Modeling Breast Cancer Risks Using Artificial Neural Network: A Case Study of Kenyatta National Hospital in Nairobi, Kenya

Rachael Wanjiru Njoroge[1], Dr. Anthony Waititu[2], Dr. Anthony Wanjoya[3]

[1, 2, 3]Jomo Kenyatta University of Agriculture & Technology, Statistics and actuarial Science, P.O. Box 62,000 – 00200 Nairobi, Kenya

Abstract: ***Background:*** ***Early diagnosis of breast cancer is crucial to the survival of breast cancer patients. In the last years, improved technology has been adopted to aid data collection and store patient's information in a database. Data mining may be used on such databases to come up with patterns that can help in predictability of diseases such as cancer. Use of statistical models may be used to aid doctors and not substitute their opinion. One search model is an artificial neural network model (ANN). Methods: Secondary data was collected from Kenyatta National Hospital (KNH), which is Kenya's national referral hospital located in the capital city, Nairobi. A total of 370 breast cancer patients' information was obtained from both the inpatient and outpatient files. A three layer feed-forward artificial neural network was trained using 320 records. The ANN model obtained was used to predict malignancy in the remaining data set. A logistic regression was used to test which independent variables were significant. Receiver operating curve (ROC) was used to evaluate the ANN's discriminative performance. Coefficient of determination (R^2) was used to evaluate the goodness of fit. Results: ANN demonstrated a superior sensitivity performance over the logistic regression. There were no false positive and no false negative, however for the logistic regression there were seven false positives and eight false negative. The full logistic regression model showed that there were 4 significant independent variables. A reduced logistic regression model was obtained which consisted of 5 independent variables down from 21. ANN was found to have better discriminative performance (AUC=1) as compared to logistic regression (AUC=0.98909). Conclusions: The authors' artificial neural network has high discriminative performance and can accurately predict breast cancer***

Keywords: artificial neural network, breast cancer risks, logistic regression, Kenyatta National Hospital, ROC.

1. Introduction

A. Background information

Breast cancer in Africa occurs at an earlier age as compared to Europe according to Abdulrahman, G. *et al* (2012)[5]. According to International Agency for Research on Cancer (2013) [7], the cause of high mortality in Eastern Africa is because cancer is presented at late stage.

This study was carried out in Kenya which is a developing country in Eastern Africa. Kenyatta National Hospital (KNH) is Kenya's national referral hospital located in the country's capital city, Nairobi. It was founded in 1901 and is the oldest hospital in the country. In the period 1981-1985, 417 new cases of breast cancer was reported in Kenyatta National Hospital. Breast cancer contributed 5% of all malignancies. The female incidence rate was 1.08 per 100 000 persons. This incidence rate has been on the increase over the years according to the Nairobi Cancer Registry [9].

Early diagnosis of breast cancer is crucial to the survival of breast cancer patient. Use of statistical models may be used to aid doctors and not substitute their point of view. One search model is an artificial neural network model (ANN). This study focused on use of a computer aided diagnosis to help radiologists and oncologists in diagnosis of breast cancer.

B. Review of the previous studies

A research done by Ayer T *et al* (2010) on Breast Cancer Risk Estimation with Artificial Neural Networks (ANN) showed that ANN can be effective in discriminating between cancerous and non-cancerous abnormalities of the breast for individuals [1]. They built a feed-forward ANN of size 1000. The authors also compared logistic regression and artificial neural network, where ANN was found to have superior discriminative ability. Their models consisted of 36 independent variables.

2. Materials and Methods

2.1 Data collection

Secondary data was collected from Kenyatta National Hospital. Data collected was for the years 2009, 2010, 2011 and 2014. Demographic factors and mammographic findings of patients at risk of breast cancer were used as independent variables. The dependent variable was biopsy outcome However, not all patients in the above period were included in the data due to missing information in their files.

2.2 Ethical Clearance

This study was approved by the KNH/ UoN-ERC committee. The committee is a joint board between KNH and University of Nairobi.

2.3 Artificial neural network model

A neural network contains three layers namely input layer, hidden layer and output layer. The input-output map consists of d input nodes, H hidden nodes and an activation function [6]. The net input value to the output node is

$$z(x) = \alpha 0 + \sum \alpha_n \phi_n(x) \qquad (1)$$

Finally, the output Z (x; θ) of the net is the value

$$Z(x:\theta) = \psi(z(x)) \quad (2)$$

Training of the network
The weights of network are chosen such that the error is minimal. Quasi Newton Method was used in this study. This method was independently developed by the authors: Broyden, Fletcher and Goldfarb. It's commonly known to as the BFGS method.

Logistic Regression

Logistic regression is a category of generalized linear models. It is used to model discrete outcomes which are binary.
The logistic function is given as:

$$f(h) = \frac{\exp(h)}{1 + \exp(h)} \quad (3)$$

$$= \frac{1}{1 + \exp(-h)} \quad (4)$$

$$h = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta kXk \quad (5)$$

3. Results

R software was used for data analysis

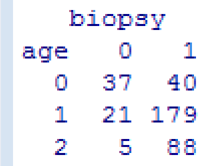
- Dependent variable was breast cancer biopsy outcome
- Independent variables were:
- Independent variable Measurement
- Age (<35, 35-55, >55)
- Sex (Male, Female)
- Menopause
- (Pre-menopause, peri-menopause, post menopause)
- Number of children (>6, 4-6, 1-3, 0)
- Age at first birth (≤30, >30, NA)
- Use of hormonal contraceptives (None Use)
- Personal history (Present, Not present)
- Family history (Present, Not present)
- Prior breast surgery (Present, Not present)
- Education level (Beyond high school, High school, Primary and below)
- Mass size (None, <2cm, 2-5cm, >5cm)
- Pain (Present, Not present)
- Nipple discharge (Present, Not present)
- Breast swelling (Present, Not present)
- Breast density (Present, Not present)
- Calcification (Present, Not present)

3.1 Descriptive statistics of the data

Biopsy outcome was coded as 0=benign and 1=malignant.

3.1.1 Age

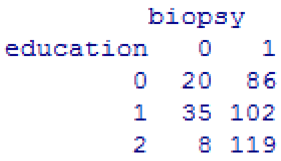
Majority of breast cancer patients were between ages 35-55 with a frequency of 179. The age 35-55 was coded as 1 as shown below. The age group<35 was coded as 0 and >55 was coded as 2.



```
    biopsy
age  0   1
  0 37  40
  1 21 179
  2  5  88
```

3.1.2 Education

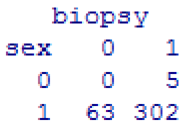
Majority of patients with breast cancer were in the category of primary education and below which was coded as 2. This category had a frequency of 119. Breast cancer patients who had an education level of beyond high school were 86 and were coded as 0. Those who had high school education were coded as 1and had a frequency of 102. The data suggest that cancer risk decreases with increase in the level of education.



```
          biopsy
education  0   1
        0 20  86
        1 35 102
        2  8 119
```

3.1.3 Sex

Male patient was coded as 0 and a female patient was coded as 1. There were more women with breast cancer with a frequency of 302. All male patients had cancer with a frequency of 5.

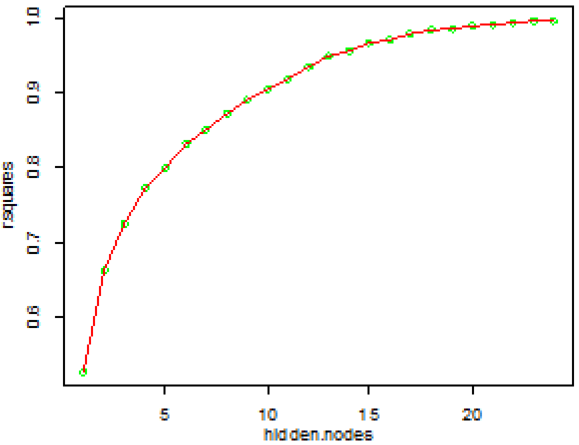


```
    biopsy
sex  0   1
  0  0   5
  1 63 302
```

3.2 Artificial neural network model

A 3 layer feed forward neural network was fitted. The specific model was defined as shown below
Training data consisted of 320 observations out of 370.
nn.nnet=nnet (newx, y, data=datatrain, size=23, entropy=T, abstol=0.0001)

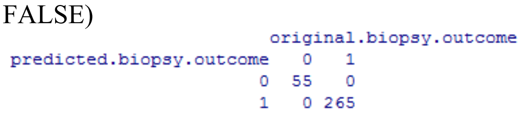
3.3 Identifying the number of hidden nodes



The optimal number of hidden nodes was the node with the highest coefficient of determination (R^2). A neural network of size [1-24] was matched with the respective R^2 value. R^2 was iterated 1000 times and the mean was obtained for each size of the ANN. Hence a curve of R^2 against the respective hidden nodes was plotted to determine the optimal number of nodes as shown above. The optimal size was found to be 23 with an R^2 of 0.996

3.4 Sensitivity analysis for the ANN

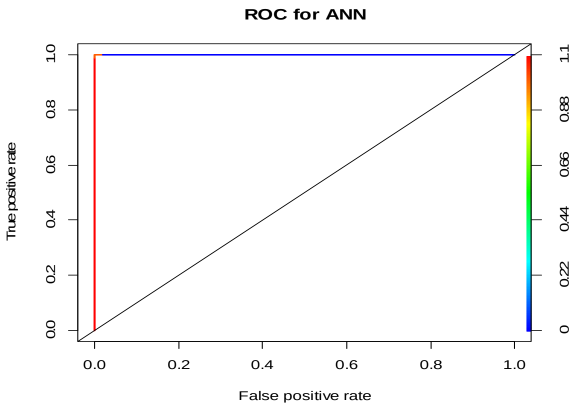
Sensitivity = 1 { (Misclassification Error when Response = TRUE)
Specificity = 1 { (Misclassification Error when Response = FALSE)



```
                       original.biopsy.outcome
predicted.biopsy.outcome   0   1
                       0  55   0
                       1   0 265
```

There were no false positives and no false negatives. Hence artificial neural network has a very good predictive capability

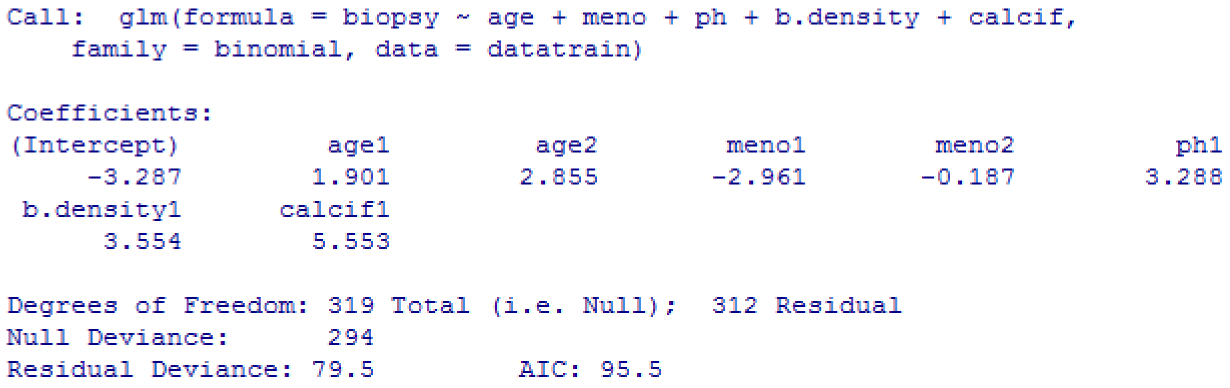
3.5 Receiver operating Curve for the neural network



3.6 Logistic regression

- Model selection
  The significant independent variables at 5% level of significance were
- Age
  Patients aged between 35-55 years were at a higher risk of breast cancer.
- Education
  Patients who had an education level of primary and below were at a higher risk of breast cancer. This could be due to ignorance
- Breast density
  Patients with dense breast are deemed to have a higher risk of breast cancer.
- Calcification
  Patients with micro-calcification are at a higher risk of breast cancer
  (As shown in appendix**[i]**)
  Out of the 21 variables, only 4 variables were significant, hence a reduced model was deemed necessary

3.7 Obtaining a reduced model using stepAIC for stepwise regression

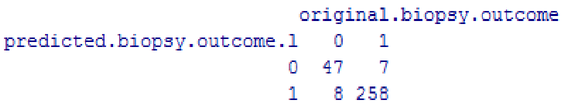


```
Call:  glm(formula = biopsy ~ age + meno + ph + b.density + calcif,
     family = binomial, data = datatrain)

Coefficients:
(Intercept)         age1         age2        meno1        meno2          ph1
     -3.287        1.901        2.855       -2.961       -0.187        3.288
 b.density1      calcif1
      3.554        5.553

Degrees of Freedom: 319 Total (i.e. Null);  312 Residual
Null Deviance:      294
Residual Deviance: 79.5        AIC: 95.5
```

Hence the reduced model was found to be: Call: glm(formula = biopsy ~ age + menopause + personal.history + breast.density + calcification, family = binomial(logit), data = datatrain) (As shown in appendix **[ii]**)

However even though personal history was not significant in the full model, it was included in the reduced model. Personal history is known to be a high risk factor for breast cancer. A patient who had had breast cancer in one breast can have recurrence of cancer in the same breast, in the other breast or in both breasts.

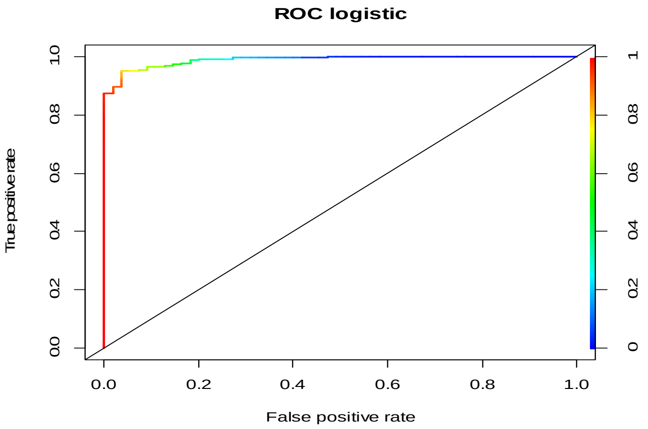
3.8 Sensitivity analysis



```
                         original.biopsy.outcome
predicted.biopsy.outcome.1   0   1
                         0  47   7
                         1   8 258
```

There were 7 false positives and 8 false negative.

3.9 Receiver Operating Curve for the logistic regression



AUC for logistic model was 0.9890909 while AUC for ANN was 1. Hence ANN was found to be a better model than logistic regression model in terms of discrimination capability.

3.10 Predicting biopsy outcome using test data

Predicted ANN output	*Original biopsy outcome*	*Predicted logistic output*
0.0000000000	0	0.0078638287
1.0000000000	1	0.9999475275
0.0000009390	0	0.0495028208
0.9999981442	1	0.9974520437

The predicted outcome was approximately the same as the original biopsy outcome for both models.

4. Summary and Conclusions

The authors' artificial neural network has high discriminative performance and can accurately predict breast cancer risks.
According to the authors' logistic regression the following were the crucial breast cancer risks; age, menopause age, personal history, education, breast density and micro-calcification.

Most breast cancer patients were aged between 35-55 years which is a younger age compared to the age of breast cancer patients in developed countries.

5. Limitations

- Missing information in patients files
- Long duration to obtain ethical clearance

6. Recommendations

The Kenyan government should increase breast cancer awareness campaign. Kenyatta National Hospital should computerize all patients' information to reduce missing information. Further research using a larger data set is recommended

7. Conflict of interest declaration

No conflict of interest.

References

[1] Ayer, T., Alagoz, O., Chhatwal, J., Shavlik, J. W., Kahn, C. E. and Burnside, E. S. (2010), Breast cancer risk estimation with artificial neural networks revisited. Cancer, 116: 3310–3321. doi: 10.1002/cncr.25081

[2] Birdwell R, Bandodkar P, Ikeda D. (2005), Computer-aided detection with screening mammography in a university hospital setting. Radiology, 236: 451-457

[3] Cook N.R. (2007), Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation, 115: 928-935

[4] Franke J., H¨ardle, W., and Hafner, C. (2004), Statistics of financial markets. Universitext, Springer-Verlag, Berlin.

[5] Ganiy Opeyemi Abdulrahman Jnr.1 and Ganiyu Adebisi Rahman2Epidemiology of Breast Cancer in Europe and Africa Journal of Cancer Epidemiology. Volume 2012 (2012), Article ID 915610, 5 pages http://dx.doi.org/10.1155/2012/915610

[6] Gichuhi, A. et al (2008) Nonparametric Changepoint Analysis for Bernoulli Random Variables Based on Neural networks, PhD, Kaiserslautern University, Germany https://kluedo.ub.uni-kl.de/files/2032/Final_Draft_October_14102008.pdf

[7] International agency for research on cancer. (2013). Latest world cancer statistics, Global cancer burden rises to 14.1 million new cases in 2012: Marked increase in breast cancers must be addressed. [Press Release]. Retrieved from http://www.iarc.fr/en/media-centre/pr/2013/pdfs/pr223_E.pdf

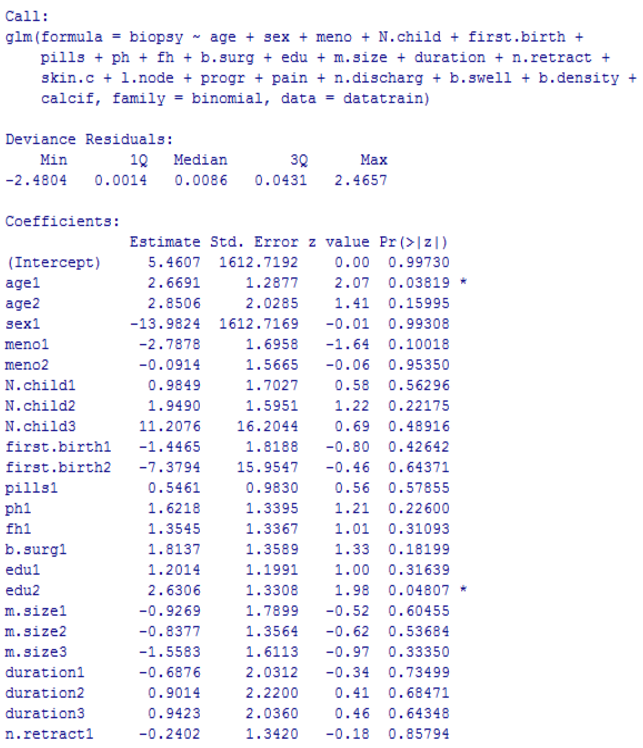
[8] McNelis, P. D. (2004), Neural networks in finance: Gaining predictive edge in the market. Academic Press, Inc., Orlando, FL, USA.

[9] Mutuma G .Z. and Korir,A. (2006). Cancer Incidence Report NAIROBI 2000-2002 . Retrieved from https://www.healthresearchweb.org/files/CancerIncidenceReportKEMRI.pdf

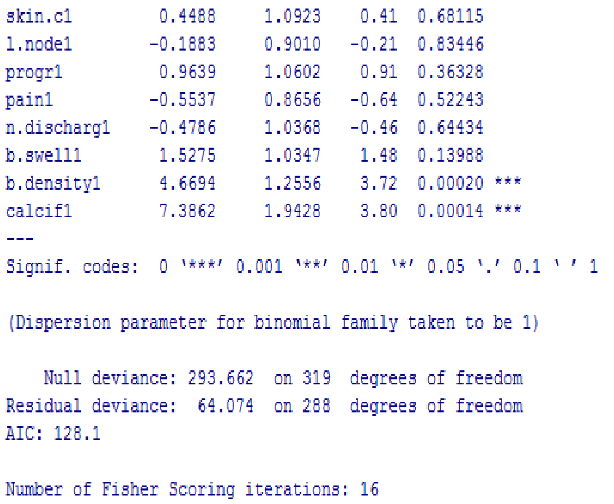
[10] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77.

Appendices

[i]

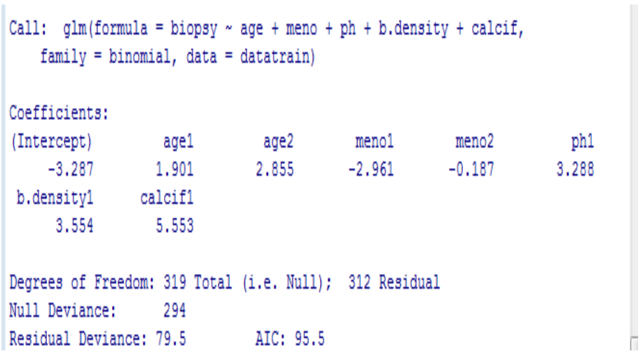


```
Call:
glm(formula = biopsy ~ age + sex + meno + N.child + first.birth +
    pills + ph + fh + b.surg + edu + m.size + duration + n.retract +
    skin.c + l.node + progr + pain + n.discharg + b.swell + b.density +
    calcif, family = binomial, data = datatrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4804   0.0014   0.0086   0.0431   2.4657

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     5.4607  1612.7192    0.00  0.99730
age1            2.6691     1.2877    2.07  0.03819 *
age2            2.8506     2.0285    1.41  0.15995
sex1          -13.9824  1612.7169   -0.01  0.99308
meno1          -2.7878     1.6958   -1.64  0.10018
meno2          -0.0914     1.5665   -0.06  0.95350
N.child1        0.9849     1.7027    0.58  0.56296
N.child2        1.9490     1.5951    1.22  0.22175
N.child3       11.2076    16.2044    0.69  0.48916
first.birth1   -1.4465     1.8188   -0.80  0.42642
first.birth2   -7.3794    15.9547   -0.46  0.64371
pills1          0.5461     0.9830    0.56  0.57855
ph1             1.6218     1.3395    1.21  0.22600
fh1             1.3545     1.3367    1.01  0.31093
b.surg1         1.8137     1.3589    1.33  0.18199
edu1            1.2014     1.1991    1.00  0.31639
edu2            2.6306     1.3308    1.98  0.04807 *
m.size1        -0.9269     1.7899   -0.52  0.60455
m.size2        -0.8377     1.3564   -0.62  0.53684
m.size3        -1.5583     1.6113   -0.97  0.33350
duration1      -0.6876     2.0312   -0.34  0.73499
duration2       0.9014     2.2200    0.41  0.68471
duration3       0.9423     2.0360    0.46  0.64348
n.retract1     -0.2402     1.3420   -0.18  0.85794
```



```
skin.c1         0.4488     1.0923    0.41  0.68115
l.node1        -0.1883     0.9010   -0.21  0.83446
progr1          0.9639     1.0602    0.91  0.36328
pain1          -0.5537     0.8656   -0.64  0.52243
n.discharg1    -0.4786     1.0368   -0.46  0.64434
b.swell1        1.5275     1.0347    1.48  0.13988
b.density1      4.6694     1.2556    3.72  0.00020 ***
calcif1         7.3862     1.9428    3.80  0.00014 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 293.662  on 319  degrees of freedom
Residual deviance:  64.074  on 288  degrees of freedom
AIC: 128.1

Number of Fisher Scoring iterations: 16
```

[ii]



```
Call:  glm(formula = biopsy ~ age + meno + ph + b.density + calcif,
    family = binomial, data = datatrain)

Coefficients:
(Intercept)         age1         age2        meno1        meno2          ph1
     -3.287        1.901        2.855       -2.961       -0.187        3.288
 b.density1      calcif1
      3.554        5.553

Degrees of Freedom: 319 Total (i.e. Null);  312 Residual
Null Deviance:      294
Residual Deviance: 79.5        AIC: 95.5
```

Author Profile

Rachael W. Njoroge. Holds a Bachelor of Science degree in Applied Statistics with Computing from Moi University, in Kenya, currently she is finalizing her Master of Science Degree in Applied Statistics at Jomo Kenyatta University of Agriculture and Technology

A. G. Waititu Holds a PhD in Applied Statistics from Kaiserslautern University, Germany and currently he is a senior lecturer in the department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology.

Wanjoya Anthony Holds a PhD in Applied Statistics from Università degli Studi di Padova and currently he is a senior lecturer in the department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology.