Collaborative Data Mining For Clinical Trial Analytics

Jay Gholap[1], Vandana P. Janeja[1], Yelena Yesha[2]
[1]Information Systems,
[2]Computer Science and Electrical Engineering,
University of Maryland, Baltimore County, USA
{jgholap1, vjaneja, yeyesha}@umbc.edu

Raghu Chintalapati[3], Harsh Marwaha[3],Kunal Modi[3]
[3]Ekagra Software Technologies,
{Raghu.Chintalapati, Harsh.Marwaha,
Kunal.Modi}@ekagra.com

Abstract- **This paper proposes a collaborative data mining technique to provide multi-level analysis from clinical trials data. Clinical trials for clinical research and drug development generate large amount of data. Due to dispersed nature of clinical trial data, it remains a challenge to harness this data for analytics. In this paper, we propose a novel method using master data management (MDM) for analyzing clinical trial data, scattered across multiple databases, through collaborative data mining. Our aim is to validate findings by collaboratively utilizing multiple data mining techniques such as classification, clustering, and association rule mining. We complement our results with the help of interactive visualizations. The paper also demonstrates use of data stratification for identifying disparities between various subgroups of clinical trial participants. Overall, our approach aims at extracting useful knowledge from clinical trial data in order to improve design of clinical trials by gaining confidence in the outcomes using multi-level analysis. We provide experimental results in drug abuse clinical trial data.**

Keywords--clinical trials, collaborative data mining, master data management, interactive visualization.

I. INTRODUCTION

Clinical trials conducted for clinical research are designed to improve medical knowledge of various diseases, drug reactions and treatment diagnosis. For the success of pharmaceutical companies, clinical trials are very important, since clinical trial data forms the basis for the approval and marketing of new drug or device they have developed. Data generated during clinical trials include information about study participants and their characteristics, interventions, adverse events and so on. This data is typically stored in variety of clinical information systems. In spite of remarkable progress in the area of data management and integration, disparate nature of clinical data makes it challenging to develop a true integration between heterogeneous data sources [1]. As pointed by Smith et al., maintaining data in disparate data stores leads to inconsistencies with data formats and values, making it difficult for organizations to identify the use of their crucial data [2]. Therefore, a central question we are trying to address is how we can tackle the challenge of using data mining methods on the scattered clinical trial data to extract knowledge that can be helpful to improve clinical trial designs.

As per the extensive review of the state of the art clinical data mining techniques by Iavindrasana et al [3], classification is the most commonly used data mining method for predictive modeling. Authors also suggest three major essentials of clinical data mining that include understanding of clinical data, assistance of domain experts and finally the implementation of data analysis methodology suitable for clinical data [3]. Another comprehensive review by Bellazi et al. states the importance of using Bayesian networks, decision trees for predictive data mining in clinical medicine. They also propose a few guidelines for constructing and assessing data mining models for clinical data [4]. We try to incorporate some of these guidelines proposed in the literature in our experiments and come up with a novel approach of collaborative data mining for analyzing clinical trial data. We believe that insights gained from mining old clinical trial data can be effectively utilized to address clinical trial design aspects such as: (a) managing participant enrollments, (b) planning clinical trial follow-ups, (c) associating demographics and medical conditions of patients with clinical trial outcomes, (d) guessing success of clinical trial and (e) identifying value of screening and interventions

Thus, appropriate clinical data mining is an essential companion to well-designed clinical trials. Assuming that clinical trials will produce data that could expose 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 insights 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

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

unwarranted assumptions. This is particularly applicable to small clinical trials [12].

We describe some of the key steps that we have implemented while attempting to build analytics solution with publicly available de-identified clinical trial data from National Institute on Drug Abuse (NIDA) data share repository. To build such clinical trial analytics solution, we propose a framework that combines data integration and data analytics with effective visualization.

Our main contributions include:

1. Integrating clinical trial data from multiple data sources with ETL driven MDM solution.
2. Applying collaborative data mining for extracting knowledge from clinical trial data to improve design of clinical trials.
3. Collating and validating results from multiple data mining techniques to support our hypotheses.

The structure of this paper is as follows:

In section II, we discuss the overall approach that integrates clinical trial data from several domain datasets. In Section III, we explain our experiments and results with specific details of tools and datasets. We also introduce use of interactive visualizations to support our analysis. Finally, in section IV we conclude with future directions for research.

II. METHODOLOGY

In this section, we present our methodology for mining clinical trial data. We propose a general framework, which can integrate data from multiple heterogeneous clinical data sources such as electronic health records, clinical notes, legacy health information systems and clinical domain datasets. This includes clinical data, subject data such as demographics, subject characteristics, substance use habits and their responses to questionnaires designed for clinical trials. For integrating dispersed clinical data into a single master data location, we propose use of master data management principles along with ETL (Extract-transform-load) concepts, which are traditionally being used for data warehouse implementation. We believe that integrated information of subjects enrolled for clinical study, with a collection of master data can give us a single point of reference, can be utilized for knowledge discovery. Applying data mining techniques on consolidated clinical trial data can give us interesting relationships between characteristics of enrolled participants and clinical trial outcomes.

Fig. 1 demonstrates our overall methodology. We use master data management techniques using ETL in order to integrate clinical trial data and then we execute data mining algorithms on the consolidated as well as stratified data and validate the results yielded from the multiple techniques. We next discuss the steps performed for clinical trial analytics.

A. *Master data management using ETL (Extract, Transform and Load)*

Master data management (MDM) consists of multiple techniques that can be used to incorporate data from various sources into a single location. Master data management for clinical data links identity data and reference data across multiple clinical sources into a single point of reference. That single point of reference could be a master record of a subject who is enrolled for clinical trials.

For building MDM solution for clinical trial data, we propose use of ETL techniques, which involve processes responsible for data extraction, transformation and load.

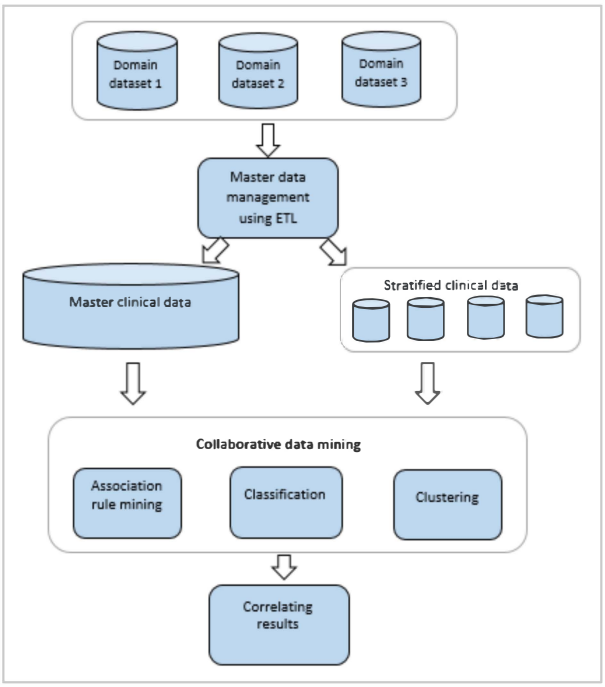


Figure 1: Overall Approach

With ETL process [14], 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 gender, ethnicity & family related information.

B. *Stratification*

Stratification is the process of sub-classifying data points in certain groups. In the context of clinical trial data, stratification is used to create different strata of subjects based on their background characteristics. Analyzing stratified clinical trial data allows us to detect quantitative trends, results in order to improve design of clinical trial. Stratifying data helps to discover existing clinical differences across different strata and then target interventions to address the disparities [5].

For clinical trial data, we select demographic attributes such as age, gender etc. to create subgroups of population.

On derived demographic strata, we execute data analytics algorithms in order to shed light on clinical trends.

C. Collaborative data mining

We adapt the concept of collaborative data mining to clinical trial data mining. In general, collaborative mining [11] is a technique where data mining is distributed to multiple collaborating agents to solve 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 first technique, users can eliminate weak patterns extracted from data and concentrate on validating patterns of interest with the help of other techniques. This is similar to ensemble learning, however here we use multiple different types of data mining techniques whereas ensembles use the same model with different datasets.

We use clustering [8], classification [7] and association rule mining [6] techniques in collaborative setting to produce and validate hypotheses from clinical trial data. We utilize proposed collaborative data mining toolkit on the master data derived with the help of ETL and validate similar outcomes with different data mining techniques working in partnership to solve a common problem. This way we harness the power of collaborative data mining methods with some amount of human involvement.

III. EXPERIMENTS & RESULTS

In our results, 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 trial study data. Further, we focus on insights gained from data mining experiments.

A. Background

We used R packages *arules* & *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 used R package called *cluster*. R provides a variety of data mining packages for advanced algorithms. Some of the packages that we propose for interactive data mining include *klaR*, *RWeka*, *extracat.* For data integration purposes, we have used Talend open studio 5.4 as our ETL toolkit [13].

National institute of drug abuse (NIDA) data share repository has datasets for completed clinical trials. Dataset for each study includes several domain datasets as specified by CDISC (Clinical Data Interchange Standards Consortium) submission format. These datasets include information about substance use, demographics & subject characteristics of enrolled participants; questionnaires designed for specific clinical trials and recorded responses during clinical trials. These datasets are publicly available to promote new research and further analyses.

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. It is important to note that our findings here are simply to validate our data mining approach and do not have clinical implications. These results should not be used to make medical conclusions for these particular trials. Following table demonstrates various domain datasets, which were utilized for our analysis.

TABLE I: DOMAIN DATASETS IN SELECTED CLINICAL TRIAL

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.

B. MDM solution using ETL

MDM schematic demonstrated in Fig. 2 was developed and translated to code with Talend data integration studio [11].

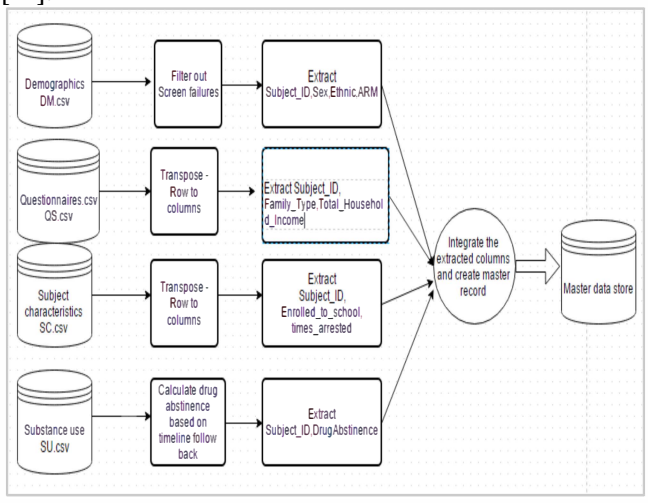


Figure 2: MDM approach

Fig. 3(b) displays attributes that were obtained in a single file with the help of data integration job. 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.

For calculating drug abstinence, we used percentage of days when subject showed abstinence to drugs such as alcohol, marijuana, tobacco etc. Different subject characteristics were stored in separate rows for each subject. As depicted in Fig. 3(a), row to column transformation was used in order to derive characteristics and their values in columnar format. 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 filtered out such records since other files did not include associated data for these participants.

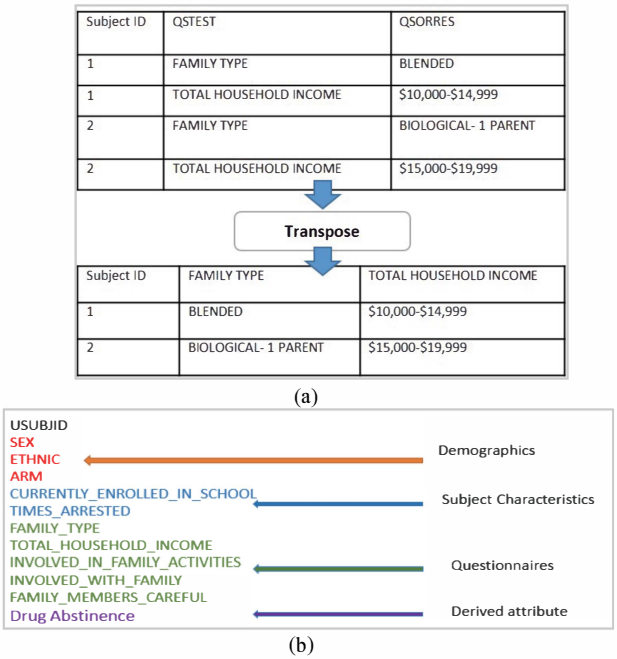


Figure 3: (a) Row to column transofrmation for questionnaire dataset (b) List of attributes in clinical master data.

C. 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 try to do collaborative analysis to find correlations in the results.

1) Exploratory analysis of association rules: We used Apriori algorithm to discover association rules from the integrated clinical trial data. Drug abstinence, being the primary outcome of selected clinical trial study, we performed selective filtering on the derived association rules in order to analyze rules which include drug abstinence label on the RHS.

{SEX=M, ETHNIC=OTHER: BOLIVIAN, FAMILY_TYPE=BIOLOGICAL-2PARENTS} => {abstinence_level=VeryHigh}

{ETHNIC=CUBAN, TOTAL_HOUSEHOLD_INCOME=$30,000-$34,999, INVOLVED_IN_FAMILY_ACTIVITIES = HARDLY EVER} => {abstinence_level=Low}

Figure 4: Sample association rules derived from clinical data

From derived association rules, we observed several interesting patterns that indicate certain combinations of demographic attributes with specific family background showing higher or lower drug abstinence. For example, first association rule in Fig. 4 suggests that male adolescents with Bolivian ethnicity and biological family with 2 parents exhibit a very high drug abstinence. Similarly, second rule indicates that Cuban adolescents with very rare involvement in family activities and total household income in the range of $30,000 to $35000 show a poor drug abstinence.

Further, we plot interactive visualization of derived association rules with the help of *arulesviz* R-package as proposed by Hahsler et al. [9]. We believe that graph based visualization of association rules in Fig. 5 is much easier to interpret and analyze the relationships between various attributes as compared to textual representation in Fig. 4.

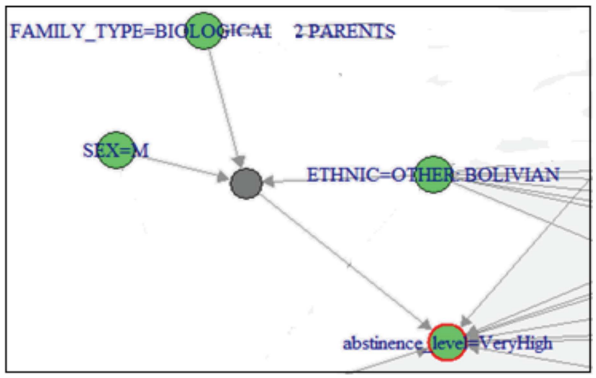


Figure 5: Interactive visualization of associ*ation rule.*

In Fig. 5, we have demonstrated a visual representation of the same association rule discussed in the Fig. 4. Incoming edges to the grey colored rule node indicate the LHS of the association rule while outgoing edge represents RHS. Color indicates important measure of association rule i.e. lift which specifies how item-sets from the LHS and RHS are correlated. For our analysis, we choose association rules with higher values of lift and confidence.

2) Classification: Before applying classification algorithms on clinical trial data, we detected imbalanced distribution of drug abstinence label. In order to build

unbiased prediction model, we eliminated class imbalance problem by sampling instances with replacement to obtain uniform distribution of class labels. Interestingly, there was an increase in prediction accuracy of C4.5 decision tree algorithm from 38.87% to 80.56% after resolving this problem. To improve predictive performance of model further, we used boosting method which uses ensemble of classifiers for the classification. Adaboost with base classifier as C4.5 decision tree, eventually increased classification accuracy to 86.17% with 10-fold cross-validation. This model was then evaluated on random subsample of 95 instances with uniform distribution of class labels. Table II displays the prediction results of our experiments.

We believe that prediction of drug abstinence can be very helpful to improve treatment of substance abuse in real clinical settings. This approach can generally be practiced to predict the outcomes of clinical trials. This would eventually help in designing clinical trials or changing existing clinical trial design.

TABLE II: CONFUSION MATRIX SHOWING PREDICTION RESULTS

Moderate	High	VeryHigh	Low	VeryLow	←Predicted/ Actual:
15	3	1	0	0	Moderate
1	16	2	0	0	High
0	4	15	0	0	VeryHigh
0	0	0	19	0	Low
0	0	0	0	19	VeryLow

3) Clustering clinical trial data: As clinical trial data merged with our MDM solution consisted categorical variables, we used K-modes clustering algorithm with K=5, in order to derive clusters of similar instances. The notion behind using clustering algorithm, was to study subjects of clinical trials clustered together to find out if there are any overlaps between their characteristics. We carefully evaluated a few clusters and noticed that one of the clusters had data instances which exhibited similar characteristics. Subjects clustered in this particular cluster showed similar demographics. Most of them were females with 'Not Hispanic' ethnicity. Also, majority of these subjects were enrolled to the school and showed high drug abstinence. Similarly, other cluster had majority of male subjects with 'Mexican American' ethnicity with rare involvement in family activities but higher drug abstinence.

We analyzed intra-cluster data instances with the help of *extracat* R package which provides visualization for categorical data in the form of parallel co-ordinates [10]. Visualization exactly supported our findings gained through manual investigation of data instances. As we outlined above, Fig. 6 indicates the similar findings indicating the majority of Mexican-American male subjects with rare involvement in family activities with tendency towards high drug abstinence.

In this case, we have performed supervised clustering as we include class label i.e. drug abstinence level while clustering the data. We strongly support an argument made by Färber et al. [18] that different or alternate cluster groupings in one data set may be possible, hence we have not evaluated clusters with extrinsic methods such as precision & recall that heavily rely on class labels for evaluation. We evaluate clusters mainly by manual inspection as suggested by Färber et al. [18].

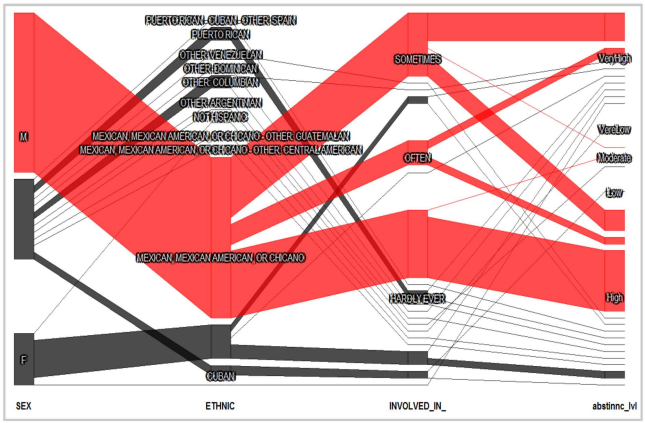


Figure 6: Visualization of intra-cluster categorical data.

D. Mining stratified clinical trial data

Demographic stratification of clinical data proves to be useful tool to understand data trends for particular population of study participants. With respect to our case study, we identified a few demographic attributes from the given set of attributes such as gender, ethnicity in order to create subgroups of clinical trial participants. Application of data mining technique such as association rule mining on individual demographic stratum itself provided interesting findings about particular group of participants. We executed simple Apriori algorithm on various strata such as male participants & female participants to understand several associations between drug abstinence and other characteristics of these adolescents. We further stratified these groups based on ethnicity.

A couple of association rules displayed in Table III imply that male adolescents with blended family and a good involvement with family starting from baseline of trial would probably demonstrate very high drug abstinence if they are treated with strategic family therapy. On the other side, those with a single biological parent and are not enrolled to school would exhibit low drug abstinence in spite of considerable involvement in family. Table IV shows number of records matching to association rules in Table III and number of trial outcomes correctly predicted by

ensemble based prediction model. All the outcomes in the classification models were in agreement with association rules.

TABLE III: OUTCOME BASED ASSOCIATION RULES DERIVED FROM MALES' STRATUM

Rule #	Association rule	Support	Confidence	Lift
1	{ARM=BRIEF STRATEGIC FAMILY THERAPY, FAMILY_TYPE=BLENDED, INVOLVED_WITH_FAMILY=OFTEN} => {abstinence_level=VeryHigh}	0.015	0.85	3.17
2	{CURRENTLY_ENROLLED_IN_SCHOOL=NO, FAMILY_TYPE=BIOLOGICAL - 1 PARENT, TOTAL_HOUSEHOLD_INCOME=$15,000-$19,999, INVOLVED_WITH_FAMILY=OFTEN} => {abstinence_level=Low}	0.005	1.00	16.43

TABLE IV: COLLABORATIVE ANALYSIS OF ASSOCIATION RULES BY PREDICTION MODEL

Association Rule #	Matching records	Correctly Predicted Outcomes
1	6	6
2	2	2

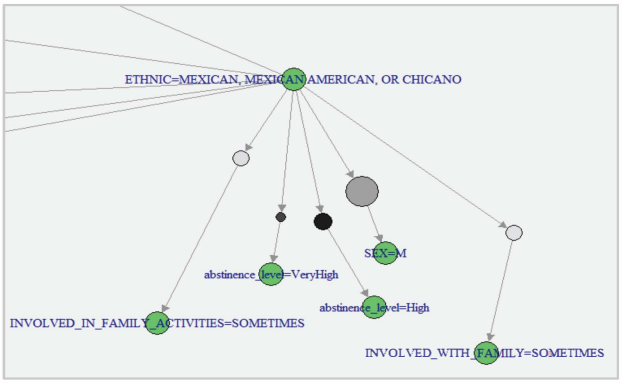
These types of collaborative validations of clinical trial outcomes based on association analysis and classification techniques may help to identify sub-cohorts of trial participants with clinical similarities, hence data stratification coupled with collaborative data mining is really powerful technique and hence can significantly expose findings about trial participants.

E. Observations & Discussion

With the application of multiple data mining techniques on the same clinical trial data, we made an attempt to associate the results and relate our findings. After careful observation, we noticed that these techniques produce matching outcomes. For example, by clustering clinical trial data, we obtained a separate cluster of male, Mexican American adolescent subjects with not-much involvement in family activities, but show high drug abstinence as verified by figure 7. We tried to investigate association rules derived from the same data. Interestingly, we found comparable rules showing higher drug abstinence associated with such population of participants. 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 predicted 50 instances out of 60 showing 83.33% of accuracy. These results are presented in Fig. 7(a), 7(b) and 7(c).

(a)

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



(b)

Test data characteristics	Precision	Recall	Correct Predictions
Male, Mexican American with 'High', 'Very High' abstinence	0.833	0.833	50 out of 60
Female, Mexican American with 'High', 'Very High' abstinence	0.9	0.875	14 out of 16

(c)

Figure 7: Validating collaborative results

F. Ground truth evaluation

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*" [15]. Most research on substance abuse among Hispanics is focused on alcohol abstinence and has confirmed that Hispanics show higher alcohol abstinence irrespective of their national origin [16], [17].

IV. CONCLUSION & FUTURE WORK

In this paper, we have proposed a general framework for analyzing clinical trial data using a novel approach of collaborative data mining. Using drug abstinence clinical trial as our case study, we have observed that multiple data mining techniques yield similar results. Results are consistent & data-driven in spite of using various techniques and can be validated with the help of several confirmatory analyses. Further, we conclude that stratification of clinical trial data proves to be useful in order to discover trends among several groups of population of clinical trial. Mining data within demographic strata can expose driving characteristics of trial participants related clinical trial outcomes. We also observe that interactive visualization of data can play a crucial role to further supplement our findings. With several techniques yielding similar implications, we can generate interesting hypotheses that can help maximize effectiveness clinical trials. Our

proposed approach can be insightful for subject enrollment and planning trial follow-ups.

In future, we plan to utilize proposed framework for mining for observational as well as clinical study data. We would like to identify and validate sub-cohorts of clinical trial participants having similar clinical characteristics with the help of data mining techniques discussed in this paper. This type of sub-cohort discovery would ultimately lead to increase in the effectiveness of treatments and interventions during clinical trials.

ACKNOWLEDGEMENTS

The information reported here results from secondary analyses of data from clinical trials conducted by the National Institute on Drug Abuse (NIDA). Specifically, data from NIDA–CTN-0014 'Brief Strategic Family Therapy for Adolescent Drug Abusers' (actual protocol number and title) were included. NIDA databases and information are available at http://datashare.nida.nih.gov.

REFERENCES

[1] "Managing and Integrating Clinical Trial Data: A Challenge for Pharma and Their CRO Partners." Http://www.bio-itworld.com/. LIAISON Healthcare Informatics. Web. 11 July 2015.

[2] H.A. Smith, and J.D. McKeen: "Developments in Practice XXX: Master Data Management: Salvation Or Snake Oil?", Communications of the Association for Information Systems, Vol. 23, 2008, pp. 63-72.

[3] J. Iavindrasana, G. Cohen, A. Depeursinge, H. Muller, R. Meyer, and A. Geissbuhler, "Clinical data mining: a review," Yearb Med Inform, vol.2009, pp. 121–133, 2009.

[4] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," international journal of medical informatics, vol. 77, no. 2, pp. 81–97, 2008.

[5]Health Research & Educational Trust. (2014, October). A framework for stratifying race, ethnicity and language data. Chicago, IL: Health Research & Educational Trust. Accessed at www.hpoe.org

[6] Kuo, M. H., Kushniruk, A. W., Borycki, E. M., & Greig, D. (2008). Application of the Apriori algorithm for adverse drug reaction detection. Studies in health technology and informatics, 148, 95-101.

[7] Chawla, Nitesh V. "C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure." Proceedings of the ICML. Vol. 3. 2003.

[8] Huang, Joshua Zhexue. "Clustering Categorical Data with k-Modes." (2009): 246-250.

[9] Hahsler, Michael, and Sudheer Chelluboina. "Visualizing association rules: Introduction to the R-extension package arulesViz." R project module (2011): 223-238.

[10] Pilhoefer, Alexander, and Antony Unwin. "New approaches in visualization of categorical data: R-package extracat." Journal of Statistical Software (2011).

[11] Moyle, Steve. "Collaborative data mining." Data Mining and Knowledge Discovery Handbook. Springer US, 2010.1029-1039.

[12] Ildstad, Suzanne T., and Charles Hawes Evans. Small clinical trials: issues and challenges. National Academy Press, 2001.

[13]Talend Open Studio, https://www.talend.com/products/talend-open-studio

[14] "Master Data Management, an Oracle White Paper." Oracle, 1 Sept. 2011. Web. 4 Oct. 2015. <http://www.oracle.com/us/products/applications/master-data-management/018876.pdf>.

[15]"Alcohol and the Hispanic Community." NIAAA, Web. 1 Oct. 2015. <http://pubs.niaaa.nih.gov/publications/HispanicFact/HispanicFact.htm>.

[16] Ames G, Mora J. Alcohol problem prevention in Mexican American populations. In: Gilbert MJ, editor. Alcohol Consumption Among Mexicans and Mexican Americans: A Binational Perspective. Los Angeles: University of California; 1988. pp. 253–281.

[17] Canino G. Alcohol use and misuse among Hispanic women: Selected factors, processes, and studies. International Journal of the Addictions. 1994;29(9):1083–1100.

[18] Färber, Ines, et al. "On using class-labels in evaluation of clusterings."MultiClust: 1st international workshop on discovering, summarizing and using multiple clusterings held in conjunction with KDD. 2010.