Computers in Biology and Medicine 106 (2019) 84–90

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# A rule-based semantic approach for data integration, standardization and dimensionality reduction utilizing the UMLS: Application to predicting bariatric surgery outcomes

Minoo Modaresnezhad[a,*], Ali Vahdati[b], Hamid Nemati[c], Ali Ardestani[d], Fereidoon Sadri[e]

[a] Dept. of Business Analytics, Information Systems & Supply Chain Management, Cameron School of Business, University of North Carolina Wilmington, 601 S College Rd, Wilmington, NC, 28403-5611, USA
[b] Department of Engineering, College of Engineering and Technology, East Carolina University, Greenville, NC, 27843, USA
[c] Department of Information Systems and Operations Management, The University of North Carolina – Greensboro, PO Box 26170, Greensboro, NC, 27402-6170, USA
[d] Division of General/GI, Department of Surgery, Brigham & Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA, 02115, USA
[e] Department of Computer Science, The University of North Carolina – Greensboro, PO Box 26170, Greensboro, NC, 27402-6170, USA

ARTICLE INFO

*Keywords:*
Medical informatics
Medical information systems
Data integration
Semantic integration
UMLS
Dimensionality reduction
Machine learning
Data standardization

ABSTRACT

Utilization of existing clinical data for improving patient outcomes poses a number of challenging and complex problems involving lack of data integration, the absence of standardization across inhomogeneous data sources and computationally-demanding and time-consuming exploration of very large datasets. In this paper, we will present a robust semantic data integration, standardization and dimensionality reduction method to tackle and solve these problems. Our approach enables the integration of clinical data from diverse sources by resolving canonical inconsistencies and semantic heterogeneity as required by the National Library of Medicine's Unified Medical Language System (UMLS) to produce standardized medical data. Through a combined application of rule-based semantic networks and machine learning, our approach enables a large reduction in dimensionality of the data and thus allows for fast and efficient application of data mining techniques to large clinical datasets. An example application of the techniques developed in our study is presented for the prediction of bariatric surgery outcomes.

## 1. Introduction

With the rising interest in utilizing patients' medical history for efficient and effective prediction of clinical outcomes, Clinical Decision Support Systems (CDSS) have become an area of research that shows tremendous potential for enhancing medical care while reducing the associated costs. CDSS aim at helping clinicians utilize the existing medical information and history of the patients for improved clinical decision making and thus improved outcomes [1–3]. CDSS can help a clinician in several ways including, 1) using patients' medical history in helping to decide the most appropriate treatment for the patients, 2) monitoring and recording the patients' medical information prior and after the start of treatment and alerting the clinician in case of any changes, and 3) using the interrelationships or findings learned from the medical data of other patients relating to a particular condition, to help in early diagnosis and treatment for future patients [4].

For successful application of a CDSS, integration, and standardization of medical data is a necessity considering the dispersed, heterogeneous nature of the existing medical data [5]. Integration of medical data is challenging because of the variations in data entry, as well as imperfections due to human error or unavailability of data [6]. This is one reason cognitive issues are of utmost important in modern medical informatics [7]. Data cleansing needs to be handled very carefully in the case of medical data because the output of this operation will later be utilized during clinical decision making. Data selection and integration can then be done on the cleaned data by choosing the relevant data, combining, and finally presenting and visualizing the data in a format suitable for computer-assisted decision making [8]. Standardization of the medical data is as important as in any other field. This will help the clinicians in their decision-making process as it generates standardized, interoperable and universally accepted medical terms. A standardization tool has to be one that exclusively coincides with standard medical terms [9–11]. The pressing need for data standardization across diverse Electronic Health Record (EHR) systems was

* Corresponding author.
*E-mail addresses:* modaresm@uncw.edu (M. Modaresnezhad), vahdatia18@ecu.edu (A. Vahdati), nemati@uncg.edu (H. Nemati), aardestani@bics.bwh.harvard.edu (A. Ardestani), sadri@uncg.edu (F. Sadri).

https://doi.org/10.1016/j.compbiomed.2019.01.019
Received 26 August 2018; Received in revised form 21 January 2019; Accepted 21 January 2019

highlighted by the President's Council of Advisors on Science and Technology (PCAST) in a 2010 report titled: “Realizing the full potential of health information technology to improve healthcare for americans: the path forward” [12]. The report argued for improved medical data standardization through a “universal exchange language whose semantics is intrinsically extensible” and for “managing and storing data for advanced data-mining techniques through breaking it down into the smallest individual pieces” [12]. One example of such an interoperable and universal exchange language for medical data is the Q-UEL [13–15]. Furthermore, with the emergence of big data in medicine, there is an ever-increasing need for methods that enable dimensionality reduction of large data sets. Efficient dimensionality reduction of big medical data is of utmost importance particularly when computationally-demanding machine learning techniques are employed to analyze the data with the goal of improving clinical decision making.

In this study, a robust semantic data integration, standardization and dimensionality reduction approach for clinical data is presented. Our approach enables the integration of clinical data from disparate sources by resolving canonical inconsistencies and semantic heterogeneity required by the National Library of Medicine's Unified Medical Language System (UMLS) to produce standardized medical data. This resulting application, henceforth referred to as RxSem, builds upon our previous preliminary research [16] and enables domain experts (i.e., healthcare and medical professionals) to semantically describe their data needs and integration requirements. Furthermore, the dimensionality reduction techniques developed in our study enable fast and efficient application of machine learning techniques to large medical datasets.

As an illustrative example, our developed approach is utilized to predict bariatric surgical outcomes employing traditional data mining techniques. The case of bariatric surgery was chosen for this study because of its relevance to and prevalence of obesity, a worldwide epidemic and major health problem facing societies. Obesity is regarded as a complex multi-factorial chronic disease that develops from an interaction of genotype and the environment and has serious health consequences. The annual health care cost attributable to obesity was about $147 billion in 2008, and it is estimated that the cost will expand to $344 billion by 2018 just in the United States alone [17]. Managing obesity-related medical data electronically has been achieved to a great extent, but the medical field is still in need of a powerful predictive tool which will assist the clinicians in making decisions regarding the optimal patient selection before the operation, during the course of treatment and during follow-up which ultimately leads to better outcomes for the patients. The patterns in which various factors impact the successful outcome of bariatric surgery are key findings of interest for surgeons and medical professionals and are critical decision-making elements for improving the nature of prognosis and treatment. Integration of the datasets based on rules provided by experts, standardization of the attributes in the integrated dataset and efficient data mining of the full and reduced datasets will be presented in this research.

## 2. Methods

### 2.1. Overview

Fig. 1 shows the architecture of the system developed in this study for integration, standardization, dimensionality reduction and data mining of bariatric surgery outcomes. The integration and standardization approach utilized here are based on the methods developed in our previous study [16]. Our improved approach in the current study utilizes machine learning for selecting the best semantic subgraph and reducing the dimensionality of the data. In this section, after a brief overview of the methods, the dataset for bariatric surgery is introduced. Next, for the sake of completeness, a brief introduction to the data integration and standardization steps is given in section 2.3. Readers interested in further details of the data integration and standardization methods may refer to our previous conference publication [16]. Subsections 2.4 to 2.5 will introduce the methods developed for dimensionality reduction through semantic networks and machine learning.

The system can be divided into three layers: data integration, semantic, and data mining layer. Our approach starts with the semantic integration of the datasets. A system is developed that relates the datasets based on defined rules and combines them in order to form a single integrated file. The data integration layer includes the multiple source schema, the semantic rules engine and target schema [16]. The semantic layer emphasizes the standardization of semantically integrated data using UMLS. The terms in the integrated file are compared to the UMLS Metathesaurus to find standardized matching terms. The final output of this part of the system is a semantic network with a higher level categorical semantic types and relationships. Next, a dimensionality reduction technique is applied to find the significant attributes of the dataset through semantic subgraphs. Finally, the data mining part of the system employs the reduced data set to identify the factors that impact the outcome of bariatric surgery.

The data mining layer generates multiple models and compares them. This step produces a set of significant attributes. Variable selection, sometimes referred to as feature selection, is one of the most fundamental areas of concern in predictive analytics model development using clinical data. The goal of variable selection is to choose a subset of variables from the pool of all available variables to be included in the predictive model. An appropriate variable selection method reduces in the dimensionality of the problem by eliminating variables that do not contribute to the performance of a given predictive model. Choosing an appropriate variable selection method is critical to the overall success of predictive modeling process since it enables reduction in the cost associated with gathering and storage of data used thus improving computational speedup while guaranteeing that the predictive model's performance is not degraded. Additionally, a model developed using a reduced set of variable is more parsimonious and would be more understandable by the end users. There are a large of number of variable selection methods available (see Refs. [18,19] or [20] for a comprehensive overview). These methods can be classified along two broad categories of statistically based or semantically based approaches. Statistically based methods reduce the number of variables used using statically based analytical methods (e.g., Principle Component Analysis, Partial Least Squares and others), while semantic based methods use semantic relationship among variables to select the final subset of the variables. Semantic based methods allow prior knowledge to be incorporated in selecting appropriate subset of variables to be used in developing predictive models. The variable selection method presented in the current study is an attempt to develop a hybrid approach that integrates the best ideas of each of the two approaches. To achieve this, a two-step process is used. First, as seen in Fig. 1, our approach allows the data mining engine to train and select the best predictive model. Once the best predictive model is selected in this step, the most significant variables from the best predictive model are extracted. It is worth noting that all of the machine learning algorithms used in the step provide us with the capability to extract the most significant variables used. In the second step our method determines the semantic relationship among the extracted set of variables. This is achieved by determining the semantic relationship of each pair of variables using the UMLS semantic network modeling capability. If there are no direct semantic relationships between any two selected variables, the full model semantic network is used to identify other variables that constitute a path between the two variables. If there is an “is-a” relationship between any two variables in the full model semantic network, the variable that is on the “from” side of the “is-a” relationship is eliminated from the set constituting the reduced set of variables. For all other relationships, all the variables in the path that connects any two variables are added to the subset. Although the reduced variable

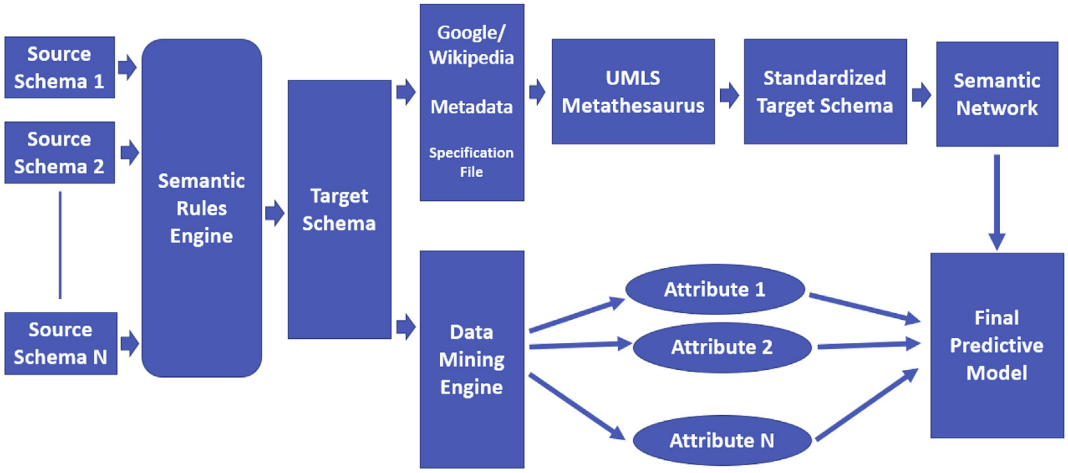


**Fig. 1.** Architecture of the RxSem system for integration, standardization, dimensionality reduction and data mining of bariatric surgery.

subset constructed in this way is not unique, since multiple paths are possible between any two variables, it consists of variables that are not only statistically significant but also are semantically related. However, we note here that using the semantic network created using the full set of variables, it possible to construct many subset semantic networks all containing the same subset of variables but each travers different paths of variable connectivity. Please see Fig. 2a and b for comparison of different sub networks containing a unique subset of variables.

### 2.2. The case of bariatric surgery

The Bariatric Surgery Information System (BSIS) containing medical data of a large number of patients who underwent the surgery was used in this study. The dataset used in this project is from the National Surgical Quality Improvement Program (NSQIP) of the American College of Surgeons which captures surgical outcomes nationwide [21]. This initiative includes over 200 institutions and, in our sample, included valuable information of more than 100,000 patients. This rich set of data chronicles patients' medical conditions from the first visit long before the surgery, all the finding during the operation and hospitalization and also a huge repertoire of follow-up visits spanning up to 3 years after the initial visit. There is also comprehensive documentation of technical failures and complications. The original data source contained five data sets with information relating to patients that underwent bariatric surgery obtained from reliable experts. The description of each data set is as follows:

- Demog – Demographic information of the patients.
- Preop – Medical information of the patient during visits before the surgery.
- Intraop – Information about the surgery that was performed on the patient.
- AE –Data relating to certain side effects and complications or adverse effects that were observed on the patient after the surgery.
- Postop - Medical information of the patient during multiple visits after the surgery.

### 2.3. Semantic integration and standardization

Data selection and integration directives (medical rules) were provided by medical experts. The integrated file contained 120,000 patient records with over 250 attributes. The algorithm for rule-based integration was implemented using Java as well as SQL. The latter was more efficient in terms of performance. To standardize the integrated file, we used the web, expert advice and the UMLS to find the best matched standard terms for the metadata in the integrated file. A specification file, provided by experts, along with the metadata of the integrated file was compared to the UMLS to find the best matched standard terms. The web was used to find terms that UMLS failed to match.

The UMLS files and data sets were stored in a MySQL database using the UMLS downloads available on the UMLS website. The search algorithm was implemented using Java which interacted with all databases to get the best match for the medical terms. Several levels of search were implemented until best possible matches were found. As the first step, the metadata of the integrated file was searched for in the UMLS Metathesaurus. The second step used keywords from the description of the term from the specification file to search the UMLS. In the third level of search, all possible permutations of the term being searched was fed into the UMLS search query. The fourth level of search, involved searching in the UMLS for all possible combinations of the words. Finally, each word in the metadata being searched for were fed into the UMLS search query. Terms that passed any one of the levels of search above did not go through the further levels of search. The application was semi-automatic which made the output obtained more relevant than that obtained using complete automation. The importance of having a human-in-the-loop in machine learning has been previously highlighted by other researchers [22].

The matching standard terms obtained from the UMLS Metathesaurus after following the above-mentioned search strategies were presented to the users. The user also had a choice of selecting an appropriate match for the term from a list of suggestions, if the default selected match was not satisfactory. After the best matched standard terms were assigned to each term in the metadata of the integrated file, the algorithm replaced the metadata of the integrated file with the corresponding standard term. The output of this operation provided an integrated and standardized file containing bariatric surgery data.

Furthermore, the relationship among the sematic types returned by UMLS for the matched standard terms was obtained. A matrix was generated with the semantic types corresponding to our dataset and the relationship among them. This semantic network is later used during data mining and discovery of significant data attributes.

### 2.4. Dimensionality reduction using semantic networks

We combined the semantic representation with the results of data mining engine. The data mining engine initially considered all the attributes of the integrated dataset to find the best fit model. The significant attributes of the best fit model were then extracted to perform a data reduction process before generating the final predictive model. Based on these significant attributes, a number of subgraphs of the semantic networks generated by the UMLS were isolated. The best of

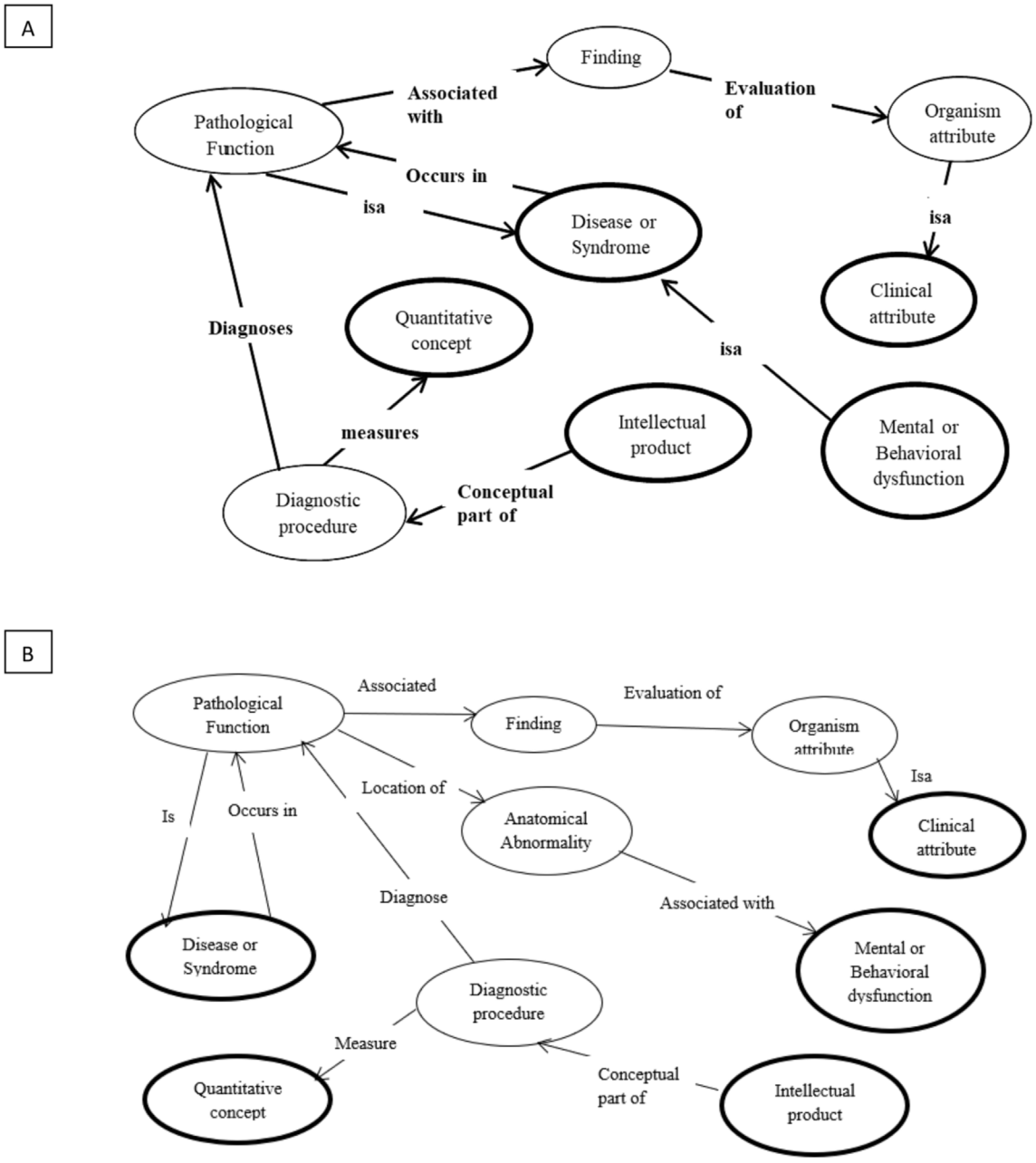


**Fig. 2.** a and b. Two sample subgraphs. The bolded nodes of the graph denote UMLS semantic types corresponding to the significant attributes generated by the data mining engine. The criterion used to find the best subgraph is the one that has the least number of nodes that does not correspond to any of the significant attributes. The number of non-matching nodes is four and in the top and five in the bottom graph respectively. So, the top graph is considered better than the bottom one.

these subgraphs was selected with the criteria being the smallest possible subgraph connecting all the concepts relating to the significant attributes. The dataset was reduced to the significant attributes from the data mining engine and the best possible subgraph. This reduced dataset was sent through the data mining engine to generate the best final predictive model. Sample subgraphs are shown in Fig. 2. The bolded nodes of the graph denote UMLS semantic types corresponding to the significant attributes generated by the data mining engine. The criterion used to find the best subgraph is the one that has the least number of nodes that does not correspond to any of the significant attributes.

### 2.5. The data mining engine

SAS Enterprise Miner@ was utilized for mining of the data. Before performing the data mining operation, the first challenge was to find the best target variable for the analysis. Since the dataset is on bariatric surgery, the target variable can be related to the impact of the surgery on the body mass index (BMI) of the patient. The integrated dataset has a number of attributes for the weight of the patient, each of which corresponded to each visit of the patient before and after the surgery. BMI of the patient was calculated by dividing the weight by the height of patient squared. The change in the BMI of the patient after the surgery was chosen as the target variable (Table 1).

The Data Partitioning node in SAS Enterprise Miner was used to partition the data into training and validation data, with training data used for preliminary model fitting and validation data used for monitoring, tuning and assessing the model. Decision trees, regression and neural networks were used as the techniques to generate the models for the dataset.

**Table 1**
Criteria for successful versus unsuccessful bariatric surgery operation.

| | |
|---|---|
| Change in BMI < 5 | Operation was not Successful |
| Change in BMI > = 5 | Operation considered Successful |

Three decision tree models were run on the dataset with variation in properties. The DTM1 was run with the default properties in Enterprise Miner, DTM2 was generated using an interactive process, where an attribute of interest was chosen to start the splitting of the tree and further branches of the tree were trained by the software. In DTM3 the depth and splitting rule were specified.

The dataset in this study was also modeled using three regression models. In the first attempt using the forward selection model, a baseline model was generated that represented an overall average of the dataset. In the next step, the best of the models with one input was chosen, followed by a model with two inputs. This sequence was continued until no significant improvement could be made. The next regression model used was backward selection model, which started off with a saturated model and continued to eliminate variables unless a change happened in the results. Stepwise selection model was used in the third regression model.

Finally, three neural network models were run for this dataset with variations in some properties and network architecture. NNM1, used a multilayer perception network architecture. NNM2, used a generalized linear model architecture and in NNM3 the network architecture was not changed from default but the number of hidden units was modified to a higher value.

## 3. Results

During the data mining phase, the integrated dataset was mined using a decision tree, regression, and neural network models. Each type of model was run based on three separate configurations. In all, nine models were compared and various statistical aspects including misclassification rate, accuracy, sensitivity, specificity, and precision were considered. Definitions of accuracy, sensitivity, specificity, and precision are given below:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False negative}}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

**Table 3**
Comparison between the original dataset and the reduced dataset in terms of file size, number of attributes and performance.

| | Size | No. of attributes | Performance |
|---|---|---|---|
| Entire Dataset | 136 MB | 250 | 15–20 min |
| Reduced Dataset | 5 MB | 20 | Less than a minute |

Table 2 shows the output of accuracy, specificity, sensitivity, precision and misclassification rate of the models generated from the data when all the attributes (full model) are included as opposed to when only significant attributes (reduced model) are considered. Both approaches using the full data and the reduced data generated fairly similar results. Table 3 highlights the volume of the data and performance (time taken for the execution of the data mining engine) before and after dimensionality reduction. As can be seen in the results in Table 3, our rule-based semantic approach for reducing data dimensionality was highly effective in reducing the volume of the data and the time needed to run the analysis. The results in Table 2 show that the reduced model performs as well as the full model. The outcome of the data mining using the nine different algorithms was not significantly changed when the reduced data was utilized instead of the full data. In particular, the misclassification rates for the reduced set are very close to the full set (Table 2).

The similarity in the data mining results after data reduction is one of the most important findings of the research. The data reduction was done by reducing the size of the dataset by applying the results of the UMLS semantic network on the output of the data mining engine while considering the whole dataset. The understandability of the data mining results increased considerably because of the smaller number of attributes that were considered during data mining. As is clearly evident from the results in Tables 2 and 3, the predictive power of the full and reduced models are similar while runtime and data volume decreased significantly in the case of the reduced model. Determining the attributes to be removed from the dataset during data reduction would be the most computationally expensive part involved in this study. This computational cost was nearly negligible in the procedure developed here due to the involvement of the semantic network generated by the UMLS.

## 4. Discussion

Utilization of clinical data for medical informatics with the goal of improving patient outcomes poses a number of challenging and complex problems. Additionally, the use of clinical data for retrospective analysis in support of medical informatics is problematic when considering that most clinical data, at the time of collection, were intended for patient care and may not have been explicitly collected for developing medical informatics. These challenges stem from the fact that clinical data come from disparate sources and therefore may lack

**Table 2**
Comparison of machine learning approaches to find the best fit model, considering attributes obtained from the Data Mining engine and the Semantic Network.

| | Accuracy | | Sensitivity | | Specificity | | Precision | | Misclassification Rate | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Full Model | Reduced Model | Full Model | Reduced Model | Full Model | Reduced Model | Full Model | Reduced Model | Full Model | Reduced Model |
| DTM1 | 74.21% | 74.29% | 15.71% | 11.53% | 95.24% | 96.86% | 54.26% | 56.86% | 25.80% | 25.71% |
| DTM2 | 74.21% | 74.29% | 15.71% | 11.53% | 95.24% | 96.86% | 54.26% | 56.86% | 25.80% | 25.71% |
| DTM3 | 74.33% | 74.21% | 16.08% | 16.31% | 95.28% | 95.03% | 55.05% | 54.11% | 25.67% | 25.80% |
| RM1 | 74.92% | 74.45% | 19.28% | 15.81% | 94.92% | 95.54% | 57.72% | 56.05% | 25.08% | 25.49% |
| RM2 | 74.91% | 74.52% | 18.73% | 15.93% | 95.11% | 95.58% | 57.91% | 56.45% | 25.10% | 25.55% |
| RM3 | 74.91% | 74.45% | 18.73% | 15.81% | 95.11% | 95.54% | 57.91% | 56.05% | 25.10% | 25.55% |
| NNM1 | 74.86% | 74.55% | 20.03% | 17.83% | 94.58% | 94.94% | 57.04% | 55.89% | 25.14% | 25.45% |
| NNM2 | 74.74% | 74.52% | 17.94% | 15.98% | 95.17% | 95.57% | 57.19% | 56.45% | 25.26% | 25.48% |
| NNM3 | 74.87% | 74.56% | 19.77% | 17.89% | 94.68% | 94.94% | 57.19% | 55.96% | 25.13% | 25.44% |

canonical consistency and semantic homogeneity making their integration particularly challenging. Furthermore, clinical data are usually large, complex, time series, nontraditional, and may require extensive data preparation prior to integration, standardization and mining.

Health care applications need semantic medical data integration as they should accept inputting medical data in the form of structured representation. Thus, the applications must address the heterogeneous nature of medical data [23] in order to allow processing of data such as analysis or mining, manipulations, and translations. In this study, we presented a robust rule-based semantic integration, standardization and dimensionality reduction procedure for medical data. Data integration is guided through a set of rules obtained from medical experts. The UMLS was used to achieve the standardization part of the dataset. A stepwise procedure for the mapping of the metadata to the UMLS was incorporated. Decision trees, regression and neural network models were used to analyze the integrated dataset and to find the significant variables and thus reduce the dataset. This dimensionality reduction of the data utilizing a semantic network allowed us to reduce the data volume and data mining runtime significantly.

It should be noted that unlike other dimensionality reduction methods (i.e. principle component analysis, linear discrete analysis, or canonical correlation analysis), the reduced model presented here is not statistically based, but it is semantically based. This is an important contribution of this study. We argue that for the medical data mining to be useful, the understandability of the model and the features used are more important than just accuracy of the models since these models are inherently used by medical professionals as decision support systems. In a decision support situation, understandability and alignment of the model with practitioners mental model are of paramount importance. Semantic based models can accomplish this task.

There are various challenges involved in the process of semantic integration, standardization and deminasionality reduction. The first and most important of these challenges is understanding the semantics of the elements involved. There are just a few ways this can be done, through gathering information from the creators of the data, from documentation or from the schema or data. The first two sources are often inaccessible because the data sources that need to be integrated might be in use for a long time and the documentation, if any exists, might be incorrect or outdated. The last source, schema or data, are often unreliable for the inferring of semantics. The second challenge involves the incompleteness and inaccuracy of the schema and the data involved. It does not provide sufficient information to determine the exact nature of the relationships. The third challenge is the cost involved in the process of matching the elements of the schema because often the datasets are very large in size. The worst of the problems is the customization required for certain elements during matching because of the subjective nature of it. The problem of matching data tuples also faces similar challenges as in schema matching. There are various schema matching techniques among which rule-based semantic matching is an important one. Some of the benefits of rule-based schema matching are that they are inexpensive, fairly fast and provide a quick and concise method to capture valuable user knowledge about the domain. Our work addresses the heterogeneous nature of target schema and enhances interoperability by mapping terms semantically to the UMLS. Standardization through a process of semantic-mapping of target schema to the standard terms was successfully achieved in this study.

Semantic network and ontologies have been utilized by researchers for analyzing medical data in the past [24–28]. The popularity of semantic network and ontologies in the field of medicine is mainly due to their power in representing data and the relationship among the dataset elements.

In summary, we believe our most important contributions in this study are as follows:

- Information Integration: We proposed a rule-based semantic integration approach where domain experts provide the integration rules in a precise logical language, and developers implement the integration in an appropriate database programming language. We demonstrated this approach using the Bariatric Surgery application.
- Data Standardization: Information from multiple heterogeneous sources should be prepared for mining. Two of the most important tasks in data preparation are data cleaning and standardization. In our study, we utilized the UMLS standard medical vocabulary for data standardization.
- Dimensionality Reduction: The integrated and standardized medical datasets can be (and usually are) very large in size. To make medical data mining feasible, we proposed to identify a small subset of parameters that play a significant role in the prediction of the target variable. In our project, the integrated data had in excess of 250 columns, and 100,000 rows. In other projects, the size and dimensionality of data can be much greater. We proposed the following approach for dimensionality reduction of the data: (1) First, we used a small subset of the data to rapidly determine a subset of significant parameters. (2) Next, we generated the semantic network graph of the significant parameters obtained in Step 1, plus other parameters that are semantically related to these significant parameters. (3) subsequently, we explored the subgraphs of the semantic network obtained in Step 2, and obtain the “best” subgraph. (4) We reduced the dimensionality of integrated data by only considering the parameters in the chosen subgraph, and (5) we performed data mining on the reduced data generating predictive models as good as the models generated using the full data, but at a fraction of the computational cost.

## 5. Limitation and future research

We have shown that our approach allows for variable subset selection using a hybrid method that integrates statistical and semantic approaches. The current paper describes this method and provides a proof of concept using a limited dataset. We have shown that while the reduced model is far more parsimonious than the full model, its predictive power is comparable to the full set model. However, we are cognizant that the method presented in this paper suffers from several limitations that are described here along with proposed future research avenues to overcome them. First, as we have acknowledged earlier that many possible subset semantic networks could be created using the same subset of variables, the current paper does not propose a method for selecting the “best” sub network. This is a computationally difficult problem to solve and requires additional insights. Second, our method is only tested on a limited dataset. We invite other researchers to replicate our experimental results using other well-known medical datasets. Third, we are aware there are a number of well-established and robust approaches (e.g., OMOP [29], PCORnet [30], i2b2 [31] and Sentinel [32] initiatives) that have been developed to facilitate data integration and standardization of clinical data. We do not claim that our proposed approach presented here should be viewed as yet another data integration and standardization approach. We note that these approaches have been developed primarily to overcome data integration and standardization problems, but not variables subset selection problem. Additionally, we note that these methods require the collected data to be curated based on specific standards and templates that facilitate the eventual integration and standardization of data among disparate sources. Out method does not require that. It assumes no such standardization of data at the point of collection. The standardization and eventual integration occur using the three of the most important components of UML: Lexicon, Meta-thesaurus, and the Semantic network engines. Therefore, we argue that our approach should not be viewed as a “competitor” of these methods, rather as a front-end preprocessing approach that could be used to augment and enhance the capabilities of these methods. We posit that a research project that

examines the efficacy of the use of our approach as a front end to these approaches would make a valuable contribution to extant literature.

## 6. Summary and conclusions

In summary, RxSem is a system which goes beyond data integration, standardization and mining of the data related to bariatric surgery by utilizing semantic networks for reducing data dimensionality thus making predictive analytics using large dataset feasible and efficient. Although our study focused on medical data, we believe our novel approach and the techniques developed in this study are general and may apply to a vast array of information systems.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2019.01.019.

## References


[1] A.X. Garg, N.K.J. Adhikari, H. McDonald, M.P. Rosas-Arellano, P.J. Devereaux, J. Beyene, J. Sam, R.B. Haynes, Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review, J. Am. Med. Assoc. 293 (2005) 1223–1238, https://doi.org/10.1001/jama.293.10.1223.

[2] H. Åhlfeldt, B. Johansson, R. Linnarsson, O. Wigertz, Experiences from the use of data-driven decision support in different environments, Comput. Biol. Med. 24 (1994) 397–404, https://doi.org/10.1016/0010-4825(94)90008-6.

[3] S. Hegenbart, A. Uhl, A. Vécsei, Survey on computer aided decision support for diagnosis of celiac disease, Comput. Biol. Med. 65 (2015) 348–358, https://doi.org/10.1016/j.compbiomed.2015.02.007.

[4] D.L. Hunt, R.B. Haynes, S.E. Hanna, K. Smith, Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review, J. Am. Med. Assoc. 280 (1998) 1339–1346, https://doi.org/10.1001/jama.280.15.1339.

[5] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artif. Intell. Med. 34 (2005) 113–127, https://doi.org/10.1016/j.artmed.2004.07.002.

[6] M. Pine, M. Sonneborn, J. Schindler, M. Stanek, J.L. Maeda, C. Hanlon, Harnessing the power of enhanced data for healthcare quality improvement: lessons from a Minnesota hospital association pilot project, J. Healthc. Manag. 57 (2012) 406–418, https://doi.org/10.1097/00115514-201211000-00007.

[7] A. Holzinger, R. Geierhofer, F. Mödritscher, R. Tatzl, Semantic information in medical information systems: utilization of text mining techniques to analyze medical diagnoses, J. Univers. Comput. Sci. 14 (2008) 3781–3795 www.meduni-graz.at/imi , Accessed date: 11 October 2018.

[8] C. Turkay, F. Jeanquartier, A. Holzinger, H. Hauser, On Computationally-Enhanced Visual Analysis of Heterogeneous Data and its Application in Biomedical Informatics, (2014), pp. 117–140, https://doi.org/10.1007/978-3-662-43968-5_7.

[9] I.M. Mullins, M.S. Siadaty, J. Lyman, K. Scully, C.T. Garrett, W. Greg Miller, R. Muller, B. Robson, C. Apte, S. Weiss, I. Rigoutsos, D. Platt, S. Cohen, W.A. Knaus, Data mining and clinical data repositories: insights from a 667,000 patient data set, Comput. Biol. Med. 36 (2006) 1351–1377, https://doi.org/10.1016/j.compbiomed.2005.08.003.

[10] E.E. Westberg, N.H. Mann, D.M. Spengler, Integrating and presenting clinical and treatment outcome data for cost-effective case management, Comput. Biol. Med. 27 (1997) 31–47, https://doi.org/10.1016/S0010-4825(96)00071-6.

[11] M. Komenda, D. Schwarz, J. Švancara, C. Vaitsis, N. Zary, L. Dušek, Practical use of medical terminology in curriculum mapping, Comput. Biol. Med. 63 (2015) 74–82, https://doi.org/10.1016/j.compbiomed.2015.05.006.

[12] President's Council of Advisors on Science and Technology, Report to the President Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: the Path Forward, White House, 2010, p. 108, https://doi.org/10.1021/acs.langmuir.8b01007.

[13] B. Robson, T.P. Caruso, U.G.J. Balis, Suggestions for a Web based universal exchange and inference language for medicine, Comput. Biol. Med. 43 (2013) 2297–2310, https://doi.org/10.1016/J.COMPBIOMED.2013.09.010.

[14] B. Robson, S. Boray, Implementation of a web based universal exchange and inference language for medicine: sparse data, probabilities and inference in data mining of clinical data repositories, Comput. Biol. Med. 66 (2015) 82–102, https://doi.org/10.1016/j.compbiomed.2015.07.015.

[15] B. Robson, Studies in using a universal exchange and inference language for evidence based medicine. Semi-automated learning and reasoning for PICO methodology, systematic review, and environmental epidemiology, Comput. Biol. Med. 79 (2016) 299–323, https://doi.org/10.1016/j.compbiomed.2016.10.009.

[16] A. Ardestani, H. Nemati, O. Eleti, F. Sadri, RxSem: a rule based semantic integration method for medical informatics, Proc. 2012 IEEE 13th Int. Conf. Inf. Reuse Integr. IRI, 2012, pp. 564–571, , https://doi.org/10.1109/IRI.2012.6303059 2012.

[17] S.B. Wyatt, K.P. Winters, P.M. Dubbert, Overweight and obesity: prevalence, consequences, and causes of a growing public health problem, Am. J. Med. Sci. 331 (2006) 166–174, https://doi.org/10.1097/00000441-200604000-00002.

[18] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182 http://www.jmlr.org/papers/v3/guyon03a.html , Accessed date: 26 December 2018.

[19] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1997) 245–271, https://doi.org/10.1016/S0004-3702(97)00063-5.

[20] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324, https://doi.org/10.1016/S0004-3702(97)00043-X.

[21] P.R. Fuchshuber, W. Greif, C.R. Tidwell, M.S. Klemm, C. Frydel, A. Wali, E. Rosas, M.P. Clopp, The power of the National Surgical Quality Improvement Program–achieving a zero pneumonia rate in general surgery patients, Perm. J. 16 (2012) 39–45 http://www.ncbi.nlm.nih.gov/pubmed/22529758 , Accessed date: 25 August 2018.

[22] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf. 3 (2016) 119–131, https://doi.org/10.1007/s40708-016-0042-6.

[23] N. Lavrač, Selected techniques for data mining in medicine, Artif. Intell. Med. 16 (1999) 3–23, https://doi.org/10.1016/S0933-3657(98)00062-1.

[24] M.-F. Sy, S. Ranwez, J. Montmain, A. Regnault, M. Crampes, V. Ranwez, User centered and ontology based information retrieval system for life sciences, BMC Bioinf. 13 (2011) S4, https://doi.org/10.1186/1471-2105-13-S1-S4.

[25] A. Rodríguez, E. Jiménez, J. Fernández, M. Eccius, J.M. Gómez, G. Alor-Hernandez, R. Posada-Gomez, C. Laufer, SemMed: applying semantic web to medical recommendation systems, Proc. 1st Int. Conf. Intensive Appl. Serv. INTENSIVE 2009, IEEE, 2009, pp. 47–52, , https://doi.org/10.1109/INTENSIVE.2009.12.

[26] I. Sim, S. Carini, S.W. Tu, L.T. Detwiler, J. Brinkley, S.A. Mollah, K. Burke, H.P. Lehmann, S. Chakraborty, K.M. Wittkowski, B.H. Pollock, T.M. Johnson, V. Huser, Ontology-based federated data access to human studies information, AMIA Annu. Symp. Proc. (2012) 856–865 2012 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3540523&tool=pmcentrez&rendertype=abstract , Accessed date: 9 October 2018.

[27] B. Chen, Y. Ding, D.J. Wild, Improving integrative searching of systems chemical biology data using semantic annotation, J. Cheminf. 4 (2012) 6, https://doi.org/10.1186/1758-2946-4-6.

[28] M.E. Holford, J.P. McCusker, K.-H. Cheung, M. Krauthammer, A semantic web framework to integrate cancer omics data with biological knowledge, BMC Bioinf. 13 (2011) S10, https://doi.org/10.1186/1471-2105-13-S1-S10.

[29] P.E. Stang, P.B. Ryan, J.A. Racoosin, J.M. Overhage, A.G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, J. Woodcock, Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership, Ann. Intern. Med. 153 (2010) 600–606, https://doi.org/10.7326/0003-4819-153-9-201011020-00010.

[30] R.L. Fleurence, L.H. Curtis, R.M. Califf, R. Platt, J.V. Selby, J.S. Brown, Launching PCORnet, a national patient-centered clinical research network, J. Am. Med. Inf. Assoc. 21 (2014) 578–582, https://doi.org/10.1136/amiajnl-2014-002747.

[31] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), J. Am. Med. Inf. Assoc. 17 (2010) 124–130, https://doi.org/10.1136/jamia.2009.000893.

[32] R. Platt, M. Wilson, K.A. Chan, J.S. Benner, J. Marchibroda, M. McClellan, The new Sentinel network — improving the evidence of medical-product safety, N. Engl. J. Med. 361 (2009) 645–647, https://doi.org/10.1056/NEJMp0905338.