MACHINE LEARNING MODELS FOR ASSESSMENT OF HIV BIOMARKERS IN MEDICINE

Dr.Anubha Dubey*

ABSTRACT: Biomarkers have gained immense scientific and clinical value and interest in the practise of medicine. Biomarkers are potentially useful in diagnosis, screening, staging and selection of initial therapy. Advances in genomics, proteomics, and emerging high-throughput technologies in medical practice are important sources to develop drugs for HIV/AIDS. In this paper assessment of HIV biomarkers is studied by machine learning techniques as it is the important step to find out the better biomarker for HIV diagnosis and treatment. In future, these novel approaches have provided opportunities to develop personalized treatment strategies for HIV/AIDS.

Keywords: Biomarker, Diagnosis, High-throughput, Treatment, Personalized

* Phd Bioinformatics

INTRODUCTION

In 2001, a consensus panel at the 'National Institutes of Health" defined the term biomarker as 'a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention or other health care intervention'. The biomarker is either produced by the diseased organ (e.g., tumour) or by the body in response to disease. Biomarkers are potentially useful along the whole spectrum of the disease process. Before diagnosis, markers could be used for screening and risk assessment. During diagnosis, markers can determine staging, grading, and selection of initial therapy. Later, they can be used to monitor therapy, select additional therapy, or monitor recurrent diseases [1]. Thus,

identifying biomarkers include all diagnostic tests, imaging technologies, and any other objective measures of a person's health status. Hence biomarkers are also used in diagnosis of HIV-AIDS [12], the world most deadly disease. AIDS is caused by HIV leading to severe immune damage to human body and even death at later stage. Potential of biomarkers lead to sufficient use in AIDS for diagnosis, monitoring and selection of therapy.

Biomarkers can also be used to reduce the time factor and cost for phase I and II of clinical trials by replacing clinical endpoints. Biomarkers span a broad sector of human health care and have been around since the understanding of HIV-AIDS biology and other diseases to evolve.

Phases of evaluation of biomarkers

A clinical trial design may be consist of following phases which was guided by the National Cancer Institute's "Early Detection Research Network" [2].

Phase I: It refers to preclinical exploratory studies. Biomarkers are discovered through knowledge-based gene selection, gene expression or protein profiling to distinguish HIV and normal samples. Identified markers are prioritized based on their diagnostic/prognostic/therapeutic (predictive) value that could suggest their evolution into routine clinical use. The analysis of this phase is usually characterized by ranking and selection, or finding suitable ways to combine biomarkers. It is preferred that the specimen for this phase of

discovery comes from well-characterized cohorts, or from a trial with active follow-ups.

Phase II: It has two important components. Upon successful completion of phase I requirements, an assay is established with a clear intended clinical use. The clinical assay could be a protein, RNA, DNA or a cell-based technique, including ELISA, protein profiles from MS, phenotypic expression profiles, gene arrays, antibody arrays or quantitative PCR. The significance of these techniques depends upon two parameters: firstly, such assays need to be validated for reproducibility and shown to be portable among different laboratories. Secondly, the assays should be evaluated for their clinical performance in terms of 'sensitivity' and 'specificity' with thresholds determined by the intended clinical use.

Phase III: During this phase, an investigator evaluates the sensitivity and specificity of the test for the detection of diseases that have yet to be detected clinically. The specimens analyzed in this evaluation phase are taken from study subjects before the onset of clinical symptoms, with active follow-up of HIV-AIDS occurrence. It is usually time- consuming and expensive to collect these samples with high quality; therefore, phase III should consist of large cohort studies or intervention trials whenever possible. This is probably when the bio-marker will be ready for clinical use.

Phase IV: It evaluates the sensitivity and specificity of the test on a prospective cohort. A positive test triggers a definitive diagnostic procedure, often invasive and that could lead to increased economic healthcare burden. Therefore, in a phase IV study, an investigator can estimate the false referral rate based on tested biomarkers and describe the extent and characteristics of the disease detected (e.g., the stage of HIV-AIDS at the time of detection). Sometimes, phase IV requires a large cohort with long-term follow-up.

Phase V: In this phase, the overall benefits and risks of the new diagnostic test is evaluated.

Table 1. Performance characteristics of Biomarkers.

Response of biomarker	Disease present	Disease absent
Biomarker positive	A	B
Biomarker negative	C	D

$$Diseaseprevalence = \frac{A+C}{A+B+C+D}$$

$$Negativelikelihoodratio = \frac{1-sensitivity}{specificity}$$

$$Negativepredictivevalue = \frac{D}{C+D}$$

If the biomarker used as a diagnostic test, it should be sensitive and specific and have a high predictive value as shown in table 1. A highly sensitive test will be positive in nearly all patients with the HIV infection, but it may also be positive in many patients without the HIV infection. To be of clinical value, most patients without the HIV infection should have negative test results.

Characteristics of an ideal biomarker and basic statistical methods for evaluation

- An ideal biomarker should be safe and easy to measure.
- The cost of follow-up tests using biomarkers should be relatively low.
- It should be consistent across genders and ethnic groups.

Diagnostic odds ratio (DOR) of a biomarker represents the comprehensive ability of the marker:

$$DOR = \frac{\frac{sensitivity}{1-specificity}}{1-sensitivity/specificity}$$

$$LR+ = \frac{sensitivity}{1-specificity}$$

$$LR- = \frac{1 - sensitivity}{specificity}$$

Information about the diagnostic test itself can be summarized using a measure called the likelihood ratio. The likelihood ratio combines information about the sensitivity and specificity. It tells how much a positive or negative result changes the likelihood that a patient would have the disease.

The likelihood ratio for a positive result (LR+) tells how much the odds of the disease increase when a test is positive. The likelihood ratio for a negative result (LR−) tells how much the odds of the disease decrease when a test is negative. The likelihood ratio can be combined with information about the prevalence of the disease, characteristics of your patient pool, and information about a particular patient to determine the post-test odds of disease. To quantify the effect of a diagnostic test, information about the patient is needed first. The pre-test odds, such as the likelihood that the patient would have a specific disease prior to testing should be specified. The pre-test odds are usually related to the prevalence of the disease, though it might be adjusted upwards or down-wards depending on characteristics of the over-all patient pool or of the individual patient. Once pre-test odds have been specified, they are multi-plied by the likelihood ratio to give the post-test odds:

$$odds_{post} = odds_{pre} \times likelihoodratio$$

The post-test odds represent the chances that a particular patient has a disease. It incorporates information about the prevalence of the disease, the patient pool, and specific patient risk factors (pre-test odds) and information about the diagnostic test itself (the likelihood ratio). Most biological markers, however, are not simply present or absent but have wide ranges of values that overlap in persons with a disease and in those without it. The risk typically increases progressively with increasing levels; few markers have a threshold at which the risk suddenly rises, so various cut-off points must be evaluated for their ability to detect disease. Cut-off points with high sensitivity, producing few false negative results, are used when the consequences of missing a potential case are severe, whereas highly specific cut-off points, producing few false

positive results, are used to avoid mislabelling a person who is actually free of the disease. Sensitivity and specificity calculated at various cut-off points generate a receiver-operating-characteristic (ROC) curve, which ideally will be highly sensitive throughout the range of specificity. The most useful clinical tests are typically those with the largest area under the ROC curve. The use of multiple tests may also be considered

for screening. When multiple tests are obtained in series and the disease is considered present when all tests are positive ('AND rule'), specificity is enhanced whereas sensitivity is diminished. When multiple tests are obtained in parallel and the disease is considered to be present when any of the tests are positive ('OR rule'), sensitivity is enhanced and specificity diminishes [3].

Specific ways for Biomarker assessment:

1. Model discrimination: The C-statistic or area under the receiver operating characteristic curve (AUC) or ROC (receiver operating curve) is a popular method to test model discrimination. C-statistic for a multivariable model reflects the probability of concordance among persons who can be compared for a given outcome of interest and represents the probability that a case has a higher or risk score (or a shorter time to event in survival analyses) than a comparable control. The C-statistic measures the concordance of the score and disease state. The value of the C-statistic ranges from 0.5 (no discrimination) to 1.0(perfect discrimination). When considering the efficacy of novel bio-markers in risk stratification, one approach is to determine to what extent entering the can-didate biomarker into standard risk prediction models will actually increase the model's C-statistic.

One can grouping of new attributes with existing ones (also be grouped according to same characteristics). Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence shows better results in classification. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. There are many machine- learning software available for analysis of data [5] i.e. WEKA software package [4].

Following are some of the methods which are inbuilt in WEKA and are good for analysis of data.

(a)Decision tree (or tree diagram) is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision trees is as a descriptive means for calculating conditional probabilities. In data mining and machine learning, a decision tree is a predictive model; that is, a mapping from observations about an item to conclusions about its target value. More descriptive names for such tree models are classification tree (discrete outcome) or regression tree (continuous outcome). In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a decision tree from data is called decision tree learning, or (colloquially) decision trees [5].

(b)Naïve Bayes Classifier: - A **naive Bayes classifier** is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or lack of presence) of a particular feature of a class is unrelated to the presence (or lack of presence) of any other feature Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood [6].
Of the various biomarkers described in literature CD4+, P24, IL10, IFN-γ can be as biomarkers for prediction and diagnosis of HIV [12].

There are several limitations to using increments in the C-statistic to determine the utility of biomarkers in risk prediction [7]. First, the C-statistic depends to a large extent on the magnitude of the association between a dichotomous exposure and outcome. Other limitations of C-statistic include low sensitivity for determining the relative importance of different risk factors in a multivariate model.

2. Model calibration: A complementary step when analyzing the efficacy of a biomarker is to assess the degree to which the biomarker improves model calibration. This can be thought of as the extent to which the expected risk (estimated by statistical models) agrees with the observed (or true) risk. This concept may be important when counselling patients with regards to their numeric risk or probability of developing a given condition. A simple statistical test to compare model discrimination with and without the biomarker of interest would fail to provide valuable information regard-ing which specific groups (i.e., which deciles or quintiles and so on) of observed and expected risk are better explained by including a biomarker of interest.

3. Risk reclassification: The utility of a biomarker may also be assessed by studying how biomarker information may lead to a reclassification of individuals in low medium- and high-risk categories based on traditional risk factors. The ultimate goal of this approach is to refine risk stratification, and it has been particularly emphasized when considering biomarker information that would serve to shift individuals who are in the intermediate-risk groups, upwards into the high-risk category or downwards into the low-risk category. Recent guidelines have recommended that the individuals in the intermediate-risk category be targeted to undergo screening for existing HIV [8].

4. Model validation: Models can be validated by 5-fold cross validation or 10 fold cross validation. Boot strapping is one of the good methods in machine learning for validating model.

5. Considering multiple biomarkers for HIV: Combinations of biomarkers with their accuracies as obtained by machine learning techniques proves which combination is better for identification, and diagnosis of HIV [12]

In near future if any of the HIV biomarker is identified, the combinations of biomarkers is improved and better one is again analysed by above discussed methods.

High throughput technologies for HIV biomarkers: These technologies are useful to assess genomic data which define the messages and the resulting protein sequences using single

nucleotide polymorphisms and different types of repeats. Transcriptomic data reveal the levels of messages present. The basic idea in transcription profiling is to measure Mrna expression levels of thousands of genes simultaneously in a cell or tissue sample under specific conditions. Proteomics could be described as a large scale study of protein structure, expression, and function (including modifications and interactions). Metabolomics is a whole cell measurement of all the metabolites and it is considered to be equivalent to transcriptomics in Mrna expression analysis.

The reason for using high-throughput technologies is that they provide a large number of correlative data on gene or protein expression in relation to disease. Such data are then analyzed for their association to the disease. The assumption is that multiple variables will be able to provide information on associations more accurately than a single variable (marker). Such strong associations provide major impetus for the molecular profiling approaches to find patterns or profiles for a clinical test based on high dimensional gene or protein expression panels [9].

Comparative genomic analyses have yielded a large number of genomic expression data in relation to disease. The patterns of gene expressions that are observed, represent novel signatures for the respective diseases and can be used to develop new clinical tests based upon gene expression patterns, and identify candidate markers for diagnosis and prognosis.

Single nucleotide polymorphisms have also been used as genetic markers of risk, treatment response, and gene and environment interactions. These high-throughput technologies have significantly increased the number of potential DNA, RNA, Protein biomarkers under study. One of the major problems with high-dimensional data derived from high-throughput genomic and proteomic technologies is overfitting of the data when there are large numbers of potential predictors among a small number of outcome events. For example, a recent study of RNA microarray analysis showed how easy it was to overfit data with a small number of samples. Simon and colleagues clearly demonstrated that expression data on 6000 genes from imaginary individuals, 10 normal and 10 cases, could be used to discover discriminatory patterns, using one common method, with 98% accuracy [10]. Many of the so-called ‘omics’ derived data are subjected to a similar over-fitting if the training and validation sets for analyses are small and not

randomized. Most commonly used approaches to analyze 'omics' data are artificial neural networks, boosted decision tree analyses, various types of genetic algorithms and support vector machine-learning algorithms. Each approach has the potential to over fit the data. Over fitting has led to strong conclusions that are likely to be erroneous. The first step, therefore, would be to determine whether the results are reproducible and portable. For this purpose, information on samples should be blinded and samples be sent to several laboratories for running the sample sets under a fixed protocol. The data from each laboratory should be analyzed by an independent data manager to learn if each laboratory reproduced a similar result. Splitting the samples randomly between 'training sets and validation sets' should minimize the over fitting. The validation set should not contain samples used in training sets [11].

ANTIRETROVIRAL THERAPY & HIV BIOMARKERS

CD4+ and viral load are used as biomarkers for diagnosis and starting treatment for HIV.

Table 2: Relation between CD4+cells,viral load and stages of HIV.

CD4+T cells	Greater than 500 per micro litre of blood	Stage I	Viral load is low.
	Less than 500 per micro litre of blood	Stage II	
	Less than 350 per micro litre of blood	Stage III	Viral load reaches to millions.
	Less than 200 per micro litre of blood	Stage IV	

The level of virus in the body continues to rise and CD4+T cell count continues to fall. Some illnesses that develop in people infected with HIV leads to the need for antiretroviral therapy. The illnesses include HIV-related kidney diseases and certain opportunistic infections. ART is a lifelong treatment that helps people with HIV live longer and healthier lives. Available drugs called Highly active anti-retro viral therapy (HAART). HAART provides effective treatment options for treatment-naive and treatment-experienced patients. Six classes of antiretroviral agents currently exist, as follows:

- Nucleoside reverse transcriptase inhibitors (NRTIs)

- Non-nucleoside reverse transcriptase inhibitors (NNRTIs)
- Protease inhibitors (PIs)
- Integrase inhibitors (INSTIs)
- Fusion inhibitors (FIs)
- Chemokine receptor antagonists (CCR5 antagonists)

Each class targets a different step in the viral life cycle as the virus infects a $CD4^+$ T lymphocyte or other target cell. The use of these agents in clinical practice is largely dictated by their ease or complexity of use, side-effect profile, efficacy based on clinical evidence, practice guidelines, and clinician preference.

Resistance, adverse effects, pregnancy, and confections with hepatitis B virus, or hepatitis C virus present important challenges to clinicians when selecting and maintaining therapy for HIV/AIDS. Combination antiretroviral therapy (cART) has significantly reduced morbidity and mortality of HIV-infected patients, yet their life expectancy remains reduced compared with the general population. Most HIV-infected patients receiving cART have some persistent immune dysfunction characterized by chronic immune activation and premature aging of the immune system. Biomarkers of T-cell activation (CD69, -25 and -38, HLA-DR, and soluble CD26 and -30) is reviewed; generalized immune activation (C-reactive protein, IL-6 and D-dimer); microbial translocation (lipopolysaccharide, 16S rDNA, lipopolysaccharide-binding protein and soluble CD14); and immune dysfunction of specific cellular subsets (T cells, natural killer cells and monocytes) in HIV-infected patients on cART and their relationship to adverse clinical outcomes including impaired CD4 T-cell recovery, as well as non-AIDS clinical events, such as cardiovascular disease are studied [13].

Drug development based on molecular biomarkers and targeted personalised medicine for HIV

In the treatment of diseases especially AIDS, there is a shift from the traditional clinical practices to novel approaches. Traditionally HIV positive patients are treated with nucleotide, nucleoside, protease, integrase inhibitors. However, recent advances in basic and clinical research have provided opportunities to develop 'personalized' treatment strategies. These novel approaches

are intended to identify individualized patient benefits of therapies, minimize the risk of toxicity and reduce the cost of treatment. The biggest challenge for researchers and clinicians today is, to decide on which type of biomarker to use across the wide spectrum of disease processes.

The evolving trend is the usage of patterns of markers instead of a single marker. This approach could, to some extent, reduce the error rate in predicting the outcome or severity of side effects during the targeted therapies. With the increasing knowledge of the molecular pathways underlying the development of various diseases, the selection of patients and their efficacy in future will be based on molecular profiling or phenotypic expression of their target molecules. These targeted drugs shut down their specific pathway or sets of pathways. The predictability of the response to targeted drugs rules out their use in all patients, which helps to avoid unnecessary drug-associated side effects.

CONCLUSION AND FURTHER SCOPE: Machine learning techniques are better to identify the correct biomarker or combination of biomarkers for prediction, diagnosis, prognosis and treatment monitoring of diseases. A large concerted effort is required to advance the field of biomarker discovery. Most current biomarkers do not satisfy the required characteristics for use among the spectrum of diseases. Validation of new biomarkers is necessary. Generation of prospective data will be necessary for validation and demonstration of clinical utility. High-throughput technologies have begun to define disease processes and other biological processes with molecular biology detail and thus offer the potential to identify and characterize novel biomarkers. Molecular biology is now seen as encouraging more 'personalized medicine' – the closer alignment of biological information (derived from molecular diagnostics) and therapy selection. Well designed efforts will be needed to develop general knowledge about the molecular history of diseases, to keep up with the progress with biomarkers development. The evolution of molecular medicine, coupled with the discovery and clinical application of new biomarkers, will play a significant role in reshaping medicine as a science. Science in India could make a significant impact on the global scene if scientists and policy makers could agree to dedicate sufficient time and resources to the field of biomarkers. This should be much beyond task-force and excellence initiatives, and should be output-driven in a defined time line.

REFERENCES

1. Atkinson A J et al 2001NCI-FDA Biomarkers Defination working group; Biomarkers and surrogate enspoints: preferred definations and conceptual framework; Clin. Pharmaco. Ther. 69 89-95.
2. Sullivan Pepe M 2001 Phases of biomarker development for early detection of cencer; J.Natl.Cancer Inst.93 1054-1061.
3. Sackett D L,Hayens R B,Guyatt G H and Tugwell P 1991 The interpretation of diagnostic data; In: Clinical epidemiology. A basic science for clinical medicine (2nd edition) (Boston: little Brown) 62-152.
4. Weka Data Mining Java Software [http://www.cs.waikato.ac.nz/~ml/weka/]
5. Han, J., and M. Kamber. Data mining: concepts and techniques .San Francisco: Morgan Kaufmann, 2001.
6. CSE5230 Tutorial: The Na¨ıve Bayes Classifier 1
7. Cook N R 2007 Use and Misuse of the receiver operating curve in risk prediction; *Circulation* 115 7 928-935.
8. Greenland P,Bonow R O,Brundage B H et al 2007 ACCF/AHA 2007 clinical expert consensus document on coronary artery calcium scoring by computed tomography in global cardiovascular risk assessment and in evaluation of patients with chest pain: A report of the American College of Cardiology Foundation Clinical Expert Consensus Task Force (ACCF//AHA Writing Committee to Update the 2000 Expert Consensus Document on Electron Beam Computed Tomography) developed in collaboration with the society of Atherosclerosis Imaging and Prevention and the Society of Cardiovascular Computed Tomography; *J.Am.coll.Cardiol*. 49 3 378-402.
9. Verma M and Srivastava S 2003 New cancer biomarkers deriving from NCI early detection research; *Recent Results Cancer Res*. 163 72-84.
10. Feng Z et al 2004 Research issues and strategies for genomic and proteomic biomarker discovery and validation: A statistical perspective; *Pharmacogenomics* 5 709-719.
11. Ransohoff D F 2004 Rules of evidence for cancer molecular marker discovery and validation; *Nat. Rev.Cancer* 5 11 845-856.
12. Dubey A, 2015 Machine learning classification for HIV Biomarkers: Online Journal of

Bioinformatics. 16(3): 344-356.

13. WWW.wikipedia for antiretroviral therapy for hiv/aids