Uncovering insights from healthcare archives to improve operations: An association analysis for cervical cancer screening

[1]*Carmen Kar Hang Lee, [2]Ying Kei Tse, [3]To Sum Ho, [4]Sai Ho Chung

[1]Singapore University of Social Sciences Singapore, Singapore

[2]Cardiff Business School, Cardiff University, United Kingdom, CF10 3EU

[3]Department of Supply Chain and Information Management, The Hang Seng University of Hong Kong, Hong Kong

[4] Department off Industrial Systems and Engineering, The Hong Kong Polytechnic University, Hong Kong

*Corresponding author

Abstract

The digitalization in healthcare opens opportunities for more effective chronic disease management. Digitalized medical records are valuable data sources for identifying high-risk patients and facilitating early clinical intervention. However, the liberation of data has plagued adoption among physicians as massive data mean more difficult to identify important knowledge from the data. In the cervical cancer context, many patients are adherence to prescription medications only when symptoms appear, beyond the earlier point-in-time of the disease progression. Regular screening is the only way to detect abnormal cells that may develop into cancer if left untreated. Yet, without a comprehensive understanding of the relationship between risk factors and healthcare outcomes, inappropriate screening procedures may be conducted, lengthening the treatment process. Delay in the treatment process may have an irreversible influence on patients' conditions as chronic diseases progress. This study demonstrates a data-mining framework which extracts knowledge that can advance cervical cancer screening processes in the form of association rules and improves the generalization potential of the rules for deployment. The knowledge discovered serves as an additional supplement for physicians' experience and uncovers appropriate screening strategies based on patients' risk factors, increasing the chance for high-risk patients to get treated for cervical pre-cancers.

Keywords: Healthcare analytics, knowledge discovery, association rules, chronic disease management, cervical cancer screening.

1. Introduction

The digitalization of the healthcare service and ecosystem has led to a significant transformation in the healthcare sector. Typical technologies that mainstream the digitalization in healthcare include wearable medical devices, Internet of Things, cloud computing and healthcare analytics. The rapid development of these technologies enables the liberation of personal health data and opens opportunities for a more cost-effective approach to chronic disease management.

Nevertheless, the opportunities for chronic disease management brought by digitalization in healthcare are limited for developing countries due to their lack of stable availabilities of information and communication technologies. As a result, the mortality rate of various diseases is higher in developing countries, compared to developed countries. Cervical cancer is one of the most common chronic diseases among females. Worldwide, it is the fourth most frequently occurring malignancy in women, resulting in an estimated more than half millions new cases per year (Small et al., 2017). The discrepancy in cervical cancer incidence and mortality between developed and developing countries has become increasingly apparent (LaVigne et al., 2017). Nearly 90% of cervical cancer mortality cases occurred in low- and middle-income countries where women have poor access to prevention, screening and treatment (World Health Organization, 2019). Such a disproportionate global cervical cancer burden is due to the lack of resources and knowledge in the developing countries.

Regarding cervical cancer, there are no symptoms observed in its early stage. Many patients are aware of their health conditions only when symptoms such as pelvic pain and vaginal bleeding appear. Upon diagnosis, the cancer has already been in its later stage. If not managed to slow the progression of the cancer, the cancer may lead to greater damage to the patients by spreading from the cervix to other parts in the body. Routine screening tests are the only way to detect any abnormal cells that may develop into cancer if left untreated. This study discovers the strong links between risk factors and four screening methods which are Hinselmann (i.e., colposcopy using acetic acid), Schiller (i.e., colposcopy using Lugol iodine), cytology and biopsy. Details of the methods are summarized in the Appendix. One of the challenges faced by the physicians is that the suggestions of screening methods highly depend on the physicians' expertise and subjective comfort on the decision process (Fernandes, Cardoso & Fernandes, 2017). Without a comprehensive understanding of the relationship between patients' risk factors and healthcare outcomes, inappropriate screening procedures may be suggested, lengthening the entire treatment process with additional time wasted on following up with more appropriate screening tests.

The medical record archive provides excellent and massive data for physicians to identify high-risk patients and prioritize healthcare resources for screening. In this study, we demonstrate the use of data-mining approaches to enhance the quality of healthcare programs and extract medical knowledge that can avoid delays in cervical cancer diagnosis. Some 'metadata' which may look not important may have an in-direct linkage to the appropriateness of screening methods. Association rule mining is a promising data mining tool to identify relationships among variables. In the context of cervical cancer detection, it can be applied to discover the relationships between risk factors (e.g. Human Papilloma Virus (HPV), smoking habit, sexually transmitted diseases (STDs), and use of contraceptives) and the screening strategies. Yet, in the absence of a systematic mechanism to manage the medical knowledge and control the quality of decisions made, it is possible that the association rules mined from the medical data are irrelevant or with poor generalization potential. In addition, it is also impractical for physicians to discover useful knowledge if the number of rules obtained is very large. This highlights the importance of rule reduction and rule validation in the medical domain. Accordingly, this paper addresses the following research questions:

1. How to extract knowledge for cervical cancer screening using association rule mining from the massive data archives in healthcare systems?
2. How to improve the generalization potential of healthcare association rules for (e.g. cervical cancer) screening program deployment?

The aim of the study is to extract useful knowledge that can avoid treatment delays in medical screening processes. There are two reasons causing the delays. The first one is that the knowledge and awareness of patients is generally low while the second one is that some physicians do not make

appropriate decisions on site (Lim et al., 2014). In line with this view, the significance of this study can be realized from two perspectives. Firstly, the proposed framework can uncover the links between risk factors and appropriate screening strategies that can improve patients' awareness and knowledge of the disease. Prior studies have shown that increasing awareness and knowledge of risk factors can increase acceptance of screening tests (Kahesa et al., 2012). For instance, in an educational intervention in India, increasing awareness about cervical cancer symptoms to 76% in an intervention community as compared with 25% in the control community was associated with significant change in stages at presentation of cervical cancer from 38% before the intervention to 51% after the intervention (Jayant et al., 1995). It is therefore suggested that action to increase patients' awareness of risk factors and symptoms is important to healthcare screening programs. Specifically, rules extracted by using our proposed framework can be used as an initial component of educational programs to encourage early medical consultation for risk factors suggestive of cervical cancer. Secondly, the knowledge discovered can serve as an additional supplement for physicians' experience in decision making. This is useful to avoid diagnosis delay or mismanaging cases with high-risk patients. The rules extracted by using the framework can be used jointly with physicians' experience to design a set of guidelines for physicians to follow when arranging patients for screening, minimizing the observed discrepancy between evidence-based treatment strategies and current practice in the healthcare domain (McGlynn, et al., 2003; Nolte et al., 2012).

The motivations of this study are twofold. First, because of the aging population (Compagna & Kohlbacher, 2015; Peine & Moors, 2015), the prevalence of chronic diseases is increasing worldwide (Schuitmaker, 2012). Chronic disease is therefore a global burden. Chronic conditions frequently go untreated until the disease has been progressed to a stage where more expensive and intensive treatments are required. Thus, an effective response to the rising burden of chronic diseases is needed to improve the quality of patients' lives while reducing healthcare costs by preventing or minimizing the effects of a disease. Second, in the era of digitalization, maximizing the use of various sources of personal health information, such as electronic health records and telehealth apps, have put us in a position to uncover important health insights via healthcare analytics (Ceccato & Price, 2019). Healthcare analytics that reveals insights from evidences has had an increasing demand for knowledge discovery, problem solving, and prediction. Information that could be obtained from healthcare analytics includes association between symptoms and the development of suitable treatment plans for individual patients (Baker et al., 2017).

The rest of this paper is organized as follows: Section 2 reviews the existing literature related to this study. Section 3 describes the methodology, constraints and validation procedures. Section 4 is the experimental results and discussion. Section 5 presents the contributions of this study. Lastly, Section 6 concludes this study and suggests future research directions.

2. Literature Review

2.1 Cervical Cancer

Arising from the cervix, cervical cancer is caused by the change in genes that affect the growth and division function of cells (Wu & Zhou, 2017). No symptoms can be observed in its early stage. Routine screening tests are the only way to detect any abnormal cells that may develop into cancer if left untreated. In developing countries, yet, resources are scarce and women tend to have poor adherence to regular screening tests due to low problem awareness (LaVigne et al., 2017; Nakisige et al., 2017). As a consequence, it is not uncommon that women there are diagnosed with cervical cancer in the

later stage where symptoms, such as pelvic pain and vaginal bleeding, appear. Cervical cancer may also spread from the cervix to other parts in the body. It has been one of the dominant causes of mortality in developing countries, especially low-income countries. On the other hand, in developed countries, cervical cancer control has been largely associated with both primary prevention and second prevention, more comprehensively approaching cervical cancer. Primary prevention can be effectively done by having Human Papilloma Virus (HPV) vaccination as HPV infection has been acknowledged as a key factor that causes cervical cancer (Marlow et al., 2007; Bao et al., 2008). Despite the discovery of strong links between HPV and cervical cancer, implementing HPV vaccination in developing countries faces more challenges than in developed countries. For instance, HPV vaccines were not available in mainland China until 2016 (Zhao, 2017) and thus Chinese women usually did not have the opportunities to be vaccinated before immigration (Seo et al., 2018).

HPV is considered as necessary but insufficient cause of cervical cancer (Bailey et al., 2016). As the types of HPV associated with cervical cancer are transmitted sexually, studies on cervical cancer often take into account other risk factors such as a woman's lifetime number of sexual partners and her age at commencement of sexual activity (Bosch et al., 1992). In addition, other factors such as smoking, nutrition, parity and use of hormonal contraceptives have also been reported (Kjellberg et al., 2000; Plummer et al., 2003). Studies have consistently found that smoking may influence the risk of progression from cervical HPV infection to cervical malignancy (Roura et al., 2014). Prolonged use of oral contraceptives increases the risk of cervical cancer among HPV positive women (Moreno et al., 2002; Smith et al., 2003).

While these studies serve as useful references to treat cervical cancer, it remains unclear whether the reported risk factors are independent risk factors or whether they may act as cofactors to HPV infection in cervical carcer (Kjellberg, et al., 2000). Therefore, apart from HPV vaccination, secondary prevention, such as having regular cervical cancer screening, becomes important. Unfortunately, in view of the lack of resources and knowledge, women in developing countries tend to have poor adherence to regular screening tests. Their awareness of cervical cancer is low as cervical cancer has no symptoms in the early stage. As a result, women are diagnosed with cervical cancer only when symptoms appear in the later stage. Preventive measures need to be prioritized to minimize the negative effects caused by the late presentation of women with advanced cervical cancer (Nakisige et al., 2017). Of urgent necessity is the discovery of relevant medical knowledge for identifying high-risk patients, the availability of and linkage to appropriate screening strategies.

2.2 Association Rule Mining in Healthcare

Routine screening tests are the only way to detect any abnormal cells that may develop into cancer if left untreated. Therefore, prediction of patients' risk level and suggesting the most suitable screening strategy is vital to provide timely diagnosis. Data-driven approaches can be applied to discover the relevant knowledge. Certain detection methods, such as data-driven approaches, have been proposed in the healthcare domain for the provision of timely diagnosis (de Braal et al., 2006; Salmeron et al., 2017; Wu & Zhou, 2017). In particular, association analysis has great potential to improve disease prediction (Ordonez, 2006).

Regarding cervical cancer, various causal factors have been studied individually (Bosch et al., 1992; Moreno et al., 2002; Smith et al., 2003; Plummer et al., 2003; Bailey et al., 2016). However, how the co-occurrence of two or more factors affects the chance of cervical malignancy remains unclear (Kjellberg et al., 2000). Association analysis, on the other hand, allows us to shift attention from

individual factors to the associative patterns among factors. This provides a more holistic view to assess the risk of patients who may possess more than one factor. Based on the combination of risk factors, appropriate screening strategies can be suggested based on the historical data stored in the medical record archive. The Apriori Algorithm proposed by Agrawal and Srikant (1994) is widely used for association analysis. It employs a downward closure property: if an itemset is not frequent, then any superset of it cannot be frequent either (Coenen, Leng & Ahmed, 2004; Song & Rajasekaran, 2006). The Apriori Algorithm performs a breadth-first search in the search space by generating candidate *k*+1-itemsets from frequent *k*-itemsets. Only frequent itemsets are used to generate association rules. As such, the number of rules increases with the number of frequent itemsets. Furthermore, threshold values for support and rule confidence have to be defined, which affects the quality of the rules obtained (Lim et al., 2012). If a rule with its support and confidence larger than or equal to the threshold values, it is considered useful or significant (Demiriz et al., 2011). On the other hand, if one sets the threshold values too low, the number of rules generated could be huge. Hence, one of the critical tasks in association rule mining is to set an appropriate support and confidence threshold values such that irrelevant rules are filtered out and only the significant rules are generated. In addition, various constraints have been incorporated in the Apriori algorithm to reduce the number of rules. Constraints imposed on the antecedent and consequents of the rules were used in Ng et al. (1998). Item constraints were incorporated to include or exclude certain items appearing in the rules in Srikant, Vu and Agrawal (1997). Constraints on the support values were used for pruning the number of candidate itemsets generated (Wang, He & Han, 2003).

In addition to rule reduction, rule validation is critical in medical knowledge discovery for cervical cancer screening because most of the screening methods highly depend on individual expertise and subjective comfort on the decision process (Fernandes, Cardoso & Fernandes, 2017). This implies that rules without validation could be with poor generalization potential or low predictive accuracy. Another reason supporting rule validation is that collecting new medical records with similar characteristics is not easy due to privacy issues (Roddick, Fule & Graco, 2003; Ordonez, 2006).

To summarize, this paper discovers medical knowledge for cervical cancer screening in the forms of association rules. It advances previous research on incorporating constraints for association rules and validating rules with train and test approaches so as to get rules with high predictive accuracy. The rules represent valuable knowledge to help countries, especially low-income countries, to more effectively allocate appropriate equipment for the treatment of cervical cancer and avoid delays in patient presentation.

3. Methodology

In the era of big data, massive data from various sources such as telehealth applications, wearable sensors and home devices become available. These sources have given us an opportunity to access to structured and unstructured data that can be leveraged to uncover important health insights. Each single source of data offers the potential to identify cost-effective treatments for chronic diseases. For example, data from telehealth applications can address mental and physical challenges of patients with chronic illnesses; data captured by wearables and home devices can be used for monitoring the status of chronic disease patients and predicting adverse events before they occur (Jiang & Cameron, 2020). However, to fully realize the benefits of big data, data from various sources have to be integrated to maximize the possibility use of the data. The public health sectors in many countries have started to digitalize old medical records and integrate them to implement a single national medical archive for residents. For example, Iceland's government appointed a private company to

create an Icelandic health sector genetic database which links genetic samples with individual health records (Winickoff, 2006). In fact, there are plentiful targeted use cases showing that improvement has been made by the application of big data in various domains, for example, in predicting newborn complications (Malak et al., 2019). If we are able to extract more interesting hidden patterns from the data to diagnose chronic diseases, there are many opportunities to capitalize on.

The research methodology for extracting knowledge for chronic disease is shown in Figure 1. In this study, we have used cervical cancer screening as the examples of extracting the screening knowledge. It starts with the collection of medical data from the digital repository. The data repository stores the integrated data which are merged from body check records, screening records, previous patient medical history, and demographic information of previous patients. In the context of cervical cancer screening, the data required can be classified into two types: patients' risk factors (e.g., age, number of pregnancies of the patients) and clinical tests (e.g., suggested screening tests). Considered that association rule mining requires the data to be categorical, data discretization is performed to convert numeric attributes into categorical attributes. For example, the age of a patient is numeric and can be converted into three categories such as {young}, {middle-aged}, and {old}. In association rule mining, each category is considered as an item. These items are used to generate association rules, each of which represents an IF-THEN relationship between items. For instance, one of the rules could be: IF the patient's age is {old} and the number of pregnancies is {many}, THEN the suggested screening test is {biopsy}.

The association rule learning algorithm embeds two search constraints which are used to achieve rule reduction. This feature is important in the era of digitalization. Without a systematic approach to filtering the rules, the number of rules could be massive, and it becomes impractical for physicians to identify critical rules for decision making. Details of the constrained-rule mining algorithm are presented in Section 3.1.

In addition, a train and test approach is adopted in the methodology for rule evaluation. To do so, the data collected is partitioned into two subsets: training data and testing data. The association rule mining is applied only on the training data. Yet, the rules obtained may not be representative or generalized enough for unseen data. Thus, before deployment, they are validated using the testing data. The objective is to ensure that the rules obtained are generalized enough for deployment even in developing countries where healthcare analytics may not be feasible. The medical significance of rules is evaluated in terms of support, confidence, and lift. Details of the train and test approach are given in Section 3.2.

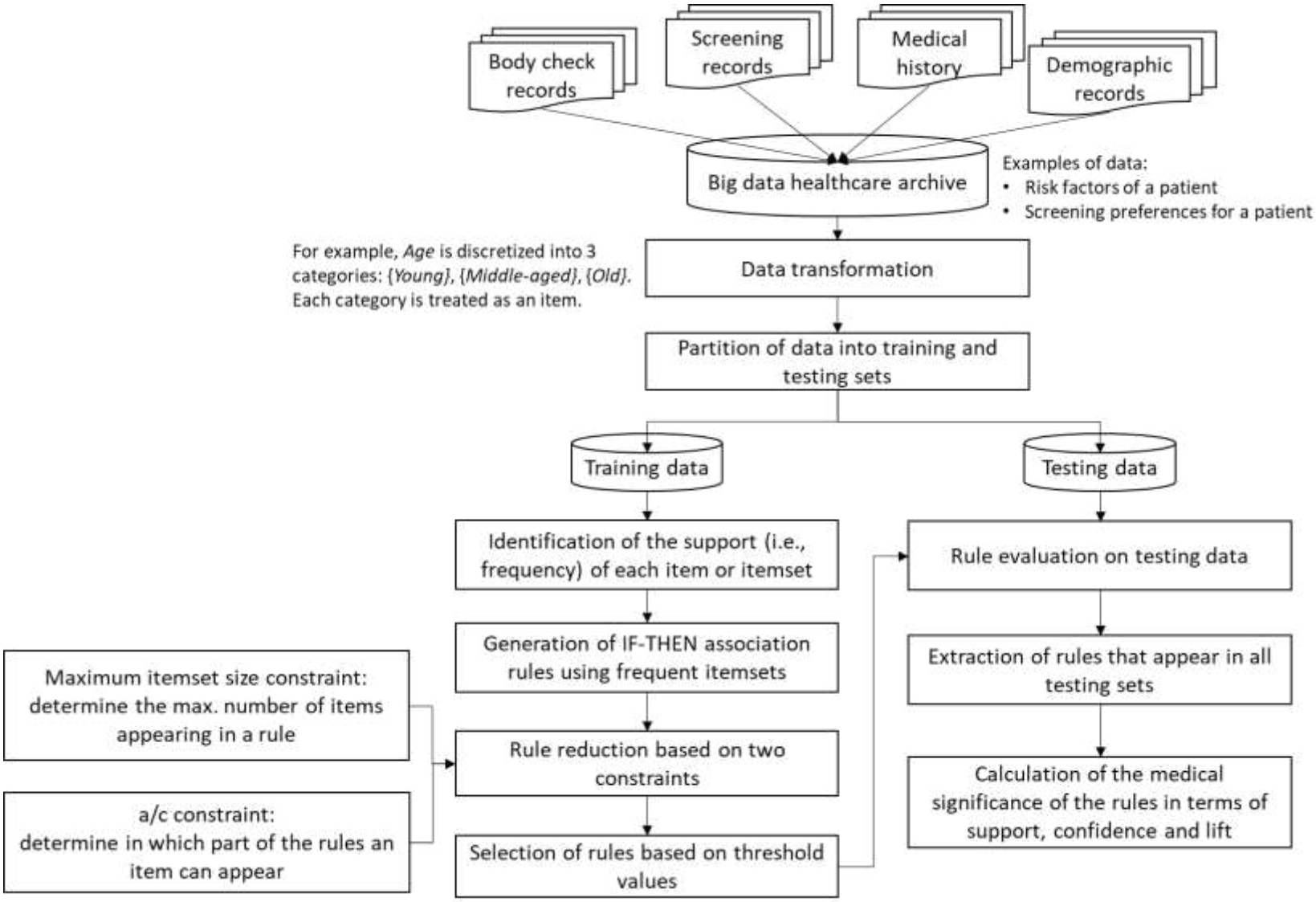

Figure 1. Methodology for extracting knowledge for chronic disease management

3.1 Constrained-rule Learning Algorithm

The constrained-rule learning algorithm is based on the standard Apriori algorithm (Agrawal & Srikant, 1994). Let $I = \{i_1, i_2, ..., i_n\}$ be a set of items, and $DB = \{T_1, T_2, ..., T_q\}$ be a medical record database, where $T_i (i \in [1...q])$ is a medical record containing a set of items in I. An itemset A is a subset of I. A containing k items is called a k-itemset. $sup(A)$, the support of A, is the number of medical records containing A in DB. It is also known as the probability of having the occurrence of A in DB, i.e. $P(A) = sup(A)$. A is a frequent itemset if $sup(A)$ is greater than or equal to a user-specified minimum support threshold λ. Let itemset A and itemset B be two subsets of I. An association rule denoted by $A \Rightarrow B$ implies that if A appears (antecedent), then B appears (consequent). The support of rule $A \Rightarrow B$ is defined as $sup(A \Rightarrow B) = sup(A \cup B)$ that is the probability of having the co-occurrence of A and B in DB. Rule confidence $conf(A \Rightarrow B)$ is used to measure the reliability of the rule that is conditional probability of the occurrence of A given the occurrence of B. Therefore, $conf(A \Rightarrow B) = P(B|A) = P(A \cup B)/P(A) = sup(A \cup B)/sup(A)$. The rule is considered reliable if its confidence is larger than or equal to the minimum confidence threshold β. Another measure called lift, $lift(A \Rightarrow B)$, is used to quantify the relationship between A and B. It is defined as $lift(A \Rightarrow B) = conf(A \Rightarrow B)/sup(B)$. In general, a lift ratio larger than 1 meaning that A and B depend on each other. The algorithm generates association rules in two phases: finding frequent itemsets (Phase 1) and generating rules (Phase 2), as summarized in Figure 2.

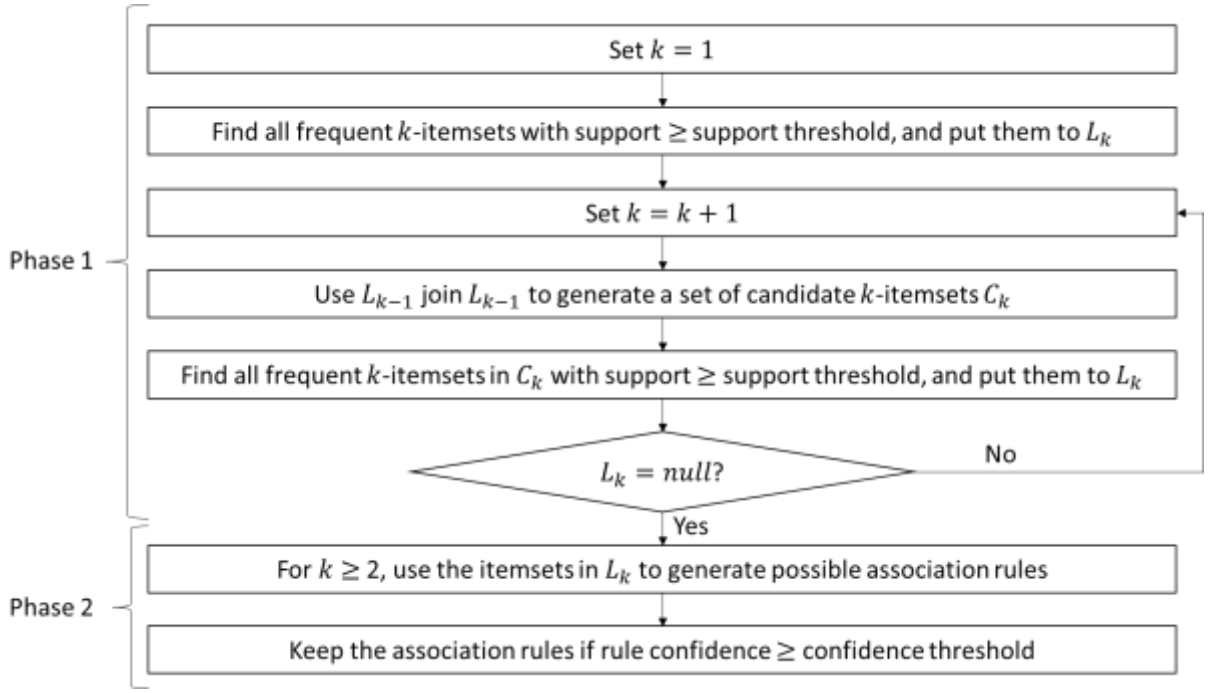

Figure 2. Outline of the Apriori algorithm

There are two search constraints incorporated in the algorithm: maximum itemset size constraint and the antecedent-consequent (a/c) constraint. To reduce the number of rules, the former is used in Phase 1 while the latter is used in Phase 2.

(i) Maximum itemset size constraint

Given m unique items in I, the number of frequent itemsets can be up to $2^m - 1$. Since m can be a very large number in disease prediction, the search space of itemsets that need to be explored is exponentially large. As association rules are generated using the frequent itemsets, the number of rules increases with the number of frequent itemsets. The total number of possible association rules

generated from all the frequent itemsets can be up to 3^m-2^{m+1}+1. This constraint restricts the size of frequent itemsets up to a maximum size ψ, terminating the search process of frequent itemsets with size larger than ψ. This constraint is applied in Phase 1 of the algorithm where the itemsets are generated. It is simple but effective in reducing the number of rules. Another reason supporting the incorporation of this constraint is that the associations among too many items may be irrelevant for quality prediction when the itemset size is large. Though this constraint can also be applied in Phase 2, it is suggested applying it as early as possible (i.e. in Phase 1) to improve the computation efficiency.

(ii) a/c constraint

The a/c constraint is used to reduce the number of rules by specifying in which part of the rules the items should appear. There are three possible conditions for an item to appear in a rule. First, an item can only appear in the antecedent of a rule. Second, an item can only appear in the consequent of a rule. Third, an item can appear in either antecedent or consequent of a rule. The a/c constraint eliminates all the rules containing items that do not appear in the correct place of the rules. For cervical cancer risk level assessment, items related to risk factors can only appear in the antecedent of a rule. In this study, available cervical cancer screening strategies for diagnosis can appear in either antecedent or consequent of a rule. This is because it is possible that even though other screening methods have been done (i.e. screening methods appearing in the antecedent of a rule), a gynecologist might still request another screening methods for further diagnosis (i.e. screening methods appearing in the consequent of a rule). For instance, some screening methods might not be as accurate as biopsy. Therefore, sometimes even if the actual decision suggested by the other methods is negative, the gynecologist may request a biopsy.

3.2 Train and Test Approach

To validate the rules, training and testing approaches are used. *DB* is partitioned into two independent subsets: training set DB_{tr} and testing set DB_{te}. $DB = DB_{tr} \cup DB_{te}$. The size of DB_{tr} is determined by a training fraction ε and is equal to $\varepsilon \times q$. In this paper, the constrained-rule mining algorithm is firstly applied in DB_{tr} so as to generate a set of training rules R_{tr}. Then, R_{tr} is validated on DB_{te} by setting the set of testing rules $R_{te} = R_{tr}$. For each rule $A \Rightarrow B$ in R_{te}, $sup(A \Rightarrow B)$, $conf(A \Rightarrow B)$ and $lift(A \Rightarrow B)$ are computed. Rule $A \Rightarrow B$ is eliminated from R_{te} if $sup(A \Rightarrow B)$, $conf(A \Rightarrow B)$ and $lift(A \Rightarrow B)$ are smaller than λ, β, and γ, respectively. Set $R_\theta = R_{test}$. This process is repeated t times on independent test samples to get a set of accurate predictive association rules. Only rules that appear in all R_θ are considered valid. Let $A \Rightarrow B$ be a valid rule appearing in all R_θ. It is measured in terms of the average support, confidence and lift values as (1)-(3).

$$Avg.sup(A \Rightarrow B) = t^{-1} \sum_{\theta=1}^{t} sup(A \Rightarrow B, R_\theta) \quad (1)$$

$$Avg.conf(A \Rightarrow B) = t^{-1} \sum_{\theta=1}^{t} conf(A \Rightarrow B, R_\theta) \quad (2)$$

$$Avg.lift(A \Rightarrow B) = t^{-1} \sum_{\theta=1}^{t} lift(A \Rightarrow B, R_\theta) \quad (3)$$

4. Experimental Results and Discussion

The data set used in this study was collected and organized by Fernandes, Cardoso and Fernandes (2017). It was collected at Hospital Universitario de Caracas in Caracas, Venezuela, and can be

accessed in the UCI Machine Learning Repository[1]. It contains 858 patients' medical records with 26 attributes that can be referred to Fernandes et al. (2017). 22 attributes are risk factors related to patients' demographic information, habits and medical history while 4 attributes are the available screening strategies (i.e. colposcopy using acetic acid – Hinselmann, colposcopy using Lugol Iodine – Schiller, Cytology and Biopsy). The experiments in this study are divided into three steps. First, the data set is transformed into a transaction database *DB* as defined as Section 3. Second, association rules are discovered from the database and the numbers of rules with and without the incorporation of the constraints are reported. Third, rule validation is performed on several independent test samples to achieve cross-validation. Details of each step are discussed in the following sections.

4.1 Transformation of Data Set

(i) Discretization of numeric attributes

To transform the data set into a transaction database for association rule mining, each numeric value in the data set is processed as an item. However, this will result in an extremely large number of items as some attributes have a large range of values. In this study, if an attribute contains more than 5 items, discretization is performed to divide the range into intervals using *k*-means clustering where *k* is a user-specified number of groups. For instance, the range of "Age" is from 13 to 84. Before discretization, the number of items related to "Age" is up to 72. Set *k*=3, "Age" is binned at 25 and 35. Consequently, there are 3 items related to "Age": {Age<25}, {25<=Age<35}, and {Age>=35}. In the database, on the other hand, there are only 4 values in "STDs: Number of diagnosis": 0, 1, 2 and 3. Therefore, discretization is not necessary to reduce the number of items associated with this attribute. For Boolean attributes, however, no discretization is required. If the value of a Boolean attribute is 1 (True), this means that the attribute item exists in the transaction.

(ii) Item filtering

Some items are filtered out from this study. It is found that the majority of patients decided not to provide their "STDs: Time since first diagnosis" and "STDs: Time since last diagnosis". This could be due to privacy concerns or lack of information. Since these two attributes contain missing values in most of the records, they are excluded from this study. Besides, no records in the database contain "1" in two of the Boolean attributes, namely "STDs:cervical condylomatosis " and "STDs:AIDS". This means that no patients were diagnosed with cervical condylomatosis or AIDs. To improve the computation efficiency, these two attributes are filtered out from this study.

(iii) Combination of items

Some items in the database could be combined. In the database, there are Boolean attributes used to indicate whether the patients have certain habits or diseases. For instance, "Smokes" contains 0 or 1 to indicate if the patient is a smoker. It is known that if the patients are non-smokers, their values of "Smokes (years)" and "Smokes (packs/year)" must be 0. In order to improve the computation efficiency by reducing the number of attributes, "Smokes" is removed from the database while two items, {SmokeYear=0} and {SmokePack=0}, are created under the attributes "Smokes (years)" and "Smokes (packs/year)", respectively. Similarly, "Hormonal Contraceptives" is removed from the database while {HormContraYear=0} is created. "IUD" and "STDs" are also removed and {IUDYear=0} and {STDNo.=0} are created.

[1] https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

The number of items created for each attribute is listed in Table 1. Overall, association rules are discovered among 62 items in 858 records, i.e. n=62 and q=858.

Table 1. Items in Data Set after Transformation

Attribute	Type	Range / Value	No. of items
Age	Integer	13 – 84	3
Number of sexual partners	Integer	1 – 28	3
First sexual intercourse (age)	Integer	10 – 32	3
Num of pregnancies	Integer	0 – 11	4
Smokes (years)	Integer	0 – 37	6
Smokes (packs/year)	Integer	0 – 37	5
Hormonal Contraceptives (years)	Integer	0 – 30	6
IUD (years)	Integer	0 – 19	5
STDs (number)	Integer	0 – 4	5
STDs:condylomatosis	Boolean	1 / 0	1
STDs:vaginal condylomatosis	Boolean	1 / 0	1
STDs:vulvo-perineal condylomatosis	Boolean	1 / 0	1
STDs:syphilis	Boolean	1 / 0	1
STDs:pelvic inflammatory disease	Boolean	1 / 0	1
STDs:genital herpes	Boolean	1 / 0	1
STDs:molluscum contagiosum	Boolean	1 / 0	1
STDs:HIV	Boolean	1 / 0	1
STDs:Hepatitis B	Boolean	1 / 0	1
STDs:HPV	Boolean	1 / 0	1
STDs: Number of diagnosis	Integer	0 – 3	4
Dx:Cancer	Boolean	1 / 0	1
Dx:CIN	Boolean	1 / 0	1
Dx:HPV	Boolean	1 / 0	1
Dx	Boolean	1 / 0	1
Hinselmann	Boolean	1 / 0	1
Schiller	Boolean	1 / 0	1
Cytology	Boolean	1 / 0	1
Biopsy	Boolean	1 / 0	1

4.2 Rule Reduction Using Search Constraints

In this section, each constraint is studied individually to measure its impact on rule reduction. Figure 3 shows the number of rules under different parameter settings. Obviously, the number of rules decreases with β. This is because only rules with confidence greater than or equal to β are considered useful. With a larger β, more rules are not considered useful and thus the number of rules obtained decreases. Figure 3(a-e) show that the impact of β on rule reduction is more important at λ =1% and λ =2%. When $\lambda \geq$3%, the increase in β does not reduce the number of rules significantly.

In addition, it can be seen from each of the sub-graphs in Figure 3 that the number of rules decreases with λ. With a smaller λ, more itemsets become frequent itemsets candidates for rule generation. It is observed that, when the β remains unchanged, the numbers of rules are relatively comparable when $\lambda \geq$3%. This shows that the number of occurrences of the majority of items or itemsets in the transaction is less than 26 times (i.e. 858 x 3%).

An outstanding rule reduction is achieved after the introduction of the a/c constraint regardless of the threshold values. In the presence of the a/c constraint, the impact of λ on the number of rules becomes less important at $\lambda \geq 3\%$, as shown in Figure 3(f-j). Results showed that less than 20 rules were obtained at $\lambda = 4\%$. The number of rules discovered is very limited at $\lambda \geq 4\%$. This illustrates a trade-off between λ and the use of the a/c constraint.

Furthermore, the number of rules increases with ψ. The impact of ψ on the number of rules is more significant when the a/c constraint is absence. However, after the a/c constraint is introduced, there is a point where the number of rules become saturated. At $\lambda = 1\%$ and $\lambda = 2\%$, the number of rules does not increase further when ψ is greater than 9. At $\lambda = 3\%$, the number of rules does not increase any more when ψ is greater than 5. Similarly, At $\lambda = 4\%$, the number of rules does not change when ψ is greater than 4. This illustrates another trade-off between λ and ψ. Figure 4 shows the distribution of rules with different sizes of itemsets with and without the a/c constraints at $\lambda = 1\%$. It is found that frequent 6-itemsets are dominant when the a/c constraint is absent while frequent 5-itemsets are dominant when the a/c constraint is present.

4.3 Rule Validation

To validate the rules, the training fraction was $\varepsilon = 70\%$ in this study. λ and β are set to be 1% and 50%, respectively. In general, γ is set to be 1. To achieve basic cross validation and to eliminate rules that cannot be generalized, a valid rule must meet the minimum threshold values λ, β and γ on both training and testing sets, in the presence of the two search constraints. Association rules are discovered in the training set, and the training rules are validated in the testing set. This process is repeated three times (i.e. $t = 3$) to get three independent sets of rules to compute the intersection of rule sets, where each rule's support, confidence and lift are computed as averages of rule metrics from all testing rule sets.

Table 2 shows the number of rules obtained in all the training and testing sets. The number of rules in each individual experiment varies and the average number of rules is calculated. However, the three testing rule sets contain rules that might be in common or different. Only the rules that exist in all the testing sets are considered valid. Totally there are 614 valid rules obtained in this study. Table 3 is the summary of the valid rules. The majority of the rules that remains valid on all testing sets has an average support count value smaller than or equal to 2%. Slightly more than half of the rules have an average confidence value greater than or equal to 85%. All the rules are with a relatively high lift ratio that is greater than or equal to 2.5. In terms of the rule size, rules containing 5 items are dominant. Around one third of the valid rules are generated from the frequent 5-itemsets. Table 4 shows some examples of the valid rules obtained in this study. The first rule in Table 4 means that if the patient has pregnancy three times or more, does not have any sexual transmitted diseases, is not a smoker, even she has done the Schiller screening test, it is still recommended that a biopsy screening test should be arranged. The second rule means that if the patients aged 16 but below 20, does not use intrauterine device, is not a smoker nor diagnosed with sexual transmitted diseases, and have done Schiller screening test, it is recommended that a Hinselmann screening test should be arranged. The rule confidence measures the reliability of the rule. In our case, the first rule has a higher accuracy and thus it is more reliable. The second rule has a relatively low confidence, indicating that the knowledge is less reliable compared to that in rule 1. Users are able to adjust the threshold values so as to focus the most reliable rules.

To summarize, the knowledge discovered by the algorithm acts as a decision support for physicians to be well-informed the appropriate screening strategy when the risk factors of patients are given. The appropriateness of screening strategies used for the detection of pre-cancerous cervical lesions has a critical influence on patients' conditions, both physically and financially. For instance, a biopsy has the highest accuracy among all the other screening strategies. However, because of the possible complications of a biopsy, it should not be suggested without considering the risk level of the patients. In addition, the knowledge discovered can be used to speed up the diagnosis process, without additional time spent on following up with more appropriate screening tests. Eventually, the overall treatment costs can be lower because the knowledge discovered opens up an opportunity to facilitate early intervention before chronic diseases progress to a stage where higher cost and more intensive treatments are required (Kohli & Tan, 2016). This is a significant contribution because treatment of chronic diseases could be very expensive (e.g. it consumes more than 80% of the healthcare costs in the U.S. (Thompson et al., 2020)).

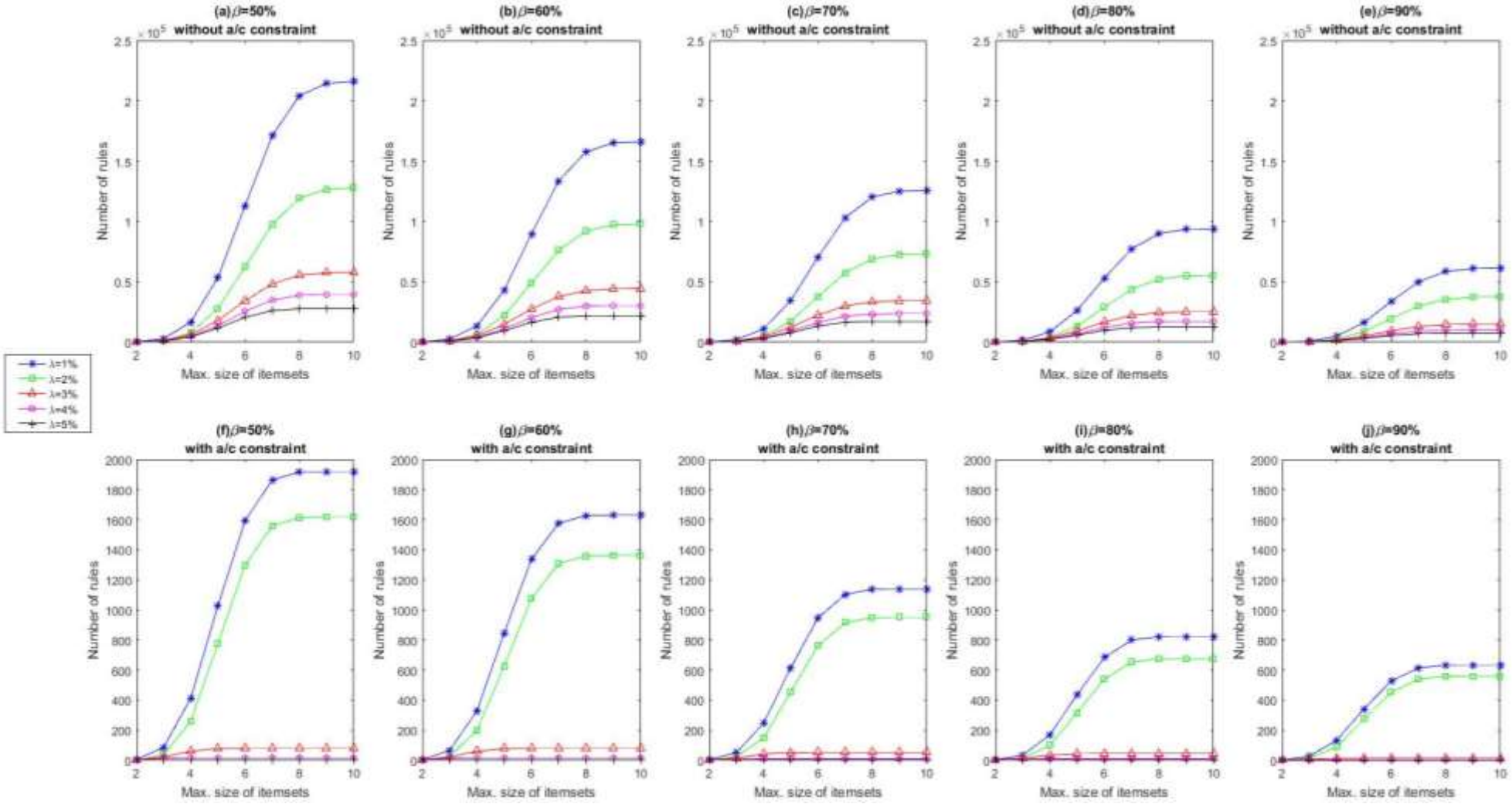

Figure 3. Number of rules with and without constraints

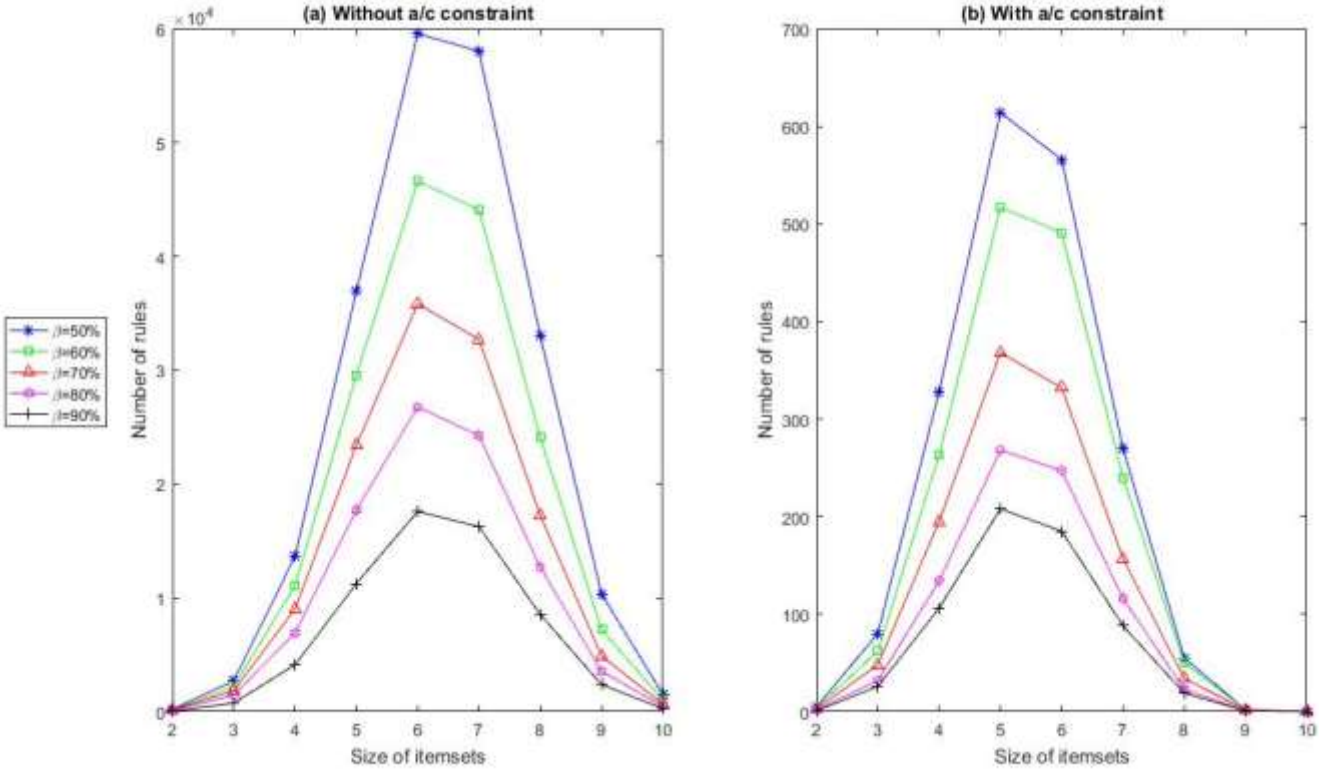

Figure 4. Distribution of rules with different sizes of itemsets with and without the a/c constraint

Table 2. Number of rules in training and testing sets

	Experiment 1		Experiment 2		Experiment 3		Average	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
ψ =2	6	4	5	5	5	4	5.3	4.3
ψ =3	97	58	79	59	74	58	83.3	58.3
ψ =4	482	255	356	229	348	260	395.3	248.0
ψ =5	1203	562	862	486	824	581	963.0	543.0
ψ =6	1907	829	1339	679	1252	863	1499.3	790.3
ψ =7	2238	955	1594	759	1461	998	1764.3	904.0
ψ =8	2310	985	1659	773	1506	1028	1825.0	928.7
ψ =9	2316	987	1664	773	1508	1030	1829.3	930.0

Table 3. Summary of rules existing in all testing sets

	No. of rules	Percentage
All	614	100%
Avg. sup ≥1%	614	100%
Avg. sup ≥2%	403	66%
Avg. sup ≥3%	109	18%
Avg. sup ≥4%	20	3%
Avg. sup ≥5%	4	1%
Avg. conf ≥50%	614	100%
Avg. conf ≥55%	605	99%
Avg. conf ≥60%	596	97%
Avg. conf ≥65%	572	93%
Avg. conf ≥70%	484	79%
Avg. conf ≥75%	424	69%
Avg. conf ≥80%	376	61%
Avg. conf ≥85%	314	51%
Avg. conf ≥90%	244	40%
Avg. conf ≥95%	128	21%
Avg. conf ≥100%	116	19%
Avg. lift ≥1.5	614	100%
Avg. lift ≥2	614	100%
Avg. lift ≥2.5	614	100%
Rules generated from 2-itemsets	4	1%
Rules generated from 3-itemsets	44	7%
Rules generated from 4-itemsets	144	23%
Rules generated from 5-itemsets	205	33%
Rules generated from 6-itemsets	149	24%
Rules generated from 7-itemsets	58	9%
Rules generated from 8-itemsets	10	2%

Table 4. Examples of valid rules

Rule	Avg. sup	Avg. conf	Avg. lift
IF Pregnancy>=3 & STDDiagnosis=0 & STDNo=0 & Schiller& SmokePack=0, THEN Biopsy	6.33	81.59	11.69
IF 16<=SexAge<20 & IUDYear=0 & STDDiagnosis=0 & Schiller & SmokePack=0 & SmokeYear=0, THEN Hinselmann	4.33	61.11	13.75

5. Contributions

5.1 Improving disease knowledge and reducing overall healthcare costs

The proposed algorithm discovers hidden medical knowledge in the form of rules. For example, based on the risk factors of a patient, it could predict whether the patient is high-risk and then generate a cervical cancer screening test suggestion. The suggestion made by the algorithm is believed to be a suitable prescription for the patient based on the association among medical-related variables stored

in the transaction database. Physicians are allowed to decide whether to accept the suggestion or to propose alternatives based on their experience. Therefore, the knowledge discovered can serve as an additional supplement for physicians' experience. With the rapid growth of digitalization, it is expected that more data will become available and can be inputted in the algorithm. As a consequence, when the data volume increases over time and more successful instances are used for healthcare analytics, physicians can be more confident that the knowledge discovered by the algorithm is of better quality.

With references to the rules, at-risk patients can be effectively identified for screening and healthcare resources can be better prioritized, shortening the entire treatment process. As the appropriate screening strategies are recommended in the first time, early intervention is facilitated before chronic diseases progress to a stage where higher cost and more intensive treatments are required. Such a cost reduction is particularly crucial in developing countries where the risk of conducting inappropriate screening tests for patients could be expensive.

In chronic disease management, it is important not only for the physicians to be able to make decisions, but also for the patients to accrue the knowledge, confidence and skills to manage their condition. The antecedent part of the rules states the risk conditions of a patient. If patients fulfil the conditions of a particular rule, they should be aware of their health status and seek medical help by, for example, arranging diagnostic check-ups. Increasing the problem awareness of the patients can increase their chance to receive treatment for cervical pre-cancers. As pre-cancer treatments are often less costly than cancer treatments, the knowledge discovered by the algorithm can reduce patient's economic burden.

5.2 Promoting patients' behavioral modifications

A number of chronic conditions are closely connected to unhealthy lifestyle behaviors. Behavior change utterly depends on individuals but is critical for chronic disease management. Hence, more attention should be put on promoting patient involvement and engagement with their health and healthcare. Participatory medicine has been seen as a mean to promote patients' behavioral modifications that are required for the prevention, management, and treatment of chronic conditions (Laranjo, 2016). Participatory medicine refers to the movement of patients from being passively receiving medical treatments decided by physicians to being actively participating in decisions related to their health (Frydman, 2010). For instance, there is an increasing number of patients who seek health information via a suite of online technologies such as social media, video and podcasts (Gallant et al., 2011). With reference to what they read online, patients are demanding personal information they can use to improve their health and wellbeing (Flore et al., 2013). One of the contributions of this study is that it provides a foundation on participatory medicine by generating rules that improve patients' disease knowledge and encourage self-manage behaviors. For instance, based on the first rule in Table 4, patients become aware of the need of biopsy screening test when they know they had done the Schiller screening test and had pregnancy three times or more, even though they are not a smoker nor having any STDs. Other rules may also make patients realize that smoking habit has a significant relationship with cervical cancer. In short, the knowledge discovered by the algorithm is useful to remind relevant patients to take certain actions such as conducting regular diagnostic check-ups and being adherence to prescription medications at an earlier point-in-time of the disease progression.

5.3 Ensuring the quality of rules and feasibility of deployment

The massive amount of data for knowledge discovery has plagued adoption in the healthcare industry as higher costs are required to collect, analyze and disseminate clinical outcomes (Kohli & Tan, 2016). This study addresses this issue by embedding search constraints in the standard Apriori algorithm. The advantages are twofold. First, it helps to reduce the number of rules, which in turn decreases costs in terms of technology investments for analytics. This is critical in the era of digitalization as more data is generated rapidly at an accelerating rate. Without a systematic approach to filtering the rules, the number of rules could be massive while a large proportion could be trivial or meaningless rules. Second, discarding irrelevant rules improve the feasibility and practicability of decision making in the treatment process. Physicians can pay attention to the most important rules for decision making.

Further, to validate the rules obtained, train and test iterations were performed to discard rules with poor generalization potential. Hence, our result generated a set of well-trained rules, uncovering the best practice of chronic disease management in developed countries. A significant contribution made by this study is that the knowledge discovered by the well-trained rules can be generally applied in countries where healthcare analytics is not mature. The valid rules can be a valuable and additional supplement for improving the healthcare programs across countries.

6. Conclusion

This paper presents a constrained-rule learning algorithm to discover knowledge for cervical cancer screening in the form of association rules. Experiments are conducted on a real data set to illustrate the impacts of constraints and the elimination of irrelevant rules with validation on independent test sets. The medical significance of the rules is evaluated in terms of support, confidence, and lift. The rules represent valuable knowledge that can help the medical industry to improve the quality of cervical cancer screening programs so that patients can receive timely treatment in an earlier stage where symptoms have not yet appeared.

A limitation of this study is that the knowledge discovered is specifically for cervical cancer screening. The extracted knowledge may not be applicable to other types of chronic diseases. Nevertheless, the methodology proposed in this study is a generic one, meaning that it can be applied to various medical datasets for knowledge discovery to advance our understanding of other chronic diseases.

Future research directions are twofold. Firstly, this paper does not explore how the rules discovered can be applied for optimizing healthcare operations in an actual environment. Currently, the paper contributes to the area of descriptive and predictive analytics. Yet, it has a potential to be extended to prescriptive analytics for finding the optimal course of action for a given situation. Future work can focus on extending this study to prescriptive analytics by integrating the algorithm with optimization techniques. For instance, when there are a group of patients opting for healthcare appointments with providers, it is important to generate an optimal course of action plan for scheduling and prioritizing healthcare resources. Another research area is to integrate fuzzy set concepts into the rule mining algorithm. In the medical domain, judgement made by physicians' is usually expressed in linguistic terms or in fuzzy ones which have no clear boundaries and cannot be precisely associated with a real number. Therefore, knowledge represented in the form of fuzzy association rules could help provide more direct knowledge support for cervical cancer screening. It would be interesting to know whether fuzzy association rule mining can help improve the medical significance of the rules.

References

Agrawal, R., Srikant, R. (1994), Fast algorithms for mining association rules in large databases, In *Proceedings of 20th international conference on very large databases*, Santiago de Chile, pp. 487–489.

Bailey, H.H., Chuang, L.T., DuPont, N.C., Eng, C., Foxhall, L.E., Merrill, J.K., Wollins, D.S., Blanke, C.D. (2016). American Society of Clinical Oncology Statement: Human Papillomavirus Vaccination for Cancer Prevention. *Journal of Clinical Oncology 34*(15)*,* 1803-1812.

Baker, S.B., Xiang, W., Atkinson, I. (2017). Internet of Things for smart healthcare: technologies, challenges and opportunities. *IEEE Access 5*, 26521-26544.

Bao, Y.P., Li, N., Smith, J.S., Qiao, Y.L. (2008). Human papillomavirus type distribution in women from Asia: A meta-analysis. *International Journal of Gynecological Cancer, 18* (1), 71-79.

Bengtsson, E., Malm, P. (2014). Screening for cervical cancer using automated analysis of PAP-smears. *Computational and Mathematical Methods in Medicine 2014,*842037.

Bosch, F. X., Muñoz, N., De Sanjosé, S., Izarzugaza, I., Gili, M., Viladiu, P., Tormo, M.J. et al. (1992). Risk factors for cervical cancer in Colombia and Spain. *International journal of cancer 52*(5), 750-758.

Ceccato, N., Price, C. (2019). When personal health data is no longer "personal". *Healthcare Management Forum 32*(6), 326-328.

Chuang, L.T., Temin, S., Camacho, R., Feldman, S., Gultekin, M., Gupta, V., Horton, S. et al., (2016). Management and care of women with invasive cervical cancer: American Society of Clinical Oncology Resource-Stratified Clinical Practice Guideline. *Journal of Global Oncology 2*(5), 311-340.

Coenen, F., Leng, P., Ahmed, S. (2004). Data Structure for Association Rule Mining: T-Trees and P-Trees. *IEEE Transactions on Knowledge and Data Engineering 16*(6), 774-778.

Compagna, D., Kohlbacher, F. (2015). The limits of participatory technology development: The case of service robots in care facilities for older people. *Technological Forecasting & Social Change 93*, 19-31.

de Braal, L., Ezquerra, N., Schwartz, E., Cooke, C. D., Garcia, E. (1996). Analyzing and predicting images through a neural network approach. In *Visualization in Biomedical Computing* (pp. 253-258). Springer, Berlin, Heidelberg.

Demiriz, A., Ertek, G., Atan, T., Kula, U. (2011). Re-mining item associations: Methodology and a case study in apparel retailing. *Decision Support Systems 52*(1), 284-293.

Fernandes, K., Cardoso, J.S., Fernandes, J. (2017). "Transfer Learning with Partial Observability Applied to Cervical Cancer Screening," Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing.

Fernandes, K., Cardoso, J.S., Fernandes, J. (2018). Automated methods for the decision support of cervical cancer screening using digital colposcopies. *IEEE Access 6*, 33910-33927.

Flores, M., Glusman, G., Brogaard, K., Price, N.D., Hood, L. (2013). P4 medicine: how systems medicine will transform the healthcare sector and society. *Personalised Medicine 10*(6), 565-576.

Frydman, G. (2010) A Patient-Centric Definition of Participatory Medicine. *Society for Participatory Medicine*. Available at: https://participatorymedicine.org/epatients/2010/04/a-patient-centric-definition-of-participatory-medicine.html (Accessed 24 August 2019).

Gallant, L.M., Irizarry, C., Boone, G., Kreps, G.L. (2011). Promoting Participatory Medicine with Social Media: New Media Applications on Hospital Websites that Enhance Health Education and e-Patients' Voices. *Society for Participatory Medicine*. Available at: https://participatorymedicine.org/journal/evidence/research/2011/10/31/promoting-participatory-medicine-with-social-media-new-media-applications-on-hospital-websites-that-enhance-health-education-and-e-patients-voices/ (Accessed 24 August 2019).

Jayant, K., Rao, R.S., Nene, B.M., Dale, P.S. (1995). Improved stage at diagnosis of cervical cancer with increased cancer awareness in a rural Indian population. *International Journal of Cancer 63*, 161-163.

Jiang, J., Cameron, A.-F. (2020). IT-enabled self-monitoring for chronic disease self-management: An interdisciplinary review. *MIS Quarterly 44*(1), 451-508.

Kahesa, C., Kjaer, S., Mwaiselage, J., Ngoma, T., Tersbol, B., Dartell, M., Rascho, V. (2012). Determinants of acceptance of cervical cancer screening in Dar es Salaam, Tanzania. *BMC Public Health 12*, 1093. https://doi.org/10.1186/1471-2458-12-1093

Kjellberg, L., Hallmans, G., Ahren, A.-M., Johansson, R., Bergman, F., Wadell, G., Ångström, T., Dillner, J. (2000). Smoking, diet, pregnancy, and oral contraceptive use as risk factors for cervical intra-epithelial neoplasia in relation to human papillomavirus infection. *British Journal of Cancer 82*(7), 1332-1338.

Kohli, R., Tan, S.S.-L. (2016). Electronic health records: How can IS researchers contribute to transforming healthcare? *MIS Quarterly 40*(3), 553-573.

Laranjo, L. (2016). Social media and health behavior change. In *Participatory Health Through Social Media* (pp. 83-111). Academic Press.

LaVigne, A.W., Triedman, S.A., Randall, T.C., Trimble, E.L., Viswanathan, A.N. (2017). Cervical cancer in low and middle income countries: Addressing barriers to radiotherapy delivery. *Gynecologic Oncology Reports 22*, 16-20.

Lim, A. H. L., Lee, C. S., Raman, M. (2012). Hybrid genetic algorithm and association rules for mining workflow best practices. *Expert Systems with Applications 39*(12), 10544-10551.

Lim, A.W., Ramirez, A.J., Hamilton, W., Sasieni, P., Patnick, J., Forbes, L.J.L. (2014). Delays in diagnosis of young females with symptomatic cervical cancer in England: an interview-based study. *British Journal of General Practice 64*(627), e602-e610.

Malak, J.S., Zeraati H, Nayeri, F.S., Safdari R, Shahraki, A.D. (2019). Neonatal intensive care decision support systems using artificial intelligence techniques: a systematic review. *Artificial Intelligence Review 52*(4), 2685-2704.

Marlow, L., Waller, J., Wardle, J. (2007). Public awareness that HPV is a risk factor for cervical cancer. *British Journal of Cancer 97*(5), 691-694.

McGlynn, E.A., Asch, S.M., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., Kerr, E.A. (2003). The quality of health care delivered to adults in the United States. *The New England Journal of Medicine 348*, 2635–2645.

Moreno, V., Bosch, F.X., Muñoz, N., Meijer, C.J., Shah, K.V., Walboomers, J.M., Herrero, R. et al. (2002). Effect of oral contraceptives on risk of cervical cancer in women with human papillomavirus infection: the IARC multicentric case-control study. *Lancet 359*(9312), 1085-1092.

Nakisige, C., Schwartz, M., Ndira, A.O. (2017). Cervical cancer screening and treatment in Uganda. *Gynecologic Oncology Reports 20*, 37-40.

Ng, R., Lakshmanan, L.V.S., Han, J., Pang, A. (1998). Exploratory Mining and Pruning Optimizations of Constrained Associations Rules. In *Proceeding ACM SIGMOD Conference*, 13-24.

Nolte, E., Conklin, A., Adams, J. L., Brunn, M., Cadier, B., Chevreul, K., Durand-Zaleski, I. et al. (2012). *Evaluating chronic disease management: Recommendations for funders and users*. RAND Corporation.

Ordonez, C. (2006). Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction. *IEEE Trans. Information Technology in Biomedicine 10*(2), 334-343.

Peine, A., Moors, E.H.M. (2015). Valuing health technology – habilitating and prosthetic strategies in personal health systems. *Technological Forecasting & Social Change 93*, 68-81.

Plummer, M., Herrero, R., Franceschi, S., Meijer, C.J.L.M., Snijders, P., Bosch, F.X., de Sanjosé,S., Muñoz, N. (2003) Smoking and cervical cancer: pooled analysis of the IARC multi-centric case–control study. *Cancer Causes & Control 14*(8), 805-814.

Randall, T.C., Ghebre, R. (2016). Challenges in Prevention and Care Delivery for Women with Cervical Cancer in Sub-Saharan Africa. *Frontiers in Oncology 6*(160), 1-7.

Roddick, J.F., Fule, P., Graco, W.J. (2003). Exploratory medical knowledge discovery: Experiences and issue. *SIGKDD Explorations Newsletter 5*(1), 94-99.

Roura, E., Castellsagué, X., Pawlita, M., Travier, N., Waterboer, T., Margall, N., Bosch, F.X. et al. (2014). Smoking as a major risk factor for cervical cancer and pre-cancer: Results from the EPIC cohort. *International Journal of Cancer 135*(2), 453-466.

Salmeron, J.L., Rahimi, S.A., Navali, A.M., Sadeghpour, A. (2017). Medical diagnosis of Rheumatoid Arthritis using data driven PSOCFCM with scarce datasets. *Neurocomputing 232,* 104-112.

Schuitmaker, T.J. (2012). Identifying and unravelling persistent problems. *Technological Forecasting & Social Change 79*(6), 1021-1031.

Sellors, J.W., Sankaranarayanan, R. (2003). *Colposcopy and Treatment of Cervical Intraepithelial Neoplasia: A Beginner's Manual*. Diamond Pocket Books.

Seo, J.Y., Li, J., Li, K. (2018). Cervical Cancer Screening Experiences Among Chinese American Immigrant Women in the United States. *Journal of Obstetric, Gynecologic & Neonatal Nursing 47*(1), 52-63.

Small, W., Bacon, M. A., Bajaj, A., Chuang, L. T., Fisher, B. J., Harkenrider, M. M., Jhingran, A., et al.. (2017), Cervical cancer: A global health crisis. *Cancer 123*, 2404–2412.

Smith, J.S., Green, J., De Gonzalez, A.B., Appleby, P., Peto, J., Plummer, M., Franceschi, S. et al. (2003). Cervical cancer and use of hormonal contraceptives: a systematic review. *Lancet 361*, 1159-1167.

Song, M., Rajasekaran, S., (2006). A Transaction Mapping Algorithm for Frequent Itemsets Mining. *IEEE Transactions on Knowledge and Data Engineering 18*(4), 472-481.

Srikant, R., Vu, Q., Agrawal, R. (1997). Mining Association Rules with Item Constraints. In *Proceedings ACM Knowledge Discovery and Data Mining Conference,* 67-73.

Thompson, S., Whitaker, J., Kohli, R., Jones, C. (2020). Chronic disease management: How IT and analytics create healthcare value through the temporal displacement of care. *MIS Quarterly 44*(1), 227-256.

Wang, K., He, Y., Han, J. (2003). Pushing Support Constraints Into Association Rules Mining. *IEEE Transactions on Knowledge and Data Engineering 15*(3), 642-658.

Winickoff, D. E. (2006). Genome and nation: Iceland's health sector database and its legacy. *Innovations: Technology, Governance, Globalization 1*(2), 80-105.

World Health Organization (2019). *Human papillomarvirus (HPV) and cervical cancer*. Available at: https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer (Accessed 24 Jun 2020).

Wu, W., Zhou, H. (2017). Data-driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches. *IEEE Access.*

Zhao, X.Y. (2017). Vaccine for HPV wins approval. *China Daily*. Available at: http://www.chinadaily.com.cn/china/2017-05/22/content_29437390.htm (Accessed 6 February 2018)

Appendix

Screening methods include cytology, colposcopy and biopsy. Both cytology and colposcopy are image-based screening processes. The former involves examination of vaginal and cervical cells under a microscope while the latter involves macroscopic examination with a naked eye (or with a magnifier lens). A biopsy involves taking a small sample of tissue so that it can be examined under a microscope. While cytology and colposcopy cannot tell whether the abnormal cells are cancerous, a biopsy can provide this information. Yet, possible complications of a biopsy may include inflection and bleeding. In addition, biopsies may increase the risk for infertility and miscarriage as changes and scarring in the cervix may happen from the procedure. Therefore, despite of its higher accuracy, a biopsy may not be suggested without considering the risk level of the patients. This study focuses on four screening methods which are Hinselmann, Schiller, cytology and biopsy. Details of the methods are summarized in Table A1.

Table A1. Cervical cancer screening strategies

Screening strategy	Description
Hinselmann	Hinselmann is the colposcopy method using acetic acid. The colposcopy examination consists in the observation of the cervix tissues after the application of 5% acetic acid solution. The change of appearance of cervix tissues after the application of acetic acid improves the discriminability of cervical regions by a human expert, and precancerous lesions can be observed.
Schiller	Schiller is the colposcopy method using Lugol iodine. Using this method, the normal cervical regions stain and become mahogany brown or black while some abnormal patterns such as cervical polyps do not stain with iodine (Sellors & Sankaranarayanan, 2003). The Schiller's strategy may help in identifying lesions that could be overlooked during examination with acetic acid, thereby facilitating treatment.
Cytology	There are the conventional and liquid based cytology. The conventional cytology involves manual smearing and staining. In some cases, the uneven distribution of cells may induce dense regions where light cannot penetrate and empty regions of the slide (Bengtsson & Malm, 2014). Other artifacts such as blood may harm the effectiveness of this screening modality. On the other hand, liquid-based cytology preparations help uniformize the distribution of cells and dilute the presence of external factors (Fernandes et al., 2018). One common method of the liquid-based cytology is to submerge the brush with the cellular materials collected from the cervix in a liquid, followed by various ways for examination.
Biopsy	A cervical biopsy is a procedure to remove tissue from the cervix to test for abnormal or precancerous conditions, or cervical cancer. After the tissue sample has been removed, it can be tested using various chemicals to see how it responds and to find out what it contains. While the cytology and colposcopy cannot tell whether the abnormal cells are cancerous, a biopsy can provide this information.