Published in final edited form as:
Methods Inf Med. 2014 ; 53(3): 173–185. doi:10.3414/ME13-01-0075.

A Complementary Graphical Method for Reducing and Analyzing Large Data Sets*:

Case Studies Demonstrating Thresholds Setting and Selection

X. Jing and **J. J. Cimino**
Laboratory for Informatics Development, National Library of Medicine and NIH Clinical Center, National Institutes of Health, Bethesda, Maryland, USA

Summary

Objectives—Graphical displays can make data more understandable; however, large graphs can challenge human comprehension. We have previously described a filtering method to provide high-level summary views of large data sets. In this paper we demonstrate our method for setting and selecting thresholds to limit graph size while retaining important information by applying it to large single and paired data sets, taken from patient and bibliographic databases.

Methods—Four case studies are used to illustrate our method. The data are either patient discharge diagnoses (coded using the International Classification of Diseases, Clinical Modifications [ICD9-CM]) or Medline citations (coded using the Medical Subject Headings [MeSH]). We use combinations of different thresholds to obtain filtered graphs for detailed analysis. The thresholds setting and selection, such as thresholds for node counts, class counts, ratio values, p values (for diff data sets), and percentiles of selected class count thresholds, are demonstrated with details in case studies. The main steps include: data preparation, data manipulation, computation, and threshold selection and visualization. We also describe the data models for different types of thresholds and the considerations for thresholds selection.

Results—The filtered graphs are 1%-3% of the size of the original graphs. For our case studies, the graphs provide 1) the most heavily used ICD9-CM codes, 2) the codes with most patients in a research hospital in 2011, 3) a profile of publications on “heavily represented topics” in MEDLINE in 2011, and 4) validated knowledge about adverse effects of the medication of rosiglitazone and new interesting areas in the ICD9-CM hierarchy associated with patients taking the medication of pioglitazone.

Conclusions—Our filtering method reduces large graphs to a manageable size by removing relatively unimportant nodes. The graphical method provides summary views based on computation of usage frequency and semantic context of hierarchical terminology. The method is applicable to large data sets (such as a hundred thousand records or more) and can be used to generate new hypotheses from data sets coded with hierarchical terminologies.

*Supplementary material published on our website www.methods-online.com

Correspondence to: Xia Jing, Assistant Professor, Grover Center W357, Department of Social and Public Health, College of Health Sciences and Professions, Ohio University, Athens, Ohio 45701, USA, jingx@ohio.edu.

Keywords

Data mining method; data filtering method; threshold setting; threshold selection; data visualization; hierarchical terminology; data analysis; clinical data repository

1. Introduction

Graphical visualization techniques can help people comprehend data sets in intuitive and straightforward ways. However, large graphs can overwhelm human beings easily and also present challenges to common printing and display devices. Researchers have applied a number of techniques for visualizing health data coded with hierarchical terminologies, such as the use of TreeMap by Murphy and colleagues [1] and SpaceTree by Plaisant and colleagues [2]. Similarly, Keller and colleagues [3] have provided graphic laboratory data via the WWW-based graphic user interface and Kopanitsa and colleagues [4] looked for a generic method for visualizing medical data. Other examples include that Stiglic and colleagues [5] have pruned large decision trees to improve comprehension, McGuffin and colleagues [6] have developed metrics for quantifying the space-efficiency of tree representation. Although Stiglic's [5] work related to pruning, no details on threshold setting were provided. Our previous paper described a method for representing and reducing large data sets coded with hierarchical terminologies [7]. The method is suitable for a wide variety of data sets, however setting the thresholds used for filtering and depicting the data is not straightforward. The purposes of this paper are: 1) to describe our methodology for setting and selecting thresholds for filtering, 2) to demonstrate the application of our method to produce graphs that compare paired data sets ("diff graphs"), 3) to apply the method to four real data sets as case studies, and 4) to validate the method preliminarily using two of the data sets.

2. Background

2.1 Data Sources

The basic requirement for our method is that the data consist of frequencies of codes from a hierarchical coding system. Data may be coded as "leaf node" instances under a hierarchical classification system, or they may be coded at any level within the hierarchy. The hierarchy itself must be a directed acyclic graph consisting of either a single, strict hierarchy or multiple hierarchies. A wide variety of data in health care and biomedical literature meet these requirements. For example, clinical data repositories, such as the Biomedical Translational Research Information System (BTRIS) at the National Institutes of Health (NIH) [8] and the Informatics for Integrating Biology and the Bedside (i2b2) repositories [9] contain data encoded with hierarchical terminologies, such as the International Classification of Diseases, Clinical Modifications (ICD9-CM) [10], the Logical Objects, Identifiers, Names and Codes (LOINC)[11], the Systematized Nomen clature of Medicine-Clinical Terms (SNOMED-CT) [12] and RxNorm [13]. The specific descriptions about each data source are in the data acquisition section (3.1).

2.2 Data Filtering

Ideally, a graphical depiction of data should seek to include as much information as possible, but the reality is that large graphs can lead to cognitive overload. We have found empirically that graphs smaller than 120 nodes can be readily managed and compared, but once the number of nodes exceeds around that number, both the size and layout of graphs challenge comprehension and available printing and display devices. We conducted empirical tests by using three real data sets. Deciding the maximum size of a filtered graph requires a subjective judgment; however, criteria can include setting sizes that are manageable for manual comparisons between filtered graphs and that are manageable by available display and printing devices. As described in our previous paper [7], our method for filtering hierarchically coded data sets consists essentially of treating data counts (i.e., frequency of occurrence of specific codes or of patients associated with specific codes over a certain period of time) for each code in the terminology as *node counts*[a]. We then calculate *class counts* for each code (calculation in section 3.2). Note that while a code with multiple parents (in a multiple-hierarchy) will contribute its node count to the class count of each of its parents, node counts are never counted more than once for each ancestor's class count, since the sum is based on all descendant node counts, not children class counts. We then calculate a *ratio value* for each term (calculation in section 3.2), to provide a measure of the relative importance of the term's contribution to its parent, as compared to its sibling terms. *Percentiles* of class counts (i.e., the percentage of class counts which are below the selected class count) are also used as a reference for comparing filtered graphs. Specific examples are presented in section 3.4. The filtering process uses a *node count threshold*, a *class count threshold*, and a *ratio threshold* or their combinations to remove terms from the set. The similar *percentiles* of class counts (or node counts) are considered for threshold selection to make two graphs of different sized data sets comparable.

When we compare two data sets to look for the most significantly different terms, we create a "diff data set" including all the terms and two class counts (or two node counts), one from each data set. In general a two-sample test of proportion can be applied to identify the differences between the two data sets for each term.

$$CCP\,(term) = CC\,(term) \,/ \sum_{k=1}^{n} CC(term)_k$$

$$NCP\,(term) = NC\,(term) \,/ \sum_{k=1}^{n} NC(term)_k$$

CCP (term) is the class count proportion of the term; CC (term) is the class count of the term; CC $(term)_k$ is a class count of a term in the data set; n is the total count of the terms in the data set. NCP (term) is the node count proportion of the term; NC (term) is the node

[a]The italic is used to refer to specific terms used in this paper

count of the term; NC $(term)_k$ is a node count of a term in the data set; n is the total counts of the terms in the data set.

However, if most of the observed data points are single digits, then an exact test may be more appropriate. In diff data sets the p value from the statistical test is used as a threshold for filtering. Lower p values indicate a significant difference between the terms (e.g., ICD9-CM codes) in the two data sets. The less significant data points (i.e., terms with higher p values) will be removed. The interpretation of p values for our Case Study 4 is in section 3.3. Varying any one threshold continuously might or might not lead to a graph within this range. Varying multiple thresholds simultaneously can be more challenging. We describe our techniques for doing so in the Methods section, below.

2.3 Graphical Display

Our methods use Graphviz [14, 15], an open source software package initiated by AT&T Labs Research, to provide visualization of our reduced data sets as hierarchical graphs. One of the Graphviz layouts, called DOT, is particularly suitable for representing graphs with hierarchical structures. The rich formatting features, such as node colors and shapes, and link colors, weights (thickness) and arrow head directions and sizes provide abundant secondary information beyond simply the names and interconnections of nodes. We use these different features to convey different frequencies, aggregated frequencies, ratios and statistical significance of differences between different groups. The data input format required by Graphviz is straightforward and a clear instruction manual is available.

2.4 Definitions of terms

Important information refers to the nodes their semantics retained in the filtered graphs, including frequently used terms (high node counts), terms with frequently used descendants (high class counts), and nodes whose data contribute disproportionally more to their parents' class counts (high ratio values). In diff data sets, important nodes are those with node or class counts that most significantly different between the two data sets.

Optimal options: thresholds, alone or in combination, can have a broad range of values. An "optimal option" is one that produces graph that most closely meets both of the following conditions: 1) the combinations of different thresholds produce graphs in manageable size, e.g., fewer than 120 nodes; 2) to select the points that minimize the errors, i.e., the closest points to the regression plane.

Code : refers to a node in a hierarchical graph, corresponding to a term in a terminology. Code count, term count and node count all refer to the frequency of a node in a data set.

3. Methods

We illustrate our method through its application to real data sets. Section 3.1 describes research questions and data acquisition. The remaining sections described the transformation of the data into graphs, including modeling and computation (Section 3.2), creating "diff data sets" from paired data sets (Section 3.3), setting and selection of thresholds (Section

3.4), formatting data for Graphviz and the implementation of our solution (Section 3.5). Our intent is to lay out processes that others can follow for analyzing their own data sets.

3.1 Research Questions and Data Acquisition

The main steps in our method include data preparation, data calculation and manipulation, and data filtering and visualization (Figure 1). Each step will be illustrated with details in the following sections.

3.1.1 Case Study 1: Annual Profile of ICD9-CM Distribution in BTRIS—The BTRIS repository contains data from many different source systems at NIH, including two electronic health record (EHR) systems at the Clinical Center (the NIH's research hospital) used from 1976 to 2004 and from 2004 to present. The terminologies used to code data in the source systems are unified in the Research Entities Dictionary (RED), a multi-hierarchical terminology that includes terms for ICD9-CM diagnoses and procedures, laboratory tests, medications, radiology procedures and clinical research concepts. In this paper, we obtained de-identified summary data from BTRIS regarding research subjects' discharge diagnoses coded by the hospital medical records department using ICD9-CM. These data were obtained with approval from the NIH Office of Human Subjects Research Protections.

We obtained the total number of ICD9-CM codes and their code counts appearing in all patient discharge diagnosis records for the year 2011 from BTRIS. This data set reflects total reasons for hospitalization without consideration of number of codes per patient or per admission. Visualization of this data set can provide users such as hospital administrators with an assessment of the most heavily used ICD9-CM codes that may be in the hospital at any given time. In total, there are 21,933 codes in our ICD9-CM coding system. Attempting to visualize all these data in a single graph would be overwhelming if it included, say, 40 codes for diabetes mellitus, 420 codes for tuberculosis, or 1110 codes for cancer diagnoses, while seeing a single code for each class of disease that summarizes the number of diseases below it may be more manageable. This higher level overview can provide useful data for setting resource priorities. In some cases, being able to see subclasses (such as malignant neoplasm of stomach, acute lymphoid leukemia or acute myocardial infarction) under large classes (such as neoplasms or diseases of circulatory system) may be more helpful.

We queried BTRIS to get a node count (i.e., ICD9-CM code counts) of each ICD9-CM code for the year 2011. To this initial set, we added all ICD9-CM terms that were ancestors of those in the actual data set and assigned each a node count of zero.

3.1.2 Case Study 2: Annual Profile of Patient Distribution in BTRIS—We also obtained the total number of unique individuals (i.e., distinct patients) who had been coded with any ICD9-CM codes in any of their patient records for the year of 2011 (across one or multiple hospitalization admissions) from BTRIS. When a code appears multiple times for an individual, the code is counted once only. This provides a general representation of the ICD9-CM codes occurring within the most individual patients in 2011. Graphs based on this type of data can be used to study variation in disease prevalence between years, seasons, or institutions (geographic variation). The use of code counts (i.e., Case Study 1) and subject

counts (i.e., Case Study 2) provides different views of disease profiles, analogous to incidence and prevalence.

We queried BTRIS to get the number of distinct subjects (i.e., no duplicate count for any patient) with each ICD9-CM code during the course of 2011. To this initial set, we added all ICD9-CM terms that were ancestors of those in the actual data set and assigned each a node count of zero. All the node counts of ICD9-CM codes are distinct patient counts in this case.

3.1.3 Case Study 3: Annual Profile of MeSH Terms Usage in MEDLINE—

MEDLINE is the world's largest database of medical literature citations. The citations are indexed using the hierarchical Medical Subject Headings (MeSH). The Unified Medical Language System (UMLS) [16] [17] provides a file (MRSAT.RRF) containing the annual counts of citations indexed by the MeSH terms, with an indication as to whether terms are "major" or "minor" topics in the related journal article. In this paper, we use major MeSH term counts from the MRSAT.RRF file to examine the active areas in biomedical publications indexed in MEDLINE.

This data set was obtained from the MRSAT.RRF file of the UMLS Metathesaurus. It consisted of major MeSH terms and the frequencies with which they were used for indexing the publications in MEDLINE in 2011. A graph with this type of data set provides an overview of frequent publication topics and can serve as an indirect indicator for current important research topics. Comparisons of different years of this type of data sets, for example an analysis of data sets in 1971, 1981, 1991, 2001 and 2011, can identify trends in publication and, presumably, research.

We queried the MRSAT.RRF from the second UMLS version in 2011 (UMLS2011AB) to get the usage data of MeSH terms, i.e. terms and frequencies. We added all MeSH terms that were ancestors (from the UMLS file MRHIER.RRF) of those in the actual data set and assigned each a node count of zero.

3.1.4 Case Study 4: ICD9-CM Data Comparisons between Pioglitazone and Rosiglitazone in BTRIS and MIMIC-II—

The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) database [18] is a free, publicly available de-identified database collected from a large population of patients from intensive care units (ICUs) [18]. For this study, we use MIMIC-II data related to patient problems (coded in ICD9-CM) and medication administration.

Pioglitazone (brand name Actos) and rosiglitazone (brand name Avandia) are medications used to treat non-insulin-dependent diabetes mellitus [19, 20]. They have similar actions, but are associated with different adverse events [21]. A possible increased risk for bladder cancer from pioglitazone [21] and an increased risk of cardiovascular events from rosiglitazone have been reported [1]. Previous researchers have shown that retrospective analysis of data sets from a clinical data repository can be used to detect some of these differences [1]. We therefore chose this as an example to demonstrate the use of our visualization method for such detection in both BTRIS and MIMIC-II.

We queried BTRIS and MIMIC-II to identify all patients with any record of administration of either pioglitazone or rosiglitazone, and the first date of administration of each medication. Patients with evidence for administration of both medications were excluded. We then obtained the unique ICD9-CM diagnoses associated with each patient that were dated after the first date of either pioglitazone (AP) or rosiglitazone (AR). This produced four sets of data: AP-BTRIS, AR-BTRIS, AP-MIMIC-II and AR-MIMIC-II. No specific dates or subject identifiers were included in the data set provided from BTRIS or MIMIC-II, i.e., only summary data of patient counts were involved in the final analysis. Although BTRIS and MIMIC-II may include multiple instances of the same ICD9-CM diagnosis for a given patient, these instances do not necessarily reflect new occurrences of disease, but rather instances in which a provider entered a diagnosis on the patient's problem list. The summary data from BTRIS and MIMIC-II, therefore, only indicated that an ICD9-CM code appeared at least once in the record after the medications of interest. The class counts in this data set are subject counts rather than ICD9-CM code counts.

3.2 Modeling the Data and Calculating Thresholds

For Case Studies 1 and 3, the data model is the same apart from the different terminology hierarchies, i.e., ICD9-CM for Case Study 1 and MeSH for Case Study 3. The frequencies (i.e., code counts) of the ICD9-CM codes in Case Study 1 and the frequencies of MeSH terms in Case Study 3 were used as node counts directly.

For Case Studies 2 and 4, we modeled the relationships between patients (i.e., subjects) and their ICD9-CM diagnoses as graphical connections between patient nodes and ICD9-CM nodes. When combined with the hierarchical connections among ICD9-CM codes, this resulted in graphs of an upper ICD9-CM tree (a strict directed tree hierarchy) and a bottom layer of leaf nodes (each one corresponds to a patient), each of which was linked as a child of one or more ICD9-CM nodes correspondingly. For our filtering method, we assigned a node count of zero to each ICD9-CM code and a node count of 1 to each patient node. The hierarchical relationships in the terminologies were used to identify ancestors, calculate class counts for each node, and produce graphs. Figure 2 shows a simple example of this arrangement.

We calculate class counts (i.e., the sum of the node counts of itself and all its descendants') using an *ancestor-descendant table* that contains one row for every node and each of its descendants. For subject counts, such as Case Studies 2 and 4, the class count of an ICD-9-CM node equals to the total number of distinct patients that are modeled as its descendants; for node term counts, such as Case Studies 1 and 3, the class count of an ICD-9-CM node (or a MeSH term) equals to the sum of its own node count and of the node counts of all its descendant terms.

In addition to node and class counts, our method can also make use of a node ratio. The ratio equals the class count of a term divided by the class count of the term's parent. This value reflects the relative contribution of each child node to their parent node, therefore to identify the important (i.e., the bigger contributors to their parent node) child nodes that might be retained after filtering, depending on threshold settings. We used Microsoft Excel to calculate percentiles for each class count in the data set: if we rank all the class counts in

descent order in the data set, a percentile of a selected class count is the percentage of the class counts in the data set lower than the selected class count.

3.3 Creation of diff Data Sets

In addition to displaying data set summaries in each of our case studies, we also wish to compare data sets; for example, Cases Studies 1, 2 and 3, we can compare data sets from different years to identify trends although these are not the focus this paper. In order to visualize these comparisons, we create “diff” tables from data sets being compared. We use Case Study 4 to demonstrate how to produce diff table in order to use both graphical and statistic methods to highlight the differences between different data sets. In Case Study 4, we want to detect different conditions associated to rosiglitazone compared to pioglitazone, so we compare data sets for pioglitazone (BTRIS-AP and MIMIC-II-AP) and rosiglitazone (BTRIS-AR and MIMIC-II-AR) respectively by using following steps:

1. Obtain all the class counts for each ICD9-CM code in two groups (e.g., BTRIS-AP and BTRIS-AR for BTRIS data sets) as described above.
2. Create a single table including three columns: ICD9-CM codes, class counts for AP and AR in BTRIS.
3. Create a fourth column that contains the p values from a two-sample test proportion (we used “prop.test” R[22]) that tests for differences between the two class counts for each ICD9-CM code. For each ICD9-CM code, the proportion equals to the patients’ count each code divided by the total patients’ count in their own group.

An unfiltered “diff” graph can now be constructed using the first and fourth columns in the table. The p values represent probability of observing a difference due to randomness in the proportions of subject counts (coded with ICD9-CM codes) between the AP and AR groups when we assume there is truly no underlying difference between the AP and AR groups. A low p-value indicates that it is very unlikely that AP and AR groups have the same underlying proportion of subject p values of 0.05, 0.01 or 0.001 can be used as thresholds for filtering diff graphs, i.e. only keep the ICD9-CM codes whose p values are lower than the threshold p value and their ancestors.

3.4 Thresholds Selection for Class Counts and Ratio Values

For a data set that can be represented in a hierarchy, every node has a class count and any value can be used as a class count threshold. The higher class count threshold means fewer nodes are included in the filtered graph and only the most important nodes are in the filtered graph. Similarly any ratio value can be used as a threshold too. The higher ratio value threshold means only the most important contributors are kept in the filtered graph. However, it is not practical to try all possible thresholds and their combinations to obtain all possible filtered graphs. Our purpose is to select rational thresholds for the optimal option of filtered graph (i.e., manageable size with the minimized errors) without losing important information. We used the following steps in setting and selecting class count and ratio thresholds:

1. Sort the data table according to class count in descending order and get the class count at about the 150th position; this is used as the initial class count threshold.

We expect a filtered graph with 120 nodes and below by using both class count threshold and ratio value threshold. By including a ratio value threshold, the class count threshold of the filtered graph will move lower than 120th position (in descending order). The initial class count threshold therefore starts from 150th. We will then adjust the class count threshold in both directions (increasing and decreasing) in order to select the optimal filtered graph.

2. Use the initial class count threshold with ratios of 0, 0.2, 0.4, 0.6, 0.8 and 1 to produce graphs and to identify the number of nodes in each filtered graph.

3. Adjust (both increase and decrease) the class counts threshold from the initial value and calculate the graph size (number of nodes) for each permutation of class count threshold and ratio;

    i. 20% of the initial class count threshold can be used as the interval for increasing or decreasing class count thresholds (we tested 10% and 5% of initial class count as intervals; these did not produce a better 3D plot for thresholds selection); for example if the initial class count is 2000, then the next increased class count threshold will be 2200 and the next decreased class count threshold will be 1800.

    ii. Stop increasing the class count threshold if the combination of a class count threshold with 0.2 produces smaller graph than 90 and stop decreasing the class count threshold if the combination of a class count threshold with 0.8 produces larger graph than 120.

    iii. For one class count threshold, if the final sizes of the graphs for all ratio values (0–1) are smaller than 90 or greater than120, then exclude the class count threshold.

4. Put all the results from step 2 and 3 into a matrix including data codes (i.e. ICD9-CM codes or MeSH codes), class count thresholds, ratios, and numbers of nodes in the filtered graphs.

5. Highlight the rows that the combinations of class count thresholds and ratio values produce graphs between 90 and 120 nodes.

6. Plot the matrix (i.e., class count thresholds, ratio values and the final graph sizes) in a 3D plot and fit a regression plane in R (scatter3dplot package) [23]; the regression plane is the 2D plane that minimizes the total of the squared distances between the observed values and the closest point on the regression plane. We used the least square distance (residual, a function in R, the difference between an observed value and the fitting value from the model (i.e., regression plane), is used as an indicator in Table 1) as one of the methods to select the combination of class count threshold, ratio value and final graph size to minimize the errors.

7. Select points that are closest to the plane in the 3D plot according to the least squared distances.

8. Get the overlap of step 5 and 7.

9. Get the percentiles for all the selected class count thresholds.

We use Case Study 3 to illustrate the process. Table 1 presents the matrix table for Case Study 3. Figure 3 is the corresponding 3D plot with regression plane of the data in Table 1. The relationships among different class count thresholds, ratio value thresholds and the final nodes in the filtered graphs have been shown in Table 1 and Figure 3 respectively. Although the observed data points in the 3D plot do not fit into a strict linear relationship, there are important linear components. The R-squared (fraction of variance explained by the model), which is from the Summary function for Linear Model in R, equals to 0.9119. More complex transformations (not shown) do not provide a significant improvement (lower than 2%). The points closest to the regression plane are indicated with * in Table 1 and highlighted in Figure 3. The overlap from step 5 and 7 are marked with * and highlighted in Table 1 and Figure 3.

There are other considerations for threshold selection. For Case Study 3 we selected a class count threshold of 10000 and a ratio threshold of 0.4 to produce the graph for further analysis. The reasons for these selections are:

1. From pilot graphs, we can see the nodes with more than 5 children are very common and the average number of children is 11, so that each child node's contribution will, on average, be less than 10%. Therefore, a ratio value 0.4 is considered above average.
2. The final node count is 95, which is within the target range of 90 to 120.
3. Comparing with another row, which is highlighted with * (8000, 0.4 row in table 1), the 10000 and 0.4 row is closer to the regression plane (−3.958578 vs −5.045806).

The main principles we use in selection thresholds are:

i. To produce a manageable-size (filtered) graph, i.e. the number of nodes is less than 120 after filtering.
ii. To include more nodes under the condition of a manageable-size graph.
iii. To avoid using extreme ratio value thresholds to keep relative important nodes.iv) To select the data points with the minimized errors.

The ratio value is used as a threshold to guarantee the filtered graph has not only the nodes with higher class counts but also the nodes that are relatively important comparing to their sibling nodes. For a large hierarchy like MeSH, the average children are 11 and the average ratio value is 0.09. The ratio value thresholds of 0.6 or 0.8 will filter out large numbers of important nodes.

Apart from keeping relatively important nodes, the ratio and percentile thresholds are used for comparison purposes, for example to compare different years' data (e.g. data sets from 2011 and 2005) or data from different sources (e.g. from BTRIS and MIMIC-II). Keeping the graphs roughly the similar size, with similar ratio values as thresholds and similar percentiles for class count thresholds (absolute class count thresholds may not equal), helps make the graphs more comparable.

3.6 Solution for Implementation and Creation of Graphviz Input Files

We queried and exported data from SQL Developer for MEDLINE, MIMIC-II data sets and Microsoft SQL Server for BTRIS data sets. We imported the data sets into MySQL and then used a home grow program pipeline (available on request, mainly involving PHP and MySQL) to carry out testing, threshold tuning, threshold setting, computation and data manipulations to get the appropriate file format for Graphviz. Microsoft Excel was used to summarize results and to get the percentiles for the class count thresholds. The summarized result in CSV format (a table similar to the Table 1 with first four columns) was used for prop.test, summary and scatter3dplot (for 3D plots and regression plane) in R.

The Graphviz input file includes the type of graph (e.g., the DOT layout), the relationships between nodes (i.e., hierarchies of the structure or path), and the characteristics of the nodes and arrows (node shapes, node styles, colors, directions of arrows, and sizes of arrow heads). When a Graphviz input file is created from the table of node counts and class counts and the path file, the different attributes for nodes are assigned according to a predetermined set of desired conditions. For example: colors of nodes can be used to represent value ranges of class counts; different types of arrows (based on colors, arrow head types, sizes and directions) can be used to represent different ratio values. For a diff graph, different colors can be assigned to different nodes according to the different p value ranges: for example: pink for nodes with p value 0.01 and blue for nodes with p value > 0.01 and < 0.05.

4. Results

In all graphs in this paper, parent-child relationships are depicted by the placement of the nodes: i.e., the parent nodes always appear above their corresponding child nodes. The arrow head directions here are used to convey the contributions from child nodes to their parent nodes; they should not be used for judging directed cycles.

4.1 Case Study 1: 2011 Profile of ICD9-CM Data in BTRIS

There are 6175 distinct ICD9-CM codes in our BTRIS 2011 data set. Figure 4 shows the resulting graph sizes for varying class count thresholds. When class count thresholds increase, the final nodes in the filtered graphs decrease. Figure 4 shows the exact filtering results based on class count thresholds. The original graph for this data set has over 6000 nodes, while Figure 4 shows the filtered graph sizes between 130 to 90 nodes. A class count threshold of 2300 results in a graph of 105 nodes (shown in Figure 5 in Appendix). Figure 5 in Appendix shows the summary view of the data set for Case Study 1 according to ICD9 codes' class counts. The size of the filtered graph is 1.7% of the origi nal one (105/6175). A class count threshold of 2300 represents the percentile of 0.98308, that is, over 98% ICD9-CM codes in this data set have a class count that is smaller than 2300. The circle area in Figure 5 highlights nodes that do not appear in the graph for Case Study 2 (below): malignant neoplasm of bone, connective tissue, skin & breast and secondary malignant neoplasm of lung.

4.2 Case Study 2: 2011 Profile of Patients Coded with ICD9-CM in BTRIS

As in Case Study 1, there are 6175 distinct ICD9-CM codes in our BTRIS 2011 data set. Figure 6 demonstrates the filtering details for the filtered graphs with the final nodes between 90 and 130 nodes, which correspond to different class count thresholds. For example, a class count threshold of 750 results in a graph of 105 nodes (Figure 7 in Appendix). Comparing with Figure 4, both figures have similar decreasing trends with increasing class count thresholds, however they do show different details due to the two different underlying data sets. Figure 6 looks much more linear comparing to Figure 4. However both Figure 4 and Figure 6 show important linear components, which support the methodology (linear model) we used to get filtered graphs. Figure 7 in Appendix shows the summary view of the data set for Case Study 2 according to subjects' class counts. The size of the filtered graph is 1.7% of the original one (105/6175). The percentile of 750 among the whole class counts is 0.98302. Table 2 compares the filtered graphs from Case Studies 1 and 2. Although the absolute class count thresholds are different, the final sizes and the percentiles of class count thresholds make the two filtered graphs comparable.

These are the particular areas that appear only in this subject counts group by comparing with the filtered graph produced from Case Study 1:

- Symptoms involving respiratory system and other chest symptoms.
- Obesity.
- Family history with psychiatric condition.
- Long-term drug use. Most of the nodes are overlapped between the two filtered graphs produced from Case Studies 1 and 2, although the rank orders of the ICD9-CM codes may be different in each group.

4.3 Case Study 3: 2011 Profile of Major MeSH Term Usage in MEDLINE

There are 23,086 MeSH terms in our data set. We chose a class count threshold of 10,000 and ratio threshold of 0.4, which results in a graph with 95 nodes (0.41% of the original graph size, 95/23086). Figure 8 in Appendix shows the whole filtered graph as the summary view and Figure 9 in Appendix shows the most heavily represented areas in an enlarged view. The filtered graphs are too large to be printed out. For example, literature about Diseases (especially neoplasms) and Chemicals and Drugs (especially Amino Acids, Peptides and Proteins) appear especially frequently in 2011. The combination of class count thresholds and ratio value thresholds shows a clear decreasing trend in terms of the final nodes in the filtered graphs.

4.4 Case Study 4: ICD9-CM Data Comparisons between Patients Taking Pioglitazone versus Rosiglitazone in BTRIS and MIMIC-II

Table 3 summarizes the subject counts in the after-pioglitazone and after-rosiglitazone groups in both BTRIS and MIMIC-II data sets. Table 3 also serves as a basic profile for each group in both data sets. Table 4 and Table 5 summarize subject counts coded with the ICD9-CM codes that are consistently and significantly different (in grey shading cells) between after-pioglitazone and after-rosiglitazone groups across data resources (Table 4 for

MIMIC-II data sets and Table 5 for BTRIS data sets). The consistent and significant codes cross data sets are an important indicator that these medical events may relate to the corresponding medications. Especially the grey shading cells in both Table 4 and Table 5 are important and interesting events that need further investigations for a better interpretation about the relationships between the medications and the medical events. In the MIMIC-II data set patients in the afterrosiglitazone group have more frequent heart-related problems and the results have been reported in Table 6.

The subject counts coded with the ICD9-CM codes with significant difference between the after-pioglitazone and the after-rosiglitazone groups in both the MIMIC-II and the BTRIS data sets are shown in the online Appendix (Figures 10 and 11 in Appendix), with filtering based on p values from a two-sample test of proportion; different colors are used to show different p values. After filtering MIMIC-II data set shows 3.1% (50/1611) of the original data set (Figure 10 in Appendix), while BTRIS data set shows 1.9% (45/2411) of the original data set (Figure 11 in Appendix).

5. Discussion

5.1 Interpretation of the Case Study Results and their Significance

Our filtering method provides manageable data sets by trimming out relatively less important data. The resulting graphical views may reveal, through the uses of intra-node relationships, patterns in the data that are not obvious from viewing numbers in a spreadsheet. For example, while any one node in the graph may be only mildly interesting, a cluster of such nodes can attract attention that leads to further analysis.

5.1.1 Case Studies 1 and 2—While the methods we used for calculating class counts in Case Studies 1 and 2 are similar, the differences in underlying data (counts of ICD9-CM codes versus counts of patients, respectively) lead to different kinds of interpretations of the results. The Case Study 2 removes all the repeated codes for each patient. The structures of the original hierarchical graphs are therefore identical but the node counts are different (numbers of patient diagnoses versus numbers of patients, respectively). As a result, the two case studies produce different graphs after filtering and are useful for drawing different conclusions from the original data sets (common patterns of disease versus common patient types).

For example: secondary malignant neoplasm of lung and anemia are only significant in the Case Study 1 filtered graph, which tells us that these are important reasons for hospitalization in 2011 due to repeated admissions for a subset of patients, rather than occurring in a large subset of all patients. Respiratory system symptoms and obesity appear only in the filtered graph from Case Study 2. This tells us there are more patients who have respiratory symptoms and obesity in 2011. These two cases are used to demonstrate the appropriate selection of class count thresholds and the use of percentiles for comparison purpose. Although our methods provide summary graphs for large data sets, Case Studies 1 and 2 demonstrate the application of our methods one step further: the comparison of filtered graphs. Because Case Studies 1 and 2 produce different filtered graphs due to different underlying data sets, valid comparison requires the graphs to be similar in size and

percentile. Case Study 1 can also be used as an indicator for diseases distribution; comparison of different years' data sets can reflect the changes (e.g., research focuses or protocols distribution) in the research hospital; Case Study 2 can reflect the real population covered by the hospital and the changes over different periods if multiple years data are available. Both cases are useful for building a precise profile for the hospital. Furthermore, comparisons among hospitals may also lead to interesting findings.

5.1.2 Case Study 3—The filtered graph from Case Study 3 clearly shows the most heavily represented areas are Proteins, Chemical Actions and Uses, Neoplasms, Investigative Techniques, and Diagnosis. Missing is Organic Chemicals, whose class count is in the top 20 of the data set (45,278); however, it is filtered out since its ratio is 12.7% (45,278/357,379, parent term Chemicals and Drugs). The filtering demonstrated in Case Study 3 provides an aggregated view about the important topics in biomedical publications. We find that the graph of Case Study 3 provides a useful overview of the nearly 1 million 2011 citations with over 23000 MeSH terms. Other possible use case for Case Study 3 includes comparisons of different years of usage of MeSH terms to show changing trends in literature and, presumably, research.

5.1.3 Case Study 4—Using our filtering method, we can see clearly from Figure 10 in Appendix that more patients in the after-rosiglitazone group in the MIMIC-II data set had cardiac diagnoses. This finding is consistent with recently discovered adverse effects related to rosiglitazone[1]. Case Study 4 is also an example of using our method establish new hypotheses. There are some interesting findings in Case Study 4 (the exact results have been listed in Table 4 and Table 5) about pioglitazone: 1) there are significantly more patients associated with chronic liver diseases, neurotic disorders, personality disorders, and nonpsychotic mental disorders in the after-pioglitazone group in both the MIMIC-II and BTRIS data sets; 2) in the MIMIC-II data set the after-pioglitazone group also had significantly more patients associated with chronic liver diseases, neurotic disorders, personality disorders and nonpsychotic mental disorders, diseases of the musculoskeletal system and connective tissue, rheumatism, diabetes with ketoacidosis and other history of health hazards when compared to the after-rosiglitazone group. It is premature to state that pioglitazone causes these conditions or to interpret the exact relationship between these medical events and pioglitazone; however, our method highlights areas that may be worthy of further investigation. Other possible use cases for Case Study 4 include comparison of adverse events related to two similar procedures or two time points (such as before and after a medication administration or a procedure).

5.2 Comparison with Other Methods

There are many possible ways to visualize large sets of health data. For example, Murphy and colleagues have visualized the data from their study [1] using TreeMap [24] as a way of identifying areas in a hierarchical terminology that show high correlation with clinical findings [Murphy SM. Personal communication]. They did not choose to filter their data set but rather allowed TreeMap to show successive levels in a tree structure using smaller and smaller rectangles nested inside larger ones. A direct comparison of the Graphviz and TreeMap presentation on the same data set (including the addition of filtering to the

TreeMap depiction) to determine which one leads to better human comprehension would be worthwhile. However, caution is needed with TreeMap when the terminology (such as MeSH) contains multiple hierarchies, as the results in multiple copies of rectangles throughout the graph, while Graphviz handles such structures with multiple parent-child links.

Plaisant and colleagues created an interactive tree browser called SpaceTree [2] for visualization of large graphs. Their study compared SpaceTree with two other tools in completing designed tasks. Although as an interactive browser, SpaceTree can be used to hide or display certain parts of large graphs easily, it is not used for comprehension of data. While Graphviz is limited in terms of interaction, it is very good at representing different attributes and complex tree structures of graphs using colors and line types. These features are helpful for communicating additional information about the data sets. The combination of our method, algorithms and filtering functionalities with an interactive tool like SpaceTree may lead to additional comprehensibility.

Munzner and colleagues [25] developed an algorithm for automatically comparing the structural differences between two graphs. Their method associates every node in one graph with the most similar node in the other graph and compares the structural differences by measuring and computing similarities between nodes. In our method, we always filter data sets first and get manageable-size graphs, and then any comparison after filtering (e.g., two data sets in two different years or similar data sets from two sites) focuses on the precise comparison, which cannot be achieved by their method. In our methods, the two comparable graphs use the same hierarchy, so the pure automatic structural comparison will benefit our purpose in a very limited way. Apart from that, the sizes of the filtered graphs are manageable, and it is easy to get precise comparison results manually.

5.3 Limitations

Graphical visualization techniques in general are well known to be useful for understanding complex data, and we believe our filtered hierarchies are intuitive and obvious to the users. However, only an objective and systematic evaluation can determine whether the approach is useful. Establishing a gold standard for “understanding the data sets” is challenging, while waiting to see if potential hypotheses stimulated by graphic presentation bear fruits will be slow. Plaisant’s method [2] (i.e. to use different tools to complete designed tasks) can be a good evaluation point from which to start. A complete evaluation of the utility of our method requires an elaborate study, which is beyond the scope of the current paper. Future evaluation about the exact relationship between reducing graph size and increasing comprehension will need a metric for comprehension and then measure the level of comprehension and review time needed to achieve that comprehension, furthermore to look for measurements of the possible impacts by using our methods.

Another limitation is we cannot use a simple formula to express the methods. We developed the method to calculate class counts and ratio values, we then use class count thresholds, ratio thresholds and least distances to the regression plane for filtering purposes. However, there are no specific weights for each type of threshold. According to the R squared value in Case Study 3, the linear model fits well to the data set. We use the thresholds in the order of

class counts, ratio values and the least distances. The range of 90–120 nodes in the filtered graphs is empirical, so these are not rigid cutoffs. In Case Study 3 (i.e. the grey rows in Table 1), we keep 88 and 121 nodes as candidates. We use the row of 10000 and 0.4 to produce the graph due to the less distance to the regression plane. However the other row with 8000 and 0.4 should be considered too if the purpose is to understand the MEDLINE data set of 2011, because the distances to the regression plane are not statistically different. Nevertheless, our threshold setting and selection principles, steps can still be used if the filtered graphs are out of 90–120.

Other limitations are not related to the methodology per se, but they are related to the underlying data sets we used in the case studies. For example: *Coding quality* of the original data is closely related to the results of the graphic method since all the further results depend on good quality coding at the first place. ICD9-CM codes were all coded by the medical record department in Case Study 1, 2 and 4 and MeSH terms were indexed by biomedical experts in Case Study 3, so we believe the coding quality is relatively reliable in our case studies. In Case Study 4, if we can get ICD9-CM codes before and after medication administration, diff table can be used to minimize the possible bias, however there are not sufficient subjects in before medication groups in both the MIMIC-II and the BTRIS data sets. So we can only use two after-medication groups without considering the ICD9-CM codes before medication administration.

Another limitation of the underlying data sets is related to the secondary data use. The ICD9-CM diagnoses associated with the patients reflect a list of prominent problems associated with admissions to the hospital. However, there is no guarantee that the lists are complete. In particular, a patient who has an event such as a myocardial infarction is not likely to be seen acutely at the NIH Clinical Center, since there is no emergency room here. The likelihood that “myocardial infarction” will be included in the patients’ problem list on a subsequent encounter at the Clinical Center is unknown. Conversely, the presence of a diagnosis on a patient’s problem list does not necessarily indicate a new event, but may be repeated for each encounter if the problem is still active. We therefore limit our use of the data to count each diagnosis only once per patient, reflecting prevalence rather than incidence. Finally, the appearance of an ICD9-CM code on a problem list does not necessarily imply that the patient has the diagnosis. These codes are often annotated with text such as “rule out” or “sibling of patient with…” All of these limitations, and others, are typical of EHR data coded with ICD9-CM. Because the data were not collected answering the specific research questions, it may inevitably have assumptions in using the data secondarily. It is necessary to be cautious in interpreting the secondary use of data results.

6. Conclusions

This paper demonstrates a filtering method that is applicable to data sets coded with hierarchical terminologies and obtained from typical data sources like clinical data repositories or citation databases. The filters remove relatively unimportant data and render them more suitable for graphical display. This paper describes approaches to setting and selection thresholds needed for the method, illustrated with real data in four case studies. We believe the method can aid knowledge validation by highlighting expected results and can

also aid in establishing new hypotheses and knowledge discovery by highlighting interesting unexpected areas or differences among different graphs. The method is applicable to large data sets and many broader application areas apart from biomedical fields.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by intramural research funds from the National Library of Medicine and the NIH Clinical Center. Authors would like to thank Drs. Olivier Bodenreider, Bastien Rance, Kin Wah Fung and Swapna Abhyankar for helpful and constructive discussions; thank Dr. Fiona Callaghan for statistical consultations and suggestions; thank Mr. Dwayne McCully, Chuck Gibbs and Kevin Sung for technical support.

References

1. Brownstein JS, Murphy SN, Goldfine AB, Grant RW, Sordo M, Gainer V, et al. Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. Diabetes Care. 2010; 33(3):526–531. [PubMed: 20009093]
2. Plaisant, C.; Grosjean, J.; Bederson, B. SpaceTree: Supporting Exploration in Large Node Link Tree, Design Evolution and Empirical Evaluation; Procedings of IEEE Symposium on Information Visualization; Boston, MA. 2002; p. 57-64.
3. Keller D, Schaller W, Wong J, de Groen P. World-Wide Web-based graphical user interfaces for laboratory data. Methods Inf Med. 2002; 41(5):411–413. [PubMed: 12501813]
4. Kopanitsa G, Hildebrand C, Stausberg J, Englmeier K. Visualization of medical data based on EHR standards. Methods Inf Med. 2013; 52(1):43–50. [PubMed: 23223709]
5. Stiglic G, Kocbek S, Pernek I, Kokol P. Comprehensive Decision Tree Models in Bioinformatics. PLoS ONE. 2012; 7(3):e33812. [PubMed: 22479449]
6. McGuffin MJ, Robert J-M. Quantifying the Space-Efficiency of 2D Graphical Representations of Trees. Information Visualization. 2010; 9(2):115–140.
7. Jing, X.; Cimino, J. Graphical methods for reducing, visualizing and analyzing large data sets using hierarchical terminologies. AMIA; Washington DC: 2011. p. 635-643.
8. Cimino J, Ayres E. The clinical research data repository of the US National Institutes of Health. Stud Health Technol Inform. 2010; 160(Pt 2):1299–1303. [PubMed: 20841894]
9. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010; 17(2):124–130. [PubMed: 20190053]
10. CDC. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Updated June 18, 2013; cited Sept. 6, 2013.Available from: http://www.cdc.gov/nchs/icd/icd9cm.htm
11. Vreeman D, McDonald C, Huff S. LOINC®–A Universal Catalog of Individual Clinical Observations and Uniform Representation of Enumerated Collections. Int J Funct Inform Personal Med. 2010; 3(4):273–291. [PubMed: 22899966]
12. International Health Terminology Standards Development Organization. SNOMED CT. Cited Sept. 6, 2013Available from: http://www.ihtsdo.org/snomed-ct/
13. Liu S, Wei M, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Professional. 2005; 7(5):17–23.
14. Graphviz–Graph Visualization Software. Cited Dec. 12, 2010Available from: http://www.graphviz.org/
15. Ellson, J.; Gansner, E.; Koutsofios, E.; North, S.; Woodhull, G. Graphviz and Dynagraph–Static and Dynamic Graph Drawing Tools. In: Junger, M.; Mutzel, P., editors. Graph Drawing Software. Springer-Verlag; 2004. p. 127-148.

16. Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. Methods Inf Med. 1993; 32(4):281–291. [PubMed: 8412823]
17. UMLS Reference Manual (Internet). NLM; Bethesda, MD: 2009. Available from: http://www.ncbi.nlm.nih.gov/books/NBK9676/
18. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMICII): A public-access intensive care unit database. Critical Care Medicine. 2011; 39:952–960. [PubMed: 21283005]
19. NIH N. PubMed Health–Pioglitazone. NLM, NIH; Cited Sept. 6, 2013.Available from: http://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0011742/
20. NCBI-NLM-NIH. PubMed Health–Rosiglitazone. NLM NIH; 2011. Cited Sept. 6, 2013.Available from: http://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0012041/
21. FDA. FDA Drug Safety Communication: Update to ongoing safety review of Actos (pioglitazone) and increased risk of bladder cancer. 2011. Cited Feb 25, 2012Available from: http://www.fda.gov/Drugs/DrugSafety/ucm259150.htm
22. The R project for statistical computing. Cited Dec. 12, 2012Available from: http://www.rproject.org/
23. Ligges U, Mächler M. Scatterplot3d–an R Package for Visualizing Multivariate Data. Journal of Statistical Software. 2003; 8(11):1–20.
24. Human-Computer Interaction Lab at University of Maryland. Treemap. Cited Mar. 14, 2012Available from: http://www.cs.umd.edu/hcil/treemap/
25. Munzner T, Guimbretiere F, Tasiran S, Zhang L, Zhou Y. TreeJuxtaposer: Scalable Tree Comparison using Focus+Context with Guaranteed Visibility. ACM Transactions on Graphics. 2003; 22(3):453–462.

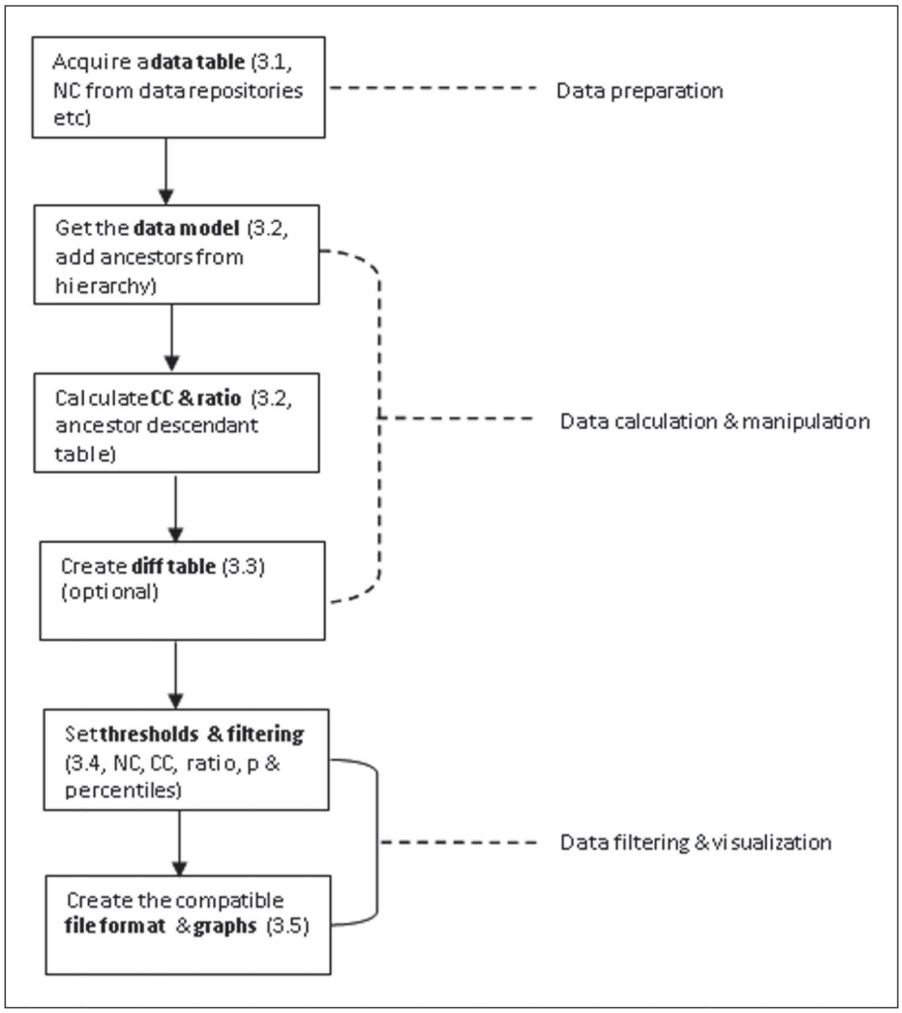

Figure 1.

Main calculation steps in the graphic method for data manipulation including node counts (NC), class counts (CC), ratios, p values and percentiles. (The bold is used to indicate the main purpose of the step.)

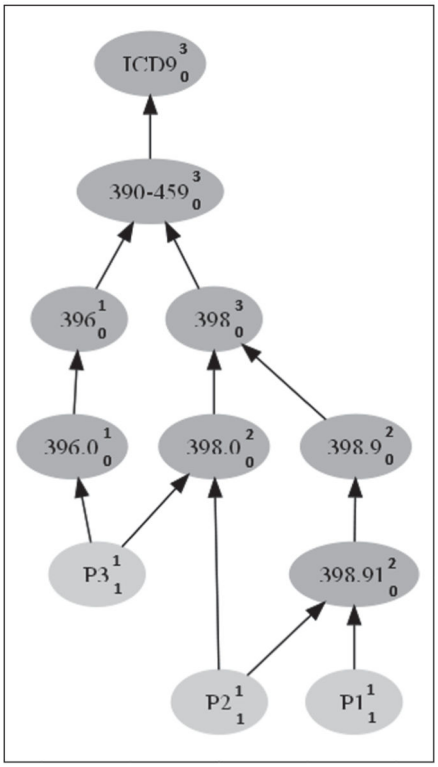

Figure 2.
Sample patients (i.e., P1, P2, and P3 in light grey), each of whom has one or multiple ICD-9-CM codes. The subscripts are node counts and superscripts are class counts.

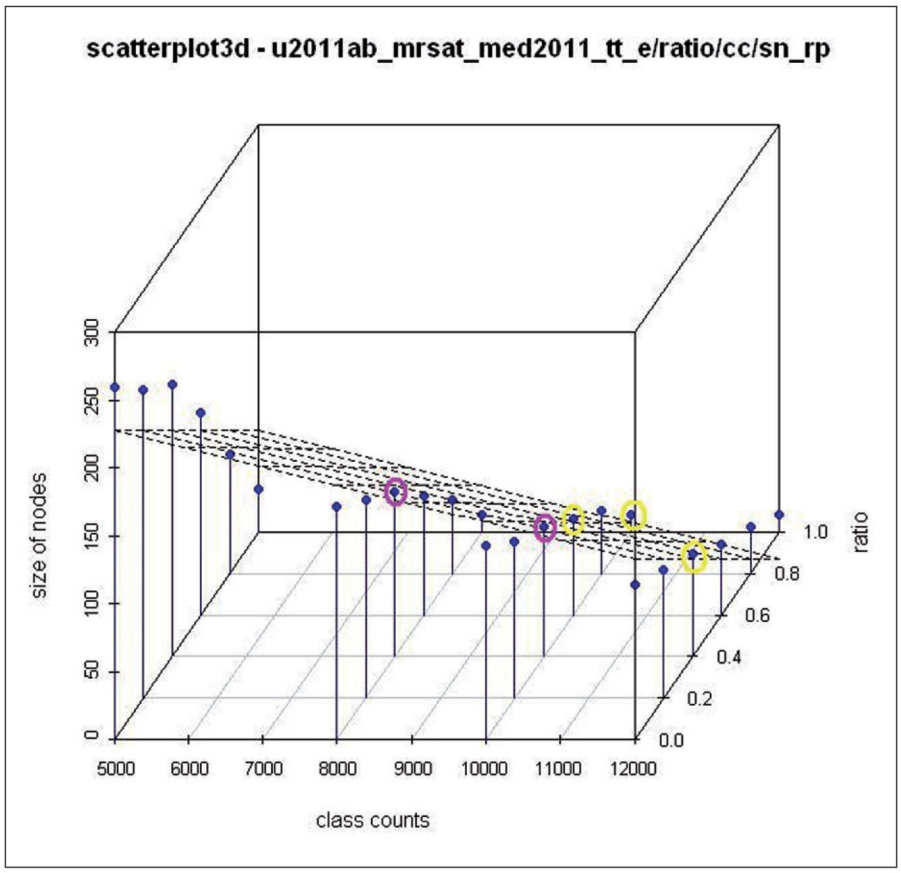

Figure 3.

3D plot with the regression plane for MeSH terms in 2011. (Data from Case Study 3, five circles are the top five observed points closest to the regression plane; two of the observed points with purple circles have final size of 90–120)

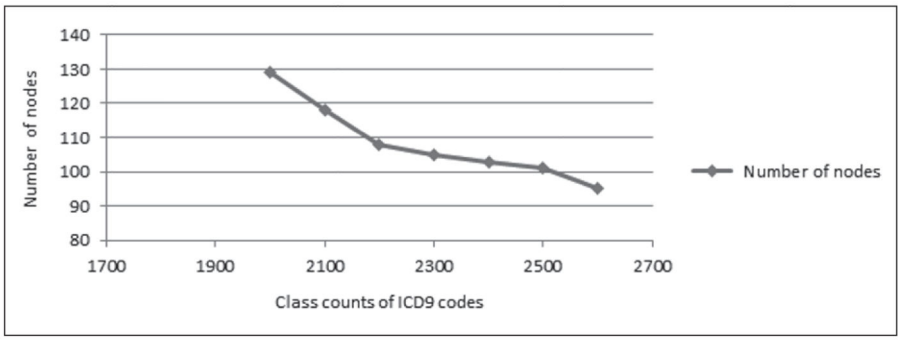

Figure 4.

Number of nodes in graph (Y-axis) for Case Study 1 filtered with different class count thresholds (X-axis)

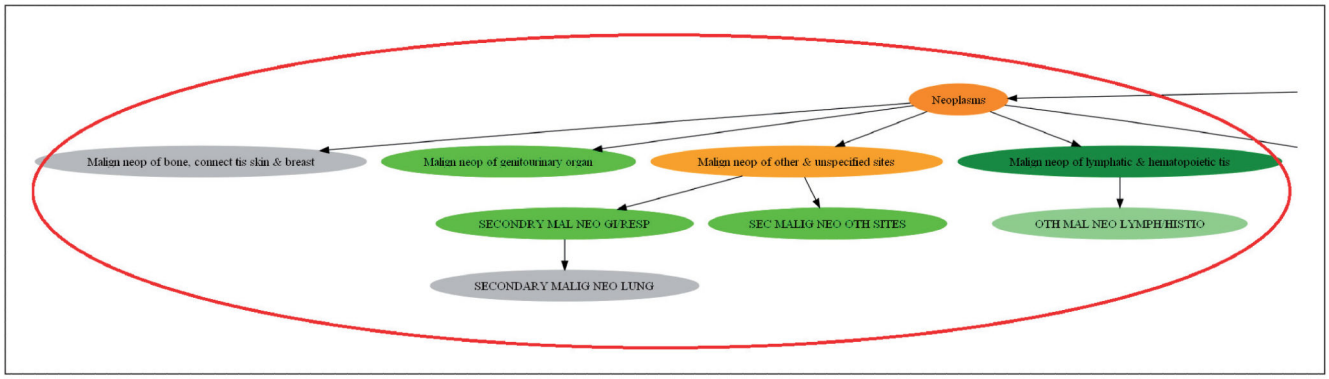

Figure 5.

The circled area is particular for Case Study 1 data set (the most common ICD9-CM codes in BTRIS), showing these are the conditions associated with readmissions in 2011. Filtered graph from Case Study 1 with class count threshold of 2300 shown in Figure 5 in Appendix. Color annotations correspond to different class count ranges: dark orange, 20000–80000; orange, 14000–20000; green, 6000–10000; lime, 4000–6000; light green, 3000–4000; and grey, 2000–3000.

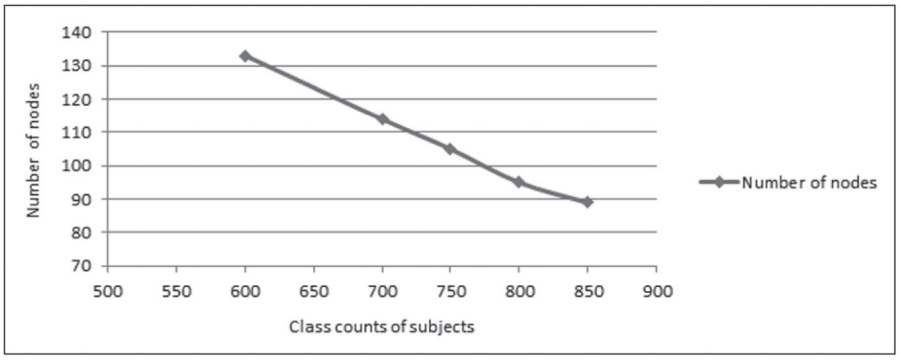

Figure 6.

Number of nodes in graph (Y-axis) for Case Study 2 filtered with different class count thresholds (subject counts, X-axis)

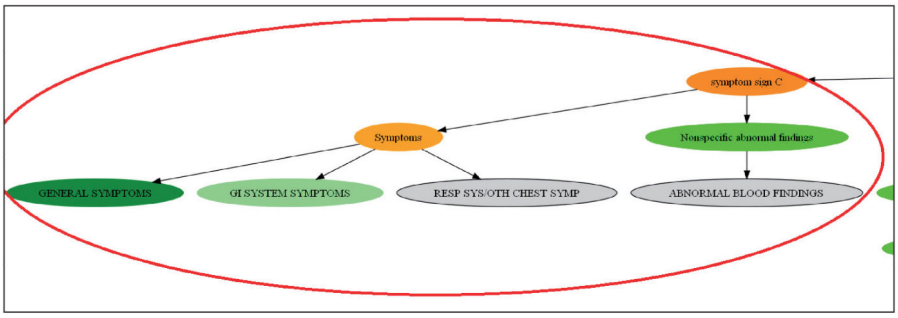

Figure 7.

The circled area is one of the particular areas for Case Study 2 data set (Case Study 2, the most patients coded with ICD9-CM codes in BTRIS 2011 with a class count threshold of 750): symptoms involving respiratory system and other chest symptoms. The whole filtered graph of Case Study 2 are shown in Figure 7 in Appendix. Color annotations correspond to different subject ranges: dark orange, 8000–10000; orange, 6000–8000; lime, 1200–2000; light green, 1000–1200; grey, 800–1000.

Table 1

Matrix for class counts (CC) threshold, ratio values threshold, number of nodes in filtered graphs and percentiles of class counts threshold (Case Study 3)

MeSH major terms indexed in 2011				
CC thresholds	**Percentile of CC**	**Ratio**	**Number of nodes**	**Residuals (lm)**
1	0.073	0	23086	
5000	0.98866	0	259	30.9
5000	0.98866	0.2	227	29.6
5000	0.98866	0.4	200	33.3
5000	0.98866	0.6	149	13.016211
5000	0.98866	0.8	88	−17.290932
5000	0.98866	1	31	−43.598075
8000	0.99247	0	171	−16.43152
8000	0.99247	0.2	146	−10.738663
*8000	0.99247	0.4	121	−5.045806
8000	0.99247	0.6	88	−7.352948
8000	0.99247	0.8	54	−10.660091
8000	0.99247	1	13	−20.967234
10000	0.99383	0	143	−17.344292
10000	0.99383	0.2	115	−14.651435
*10000	0.99383	0.4	95	−3.958578
*10000	0.99383	0.6	71	2.734279
10000	0.99383	0.8	46	8.427136
*10000	0.99383	1	13	6.119993
12000	0.99519	0	114	−19.257065
12000	0.99519	0.2	94	−8.564208
*12000	0.99519	0.4	76	4.128649
12000	0.99519	0.6	52	10.821506
12000	0.99519	0.8	34	23.514364
12000	0.99519	1	13	33.207221

* Grey rows are selected by the number of nodes in filtered graphs and rows are selected by the top 5 minimum absolute residuals; residual reflects the difference between an observed value and the fitting value from the model-regression plane; lm: linear model

Table 2

Comparison of two filtered graphs from Case Studies 1 and 2

	Class count (CC) threshold	Percentile of CC	Final size
Case Study 1	2300	0.98308	105
Case Study 2	750	0.98302	105

Table 3

Profiles of the BTRIS & MIMIC-II data sets: comparison of pioglitazone and rosiglitazone in Case Study 4

	MIMIC-II	BTRIS
After-pioglitazone group subjects	186	251
After-rosiglitazone group subjects	234	224
Filtering rate (nodes in filtered graph/nodes in original graph)	3.1% (50/1611)	1.9% (45/2411)

Table 4

Significantly different subject counts coded with the ICD9-CM codes between the after-pioglitazone and after-rosiglitazone groups in the MIMIC-II data set (prop.test; Case Study 4)

	After- pioglitazone (n = 186)	After- rosiglitazone (n = 234)	p
Chronic liver disease & cirrhosis	9 (9/186)	3 (3/234)	0.060
Neurotic disorders, personality disorders and nonpsychotic mental disorders	23 (23/186)	13 (13/234)	0.005
Diseases of the musculoskeletal system and connective tissue	37 (37/186)	23 (23/234)	0.005
Rheumatism, excluding the back	9 (9/186)	2 (2/234)	0.026
Diabetes with ketoacidosis	20 (20/186)	12 (12/234)	0.048
Other history of health hazards	20 (20/186)	12 (12/234)	0.048

The data in grey shading cells are the ICD-9CM codes that have more subjects in the after-pioglitazone group than in the after-rosiglitazone in both the MIMIC-II and the BTRIS data sets; this table list all the ICD9-CM codes that have more subjects in the after-pioglitazone group data set even though the after-pioglitazone group has fewer subjects in total: 186 vs 234 in the MIMIC-II.

Table 5

Significantly different subject counts coded with the ICD9-CM codes between the after-pioglitazone and after-rosiglitazone groups in the BTRIS data set (prop.test; Case Study 4)

	After- pioglita-zone (n = 251)	After- rosiglita-zone (n = 224)	p
Chronic liver disease and cirrhosis	55 (55/251)	12 (12/224)	0.0000
Neurotic disorders, personality disorders and nonpsychotic mental disorders	65 (65/251)	31 (31/224)	0.0016
Closed biopsy of liver	63 (63/251)	6 (6/224)	0.0000
Other chronic nonalcoholic liver diseases	48 (48/251)	3 (3/224)	0.0000
Esophageal reflux	20 (20/251)	1 (1/224)	0.0002
Viral hepatitis	25 (25/251)	4 (4/224)	0.0004
Other diseases of digestive system	76 (76/251)	26 (26/224)	0.0000
Digestive disorders	113 (113/251)	56 (56/224)	0.0000
Disease of lipoid metabolism	107 (107/251)	61 (61/224)	0.0006
Other metabolic and immunity disorders	149 (149/251)	100 (100/224)	0.0018
Symptom sign and ill-defined conditions	100 (100/251)	61 (61/224)	0.0051

The data in grey shading cells are the consistent results in both the MIMIC-II and the BTRIS data sets; this table lists all the ICD9-CM codes that have more subjects in the after-pioglitazone group than in the after-rosiglitazone group in the BTRIS data set.

Table 6

The comparison of the subject counts coded with the heart-related ICD9-CM codes in the after-pioglitazone and after-rosiglitazone groups in the MIMIC-II data set (prop.test; Case Study 4)

	After- pioglita-zone (n = 186)	After- rosiglita-zone (n = 234)	p
Unspecified chronic pulmonary heart diseases	3 (3/186)	13 (13/234)	0.066
Other forms of heart diseases	103 (103/186)	1 51 (151/234)	0.071
Other pericardial diseases	0 (0/186)	6 (6/234)	0.074
Heart failure	60 (60/186)	96 (96/234)	0.081