Collaborative Data Mining for Clinical Trial Analytics

Vandana P. Janeja[*1], Jay Gholap[1], Prathamesh Walkikar[1], Yelena Yesha[1], Naphtali Rishe[2], Michael A. Grasso[4],
[1]University of Maryland Baltimore County, Baltimore, USA,
[2]Florida International University, Miami, USA
[4]University of Maryland School of Medicine, Baltimore, USA

Abstract. Clinical research and drug development trials generate large amounts of data. Due to the dispersed nature of clinical trial data across multiple sites and heterogeneous databases, it remains a challenge to harness these trial data for analytics to gain more understanding about the implementation of studies as well as disease processes. Moreover, the veracity of the results from analytics is difficult to establish in such datasets. We make a two-fold contribution in this paper: First, we provide a mechanism to extract task-relevant data using Master Data Management (MDM) from a clinical trial database with data spread over several domain datasets. Second, we provide a method for validating findings by collaborative utilization of multiple data mining techniques, namely: classification, clustering, and association rule mining. Overall, our approach aims at extracting useful knowledge from data collected during clinical trials to enable the development of faster and cheaper clinical trials that more accurate and impactful. For a demonstration of the efficacy of our proposed methods, we utilized the following datasets: (1) the National Institute on Drug Abuse (NIDA) data share repository and (2) the data from the Osteoarthritis initiative (OAI), where we found real-world implications in validating the findings using multiple data mining methods in a collaborative manner. The comparative results with existing state of the art techniques show the usefulness and high accuracy of our methods.

Keywords: Collaborative data mining, clinical trial analytics, Master Data Management

---
[*]Corresponding author email: vjaneja@umbc.edu

1   Introduction

In this paper we propose a collaborative data mining approach, not only to optimize performance of data mining application but also to facilitate extraction of strong patterns from heterogeneous and modular data, which can be verified through multiple approaches. Traditionally, data mining is associated with finding interesting, non-trivial patterns in large datasets. However, extracting useful knowledge from data, whether small or very large, is a challenging task, considering that data mining models can be biased due to small sized training data or results can be inaccurate due to highly diverse training sets. To address this challenge, we propose a collaborative data mining framework to optimize the mining process. Here we combine the output of multiple data mining algorithms to generate a novel discovery in heterogeneous and disparate datasets.   Our collaborative approach is an ensemble of multiple data mining algorithms to validate and provide higher confidence in the veracity of the results.

Motivation: An important challenge faced by clinical trial analytics is the size of data. Due to sample sizes of both small and extremely large clinical trials, it is important to check consistency and robustness of results from mining of clinical trial data. Assuming that a clinical trial produces data that could reveal associations between patient characteristics, interventions and trial outcomes, clinical data analyses are used to validate these relationships. In the context of clinical trials with smaller number of participants, it is essential for clinical researchers to identify peculiarity between clinical outcome and confirmatory data analysis. It is significant to collect considerable preliminary insight on subjects based on historical data before the trial is conducted. For such trials, hypothesis testing might be challenging. Thus, it is logical that several different data mining techniques should be applied collaboratively on such clinical trial data. If multiple data mining techniques produce consistent results, one can be more confident that results are not due to unwarranted assumptions [1]. Hypotheses generated based on mining clinical trial data can enrich the design of clinical trials, hence reducing the duration and cost associated with trials.

We believe that insights gained from mining clinical data can be effectively utilized to address clinical trial design aspects such as: (a) managing participant enrollment (b) planning clinical trial follow-up (c) associating demographics and medical conditions of patients with clinical trial outcomes, and (d) predicting success of clinical trials based on the features associated with the trial.

Fig.1 depicts several steps involved in clinical trial planning. In early stages, a clinical team develops a clinical development plan for the trial. Once a formal protocol is finalized, investigators are selected to conduct the clinical trial in formal settings to test a newly developed drug or intervention. As a part of strategic planning, secondary data analyses of previous clinical trials are utilized in order to gain some interesting insights on clinical response trends. This is where our approach of collaborative data mining can be applied. With limited amounts of data, it is often challenging to generate hypotheses that may be helpful for deciding on patient enrollment, inclusion or exclusion criteria, prediction of clinical outcomes, etc.

Clinical trials are conducted to enhance medical knowledge related to certain treatments, disease conditions or testing new drugs on trial participants before regulatory approval and release to the market. In clinical trials, participants receive specific interventions in the form of medical treatment and procedures or change in the behavior. Several phases involved in clinical trials tend to generate a lot of data [2,3,4]. Let us consider two types of clinical trial datasets: (1) the National Institute on Drug Abuse (NIDA) data share repository and (2) data from the Osteoarthritis initiative (OAI) [5]. These clinical trial datasets have been completely de-identified to prevent linkage to individual participants [6].

National Institute on Drug abuse (NIDA)

The NIH-supported data share on the NIDA website promotes scientific research and encourages researchers to conduct secondary analyses of drug abuse clinical trial data. Datasets on NIDA are good learning tools for students and researchers. Most of the datasets are in Clinical Data Interchange Standards Consortium (CDISC) format. Detailed documentation regarding the dataset and corresponding study protocol are also available under each study. Prior to data download, users have to provide email, name, affiliation, etc. details in order to access data under the NIDA data use policies [6].

Data from Osteoarthritis Initiative (OAI)

The Osteoarthritis Initiative (OAI) is an ongoing multicenter clinical trial supported by NIH in order to study progression of knee osteoarthritis (OA). The OAI has about 5000 participants with clinically significant knee OA or having risk of developing knee OA. The OAI has made the trial data available on their website by accepting a data

use agreement. OAI aims at multiplying scientific research contributions by providing study data to the OA research community. The primary aim of the OAI is to facilitate research that could ultimately prevent progression of knee OA pain and disability. Datasets include a number of data files with information about biomarkers, subject characteristics, joint symptoms, medical history, medications etc. [7]. The data used in the preparation of this article were obtained from the Osteoarthritis Initiative (OAI) database, which is available for public access at http://www.oai.ucsf.edu/. Specific datasets used in our experiments are 0.2.2 baseline clinical datasets.

In these datasets the following challenges emerge:
1. Data are scattered in several domain datasets such as demographics, medical history, subject characteristics etc.
2. Clinical outcomes are not explicitly stored to perform secondary data analysis.
3. Clinical trials do not have more than a few thousand participants, so we have access to a limited size of data. Alternatively, there may be several trials that are similar in nature but highly scattered with large amounts of data.

To address these challenges within the domain of clinical trials, we make the following contributions in this paper:
1. Integrating clinical trial data from multiple data sources with ETL (extract-transform-load) driven Master Data Management (MDM) solution.
2. Applying collaborative data mining for extracting knowledge from clinical trial data to improve the design of clinical trials.
3. Collating and validating results from multiple data mining techniques to support our hypotheses.

The remainder of the paper is organized as follows. In Section 2 we discuss the related work that focuses on clinical data integration and collaborative learning based clinical data mining. In Section 3 we discuss the overall approach that integrates clinical trial data from several datasets and utilizes collaborative data mining. In Section 4 we explain our experiments and results with specific details of tools and datasets. Finally, in Section 4 we conclude with future directions for this research.

2 Related work

We study the related literature in clinical data integration, data mining for clinical data and collaborative data mining.

Clinical study data is generally scattered among multiple systems in various formats across various environments and organizations. Many organizations use manual, resource-intensive approach to consolidate key clinical information. Significant research effort has been directed towards consolidating data, removing inconsistency, reducing redundancy and integrating information from disparate clinical sources [8, 9]. Strategies and guidelines proposed by Cleven & Worthmann [10] for master data management (MDM) are very helpful to form a single version of integrated data from multiple data sources. Chow & Liu [11] also suggest the use of master data management techniques to combine clinical trial data to build a unified view of trial data appropriate for performing secondary analyses. Palmer [12] shows that MDM helps to store master data related to process-centric entities at a single repository preventing a significant amount of manual data cleansing and sourcing.

Clinical data mining refers to the science of extracting interesting patterns from clinical databases. Typically, state of the art data mining techniques (shown in Fig. 2) are widely categorized into three types: classification (supervised learning), clustering (unsupervised learning) and association rule mining.

We also summarize scholarly articles related to clinical trial analytics in Table 1 under the three broad categories of data mining discussed earlier.

According to Bose & Das [22], clinical trial analytics (CTA) help clinical research organizations to improve clinical trial design. Proposed CTAs use a variety of tools and techniques for decision support, querying, data mining and data visualization to gather and analyze data for clinical research. CTAs can be used for predicting outcomes of clinical trials and identifying trends of participants and hence can be useful for addressing challenging tasks such as patient enrollment and planning clinical trial follow-ups. Relyea [23] discusses analytic approaches to designing clinical trials for cancer; using predictive statistical modeling, they try to identify clinical trials that are unlikely to succeed, thus avoiding cost and time required to complete a clinical trial.

Collaborative data mining uses a cooperative learning approach, where multiple distinct data mining algorithms work together to solve a given data mining problem by utilizing strengths and weaknesses of each other as summarized in Table 2.

Data mining systems, such as [27,28,29], typically use a single type of machine learning technique to improve results. However, the collaborative data mining system that we have proposed in this paper makes hybrid use of entirely different techniques such as classification, association rule mining and clustering which are not a key characteristic of any of existing - collaborative methods.

For any hybrid collaborative data mining system, interaction between its agents is based upon certain communication protocol. The major challenge in such systems is about facilitating such inter-agent interaction to solve the given problem jointly. Panait & Luke [39] discuss *direct communication* as an important aspect of co-operative team learning where agents with similar or dissimilar behaviors communicate learned information amongst each other to boost the team performance. Our approach closely resembles to *heterogeneous team learning* proposed in [39], where we use classification based interaction for enabling such team learning within three diverse data mining agents. Existing literature related to such data mining techniques focuses on only homogenous team learning where only single type of learning agent is utilized. [41,42] discuss the application of ensemble learning techniques to increase the prediction accuracy. Multi-classifier scheme 'vote' [40] is another such scheme that uses group of classifiers and combination rule to finalize the combined prediction. However, use of diverse agents such as classification, association rules and clustering is what makes our collaborative data mining approach a novel and unique contribution to the subject area of heterogeneous team learning.

Our approach not only makes an effort to improve cooperation between learners but also enhances knowledge discovery leading to synergetic results. The proposed framework provides a two-in-one solution for integrating scattered clinical data and mining it collaboratively with the help of dissimilar machine learning techniques.

3 Collaborative Data Mining Methodology

Fig. 3 above illustrates our overall approach. We use master data management techniques using ETL (Extract-

transform-load) in order to integrate clinical trial data and then execute collaborative data mining on the data and validate the results yielded from the multiple techniques. We next discuss the steps performed for clinical trial analytics.

3.1 Master data management using Extract-Transform-Load (ETL)

For building Master data management (MDM) solution for clinical trial data, we use ETL techniques, which involve processes responsible for data extraction, data transformation and data load. Fig. 4 explains how reference data is collected from multiple sources and a master record is formed. As depicted in Fig. 4, there are three data sources A, B and C. Attributes with one-to-one mappings (Attribute A_1 and Attribute B_2) across data sources are considered directly in the master data record. Value of attribute A_1 from data source A and value of attribute B_1 are extracted directly. However, for attributes with one-to-many mapping, their aggregations are separately computed and added to the master data record as a single value one-to-one mapping value for example, aggregated value (such as average value) of attribute C_1 from data source C is added to master record of identifier 1. These master records can be further loaded to another data source and can be analyzed from the single point of reference. The single point of reference could be a master record of subject who is enrolled for clinical trials.

With ETL process, we integrate relevant data from domain datasets of clinical trial and form a master record of each subject with reference data including demographics attributes such as sex, ethnicity & family related information.

3.2 Collaborative data mining

In general, collaborative mining [30] is a technique where data mining is distributed to multiple collaborating agents to solve a given data mining problem. Our primary goal in using collaborative data mining is to take advantage of dissimilar data mining techniques to produce better and more validated solutions. Results obtained by one technique can be validated with the help of other data mining collaborators in a team. Similarly, by analyzing results of the first technique, users can eliminate weak patterns extracted from the data and concentrate on validating patterns of interest with the help of other techniques. This is similar to ensemble learning, however here we use mul-

tiple different types of data mining techniques whereas ensembles use the same model with different datasets.

The pseudo code in Fig. 5 demonstrates our generalized approach with the help of detailed algorithm for collaborative data mining. We adopt this generalized approach specific to datasets to mine interesting knowledge from the clinical trials as demonstrated through the two datasets in experimental results. In our approach, Classification is generally the default choice for building the prediction model. Similarly, class based association rules (CAR) extracted from the dataset can be utilized to produce a classification for a new instance. Classification via clustering is also a well-known technique to perform classification. We perform 'cluster to class' evaluation by assigning mode value of class label to individual cluster. As depicted in Fig. 6, we use these techniques to build prediction models and generate predictions by using each of these techniques. We utilize combined knowledge from each of these three distinct approaches which are integrated using the combiner module (voting module) as shown in Fig. 6. Here, we combine individual predictions with the help of majority voting and assign it as a final collaborative classification. Picking a single method does not necessarily provide the best results and thus it becomes a random choice to decide to some extent. Using a collaborative majority voting rule however accumulates individual quality results from multiple methods and thus, leads to elimination of chances of poorly performing methods. Secondly, this also adds veracity to the results since these results have been validated from multiple methods and hence would be accurate with a higher likelihood. As this paper discusses case studies in clinical trial analytics, schematic in Fig. 6 shows categorical features with clinical outcome as a class attribute. Any suitable algorithms for classification, association rules and clustering can be utilized for the framework.

Our approach calculates the majority vote by looking at individual predictions. For example, in Fig. 7, classification and class association rules both produced 'High' prediction where clustering generated 'Low' prediction. In this case, final collaborative prediction assigned was 'High' due to majority vote.

If classification, association rule mining and clustering produce the same prediction, then we identify it as an absolute majority. In the example shown in Fig. 8, all three techniques classication, CAR and clustering produce prediction 'X' for instance with ID=1. CAR rule

associated with this instance provides useful insights such as co-occurrence of prediction 'X' along with attribute A with value a and attribute B with value b.

4 Experimental Results

In this section, we outline the software tools and packages that we have used for our experiments, the datasets used and MDM solution using ETL to build data mining prototypes on a shortlisted clinical trials' study data. Further we focus on insights gained from data mining experiments. In the two datasets, we focus on specific mining tasks (a) In NIDA data we want to predict the trial outcome by using our approach of collaborative data mining and (b) In OAI data we want to predict the specific Osteoarthritis (OA) phenotype such as bone-driven OA, traumatic OA or cartilage-driven OA, as a function of different levels of pain, stiffness and disability in OA patients. These datasets have a limited and scattered data making it difficult to be consumed by traditional data mining tools. Our framework facilitates the data integration and proves to be useful in order to generate confident hypotheses with the help of a novel technique of collaborative data mining.

4.1 Software tools and packages

We used R packages arules and arulesviz for running association analysis on clinical trial data. For prediction of drug abstinence, we experimented with Weka to build the ensemble based prediction model. Weka offers a several clustering algorithms such as k-means, DBSCAN, however for clustering categorical data with similarity measures such as Gower's coefficient or Jaccard coefficient, we recommend R package called cluster. R provides a variety of data mining packages for advanced algorithms. Some of the packages that we suggest for interactive data mining include klaR, RWeka, extracat. For data integration purposes, we have used Talend open studio 5.4 as our ETL toolkit [31]. Collaborative framework was developed in Java and collated results were stored in MySQL database.

4.2 Datasets

As described in the introduction we use NIDA and OAI data for our experiment.

NIDA data: For building data mining prototype, we selected a clinical trial data gathered for the study on 'Brief Strategic Family Therapy (BSFT) For Adolescent Drug Abusers'. This study compares BSFT to treatment as usual (TAU) in reducing adolescent drug abuse. The study also examines the effect of family interventions such as involvement of adolescents in family activities on drug abstinence, which is the primary outcome of the study. Table 3 lists domain datasets, which were utilized for our analysis.

Osteoarthritis Initiative (OAI) Data: OAI datasets include information of 5000 OA patients & their joint symptoms, medical history, biomarkers, demographics of enrollees, medications, physical examinations, MRI and X-ray [5]. For our experiments, we used information from domain data sets mentioned in table 4 below. We extracted attributes indicating pain, stiffness and disability levels of patients. We then looked at attributes that were significant to identify different phenotypes of OA and developed a rule based expert system that consumed these clinical features and identified the particular type of OA phenotype.

According to NIAMS (National Institute of Arthritis and Musculoskeletal and Skin Diseases), Osteoarthritis (OA) is a joint disease with a variety of disorders leading to structural or functional failure of joints. We aim to identify several phenotypes such as bone-driven OA, cartilage-driven OA and traumatic OA. We used known, prevalent clinical rules to identify each of the phenotypes as displayed in the table 5.

A primary aim of OAI is to study different levels of pain, stiffness and disability amongst patients suffering from knee OA. We extracted WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index) scores for pain, stiffness and disability from joint symptoms data file. WOMAC score is one of the widely used proprietary questionnaire based health status indices which is used to understand function of pain, stiffness and disability amongst hip and knee OA patients [32].

4.3 MDM solution using ETL

4.3.1 Processing NIDA Data

  MDM schematic demonstrated in Fig. 9 was developed and translated to code with Talend data integration studio

[31]. Master data record included attributes derived from demographics, subject characteristics, questionnaires and substance use file. Drug abstinence for each subject in clinical trial was calculated from substance-use file based on timeline follow back method as shown in Fig. 10.

The selected drug abuse clinical trial uses timeline follow back (TLFB) method [33] to obtain quantitative estimates of drug use. TLFB process uses a regular calendar to help people remember what substances were used on each day over a specified time period. For calculating drug abstinence, we used percentage of days when subject showed abstinence to drugs such as alcohol, marijuana, tobacco etc. We further discretize percentage drug abstinence into five levels such as low, very low, moderate, high and very high. Different subject characteristics were stored in separate rows for each subject. Row to column transformation was used in order to derive characteristics and their values in columnar format as shown in Fig. 11. Similar transformation was required for questionnaires dataset as well. Demographics dataset had information of subjects who failed the screening or were not able to complete the study due to some reason. We had to filter out such records since other files did not include associated data for these participants. Finally, the remaining attributes are shown in Fig. 12.

4.3.2 Processing OAI data

Before we mine the data we had to label the data with different types of OA phenotypes which was a key challenge since we needed to integrate related phenotype data from different related data sources which include biomarkers, medical history and joint symptoms. We developed an expert system for identification of OA phenotypes. It uses several clinical rules. It integrates different clinical features from biomarkers, medical history and joint symptom datasets as depicted in the Fig. 13 and finds out a specific type of OA phenotype by applying clinical rules. Clinical rules were collected from the domain experts and were translated into the knowledge base. Master data in this case was also derived with the help of ETL data integration job. Phenotype labels were obtained with clinical rule base knowledge. WOMAC scores for pain, stiffness and disability were discretized to obtain categorical labels as shown in the below table. WOMAC score ranges in the Table 6 were captured from [34]. Discretization was performed based on equal width binning. Scores with 0 values were assigned label - 'None'.

In our preliminary data analysis at an aggregate level, we look at the distribution of knee-OA patients with the presence of multiple phenotypes such as bone-driven OA,

traumatic OA and cartilage-driven OA from the data derived from the Osteoarthritis Initiative (OAI) and study the overlaps between different types of OA phenotypes as shown in Fig. 14. Interestingly, we noticed that highest number of patients reported pain, stiffness and disabilities when all three phenotypes were identified. This also validates our clinical rules used for the phenotype identification.

4.4 Data mining results

In this section, we discuss our findings on applying individual data mining techniques on clinical trial data integrated by MDM solution. Then, we do collaborative analysis to find correlations in the results.

4.4.1 Exploratory analysis (NIDA Data)

Results of individual data mining techniques are presented in Fig. 15. We investigate the association rules for subjects with similar demographic characteristics. Interestingly, we found comparable rules showing higher drug abstinence associated with such participants as shown in Fig. 15(c). In order to validate this further using our prediction model, we created a test dataset of subjects with similar demographic characteristics & outcomes and evaluated our prediction model on this test dataset. Our prediction model correctly classified 63 instances out of 85 showing 74% of accuracy as shown in Fig. 15(a). Visualization of classifier errors in Fig. 15(b) clearly shows the same. By clustering clinical trial data, we obtained a separate cluster of male, Mexican American adolescent subjects who do not involve much in the family activities, but show high drug abstinence. Parallel co-ordinate plot of categorical data in Fig. 15(d) from this cluster makes it easy to understand this fact.

As discussed in previous sections, one of our collaborative analyses indicated that Mexican American adolescent population typically show higher drug abstinence. This was confirmed by several research articles and the fact sheet given on NIAAA which claims: "Hispanics have high rates of abstinence from alcohol" [35]. Most research on substance abuse among Hispanics is focused on alcohol abstinence and has confirmed that Hispanics show higher alcohol abstinence [36,37].

4.4.2 Collaborative Mining in Drug abstinence prediction (NIDA Data)

Based on results shown in Table 7 and Fig. 16, it is clear that majority voting outperformed all other individual techniques by collaborative use of multiple data mining techniques. It is interesting to note that collaborative data mining yielded accurate results as compared to ensemble based Adaboost algorithm with a single type of classifier and multi-classifier technique 'vote' [40]. For our experiments, we used C4.5 decision tree as a base classifier with Adaboost. For multi-classifier based 'vote', we utilized C4.5, Naïve Bayes and Random forest algorithm. To combine predictions, majority voting was used as a combination rule since the same combination rule was used in collaborative data mining. Our motivation here was to evaluate the performance of team of homogeneous learners such as 'vote' [40] versus collaborative data mining agents with heterogeneous and diverse learning capabilities. Although we use three completely different classifiers with 'vote' algorithm, their similar way of building prediction model by learning trends in the training data makes them less versatile as compared to our collaborative data mining. At the same time, Adaboost with single type of classifier produces much better results due to its iterative weighting process.

Our aim here is not only to improve the predictive performance but also to facilitate extracting strong patterns by looking at predictions generated with absolute majority as shown in Fig. 17. For example, the second rule in Fig. 17 clearly validates the finding that Mexican American population shows high abstinence level. These rules also exhibit the ground truth that Hispanic show better abstinence as compared to non-Hispanic population.

4.4.3 Collaborative mining for Prediction of bone-driven osteoarthritis phenotype (OAI Data)

Fig. 18: Graph for bone-driven OA prediction evaluation measures

After evaluating our classification model on preprocessed OAI data, we noticed that there was a significant scope for the improvement of classification accuracy as shown in Table 8. This was due to the co-occurrence of multiple phenotypes of OA in the patient records. If multiple phenotypes are present at the same time, it is likely that a patient will report pain, stiffness or disability due to OA as there is overlap of characteristics across phenotypes. In order to rectify this, we eliminated patient records having either cartilage driven OA or traumatic OA from the training set. Additionally, we added 122 patient records from

controlled sub-cohort of OAI study with no risks or any type of knee OA. We noticed around a 10% improvement in the accuracy as shown in Table 9. Since the classification model was one of the driving learners of the collaborative system, improvement in the accuracy of majority voting was also noticed. Fig. 18 represent the results with the improved performance in the training data sets. It is important to note that OA phenotypes are evolving as we gain more understanding of the causes and progression of OA. For this analysis, our assumptions about OA phenotypes have evolved from the understanding that this is a computational demonstration. Our hope is that computational methods will help to develop accurate phenotypes for earlier and more accurate diagnosis.

We can observe in Table 11 that majority voting had better accuracy as compared to classification and association rules because it collects votes from all individual data mining agents and assigns majority vote as prediction. Classification via clustering had the best accuracy amongst all because there were only few categorical variables in the training set and we had binary class problem.

In the above experiments, training set had imbalanced distribution of positive and negative cases of bone-driven OA patients. We addressed the class imbalance problem and reduced positive cases to make to uniform distribution with equal number of 534 for positive and negative instances. We evaluated collaborative data mining to predict bone-driven OA on this dataset. Results are displayed in Table 10 and Fig. 19. Fig. 20 shows the rules corresponding to absolute majority in the balanced data. As we can see, the first rule indicates that mild level of pain and disability typically hints bone-driven OA and second rule indicates the obvious truth that absence of pain, stiffness and disability is likely to show the absence of bone-driven OA.

Results obtained here are somewhat consistent in a way that collaborative data mining technique gives accuracy close to accuracy of the best method in the group of data mining agents. In this case, clustering produced the best results. This highlights the advantage of using collaborative data mining with multiple techniques. We cannot guarantee that specific individual technique would always work the best for given training data, so majority voting in collaborative data mining in this case guarantees that results are close to the best possible method.

4.5 Collaborative Data Mining in Big Data Environment

Collaborative data mining approach proposed in this paper can also be implemented in a big data environment to mine a very large clinical data warehouse. We propose big data analytics architecture shown in Fig. 21 for mining a large amount of heterogeneous clinical data store, to build enhanced clinical data analytics solution. We used SQL-like big tools to extract a cohort of clinical data points using interactive SQL queries on data residing on Hadoop Distributed File System (HDFS). Extracted data related to specific clinical decision making questions is passed to data analytics environment using R to execute collaborative data mining.

For implementing the proof of concept of our big data analytics solution for collaborative data mining, we utilized Cloudera based single node setup of Hadoop. This is also compatible to a multi-node setup. From a large clinical data store, we extracted analysis specific dataset into R environment using the package called RImpala [38] that allows us to execute distributed SQL queries in SQL-on-Hadoop tools such as Cloudera Impala and Apache Hive. RImpala [38] is developed by Mu-Sigma and is being extensively used in the big data industry for integrating the statistical power of R language in enterprise big data applications. We tested this solution on NIDA datasets by replicating different study datasets to simulate a typical big data store. Due to integration of Hadoop technologies with R, the presented architecture is easy to implement and adds a great value to the area of big data analytics in healthcare domain.

4.6 Discussion of Findings

We highlight some key findings about the collaborative mining and specific findings in the two datasets:
1. We presented a consolidated view of clinical trial data in case of NIDA as well as OAI facilitating the knowledge discovery.
2. Our collaborative data mining technique predicted clinical trial outcomes with the highest accuracy for NIDA drug abuse clinical trial dataset. Predicted outcomes and patterns were also validated with the ground truth found in the literature.
3. In NIDA and OAI datasets, primary clinical outcomes were not readily available. Extensive data processing with the help of ETL helped us with the quantification of outcomes before they could be used for pre-

dictive modeling. This type of preprocessing shows the data related challenges in real world clinical trial datasets.

4. In OAI datasets, data was first labelled with different phenotypes by the expert system built on the top of knowledge base collected from the domain experts. Collaborative data mining not only boosted the predictive performance of phenotype identification but also yielded several interesting rules indicating useful combinations of pain, stiffness, disability associated with presence of OA phenotypes.

5. Collaborative data mining produced better prediction results as compared to traditional multi-classifier based ensemble methods such as Adaboost and 'vote'.

6. Proposed big data architecture for collaborative data mining makes it easy to extract and mine study specific trial datasets from a repository of very large clinical trial data.

Based on our survey and experimental findings, we strongly believe that work proposed in this paper lays the foundation for diverse based collaborative data mining. We would like to highlight the following major challenges addressed in our work on collaborative data mining that address research topics in the data science and analytics community:

1. Facilitating inter-agent communication between diverse types of mining agents to learn shared information.

2. Existing collaborative data mining techniques mainly use only classification based learners. Clustering algorithms as well as association rule mining are fairly underused in the collaborative setting.

3. Merging common prediction rules from diverse learners needs to be addressed in order to build scalable and accurate algorithm.

5   Conclusion

In this paper, we have proposed a framework for analyzing clinical trial data using a novel approach of collaborative data mining. Using osteoarthritis clinical trial and drug abstinence clinical trial data as our case studies, we have observed that results are consistent and data-driven across various techniques and can be validated with the help of several confirmatory analyses. However, it is important to note that the findings in the datasets are not necessarily validations of clinical findings but validation of our proposed data mining approach. The proposed approach, when combined with clinical knowledge, can be insightful for subject enrollments, planning trial follow-ups, and thus can provide an early predictor as to whether clinical trials would succeed. The proposed big data architecture for implementing collaborative data mining looks promising and can be very useful in mining smaller study specific datasets within a large clinical data warehouse.

In our future work, we plan to extend this framework to combine common prediction rules captured with multiple techniques and to evaluate them on separate test datasets with a hold-out method. In this paper, we have presented an evaluation of a collaborative data mining framework on smaller clinical trial datasets.

6   Acknowledgement

References

[1]   J. and S. T. I. Charles H. Evans, *Small clinical trials issues and challenges*. Washington, D.C. : National Academy Press, 2001.

[2]   U.S. National Institutes of Health, "ClinicalTrials.gov," *clinicaltrials.gov*, 2013. [Online]. Available:

http://clinicaltrials.gov/.

[3]   D. a Zarin, T. Tse, R. J. Williams, R. M. Califf, and N. C. Ide, "The ClinicalTrials.gov results database--update and key issues.," N. Engl. J. Med., vol. 364, no. 9, pp. 852–60, 2011.

[4]   "Learn About Breast Cancer Trials at BreastCancerTrials.org." [Online]. Available: https://www.breastcancertrials.org/bct_nation/home.seam. [Accessed: 16-Feb-2016].

[5]   "OAI:Home." [Online]. Available: http://www.oai.ucsf.edu/datarelease/. [Accessed: 16-Feb-2016].

[6]   "NIDA Data Share | datashare.nida.nih.gov." [Online]. Available: https://datashare.nida.nih.gov/. [Accessed: 16-Feb-2016].

[7]   "OAI Study Protocol." [Online]. Available: http://www.oai.ucsf.edu/datarelease/docs/StudyDesignProtocol.pdf. [Accessed: 16-Feb-2016].

[8]   P. Lopes, L. B. Silva, and J. L. Oliveira, "Challenges and opportunities for exploring patient-level data," Biomed Res. Int., vol. 2015, 2015.

[9]   K. Jiang, "Integrating clinical trial data for decision making via web services." Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf., vol. 5, pp. 3346–9, Jan. 2004.

[10]   A. Cleven and F. Wortmann, "Uncovering four strategies to approach master data management," in Proceedings of the Annual Hawaii International Conference on System Sciences, 2010.

[11]   S.-C. Chow and J.-P. Liu, Design and analysis of clinical trials: concept and methodologies. 1998.

[12]   J. Palmer, "The Clinical Data Warehouse – a New Mission-Critical Hub." [Online]. Available: http://www.oracle.com/us/industries/health-sciences/clinical-data-warehouse-hub-2272366.pdf.

[13]   I. A. Pilih, D. Mladenic, N. Lavrac, and T. S. Prevec, "Using machine learning for outcome prediction of patients with severe head injury," in Proceedings of Computer Based Medical Systems, pp. 200–204.

[14]   G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," IEEE J. Biomed. Heal. Informatics, vol. 18, no. 6, pp. 1750–1756, 2014.

[15]   L. Tanner, M. Schreiber, J. G. H. Low, A. Ong, T. Tolfvenstam, Y. L. Lai, L. C. Ng, Y. S. Leo, L. T. Puong, S. G. Vasudevan, C. P. Simmons, M. L. Hibberd, and E. E. Ooi, "Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness," PLoS Negl. Trop. Dis., vol. 2, no. 3, 2008.

[16]   J. P. Feighner, L. Sverdlov, G. Nicolau, and J. F. Noble, "Cluster analysis of clinical data to identify subtypes within a study population following treatment with a new pentapeptide antidepressant.," Int. J. Neuropsychopharmacol., vol. 3, no. 3, pp. 237–242, Sep. 2000.

[17]   T. Hao, A. Rusanov, M. R. Boland, and C. Weng, "Clustering clinical trials with similar eligibility criteria features," J. Biomed. Inform., vol. 52, pp. 112–120, 2014.

[18]   G. Bruno, T. Cerquitelli, S. Chiusano, and X. Xiao, "A Clustering-Based Approach to Analyse Examinations for Diabetic Patients," in 2014 IEEE International Conference on Healthcare Informatics, 2014, pp. 45–50.

[19]   A. Wright, A. McCoy, S. Henkin, M. Flaherty, and D. Sittig, "Validation of an association rule mining-based method to infer associations between medications and problems.," Appl. Clin. Inform., vol. 4, no. 1, pp. 100–9, 2013.

[20]   S. Stilou, P. D. Bamidis, N. Maglaveras, and C. Pappas, "Mining association rules from clinical databases: An intelligent diagnostic process in healthcare," in Studies in Health Technology and Informatics, 2001, vol. 84, pp. 1399–1403.

[21]   C. Wang, X.-J. Guo, J.-F. Xu, C. Wu, Y.-L. Sun, X.-F. Ye, W. Qian, X.-Q. Ma, W.-M. Du, and J. He, "Exploration of the association rules mining technique for the signal detection of adverse drug events in spontaneous reporting systems." PLoS One, vol. 7, no. 7, p. e40561, 2012.

[22]   A. Bose and S. Das, "Trial analytics - A tool for clinical trial management," Acta Poloniae Pharmaceutica - Drug Research, vol. 69, no. 3. pp. 523–533, 2012.

[23]   Stephen L. Relyea, "An analytics approach to designing clinical trials for cancer." Massachusetts Institute of Technology, 2013.

[24]   Q. Li and R. Khosla, "Performance optimization of data mining applications using a multi-layered multi-agent data mining architecture," in CIMSA. 2005 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, 2005, pp. 227–231.

[25]   X. Zhu, Z. Huang, and H. Zhou, "Design of a

Multi-agent Based Intelligent Intrusion Detection System," in 1st International Symposium on Pervasive Computing and Applications, 2006, pp. 290–295.

[26]   M. F. Santos, F. Portela, and M. Vilas-Boas, "INTCARE : multi-agent approach for real-time intelligent decision support in intensive medicine," 2011.

[27]   L. F. Schroeder and A. L. C. Bazzan, "A Multi-agent System to Facilitate Knowledge Discovery: an Application to Bioinformatics," Proc. Work. Bioinforma. Multi-Agent Syst. {(BIXMAS'2002)}, pp. 44–50, 2002.

[28]   H. L. Viktor, "Cooperating to learn: knowledge discovery through intelligent learning agents," in Proceedings Fourth International Conference on MultiAgent Systems, 2000, pp. 453–454.

[29]   J. Gao, J. Denzinger, and R. C. James, "CoLe: A cooperative data mining approach and its application to early diabetes detection," in Proceedings - IEEE International Conference on Data Mining, ICDM, 2005, pp. 617–620.

[30]   S. Moyle, "Collaborative Data Mining," in Data Mining and Knowledge Discovery Handbook, 2010, pp. 1029 –1039.

[31]   "Talend Open Studio: Open Source ETL & Data Integration."   [Online].   Available: https://www.talend.com/products/talend-open-studio.

[32]   "Western Ontario & McMaster Universities Osteoarthritis Index (WOMUOI)." [Online]. Available: http://www.rheumatology.org/I-Am-A/Rheumatologist/Research/Clinician-Researchers/Western-Ontario-McMaster-Universities-Osteoarthritis-Index-WOMAC.

[33]   "Instrument: Timeline Followback Method Assessment | NIDA CTN Common Data Elements." [Online].   Available: http://cde.drugabuse.gov/instrument/d89c8e23-16e5-625a-e040-bb89ad43465d. [Accessed: 16-Feb-2016].

[34]   "WOMAC Osteoarthritis Index."

[35]   "Alcohol and the Hispanic Community." [Online].   Available: http://pubs.niaaa.nih.gov/publications/HispanicFact/HispanicFact.htm. [Accessed: 16-Feb-2016].

[36]   Alcohol Consumption among Mexicans and Mexican Americans: A Binational Perspective. Spanish Speaking Mental Health Center, University of California, 1988.

[37]   G. Canino, "Alcohol use and misuse among Hispanic women: selected factors, processes, and studies," Int J Addict, vol. 29, no. 9, pp. 1083–1100, 1994.

[38] "CRAN - Package RImpala." [Online]. Available: https://cran.r-project.org/web/packages/RImpala/index.html.

[39] L. Panait and S. Luke, "Cooperative Multi-Agent Learning: The State of the Art," Auton. Agent. Multi. Agent. Syst., vol. 11, no. 3, pp. 387–434, Nov. 2005.

[40] Ludmila I. Kuncheva (2004). Combining Pattern Classifiers: Methods and Algorithms. John Wiley and Sons, Inc.

[41] Pourhomayoun, M., Alshurafa, N., Mortazavi, B., Ghasemzadeh, H., Sideris, K., Sadeghi, B., Sarrafzadeh, M. (2014). Multiple model analytics for adverse event prediction in remote health monitoring systems. 2014 IEEE Healthcare Innovation Conference (HIC).

[42] West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. European Journal of Operational Research,162(2), 532-551.

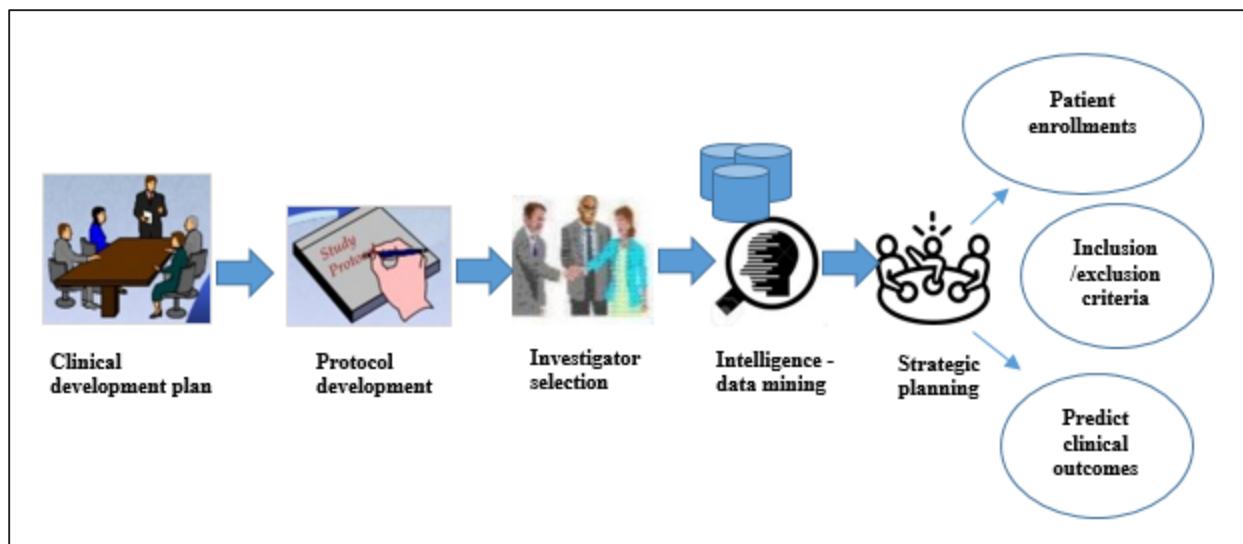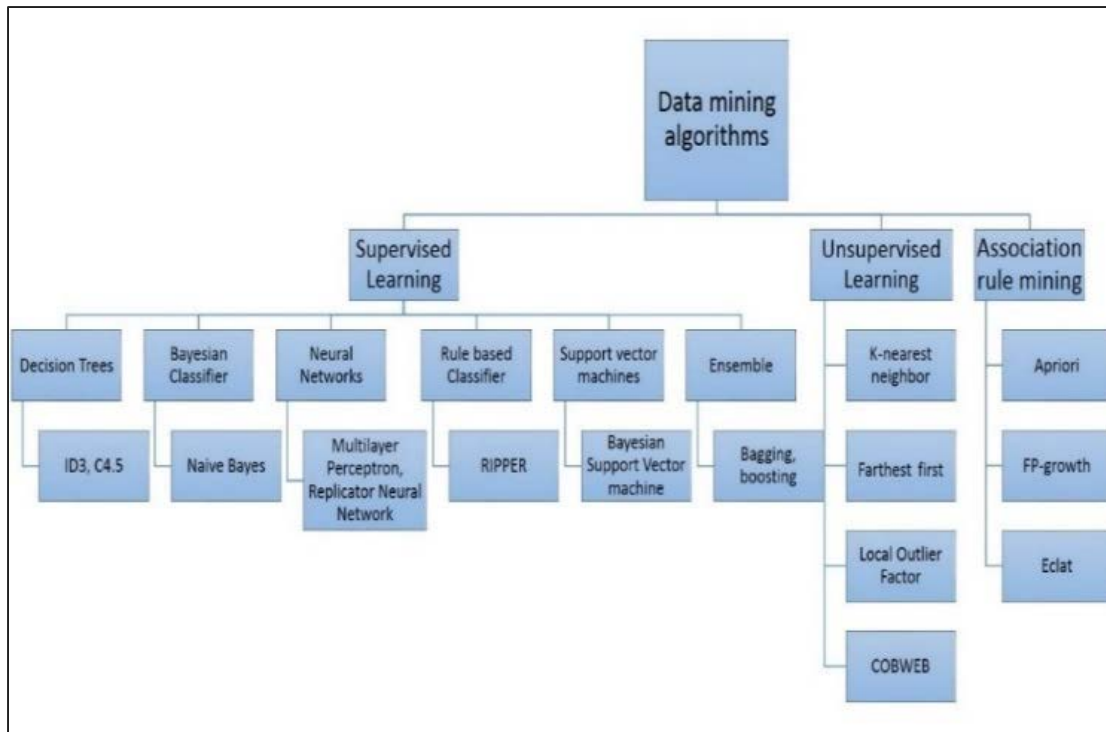**Fig. 1: Clinical trial planning**
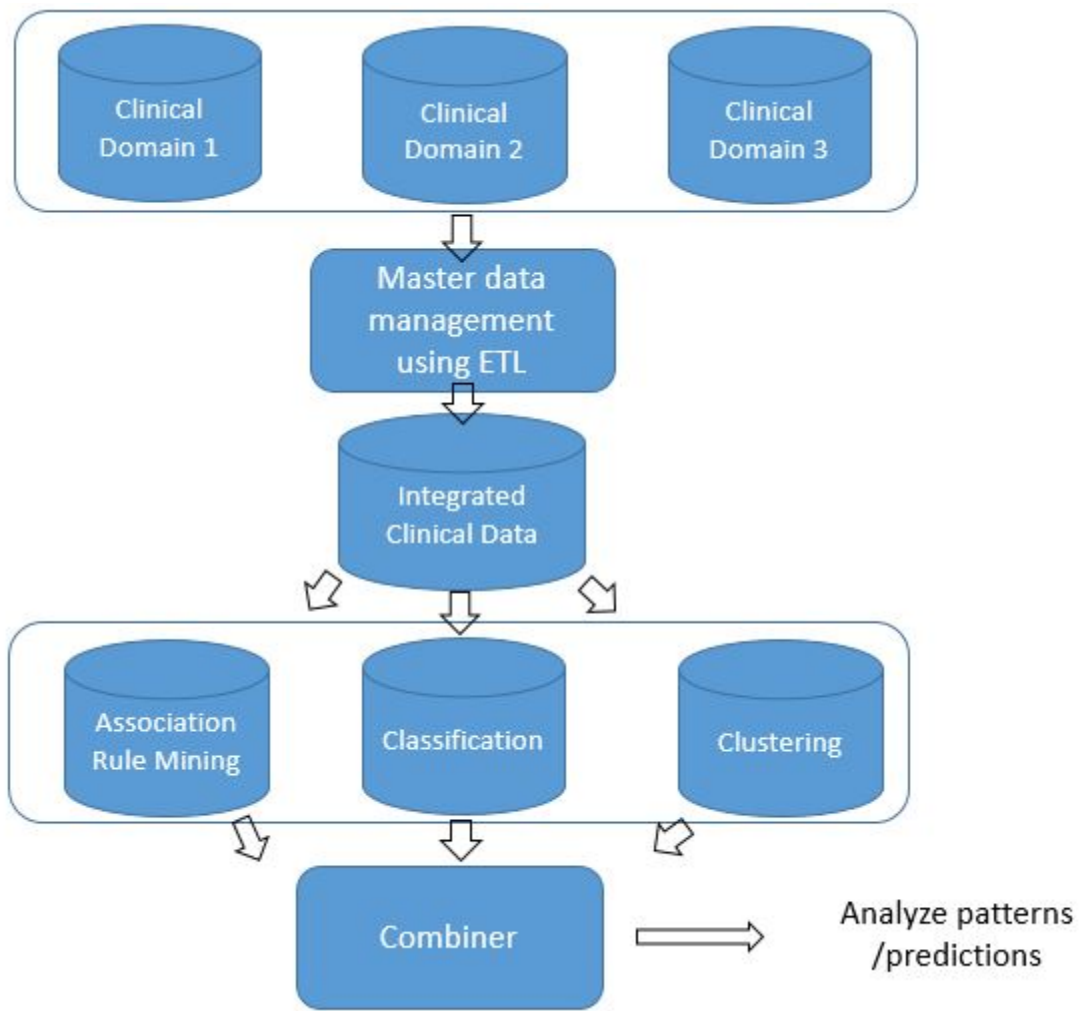
Fig. 2: Various data mining techniques
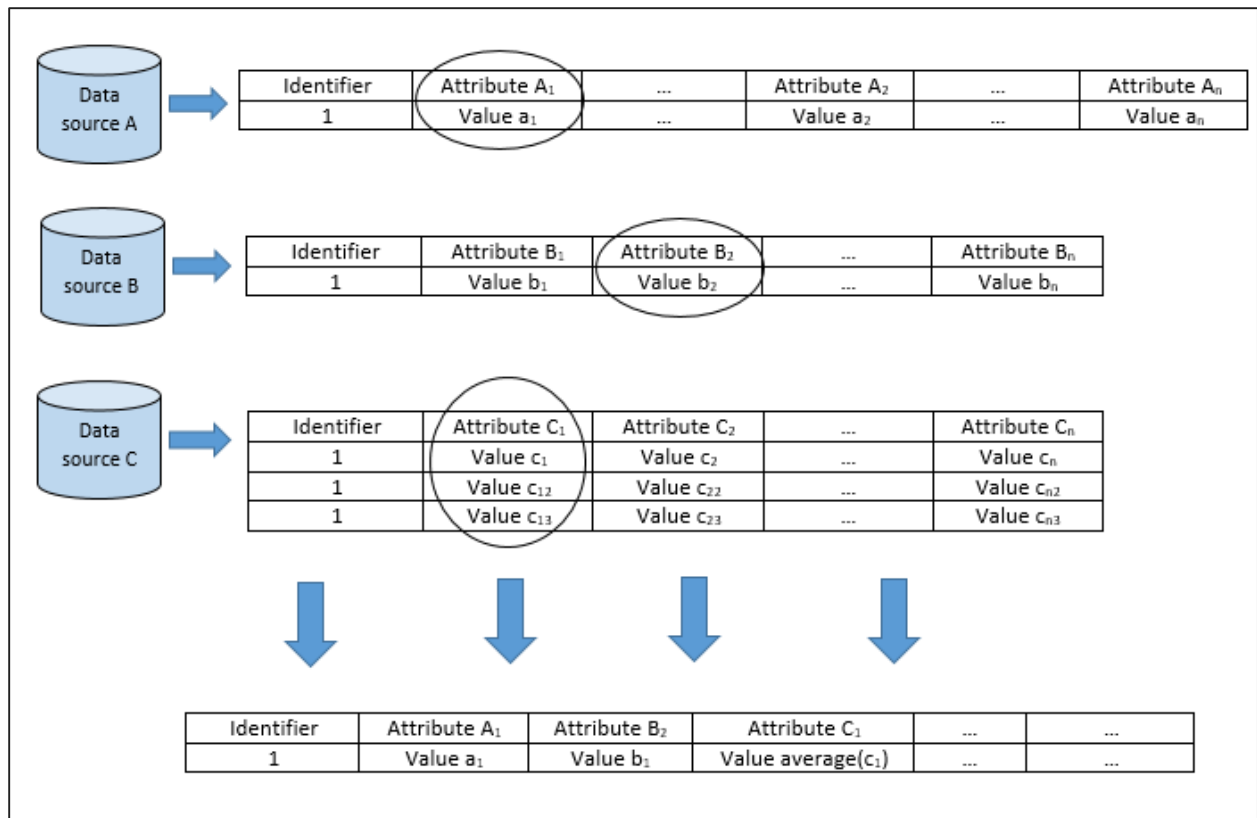
Fig. 3: Overall approach

Fig. 4: Formation of master record using MDM

Input:

D_{in}: input dataset with categorical features

minSup: minimum support value for extracting association rules

minConf: minimum confidence value for extracting association rules

C_{Class}: class attribute from D_i

Output:

$O_{majority}$: prediction with majority voting per instance in dataset

start;

classification(D_{in});

class_association_rules(D_{in});

clustering (D_{in});

cluster_car(D_{in});

for each record rec in Din

Omajority = majority_vote
(rec[classifierclass],rec[carclass],rec[clusteringclass],rec[cluster_carclass]);

end for;
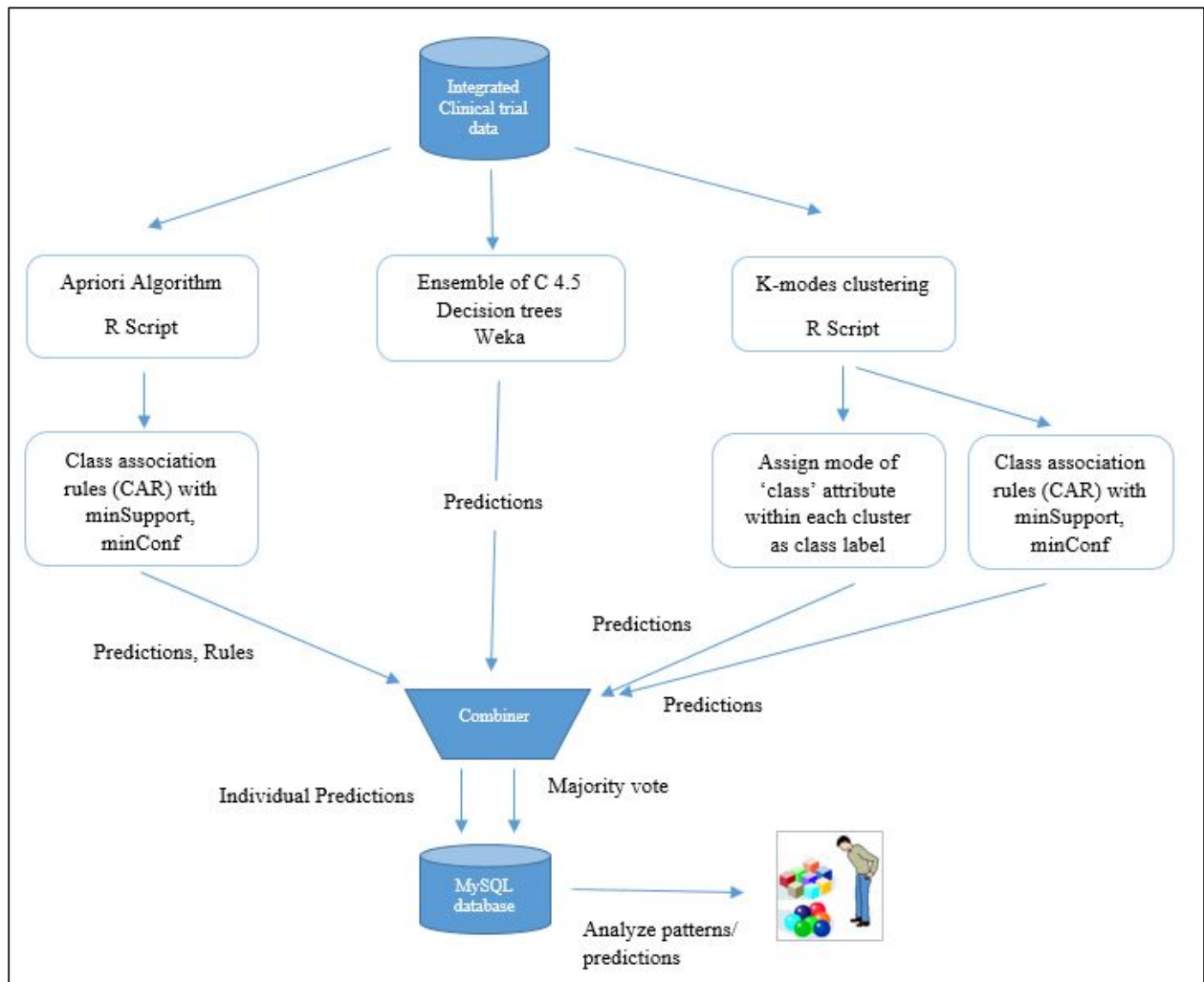
end;

Fig. 5: Pseudo code for collaborative data mining

Fig. 6: Schematic for Collaborative Data Mining

ID	Classification	CAR	Clustering	CAR_rule
1	X	X	X	{A=a, B=b -> Class= X}
2	Y	Y	Y	{A=a, C=c -> Class= Y}

Fig. 8: Absolute majority
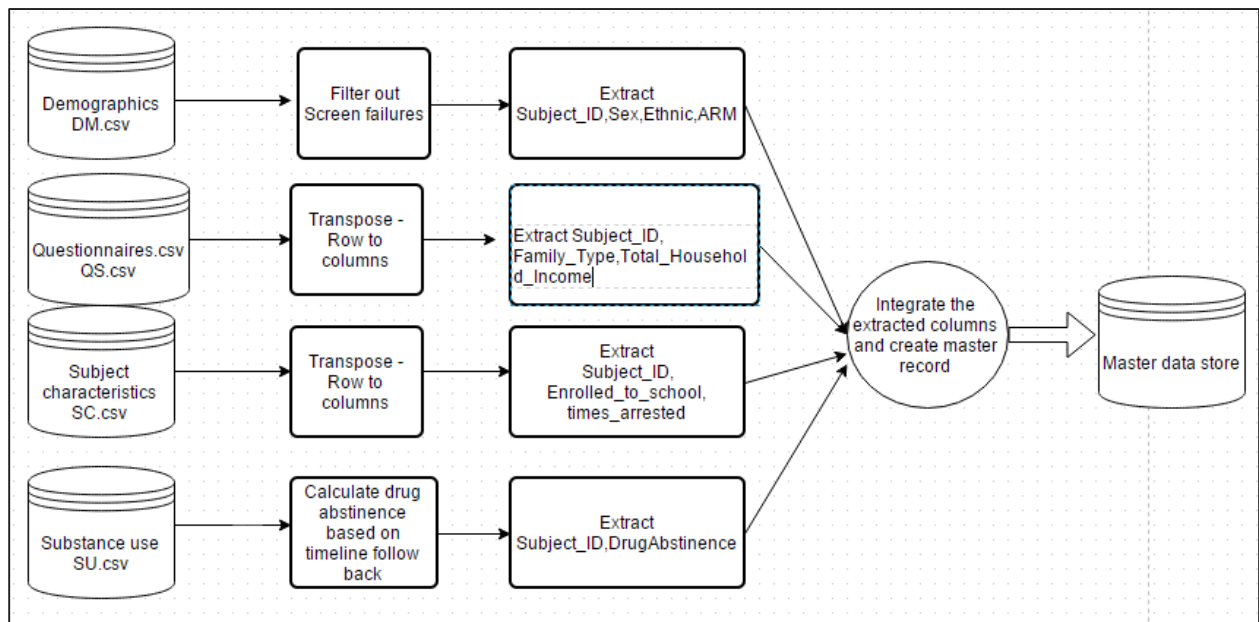
Fig. 9: MDM approach

Day #	Substance Use
1	Yes
2	Yes
3	No
…	…
…	…

Drug Abstinence = no. of instances when subject abstained from drugs/ total no. of instances per subject

Timeline follow back method

Fig. 10: Timeline follow back and calculation of drug abstinence

Subject ID	QSTEST	QSORRES
1	FAMILY TYPE	BLENDED
1	TOTAL HOUSEHOLD INCOME	$10,000-$14,999
2	FAMILY TYPE	BIOLOGICAL- 1 PARENT
2	TOTAL HOUSEHOLD INCOME	$15,000-$19,999

Transpose
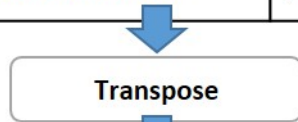
Subject ID	FAMILY TYPE	TOTAL HOUSEHOLD INCOME
1	BLENDED	$10,000-$14,999
2	BIOLOGICAL- 1 PARENT	$15,000-$19,999

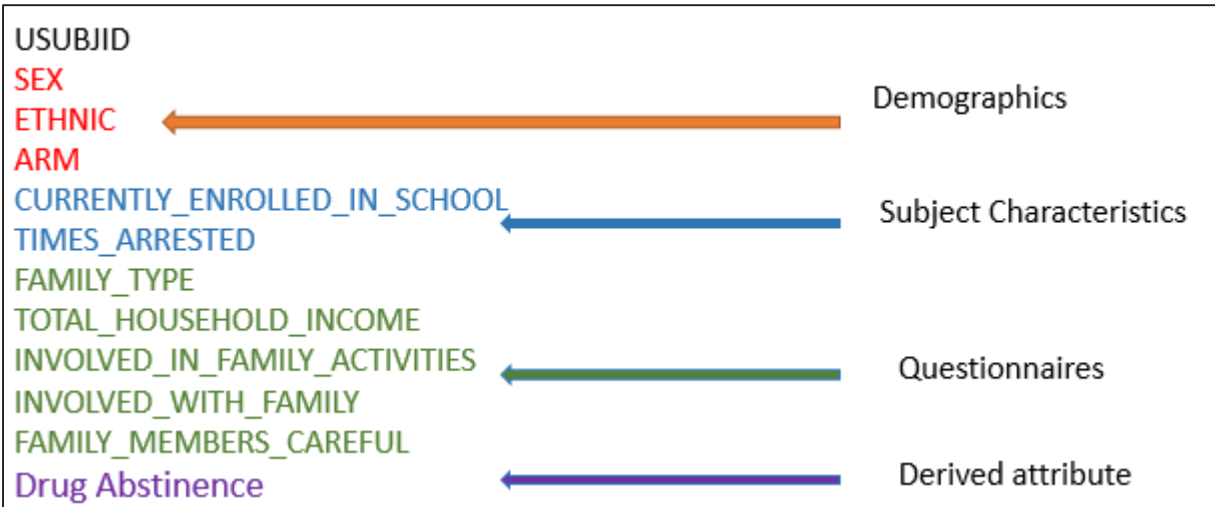**Fig. 11: Row to column transformation for questionnaires dataset**

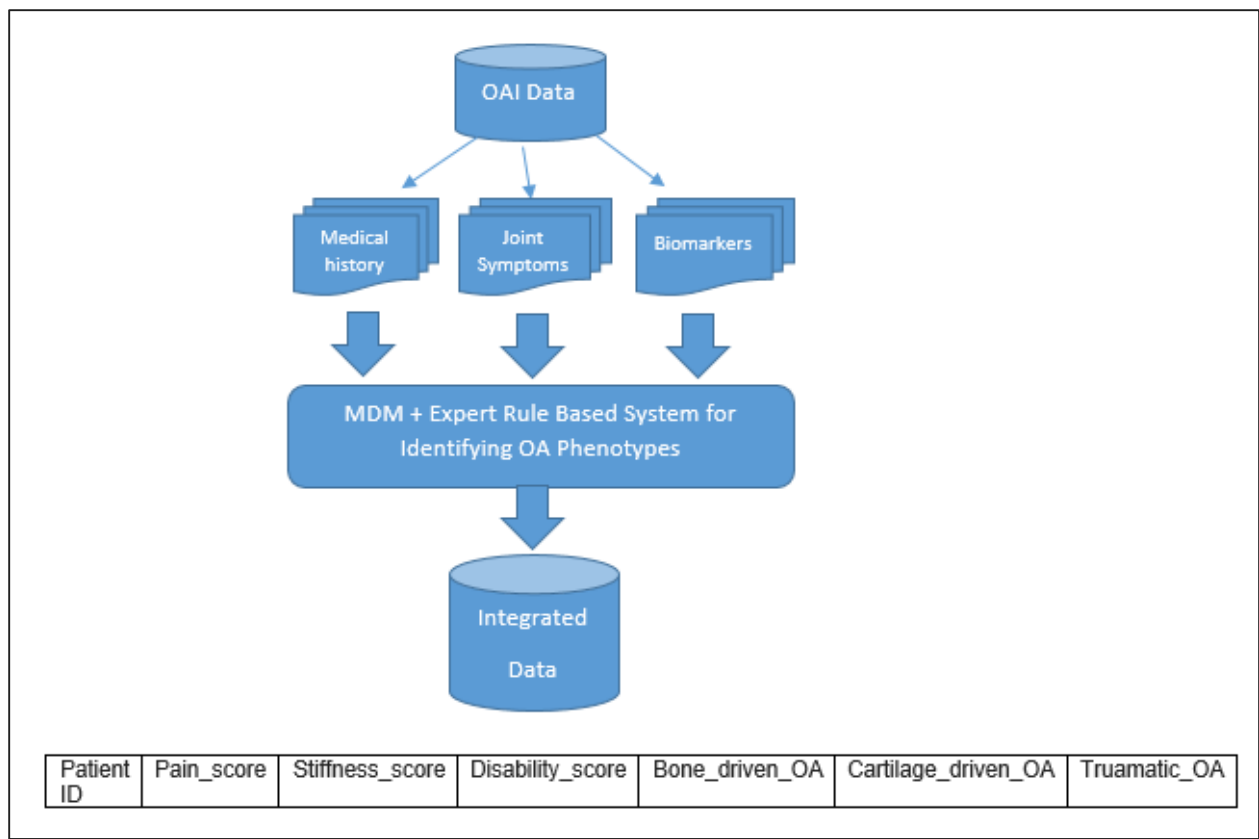**Fig. 12: List of attributes in clinical master data.**

Fig. 13: Approach for processing OAI data

Total Positive Cases	4321
Positive cases of Bone-driven OA (B)	4078
Positive cases of Cartilage-driven OA (C)	2142
Positive cases of Traumatic OA (T)	2269
Positive cases of Bone-driven OA & Cartilage-driven OA (B ∩ C)	2085
Positive cases of Bone-driven OA & Cartilage-driven OA but not Traumatic OA (B ∩ C ∩ T^c)	916
Positive cases of Bone-driven OA & Traumatic OA (B ∩ T)	2058
Positive cases of Bone-driven OA & Traumatic OA but not Cartilage-driven OA (B ∩ T ∩ C^c)	889
Positive cases of Cartilage-driven OA & Traumatic OA (C ∩ T)	1194
Positive cases of Cartilage-driven OA & Traumatic OA but not bone-driven OA (C ∩ T ∩ B^c)	25
Positive cases of Bone-driven OA & Cartilage-driven OA & Traumatic OA (B ∩ C ∩ T)	1169

$$n(B \cup C \cup T) = n(B) + n(C) + n(T) - n(B \cap T) - n(B \cap C) - n(C \cap T) + n(B \cap C \cap T)$$
$$= 4078 + 2142 + 2269 - 2058 - 2085 - 1194 + 1169$$
$$= 4321.$$

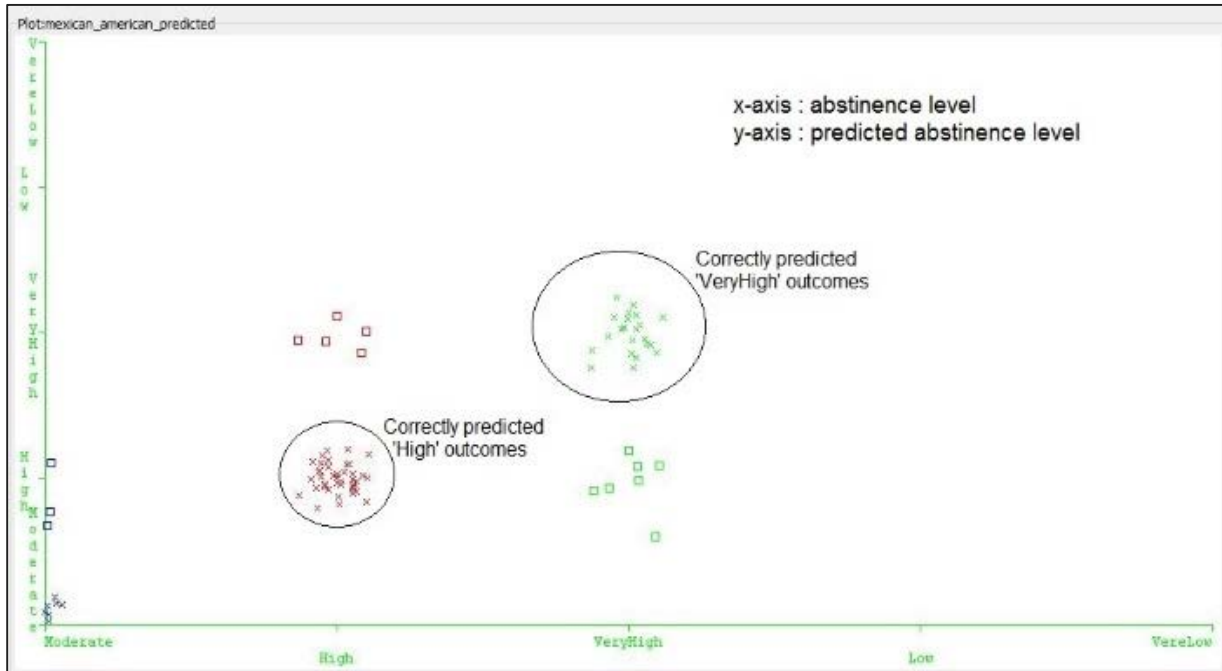| Phenotypes	Pain	Stiffness	Disability
Bone driven OA + Cartilage driven OA (916 patients)	62.12 % Mild 27.51 % None 09.49 % Moderate 00.65 % Severe	50.87 % Mild 26.85 % None 20.41 % Moderate 01.85 % Severe	66.15% Mild 21.83% None 11.35% Moderate 0.65 % Severe
Bone driven OA + Traumatic OA (889 patients)	63.66 % Mild 28.12 % None 7.76 % Moderate 01.85 % Severe	57.87 % Mild 24.07 % None 16.64 % Moderate 01.46 % Severe	68.50 % Mild 23.28 % None 07.53 % Moderate 00.67 % Severe
Cartilage driven OA + Traumatic OA (25 patients)	52.00 % Mild 32.00 % None 16.00 % Moderate 00.00 % Severe	28.00 % Mild 36.00 % None 36.00 % Moderate 00.00 % Severe	58.00 % Mild 28.00 % None 20.00 % Moderate 00.00 % Severe
Bone driven OA + Cartilage driven OA + Traumatic OA (1169 patients)	70.57 % Mild 16.59 % None 11.97 % Moderate 00.85% Severe	53.97 % Mild 23.86 % None 19.84 % Moderate 02.30 % Severe	70.40 % Mild 15.73 % None 13.23 % Moderate 00.59 % Severe

Fig. 14: Distribution of OA patients based on overlaps of phenotypes.

Evaluation of prediction model on test dataset with instances having ethnicity
'MEXICAN, MEXICAN AMERICAN, OR CHICANO' and abstinence level
Either 'High' or 'Very High'

Correctly predicted instances : 63 out of 75   74%

a.

Plot:mexican_american_predicted

x-axis : abstinence level
y-axis : predicted abstinence level

Correctly predicted
'VeryHigh' outcomes

Correctly predicted
'High' outcomes

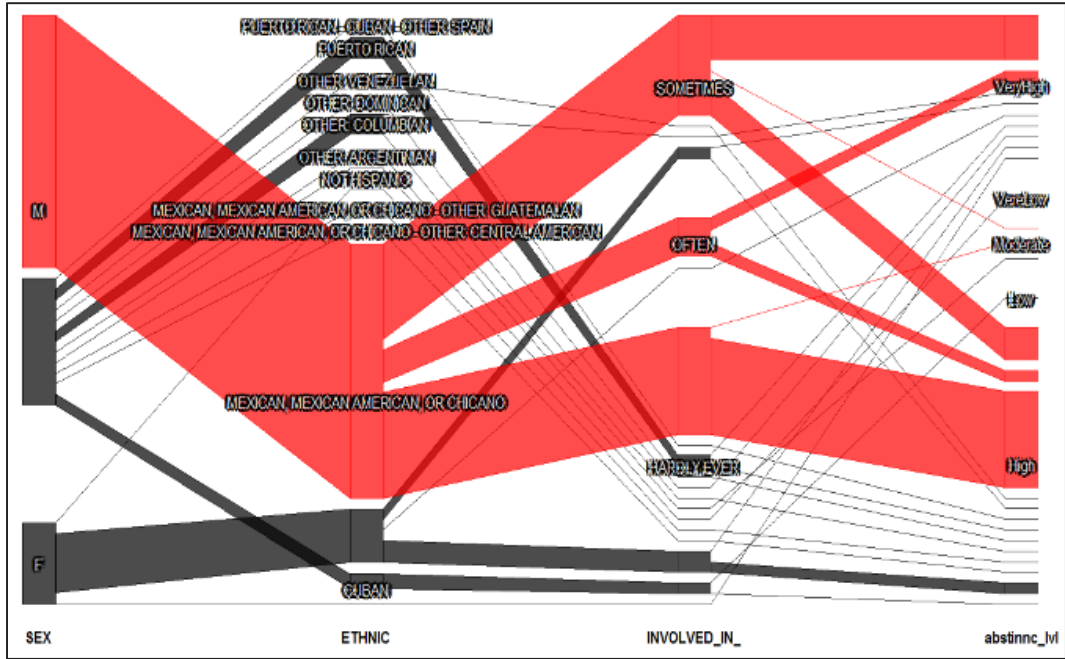Moderate   High   VeryHigh   Low   VereLow

b.

{ETHNIC=MEXICAN, MEXICAN AMERICAN, OR CHICANO} =>
{abstinence_level=High}
{ETHNIC=MEXICAN, MEXICAN AMERICAN, OR CHICANO} =>
{abstinence_level=VeryHigh}

c.

d.

Fig. 15: (a) classification results (b) Visualizing classification results (c) textual representation of association rules (d) visualizing cluster data with parallel co-ordinates
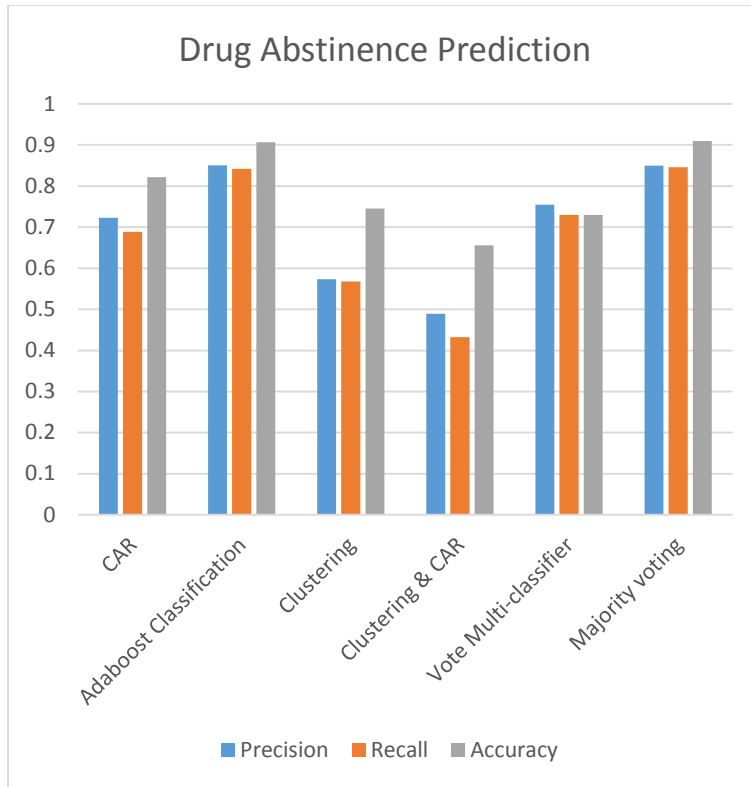
Fig. 16: Graph for drug abstinence prediction evaluation

Rule	Records
{ethnic=NOT HISPANIC,total_household_income=$10,000-$14,999,involved_in_family_activities=OFTEN,family_members_careful=ALMOST NEVER} => {abstinence_level=High}	8
{ethnic=MEXICAN, MEXICAN AMERICAN, OR CHICANO,family_type=BIOLOGICAL - 1 PARENT,total_household_income=$10,000-$14,999,involved_in_family_activities=SOMETIMES} => {abstinence_level=High}	3
{sex=F,ethnic=NOT HISPANIC,involved_in_family_activities=OFTEN,family_members_careful=SOMETIMES} => {abstinence_level=Moderate}	3
{sex=F,family_type=BLENDED,total_household_income=$50,000 OR MORE,involved_with_family=OFTEN} => {abstinence_level=High}	3
{ethnic=PUERTO RICAN,total_household_income=LESS THAN $5,000,involved_in_family_activities=OFTEN} => {abstinence_level=VeryHigh}	3

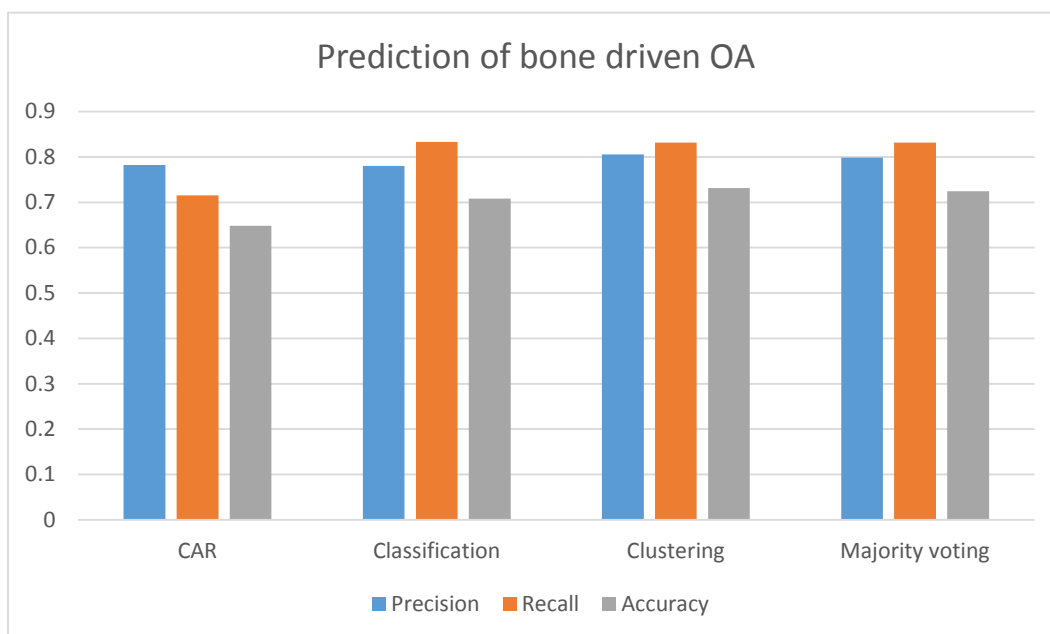**Fig. 17: Strong rules corresponding to absolute majority.**

Fig. 18: Graph for bone-driven OA prediction evaluation measures
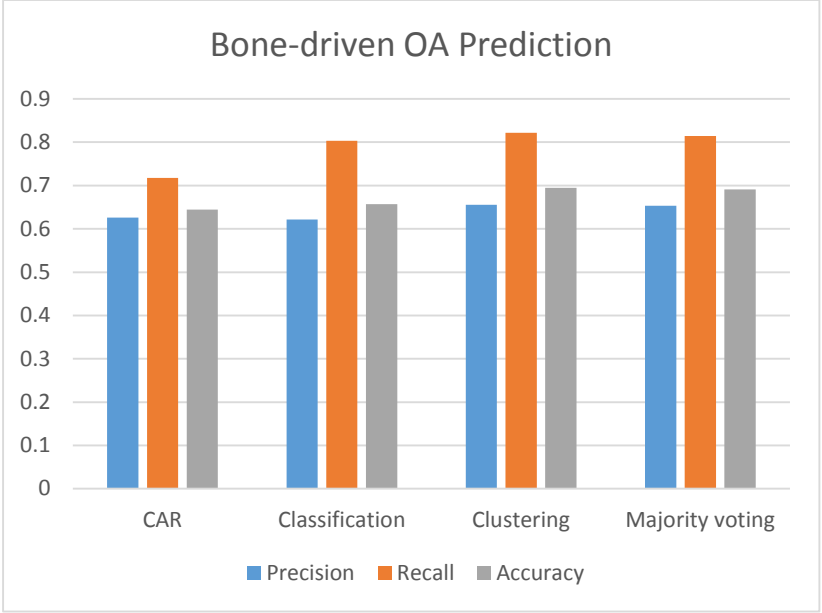
Fig. 19: Evaluation measures for prediction of bone driven OA with balanced training dataset

Rule	Records
{pain_score=MILD,disability_score=MILD} => {bone_driven_oa=YES}	527
{pain_score=NONE,stiffness_score=NONE,disability_score=NONE} => {bone_driven_oa=NO}	152
{stiffness_score=MODERATE,disability_score=MILD} => {bone_driven_oa=YES}	99
{stiffness_score=MILD,disability_score=MILD} => {bone_driven_oa=YES}	49
{pain_score=MODERATE,stiffness_score=MODERATE} => {bone_driven_oa=YES}	39
{pain_score=MILD,stiffness_score=MODERATE} => {bone_driven_oa=YES}	19

Fig. 20: Strong rules corresponding to absolute majority

Table 1. List of data mining techniques and related research in clinical domain

Clinical data mining technique	Research articles
Classification	[13], [14], [15]
Clustering	[16], [17], [18]
Association rule mining	[19], [20], [21]

Table 2. List of research articles on collaborative data mining and their applications

Research article	Contribution	Application domain
[24]	Multi-agent data mining framework to optimize prediction accuracy	Banking, Finance.
[25]	Multi-agent based intelligent intrusion detection system	Cybersecurity
[26]	Real-time decision support system for intense care unit using cooperative agents	Healthcare
[27]	Automated annotation of protein sequences	Bioinformatics

Table 3: Domain datasets in selected clinical trial study

Domain dataset file	Information	Description
SU.csv	Substance use	Subjects with substance/drug use details: alcohol, cocaine
SC.csv	Subject characteristics	Subject characteristics: Criminal background, educational background
DS.csv	Disposition	Whether subject was given BSFT or treatment as usual.
QS.csv	Questionnaires	Responses to specific questions designed as per study.

Table 4: Several OAI domain datasets that we have used for experiments

DOMAIN DATASET	DESCRIPTION
Joint Symptoms	Questionnaire results regarding arthritis symptoms in the knee, hip, back, and other joints; arthritis-related joint function and disability
Biomarkers	Readings from images, assay results, and metadata on biospecimens.
Medical history	Questionnaire data regarding a participant's arthritis-related and general health histories

Table 5: Clinical rules and corresponding features used for phenotype identification

Phenotype	Symptoms derived from domain experts	Mapping columns from OAI data	Source data file	Decision
Bone driven OA	Presence of osteophytes in left/right knee	P01SVLKOST,P01SVRKOST	biomarkers	Yes/ No
Cartilage driven OA	Left/right knee lateral or medial joint space narrowing	P01SVLKJSL,P01SVRKJSL, P01SVLKJSM,P01SVRKJSL	biomarkers	None/ Narro wed/D efinite
Traumatic OA	Past knee injuries, surgeries such arthroscopy, meniscectomy, knee replacements	P01ARTL, P01ARTR, P01INJL, P01INJR, P01KRSL, P01KRSR, P01KSURGL, P01KSURGR, P01LRL, P01LRR, P01MENL, P01MENR, P01ARTRINJ, P01ARTLINJ, P01MENRINJ, P01MENLINJ, P02KSURG	medical history	Yes/ No

Table 6: Discretized values of WOMAC scores.

WOMAC Score	Range	Discretized values	
Pain	0-20	[0-6.67]	Mild
		[6.67-13.34]	Moderate
		[13.34-20]	Severe
Stiffness	0-8	[0-2.67]	Mild
		[2.67-5.34]	Moderate
		[5.34-8]	Severe
Disability	0-68	[0-22.67]	Mild
		[22.67-45.34]	Moderate
		[45.34-68]	Severe

Table 7: Evaluations for drug abstinence prediction across multiple approaches

Method	Precision	Recall	Accuracy
CAR	0.722861	0.688172	0.822002
Adaboost Classification	0.850259	0.841996	0.906756
Clustering	0.573328	0.567568	0.745605
Clustering & CAR	0.489502	0.432432	0.656018
'Vote' Multi-classifier	0.755	0.73	0.72973
Majority voting	0.850024	0.846154	0.909747

Table 8: Evaluation on training set with co-occurrence of multiple phenotypes and no controlled cohort.

Method	Precision	Recall	Accuracy
CAR	0.504424	0.519869	0.607775
Classification	0.285455	0.534279	0.612345
Clustering	0.706851	0.796507	0.860067
Majority voting	0.578785	0.601965	0.676774

Table 9: Evaluation on training set with only patients having bone driven phenotypes and controlled cohort patient records without any knee OA.

Method	Precision	Recall	Accuracy
CAR	0.782565	0.715201	0.648
Classification	0.780446	0.833333	0.708
Clustering	0.805679	0.831502	0.731333
Majority voting	0.798593	0.831502	0.724667

Table10: Evaluation on training set with balanced class distribution

Method	Precision	Recall	Accuracy
CAR	0.625817	0.717228	0.644195
Classification	0.621739	0.803371	0.657303
Clustering	0.655224	0.822097	0.694757
Majority voting	0.653153	0.814607	0.691011