Scientific Research and Essays Vol. 5(17), pp. 2384-2390, 4 September, 2010
Available online at http://www.academicjournals.org/SRE
ISSN 1992-2248 ©2010 Academic Journals

Full Length Research Paper

Support vector machines to forecast changes in CD_4 count of HIV-1 positive patients

Yashik Singh* and Maurice Mars

[1]Department of Telehealth, Nelson R Mandela School of Medicine, Durban, South Africa.
[2]School of Information Systems and Technology, University of Kwa-Zulu Natal, South Africa.

Accepted 18 February, 2010

There are currently 5.5 million confirmed cases of HIV/AIDS in South Africa. HIV infection can be effectively managed with antiretroviral (ARV) drugs, but close monitoring of the progression of the disease is vital. One of the better surrogate markers for disease progression is the use of CD_4 cell counts. Forecasting CD_4 cell count will help clinicians with treatment management and resource allocation. The aim of this paper was to investigate the application of machine learning to predict future CD_4 count change. A support vector machine classification model that predicted the degree of CD_4 count change was built. The model took as input the genome, current viral load and number of weeks from baseline CD_4 count and predicted the range of CD_4 count change. The model produced an accuracy of 83%. This pilot project shows that a change in CD_4 count may be accurately predicted using machine learning on genotype, viral load and time. Clinical studies to validate this are required. The aim of this paper was to investigate the application of machine learning to predict future CD4 count change. A Support Vector Machine classification model that predicted the degree of CD4 count change was built. The model took as input the genome, current viral load, and number of weeks from baseline CD4 count and predicted the range of CD4 count change.

Key words: HIV, e-health, CD_4 lymphocyte count, forecasting.

INTRODUCTION

In 2008, there were between 30 and 36 million HIV positive patients around the world, two-thirds of whom are in sub-Saharan Africa. This number has steadily increased due to the high incidence rate of HIV. HIV/AIDS is the leading cause of death in sub-Saharan Africa (Campbell et al., 2008) and is currently the fastest growing epidemic in South Africa (Simon-Meyer and Adallo, 2002) with 5.5 million confirmed cases of HIV/AIDS (Giarelli and Jacabs, 200; Rispel and Metcalf, 2009).

HIV infection can be effectively managed with antiretroviral (ARV) drugs but close monitoring of the disease progression is vital. Laboratory (Fahey et al., 1990; Moss et al., 1988) and clinical (Cahn et al., 1991; Montaner et al., 1992) markers of disease progression are used to monitor HIV infection. Common laboratory tests include viral load (measurement of HIV nucleic acid concentration), CD_4 lymphocyte counts or CD_4/CD_8 ratio (quantitative measurement of particular lymphocyte cells that indicate the strength of the immune system), HIV p24 antigenaemia (p24 antigen count increases one to three weeks after initial infection and is present before HIV antibodies are formed), erythrocyte sedimentation rate (a non-specific marker of inflammation) and serum-β-2-microglobulin (high concentrations are a good predictor of AIDS) (Morfeldt-maringnson et al., 1991). Common clinical markers of disease progression are weight loss, mucocutaneous manifestations, bacterial infections, chronic fever, chronic diarrhoea, herpes zoster, oral candidiasis and pulmonary tuberculosis (Morgan et al., 2002).

One of the best available surrogate markers for HIV progression is the use of CD_4 lymphocyte cell counts

*Corresponding author. E-mail: singhy@ukzn.ac.za. Tel: +27829408032. Fax: +27312604737.

(Post et al., 1996; Schechter et al., 1994). CD_4 count is recognized as a standard measure of immunodeficiency in HIV positive patients in developed countries and areas where there is a low rate of HIV infection (Mwamburi et al., 2005). Determining a patient's CD_4 count is important in CD_4–guided treatment of HIV (Osmond et al., 1993). It thus follows that the ability to forecast CD_4 count changes would be valuable to physicians. If one can predict the short-term change of the CD_4 count, a physician may change treatment in order to take steps to prevent opportunistic infection such as pneumocystis pneumonia and delay the onset of AIDS (CD_4 count < 200).

Although the use of CD_4 count is part of the standard of care in developing countries, the measurement of CD_4 count requires many complex and expensive flow cytometric procedures which burden the minimal resources available (Schechter et al., 1994). Knowing possible changes to CD_4 count may help with resource allocation. There have been previous attempts to measure current CD_4 count using cheaper chemical assays and even correlating a patient's total lymphocyte count (TLC) with current CD_4 counts using logistic and linear regression (Schechter et al., 1994; Osmond et al., 1993). In developing countries, TLC may be used to make treatment decisions for symptomatic patients when CD_4 count is not available (Daka and Loha, 2008).

Machine learning may be used to develop a model that predicts virological response. Machine learning is an artificial intelligence computer science technique that tries to find a mathematical model that map between inputs and outputs of a domain problem. There are two stages to applying machine learning techniques: creating the mathematical model by learning mappings between given input and output and using the model to predict an output, given unseen input. (Wang et al., 2002) developed a prototype neural network machine-learning algorithm on a small dataset and predicated the viral load of HIV patients from genotype and treatment data with 75% accuracy. Larder et al. (2007) later created a genotype and treatment history based input neural network model that predicted viral load with 69% accuracy. Altman et al., 2008 created a machine-learning algorithm that predicts a dichotomized virological response, that is, success or failure of therapy with 80% success. This was later changed by predicting the probability of treatment success (Altman et al., 2009) based on a degree of predicted HIV drug resistance. There has however been no chemical test or computer model developed to forecast changes to the CD_4 count.

Objective

The aim of this pilot study was to apply a machine-learning technique to investigate the feasibility of forecasting CD_4 count change. The objective was to produce a mathematical model that can predict the range of change of an individual HIV-1 positive patient's CD_4 count.

METHODOLOGY

Dataset

Datasets were obtained from the Stanford HIV drug resistance database (http://hivdb.stanford.edu/), which is publically available and contains data from clinical trials. The ACTG333, ACTG384 and GART clinical trial data was obtained and the separate protease genome sequences, CD_4 counts and viral loads were extracted.

Protease is a viral enzyme that facilitates the polyprotein precursor cleavage during HIV transmission. Protease inhibition is one form of HIV treatment.

Pre-processing

Subtype B consensus protease (PR) genome sequences, CD_4 count, viral load and the number of weeks from the baseline measure of CD_4 count for each patient sample was determined by joining individual datasets using the sample identifier (the unique number that identifies a sample) and date. The protease dataset consisted of approximately 3000 data elements. Each protease sequence consisted of 99 amino acids from position 1 to 99. The sequences were processed such that at any position a 1 represented an amino acid mutation and a 0 represented no mutation.

The PR datasets were further processed to reduce the number of amino acid positions that would be used for learning and testing. The number of mutations that occurred in the entire dataset at each position was calculated. All positions where the total number of mutations were less than 20%, were removed. The 20% threshold cut-off value was chosen after running prediction accuracy tests at 5, 10 and 20%. Highest accuracies were obtained with 20%. Literature reports the use of a 5% threshold (Rabinowitz et al., 2006) and this was used to counter the possibility of random mutations in order to unbiasly determine the effect of each mutation with respect to drug resistance. Subsequently, it has been documented that a small percentage of mutations affect ARV therapy (Rhee et al., 2006). Thus using a higher threshold cut-off value will not adversely affect the forecasting model. The 20% threshold cut-off value resulted in the PR dataset having genome sequences with a total number of 25 amino acids, at positions 10, 12, 13, 14, 15, 19, 23, 24, 30, 32, 33, 46, 47, 48, 50, 53, 54, 69, 74, 76, 82, 84, 88, 90 and 93. As expected, the mutations at these positions are important precursors to PI resistance.

Data of Patient's genome sequences and associated viral load and CD_4 count data at different time points were extracted. These time points ranged from 1 to 183 weeks apart with an average of 32 weeks. This produced 540 data elements that consisted of repeat patients with viral load, CD_4 count and genome sequences taken at different times. The distribution of the times that the viral load, CD_4 count and genome sequences were measured are shown in Figure 1. This dataset was further processed to determine the change in CD_4 count between these measurements. 65 random data elements were removed from each dataset and formed a testing set. Training was done on the remaining data elements of the protease datasets.

Input of the machine learning algorithms

Three different groups of inputs where created and each was feed into the machine learning algorithm separately, forming three models. These input groups were: Input1, consisted only of genome sequence; Input 2, consisted of genome sequence and current viral

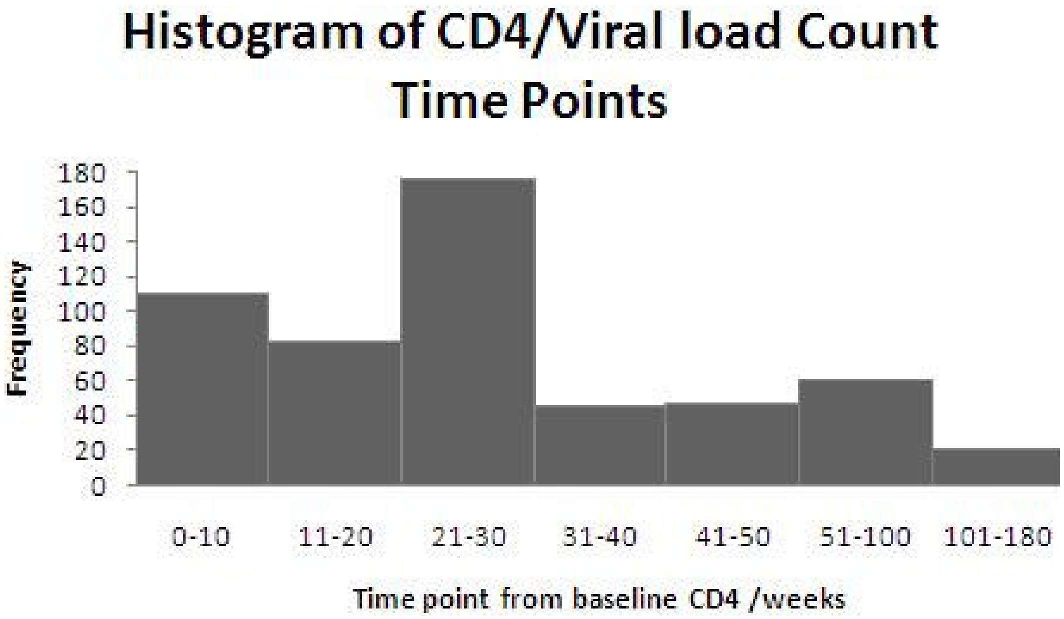

Figure 1. Frequency of the number of weeks from baseline that CD4 counts and viral loads were taken.

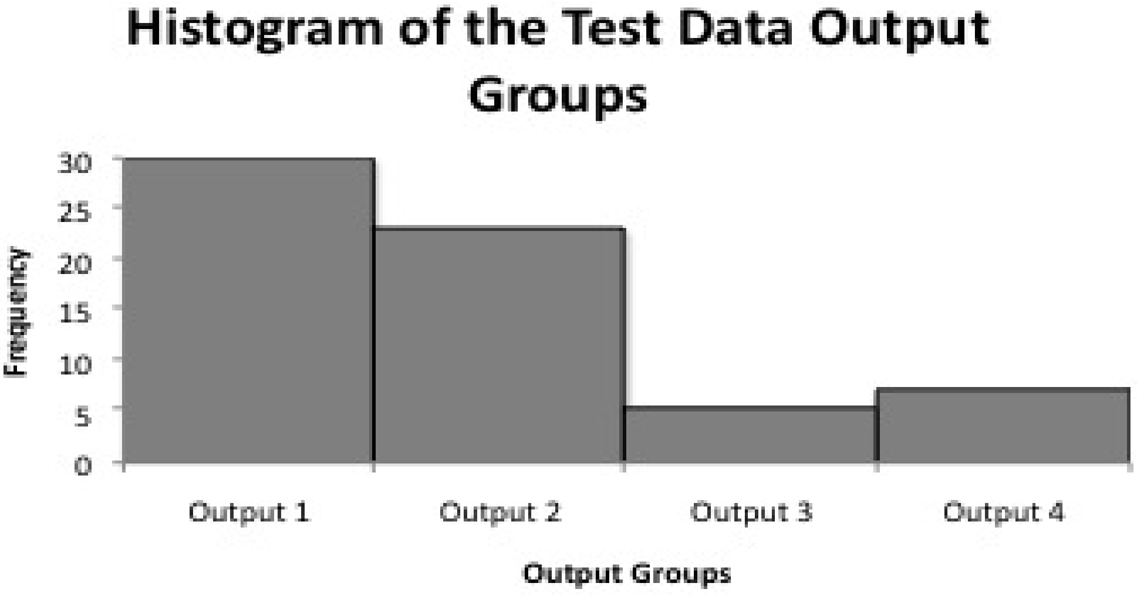

Figure 2. Frequency of the output groups for the testing dataset.

load; Input3, consisted of genome sequence, current viral load and number of weeks from the current CD_4 count to baseline CD_4 count.

Output of the machine learning algorithms

A classification model was built based on the changes (Δ) in CD_4 count. The changes in CD_4 count were grouped into four categories as shown in Equation 1. The distribution of the output groups for the training and testing set is shown in Figures 2 and 3.

$$Classification = \left\{ \begin{array}{ll} Output1, & \Delta CD4 < 0 \\ Output2, & 0 \le \Delta CD4 \le 50 \\ Output3, & 51 \le \Delta CD4 \le 100 \\ Output4 & \Delta CD4 > 100 \end{array} \right\} \quad (1)$$

Technology description

A statistical technique called support vector machines was used to

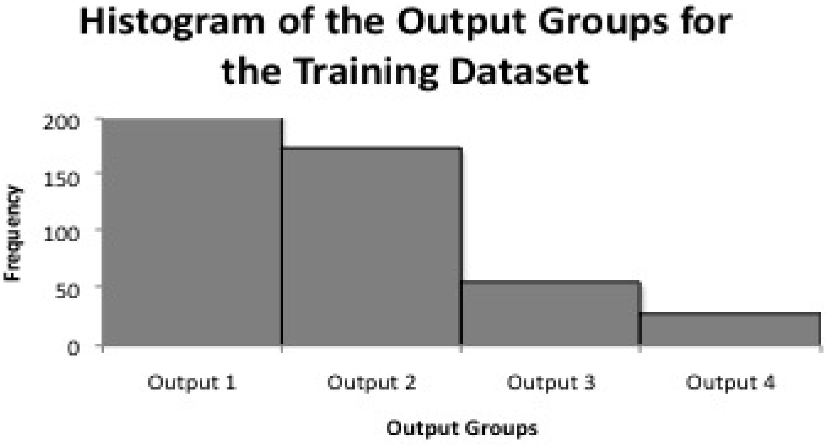

Figure 3. Frequency of the output groups for the training dataset.

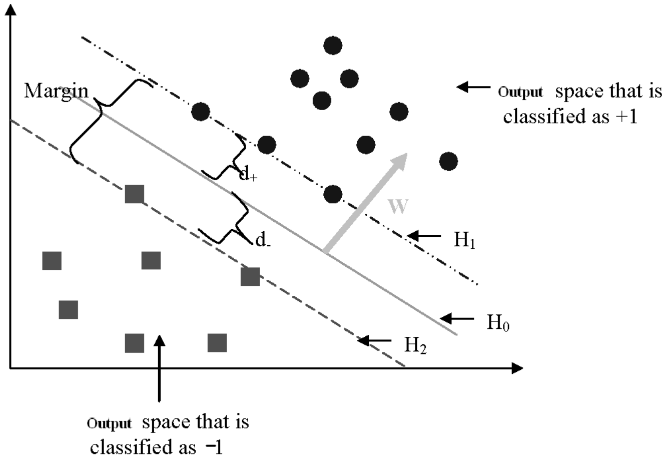
Margin
d+
d-
W
Output space that is classified as +1
H1
H0
H2
Output space that is classified as −1

Figure 4. Graphical representation of SVMs.

find a mapping between the three input groups and the four output groups. Support vector machines (SVMs) were chosen due to their ability to learn well with high dimension inputs and because they are considered to be very robust (Daka and Loha, 2008; Navia-Vazquez, 2007; Bugers, 1998).

Support vector machines

SVMs work by embedding data into a higher dimensional vector space and then attempt to find linear relations in that space. SVMs have been described as having the properties of duality, ability to incorporate kernels, margin maximization, convexity and sparseness. The simplest possible SVM is one that represents data that is linearly separable. This means that there must exist a separating hyperplane (H0) that completely separates the input space from its output space. This separating boundary divides the input space in such a way that all input space elements that lie on one side of the boundary have the output space value of +1, while those on the opposite side have the value -1. This separation is shown in Figure 4.

SVMs try to maximize the margin between (H0) and (H1), that is, ||W||. After incorporating the Wolf dual, Lagrangian multipliers and kernels (K() which are used to convert a non-linear search space into a linear one by increasing dimensionality), SVMs reduce these to a single equation (Equation 2). SVM perform multiple-class classification by a 'one-against-one technique', which applies the model to binary sub-classifiers and then determines the correct class by means of a voting system.

$$L_{dual} = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{N} \alpha_i \alpha_j \gamma_i \gamma_j \mathrm{K}(X_i, \mathrm{X}_j) \tag{2}$$

$$W = \sum_{i=1}^{N} \alpha_i \gamma_i X_i$$

$$\sum_{i=1}^{N} \alpha_i \gamma_i = 0$$

$$0 \leq \alpha_i \leq \zeta$$

Where, ζ is a user defined parameter that assigns the importance of an error, α_i are Lagrangian multipliers, X_i is the input space and y_i is the output space for i = 1 … number of data elements.

LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used to create the SVM models in this study. Linear, quadratic (polynomial with degree two) and radial base function (RBF) kernels were used. The radial base function kernel cost of 3 and gamma of 0.2, and the polynomial kernel cost of 10, co-efficient of 1 and gamma of 1 was determined by a coarse grid search. The default linear kernel was used.

RESULTS

The accuracy of the machine learning models is shown in Table 1. Analysis was performed on the raw results of the algorithms. In an attempt to further assess the predictive capabilities of the algorithms, each predicted and desired

Table 1. The accuracies of the different models.

Input space model	RBF	Quadratic	Linear
Input 1	66	66	66
Input 2	68	68	66
Input 3	83	72	71

Table 2. The desired and predicted output pairs for the different discrepancy groups.

None	Low	Intermediate	High
Output 1-1	Output 1-2	Output 1-3	Output 1-4
Output 2-2	Output 2-3	Output 2-4	
Output 3-3	Output 3-4		
Output 4-4			

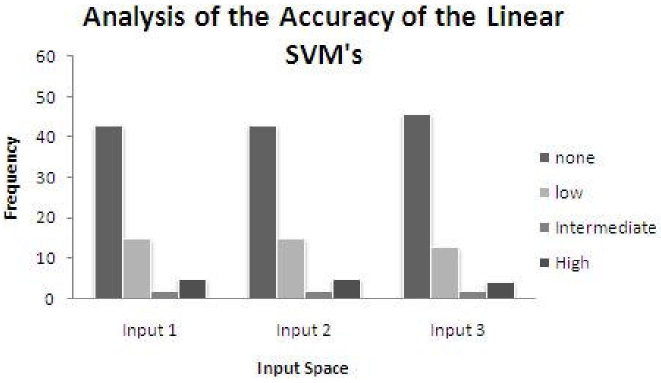

Figure 5. Discrepancy group distribution for linear SVM models.

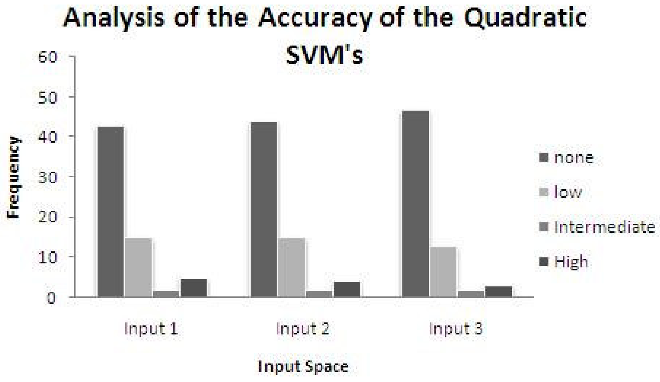

Figure 6. Discrepancy group distribution for the quadratic SVM models.

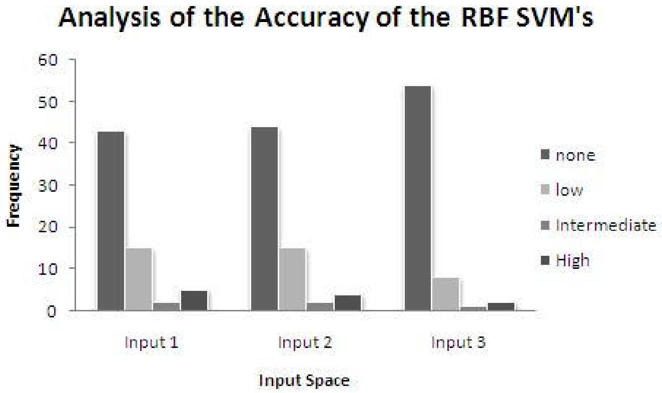

Figure 7. Discrepancy group distribution for the RBF SVM models.

output of the test dataset was compared to find the extent of discrepancies. The desired and predicted output pairs were divided into the categories of none, low, intermediate and high discrepancies defined in Table 2 where output X-Y indicates that output group X was desired but output group Y was predicted or vice versa. Results of this analysis are shown in Figures 5, 6 and 7.

DISCUSSION

Results shown in Table 1 indicate that for the RBF, linear and quadratic SVM's, there are no differences between the Input 1 and 2 models, but there are differences between Input 1 and 3 models as well as the Input 2 and 3 models. The Input 3 model is more accurate than the other two models. This result was expected due to the fact that the longer a patient is on effective ARV therapy, the more the immune system reconstitutes, resulting in a higher CD_4 count. Thus, the time component is a valuable predictor.

There is no difference between the quadratic and linear SVM algorithms as shown in Table 1, while the RBF model outperforms both the quadratic and linear SVMs with Input 3. It was unlikely that the data would be linearly separable hence the poorer performance of the linear SVM models. The superior performance of the RBF kernel is due to its localized and finite responses across the entire range of predictors (Cherkassky and Ma, 2004).

The ideal distribution of discrepancies should be a high frequency for the no (none) discrepancy group and a progressively lower frequency for the low, intermediate and high groups, respectively. The frequency of the high discrepancy group is greater than the intermediate group for the RBF SVM model with Input 1 and the distributions of the discrepancy groups are similar for Input 1 and 2 (Figure 7). Input 3 for the RBF SVM, has a higher frequency for the no (none) discrepancy group as compared to Input 1 and 2. There are fewer high, intermediate and low discrepancies between the desired and predicted outputs of the test dataset when using the Input 3 model as compared to Input 1 and 2 for the RBF models.

The distributions of the discrepancy groups are similar for the quadratic and linear SVMs as seen in Figures 5 and 6. This again indicates that there is no difference in the accuracy of these two models with respect to the Input Space.

Limitations

This pilot study shows that it is possible to forecast changes to CD_4 count. There are however several limitations to this study:

1. The most important is the low number of data elements used for the machine learning and testing. A much larger dataset is required to create a machine learning algorithm that can better predict CD_4 count changes. There is at present no dataset available in the public domain larger than that used.
2. The misclassification of the machine learning algorithm may be due to the fact that only the protease genome sequence was used. In order to create a more accurate and encompassing algorithm, a future study should include the reverse transcriptase genome sequences.
3. It is acknowledged that genome sequencing and determining viral load is costly for developing countries. The input space should be expanded to include models that use more common treatment information and less laboratory tests results.
4. More machine learning algorithms should be tested such as committee networks, neural networks, random forests and boosting which have been shown to learn medical related data well.
5. A regression model should be built to try and predict the actual CD_4 count as opposed to a range predicated in this pilot study.

Conclusion

This pilot study shows that it is possible to mathematically forecast a change in CD_4 count using SVM's. The best accuracy achieved was 83% using genome, current CD_4 count and number of weeks from baseline CD_4 count. The remaining 17% of misclassified CD_4 count changes are distributed such that the majority of the discrepancies are one category/CD_4 change group apart.

This pilot study forms part of a larger study to create a web-based HIV resistance portal. It is envisioned that this portal will be used to guide treatment of complicated HIV resistant patients by prompting clinicians to enter genomic, virological and treatment history data and then providing them with information about the specific patient's current resistance profile, future resistance profiles, the effect of changes in treatment and the prediction of the onset of AIDS, opportunistic diseases and mortality. Forecasting the CD_4 count of a patient from genotypic information is thus vital to the creation of the resistance portal, which will guide the clinician in determining the optimal therapy for individual patients.

REFERENCES

Altman A, Däumer M, Beerenwinkel N, Peres Y, Schülter E, Büch J, Rhee SJ, Sönnerborg A, Fessel WJ, Shafer RW, Zazzi M, Kaiser R, Lengauer T (2009). Predicting the Response to Combination Antiretroviral Therapy: Retrospective Validation of geno2pheno-THEO on a Large Clinical Database, JID. 199: 999-1006.

Altmann A, Rosen-Zvi M, Prosperi M, Aharoni E, Neuvirth N, Schülter E, Büch J, Struck D, Peres Y, Incardona F, Sönnerborg A, Kaiser R, Zazzi M, Lengauer T (2008). Comparison of Classifier Fusion Methods for Predicting Response to Anti HIV-1 Therapy. PLoS ONE 3(10): e3470.

Bugers CJC (1998). A tutorial on support vector machines for pattern recognition, Data mining and Knowledge discovery 2:121-167.

Cahn P, Perez H, Casiro A, Crinberg N, Muchinik G (1991). Progression of HIV-disease: the Buenos Aires cohort study. International conference on AIDS. 8: 157 (Abstract no. PuC 8029).

Campbell C, Nair Y, Maimane S, and Sibiya Z (2008). Supporting people with aids and their careers in rural South Africa: Possibilities and challenges, Health and Place 14: 507-518

Cherkassky V, Ma Y (2004). Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw. 17(1): 113-126.

Daka D, Loha E (2008). Relationship between Total Lymphocyte count (TLC) and CD4 count among peoples living with HIV, Southern Ethiopia: a retrospective evaluation. AIDS Res Ther. 5: 26.

Fahey JL, Taylor JM, Detels R, Hofmann B, Melmed R, Nishanian P, Giorgi JV (1990). The prognostic value of cellular and serologic markers in infection with human immunodeficiency virus type 1. N. Engl. J. Med. 322: 166-172.

Giarelli E, Jacabs AL (2002). HIV/AIDS care in Kwazulu-Natal, South Africa: An interview with Dr. Leana Uys, JANAC, 11(4): 52-67.

Larder B, Wang D, Revell A, Montaner J, Harrigan R, De Wolf F, Lange J, Wegner S, Ruiz L, Pérez-Elías MJ, Emery S, Gatell J, Monforte AD, Torti C, Zazzi M, Lane C (2007). The development of artificial neural networks to predict virological response to combination HIV therapy. Antivir. Ther. 12(1): 15-24.

Montaner JSG, Le TN, Le N, Craib KJP, Schechter MT (1992). Application of the World Health Organisation system for HIV infection in a cohort of homosexual men in developing a prognostically meaningful staging system. AIDS, 6: 719-724.

Morfeldt-maringnson L, Boumlttiger B, Nilsson B, von Stedingk L (1991). Clinical signs and laboratory markers in predicting progression to AIDS in HIV-1 infected patients, Scand J. Infect. Dis., 23(4): 443-449.

Morgan D, Mahe C, Mayanja C, Whitworth JAG, Kilmarx PH (2002). Progression to symptomatic disease in people infected with HIV-1 in rural Uganda: prospective cohort study. BMJ; 324: 193-197.

Moss AR, Bacchetti P, Osmond D, Krampf W, Chaisson RE, Stites D, Wilber J, Allain JP, Carlson J (1988). Seropositivity for HIV and the development of AIDS and AIDS related condition: three year follow up of the San Francisco general hospital cohort. Br. Med. J., 296: 745-750.

Mwamburi DM, Ghosh M, Fauntleroy J, Gorbach SL, Wanke CA (2005). Predicting CD4 count using total lymphocyte count: A sustainable tool for clinical decisions during HAART use. Am. J. Trop. Med. Hyg., 73(1): 58-62.

Navia-Vazquez Z (2007). Support vector perception, Neurcomputing, 70: 1089-1095.

Osmond D, Charlebois E, Lang W, Shiboski S, Moss A (1993). Factors affecting changes in time from CD4 = 200 to death in two San Francisco cohorts, 1983-1992, Int. Conf. AIDS. 9: 656 (abstract no. PO-C04-2632).

Post FA, Wood R, Maartens G (1996). CD4 and total lymphocyte counts as predictors of HIV disease progression. Q. J. Med., 89: 505-508.

Rabinowitz M, Myers L, Banjevic M, Chan A (2006). Sweetkind-Singer J. Haberer J. McCann K.

and Wolkowicz R., Accurate prediction of hiv-1 drug response from the reverse transcrip-tase and protease amino acid sequences using sparse models created by convex optimization, Bioinfo, 22(5): 541-549.

Rhee SY, Taylor J, Wadhera G, Ben-Hur A, Brutlag DL, Shafer RW (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. Proc. Acad. Sci. USA.

Rispel CL, Metcalf AC (2009). Breaking the silence: South African HIV policies and the needs of men who have sex with men, Reprod. Health, 33: 133-142.

Schechter M, Zajdenverg R, Machado LL, Pinto ME, Lima LA, Perez MA (1994). Predicting CD4 Counts in HIV-infected Brazilian Individuals: A Model Based on the World Health Organization Staging System. J. Acquir. Immune Defic. Syndr., 7(2): 163-168.

Simon-Meyer J, Odallo D (2002). Greater involvement of people living with HIV/AIDS in South Africa, Eval. Prog. Plan, 25: 3850-3855.

Wang D, DeGruttola V, Hammer S, Harrigan R, Larder B, Wegner S, Winslow S, Zazzi M (2002). A Collaborative HIV Resistance Response Database Initiative: Predicting Virological Response Using Neural Network Models. Poster presentation at: The XI International HIV Drug Resistance Workshop, Seville.