Drug Exposure Side Effects from Mining Pregnancy Data

Yu Chen[1] Lars Henning Pedersen[2] Wesley W. Chu[1] Jorn Olsen[2, 3]

[1]Department of Computer Science
University of California, Los Angeles
{chenyu, wwc}@cs.ucla.edu

[2]Danish Epidemiological Science Center
University of Aarhus, Denmark
lhp@soci.au.dk

[3]Department of Epidemiology
University of California, Los Angeles
jo@ucla.edu

ABSTRACT

This paper presents an interdisciplinary collaborative research project between the Epidemiology Department and the Computer Science Department for using data mining technique to analyze data from pregnant women. Specifically, the authors use association rule mining approach to derive possible side effects due to exposure to multiple drugs at different duration of the pregnancy. The derived temporal sequential rules discover new information that would not be detected by the traditional analysis method that is currently used in pharmaco-epidemiology.

Keywords

Association rule mining, epidemiology data analysis, side effect of drug exposure.

1. INTRODUCTION

More than half of all pregnant women use some sort of medication during pregnancy [1]. The exposure varies from accidental use of known teratogens such as retinoic acid to planned use of assumed safe drugs, e.g. acetaminophen. However, the fetus is potentially more vulnerable partly due to incomplete metabolic pathways that normally detoxify drugs and other chemicals in adults or children. Furthermore, the complexity of fetal biology makes it very difficult to predict what kind of side effects certain drugs might possibly have in the developing organism [2].

The knowledge about the safety of the various types of drugs is from different sources, initially from animal research, as all drugs released to the market have to be tested on experimental animals. However, the results cannot be easily extrapolated from animals to humans, and the main information on the safety in humans must be derived from observational studies, including post-marketing surveillance and register based studies. The major problem in the observational studies is to recognize the causal relationship between drug exposure and its effect in these data sources. As an example, thalidomide, one of the most well known human teratogens, may cause specific malformation in one third of infants exposed in first trimester. Despite that, thousands of malformed babies were born before the relationship was discovered illustrating the problems in not using available information before someone comes up with a specific hypothesis. In the case of thalidomide, the signal was strong enough to alert a clinician. Most associations are expected to be much weaker and occur over a longer time span, therefore will be even harder to identify.

The traditional approach in epidemiology has been to use a deductive approach for data analysis and the results have been that large data files remain un-analyzed for years. Setting up a deductive hypothesis usually means waiting for reports on side effects to be published that can activate research findings. Then follows testing of specific associations using existing data or by generating new data sources; this usually takes at least months and more often years. We need a screening tool to use on available data sources in order to identify drugs or combinations of drugs that deserve further scrutiny. We need inductive methods because in principle, only a very small number of cause-effect mechanisms can be ruled out using biological reasoning alone.

In the light of the amount of research invested in developing new drugs, it is truly surprising how little time, money and resources are invested in finding out how the drugs work in normal clinical practice.

There are many technical reasons that pregnancy data is difficult for traditional data analysis methods. For example: 1) S*ubtle side effect*: In pregnancy data, only a very small number of cases may reflect the subtle influences of drug exposure. For example, among a large number of patients who took a particular type of drug, only 1% may be a susceptible preterm birth. However, this 1% is still significant side effect of the drug to discover. 2) *Temporal sensitive*—Timing is an important aspect, as the susceptibility of the fetus varies in the developmental processes. Thalidomide was linked to serious malformations of the limbs; however, these specific malformations only resulted from exposure in the first eight weeks after conception. With a different timing, other malformations may have resulted. 3) *Data sequences*—The sequence of taking different types of drugs is related to the timing, but also involves the possible interaction between the different drugs. Almost nothing is known about the potential time-dependent interaction between drugs in pregnancy, partly due to the lack of analytical screening tools.

The aspects described above, i.e. the problem with the low number of cases, the timing and sequence of the use of drugs motivates us to apply data mining techniques for a large nationwide dataset. Our goal is to use the mining results to provide an early warning for drugs that is harmful to the unborn baby.

2. DATA MINING METHODOLOGY

Data mining may be an alternative that can discover more patterns of drugs and health effects than that can be scrutinized using more traditional techniques. Since the pregnancy drug exposure data is usually presented in tabular format, we develop a mining technique SmartRule that derives association rules directly from tabular data [5]. The generated association results can efficiently represent the side effects (such as preterm birth or malformation) of taking a certain type of drug during pregnancy. In this section, we will first introduce the design and features of SmartRule technique and then discuss how to apply and extend it to discover the temporal sequential patterns in pregnancy data.

2.1 SmartRule Association Rule Mining

Different from many other association rule mining algorithms, SmartRule is specially designed for mining tabular data, such as spreadsheets. SmartRule stands out for mining the pregnancy data based on the following features.

- SmartRule can generate Maximum Frequent Itemsets (MFIs) directly from tabular data, eliminating the need for conversion to transaction-type datasets. In addition, by taking advantage of column structures in tabular data, this method can significantly reduce the search space and the support counting time.
- SmartRule uses subset of MFIs for targeted rule mining. A user can select a subset of MFIs to include certain attributes known as targets (e.g., drug safety outcomes) in rule generation. Therefore, domain experts can filter out those uninterested associations and only generate rules for the targets.
- To handle the large number of rules generated, SmartRule hierarchically organize rules into trees and use spreadsheet to present the rule trees. In a rule tree, general rules can be extended into more specific rules. A user can first exam the general rules and then extend to specific rules in the same tree. Based on spreadsheet's rich functionality, domain experts can easily filter or extend branches in the rule trees to create the best view for their interest.

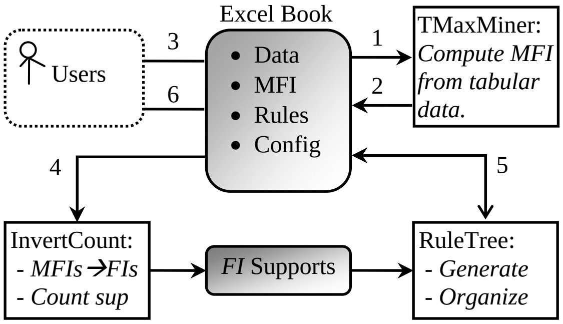


Figure 1. System overview of SmartRule

SmartRule uses an Excel book (or other spreadsheet software) to store both source data and mining results. As shown in Figure 1, the process of generating association rules consists of the following steps. First, the tabular dataset is fed to the TMaxMiner module to directly compute Maximum Frequent Itemsets (MFIs) for a user-chosen minimum support. Once the MFIs are generated, the TMaxMiner outputs all MFIs back to the Excel book. Then the user can select a subset of MFIs containing the targeted attributes and InvertCount module builds the Frequent Itemset (FI) list for the user-selected MFIs and counts the supports for the FIs. The RuleTree module can use these FIs and their support to generate association rules and organize them into hierarchical trees. Finally, the rules are written back to the Excel book to present to the users.

2.2 Derive Drug Side Effects via SmartRule

The SmartRule mining technique can be applied to the pregnancy data to handle the problems mentioned in Section 1.

SmartRule can discover the small number of cases that reflects the subtle influence of drug exposure. Different from traditional analysis methods, this mining approach can generate all possible rules with a very low support and confidence level. However, such low support or low confidence rules could still be significant because of their contrast to normal pregnant woman.

The derived rules can include the relationship between timing of drug exposure and the safety outcome. Our proposed approach bypasses the problem of a large combination of drug exposure sequences by dividing the pregnancy period into time slots. The drug exposure information is represented for each time slot based on patient pharmacy record. By treating each drug exposure in a certain time slot as a single independent attribute, the rules generated contain both drug type and timing information. Such a method is very flexible in terms of timing slot division. The user can control the granularity of time sequences to study specific effects — for example, a drug taken in each trimester for a big picture, or drug exposure in every week for finer granularity rules.

2.3 Computation Complexity and Scalability of SmartRule

Since the SmartRule MFI mining algorithm does not require superset checking and reduces the computation for counting support, it is very efficient in mining rules. To further improve the performance of mining MFI, we use a technique to gather past tail information to determine the next node to explore during the mining process. Our experimental results in [6, 7] reveal that it is significantly faster than that of Mafia [8] and GenMax [9]. Compared to the recent frequent itemset mining implementations in FIMI repository [12], SmartRule is still reasonably efficient in mining MFIs for generating association rules. When the dataset exceeds the spreadsheet size limit (for example, the current Microsoft Excel spreadsheet size is 65,536 rows in one spreadsheet), we can partition the dataset into multiple groups of the maximum spreadsheet size to derive MFIs for each spreadsheet, and then join these MFIs for generating association rules. Therefore the data mining method is scalable for large datasets that contains millions of records.

3. MINING DANISH NATIONAL BIRTH COHORT DATASET

The large-scale Danish National Birth Cohort (DNBC) study's focus is to describe many aspects of pregnancy. The DNBC data collection is nationwide in Denmark. It started in 1996 and is ongoing. The inclusion of pregnant women stopped in 2002, and approximately 100,000 women were recruited. Exposure information and later developmental information has mainly been collected by telephone interview. Women were recruited through their general practitioner. Following consent, the women were contacted twice during pregnancy at 17 and 32 gestational weeks.

In the DNBC dataset, each patient's exposure status, possible confounders, and endpoint are registered. Exposure status will be drug exposure according to category of drug, timing, and sequence of different drugs. Possible confounders include vitamin intake, smoking, alcohol consumption, socio-economic status and psycho-social stress. Endpoints will be preterm birth, malformations and prenatal complications, e.g., low Apgar score (a score to summarily assess the health of newborn children immediately after childbirth based on Appearance, Pulse, Grimace, Activity, and Respiration) and low birth weight.

Since depression during pregnancy affects approximately 10 % of pregnant women in the United States, knowledge about potential side effects on the fetus is important in proper evidence based treatment of the disease during pregnancy [3, 4]. We have applied the SmartRule data mining technique to investigate part of the data from the DNBC dataset. Although it is part of the large DNBC cohort study, this sub-dataset focuses particularly on investigating the safety of different treatments of depression in pregnancy. The structure of this sub-dataset is identical to the large dataset on all births and the use of drugs in the nationwide register, but in a smaller scale.

3.1 Side Effect of Antidepressant

The data in this study includes multiple types of antidepressants (the four major SSRIs (Selective Serotonin Reuptake Inhibitors) fluoxetine, citalopram, paroxetine, sertraline, and various tricyclic antidepressants (e.g. clomipramine)) and a few confounders such as alcohol and smoking. Our goal is to study how these antidepressants and confounders influence the preterm birth of the new-born. To represent the time of drug exposure, we divide the pregnancy period into three trimesters and record the exposure in each trimester. There are about 4454 pregnant patients in this sub-dataset, which comprised women with psychiatric disease and a control group of women with no psychiatric disease. Among these patients, approximately 1000 women were depressed and/ or exposed to various active drugs used for depression or co-morbid conditions, e.g. antidepressants, with variation in timing and sequence. With the SmartRule data mining technique, we can generate a large number of rules in terms of many aspects of this sub-dataset. Due to the prevalence of drug side effects, the confidence level of generated rules is usually low. However, in pharmaco-epidemiological studies, drug side effects may be important even at low confidence levels if the nature of the effect is serious enough and the association is causal. In the following examples, we only demonstrate those rules that are related with the exposure to citalopram (one of many specific antidepressants) during pregnancy to show its effect on preterm birth.

In the rules generated, we first confirmed that exposure to citalopram (”*cita*”) significantly increases the risk of preterm birth as shown in the following. These rules match the results derived from traditional logistic regression analysis.

1) In general, patients have preterm birth with support=0.0454, and confidence=0.0454

2) If patients exposed to *cita* in the 1st trimester, then have preterm birth with support=0.0016, confidence=0.0761

3) If patients exposed to *cita* in the 2nd trimester, then have preterm birth with support=0.0013, confidence=0.1714

4) If patients exposed to *cita* in the 3rd trimester, then have preterm birth with support=0.0011, confidence=0.1786

5) If patients not exposured to *cita*, then have preterm birth with support=0.0433, confidence=0.0444

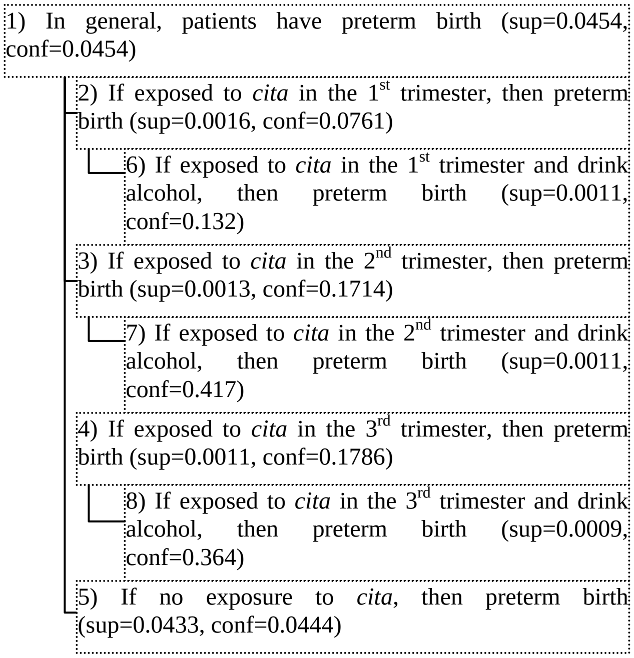
The rules suggest that exposure in the later part of pregnancy may be associated with a higher risk of preterm birth, as described in earlier epidemiological studies. Further, as shown in the following rules, we found that for all three trimesters of citalopram exposure, “alcohol” is a factor that combines with the drug exposure to cause preterm:

6) If patients exposed to *cita* in the 1st trimester and drink alcohol, then have preterm birth with support=0.0011 and confidence=0.132

7) If patients exposed to cita in the 2nd trimester and drink alcohol, then have preterm birth with support=0.0011 and confidence=0.417

8) If patients exposed to cita in the 3rd trimester and drink alcohol, then have preterm birth with support=0.0009 and confidence=0.364

The confidence levels suggest that combined exposure to citalopram and alcohol in pregnancy may be associated with an increased risk of preterm birth. Further, as we compare these three rules with other combinations of attributes with cita exposure, we find that among all single attributes, alcohol is the most important modifying factor in the association between citalopram exposure and preterm birth. Such findings were not initially discovered by our co-authors in epidemiology study, but were later confirmed by a traditional interaction analysis. However, because of the large number of combinations among all the attributes and their values, traditional methods cannot support the testing of all the potential interactions to locate the significant ones.



1) In general, patients have preterm birth (sup=0.0454, conf=0.0454)

2) If exposed to *cita* in the 1st trimester, then preterm birth (sup=0.0016, conf=0.0761)

6) If exposed to *cita* in the 1st trimester and drink alcohol, then preterm birth (sup=0.0011, conf=0.132)

3) If exposed to *cita* in the 2nd trimester, then preterm birth (sup=0.0013, conf=0.1714)

7) If exposed to *cita* in the 2nd trimester and drink alcohol, then preterm birth (sup=0.0011, conf=0.417)

4) If exposed to *cita* in the 3rd trimester, then preterm birth (sup=0.0011, conf=0.1786)

8) If exposed to *cita* in the 3rd trimester and drink alcohol, then preterm birth (sup=0.0009, conf=0.364)

5) If no exposure to *cita*, then preterm birth (sup=0.0433, conf=0.0444)

Figure 2. A part of the rule hierarchy for the exposure to the antidepressant *citalopram* and alcohol at different time period of pregnancy with preterm birth

Using our proposed temporal data mining method, we are able to derive associations between drug exposure in different periods and the preterm outcome. The rules suggest an association between exposure to citalopram in late pregnancy and preterm birth in accordance with earlier studies [3, 4]. Additionally, we found a negative interaction with use of alcohol in this dataset.

The generated rule hierarchy (Figure 2) provides an organized way to discover useful information.

3.2 Side Effect of Multiple Types of Drug

Treatment with antidepressants may be combined with other neuroleptic treatment including sedatives and antipsychotic medication, or the pregnant women may suffer from other diseases that mandate treatment, e.g. migraine. Therefore, it is important to reveal: 1) if exposure to these treatments and drugs causes preterm birth and which time period is the most vulnerable; 2) the confounders that can significantly increase preterm rate with particular type of drug exposure; 3) the interaction between different types of drugs and different time period of the same drug. Almost no human data exist on the possible pharmacological interaction between the types and sequences of medication investigated in this study [1].

With the above goals, we apply SmartRule mining technique to generate rule based on a sub-dataset that contains four types of drugs (antidepressant, migraine medication, sedative medicine and antipsychotic medication) and three confounder information (alcohol, tobacco and symptoms of depression). For each type of the drugs, the exposure time was divided into three time periods: exposure before conception ("*pre*"), exposure during the period of the main development of the organs (*"in"*) or exposure after the this period ("*post*"). This sub-dataset represents 6231 patients, in which 414 patients (6.64%) experienced preterm birth.

First of all, we investigate the rules to prove that exposure in the late stage of pregnancy is the most dangerous time for causing preterm. For example, rules for sedative medicine ("*Anxio*") demonstrate the effect:

1) If exposed to Anxio in *post* stage, then have preterm birth with support=0.17%, confidence=0.129
2) If exposed to Anxio in *in* stage, then have preterm birth with support=0.14%, confidence=0.103
3) If exposed to Anxio in *pre* stage, then have preterm birth with support=0.09%, confidence=0.097

And antidepressant ("*Ad*") rules prove it too:

4) If exposed to Ad in *post* stage, then have preterm birth with support=0.35%, confidence=0.116
5) If exposed to Ad in *in* stage, then have preterm birth with support=0.56%, confidence=0.096
6) If exposed to Ad in *pre* stage, then have preterm birth with support=0.56%, confidence=0.097

Note that the confidence of the *post* rule is higher than the *in* and *pre* rules, which means for the same type of drug, the exposure in *post* time is more likely to cause preterm birth than the earlier time. As seen with citalopram, these rules suggest that exposure in the later part of pregnancy appears to have the strongest association with preterm birth. From a biological point of view the association seems plausible; however, it may be related to other factors not included in the analyses. Despite these cautions the rules may be very important in hypotheses generating research and are difficult to extract from traditional methods.

Next, we investigate the effect of confounders that could increase the preterm rate with drug exposure. Our results show that tobacco is the most significant confounder for drug *Anxio* at all times. For example, in the *pre* stage,

7) If exposed to Anxio in *pre* stage and take tobacco, then have preterm birth with support=0.08%, confidence=0.143
8) If exposed to Anxio in *pre* stage and drink alcohol, then have preterm birth with support=0.03%, confidence=0.077
9) If exposed to Anxio in *pre* stage and have symptoms of depression, then have preterm birth with support=0.05%, confidence=0.13

Similarly, in the *in* and *post* stage of pregnancy, tobacco is more significant than alcohol and depression symptoms.

10) If exposed to Anxio in *in* stage and take tobacco, then have preterm birth with support=0.13%, confidence=0.174
11) If exposed to Anxio in *post* stage and take tobacco, then have preterm birth with support=0.11%, confidence=0.149

Our investigation also shows that some combination of multiple confounders could cause a higher preterm rate than single confounder, but some other combination could lower the preterm rate. For example,

12) If exposed to Anxio in *pre* stage and take tobacco and have symptoms of depression, then have preterm birth with support=0.05%, confidence=0.231
13) If exposed to Anxio in *pre* stage and take tobacco and alcohol, then have preterm birth with support=0.03%, confidence=0.133

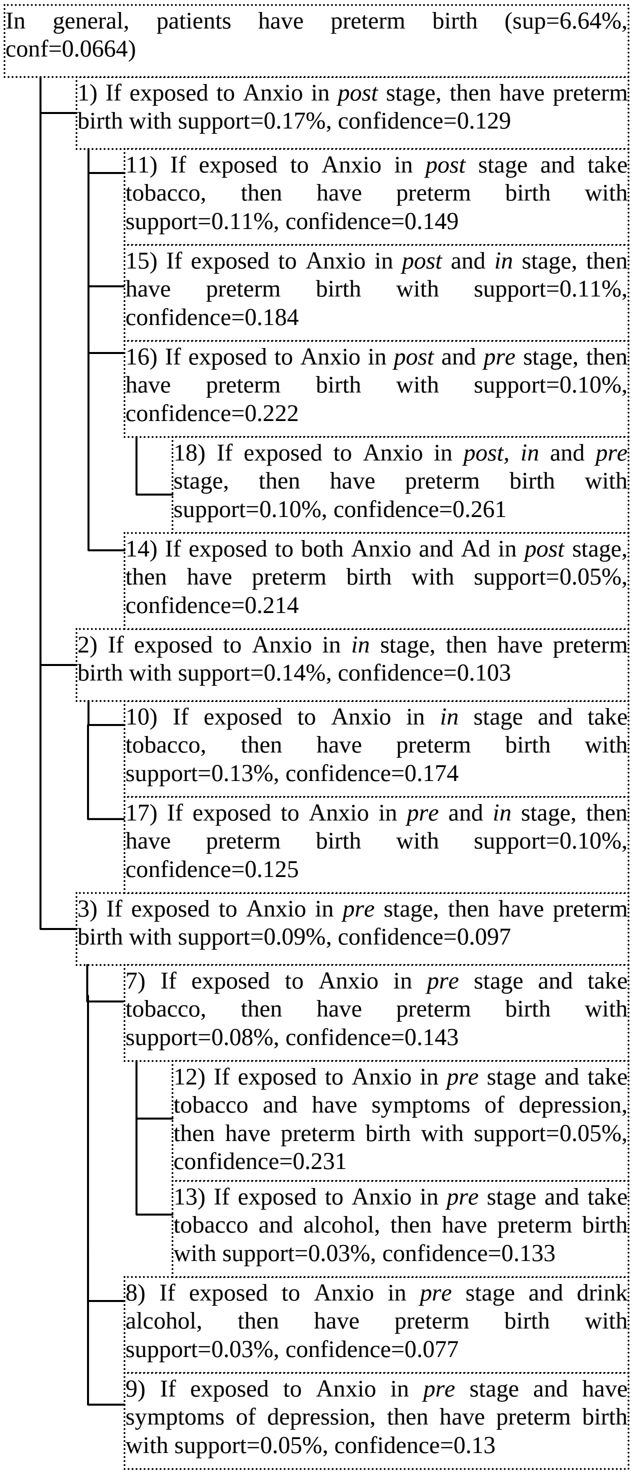
The above rules suggest an association with exposure to sedatives before or at time of implantation and preterm birth given tobacco smoke and symptoms of depression. It is not possible to determine whether or not the association represents a causal relation from the given dataset. An association with early exposure to some sedatives and preterm birth has been suggested but is disputed.

From the interaction of the drug Ad and Anxio, we find that the exposure of two drugs in most time period combinations results in higher preterm birth than any of one drug can cause. The largest increase of preterm rate happens when the patient exposes to both Ad and Anxio in the *post* stage.

14) If exposed to both Anxio and Ad in *post* stage, then have preterm birth with support=0.05%, confidence=0.214

For the sequence of exposure of one drug in different time periods, we find that exposure to Anxio in multiple time periods increases the preterm rate.

15) If exposed to Anxio in *post* and *in* stage, then have preterm birth with support=0.11%, confidence=0.184
16) If exposed to Anxio in *post* and *pre* stage, then have preterm birth with support=0.10%, confidence=0.222
17) If exposed to Anxio in *pre* and *in* stage, then have preterm birth with support=0.10%, confidence=0.125
18) If exposed to Anxio in *post, in* and *pre* stage, then have preterm birth with support=0.10%, confidence=0.261



In general, patients have preterm birth (sup=6.64%, conf=0.0664)

1) If exposed to Anxio in *post* stage, then have preterm birth with support=0.17%, confidence=0.129

11) If exposed to Anxio in *post* stage and take tobacco, then have preterm birth with support=0.11%, confidence=0.149

15) If exposed to Anxio in *post* and *in* stage, then have preterm birth with support=0.11%, confidence=0.184

16) If exposed to Anxio in *post* and *pre* stage, then have preterm birth with support=0.10%, confidence=0.222

18) If exposed to Anxio in *post*, *in* and *pre* stage, then have preterm birth with support=0.10%, confidence=0.261

14) If exposed to both Anxio and Ad in *post* stage, then have preterm birth with support=0.05%, confidence=0.214

2) If exposed to Anxio in *in* stage, then have preterm birth with support=0.14%, confidence=0.103

10) If exposed to Anxio in *in* stage and take tobacco, then have preterm birth with support=0.13%, confidence=0.174

17) If exposed to Anxio in *pre* and *in* stage, then have preterm birth with support=0.10%, confidence=0.125

3) If exposed to Anxio in *pre* stage, then have preterm birth with support=0.09%, confidence=0.097

7) If exposed to Anxio in *pre* stage and take tobacco, then have preterm birth with support=0.08%, confidence=0.143

12) If exposed to Anxio in *pre* stage and take tobacco and have symptoms of depression, then have preterm birth with support=0.05%, confidence=0.231

13) If exposed to Anxio in *pre* stage and take tobacco and alcohol, then have preterm birth with support=0.03%, confidence=0.133

8) If exposed to Anxio in *pre* stage and drink alcohol, then have preterm birth with support=0.03%, confidence=0.077

9) If exposed to Anxio in *pre* stage and have symptoms of depression, then have preterm birth with support=0.05%, confidence=0.13

Figure 3. A part of the rule hierarchy for the exposure to the sedative medication and other confounders at different time period of pregnancy with preterm birth

However, the exposure to Ad in multiple time periods does not increase the rate of preterm.

19) If exposed to Ad in *post* and *in* stage, then have preterm birth with support=0.27%, confidence=0.108

20) If exposed to Ad in *post* and *pre* stage, then have preterm birth with support=0.22%, confidence=0.107

21) If exposed to Ad in *pre* and *in* stage, then have preterm birth with support=0.45%, confidence=0.092

22) If exposed to Ad in *post, in* and *pre* stage, then have preterm birth with support=0.21%, confidence=0.103

In this study, the data mining tool suggested association with various sequences and combinations of drugs not previously described. However, the traditional methods were not able to evaluate some of the results due to limited sample size.

4. KNOWLEDGE DISCOVERY FROM DATA MINING RESULTS

After the association rules have been generated, our goal is to discover novel knowledge that is valuable to the users. However, there are two challenges for the knowledge discovery process: 1) examining the vast number of rules manually is too labor-intensive; and 2) exploring knowledge (rules) without specific goal.

Techniques have been proposed to organize and summarize the discovered rules. For example, in [10, 11], association rules are represented in general rules, summaries and exception rules (GSE patterns). The GSE pattern presents the discovered rules in a hierarchical fashion. Users can browse the hierarchy from top down to find interesting exception rules. However, in our case, due to the low support of drug side effects, interesting rules are exception rules and usually reside at the lower level of the hierarchy. Without user guidance, it requires exploration of the entire GSE hierarchy to locate the interesting exception rules.

In the early stage of discovery, it is quite usual that even domain experts cannot specify exactly what they are looking for. In such case, we can derive a set of *seed attributes* from high-confidence rules and then explore more rules based on these *seed attributes* in the rule hierarchies. Thus the knowledge discovery process starts from the bottom of the rule hierarchy rather than the top-down approach in [10].

More specifically, we start with high-confidence rules that contain the seed attributes in the rule body. Although such complex rules may be difficult to interpret directly, they contain the seed attributes that contribute to the rule head. For example, the following rule has high confidence:

> If exposed to Anxio in the pre, in and post time and use tobacco and have symptoms of depression, then have preterm birth with confidence = 0.6

From the above rule, we can learn that the exposure to sedatives at pre, in and post time may cause the preterm birth. Tobacco or symptoms of depression may also increase preterm rate. Therefore, we obtain the following list of seed attributes for further exploration: Anxio_pre, Anxio_in, Anxio_post, tobacco and symptoms of depression. We can first look for the rules that represent the effect of each single seed attribute on preterm birth, and then we can further explore the combination of multiple seed attributes. For the above example, we find single-attribute rules as:

> If exposed to Anxio in the post time, then have preterm birth with confidence =0.129

> If exposed to Anxio in the in time, then have preterm birth with confidence =0.097

And rules with multiple attributes as:

> If exposed to Anxio in the post time and use tobacco, then have preterm birth with confidence = 0.149

Therefore, given a set of high-confidence rules with multiple attributes, we can search for more general rules from the rule hierarchy using a bottom-up approach for knowledge discovery.

Since the SmartRule outputs the association rules in an Excel worksheet, each column represents an attribute and each rule is represented in a row. Therefore, when searching for rules in the hierarchy, we can utilize the rich functionality in Excel to sort and filter the rules based on different columns. For example, when we want to find the most relevant factor that causes preterm birth with seed attribute citalopram (”*cita*”) exposure in the 3rd trimester, we first use filters to select all the rules with column ”*cita_3*” has value ”1” and the rule body contains two attributes (i.e. the column “Length” of the rule body equals ”2”). This will show all rules with citalopram exposure in the 3rd trimester and a single other attribute. Then we sort the filtered rules by their confidence in descending order as shown in Figure 4. The most significant factor should appear in the first rule after the sort. By filtering and sorting tools of the spreadsheet, we can easily find the rules that show the effect of each seed attribute and their combined effects. As a result, the users can gradually learn the multiple aspects of the knowledge through the rule hierarchy.

5. DATA MINING VS. TRADITIONAL STATISTICAL ANALYSIS

Statistical based analysis, e.g. logistical regression, is the standard approach to process data in epidemiology. Thus, we need to compare the association rules derived from data mining against the results obtained from the standard methods.

Because of the large number of combinations among all the attributes and their values, it becomes infeasible to test all the potential interactions among the attributes and to locate the significant ones. However, data mining does not require domain experts to propose a hypothesis and is capable of mining side effects in large dataset with multiple temporal attributes. In hypothesis generating studies and in monitoring large health registries, this is a crucial advantage compared to traditional analytical approaches. In our experiments, we are already able to discover important associations (e.g. alcohol intake and anti-depressants) that can be validated by the statistical methods, but were missed by the traditional methods.

Further, testing hypotheses with small sample size has limited statistical power. On the other hand, data mining can generate association rules independent of the sample size, which represents an advantage in our study since the side effects of drugs are usually represented by small sample sizes.

6. CONCLUSION

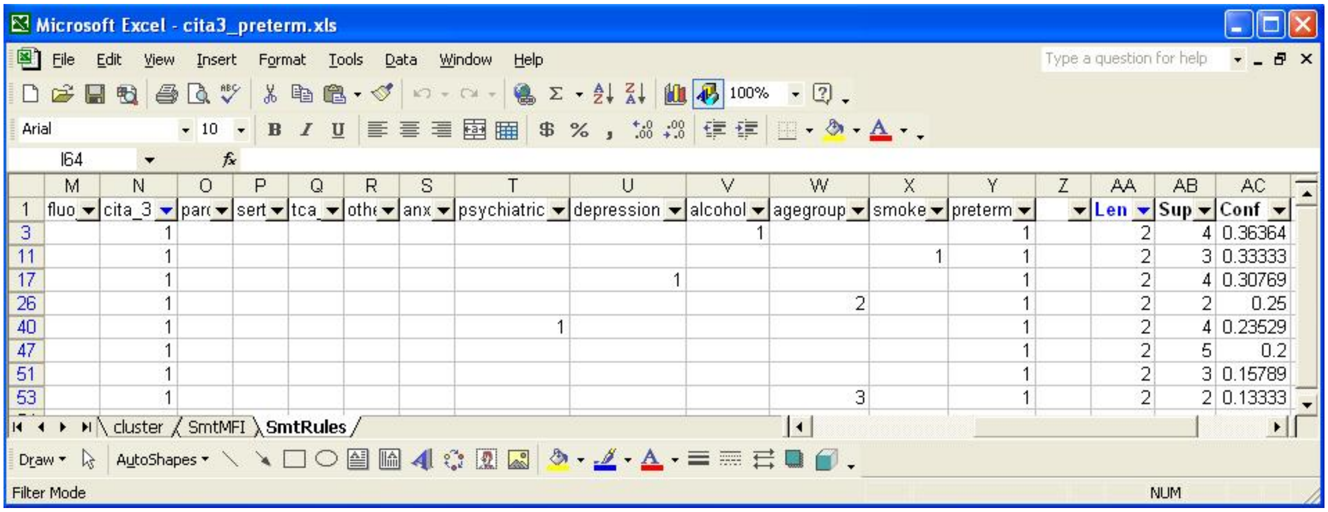
In this paper, we have proposed to use data mining technique to discover side effects of drug exposure to pregnant women from real data.

Specifically, we develop SmartRule to generate MFIs directly from tabular data sources. Further, the users can select a subset of the MFIs to derive targeted association rules that is relevant with the specified attributes. We found that the interactive capability of our data mining tool is a valuable feature for the user to explore useful rules. The filtering and sorting facility of the spreadsheet provides a very effective user interface in the discovery of the side effects. When the user wants to explore new findings, he/she can derive a set of seed attributes from high-confidence rules and search for additional rules based on this set of attributes from the rule hierarchy using a bottom-up approach.

Our data mining tool has the advantage over traditional analytical approaches that potentially important data sources can be examined for ‘signals’ of importance for clinical or public health practice without having to wait for a proper hypothesis to come by. Data mining is independent of hypotheses and is capable of mining side effects in large datasets with multiple temporal attributes. Further, it also alleviates the problems of statistical hypotheses testing for very small sample size as in our case. The accuracy of our data mining technique was validated by regression analysis with known published results where regression analysis is valid. Our data mining is able to generate new discovery from the clinical dataset that the traditional method are missed or unable to derive due to its limitation. Therefore, data mining represents a useful tool in bio-medical research for the discovery of new knowledge from epidemiological data.

7. ACKNOWLEDGMENTS

This research is in part supported by the NSF IIS-0313283, the Lundbeck Foundation and the National Danish Research Council.



	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	fluo	cita_3	parc	sert	tca	othe	anx	psychiatric	depression	alcohol	agegroup	smoke	preterm		Len	Sup	Conf
3		1								1			1		2	4	0.36364
11		1										1	1		2	3	0.33333
17		1							1				1		2	4	0.30769
26		1									2		1		2	2	0.25
40		1						1					1		2	4	0.23529
47		1											1		2	5	0.2
51		1											1		2	3	0.15789
53		1									3		1		2	2	0.13333

Figure 4. To find the most relevant factor that causes preterm birth with citalopram exposure in the 3rd trimester (”*cita_3*”), we first use filters to select all the rules with column ”*cita_3*” has value ”1” and the column ”Length” equals ”2”; then sort the filtered rules by their confidence in descending order. In this example, eight rules were derived. To interpret these rules, value “1” in column “*cita_3*” represents exposure to citalopram in the third trimester; value “1” in column “*psychiatric*” represents patient with psychiatric disease; value “1” in column “*depression*” represents patient with depression; value “1” in column “*alcohol*” represents patient with alcohol intake; in column “*agegroup*”, value “2” represents patient 20 to 25 years old; value “3” represents patient 25 to 30 years old; value “1” in column “*smoke*” represents patient with tobacco exposure; and value “1” in column “*preterm*” represents birth before 37 full weeks of gestation.

8. REFERENCE

[1] Briggs, GG et al. Drugs in Pregnancy and Lactation. Ed. Williams & Wilkins Lippincott. 7 ed. 2005

[2] Carlson, B. M. Human Embryology and Developmental Biology. 3 ed. Mosby, 2004.

[3] Chambers CD, Johnson KA, Dick LM, Felix RJ, Jones KL. Birth outcomes in pregnant women taking fluoxetine. N.Engl.J.Med. 1996;335:1010-15.

[4] Wen SW, Yang Q, Garner P, Fraser W, Olatunbosun O, Nimrod C et al. Selective serotonin reuptake inhibitors and adverse pregnancy outcomes. Am.J.Obstet.Gynecol. 2006;194:961-66.

[5] Q. Zou, Y. Chen, W. W. Chu and X. Lu. Mining Association Rules from Tabular Data Guided by Maximal Frequent Itemsets. Book Chapter in “*Foundations and Advances in Data Mining*”, edited by Wesley W. Chu and T.Y. Lin, Springer, 2005.

[6] Q. Zou, W.W. Chu, and B. Lu. SmartMiner: A depth-first search algorithm guided by tail information for mining maximal frequent itemsets. In Proc. of the IEEE Intl. Conf. on Data Mining, 2002.

[7] Q. Zou, W. Chu, D. Johnson, and H. Chiu: Pattern Decomposition Algorithm for Data Mining of Frequent Patterns. Journal of Knowledge and Information System, 2002.

[8] D. Burdick, M. Calimlim, and J. Gehrke. MAFIA: a maximal frequent itemset algorithm for transactional databases. In Intl. Conf. on Data Engineering, April 2001.

[9] K. Gouda and M. J. Zaki. Efficiently mining maximal frequent itemsets. In Proc. of the IEEE Int. Conference on Data Mining, San Jose, CA, April 2001.

[10] B. Liu, M. Hu, and W. Hsu, "Multi-level organization and summarization of the discovered rules," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Aug, 2000, Boston, USA.

[11] B. Liu, M. Hu, and W. Hsu, "Intuitive representation of decision trees using general rules and exceptions." *Proceedings of Seventeeth National Conference on Artificial Intellgience (AAAI-2000)*, July 30 - Aug 3, 2000, Austin, Texas, USA.

[12] Frequent Itemset Mining Implementations Repository, http://fimi.cs.helsinki.fi/

About the authors:

Yu Chen is a Ph.D. candidate in the Computer Science Department at the University of California, Los Angeles. She received her master’s degree in Computer Science from Northeastern University, Boston in 2001. Her research interests include security and privacy in information systems, data mining and knowledge discovery.

Lars Henning Pedersen is a Ph.D. student at the Danish Epidemiological Science Centre, University of Aarhus, Denmark. He received his MD in 2003 from University of Aarhus, and has had his residency in general medicine and the first year of clinical pharmacology. His interests include teratology, obstetrics, and epidemiological methods. He is a member of the Teratology Society.

Dr. **Wesley W. Chu** is a distinguished professor of Computer Science and was the past chairman (1988-1991) for the Computer Science Department at the University of California, Los Angeles. He researched computer communications and distributed databases at Bell Laboratories, Holmdel, New Jersey (1966-1969). He joined the University of California, Los Angeles in 1969. His current research interest is in the areas of knowledge-based medical information systems, intelligent information systems and security and privacy in information systems. Dr. Chu was the recipient of the 2003 IEEE Computer Society Technical Achievement Award for contributions to Intelligent Information Systems. He is also a member of the Editorial Board for the Journal on Applied Intelligence and an Associate Editor for the Journal of Data and Knowledge Engineering. Dr. Chu is a Fellow of IEEE. (http://www.cs.ucla.edu/~wwc/)

Dr. **Jorn Olsen** is professor and chair at Department of Epidemiology, School of Public Health, University of California, Los Angeles, head of the Danish National Birth Cohort and is president of the International Epidemiological Association (IEA).