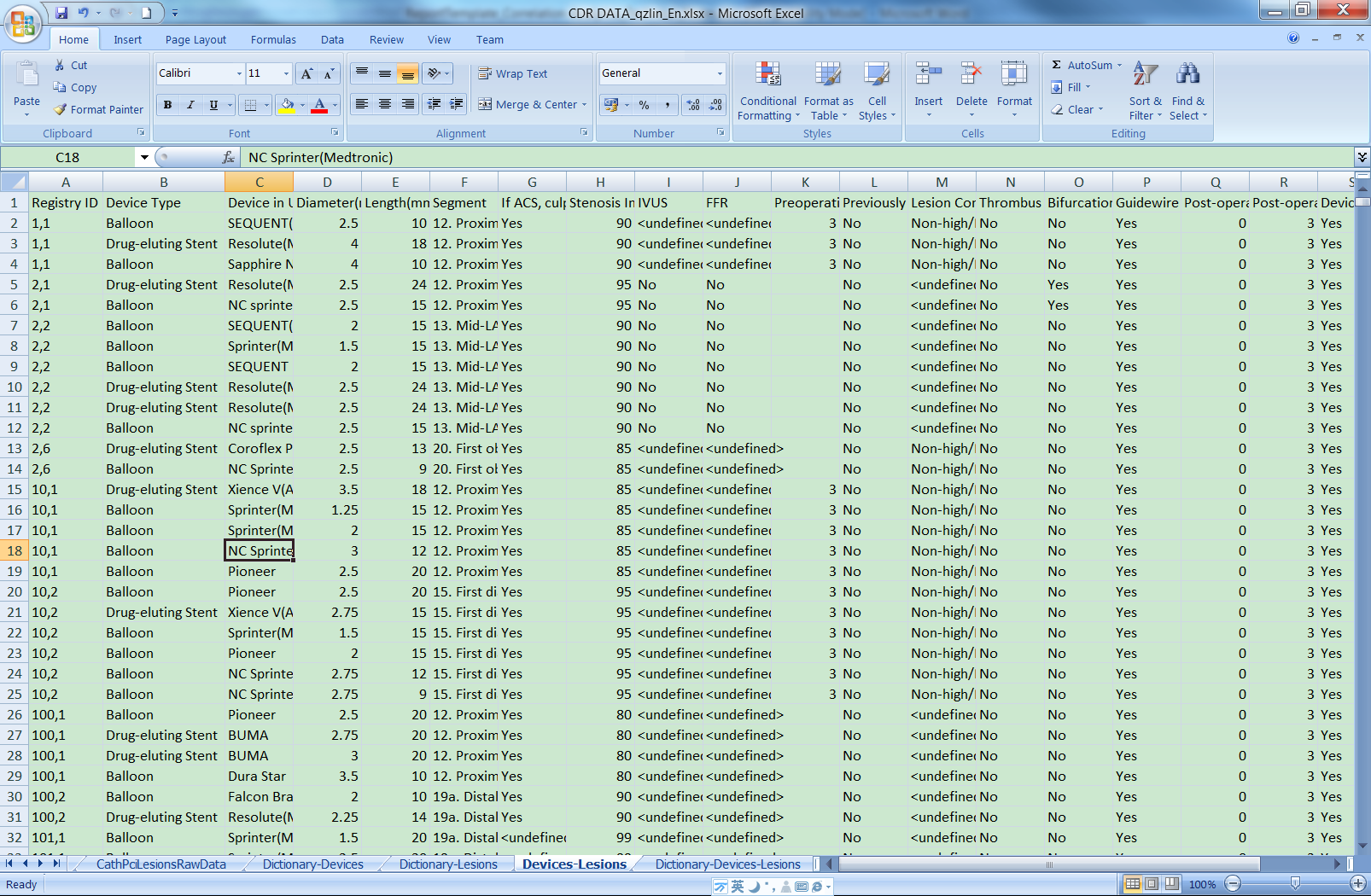


Figure 3 Join tables: the up-left is device table, the up-right is lesion table, the middle-left is standardized device, the middle-right is standardized lesion and the bottom is the joined device-lesion table

## Data Correlation

After data preprocessing, seven tables including patients, devices, lesions, patients-devices, patients-lesions, devices-lesions, patients-devices-lesion are generated and output. There are three steps to calculate data correlation as follows:

* First, calculate P-value and Phi value of paire-wise attributes.

Each pair of attributes is selected from one table of seven tables, then chi-square test[1] of them is calculated, finally the correlation results including P-values and Phi values (Cramer’s V) [1] of pair-wise attributes from all tables is obtained. P-value measures how unlikely the result is due to chance, which means in statistical significance level whether the pair-wise attributes have correlation is meaningful. In practice, the P-value of pair-wise attributes is lower than threshold like 0.05 (significance level), which means this pair-wise attributes are likely correlated. Then Phi value of pair-wise attributes is calculated to denote the strength of correlation. Please pay attention to the pre-condition of Chi-square test. For example, the number of patient in each category should be over 10 (or 5).

* Second, clean the results.

the resultant P-value and Phi value matrices of all pair-wise attributes above require removal of insignificant results. In particular, pairs with P-values higher than the significance level should be removed and the corresponding Phi value should be removed too. Then the targeted (selected) attributes are sorted according to how many related attributes in descending way.finally the targeted attributes without significantly correlated attributes are removed.

* Third, format the results.

different attribute of PCI registry belong to different subcategory, for example, Age Group and sex belong to DEMOGRAPHICS, CABG History and BMI Group belong to HISTORY AND RISK FACTORS. In PCI registry there are totally 11 subcategories, from correlation matrix of pair-wise attributes from all tables, the correlation matrix of pair-wise subcaterogies could be calculated and obtained.

From the correlation matrix of pair-wise attributes, any remaining targeted attribute and its related attributes and inter relationship of its related attributes could be searched and obtained.

## Data Visualization

### Visualization in graph connection (in Matlab Environement)

* As shown in figure 4, the left panel shows a list of sorted data attributes which have related attributes. Here the attributes are sorted according to how many related attributes there are in a desceding way

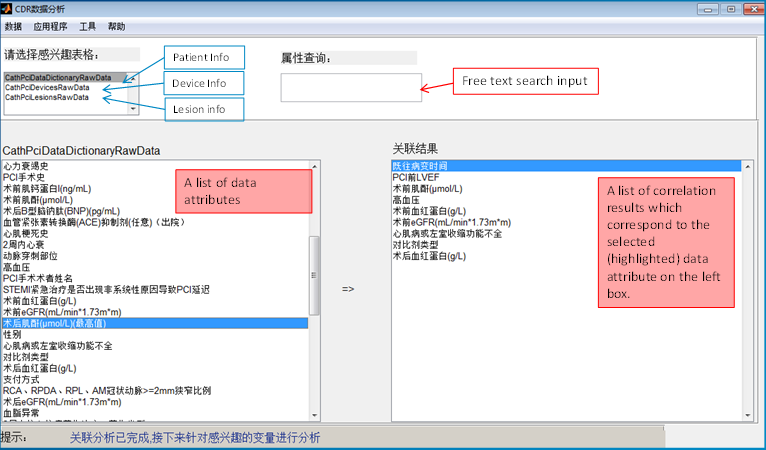


Figure 4 Demo inferface for attributes which have related attributes

* after the user selects the targeted attribute like “术后肌酐(μmol/L)(最高值)”, the right panel shows the related attributes in sorted way like first is “既往病变时间”, second is “PCI前LVEF” and others, here sorted way is according to the correlation strength in descending way.
* to view targeted attribute and its related attributes and inter relationship of its related attributes, visualization is the graph connection as shown in figure 5.

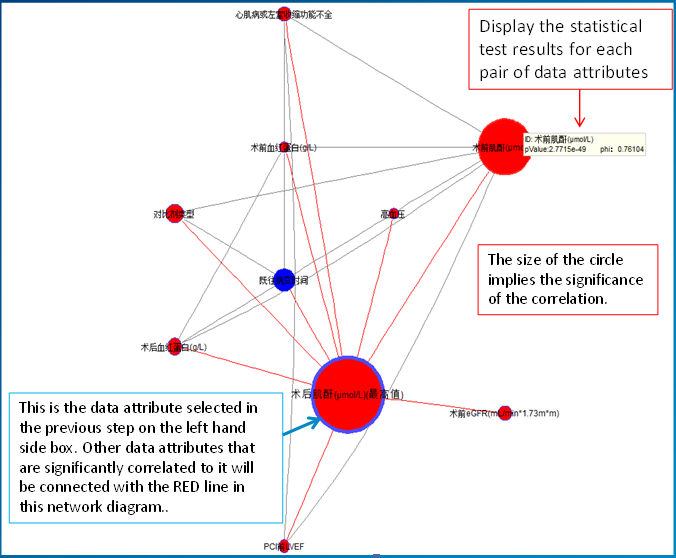


Figure 5 targeted attribute and its related attributes in graph

Bar charts of patient counts are employed to view detailed information of targeted attribute and its related attributes.Further, the bar charts provide better engagement between the abstract statistical results and the original patient data. Figure 6 shows the bar charts of two attributes and three attributes.

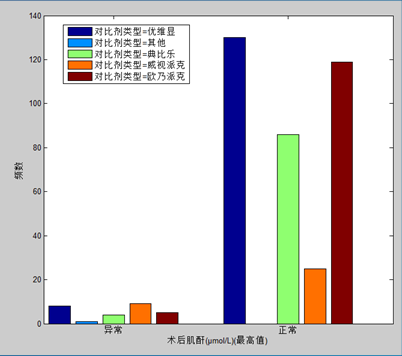
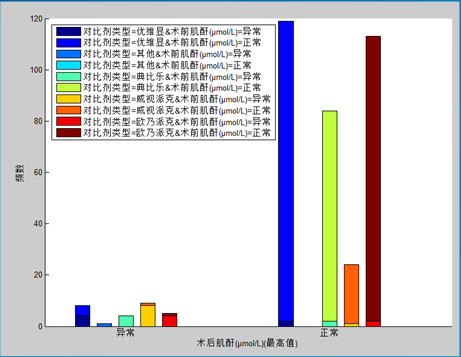
 

Figure 6 The bar charts of attributes: the left figure shows the bar chart of two attributes, the right figure show the bar chart of three attributes

### Data Visualization with intuitive GUI(final product)

To better present the analytics results and support user interactions, the GUI was re-designed by Philips Design and project team in Shanghai, The flow chart of the correlation analytics tool with new visualization is illustrated in Figure 7. Running examples of the actual GUI can be found in the Chapter 4.

Figure 7 The flow chart of the correlation analytics tool.

# Implementation

## System Structure

Figure 7 is the block diagram of Correlation Analytics, there are three modules including pre-processor, correlation analysis and visualization, visualization is implemented in front-end, others are implemented in back-end via java code.

The user interacts with the web UI, triggering an event with parameters in the front-end. The correlation analytics servlet as a controller receives the event and delegates task to module correlation analytics in the domain layer. Correlation analytics imports required data, assisted by the module pre-processor, completes the task and outputs the result in the json format to front-end, and then the web-UI is updated

**Back-end**

**Front-end**

CNCDR data

Pre-processor

Correlation

Analytics

Visualization

Graph matrix in json format

Rule file

CDR header file

post with json data

Data

Domain layer

Correlation

Analytics

Servlet

controller

Web UI

Figure 7 System diagram

## Data interface

1. CNCDR data: patient, device, lesion, and their joined tables.

The data structure for CNCDR data is: *List<CNCDRTableProxy> tableList*

*CNCDRTableProxy* model one table, which has interfaces: *String tableName, List<String> columnNames, ArrayList<ArrayList<String>> data*.

Attention:

1. the first 3 tables should be basic tables like patient, device and lesion. The last 4 tables are joined tables. The joined table has table name which is composed of basic table name via connection “-“.
2. in our API, the CNCDR data could be imported via 7 text files (patient, device, lesion, patient-device, patient-lesion, device-lesion and patient-device-lesion or one excel file with 3 sheets (patient, device and lesion)
3. the data is organized in a column-wise way to provide a convenient way for data analytics. Each column is identified by its attribute name.

For example:

*tableNameList = {“patient”, “device”, “lesion”, “patient-device”, “patient-lesion”, “device-lesion”, “patient-device-lesion”}*

*The patient table:*

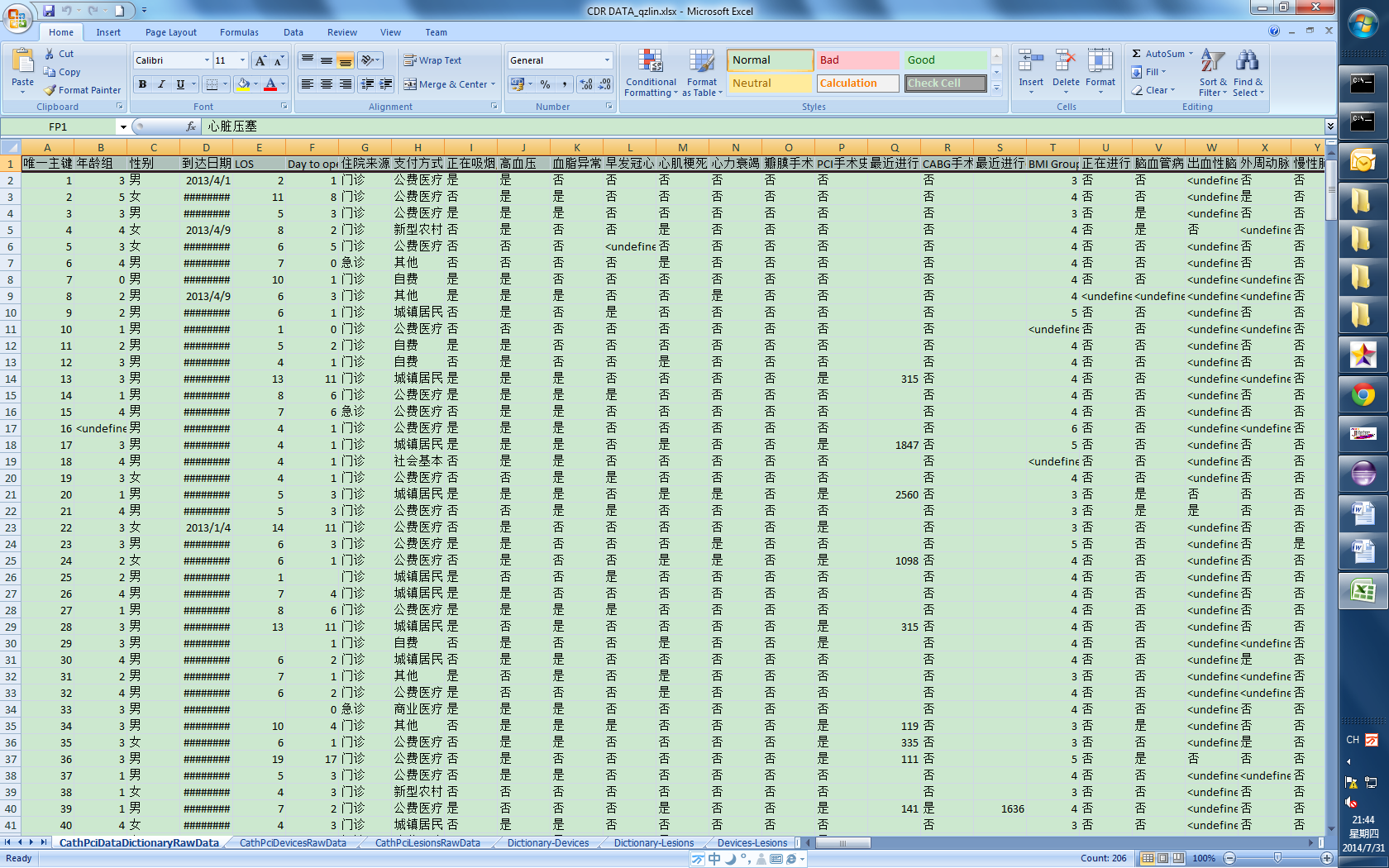


Figure 8 Patient table, other tables are similar in data structure

1. Rule file

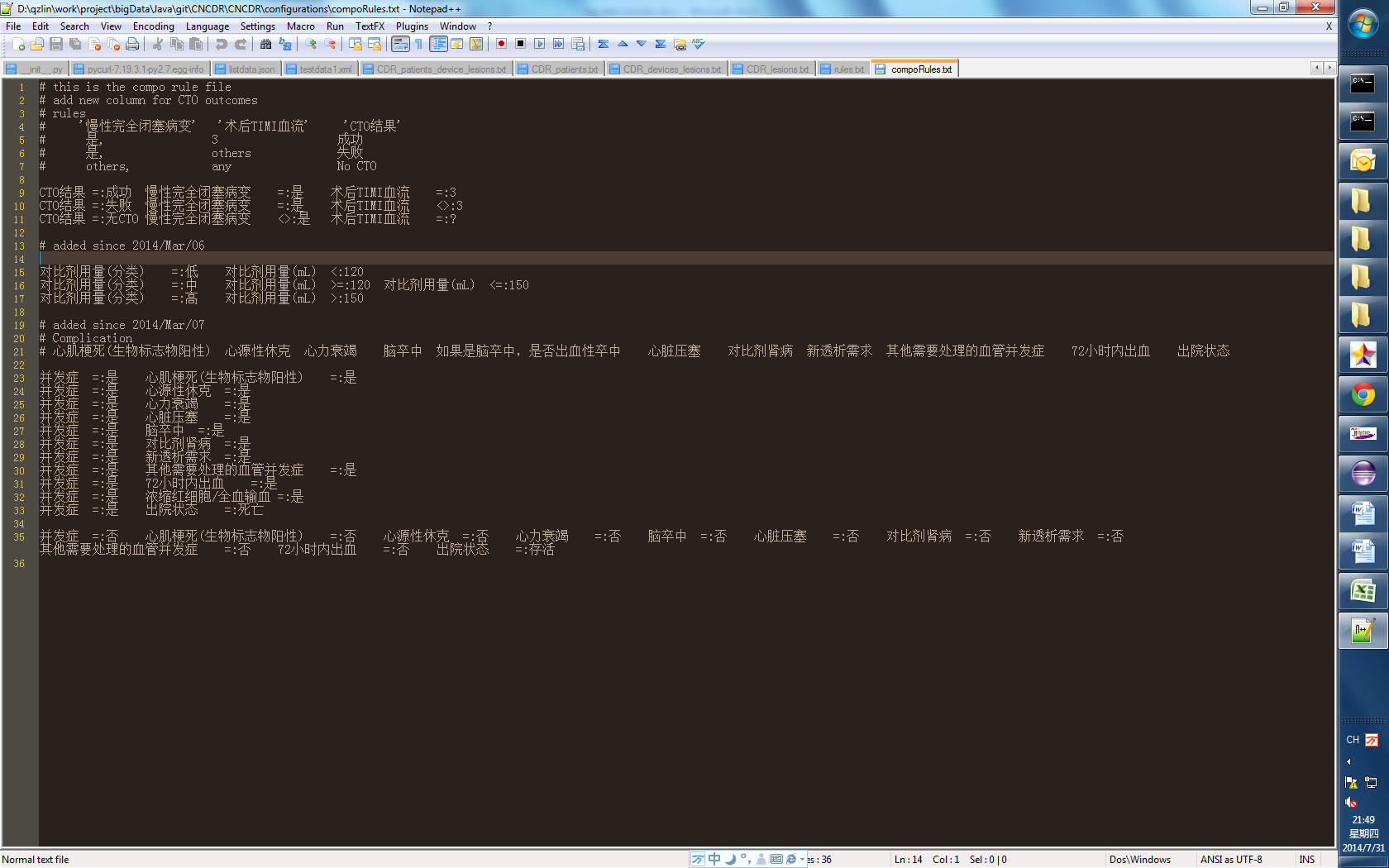
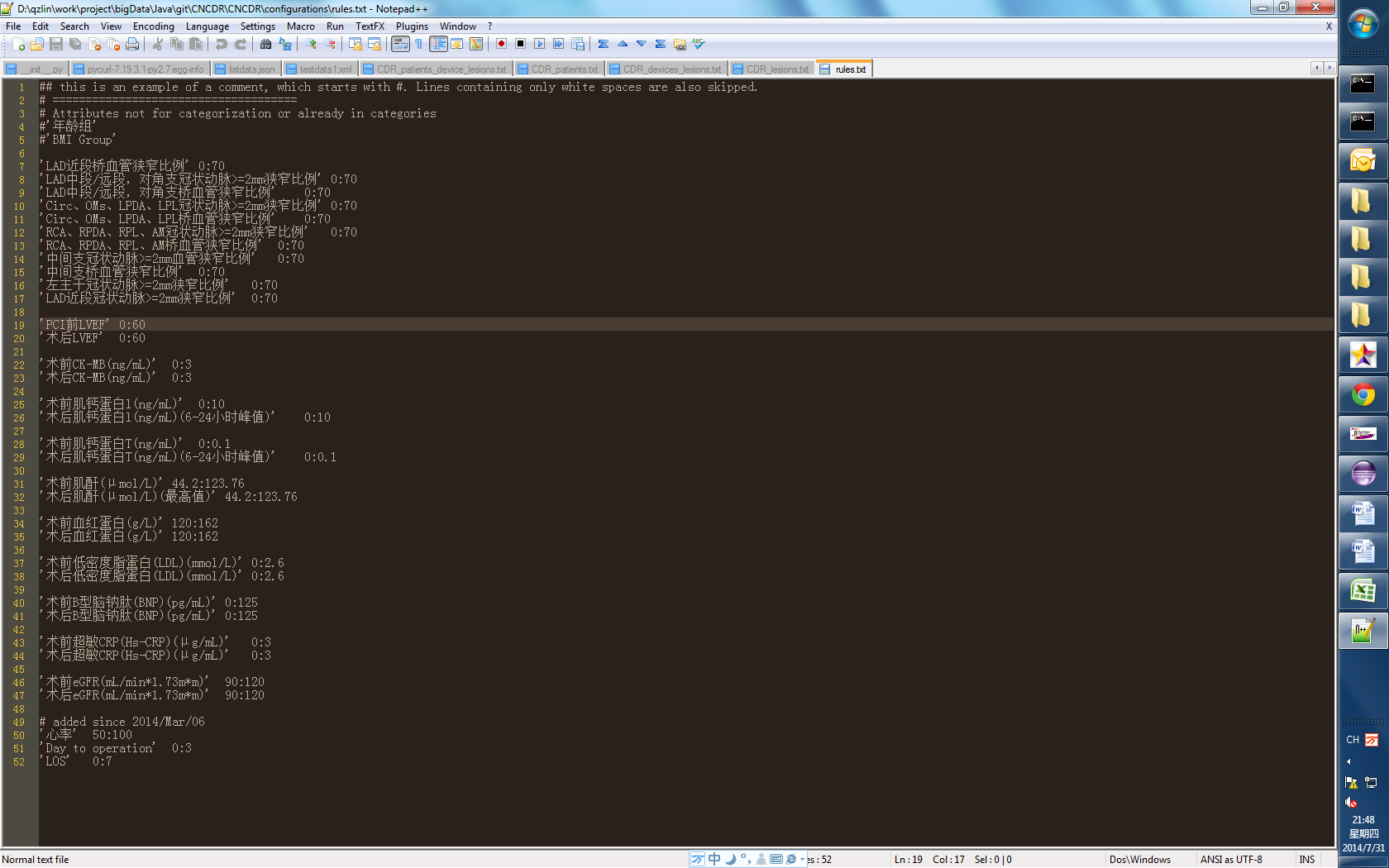


Figure 9 rule files, the left figure is single rule file, the right figure is compose file

The left file is the single rule file. In a single rule, an attribute is transformed into “normal” or “abnormal” according to domain knowledge. In Figure 9’s example, the left column indicates the attributes, and the right column indicates the numeric ranges for “normal”, numeric values otherwise are for “abnormal”.

The right file is the compose rule file, the right two columns (meaning two attributes) generate the left one column (meaning a newly derived attribute)

1. CDR header file

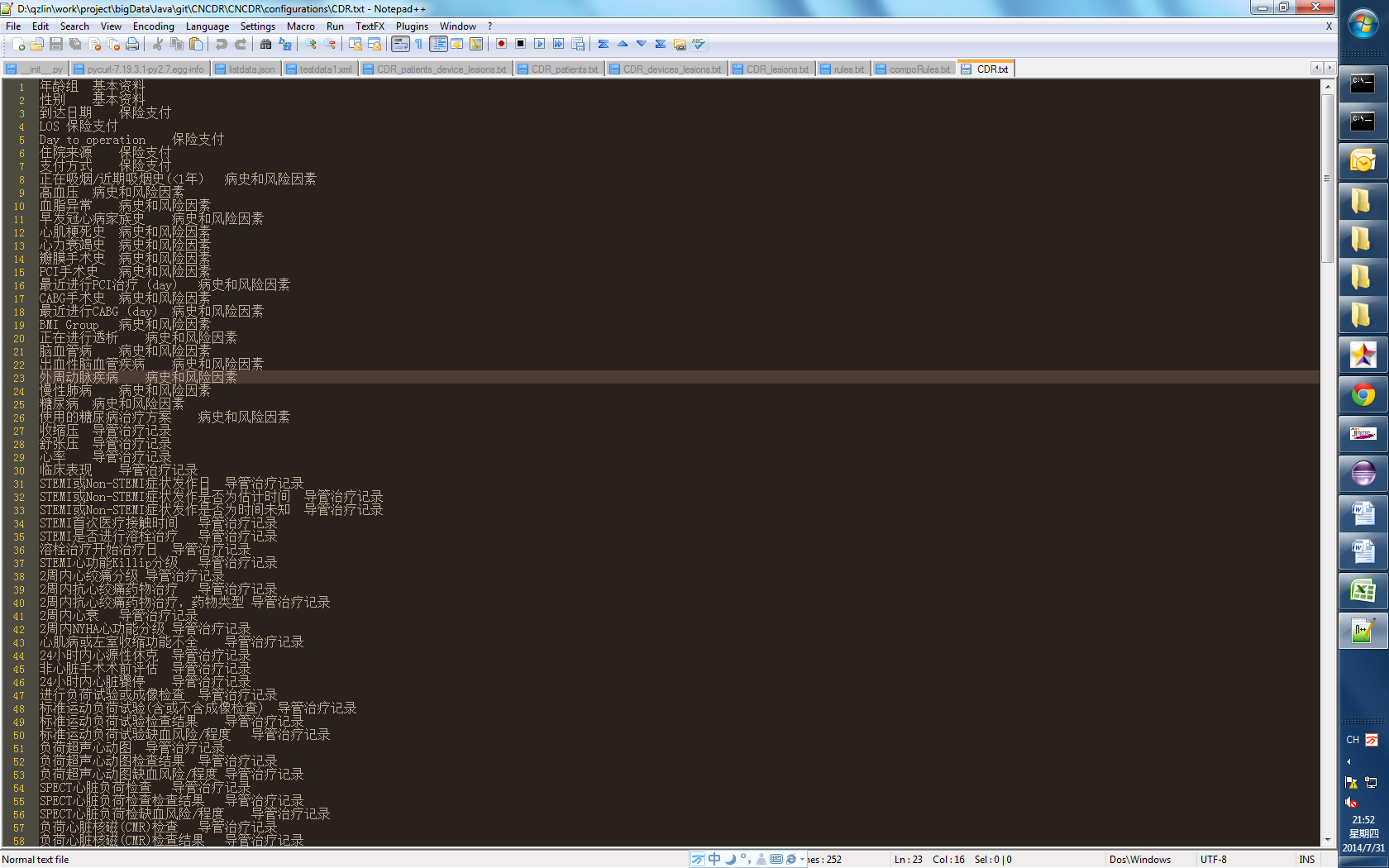


Figure 10 Attribute and its corresponding category

The left column indicates the attributes, and the right column indicates corresponding sub-categories.

## Component Diagram

The component diagram is described as follows, inside there are five packages including com.philips.ci.cdr.bigdata.correlation.servlet, com.philips.ci.cdr.bigdata.correlation, com.philips.ci.cdr.bigdata.correlation.foundation, com.philips.ci.cdr.bigdata.correlation.json and com.philips.ci.cdr.bigdata. correlation is dependent on commons-math3-3.2.jar, correlation.json is dependent on gson-2.2.2.jar. The component requires configurations file including single rule file (rules.txt), compose rule file (compoRules.txt) and CDR.txt

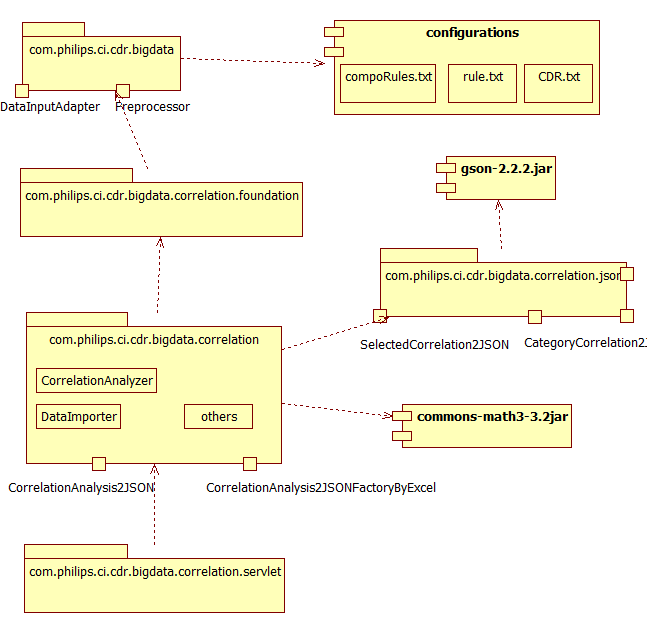


Figure 11 Component diagram

As shown in Figure 11, the component diagram exposes to the client the interface including CorrelationAnalysis2JSON, and CorrelationAnalysis2JSONFactoryByExcel.The package com.philips.ci.cdr.bigdata exposes to the package

com.philips.ci.cdr.bigdata.correlation.foundation. The interface includes Preprocessor, DataInputAdapter.

The package com.philips.ci.cdr.bigdata.correlation.json exposes to the package com.philips.ci.cdr.bigdata.correlation the interface including CategoryCorrelation2JSON, SelectedCorrelation2JSON, Chart2D2JSON, Chart3D3JSON.

Detailed class information in each package is referred to class diagram.

## Class diagram

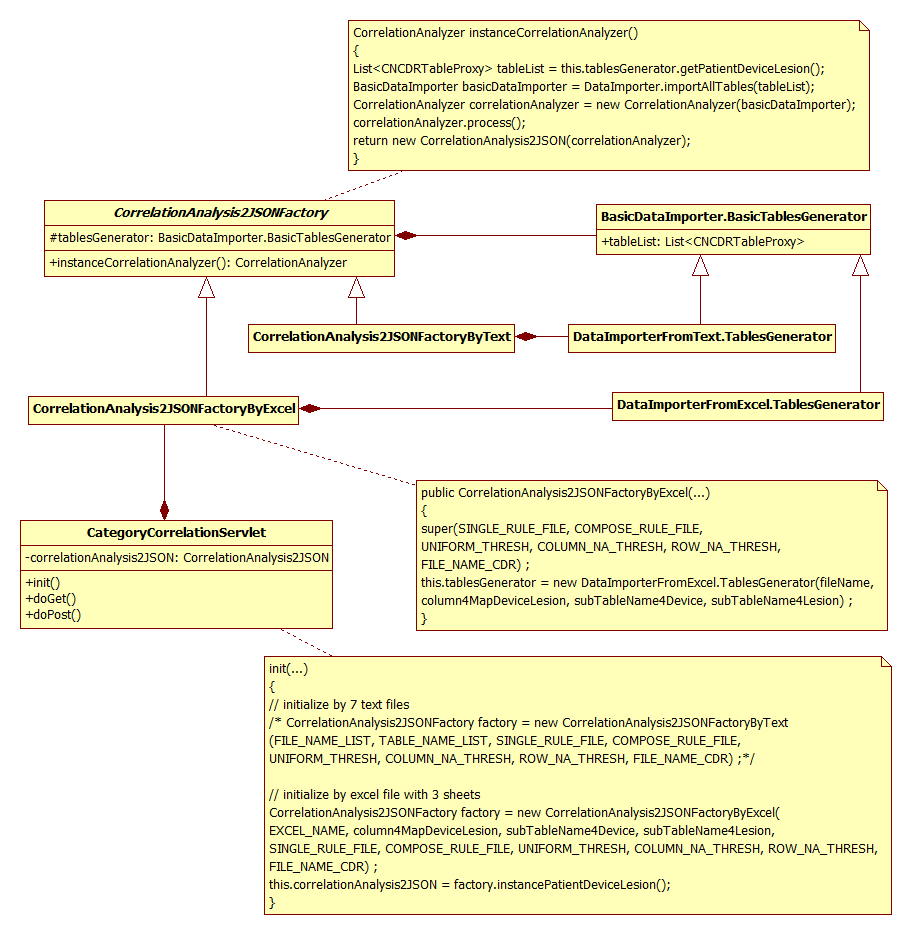


Figure 12 the class diagram in client view

## sequence diagram in client view

In client view, input the data to module, then call the service of correlation analytics, finally get the result in json format. please refer to figure 13 for more detailed interactive behaviour

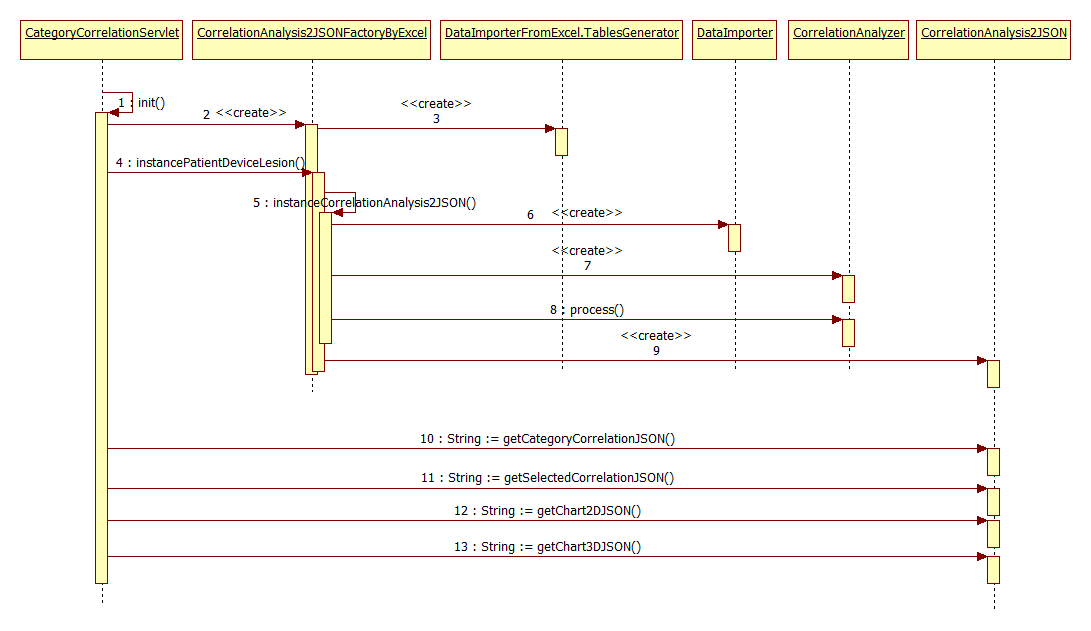


Figure 13 Sequence diagram in client view

# Results

the user navigates CNCDR platform, then filters CNCDR data and clicks correlation analysis module to view the page which shows the correlation information inside CNCDR data. Figure 14 shows correlation information after user inaction. More detailed work flow is listed in the next section.

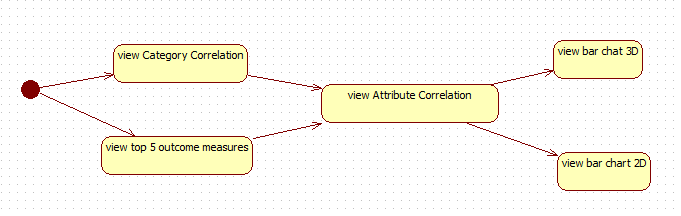


Figure 14 correlation information from coarse to fine

## view category correlation and view top 5 outcome measures

workflow inside:

1. access the URL using the “get” method in passing data
2. run correlation analysis module, get category correlation in json format.

Currently, input data is generated via file. For integration into the CDR platform, the input data will be extracted from database

1. visualization of category correlation, as shown in figure 15

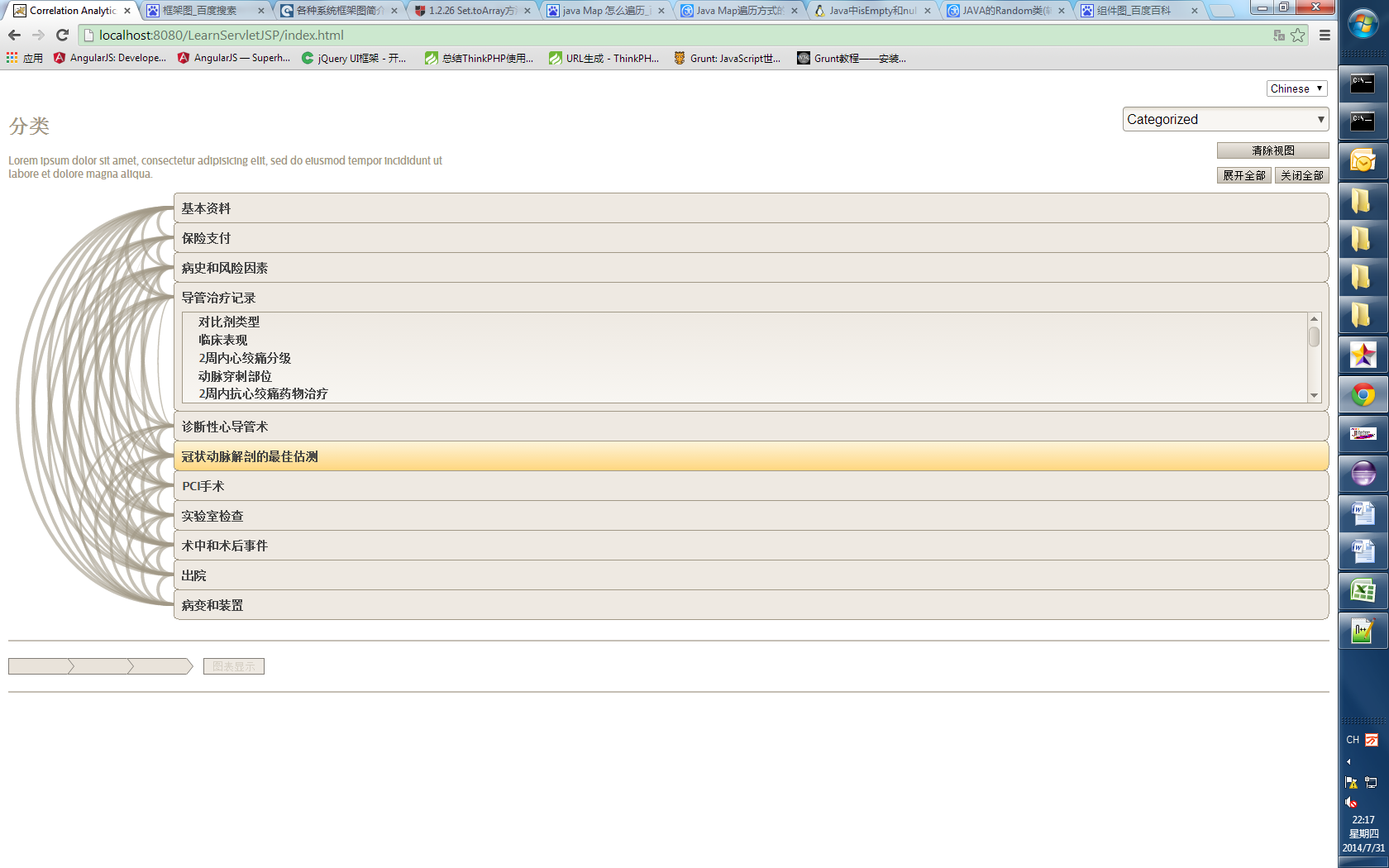


Figure 15 Category correlation

1. visualization of top 5 outcome measures, as shown in figure 16

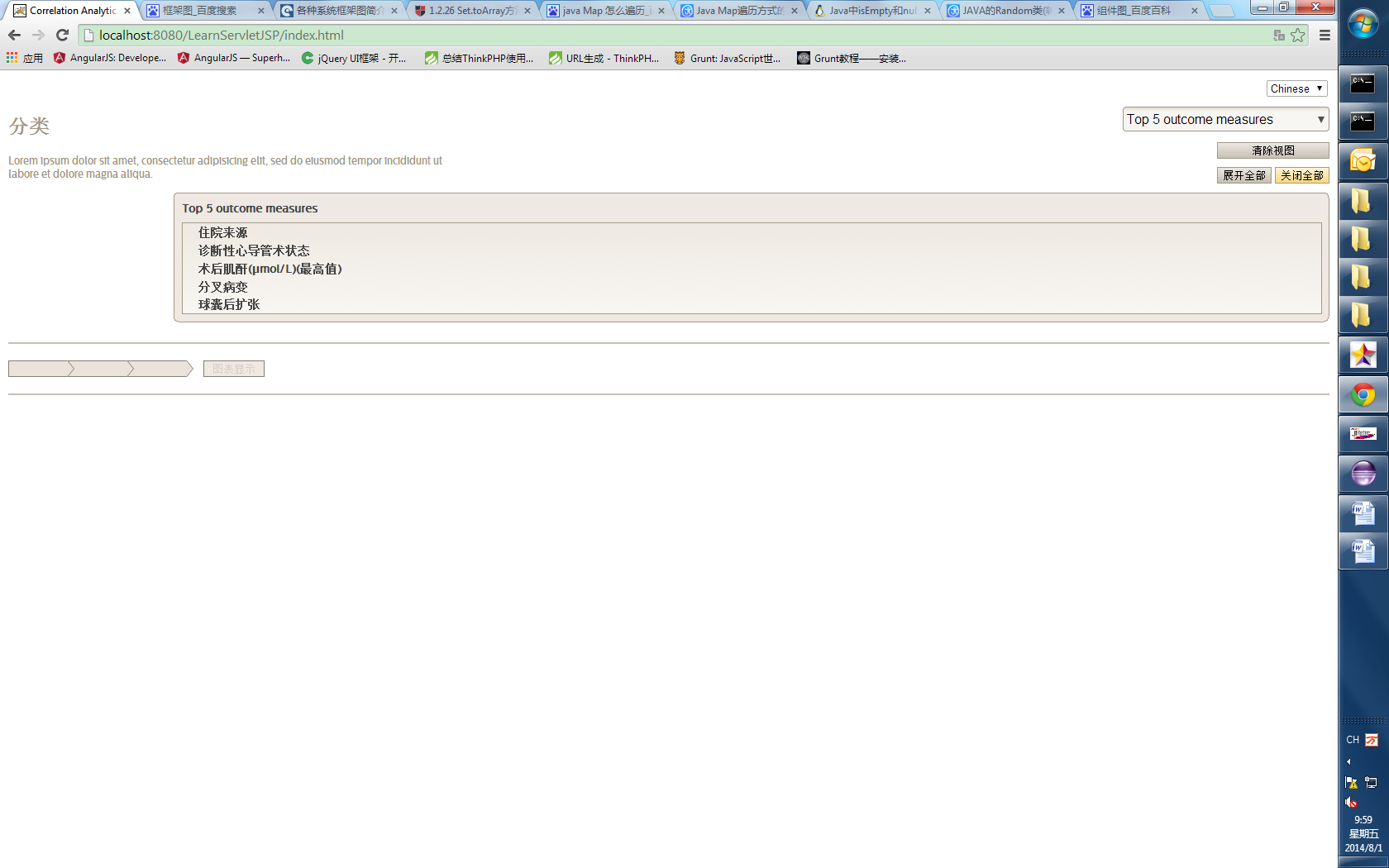


Figure 16 Top 5 outcome measures

## view Attribute correlation

workflow inside:

1. click targeted attribute from category correlation view or top 5 outcome measures
2. submit the data using the “post” method, where the post data is {selected: attribute}
3. parse data from URL, call correlation analysis module, get selected correlation in json format
4. visualization of attribute correlation, as shown in figure 17

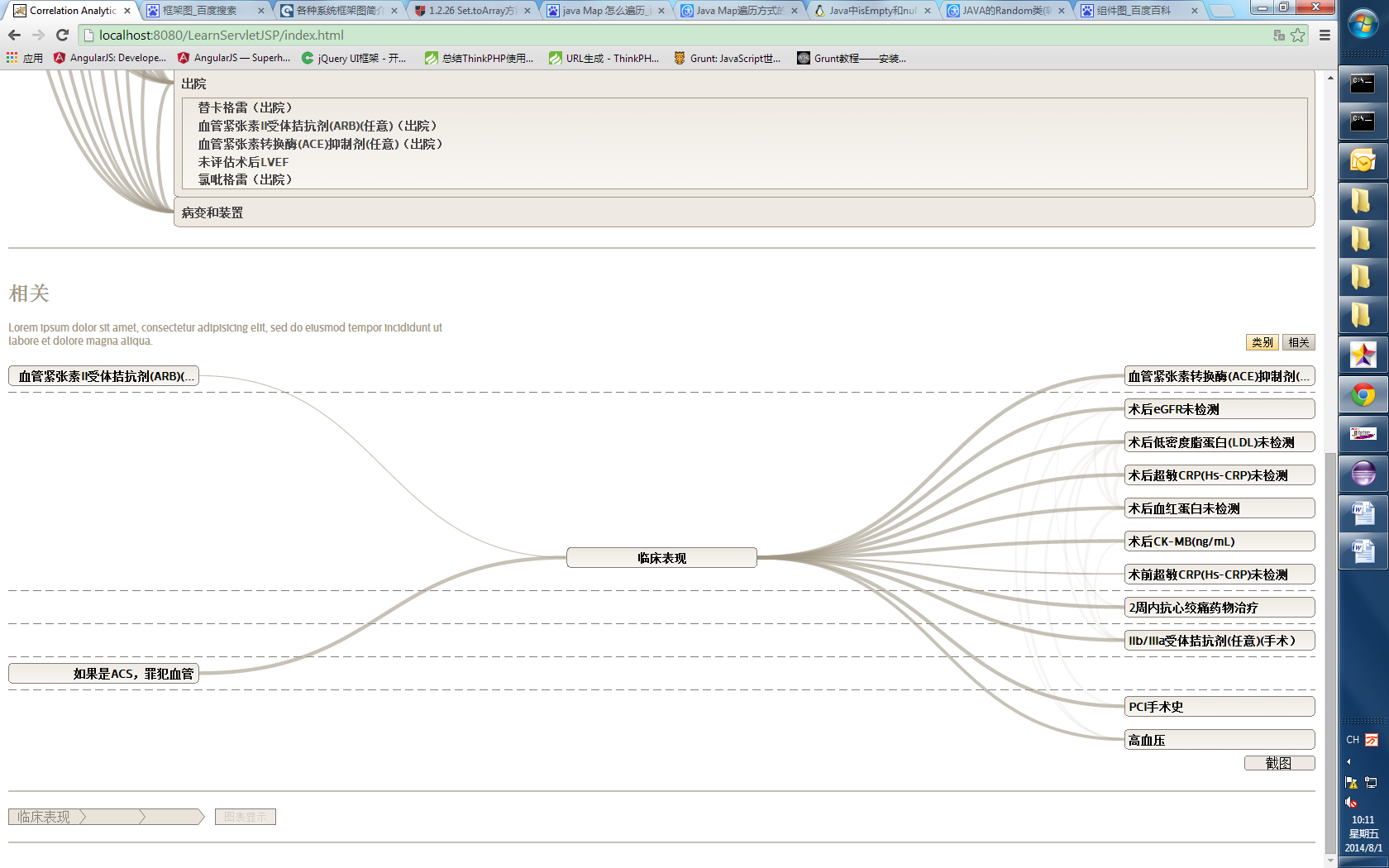


Figure 17 attribute correlation

## view bar chart 2D

1. click one related attribute from attribute correlation
2. click the button “show graph”, and then submit data using post, where the post data is {selected: attributes}
3. parse data from URL, call correlation analysis module, get the 2D bar chart in json format
4. visualization of bar chart 2D, as shown in figure 18

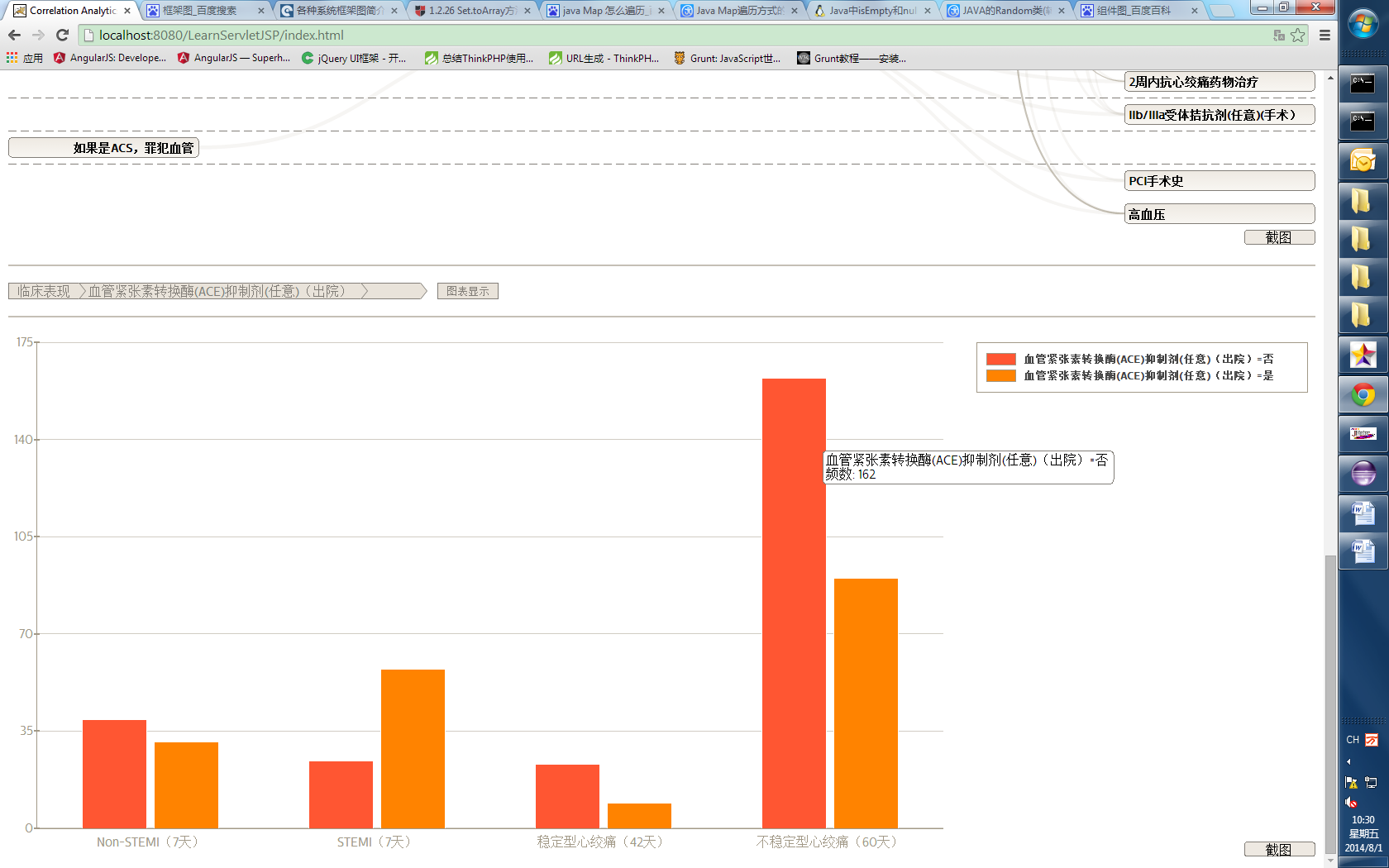


Figure 18 bar chart of two attributes

## view bar chart 3D

1. click two related attributes from attribute correlation
2. click the button “show graph” and then submit data using post, where the post data is {selected: attributes}
3. parse data from URL, call correlation analysis module, get the 3D bar chart in json format
4. visualization of bar chart 3D, as shown in figure 19

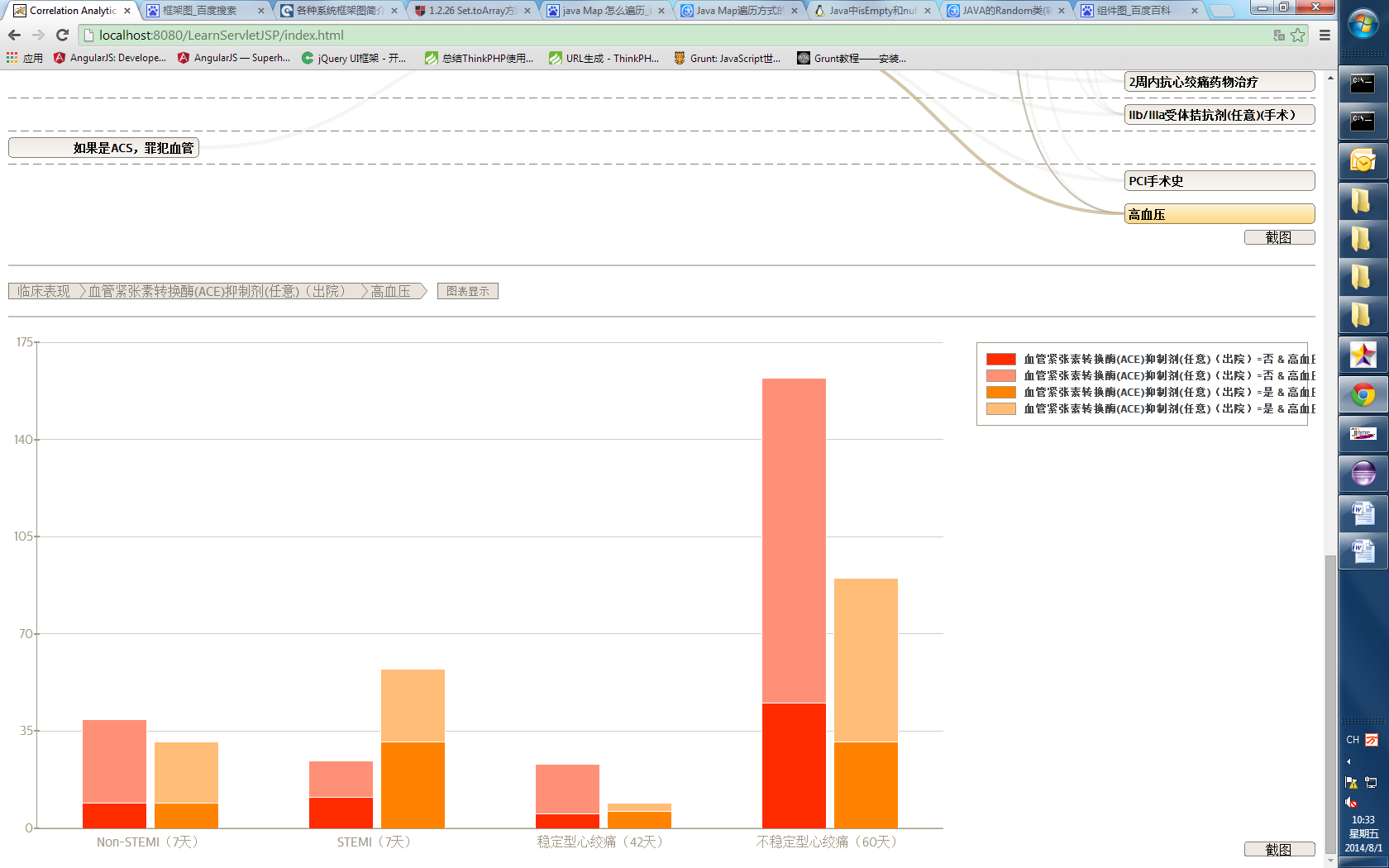


Figure 19 bar chart of three attributes

## view a list of patients

1. click a segment in the bar chart
2. patients are listed for the segment, as shown in Figure 20

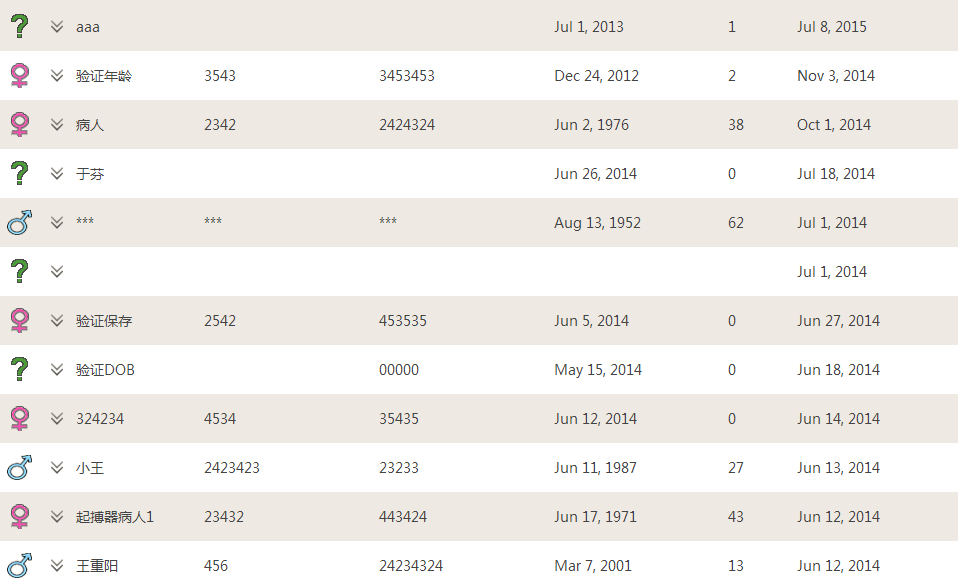


Figure 20 patient list

# Discussion and Conclusions

In physician’s view, current correlation analytic module is meaningful, and it has been accepted and transferred to BIU, BIU is integrating the correlation analytic module into the CDR platform. The correlation analytic module will also be integrated into DHP.

In future work, based on the result of correlation analytic module, pattern discovery and prediction models will be researched and developed to create more values from information from CNCDR.

# Acknowledgements

We would like to thank the following colleagues for the co-creation of the research product and their valuable insights from the market and product:

Data Visualization team from Design at Einthoven: Jeanne de Bont, Niels Laute and Jurrien Gosselink.

PINS at Einthoven: Niels Laute, Jurrien Gosselink, Jeanne de Bont

BU-HIT of Intellispace Cardiovascular Data Repository (CDR) solution team in Shanghai: Steven Bunting, William Lai, Susan Hu, Gary Jiao.

References

1. David Freedman, Robert Pisani, Roger Purves, Ani Adhikari. Statistics.

2. Pana-Ning Tan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining.