An Algorithm for Notifiable Disease Modeling and Prediction Using Artificial Intelligence Techniques: A Case of Kenya.

Nicodemus Nzoka Maingi (nicomaingi@gmail.com)
Strathmore University https://orcid.org/0000-0002-9796-4804
Ismail Ateya Lukandu
Strathmore University
Matilu MWAU
Kenya Medical Research Institute

Research

Keywords: Artificial Inteeligence, Decision Tree, Disease Burden, Disease Surveillance, Disease Symptom, Decision Tree Theory, Entropy, Information Gain

Posted Date: August 31st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-839844/v1

Abstract

Background

The disease outbreak management operations of most countries (notably Kenya) present numerous novel ideas of how to best make use of notifiable disease data to effect proactive interventions. Notifiable disease data is reported, aggregated and variously consumed. Over the years, there has been a deluge of notifiable disease data and the challenge for notifiable disease data management entities has been how to objectively and dynamically aggregate such data in a manner such as to enable the efficient consumption to inform relevant mitigation measures. Various models have been explored, tried and tested with varying results; some purely mathematical and statistical, others quasi-mathematical cum software model-driven.

Methods

One of the tools that has been explored is Artificial Intelligence (AI). AI is a technique that enables computers to intelligently perform and mimic actions and tasks usually reserved for human experts. AI presents a great opportunity for redefining how the data is more meaningfully processed and packaged. This research explores AI's Machine Learning (ML) theory as a differentiator in the crunching of notifiable disease data and adding perspective. An algorithm has been designed to test different notifiable disease outbreak data cases, a shift to managing disease outbreaks via the symptoms they generally manifest. Each notifiable disease is broken down into a set of symptoms, dubbed symptom burden variables, and consequently categorized into eight clusters: Bodily, Gastro-Intestinal, Muscular, Nasal, Pain, Respiratory, Skin, and finally, Other Symptom Clusters. ML's decision tree theory has been utilized in the determination of the entropies and information gains of each symptom cluster based on select test data sets.

Results

Once the entropies and information gains have been determined, the information gain variables are then ranked in descending order; from the variables with the highest information gains to those with the lowest, thereby giving a clear-cut criteria of how the variables are ordered. The ranked variables are then utilized in the construction of a binary decision tree, which graphically and structurally represents the variables. Should any variables have a tie in the information gain rankings, such are given equal importance in the construction of the binary decision-tree. From the presented data, the computed information gains are ordered as; Gastro-Intestinal, Bodily, Pain, Skin, Respiratory, Others. Muscular, and finally Nasal Symptoms respectively. The corresponding binary decision tree is then constructed.

Conclusions

The algorithm successfully singles out the disease burden variable(s) that are most critical as the point of diagnostic focus to enable the relevant authorities take the necessary, informed interventions. This algorithm provides a good basis for a country's localized diagnostic activities driven by data from the reported notifiable disease cases. The algorithm presents a dynamic mechanism that can be used to

analyze and aggregate any notifiable disease data set, meaning that the algorithm is not fixated or locked on any particular data set.

1. Background

Disease surveillance is an information-based activity involving the collection, analysis and interpretation of large volumes of disease outbreak data from a variety of sources in order to inform and drive objective and informed intervention. The Disease Surveillance and Response Unit (DSRU) is the entity mandated (in Kenya) to monitor and undertake response and mitigation measures in the event of a notifiable disease outbreak; a notifiable disease refers to any disease in a country or community whose occurrence must be reported to the authorities (WHO, 2006). Each time a notifiable disease is reported, the DSRU undertakes the necessary response activities (CDC, 2012; DSRU, 2014).

In Kenya, disease outbreaks are mostly tackled from two perspectives; reactive measures - in the event a notifiable disease outbreak is reported, mitigating steps are only undertaken in response to the particular incident(s) to minimize the potential consequent adverse effects; not much is learned or information utilized in the aftermath that could meaningfully, incrementally and objectively inform future outbreaks and; proactive measures – anticipatory measures are put into play such that should an outbreak occur or recur, its adverse effects are greatly minimized with health personnel taking informed, premeditated and experience-driven steps as a better approach to empower the health personnel be better prepared to cope with every subsequent outbreak.

The infectious diseases of the past have been known to have included some of the most contagious and feared plagues of the past, with new strains continuing to emerge over time; this warrants a widely and greatly co-operative and proactive approach even when the disease outbreak responses and intervention efforts remain the prerogative of the concerned national government. Global partners (such as the Centre for Disease Control and Prevention [CDC], the United Stated Agency for International Development [USAID], the World Health Organization [WHO] among others) have also been seen to play a great role by working in close collaboration to offer the much needed medico-technical and social support from its battery of experienced and seasoned teams cutting across numerous medical specialities and vast geopolitical backgrounds (Brownstein et al, 2009; Martinez, 2000).

To enable each country’s concerned teams better manage its disease outbreaks more efficiently, a notifiable disease list and its epidemiological week (Epi-Week) must be defined; an Epi-week is a weekly period in a country within which notifiable disease outbreak data must be recorded and reported to the relevant health authorities. Kenya’s epi-week runs from Monday through Sunday (DSRU, 2014; WHO, 2006).

The efforts to manage disease outbreaks have become a very complex endeavour; historically, it was easier due to smaller populations and the limited, minimized yet localized cross-border and cross-territorial movements and interactions that curtailed the cross-pollination or dissemination of infectious

diseases the concerned population may have been harbouring - this has greatly changed in the advent of globalization (Wagner, 2001).

The effects of globalization have brought forth new dynamic risk factors in disease spread and management. Such factors include: faster and easy cross-border movements of people and animals, making diseases spread faster - for instance, urbanization remains one of the greatest factors of disease spread: new urban settlements and availability of a huge community of commuting skilled and readily available labour across geopolitical boundaries having the ability to create some infection epicentres that if not well-managed, could easily become incubators for new epidemics, and zoonotic diseases, which can spread in a more rapid manner, quickly elevating them to global levels of interest and concern (Nsubuga et al, 2010).

Next comes means of transporting goods or parcels. The efficient and rapid movement of goods also presents a possibility of enabling and enhancing the spread of diseases since the goods may be harbouring and transporting whatever existing disease strain to wherever they are transported or delivered (Mack, et al, 2010). Additionally, there is also the new, modern practice of families frequently eating out where they get more exposed to different infectious disease strains, among other exposures (Zhong et. al., 2021). Suddenly, one nation's (seemingly localized) epidemic challenges quickly become other nations', regions' and partners' health concerns – pathogens are not known to commonly follow or respect geopolitical and human boundaries.

Additionally, in economic and industrial competitive terms, other factors could also kick in - for instance, the economic empowerment or disempowerment of the notifiable disease-affected populations when skilled, experienced and knowledgeable working personnel get grossly affected by a disease (Kulldorff, 2001; Morse, 2001; Neiderud, 2015; Pillai et al, 2014). The push and pull factors for disease surveillance also touch on the socio-economic activities of a nation; disease outbreaks have been known to decimate the knowledgeable, skilled and able-bodied working populations of any nation to a point of economic near-standstill if not total collapse (Roser, 2015).

Further, it is has been observed that the progression or retrogression of the economic well-being of a community can now be greatly tied to proper disease outbreak management; if the adverse effects slow down economic activity, then all measures, (including the improvement of the health infrastructure and the response and mitigation apparatus of a country) must be called upon to prevent or deal with the adversity of the disease outbreaks (Baker et al, 2002; Roser, 2015). To combat such disease strains, concerted efforts and clear-cut strategies need to be employed; the enhanced use of ICT software and tools has been seen as a great driver and catalyst to enable the quick aggregation, packaging and dissemination of disease data through to the relevant personnel for easier, faster and better-informed interventions (Weinberg et al, 2003).

The disease outbreak data used here is subjected to AI's machine learning theory. Machine learning is a technique that provides systems with the ability to automatically learn and improve from experience (Neiderud, 2015; Roser, 2015). Whilst traditional disease outbreak management assumes the method of

relying on past disease data that is seen to point towards what infectious disease strains manifest, this research looks to dig deeper. Using AI, the researcher hopes to drive a different perspective to notifiable disease outbreak management.

Of the two disease outbreak management perspectives outlined earlier, the researcher looks to build on the proactive disease outbreak measures. The main driving question or hypothesis here is whether a different approach could be employed to the processing and packaging of notifiable disease data in order to better inform and drive proactivity in the disease surveillance and response practice.

2. Method

The methodology used here employs various techniques; quantitative and qualitative research analysis blended with evolutionary and iterative prototyping. The C4.5 decision tree theory in artificial intelligence has been used in the diagnostic analysis efforts, with the computed information gains consequently becoming reliable determinants in informing the structure of the resultant binary decision tree(s).

Post validation, the algorithm could be further applied to the general notifiable disease-list across many counties and regions to handle the variation of the disease outbreak footprints as an additional test measure of the algorithm's efficacy; it is expected that any challenges experienced in the process of the development of the algorithm will be used as a basis for future improvement and to inform policy development and assist in better planning efforts (Childs et al, 2007; Moncayo et al, 2009).

The eight symptom clusters adopted are listed below:

Table 1
Disease Symptom Clusters Legend

Symptom Cluster Code	Symptom Cluster	Brief Description
B	*Bodily*	Those symptoms that are generally manifested through the general human body organs and parts e.g. fever.
G	*Gastro-Intestinal*	Those symptoms that are generally manifested through the human body's digestive system e.g. vomiting, running stomach etc.
M	*Muscular*	Those symptoms that are manifested via the human body's muscular tissues.
N	*Nasal*	Those symptoms that are generally manifested through the human nasal cavity e.g. running nose, sneezing etc.
P	*Pain*	Those symptoms that manifest in form of various forms of human body pain e.g. headache.
R	*Respiratory*	Those symptoms that manifest through the human body's respiratory processes or apparatus e.g. shortness of breath, coughing.
S	*Skin*	Those symptoms that manifest through a human body's skin tissue e.g. skin rush, skin peeling or inflammation.
O	*Other*	Those symptoms that generally fall outside the other seven defined symptom clusters e.g. blurred eye sight.

Entropy:

Equation 1 - Entropy Computation

(Russel et al, 2009)

$$H(X) = H(p_1, ..., p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Information Gains:

Equation 2 - – Information Gains Computation (Russel et al, 2009)

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

Once the information gains are computed, their rankings are used to determine the order of the symptom variable clusters in the construction of the binary decision tree, yielding a structure that helps to graphically and visually break down the notifiable disease outbreak data into a meaningful form to guide intervention and proactive action. The ranked symptom burden variables can assist the health personnel

in easily mapping what disease symptom variable(s) to lay emphasis upon in their efforts to combat outbreaks.

This means that there is a deviation from the traditional practice of the focus being laid upon the singular diseases themselves; the combating of disease outbreaks would mainly be driven by the symptom cluster variables i.e. it is possible to focus on only those diseases that manifest certain symptom cluster variables that are highly ranked via the algorithm using the computed information gains. Thus, the planning and mitigating measures would mainly be on the diseases symptom variables, and not necessarily the raw disease(s) themselves.

Each notifiable disease data set follows the information gains ranking. For instance, if the Pain symptom variable ranks first in the information gains computation, then it will become the root node in the resultant binary decision tree. The rest of the symptom variables will follow accordingly. If two or more variables tie in the information gain ranks, then they shall jointly be part of a leaf node (or the root node, if they tie on rank one) as the decision tree gets defined and constructed.

For purposes of the validation of the algorithm, the researcher chose to use the C4.5 Machine Learning Binary Decision Tree theory in the computation of the entropies and information gain values prior to the definition and construction of the binary decisions trees. The data used in the validation of this algorithm is aggregated from Kenya's Nairobi County over the 2015–2018 notifiable disease reporting period.

Below is the algorithmic process flow of the various activities:

3. Results

The information gain scores tabulated here are derived from the data sets prepared from the primary data. The information gain scores are ranked from the largest to the smallest; with the highest information gain score pointing to a particular symptom variable(s) that is the most critical in the decision tree construction, whilst the smallest information gain score points to the variable that is the least important as a binary decision tree determinant variable.

Table 2

Nairobi County 2015 – 2018 Aggregated Information Gain Scores and Rankings

Nairobi County Information Gains - Disease Symptom Burdens								
C4.5 Technique Data								
Gain (Decision, Variable) AND Gain Rankings								
	Overall Aggregated Data							
Variable	B	G	M	N	O	P	R	S
Information Gain	4.3496	4.4366	1.0144	0.7654	2.2060	3.4801	2.3451	2.8691
Rankings	2	1	7	8	6	3	5	4

The binary decision tree shall then be constructed according to the information gain rankings tabulated above.

4. Discussion

It is a prerogative of every nation to focus on the strengthening of its public health infrastructure to protect its citizen's health, and especially in combating disease outbreaks (Baker et al, 2002). Thus, all disease outbreak factors can easily be dealt with.

This research looks to present a good push in innovating new approaches and methodologies in the development of a proactive, early warning system in the response and intervention efforts to support some medium- and long-term mechanisms for the processing of the disease data with the focus on specific trends to inform policy development and planning, thereby boosting decision-making at the DSRU in collaboration with other concerned partners (Bernardo et al, 2013).

The algorithm demonstrates interesting results. Of the eight disease symptom burden cluster variables, the gastro-intestinal variable emerges as the most prominent, having registered an information gain score of 4.4366. It goes on to form the root node (the first node of a binary decision-tree). It is closely followed by the bodily symptom variable with an information gain score of 4.3496. The others follow in the following order (based on the information gain scores): Pain (3.4801), Skin (2.8691), Respiratory (2.3451), Others (2.2060), Muscular (1.0144) and finally Nasal manifestations (0.7654). With the gastro-intestinal variable emerging as the most highly ranked variable, this means that the disease mitigation efforts and focus should be laid upon those diseases that manifest any gastro-intestinal symptoms. Once these have been exhaustively addressed, next will come those with bodily symptoms. Consideration should be taken right from the most highly ranked symptom variable cluster to the lowest in order to objectively guide the diagnostic preparedness of an entity (be it a nation, a province, county or any other geographical demarcation possible)

The proposed shift here means there is a deviation from the traditional diagnostic practice of focus and diagnostic emphasis being laid upon diseases individually; instead, each disease is defined as a set of symptoms within the defined disease symptom clusters. The algorithm can then be applied; simply determine the information gains of each of the variable, the rankings and consequent variable classification and finally, the binary decision tree construction. Learning is a multi-faceted occurrence, with learning processes involving the acquisition of new declarative knowledge, the development of motor and sensitive skills through instruction and practice ordering of the new knowledge. Tools to drive such new knowledge include artificial intelligence branches such as machine learning, as utilized in this research (Michalski et al, 2013).

The essence of this algorithm is to derive a seemingly localized diagnostic framework to enable local medical personnel easily manage disease outbreaks by predicting what disease symptom variables should be given priority in the fight against outbreaks. This approach assumes that in order to manage disease outbreaks on an ongoing basis, all the diseases' should be classifiable within the eight symptom clusters. The algorithm then goes on to cluster the diseases based on their most critical symptom burdens. Emphasis is laid on the disease symptom variables i.e. the disease(s) that manifest(s) a certain highly ranked symptom variable is given more prominence in the diagnostics and interventional process.

The algorithm's ranking of symptom variables is purely data-driven i.e. as new data is posted, the symptom variables' information gains are expected to keep changing and assuming new ranks, thereby dynamically changing the order of importance of the symptom variables of focus. As such, the fight against disease outbreaks focuses not on the diseases, but by the symptoms that drive these diseases.

Of great importance is the management of disease outbreaks by providing an objective basis for crunching and aggregating the data in a novel and objective way to easily inform decision-making.

5. Conclusion

In conclusion, it has been demonstrated that the disease management efforts of an entity can be purely driven by the disease data presented aided by the just defined and validated algorithm. This research study ends up creating a case for disease diagnostics mainly using symptom burden variables. Notably, a case for the machine learning driven algorithm has been presented together with its validation process. Additionally, the algorithm has been used in the computation of the information gains and their rankings. Finally, the just defined, computed and ranked information gains have been shown to form a basis for the definition and construction binary decision tree.

In the end, the algorithm has been designed, constructed and validated. The whole process easily enables the disease outbreak management exercise of any local authority be home-grown i.e. the basis of the disease outbreak management can be guided and driven by the local disease data being captured and continuously crunched to keep the disease diagnostics exercise as fluid and as objective as the data that drives it.

6. List Of Abbreviations

AI Artificial Intelligence.

Epi-Week Epidemiological Week.

CDC Centre for Disease Control and Prevention.

DDSR Division of Disease Surveillance and Response.

ML Machine Learning.

USAID United Stated Agency for International Development.

WHO World Health Organization.

7. Declarations

Ethics Approval and Consent to Participate

This piece of research has been cleared to undertake this research and publications by both the participating entities, Strathmore University (SU) and the Kenya Medical Research institute (KEMRI).

Consent for Publication

The researcher has granted consent for this publication to be undertaken in the cover letter. He agrees to be bound by the journal's rules and regulations as the case may be.

Availability of Data and Materials

The data sets used in this research have been provided as part of the uploaded files.

Competing Interests

The researcher wishes to declare that there are no known competing interests to this research article.

Funding

This research has been undertaken with o funding whatsoever.

Authors' Contributions

IAL undertook the study, critique and review of the AI techniques and principles that were eventually used in the research study. Finally, he helped get the ethical review clearance from Strathmore University for this study to be carried out.

MM assisted in the study, classification and documentation of notifialbe diseases and breakdown into their symptom cluster variables that were eventually used in the study. He was also very instrumental in helping this research team with the engagement with health personnel from KEMRI as well as getting clearance for us to get the access to the data used in the study.

NNM undertook the study, lead discussions and the design of the multidisciplinary approach that brought together medical scholars and practioners (like Prof. Matilu) and ICT and computer science scholars and practioners (like Prof. Ateya) to undertake this research. He also assisted in the design of the research and the processes that informed the successful completion of this study. He was also responsible for the design and implementation of the data models that helped in sifting through the data that informed the eventual results presented in this study. Finally, he was also responsible for the design, implementation and validation of the algorithm as well as documentation of the processes and outcomes that informed this research writeup.

Acknowledgements

The researcher wishes to acknowledge the support received fro the Kenya Medical Research Institute (KEMRI) as well as that from Strathmore University.

8. References

1. Baker, E. L., & Koplan, J. P. (2002). Strengthening the Nation's Public Health Infrastructure: Historic Challenge, Unprecedented Opportunity. Health Affairs, 21(6), 15-27.
2. Bernardo, T. M., Rajic, A., Young, I., Robiadek, K., Pham, M. T., & Funk, J. A. (2013). Scoping Review on Search Queries and Social Media for Disease Surveillance: A Chronology of Innovation. Journal of Medical Internet Research, 15(7), e147.
3. Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital Disease Detection - Harnessing the Web for Public Health Surveillance. New England Journal of Medicine, 360(21), 2153-2157.
4. Centre for Disease Control (CDC). (2012). Summary of Notifiable Diseases (SND) -United States, 2010. Morbidity and Mortality Weekly Report (MMWR). 59(53), 1.
5. Childs, J. E., Krebs, J. W., Real, L. A., & Gordon, E. R. (2007). Animal-Based National Surveillance for Zoonotic Disease: Quality, Limitations, and Implications of a Model System For Monitoring Rabies. Preventive Veterinary Medicine, 78(3), 246-261.
6. Disease Surveillance and Response Unit (DSRU) Manual, 2014.
7. Kulldorff, M. (2001). Prospective Time Periodic Geographical Disease Surveillance Using a Scan Statistic. Journal of the Royal Statistical Society: Series A (Statistics in Society), 164(1), 61-72.
8. Mack, A., Choffnes, E. R., & Relman, D. A. (Eds.). (2010). *Infectious Disease Movement in a Borderless World: Workshop Summary*. National Academies Press.
9. Martinez, L. (2000). Global Infectious Disease Surveillance. International Journal of Infectious Diseases, 4(4), 222-228.

10. Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). Machine learning: An artificial intelligence approach. Springer Science & Business Media.
11. Moncayo, Á., & Silveira, A. C. (2009). Current Epidemiological Trends for Chagas Disease in Latin America and Future Challenges In Epidemiology, Surveillance and Health Policy. Memórias do Instituto Oswaldo Cruz, 104, 17-30.
12. Neiderud, C. J. (2015). How Urbanization Affects the Epidemiology of Emerging Infectious Diseases. Infection Ecology & Epidemiology, 5(1), 27060.
13. Nsubuga, P., Nwanyanwu, O., Nkengasong, J. N., Mukanga, D., & Trostle, M. (2010). Strengthening Public Health Surveillance and Response Using the Health Systems Strengthening Agenda in Developing Countries. BMC Public Health, 10(1), S5.
14. Pillai, S. K., Nyenswah, T., Rouse, E., Arwady, M. A., Forrester, J. D., Hunter, J. C., & Poblano, L. (2014). Developing an Incident Management System to Support Ebola Response, July–August 2014. Mortality and Morbidity Weekly Report (MMWR), 63(41), 930-3.
15. Roser, M. (2015). Life Expectancy. Our World in Data. Available at https://ourworldindata.org/life-expectancy (Accessed on 12th July 2020)
16. Russell J. R., Norvig P., (2009) Artificial Intelligence: A Modern Approach (AIMA).
17. Wagner, M. M., Tsui, F. C., Espino, J. U., Dato, V. M., Sittig, D. F., Caruana, R. A., ... & Fridsma, D. B. (2001). The Emerging Science of Very Early Detection of Disease Outbreaks. Journal of Public Health Management and Practice, 7(6), 51-59.
18. Weinberg, M., Waterman, S., Lucas, C. Á., Falcon, V. C., Morales, P. K., Lopez, L. A., & Bryan, R. (2003). The US-Mexico Border Infectious Disease Surveillance Project: Establishing Bi-National Border Surveillance. Emerging Infectious Diseases, 9(1), 97.
19. World Health Organization (WHO). (2006). Communicable Disease Surveillance and Response Systems: Guide to Monitoring and Evaluating. Available at: http://apps.who.int/iris/bitstream/10665/69331/1/WHO_CDS_EPR_LYO_2006_2_eng.pdf (Accessed on 12th July 2020)
20. Zhong, Y., Oh, S., & Moon, H. C. (2021). What Can Drive Consumers' Dining-Out Behavior in China and Korea during the COVID-19 Pandemic?. *Sustainability*, *13*(4), 1724.

Figures

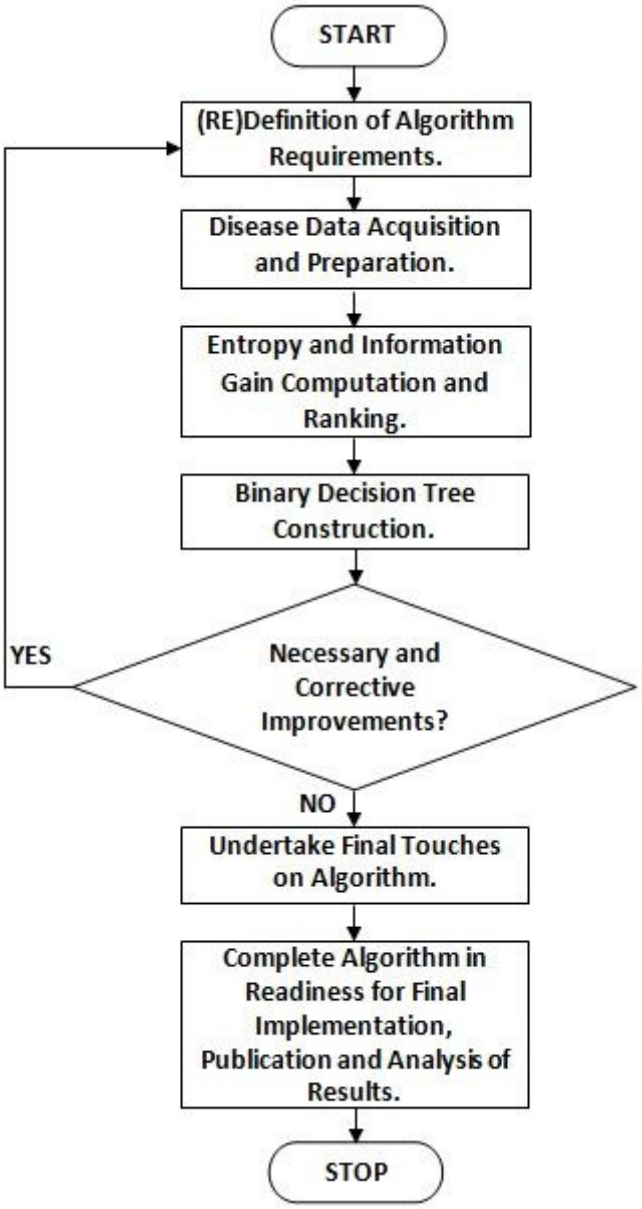

Figure 1

Algorithmic Process Flow.

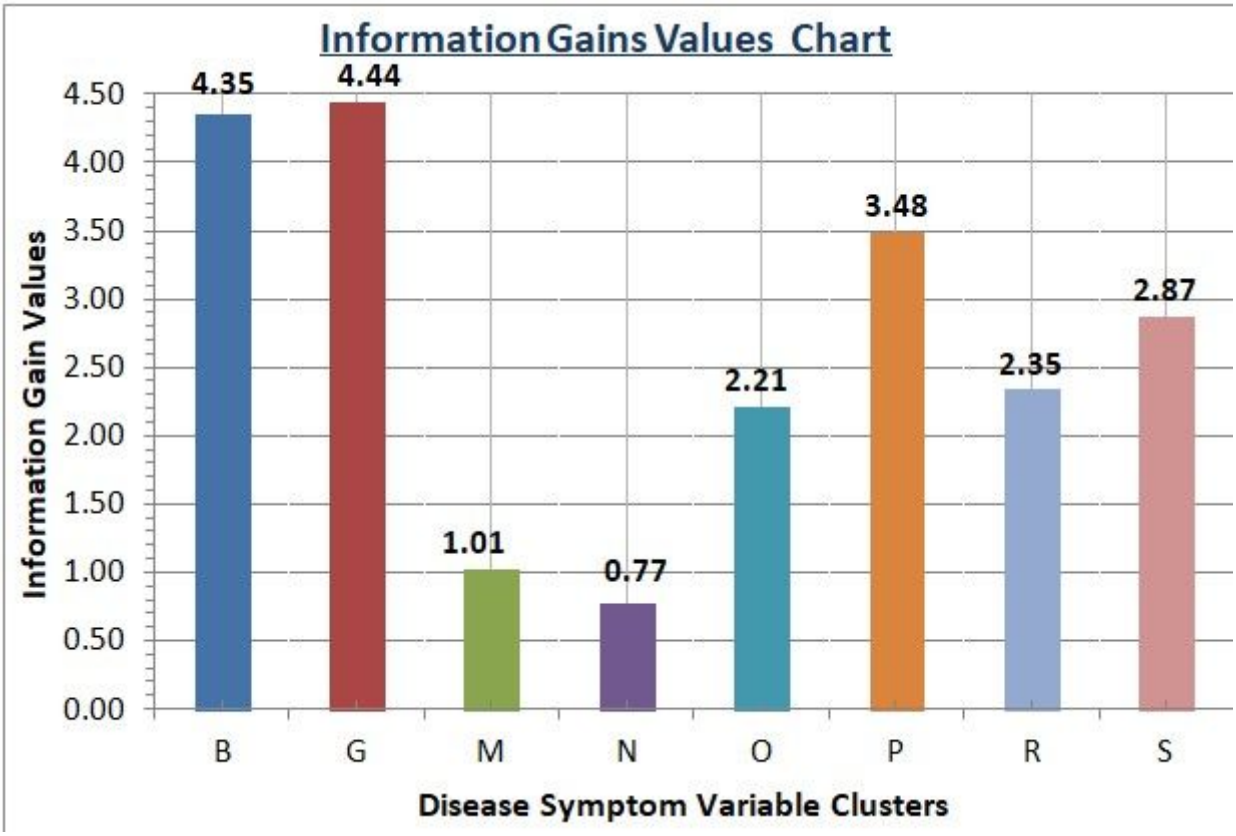

Figure 2

Symptom Variable Clusters' Information Gains Chart

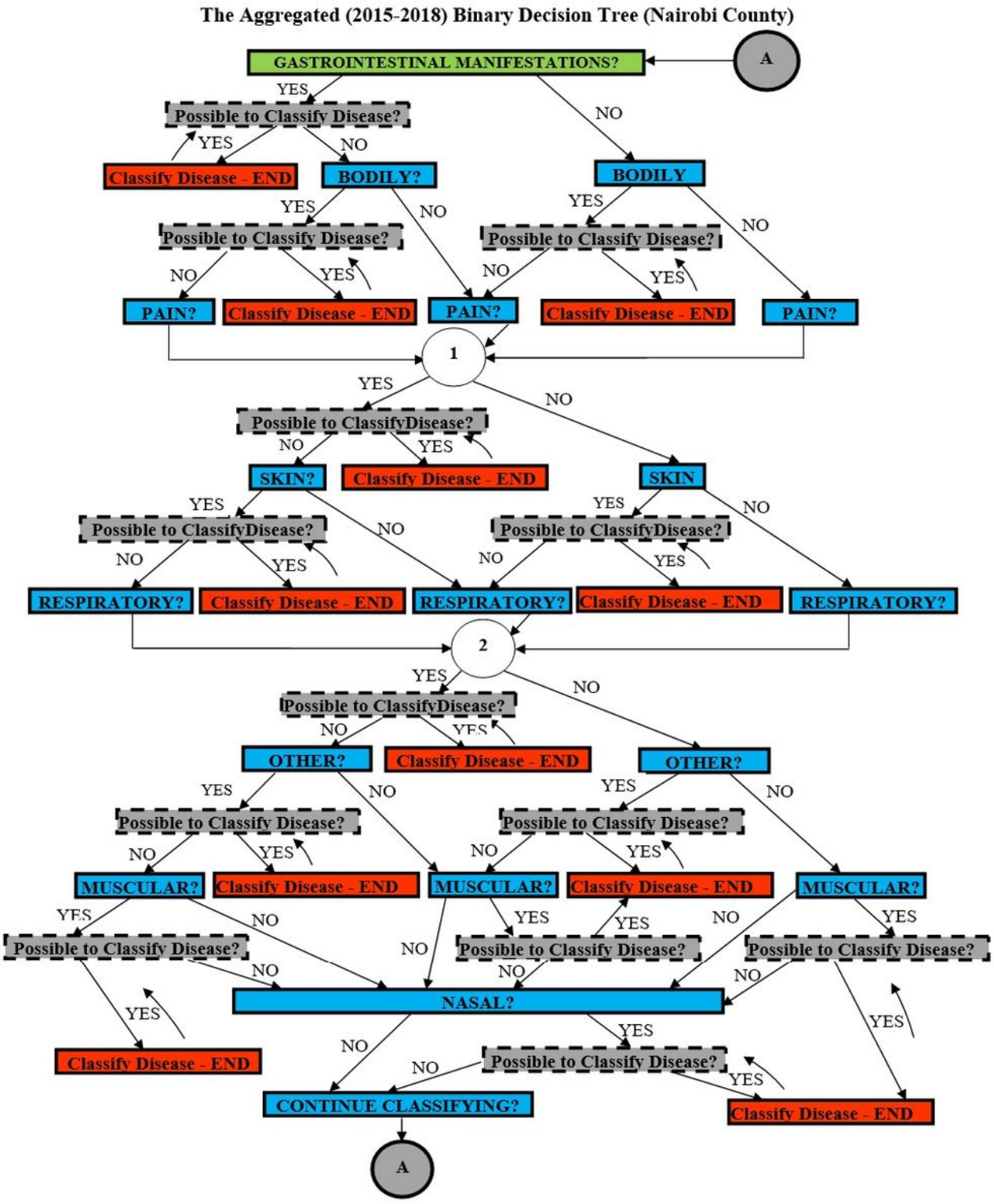

Figure 3

C4.5 Nairobi County Binary Decision Tree (1)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NairobiCountyRawData20152018.xls
- attachment.pdf