American Journal of Applied Sciences

Original Research Paper

A New Proposed Statistically-Derived Compromise Cut-off CD4 Count Value for HIV Patients to Start using ARVs

Maria Mokgadi Lekganyane and Solly Matshonisa Seeletse

Department of Statistics and Operations Research, Sefako Makgatho Health Sciences University, PO Box 107, MEDUNSA, 0204, Gauteng Province, South Africa

Article history
Received: 08-01-2016
Revised: 27-06-2016
Accepted: 28-06-2016

Corresponding Author:
Maria Mokgadi Lekganyane
Department of Statistics and Operations Research, Sefako Makgatho Health Sciences University, PO Box 107, MEDUNSA, 0204, Gauteng Province, South Africa
Email:maria.lekganyane@smu.ac.za

Abstract: The study investigated the relationship between statistical outliers and the inconsistent values of the CD4 count recommended for starting the Antiretroviral Therapy (ART) by HIV-positive patients. Low CD4 counts imply a low immune system. It could be due to AIDS existence or closeness to death. An effective treatment to curb HIV impact is ART, which is recommended for low CD4 counts. However, countries differ in the values used. Developed nations recommend start of ART when the CD4 is still high in order to curb its development. Poor countries use very low CD4 counts. Some countries keep changing the CD4 value for this purpose. The problem is that when CD4 counts are too low, HIV may already be too advanced, making it difficult to save the patient from progression to full AIDS. There should be CD4 values derived using scientific methods to assist in standardizing the CD4 count value for ART commencement. Using a retrospective single-site cohort study design, the study analyses the CD4 counts using robust statistical methods together with conventional statistical methods to study outliers and then derive a compromise CD4 count value that could serve as the standard cut-off starting point for using ART. The CD4 count confidence limits showed outliers to be above 300 cells/mm^3. The lower bound of 0 cannot happen to any living person. The 300 upper bound is a value within the manageable outliers that has not reached critical HIV state. However, this value is near risky CD4 count values. The CD4 count of 300cells/mm^3 indicates deteriorating HIV. If it is set as the starting point of taking ARVs, the patients involved can be saved from reaching painful states of lower CD4 counts.HIV patients' immune systems at this level can still be boosted without them showing physical weakness from eye inspection.

Keywords: ART, CD4, HIV, Outlier

Introduction

Curbing HIV/AIDS is undoubtedly desirable. ART was suggested for every HIV positive patient (Lichterfeld and Rosenberg, 2011). CD4 elicits ART use because low values signal HIV spread and immune system harm. ART are useful for HIV positive patients, especially for CD4 below 350 cells/mm^3. Poor countries recommend starting ART at 200 cells/mm^3. These values have also changed many times. Studying CD4 outliers is vital to manage HIV. Large outliers indicate HIV severity. Statistical modelling can help to detect severities in HIV, while robust methods help to understand the bulk of the data such that outliers do not distort meanings (Andersen, 2008). Robust statistical methods are specialised methods that can resist influences of outliers (Tofallis, 2008).

CD4 Outliers

An *outlier* is an observation lying very far from other values in a random sample, either being unreasonably low or high (Alqallaf *et al*., 2009). CD4 outliers are particularly important in identifying severe HIV status. Robust methods enable understanding the data even when outliers influence analyses. Outliers are not always clearly visible. When using traditional statistical methods, outliers may be masked and the estimates of residuals may be inflated (Maronna *et al.*, 2006). In CD4 counts, outliers can distort actualities and severities may be obscured. Hence, the use of robust methods maintains

Science Publications

© 2016 Maria Mokgadi Lekganyane and Solly Matshonisa Seeletse. This open access article is distributed under a Creative Commons Attribution (CC-BY) 3.0 license.

meanings from bulks of data while identifying anomalies that may be revealing severities.

Robust Statistical Methods

Some efficient robust statistical methods developed along common traditional methods are: Least Absolute Deviation (LAD), Least Trimmed Squares (LTS), S-estimation and M-estimators (Strutz, 2010). Upgrading robust methods includes MM-estimation which combines the robustness of S-estimation with the efficiency of M-estimation. Robust methods can assist in outlier identification, which is sometimes difficult and ensuring that outliers do not distort the results (Dawson, 2011).

Outlier Detection

Let Q_1 and Q_3 be the first and third quartiles of a data set and k a fence constant normally chosen to be either 1.5 or 3. According to Dovoedo (2011), Tukey's boxplots boundaries' method defines outliers as observations outside the interval with lower and upper boundaries:

$$L = Q_1 - k(Q_3 - Q_1) \tag{1}$$

$$U = Q_3 + k(Q_3 - Q_1) \tag{2}$$

HIV Treatment and Health

The CD4 count of a healthy person ranges from 500 to 1,200 cells/mm^3. Below these values for HIV positive patients are signs for requiring ARVs. The U.S. Department of Health and Human Services (HHS) provides guidelines on using ARVs and other HIV medicines to treat HIV infection (Sieleunou and Souleymanou, 2009). These guidelines recommend that all HIV positive people should take ART and emphasise that everyone with a CD4 count below 350 cells/mm^3 should use ARVs. If an HIV-infected person's CD4 count drops rapidly or is below 200 cells/mm^3, starting ARV is imminent (Lichterfeld and Rosenberg, 2011). A CD4 increase is a sign of immune system recovery.

Statistical Techniques

Robust Statistical Methods

Least Absolute Deviation

Least Absolute Deviation (LAD) is a robust statistical optimization procedure comparable to the general Ordinary Least Squares (OLS) technique that approximates a dataset. To design the LAD problem, Wilcox (2011) proposes a data set of points (x_i, y_i) with i = 1, 2, ..., n in which the problem is to find a function f of a specific form containing parameters to be determined such that $y_i \approx f(x_i)$. The approach is to search for estimated values of the unknown parameters that minimise the sum of the absolute values of the residuals:

$$S = \sum_{i=1}^{n} \left| y_i - f(x_i) \right| \tag{3}$$

M-Estimator

M-estimators are a robust comprehensive class of estimators obtained as the minima of sums of data functions (van de Geer, 2000). Considering a family of probability density functions f parameterized by θ, a MLE of θ is calculated for each data set by maximising the likelihood function over the parameter space $\{\theta\}$. When the observations are independent and identically distributed, a MLE $\hat{\theta}$ satisfies:

$$\hat{\theta} = \arg\max_{\theta} \left(\prod_{i=1}^{n} f(x_i, \theta) \right) \tag{4}$$

Least Trimmed Squares

Least Trimmed Squares (LTS) is a robust statistical method that fits a function to a data set (Li, 2005). In OLS, the estimated parameter values β are those values that minimise the objective function $S(\beta)$ of squared residuals:

$$S = \sum_{i=1}^{n} r_i(\beta)^2 \tag{5}$$

For a LTS analysis, this objective function is replaced as follows: For a fixed value of β, let $r_{(j)}(\beta)$ denote the set of ordered absolute values of the residuals. It is sensible to identify the outliers by using Equation 1 and 2. LTS is obtained from Equation 5 by removing the outliers.

Measures of Accuracy

Consider the observations y_1, y_2,...y_n and the corresponding estimates $\hat{y}_1, \hat{y}_2, ...\hat{y}_n$ generated from a prediction model. The errors are:

$$e = y_i - \hat{y}_i \ \ for\ i = 1, 2, ...n \tag{6}$$

Thus, large deviations from the estimates will give large error values. Accuracy measures are defined using these errors. Hence, large measure values imply less accuracy of the prediction model. The measures are the Cumulative Forecast Error (CFE), Mean Absolute Deviation (MAD), Mean Squared Error (MSE), Root Mean Squared (RMSE), standard error (s_e), Mean Absolute Percentage Error (MAPE) (Elamir, 2012) defined as:

$$CFE = \sum_{i=1}^{n} e_i \tag{7}$$

This measure adds the errors together and can be used to show if the model is over forecasting above the actual values (CFE<0) or under forecasts (CFE>0):

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|e_t| \tag{8}$$

The MAD measure is an average of the absolute errors. Smaller values of MAD show more accuracy of the predictor model while larger ones imply less accuracy:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}e_t^2 \tag{9}$$

The MSE measure is an average of the squared errors and has useful mathematical properties related to the variance. Also, smaller MSE values show more accuracy of the predictor model:

$$RMSE = \sqrt{MSE} \tag{10}$$

The RMSE measure is the square root of the MSE and the two measures are interpreted in the same way:

$$s_e = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2} \tag{11}$$

The s_e measure is a standard deviation of the errors. Smaller values are also indicators of more model accuracy:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|e_i|}{A_i} \tag{12}$$

The MAPE measure is a percentage of the absolute errors, with smaller values also showing more predictor model accuracy.

Statistical Inference

Statistical inference refers to the process based on analysed data to deduce properties of an underlying distribution (Held and Bové, 2013). It makes propositions about a population using data. Problems in statistical inference are often related to statistical modeling. Basically, statistical inference conclusions are statistical propositions such as point estimate, interval estimate, rejection of a hypothesis and clustering of data points into groups, among others. In this study, estimation is used and the cut-off measure derived is a result of this system.

Methods

This was a retrospective cohort study on CD4 counts of HIV positive patients from January 2006 to December 2013 attending treatment at the Tshepang Clinic of Dr. George Mukhari Academic Hospital in Gauteng Province, South Africa. Tshepang Clinic receives patients from clinics in surrounding townships and villages. The clinic aims to curb HIV advancement and improve ART. The HIV patients' records for the stated period were about 350, which made the study population. The study used all the useful records of the entire population. Thus the sample consisted of the remaining records (318) after eliminating the unsatisfactorily ones.

Results

CD4 Count Statistics

Table 1 shows abridged CD4 counts. The bulk of the values range from 0-200 and a notable amount is seen from 201-400. Few other values appear to be higher than 401. These may be considered candidates to be outliers.

Figure 1 shows the spread in the CD4 counts and the number of patients. The tail of values is shown from 401 upwards.

Initial Outlier Identification

Equation 1 and 2 provide the lower and upper limits for outlier identification. Let k = 1.5 to enhance stable confidence intervals and to avoid L that is too deep in the negatives. Using the crude quartiles from Table 1, the lower and upper limits are:

$$L = Q_1 - k\left(Q_3 - Q_1\right) = 50 - 1.5\left(150 - 50\right) = -100$$
$$U = Q_3 + k\left(Q_3 - Q_1\right) = 150 + 1.5\left(150 - 50\right) = 300$$

There are no small outliers since the L = –100 cells/mm^3 is below all the CD count values in Fig. 1 and Table 1. Actually, all CD4 counts ≥ 0. Since L = 300, there are 28 CD4 counts identified as (potential) outliers in class intervals above lower limit 300 in Table 1.

Statistics Grouping for ART Purpose

Table 2 indicates 247 (77.7%) bulk of patients below 200 cells/mm^3, who should take ARVs. The next larger group were between 200 and 350 cells/mm^3, which were also recommended to take ARVs. Outliers in this case are only 14, being the healthy ones (>500 cells/mm^3) and the less severe one sat 350-500 cells/mm^3.

Table 3 displays the descriptive statistics of the original CD4 count values with their potential outliers.

Table 4 displays the original mean deviation measures of the CD4 counts.

Table 5 lists the outliers identified in the box plot in Fig. 2.

Identifying Outliers

Robust methods are used as there was evidence of outliers. The analysis identified 18 outliers which were CD4 counts above 300 cells/mm^3. The box plot Fig. 2 identifies them.

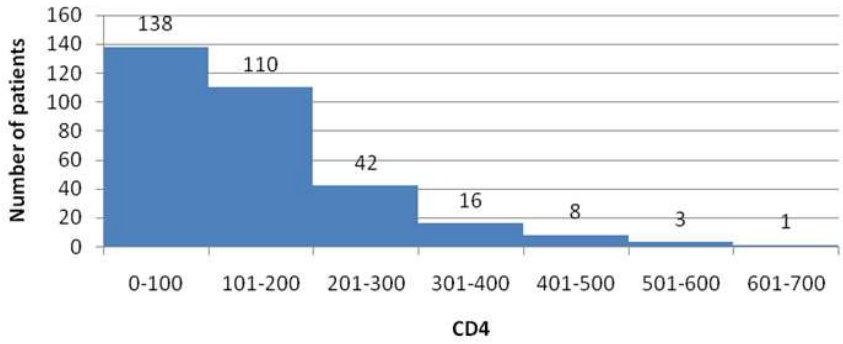

Fig. 1. Shows the spread in the CD4 counts and the number of patients. The tail of values is shown from 401 upwards

Table 1. CD4 counts

CD4	Frequency	Percent
0-100	138	43.4
101-200	110	34.6
201-300	42	13.2
301-400	16	5.0
401-500	8	2.5
501-600	3	0.9
601-700	1	0.3
Total	318	100.0

Table 2. Grouped CD4 count statistics

CD4	Frequency
0 to 200	247
200 to 350	57
350 to 500	10
500 or more	4
Total	318

Table 3. Descriptive CD4 count statistics

Mean	137.927673
Standard error	6.238083763
Median	117.5
Mode	30
Standard deviation	111.2409687
Sample variance	12374.55311
Kurtosis	2.28654435
Skewness	1.35006011
Range	619
Minimum	2
Maximum	621
Sum	43861
Count	318
Confidence level (95.0%)	12.27327796

Table 4. Original mean deviation measures of CD4 counts

CME	MAD	MSE	MAPE	SE
-1.4E-05	26949.67	3922733	930.51	6.22

Figure 2 displays the Box plot to identify potential CD4 outliers.

The box plot confirms the possibility of outliers. It also confirms that there are no outliers below the lower limit. However, instead of 18 outliers, 12 outliers identified.

Table 5. CD4 outliers

#227	=	621
#309	=	539
#32	=	547
#268	=	509
#304	=	489
#222	=	479
#226	=	495
#29	=	439
#38	=	460
#303	=	430
#160	=	405
#253	=	408

Summary Statistics

The next presentation considers new 'resampled' data when outliers have been removed.

Table 6 uses fewer 'responses'. According to Equation 2, there are still CD4 values above 300 cells/mm^3.

Table 7 displays summary statistics without outliers.

Table 8 displays the original mean deviation of CD4 counts.

Table 9 displays the original mean deviation of less CD4 counts.

Table 10 displays the original mean deviation of less CD4 counts and VL outliers.

Robust Statistical Measures

The robust measures assessed from the above analyses are LAD, M-estimator and LTS.

Analysis of Results

Twelve (12) out of 14 measures were affected. Outliers therefore affected 85.7% (12/14) of the measures. This indicates the effect of outliers on the descriptive statistics.

Estimating the Robust Measure Values

Table 12 displays the original mean deviation of CD4 counts less outliers and leverage points.

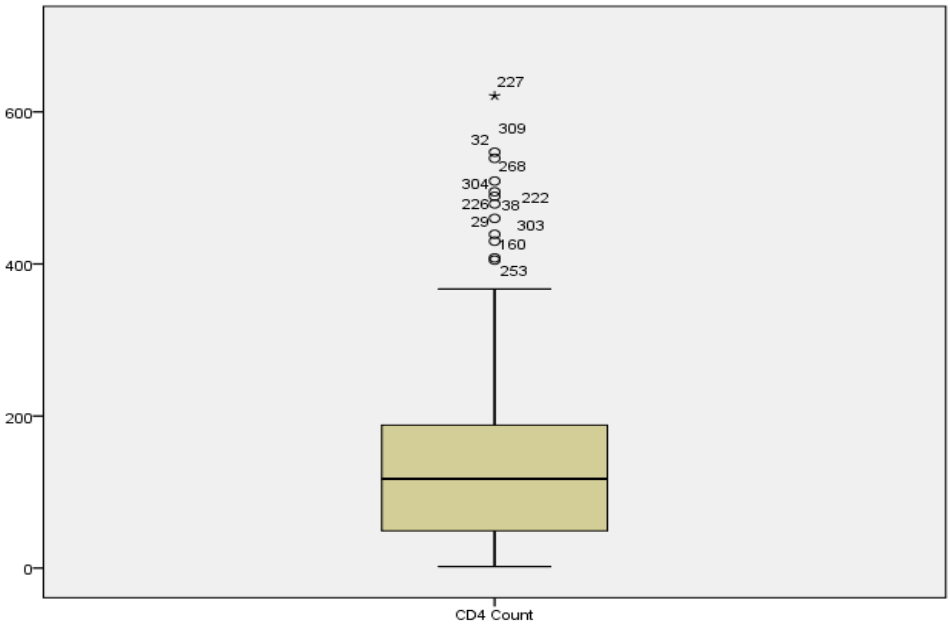

Fig. 2. Box plot of CD4 counts

Table 6. Grouped CD4 counts: Outliers removed

CD4	Frequency
0 to 200	247
200 to 350	57
350 to 500	2
Total	306

Table 7. Descriptive CD4 counts

CD4	
Mean	124.3137255
Standard error	5.045316852
Median	111
Mode	30
Standard deviation	88.25699958
Sample variance	7789.297975
Kurtosis	-0.337132788
Skewness	0.64131174
Range	365
Minimum	2
Maximum	367
Sum	38040
Count	306
Confidence level (95.0%)	9.928035022

Table 8. Original mean deviation measures of CD4 counts

CME	MAD	MSE	MAPE	SE
-1.4E-05	26949.67	3922733	930.51	111.24

Table 9.Mean deviation measures of CD4 counts less CD4 outliers

CME	MAD	MSE	MAPE	SE
-3E-06	22216.15686	2375735.88	826.9120	5.0453

Table 10. Mean deviation measures of CD4 counts less CD4 and VL outliers

CME	MAD	MSE	MAPE	SE
-0.008	21735.8	2380897	821.8535	5.2217

Table 11. CD4 descriptive statistics of different condition

	Original	No outliers
Mean	137.9277	129.8738
Std Error	6.238084	5.550239
Median	117.5	117
Mode	30	30
Std Dev	111.241	96.29305
Sample Var	12374.55	9272.351
Kurtosis	2.286544	1.042803
Skewness	1.35006	0.963119
Range	619	545
Minimum	2	2
Maximum	621	547
Sum	43861	39092
Count	318	301
95% c. lev	12.27328	10.92233

Table 12. Mean deviation measures of CD4 counts less outliers and leverage points

	MAD	MSE	MAPE	SE
Data	26949.7	3922733.0	930.5	6.2381
No outliers	22216.2	2375735.9	826.9	5.0453

Discussion

Least Absolute Deviations

LAD is the lowest sum of the absolute residual values. From Equation 3, removing outliers leads to the lowest MAD in Table 12. Thus, LAD = 22216.2.

M-Estimators

M-estimators are estimators obtained as the minima of sums of functions of the data, obtained using Equation 4. These are equivalents of minimum MAPEs using the

observed values as the divisors, identified in Table 12. Hence, is M-est = 816.9.

Least Trimmed Squares

LTS is a robust statistical method that fits a function to a data set of data, which curtails the sum of squared residuals in Equation 5, also in Table 12. Then, LTS = 2375735.88. The standard error is smaller for 'No outliers' row in Table 11, which seems to confirm that the idea of removing outliers leads to unaffected statistical results.

Discussion and Conclusion

The values obtained for robust measures are LAD = 22216.2; LTS = 2375735.88 as well as M_{est} = 826.9. The statistical outliers obtained from robust methods and extreme values that are prescribed in the starting cut-off for starting to use ARVs did not indicate to have any relationship. However, the adjustment based on statistical guidelines is possible and can be a useful adjustment in increasing ARV usage by HIV patients in order to manage HIV. More specifically, the value of 300 cells/mm^3 was obtained for an upper bound.

Recommendations

The study recommends that:

- Health Authorities should revise the starting level of using ARVs to 300 cells/mm^3

The study also recommends that a study:

- Should be conducted determine the CD4 counts that relate to death stage of a HIV positive patient
- Similar to this one should be undertaken on a bivariate sample of CD4 counts and viral load and compare the results

Acknowledgement

The authors would like to acknowledge Dr M.M Motshwane for recruiting the first author to postgraduate studies in the Department. They would also like to express their appreciations to the unknown referees for their comments and suggestions which improved the presentation of the paper.

Funding Information

This research was supported and funded by Sefako Makgatho Health Sciences University.

Author's Contributions

Maria Mokgadi Lekganyane: Conducted the study, contributed to the study methodology, data analysis and interpretation of results. Wrote the paper, reviewed, revised edited and finalised the manuscript.

Solly Matshonisa Seeletse: Supervised the study, collected and developed the initial manuscript. Designed the research plan and organised the study.

Ethics

This article is original and contains unpublished material. The authors have read and approved the manuscript and no ethical issues are involved.

References

Alqallaf, F., S. Van Aelst, V.J. Yohaiand R.H. Zamar, 2009. Propagation of outliers in multivariate data. Ann Stat., 37: 311-331. DOI: 10.1214/07-AOS588

Andersen, R., 2008. Modern Methods for Robust Regression. 1st Edn., SAGE Publications, London, ISBN-10: 1412940729, pp: 107.

Dawson, R., 2011. How significant is a boxplot outlier. J. Stat. Educ., 19: 1-12.

Dovoedo, Y.H., 2011. Contributions to outlier detection methods: Some theory and applications. Ph.D. Thesis, University of Alabama, Tuscaloosa, USA.

Elamir, E.A.H., 2012. On uses of mean absolute deviation: Decomposition, skewness and correlation coefficients. METRON, 70: 145-164. DOI: 10.1007/BF03321972

Held, L. and D.S. Bové, 2013. Applied Statistical Inference: Likelihood and Bayes. 1st Edn., Springer Science and Business Media, Berlin, ISBN-10: 3642378870, pp: 376.

Li, L.M., 2005. An algorithm for computing exact least-trimmed squares estimate of simple linear regression with constraints. Comput. Stat. Data Anal., 48: 717-734. DOI: 10.1016/j.csda.2004.04.003

Lichterfeld, M. and E.S. Rosenberg, 2011.Antiretroviral combination therapy markedly reduces risk of heterosexual HIV-1 transmission. Evidence-Based Med., 17: 95-96. PMID: 22108078

Maronna, R.A., R.D. Martin and V.J. Yohai, 2006. Robust Statistics: Theory and Methods. 1st Edn., Wiley, Chichester, ISBN-10: 0470010924, pp: 436.

Sieleunou, I. and M. Souleymanou, 2009. Determinants of survival in AIDS patients on antiretroviral therapy in a rural centre in the Far-North Province, Cameroon. Tropical Med. Int. Health, 14: 36-43. DOI: 10.1111/j.1365-3156.2008.02183.x

Strutz, T., 2010. Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond. 1st Edn., Vieweg + Teubner Verlag, Wiesbaden, ISBN-10: 3834810223, pp: 244.

Tofallis, C., 2008. Least squares percentage regression. J. Modern Applied Stat. Meth., 7: 526-534.

Van de Geer, S.A., 2000. Empirical processes in M-estimation: Applications of empirical process theory. Cambridge University Press.

Wilcox, C.W., 2011. Bias: The Unconscious Deceiver. 1st Edn., Xlibris Corporation, ISBN-10: 1465342621, pp: 280.