CLUSTERING AS AN ARTIFICIAL INTELLIGENCE TECHNIQUE IN DRUG RESISTANCE OF HIV/AIDS PATIENTS: CASE STUDY BOTSWANA

Rajeswari Chandrasekaran, Dr.Audrey Masizana
University of Botswana
Gaborone, Botswana
kavisekar06@gmail.com, masizana@mopipi.ub.bw,

ABSTRACT
HIV/AIDS is a global pandemic that has affected national economies and devastated families. In Botswana, the HIV/AIDS situation is pronounced. The HIV/AIDS infection rate in the country is one of the highest in the world. Annually, Botswana spends a significant amount of its GDP on HIV/AIDS related challenges and initiatives. One such initiative is the Messiah ARV program, which provides free HIV/AIDS treatment for every citizen. While it is being credited as one of the first such initiatives in the world, the program is however faced with challenges. Key among them is the issue of drug resistance. Drug resistance access when existing HIV drugs—are not effective and do not stop the virus from multiplying. These kinds of characteristics are known as drug resistance. When HIV isn't fully controlled by ARV drugs, the virus makes copies of itself at a very rapid rate. Because this replication is occurring so fast, HIV often makes mistakes in the copies. If these "mistaken copies" are able to reproduce themselves, they are called mutations—which creates new forms of the virus. This concept paper presents an ongoing research which aims to investigate the challenge of drug resistance among HIV/AIDS patients in Botswana. The research will employ data mining techniques to monitor the CD Count, as it affects a patient HIV aids cell mutation over time. This will improve the doctor's decision in recommending the right therapy for patients at any particular given time. The research will confirm that data mining techniques have been used successfully in education and health sectors to assist in the detection of patterns in data, and prediction of the best approaches in addressing challenges. As such, it is seen as a useful tool to apply to this problem.

KEY WORDS
HIV/AIDS, Drug Resistance, CD count and Data Mining.

1. INTRODUCTION

Botswana, officially the Republic of Botswana (Tswana: Lefatshe la Botswana), is a landlocked country located in Southern Africa. The citizens refer to themselves as "Batswana". Formerly the British protectorate of Bechuanaland, Botswana adopted its new name after becoming independent within the Commonwealth on 30 September 1966. A mid-sized country of just over two million people, Botswana is one of the most sparsely populated countries in the world [1]. The geography is flat and up to 70% of Botswana is covered by the Kalahari Desert thus making it a sparsely populated country with little agricultural activity. The government system is a parliamentary republic; the chief of state and head of government is the President. Botswana has a mixed economic system in which the economy includes a variety of private freedom, combined with centralized economic planning and government regulation. Botswana is a member of the African Union (AU) and the African Economic Community (AEC) [2].

Botswana has more than four decades of uninterrupted civilian leadership, progressive social policies, and significant capital investment have created one of the most stable economies in Africa. Mineral extraction, which is dominated by diamond mining, tourism is also a growing sector due to the country's conservation practices and extensive nature preserves [3]. Botswana has one of the world's highest known rates of HIV/AIDS infection, but also one of Africa's most progressive and comprehensive programs for dealing with the disease [3].

The IMF report of 2014 [4] asserts that the country is making explicit strides in taking into account the effects of excess mortality due to AIDS, which result in lower life expectancy, higher infant mortality, higher death rates, lower population growth rates, and changes in the distribution of population by age and sex than would otherwise be expected. This Ministry of Health is the key stakeholder assigned to oversee these strides.

2. BOTSWANA HEALTH MINISTRY SYSTEM

The Ministry of Health in Botswana has the responsibility to provide leadership on health matters. It does this by formulating Health policies, ensuring their correct interpretation and implementation throughout the health care delivery system.

Some of the objectives of the ministry include [5]:

DOI: 10.2316/J.2014.216.815-0005

1. Coordinate and integrate long and medium term national health sector planning at interdepartmental, inter-ministerial and national level to achieve the overall national health objectives.

2. Provide health policy, planning and research expertise to all departments within the Ministry to ensure that health policies and plans are based on sound social and economic principles, reflect the identified health needs and priorities of the country, and are pragmatic in terms of technology applied and implementation strategies used.

3. Coordinate the provision of health management information systems to all departments within the Ministry and to other external users for purposes of effective health planning and evaluation, as well as health and operational management decision making.

4. Develop and review health financing policies and strategies.

5. Identify research needs, undertake, commission and coordinate research, examine and propose policy changes and new initiatives.

6. To provide leadership in all matters pertaining to public health services and ensures that the objectives of public health are effectively attained.

7. To facilitate the availability, interpretation and elaboration of relevant public health policies, including legislation, standards and regulations.

8. To develop and implement a system for planning, resource mobilization and coordination of public health services.

The country's national health financing is done through the National Health Accounts. A large proportion (Over 80% of Total Health Expenditure (THE) is provided by the Government. THE as a percentage of Gross Domestic product is expected to gradually increase to more than 10.54% by the year 2021. The international agencies also contribute modestly to health care in Botswana [6].

The adequate funding of the health sector has led to several initiatives by both the public and private sectors. Such initiatives contribute significantly to improve the health sector in Botswana. Those include the programs designed for HIV and Aids prevention and Control.

3. BOTSWANA HIV AND AIDS INITIATIVES

Several health initiatives are developed in Botswana to coordinate programs that will encourage good health practices. Programs that members involve themselves in include HIV/AIDS awareness, blood drives, prostate cancer awareness, and sickle cell anemia awareness programs. Some of the initiatives which are developed in Botswana are given below.

The AIDS situation has put a strain on the Botswana health system. As a response to the challenges posed by HIV/AIDS, the country has introduced a number of initiatives. The initiatives fall into two categories, those designed for prevention and creating awareness and those designed for providing treatment.

Baylor International pediatric AIDS Initiative (BIPAI)
BIPAI [7] Botswana started its adolescent clinic in 2005 with 23 teenagers. The adolescent population has continued to grow and by the end of 2010 included more than 600 teenagers. Based on age-stratification data collected from a representative sub-section of our client population, the average age of our pediatric patients is just over 9 years old.

Harvard School of Public Health AIDS Initiative (HAI)
[8] Is dedicated to research and education to end the AIDS epidemic in Africa and developing countries. For two decades, HAI has been at the forefront of HIV/AIDS laboratory research, clinical trials, education, and leadership. The Botswana–Harvard AIDS Institute Partnership (BHP) is a collaborative research and training initiative between the Government of Botswana and HAI.

Botswana-UPenn Partnership was established in 2004 Penn [9] and received funding from the US Public Health Service through PEPFAR (President's Emergency Plan for AIDS Relief), which is a program that provides US Public Health Service funding to 16 affected nations, including Botswana.

Merck Global Health Initiatives is a case series that focuses on Merck's drug donation program [10] and then raises new issues facing management about what to do about HIV/AIDS in Africa given the company's development of a new therapy. It also describes collaboration among many parties, including the Gates Foundation, other pharmaceutical companies, and the government of Botswana.

CIEE – Botswana The CIEE [11] is a Community Public Health program designed for students with an interest in public health issues in developing countries; the governmental, societal, and environmental factors that influence health care; and international health issues. Students learn about overcoming challenges in healthcare from the context of a developing nation.

Others include Community Home based Care, Safe Male Circumcision, Mother to Child treatment and the MASA ARV program which are publicly availed through the ministries information systems.

The most significant is the MASA program which was established as a treatment response, the government offers any citizen with HIV/AIDS free treatment at any public

hospital. The Government conducted a study in year 2000 to determine the macroeconomic impact of HIV/AIDS in the country (BIDPA, 2000). The study predicted devastating economic impact the epidemic would have on the lives of Batswana and the magnitude of human suffering AIDS would cause. These underpinned the decision in 2001 to provide ARV medication through the Botswana public health care system and gave birth to programs such as the Masa program.

Botswana was the first country in Africa to establish a national antiretroviral therapy program. "Masa," a Setswana word meaning "a new dawn," was the name given to the program to signify the hope that ARVs offer to people living with HIV and AIDS to live longer, healthier lives by providing them with time to nurture their families and build a future for the nation [12].

Antiretroviral therapy is a combination of drugs used to suppress the multiplication of HIV in the body, thereby allowing the body time to repair the immune system. In Botswana, HAART (highly active antiretroviral treatment) is offered, which is a combination of three or more ARV drugs. ARV therapy is not a cure for HIV; therefore, it is important for someone to take it in the right way, at the right time for the rest of his/her life, in order to benefit fully from the drugs. This is called good adherence. When the ARV therapy is taken correctly, most patients become healthier and even return to full productivity. Without ARV therapy, chances are very high that HIV-positive people will develop AIDS and die. [13]

ARV [14] drug types are named according to how they act to interrupt the HIV life cycle, thereby helping to reduce viral load levels. There are four types of anti-retroviral drugs used in Botswana. These are:

a) Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTI)
b) Nucleoside Reverse Transcriptase Inhibitors (NRTI)
c) Protease Inhibitors (PI)
d) Integrals Inhibitors

For each individual patient a well suited regimen has to be selected for successful viral suppression virus with fewer or less side effects and to prevent drug failure. Several factors are to be considered and used for selecting new regimen from existing regimens [15]. Providing first line regimen (initial) is not so difficult as standardized national guidelines are used. For example the first line regimen for:

– Adult men and women with no risk of pregnancy and children older than 5 years is (zidovudine + lamivudine + efavirenz)
– Pregnant women or women likely to become pregnant and children younger than 5 years is (zidovudine+ lamivudine + nevirapine)

The problem emerges when the first line regimen fails and the doctor has to select a second line regimen. Moreover, the doctor has to use the practice guidelines which are directed at a large group of patients and focus these on one patient at a time. He has to consider all factors influencing drug failure

Table 1. Some Recommended Regimes for ART (Adapted from Clercq and Vandamme, 2004)

NNRTI-based Regimens		**No. of Pills per day**
Regimens Preferred by the author	Efavirenz * (Nevirapine) + Lamivudine + (Zidovudine or Tenofovir)	3-5
	Efavirenz*(Nevirapine) + Lamivudine + (Stavudine or Didanosine)	3-5
PI-based Regimens		**No.of pills per day**
Regiments preferred by the author	Lopinar + Ritonavir + Lamivudine + (Zidovudine or Stavudine)	8-10
Alternative regimens	Amprenavir + Lamivudine + (Zidovudine or Stavudine)	12-14
	Indinavir + Lamivudine + (Zidovudine or Stavudine)	8-10
	Indinavir +Ritonavir + Lamivudine + (Zidovudine or Stavudine)	8-12
	Nelfinarvir + Lamivudine + (Zidovudine or Stavudine)	12-14
	Saquinavir + Ritonavir + Lamivudine + (Zidovudine or Stavudine)	14-16
Triple NRTI-based regimens		**No of pills per day**
Alternative regimen	Abacavir + Lamivudine + Zidovudine	2
***Except for pregnant women with child bearing potential**		

De Clercq and Vandamme [15] defines drug failure as a multifactorial phenomenon that occurs due to a number of reasons that gather simultaneously in a single patient. It depends on both favoring factors and limiting factors i.e.

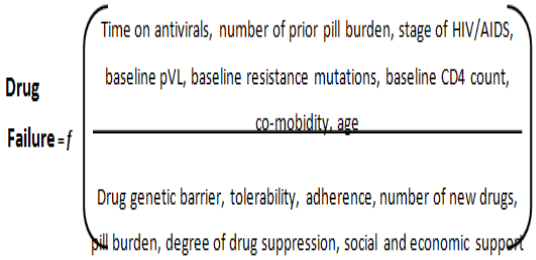


Figure 1. Drug failure as a function of favoring over limiting factors (adapted from Combination Therapy of Aids et al)

Likelihood of drug failure has to be considered comprehensively taking all factors into consideration simultaneously. The numerator consists of limiting or unmodifiable factors and the denominator is made up of favoring or controllable factors considered when planning a new ARV regimen. The numerator has clinical factors that the doctor cannot change even if he wanted to, so it is the denominator he has to work around. It is therefore the task of the doctor or clinician to design a drug combination or a regimen that seeks [15]:

- The lowest pill burden
- The higher genetic barrier
- The higher adherence to treatment
- The higher number of new drugs included
- And the higher degree of viral suppression attainable

These factors are compared and used to select a second line regimen for a specific patient with the goal achieving best treatment based on these competing factors. We can hence conclude that selecting a specific regimen for an individual patient is complex with multiple attributes that the doctors need to select to consider.

4. CD COUNT AND DRUG INTERVENTION

In order to decide whether or not an HIV positive patient should start treatment, the clinical tests that need to be carried out will determine the stage of HIV infection and the individual patient's readiness for drug treatment, which may depend on a variety of factors such as the patient's medical history. One such factor is the CD count level.

For the HIV Drug intervention to be successful, there should be consistent CD count monitoring, especially for those who are not yet eligible for the ARV treatment. The "c" and the "d" in CD stands for "cluster of differentiation," and refers to the cluster of proteins that make up a cell surface receptor. There are dozens of different types of clusters, but CD4s and CD8s are the ones talked about most often [16]. CD4 cells are the most important cells in the immune system. A type of test called a CD4 count tells you how many CD4 cells, there are in a small sample of blood. The more there are, the better. Decisions about when to start treatment are often based on the CD4 count [17]. CD4 cells are responsible for signaling other immune system cells to fight an infection in the body. Too few CD4 cells means that the immune system will no longer functions like it is supposed to [18].

The normal CD4 count is somewhere between 500 and 1500 cells per cubic millimeter of blood (a drop, more or less). In the absence of anti-HIV treatment, the CD4 cell count decreases, on average, about 50 to 100 cells each year. A large number of other infections can occur if it drops below 50 to 100 cells. AIDS-related diseases (opportunistic infections) such as Pneumocystis jiroveci pneumonia (PCP) can occur if your CD4 count falls below 200 [19]. HIV can destroy entire "families" of CD4 cells. Then the diseases these "families" were designed to fight can easily take over.

Acquired immunodeficiency syndrome (AIDS) is the most severe form of HIV infection. HIV infection is considered to be AIDS when at least one serious complicating illness develops or the number (count) of CD4+ lymphocytes decreases substantially

Hence it can be concluded monitoring trends in changes to CD4 cell count over time is very important to determine the patient's status and susceptibility. Hence the determination to contribute towards methods that are employed to monitor CD Count trends in any given patient. Hence the proposal to use a data mining technique to uncover patterns that emerge during the cell mutations as affected by HIV.

5. DATA MINING

Data mining (knowledge discovery in databases) is the process of extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases [20]. Data mining has been applied with success to different fields of human endeavor, including marketing, banking, customer relationship management, engineering and various areas of science.

Data Mining is a multidisciplinary field, drawing work from areas including database management systems, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing, and data visualization.

The process of performing data analysis may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical

research. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions on future data.

To achieve these goals, data mining solutions employ a wide variety of techniques of machine learning, artificial intelligence, statistics, and database query processing. These algorithms are also based on mathematical approaches such as

5.1 Decision trees

Decision trees are tree shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset[21]. A specific decision tree methods include Classification and Regression Trees. CART and Chi Square Automatic Interaction Detection (CHAID) are decision tree techniques used for classification of a data set. They provide a set of rules that can be applied to new (unclassified) dataset to predict which records will have a given outcome. CART segments a data set by creating two-way splits while CHAID segments datasets using the chi squared test to create multi-way splits. CART typically requires less data preparation than CHAID.

5.2 Genetic algorithms

These are optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution [21]. These algorithms solve problems by borrowing a technique from nature. GAs use Darwin's basic principles of survival of the fittest, mutation, and crossover to create solutions for problems. When a GA finds a good solution, it percolates some of that solution's features into a population of competing solutions. Over time, the GA "breeds" good solutions. Genetic Algorithms are optimization techniques. They are used for Classification. Another important use for them is in finding the best possible combination of link weights for a given neural network architecture [22].

5.3 Nearest neighbor method

This is a technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. It is sometimes called the k-nearest neighbor technique. [21] The K-Nearest Neighbor method is a Predictive technique. In order to predict what a prediction value is in one record, look for records with similar predictor values in the historical database, and use the prediction value from the record that is "nearest" to the unclassified record. In other words, it performs prediction by finding the prediction value of records similar to the record to be predicted. The data used by the K-nearest algorithm is numeric [22].

5.4 Clustering algorithms

Clusters are data items which have been grouped according to logical relationships or consumer preferences. Clustering algorithms are used to cluster or segment data. They include K-Means, BFR Algorithm, BIRCH Algorithm, CURT Algorithm, Chamelon Algorithm, Incremental Clustering, DBSCAN Algorithm, OPTICS Algorithm, DENCLUE Algorithm, Fast Map Algorithm, GRGPF Algorithm, STING Algorithm, Wave Cluster Algorithm, CLIQUE Algorithm, and COBWEB Algorithm [22].

5.5 Associations algorithms

Association in Data mining refers to when a simple correlation is made between two or more items, often of the same type, to identify patterns. Association algorithms are therefore used to solve Association Models. Association Algorithms include The Apriori Algorithm, PCY Algorithm, Iceberg Algorithm, AIS Algorithm, STEM Algorithm, AprioriHybird Algorithm, Toivonen Algorithm, and Frequent Pattern Growth Algorithm [22].

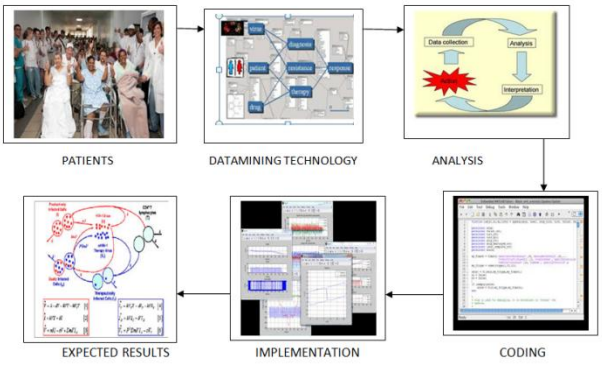
Over the last few years, the term 'data mining' has been increasingly used in the medical literature. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain

The discovered knowledge can be used by the health care administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives [23] .

In recent times, researchers in the field of Drug Resistance have developed new data mining techniques to use in this field.

Predictive data mining methods such as clustering may be applied to the construction of decision models for procedures such as prognosis, diagnosis and treatment planning, which once evaluated and verified–may be embedded within clinical information systems [24]. Hence the research proposes to support this by using the Clustering technique to uncover previously unknown patterns in the HIV mutation process.

6. PROPOSED SYSTEM: (CONCEPTUAL DESIGN)



The main aim of this research is to predict an accurate drug resistance level of the HIV/AIDS infected patients

and contribute to the existing work. Here, we intend to propose a drug resistance prediction technique to find the rate of drug resistance to the patients in relation their CD4 cell counts.

The proposed technique will collect the patient's medical information. Thus the collected data will be given to the feature extraction process which extracts the patient's data that will be utilized in the drug resistance prediction process. Based on the extracted features the data will be clustered by the K-means clustering algorithm. Based on the clustered results the drug resistance level will be predicted by the statistical methods. After that the CD4 cell counts will be predicted by the ANN technique.

The patient's drug resistance level and their corresponding CD4 cell counts will be given to the ANN for the training process, thus the well trained ANN will be utilized to find the patients future CD4 cell counts in their immune system. Hence, the proposed technique will successfully predict the patient's drug resistance level and their CD4 cell counts via clustering and ANN. The technique will be implemented in the working platform of MATLAB and the results will be analyzed to demonstrate the performance of the proposed drug resistance prediction technique. This whole pattern recognition process will be labeled as a data mining process.

7. CONCLUSION

The main purpose of the project was to investigate an analytic way of supporting doctors in determining the best ARV treatment for patients given a number of available alternative regimens and the large number of patients enrolled in these programs. It involved the use of Botswana HIV/AIDS and ARV therapy as a case study. It was concluded that the selection process was enormous due to the number of patients and the number of available combinations and criteria used in the selection process.

One of the major contributing factors is the CD count level, which could contribute to drug resistance if the CD count is lower the standards set by the medical field. Hence the need to monitor its level over time for patients to be provided with the right therapy.

Research has shown that electronic information systems are continually being proposed as one means of reducing medical errors of commission (doing the wrong thing) and omission (not providing indicated care). The research proposes the use of data mining to study a patient's mutation process in order to improve the decision making for the right therapy.

REFERENCES

[1] Botswana History http://en.wikipedia.org/wiki/Botswana retrieved on August 2013.

[2] Botswana Statistics and global business knowledge http://globaledge.msu.edu/countries/botswana retrieved on August 2013.

[3] Central Intelligence Agency World Factbok https://www.cia.gov/library/publications/the-world-factbook/geos/bc.html retrieved on August 2013.

[4] Data from the international monetary fund http://www.indexmundi.com/cote_d_ivoire/population.html retrieved on February 2014.

[5] http://www.moh.gov.bw/index.php?option=com_content1&id=2 . Retrieved on 6th March 2014

[6] Country cooperation strategy at a glance, 2009

[7]Programs for Botswana's HIV infected – adolescent http://www.bipai.org/Botswana/adolescent-programs.aspx. Retrieved on March 2014.

[8] Harvard school of public health AIDS initiatives http://www.aids.harvard.edu. Retrieved on March 2014.

[9] BUP Health care provision. http://www.med.upenn.edu/botswana/BUPProgramsClinicalCare.shtml. Retrieved on April 2014.

[10] Merck Global Health Initiatives (B): Botswana By Jamees E.Austin, Diana Barrett, James Weber Source: Harvard Business School 19 pages. Publication Date: Jan 26, 2001. Prod.#: 301089-PDF-ENG

http://hbr.org/product/a/an/301089-PDF-ENG?cm_sp=doi-_-case-_-301089-PDF-ENG&referral=00103. Retrieved on April 2014.

[11] CIEE community public health program. http://global.arizona.edu/study-abroad/program/ciee-botswana-summer. Retrieved on April 2014.

[12] Public-Private Partnerships And Antiretroviral Drugs For HIV/AIDS: Lessons From Botswanahttp://content.healthaffairs.org/content/24/2/545.full

[13] HIV and AIDS in Botswana http://www.avert.org/hiv-aids-botswana.htm

[14] Botswana HIV/AIDS Programs. http://www.hiv.gov.bw/content/anti-retroviral-drugs-arv. Retrieved on April 2014.

[15] De Clercq E. D. A., and Vandamme A. I., (2004) Combination Therapy of AIDS, Birkhauser Verlag, Basel , Switzerland

[16] CD 4 cell test with the result. http://www.aidsmeds.com/articles/TCellTest_4727.shtml. Retrieved on April 2014.

[17] CD 4 count news. http://www.aidsmap.com/CD4-count/cat/1656/ Retrieved on April 2014.

[18] Megan Michelle Pinkston, "Increasing Empathy in Medical School Students towards Individuals Living with HIV/AIDS and medication adherence", 2008

[19] Malik Adeel Umer, "Human_Physiology", pp. 1-432

[20] Hemlata Sahu, Shalini Shrma and Seema Gondhalakar, "A Brief Overview on Data Mining Survey", International Journal of Computer Technology and Electronics Engineering (IJCTEE), Vol. 1, No. 3, pp. 114-121, 2011

[21] Data Mining, CSCI-453, Dr.Khalil Research paper, Presented by Tarek El-Gally and Ahmed Gamad el-Din.

[22] Data Mining Presentation by Robert James, Eastern Michigan University. http://www.emich.edu/ia/pdf/research/Data%20mining%20use%20in%20Intrusion%20detection,%20Robert%20James.pdf. Retrieved on February 2014.

[23] Satbir Jain and Shweta, "Cluster based mining of HIV drugs in Chemoinformatics", pp. 1-10

[24] Riccardo Bellazzi and Blaz Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines", international journal of medical informatics, Vol. 77, pp. 81-97, 2008