Journal of Health Informatics in Africa

Official Journal of the Pan African Health Informatics Association (HELINA)
published by Koegni-eHealth Innovation for Development
HELINA is the Africa Region of the International Medical Informatics Association (IMIA)

Volume 7 (2020) Issue 2

Eds: Nicky Mostert, Ulrich Kemloh

ISSN: 2197-6902
DOI: 10.12856/JHIA-2020-v7-i2

JHIA : www.jhia-online.org
HELINA : www.helina-online.org
IMIA : www.imia-medinfo.org

Published in the Journal of Health Informatics in Africa
Volume 7 (2020) Issue 2

ISSN: 2197-6902
ISSN/NLM abbreviation: J Health Inform Afr
DOI: http://dx.doi.org/10.12856/JHIA-2020-v7-i2

Publisher
Koegni eHealth, Innovation for Development e.V. Germany
Postfach 652166
D-22372 Hamburg, Germany
www.koegni-ehealth.org
E-mail: info@koegni-ehealth.org

Table of Contents

JHIA - online.org
Journal of Health Informatics in Africa

Editorial to JHIA Vol. 7 (2020) Issue 2

Nicky Mostert

Nelson Mandela University, Gqeberha, South Africa

The Journal of Health Informatics in Africa is the official journal of the Pan African Health Informatics Association (HELINA) and publishes the proceedings of the HELINA conferences, as well as open-call issues. This issue is the second open-call issue for 2020 comprising of five research papers submitted directly to the journal. These papers have been double blind peer-reviewed before being accepted for publication. Although papers written in French are also published by the journal, all five papers in this issue was written in English.

In addition to the five research papers, this issue also includes a Letter to the Editor that the JHIA Editorial Board considered very relevant and important to publish. The open letter was submitted by Neumann, Dunbar, Espino, Mtonga, and Douglas and highlights six themes that the authors consider important when developing Electronic Medical Records (EMRs) in low- and middle-income countries (LMIC). The authors believe that the consideration of these themes may support more careful exploration of creative, innovative, and sustainable EMR solutions that could ultimately improve patient outcomes through a positive impact on care delivery processes.

The paper by Doualla, Bediang, Nganou-Gnindjio, Boombhi, and Kingue presents the use of an electronic stethoscope coupled with tele-transmission and remote interpretation by a distant physician as a proof of concept of a tele-auscultation system in the assessment of cardiovascular diseases in remote areas of developing countries.

Msendema investigates the question of how research, policy and practice in health information systems interface to guide design and implementation of health information systems.

Authors Asah, Kanjo, Addo, Logo, Msendema explored the usability factors that influence the use of mobile technologies among healthcare staff at the point-of-care.

Bbosa, Wesonga, Nabende, and Nabukenya reports on a hybrid data mining technique for predicting reliable malaria incidence rate thresholds that they developed.

In their paper authors Olwendo, Otieno, and Rucha employed a retrospective cross-sectional study design in order to investigate the prevalence and complications associated with Diabetes Mellitus at the Nairobi Hospital in Kenya.

As we continue to live through the COVID 19 pandemic and its impact on both our work- and private lives, I realise the tremendous effort that went into completing this issue. The admirable commitment shown by authors that still managed to carefully prepare and submit manuscripts to JHIA, as well as the dedication of reliable peer-reviewers that took time out of their already busy schedules to assist with the review process. The editorial team also persevered under very difficult circumstances to manage the review and publication process and showed commendable determination to get this issue published. I would thus like to take the time to thank the editorial team, authors and peer reviewers that made this issue of JHIA possible. Your dedication, commitment, and perseverance is very much noted and appreciated.

Nicky Mostert
11.06.2021

JHIA - online.org
Journal of Health Informatics in Africa

A LMIC-First Manifesto to Developing Electronic Medical Record Systems

Christian Neumann[1], Elizabeth L. Dunbar[2], Jeremy U. Espino[1,3], Timothy M. Mtonga[1], Gerald P. Douglas[1,3]

[1]Global Health Informatics Institute, Lilongwe, Malawi
[2]Department of Human Centered Design and Engineering, University of Washington, Seattle, Washington, USA
[3]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, USA

Letter to the Editor

Having spent almost two decades developing electronic medical record (EMR) systems in low- and middle-income countries (LMIC), the inability to quickly design, develop systems and deploy new clinical guidelines remain bottlenecks to successful scale up and continued use. Significant gaps persist in EMR designs for LMIC settings. These gaps result from failure to understand how the environments, models and processes of delivering healthcare in LMICs are fundamentally different from high income countries [1], [2]. Additionally, EMRs are often planned in top-down fashion, driven by an implementers or external funder's vision. EMR systems for LMICs model those used in the global north, requiring heavy text data entry, complex data models, and heavyweight hardware to support. Here we outline a "LMIC-First" approach to designing EMR systems intended for LMICs. This approach considers the EMR system, its environment and how these interact when the system is implemented. The following six themes describe the proposed LMIC-First approach to EMR design:

Democratised EMR Development

EMR implementation in LMICs has mostly followed a project-based approach with independent contractors or implementing partners funded by donors working independently with little-to-no direct involvement from the Ministry of Health (MoH) staff in the host country [3], [4]. While this was done to fasttrack EMR implementation, it has led to reduced country ownership, increased dependency on external organizations, and protracted development cycles, sometimes including transitions between implementing partners and often exceeding donor funding cycles [5]. This approach has led to a graveyard of failed digital health systems where software or hardware breaks lead to their abandonment. Furthermore, reliance on independent contractors increases the risk of 'over-engineered' systems with inherent complexity that requires highly-qualified and specifically-trained technical staff to support and maintain. To reverse this worrying trend, MoH staff must be empowered with EMRs that have easy-to-use, built-in tools for addressing continually changing clinical landscapes and guidelines.

Process- & guideline-centric

Many attempts to build EMR systems for LMIC settings focus on data collection with limited consideration of the care process. Healthcare is delivered through a sequence of physical and mental tasks performed by multiple people in one or more work environments i.e. workflows [6], [7]. The sequences of tasks form the basic building blocks of healthcare delivery and define the data elements needed to complete and document the performance of an activity. These activities roll up into a care visit where ideally a clinician is presented with information to support best clinical decisions for a patient. Given the complexity of care, workflows further describe the different paths, through one or more branches, for completing an activity. Electronic

J Health Inform Afr. 2020;7(2):1-3. DOI: 10.12856/JHIA-2020-v7-i2-300

systems are most beneficial when they help the healthcare provider successfully navigate the care delivery workflow [8]. To provide value for the healthcare provider and patients, EMR systems must prioritise displaying accurate and timely information through well-designed workflows rather than being designed for data collection, billing, and reporting.

Point-of-care

EMRs hold promise for improving the quality and delivery of care when used by frontline healthcare workers during care delivery. Point-of-care use has the dual benefit of supporting improved quality of care through clinical decision support in addition to not requiring additional staff/time to perform data entry required for reporting. Point-of-care EMRs must offer value to clinicians to be used consistently and be designed in a way that does not interfere with care provision.

Touchscreen-first

Touchscreen user interfaces greatly reduce the need for hand-to-eye coordination over traditional mouse and keyboard interfaces and offer an intuitive user experience that facilitates learning and reduces the time it takes to gain proficiency with electronic systems [9]. Software initially designed without a touchscreen user interface may have significant limitations when later adapted for touchscreen [10]. Developing a native user interface around a well-defined set of criteria improves the usability and user experience of systems, which are critical factors in the success of EMR implementations [11].

Low cost

LMICs have limited healthcare resources. While electronic systems can improve the delivery and quality of care, this cannot be done at the expense of providing essential commodities such as medicine, diagnostic capabilities, or human resources. If the precious resources are to be spent on electronic systems, the total cost of buying and owning the system must be low and show positive return on investment. Furthermore, many EMR projects in LMIC start as pilot projects to assess the feasibility of different systems. Donors frequently pay for demonstration/pilot projects with the expectation that the host country governments will pay for nationwide scale-up and future maintenance of successful projects. However, little attention is paid during design-time to the overall cost of ownership of these solutions and what they entail for the limited resources available.

Low power

Continuous availability of electricity is a problem in many LMICs and their health facilities [12]. EMRs require electricity, preferably uninterrupted, to function. Frequent power outages remains a challenge to EMR use in many LMICs [4], [12]. As such, power backups are essential to ensure uninterrupted service. The cost of power backup can often exceed the cost of the computing hardware when power consumption is not considered. To make economical use of available power during prolonged power outages, low power devices (especially for workstations, servers, and network equipment) are preferable. Further consideration must be paid to the choice of power backups as traditional UPS systems are designed for infrequent and brief power outages unlike the frequent, prolonged outages that are common in LMICs.

We believe careful consideration of these themes will spark a rethink of choices and approaches when developing EMRs in LMICs. We acknowledge that the LMIC-first approach may result in different interpretations and implementations based on the context and the different problems that the implementations address. However, these themes support more careful exploration of creative, innovative, and sustainable EMR solutions that positively impact care delivery processes and ultimately, patient outcomes.

*Corresponding author address: christian.neumann@gmail.com

J Health Inform Afr. 2020;7(2):1-3. 10.12856/JHIA-2020-v7-i2-300.

References

[1] L. M. Puchalski Ritchie *et al.*, “Low- and middle-income countries face many common barriers to implementation of maternal health evidence products,” *Journal of Clinical Epidemiology*, vol. 76, pp. 229–237, Aug. 2016, doi: 10.1016/j.jclinepi.2016.02.017.

[2] L. Dornan, K. Pinyopornpanish, W. Jiraporncharoen, A. Hashmi, N. Dejkriengkraikul, and C. Angkurawaranon, “Utilisation of Electronic Health Records for Public Health in Asia: A Review of Success Factors and Potential Challenges,” *Biomed Res Int*, vol. 2019, Jul. 2019, doi: 10.1155/2019/7341841.

[3] F. F. Odekunle, R. O. Odekunle, and S. Shankar, “Why sub-Saharan Africa lags in electronic health record adoption and possible strategies to increase its adoption in this region,” *Int J Health Sci (Qassim)*, vol. 11, no. 4, pp. 59–64, 2017.

[4] M. O. Akanbi *et al.*, “Use of Electronic Health Records in sub-Saharan Africa: Progress and challenges,” *J Med Trop*, vol. 14, no. 1, pp. 1–6, 2012.

[5] P. Littlejohns, J. C. Wyatt, and L. Garvican, “Evaluating computerised health information systems: hard lessons still to be learnt,” *BMJ*, vol. 326, no. 7394, pp. 860–863, Apr. 2003, doi: 10.1136/bmj.326.7394.860.

[6] C. Cain and S. Haque, “Organizational Workflow and Its Impact on Work Quality,” in *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*, R. G. Hughes, Ed. Rockville (MD): Agency for Healthcare Research and Quality (US), 2008.

[7] “What is workflow? | AHRQ Digital Healthcare Research: Informing Improvement in Care Quality, Safety, and Efficiency.” https://digital.ahrq.gov/health-it-tools-and-resources/evaluation-resources/workflow-assessment-health-it-toolkit/workflow (accessed Feb. 08, 2021).

[8] C. P. Friedman, “What informatics is and isn’t,” *J Am Med Inform Assoc*, vol. 20, no. 2, pp. 224–226, 2013, doi: 10.1136/amiajnl-2012-001206.

[9] Z. L. Lewis, G. P. Douglas, V. Monaco, and R. S. Crowley, “Touchscreen task efficiency and learnability in an electronic medical record at the point-of-care,” *Stud Health Technol Inform*, vol. 160, no. Pt 1, pp. 101–105, 2010.

[10] G. P. Douglas, Z. Landis-Lewis, and H. Hochheiser, “Simplicity and usability: lessons from a touchscreen electronic medical record system in Malawi,” *interactions*, vol. 18, no. 6, pp. 50–53, Nov. 2011, doi: 10.1145/2029976.2029990.

[11] M. Kavuma, “The Usability of Electronic Medical Record Systems Implemented in Sub-Saharan Africa: A Literature Review of the Evidence,” *JMIR Hum Factors*, vol. 6, no. 1, Feb. 2019, doi: 10.2196/humanfactors.9317.

[12] S. Chawla *et al.*, “Electricity and generator availability in LMIC hospitals: improving access to safe surgery,” *J Surg Res*, vol. 223, pp. 136–141, Mar. 2018, doi: 10.1016/j.jss.2017.10.016.

*Corresponding author address: christian.neumann@gmail.com

JHIA - online.org
Journal of Health Informatics in Africa

Proof of Concept of the Contribution of Tele-auscultation in the Screening of Heart Disease: A Cross Sectional Study

Fred-Cyrille Goethe Doualla [a], Georges Bediang [a,*], Chris Nganou-Gnindjio [a], Jérôme Boombhi [a], Samuel Kingue [a].

[a] Faculty of Medicine and Biomedical Sciences, University of Yaoundé I, Cameroon

Background and purpose: Cardiovascular diseases are the leading cause of death worldwide, accounting for 31% of all deaths in 2016. Echocardiography is the reference screening tool, but its widespread use in developing countries is limited by its high cost. The objective of this study was to evaluate auscultation of heart sounds using an electronic stethoscope coupled with tele-transmission and remote interpretation by a distant physician as a less expensive alternative.

Methods: This was a descriptive cross-sectional study. Participants meeting inclusion criteria were examined face-to-face by a cardiologist (A) using a traditional stethoscope and following a well-defined protocol. Heart sounds were then recorded by a health professional using an electronic stethoscope. A part of these digital auscultation records were randomly selected and evaluated by cardiologist (A) and remotely (via a telemedicine platform) by cardiologist (B), then rated as normal or abnormal. Diagnostic findings of cardiologists A (digital-based) and cardiologist B (remote) were compared to those found by the cardiologist (A) during face-to-face consultation by using the Cohen's Kappa coefficient.

Results: We enrolled 22 patients in the study and ten (n=10) were randomly selected for analysis. The level of agreement between face-to-face and digital-based auscultatory findings by cardiologist A was moderate (K = 0.583). It was satisfactory (K= 0.615) between face-to-face auscultatory findings by cardiologist A and tele-auscultation findings by cardiologist B.

Conclusions: This study highlights the potential of using a tele-auscultation system in the assessment of cardiovascular diseases in remote areas (developing countries) where there is a shortage of qualified personnel.

Keywords: Telemedicine, Tele-auscultation, Screening, Heart disease, Electronic stethoscope

1 Introduction

Cardiovascular diseases are the leading causes of death worldwide, accounting for 31% of all deaths in 2016 [1]. Their high mortality rates (17.9 million deaths per year) and morbidity, particularly cardiac, are associated with high levels of disability and loss of productivity, exacerbating poverty and increasing health inequalities [2–5]. Reducing the impact of these diseases requires not only better care for known patients, but also early diagnosis of patients who ignore their condition [6,7].

Screening is of public health interest for these diseases because it reduces the associated burden. In resource-limited countries, due to the lack of infrastructure, qualified personnel, and financial resources, it is difficult for this screening to be based on the use of reference diagnostic tests such as echocardiography [4–8].

In such contexts, the practice of clinical auscultation remains the most frequently used alternative. It is a low-cost, non-invasive diagnostic technique that derives its effectiveness from the usual association between the sounds emitted by the heart, including when it is in good health, and the underlying clinical lesions [9,10]. Auscultation could, therefore, be considered as a tool for early detection of heart disease [10,11].

*Corresponding author address: Faculty of Medicine and Biomedical Sciences, University of Yaoundé I, Cameroon BP: 1364 Yaoundé, Cameroon, E-mail: bediang@yahoo.com

J Health Inform Afr. 2020;7(2):4-11. DOI: 10.12856/JHIA-2020-v7-i2-263

In addition, with the help of an electronic stethoscope that allows recordings even by a junior doctor, well-conducted auscultation coupled with the remote transmission of these recordings to a cardiologist would be an alternative. This application of telemedicine in auscultation (tele-auscultation) has many advantages in terms of effectiveness and efficiency [8]. This makes it possible to reduce the cost of diagnosis and patient follow-up, even in remote areas, while maintaining similar efficiency of the conventional face-to-face auscultation [13–15].

However, the effective implementation of such a remote monitoring system capable of providing assistance for diagnosis (through remote experts) and patient follow-up requires validation before being deployed for screening purposes. To date, there are few studies [13–16] conducted in developing countries in general, and in African countries in particular, on the possibilities offered by tele-auscultation. This study was conducted in order to implement such a remote monitoring system for the screening of heart disease.

The objective of this study is to evaluate the relevance of a remote auscultation monitoring system for its possible use in the screening of heart disease in a resource-limited country such as Cameroon.

2 Materials and methods

2.1 Design, period and setting

This was a descriptive cross-sectional study (fig. 1), which was done at the outpatient cardiology unit of the Yaoundé Central Hospital over one month (April 2017). Participants who met the inclusion criteria (22 cases) and gave their informed consent were examined (face-to-face cardiac auscultation) by a cardiologist (cardiologist A) using a traditional (non-electronic) stethoscope and following a well-defined protocol. These participants were auscultated a second time and their respective heart sounds were recorded by a health professional (non-specialist) using an electronic stethoscope. Subsequently, the recordings of some of these participants (10 cases) were randomly selected. These digital recordings were listened to and evaluated by cardiologist A and then sent to a second remote cardiologist (cardiologist B) via a telemedicine platform. The diagnostic conclusions of face-to-face auscultation and those from digital-based and remote auscultation (Cardiologist A and Cardiologist B respectively) were compared. The degree of intra and inter-rater agreement were estimated using Cohen's Kappa coefficient.

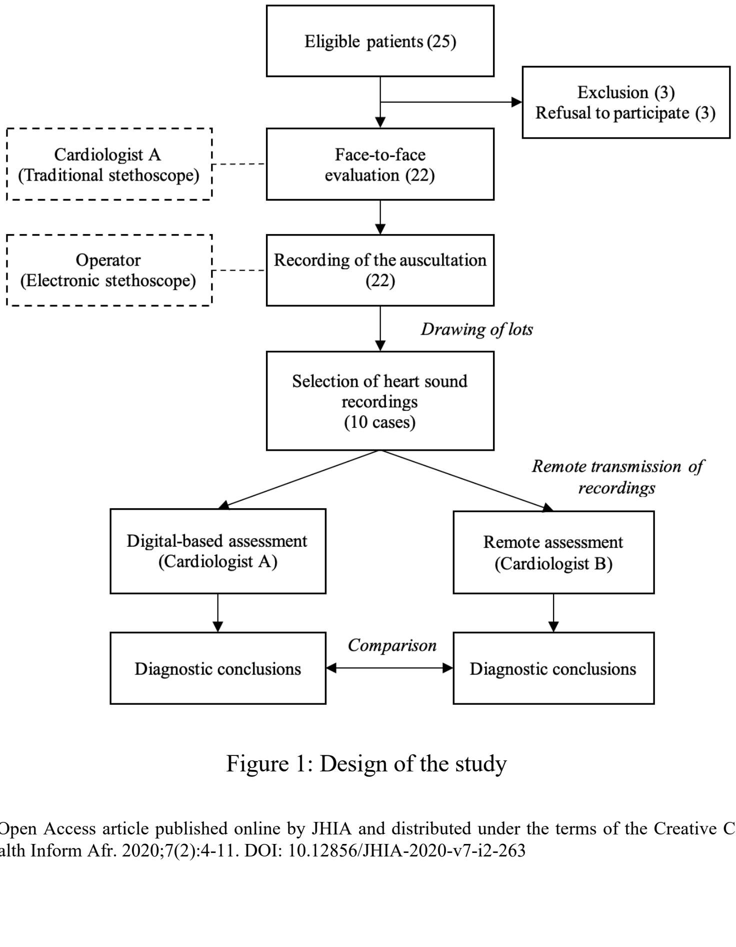

Eligible patients (25)
Exclusion (3)
Refusal to participate (3)
Cardiologist A
(Traditional stethoscope)
Face-to-face evaluation (22)
Operator
(Electronic stethoscope)
Recording of the auscultation (22)
Drawing of lots
Selection of heart sound recordings (10 cases)
Remote transmission of recordings
Digital-based assessment (Cardiologist A)
Remote assessment (Cardiologist B)
Diagnostic conclusions
Comparison
Diagnostic conclusions

Figure 1: Design of the study

2.2 Participants

All participants aged 21 years and above who came for a consultation in the cardiology outpatient department at the Central Hospital of Yaoundé, who did not have known heart disease and agreed to participate in the study were included. Participants whose audio recordings of the cardiac auscultation were of poor quality were excluded.

2.3 Sample and case description

Sampling was consecutive. Of the 25 patients who were eligible, twenty-two (22) adult patients were selected for the face-to-face cardiac auscultation phase, we had 11 patients with normal auscultation and 11 patients with auscultation abnormalities. Subsequently, 10 cases were randomly selected (by drawing lots) for digital-based and remote assessments by two cardiologists. The average age of the patients randomly selected was 47.3 years. Sixty (60) percent of the participants were of the male sex. The most common pathology found was high blood pressure (90%).

2.4 Selection of evaluators (judges)

Two evaluators (judges) were selected on the basis of their comparable experience in the medical field in general and in cardiology in particular. Cardiologist A has nine years of experience in the medical field, including eight years in cardiology. Cardiologist B has 11 years of experience in the medical field, including eight years in cardiology.

2.5 Procedure

Face-to-face evaluation and recording by an operator.
The twenty-two (22) patients selected were examined (cardiac auscultation) by a cardiologist (cardiologist A). This examination was done in person using a traditional stethoscope (non-electronic 3M Littmann® 2 178 Master brand). Five (5) cardiac areas (aortic, pulmonary, Erb, tricuspid and mitral) were systematically auscultated in all patients in the seated position. At the end of each auscultation, cardiologist A had to interpret and provide a diagnostic conclusion: either the auscultation was normal or abnormal. Using the same auscultation protocol, these patients were subsequently examined by a health professional (non-specialist) using an electronic stethoscope. The digital audio files obtained as a result of these tests were recorded on a computer.

Drawing of lots.
Patients were divided into 2 groups. The first group included patients whose cardiologist A's diagnostic conclusion at the end of the auscultation was normal, while the second group included patients whose diagnostic conclusion was abnormal. In each group, each patient received an identifier based on an alphanumeric code ranging from 1 to 11, which was associated with the letter "N" for patients with normal auscultations and "A" for patients with abnormal auscultations. Each code was written on paper and placed in two different ballot boxes. A person independent of the research team was invited to randomly draw 10 patients: 6 patients in the box containing patients with abnormal auscultations and 4 patients in the box containing patients with normal auscultations.

Digital-based and remote assessment of heart sound recordings.
Two weeks later, the recordings of the patients whose numbers were drawn at random were listened to by cardiologist A from a computer (digital-based assessment). At the end of this assessment, he had to conclude by stating if the examination was normal or abnormal. Subsequently, these recordings were transmitted for interpretation to another cardiologist (cardiologist B) remotely (remote assessment) via the Bogou telemedicine platform[17] (the telemedicine platform developed by the "Réseau en Afrique Francophone pour la Télémédecine") [18]. Cardiologist B interpreted these recordings, without knowing the diagnostic conclusions of cardiologist A. At the end of his assessment, cardiologist B was required to provide a diagnostic conclusion: either the auscultation is normal or abnormal.

Concordance between face-to-face, digital-based and remote assessments.

Table 1: Interpretation of the Kappa coefficient

Value	Degree of agreement
<0	Strong disagreement
0.00—0.20	Very poor agreement
0.21—0.40	Low agreement
0.41—0.60	Average agreement
0.61—0.80	Satisfactory agreement
0.81—1.00	Excellent agreement

The diagnostic conclusions from the face-to-face auscultation of cardiologist A were compared with those from the digital-based auscultation of cardiologist A (intra-rater agreement). The same applies to the face-to-face auscultation of cardiologist A and those from the remote auscultation of cardiologist B (inter-rater agreement). The degrees of intra and inter-rater agreement were thus assessed using Cohen's Kappa coefficient (Table 1) [19].

2.6 Data collection

The clinical diagnostic findings of each cardiologist were recorded on a specially designed data sheet.

2.7 Outcomes

The main outcomes were (i) the presence or absence of abnormalities on auscultation and (ii) the degrees of agreement within (intra-rater) and between (inter-rater) evaluators.

2.8 Data analysis

Data were entered using Epidata v.3.1 software and analysed using IBM-SPSS v.20 software for Windows. The categorical variables were represented by frequencies and percentages. The continuous variables were represented by their mean and standard deviation.

3 Results

3.1 Sociodemographic profile of participants and diagnostic conclusions (cardiologist A)

In total, cardiac auscultations of 22 patients were performed by cardiologist A (Table 2).

Table 2: Socio-demographic profile of participants

Variable		n (22)	Frequency (%)
Gender	Male	11	50
	Female	11	50
Profession of the participants	Retired	9	40.9
	Private (informal)	7	31.8
	Private (formal)	4	18.2
	Public	2	9.1
Main complain	HBP* monitoring	15	68.2
	Dyspnea	3	13.6
	Edema	3	13.6
	Other	1	4.5
Cardiac Auscultation	Normal**	11	50
	Abnormal***	11	50

Age (years)			**Mean (SD)**
			55.36 (15 683)

**High Blood Pressure; **Absence of pathological sound; ***Presence of pathological sound*

There were as many men as women. The average age was 55 years. The majority of participants were retired and the main health complaint was about high blood pressure. The most common auscultatory anomaly found was a systolic murmur at the mitral area.

3.2 Diagnostic conclusions and degree of agreement within and between judges

Of the 10 patients randomly selected, cardiologist A's diagnostic findings (face-to-face auscultation) were as follows: 6 patients had abnormal auscultations and 4 had normal auscultations. The anomalies found were essentially: decrease in intensity of the 1st and 2nd heart sounds, mainly in the aortic area; the presence of murmurs most often perceived in systole and audible in the mitral area in 46.1% of cases, in the tricuspid area in 30.8% of cases, and in the aortic area in 23.7% of cases.

Subsequently, the recordings of these patients were re-evaluated (digital-based assessment) by cardiologist A and the diagnostic findings were: 6 patients with normal auscultations and 4 with abnormal auscultations. A comparison of cardiologist A's diagnostic findings (face-to-face cardiac auscultation) with his own findings during the digital-based assessment (interpretation of audio recordings) showed that they were consistent for 3 normal and 5 abnormal auscultations (Table 3). The degree of agreement between these judgments was average (Kappa = 0.583).

Table 3: Intra-rater agreement of judgments (face-to-face versus digital-based assessment)

		Face-to-face auscultation (Cardiologist A)		**Kappa**
		Normal	Abnormal	
Digital-based auscultation (Cardiologist A)	Normal	3	1	0.583
	Abnormal	1	5	

For cardiologist B remote assessment (auscultation), the diagnostic conclusions after evaluation were as follows: 6 normal auscultations and 4 abnormal auscultations. Comparison of the diagnostic findings of cardiologist A (face-to-face cardiac auscultation) and cardiologist B during the remote assessment (tele-auscultation) showed that they were consistent for 4 patients in whom both did not find any abnormalities on auscultation and for 4 others in whom they all found abnormalities (Table 4). The degree of agreement between these judgments was satisfactory (Kappa = 0.615).

Table 4: Inter-rater agreement of judgments (face-to-face versus remote assessment)

		Face-to-face auscultation (Cardiologist A)		**Kappa**
		Normal	Abnormal	
Remote auscultation (Cardiologist B)	**Normal**	4	2	0.615
	Abnormal	0	4	

4 Discussion

Tele-auscultation is a diagnostic method that combines face-to-face auscultation with the use of information and communication technologies to remotely transmit auscultation records for diagnostic purposes or for patient follow-up. It is based on the use of a connected electronic stethoscope.

Several studies have shown that this approach can be effectively implemented to diagnose heart diseases [15] [20–22].

In our study, the degree of intra-rater agreement between the diagnostic conclusions obtained via face-to-face auscultation and those resulting from digital-based interpretation of heart sound records were average (Kappa = 0.583). This degree of agreement is lower than that of Dahl and al, who found a Kappa of 0.87 [22]. This difference may be due to the fact that in our study, cardiologist A had no clinical information about the participants whose digital heart sound recordings he was evaluating, unlike Dahl and al where the age, symptoms and clinical signs of the patients were associated with the audio recordings. During auscultation, the fact that a physician has access to a patient's clinical information improves decision-making (diagnostic, therapeutic, etc.) [23]. In addition, the electronic stethoscope is a sensitive tool and can perceive sounds that are not perceptible when using a conventional stethoscope.

The degree of inter-rater agreement was based on the comparison between the diagnostic conclusions obtained through the remote analysis of digital heart sound recordings (remote auscultation) by cardiologist B and the face-to-face auscultation by cardiologist A. This degree of agreement was satisfactory (Kappa = 0.615). This level of agreement is similar to that found by Belmont et al who found a Kappa equal to 0.77 although he only evaluated the presence or not of a systolic regurgitation murmur [21]. However, it is lower than that found by Dahl et al. whose Kappa was 0.81 (inter-rater) [15]. These results can be explained by several factors.

First, there is a lack of training and education in tele-auscultation. The introduction of this innovative approach (use of the electronic stethoscope and remote analysis), requires carrying out capacity building sessions beforehand, in order to give doctors all the skills they need to better appropriate the device. In this study, both cardiologists A and B did not receive any prior training sessions on how to listen to digital heart sound recordings. Given the similarities with face-to-face auscultation using a conventional stethoscope, we hypothesised that physicians' appropriation of this digital-based or remote auscultation would be natural and intuitive.

In addition, there are artefacts that may compromise the digital-based or remote analysis of heart sound recordings. Comments were made by the various cardiologists on this subject. Subject to excellent suppression of ambient noise (and therefore better quality of heart sound recordings), a better agreement can be achieved between the conclusions from the face-to-face assessment of patients using the conventional stethoscope and the conclusions from the digital-based or remote assessment (tele-auscultation) of heart sound recordings using the electronic stethoscope [24].

Finally, the residual variability between evaluators can also be explained by the difference in academic background and professional experience between cardiologists [21]. Experienced clinicians will tend to agree strongly when the diagnosis is simple and obvious [21]. However, an agreement between evaluators decreases significantly as the clinical signs to be evaluated become hard to perceive, as this requires the use of personal experience [21]. In this study, the two cardiologists selected had almost the same length of practice in the profession: 9 years for cardiologist A and 11 years for cardiologist B.

5 Conclusion

Although it is not the standard for the diagnosis of cardiac pathologies as it is for cardiac ultrasound, face-to-face auscultation remains the most cost-effective diagnostic method, especially in the context of resource-limited countries. It is a non-invasive and inexpensive diagnostic approach. With the continuous development of information and communication technologies, this face-to-face auscultation can now be done remotely (remote auscultation) through the use of electronic stethoscopes and telemedicine platforms. This study highlighted the possibility of using remote auscultation as a means of assessing patients in a context of limited resources. The degree of agreement between the judgement of the remote cardiologist and that of the face-to-face cardiologist underlines the potential of this method. However, additional studies on a larger scale with a gold standard (cardiac ultrasound) should be considered for better evidences.

Study Limitations

This study has some limitation: confirmatory cardiac ultrasound was not performed for cases evaluated by cardiologist A during the face-to-face cardiac auscultation. This was not possible because of the high cost of this assessment in our context and the budgetary constraints of our study. To mitigate those limitation, we considered the clinical advice of the cardiologist at the participant's bedside to be close to reality and the audio recordings of the heart sounds were transmitted unchanged to the remote cardiologist using a secure platform.

Acknowledgements

None.

Statement on conflicts of interest

The authors declare no competing interest.

References

[1] WHO. Cardiovascular diseases [Internet]. [cited 2019 Dec 20]. Available from: https://www.who.int/westernpacific/health-topics/cardiovascular-diseases

[2] Cappuccio FP, Miller MA. Cardiovascular disease and hypertension in sub-Saharan Africa: burden, risk and interventions. Intern Emerg Med. 2016 Apr;11(3):299–305. https://doi.org/10.1007/s11739-016-1423-9

[3] Burroughs Pena MS, Bloomfield GS. Cardiovascular Disease Research and the Development Agenda in Low- and Middle-Income Countries. Glob Heart. 2015 Mar;10(1):71–3. https://doi.org/10.1016/j.gheart.2014.12.006

[4] Gheorghe A, Griffiths U, Murphy A, Legido-Quigley H, Lamptey P, Perel P. The economic burden of cardiovascular disease and hypertension in low- and middle-income countries: a systematic review. BMC Public Health. 2018 Dec;18(1):975. https://doi.org/10.1186/s12889-018-5806-x

[5] Di Cesare M, Khang Y-H, Asaria P, Blakely T, Cowan MJ, Farzadfar F, et al. Inequalities in non-communicable diseases and effective responses. The Lancet. 2013 Feb;381(9866):585–97. https://doi.org/10.1016/S0140-6736(12)61851-0

[6] Reddy KS. Cardiovascular diseases in the developing countries: dimensions, determinants, dynamics and directions for public health action. Public Health Nutr. 2002;5(1a):231–7. https://doi.org/10.1079/phn2001298

[7] Ezzati M, Pearson-Stuttard J, Bennett JE, Mathers CD. Acting on non-communicable diseases in low- and middle-income tropical countries. Nature. 2018 Jul;559(7715):507–16. https://doi.org/10.1038/s41586-018-0306-9

[8] Prabhakaran D, Anand S, Watkins D, Gaziano T, Wu Y, Mbanya JC, et al. Cardiovascular, respiratory, and related disorders: key messages from Disease Control Priorities, 3rd edition. The Lancet. 2018 Mar;391(10126):1224–36.

[9] Cardiol CP. Cardiac auscultation. A cost-effective diagnostic skill. Curr Probl Cardiol. 1995;20(7):447–530.

[10] Ansa VO, Odigwe CO, Agbulu RO, Odudu-Umoh I, Uhegbu V, Ekripko U. The clinical utility of echocardiography as a cardiological diagnostic tool in poor resource settings. Niger J Clin Pract. 2013 Mar;16(1):82–5. https://doi.org/ 10.4103/1119-3077.106772

[11] Nellen M, Maurer B, Goodwin JF. Value of physical examination in acute myocardial infarction. Br Heart J. 1973;35(8):777–80. https://doi.org/10.1136/hrt.35.8.777

[12] Peacock WF, Harrison A, Moffa D. Clinical and economic benefits of using AUDICOR S3 detection for diagnosis and treatment of acute decompensated heart failure. Congest Heart Fail. 2006;12 Suppl 1(April):32–6. https://doi.org/10.1111/j.1527-5299.2006.05772.x

[13] Jerant AF, Azari R, Nesbitt TS. Reducing the cost of frequent hospital admissions for congestive heart failure: A randomized trial of a home telecare intervention. Med Care. 2001 Nov;39(11):1234–45. https://doi.org/10.1097/00005650-200111000-00010

[14] Zenk BM, Bratton RL, Flipse TR, Page EE. Accuracy of detecting irregular cardiac rhythms via telemedicine. J Telemed Telecare. 2004;10(1):55–8. https://doi.org/10.1258/135763304322764211

[15] Dahl LB, Hasvold P, Arild E, Hasvold T. Heart murmurs recorded by a sensor based electronic stethoscope and e-mailed for remote assessment. Arch Dis Child. 2002;87(4):297–300. https://doi.org/10.1136/adc.87.4.297

[16] Johanson M, Gustafsson M, Johansson LÅ. A Remote Auscultation Tool for Advanced Home Health-Care. J Telemed Telecare. 2002 Aug 10;8(2):45–7. https://doi.org/10.1258/135763302320301975

[17] RAFT. Bogou-Logiciel de télédiagnostic [Internet]. [cited 2020 January 30]. Available from: raft1.unige.ch/bogou/

[18] RAFT. Présentation du Réseau en Afrique Francophone pour la Télémédecine [Internet]. [cited 2020 January 30]. Available from: raft.g2hp.net/presentation/

[19] Bergeri I, Michel R, Boutin J-P. Pour tout savoir ou presque sur le coefficient de Kappa. Med Trop. 2002;62(December):634–6.

[20] Finley JP, Warren AE, Sharratt GP, Amit M. Assessing children's heart sounds at a distance with digital recordings. Pediatrics. 2006;118(6):2322–5. https://doi.org/10.1542/peds.2006-1557

[21] Belmont JM, Mattioli LF, Goertz KK, Ardinger RH, Thomas CM. Evaluation of Remote Stethoscopy for Pediatric Telecardiology. Telemed J. 1995;1(2):133–49. https://doi.org/10.1089/tmj.1.1995.1.133

[22] Dahl LB, Hasvold P, Arild E, Hasvold T. May heart murmurs be assessed by telemedicine? Tidsskr Den Nor Laegeforening Tidsskr Prakt Med Ny Raekke. 2003 Nov 6;123(21):3021–3.

[23] Chantepie A, Soulé N, Poinsot J, Vaillant MC, Lefort B. Souffle cardiaque chez l'enfant asymptomatique : Quand demander un avis cardiologique ? Arch Pediatr. 2016;23(1):97–104. https://doi.org/10.1016/j.arcped.2015.10.006

[24] Gigstad L. A comparison of an acoustic stethoscope and an amplified stethoscope in white noise and cafeteria noise during cardiac auscultation. Portland, OR; 2000 Jan. https://doi.org/10.15760/etd.5855

JHIA - online.org Journal of Health Informatics in Africa

Interfacing Research and Policy in Informing Data Management for Quality Data in Health Information Systems: Case of DHIS2 in Malawi

Martin Bright Msendema

PhD student (Applied Information Technology)
Faculty of Applied Sciences, Computing and Information Technology Department,
University of Malawi, The Polytechnic.

Introduction
Like in many other fields, there is an ever-growing need for new knowledge in health services delivery on how health service providers should improve their decisions in delivering health care. In attempting to respond to that desire for knowledge, universities and other institutions have been funding and motivating students and academicians carry out research in health management information systems. The expectation is that the knowledge generated from the research will inform improvement in HIS practices.

Method
This qualitative desk research attempted to answer the question how research, policy and practice in health information system interface to inform design and implementation of health information systems. The study adopted an intrinsic case study methodology. The study used District Health Information Systems 2 in Malawi as a case. It reviewed PhD and masters students' theses, Health Information System Policy and Strategy documents to explore how these resources inform each other. Data analysis was done through thematic tabular analysis, and themes were derived from a predefined set of criteria.

Results
The findings showed that there is considerable effort by researchers to publish and share their findings with practitioners through conferences, journals and working together in workshops. It also shows that participation in a policy formulation workshop has been a key means by which researchers directly contribute to health information systems management practices.

Conclusion
The paper has attempted to answer the question how research, policy and practice in health information system interface to guide design and implementation of health information systems. The findings have shown that there are considerable efforts by stakeholders in health service delivery to create and make use of platforms that should enable the interaction between researchers and practitioners. Apart from that, the findings have also stimulated a need to conduct a detailed field study to ascertain how actually the researchers inform practitioners on the ground.

Keywords: Research, Health Information System Policy, Practice, DHIS2

1 Introduction

One of the areas where Information and Communications Technology for Development has dominated debate is health service delivery. Access to quality data for decision making and telemedicine are among notable themes in literature [1], [2]. In data management, the work of Weiskopf and Weng [3] has led to the development of a data quality model in health information systems. This model and others have been benchmarks for defining practices and implementation of systems for data management with aim of attaining quality data for decision making. Sahay and Walsham [4] demonstrated the various ways through which health information systems contribute towards a better world for all. In their study titled *"Building a Better World: Frugal Hospital Information Systems in an Indian State*" they highlighted ability of health

*Corresponding author address: mmsendema@poly.ac.mw

J Health Inform Afr. 2020;7(2):12-17. DOI: 10.12856/JHIA-2020-v7-i2-272

information systems in strengthening processes to include the disadvantaged, empowering patients through access to information and use of technology to make voice of the voiceless in the rural areas be heard.

Researchers have also argued that through e-health, for example, there is potential for Information and Communication Technology (ICT) to expand access and improve efficiency in health service delivery especially in rural areas [5], [6]. Similarly, work of Thapa and Sein [2] articulated how affordances of ICT are actualised using a telemedicine case in Nepal. The literature above demonstrates that there is a lot of research work towards understanding how ICT is important in health service delivery. However unlike in other disciplines like education, for example, not much has been documented in relation to the link between research, policy and practice.

Through Health Information Systems Program (HISP), under the leadership of University of Oslo in Norway, there has been active research by students and professors which has seen the inception and evolution of the District Health Information System (DIHS) across the developing countries. DHIS, now at version 2, provides a platform for data management for Health Information Systems in many countries across the globe from Africa to Asia [7]. Despite that the research findings and recommendations have been published through conferences, journals and books, there has not been substantial research and reports on how actually these findings inform policy formulation and indeed the development of features in the DHIS as it evolves. This desk study was motivated by an attempt to explore how researchers and practitioners interface to translate research knowledge into an improvement in health information systems. The research uses a case study of DHIS in Malawi.

1.1 Problem statement

Literature shows there has been a lot of research on the role information technology in health service delivery [6], [8] [9]. Health Information Systems, data quality and eHealth are among areas which have been debated by many scholars in journals and conferences. Similarly governments' health departments and health partners have been articulating health information systems policies and strategies to govern the use and application of information technology in delivery of quality health services [10], [11], [12]. However, despite the enriching research findings, policies and strategies, not much has been researched or say documented on how these resources interface each other in attempt to achieve a common goal of quality and accessible health service to all. This concern motivated a systematic tracking of research and related health information system policies and strategies so as to understand how they talk to each other in trying to achieve a common goal, particularly of attaining quality data which should inform decision making for improved access to quality health for all.

1.2 Aim of the study

This desk research aimed to explore how research work practically contribute towards policy and strategy formulation and implementation. Specifically; it aimed to find out how the development of policies and strategies benefits from research findings. It further aimed to find out which features in DHIS2 have indeed be implemented in response to the policies and strategies.

1.3 Conceptual framework

In an attempt to shape and structure the relationship among the concepts being studied, the researcher developed a conceptual framework shown in *Figure 1.* This artefact was drawn from the work of Miles and Huberman [13] who defined a conceptual framework as "*a visual or written product, one that "explains, either graphically or in narrative form, the main things to be studied".*

The main concepts in the framework are *research*, *practice* and *product.* Where research points to the a systematic investigation to discover facts or collect information [14] and practice is being looked at as the actual application or use of an idea, belief, or method, as opposed to theories relating to it [15]. From the same dictionary the researcher defines product as a thing or person that is the result of an action or process [15], which in this study relates to such things as agreed practices and information systems.

J Health Inform Afr. 2020;7(1):12-17. DOI: 10.12856/JHIA-2020-v7-i2-272

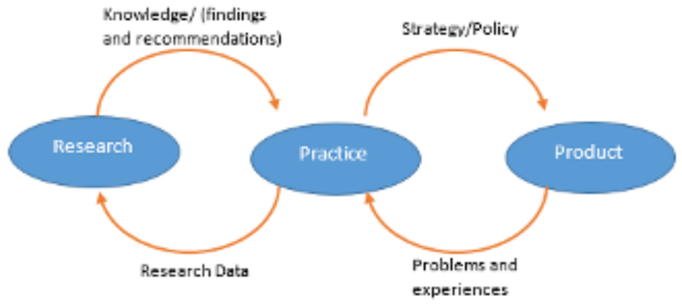


Figure 1: *Conceptual Framework*

2 Methodology

This study was a qualitative desk research which used intrinsic case study methodology. It attempted to answer the question how research, policy and practice in health information system interface to inform design and implementation of health information systems. Purposive random sampling was used to select journal or conference articles and policy or strategic documents. PhD and masters students theses also formed part of the data sources consulted where the main inclusion criteria was that studies or documentation should relate to health information systems development and implementation with reference to District Health Information Systems (DHIS 2) in Malawi. The researcher focused on documents within 2010 and 2016 timeline and 2015 to 2018 research work and strategy documents respectively. The systematic search was however not limited to local sources and it extended to international journals as long as they conformed to the set criteria. Table 1 gives a summary of the artefacts that were consulted in the study.

Table 1: *Summary of data sources*

	Thesis		***Article***		***Reports***	
Title	PhD	MSc.	Journal	Conference	Strategic	Policy
1. Monitoring, Evaluation and Health Information Systems Strategy (MEHIS)-2018					√	
2. Natinal Health Information System Policy-2015						√
3. Implications of Integrating Information Systems in Healthcare at District Level in Malawi: A Case of DHIS and Drug LMIS-2010		√				
4. Management and Use of Health Information in Malawi and Burkina Faso: The Role of Technology-2016				√		
5. The information transparency effects of introducing league tables in the health system in malawi-2016			√			
6. Strengthening Health Management Information Systems in Malawi: Gaps and Opportunities-2015				√		
7. Developing Integrated National Health Information Systems in Malawi: Challenges and South-South Collaboration-2011				√		
8. In Search of the Missing Data :The case of maternal and child health data in Malawi-2010	√					
9. Transformational Feedback: Breaking the vicious cycle of information use in Health Information Systems-A case from Malawi-2016	√					

Nine (9) documents were collected online and the researcher focused on the recommendations. The researcher summarised the crucial recommendations which were then used as a benchmark for evaluating the health information systems strategy or policy documents and the DHSI 2. The evaluation was to ascertain if these three resources: research recommendations, strategy or policy and technology (DHIS 2) implementation talk to each other. For each specific recommendation, the researcher checked for either a corresponding policy or strategic item in the strategic or policy document. Wherever relevant, the implementation of the same was checked in DHIS 2 in form of either feature or functionality. The same approach was also used to check practices in terms of use along the data processing chain in the DHIS 2.

3 Findings

The paper titled *"Strengthening Health Management Information Systems in Malawi: Gaps and Opportunities"* [16] provided a very good structure of the findings in this study. Without being specific as to which paper, the finding in that paper guided the thematic structuring of the findings in this study. Three themes were hence drawn from the study of the various data sources presented in methodology section: (i.) improving data processing practices and use (ii.) inclusion of informal data sources (iii.) capitalising on the strength of integration in data management systems.

The proponents of improving data processing practices and use recommended support to all involved in the data processing. The support highlighted included training of personnel or providing self-explanatory kit that users can train independently, introducing mobile app for reporting, peer based reviews, league tables, transformational feedback and information behaviour culture and use especially among managers [1] [16] [17] [18] as the aim of introducing health information systems and platforms like DHIS 2 is to get quality data, others argued that data cannot be complete if informal sources are not included. This led the researchers to recommend inclusion of informal sources of data like births at Traditional birth attendants [19]. Another interesting finding was about capitalising on the strengths of integration in data management systems. In their studies [20] [21] argued that several information systems especially at district levels operate independently. Duplicate data within the same institution was reiterated as a common concern arising from such a situation. Although both authors focus was more leaned towards integration approach, they articulated the need for integrating health information system and this work focused on the latter than the former.

Moving on to the findings in relation to practitioners, the paper starts by defining strategy and policy. Among the many definitions, strategy is defined as "*a plan of action designed to achieve a specific goal* [22] and policy is defined as *"a set of ideas or plans that is used as a basis for making decisions, especially in politics, economics, or business"* [23]. Although there are technical differences but for purpose of our study, we focused on plan of actions or ideas meant to achieve a certain goal, which in this case is the goal is to improve data quality for decision making.

Common findings between the two documents included actions to: improve interoperability (this was meant to reduce increased independent data management platforms), enhance continuous support across all levels of data management (through training and mentorship), strengthening capacity to use data, actions to reduce workload especially to those in data collection and finally strengthening community structures including chief in supervision of data collection and submission [10] [12].

Another finding that is worth highlighting was that the strategy and policy documents demonstrated evidence of a collaboration culture among the practitioners and the researchers, where minutes of policy and the strategy development activities indicated presence of some of the researchers in the formulation workshops and DHIS 2 platform development processes. ***Figure 2*** summaries the overall findings showing the alignment of the three artefacts.

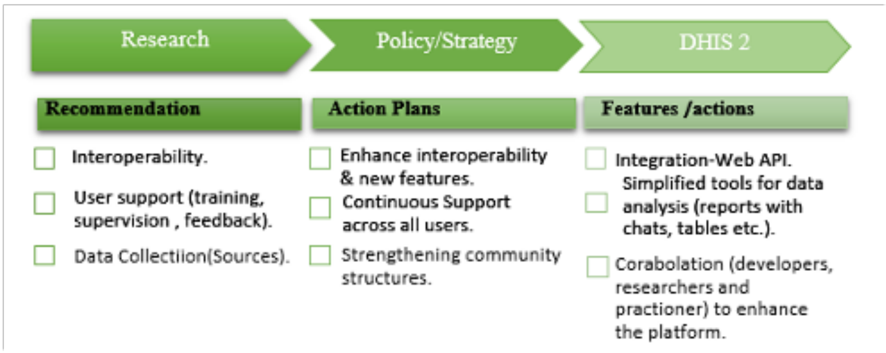

Research
Policy/Strategy
DHIS 2
Recommendation
Action Plans
Features /actions
□ Interoperability.
□ User support (training, supervision , feedback).
□ Data Collectiion(Sources).
□ Enhance interoperability & new features.
□ Continuous Support across all users.
□ Strengthening community structures.
□ Integration-Web API.
□ Simplified tools for data analysis (reports with chats, tables etc.).
□ Corabolation (developers, researchers and practioner) to enhance the platform.

Figure 2: *Graphical illustration of research, policy and strategy and DHIS2 platform alignment*

4 Discussion and conclusion

Looking at how the policy and strategic action plans are well aligned to researchers' recommendations gives a substantial evidence that the practitioners work is being guided by the researchers' findings and recommendations. There is considerable effort by stakeholders to create and make use of platforms that should enable the link between researchers and practitioners [24]. The symbiotic relationship between researchers and practitioners propels the vehicle of knowledge to new destinations. Researchers formulate research problems from practitioners' experiences. Through systematic studies, the researchers are able to find answers to problems faced by practitioners [25]. However, publishing research findings in journal articles, or participating in conferences or other research dissemination seminars, does not always guarantee that the knowledge will be applied, there are so many factors which may motivate practitioners to put the new ideas into use. Also as illustrated in the conceptual model, product of policy and research, may not always be tangible like an information system. At times it could be a set of practices which are difficult to observe on a desk research if they are really being implemented. Although the findings cannot be generalised as in any intrinsic case study [26] , I posit that research, policy and health information systems developers interface mostly through workshops than in conferences or reading journal articles in Malawi. The findings are evident in other studies which have highlighted how workshops have been integral in knowledge transfer between researchers and practitioners [27] [28]. For example, studies in Burkina Faso showed that workshops involving researchers and practitioners have helped to increase chances of knowledge to be put into practice [29]. However this differs from Kenyan experiences where publications in peer reviewed journals and conferences have been acknowledged as common means of dissemination [30]. The researcher call for further research to ascertain how local conferences can best be capitalised to enhance the existing means of sharing knowledge and implementing new ideas in Malawi Health Management Information Systems.

Acknowledgements

I am thankful to Mr. Maclan Kanyang'wa (University of Malawi, Journalism Department) for editing my work and Professor Jens Kaasbøll (University of Oslo) for his encouragement.

References

[1] P. A. Chikumba and S. L. Ramussen, "Management and Use of Health Information in Malawi and Burkina Faso:The Role of Technology," in *ICT Africa*, Durban, 2016.

[2] D. Thapa and K. M. Sein, "Trajectory of Affordances: Insights from a case of telemedicine in Nepal," *Information Systems Journal,* pp. 796-819, 2018.

[3] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of the American Medical Informatics Association,* pp. 144-151, 2013.

[4] S. Sahay and G. Walsham, "Building a Better World: Frugal Hospital Information Systems in an Indian State," in *Thirty Fifth International Conference on Information Systems*, Auckland, 2014.

[5] S. Sahay, P. Nielsen, D. Faujdar, R. Kumar and A. Mukherjee, "Frugal Digital Innovation and Living Labs: A Case Study of Innovation in Public Health in India," in *Thirty Ninth International Conference on Information Systems*, San Francisco, 2018.

[6] T. Mettler, P. Rohner and L. Baacke, ""Improving Data Quality of Health Information Systems: A Holistic DesignOriented," in *European Conference on Information Systems*, 2008.

[7] HISP, "Who is HISP," 2015. [Online]. Available: https://www.hisp.org/. [Accessed 09 July 2019].

[8] C. N. Chaulagai, C. M. Moyo, J. Koot, H. B. Moyo, T. C. Sambakunsi, F. M. Khunga and P. D. Naphini, "Design and implementation of a health management information system in Malawi: issues, innovations and results," *The London School of Hygiene and Tropical Medicine.,* 2005.

DOI: 10.12856/JHIA-2020-v7-i2-272

[9] B. Chaudhry, J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S. C. Morton and P. G. Shekelle, "Systematic Review: Impact of Health Information Technology on Quality, Efficiency, and Costs of Medical Care," *Annals of Internal Medicine,* vol. 144, pp. 742-752, 2000.

[10] Ministry of Health-Malawi, "Malawi National Health Information System Policy," Ministry of Health, Lilongwe, 2015.

[11] Ministry of Health, "Malawi National Informaiton System Policy," Ministry of Health, 2015, 2015.

[12] Ministry of Health -Malawi, "Monitoring, Evaluation and Health Information Systems Strategy (MEHIS)," Minitsry of Health, Lilongwe, 2018.

[13] M. B. Miles and A. M. Huberman, Qualitative data analysis: An expanded sourcebook (2nd ed.), Thousand Oaks, CA: Sage, 1994.

[14] C. E. Dictinary, Collins English Dictionary, Bangor: HarperCollins, 2010.

[15] O. Dictionary, "Online Dictionary," [Online]. Available: the actual application or use of an idea, belief, or method, as opposed to theories relating to it. [Accessed 2019 July 2019].

[16] M. Monawe, M. G. Chawai, G. Kapokosa and C. Moyo, "Strengthening Health Management Information Systems in Malawi: Gaps and Opportunities," in *IST-Africa*, Durban, 2016.

[17] C. Moyo, J. easboll, P. Nielsen and J. Saebo, "The information transparency effects of introducing league tables in the information transparency effects of introducing league system in Malawi," *The Electronic Journal of Information Systems in Developing Countries,* vol. 75, no. 2, pp. 1-16, 2016.

[18] C. Moyo, "Transformational Feedback: Breaking the vicious cycle of information use in Health Information Systems-A case from Malawi," University of Oslo, Olso, 2016.

[19] C. Kanjo, "In Search of the Missing Data , The case of maternal and child health data in Malawi," University of Oslo, Oslo, 2012.

[20] C. Kanjo, J. Braa, E. Kossi and C. M. Moyo, "Developing Integrated National Health Information Systems in Malawi: Challenges and South-South Collaboration," Cape Town, 2010.

[21] P. A. Chikumba and A. N. Kaunda, "Implications of Integrating Information Systems in Healthcare at District Level in Malawi: A Case of DHIS and Drug LMIS," in *AFRICOMM 2012: e-Infrastructure and e-Services for Developing Countries* , 2012.

[22] Simply Strategic Planning, "What is strategy," Simply Strategic Planning, 2017. [Online]. Available: http://www.simply-strategic-planning.com/what-is-strategy.html. [Accessed 23 September 2019].

[23] Collins, "Collins," Collins, 2008. [Online]. Available: https://www.collinsdictionary.com/dictionary/english/policy. [Accessed 23 September 2019].

[24] F. A. J. Korthagen, "The Relationship Between Theory and Practice in Teacher Education," *International Encyclopedia of Education,* vol. 7, pp. 669-675, 2010.

[25] Veloza-Gomez M, "The Research-Theory-Practice Relationship a Reference for the Discipline of Nursing," *Annals of Nursing Research and Practice,* 02 July 2016.

[26] R. E. Stake, "Qualitative Case Studies. In N. K. Denzin & Y. S.," in *Handbook of qualitative research*, Sage Publications Ltd., 2005, p. 443–466.

[27] T. K. McGee, A. Curtis, B. . L. McFarlane, B. Shindler, A. Christianson, C. Olsen and S. McCaffrey, "Facilitating knowledge transfer between researchers and wildfire practitioners about trust: An international case study," *The Forestry Chronicle,* vol. 92, no. 2, pp. 161-171, 2016.

[28] Y. Higuchi and Y. Yamanaka, "Knowledge sharing between academic researchers and tourism practitioners: a Japanese study of the practical value of embeddedness, trust and cocreation," *Journal of Sustainable Tourism,* vol. 25, no. 10, p. 1456–1473, 2017.

[29] E. Mc Sween-Cadieux, C. Dagenais, P.-A. Somé and V. Ridde, "Research dissemination workshops:observations and implications based on an experience in Burkina Faso," *Health Research Policy and Systems,* vol. 15, no. 43, pp. 1-12, 2017.

[30] J. N. Kariuki, J. Kaburi, R. Musuva, D. W. Njomo, D. Night, C. Wandera, J. Wodera and P. N. Mwinzi, "Research Dissemination Strategies Used by Kenya Medical Research Institute Scientists," *East African Health Research Journal,* vol. 3, no. 1, pp. 70-78, 2019.

JHIA
online.org
Journal of Health Informatics in Africa

The Digitalization of Routine Data Management Processes at the Point-Of-Care: The case of e-Tracker in Ghana

Flora Asah[1*], Chipo Kanjo[2], Hillar Addo[3], Darlington Divine Logo[4] Martin Bright Msendema[5]

[1]University of Oslo, P.O. Box 1080 Blindern, Oslo, 0316, Norway
[2]University of Malawi, Department of Computer Science, Malawi
[3]LUCAS College, Accra, Ghana
[4]Health Researcher Officer, Research and Development Division, Ghana Health Services,
P.O. Box MB190, Accra.
[5]University of Malawi, the Polytechnic I, Private Bag 303, Blantyre 3 1 Malawi

Purpose: This study explored the usability factors that influence the use of mobile technologies among healthcare staff at the point-of-care.
Methods: The study adopted an interpretive approach and employed a combination of qualitative and quantitative data collection methods, and inductive and deductive analysis. A questionnaire was adapted, and data was collected from 52 health facilities, with four interviews conducted between April and June 2018.
Results: The study found that the digitalization of data collection registers and forms has streamlined data management processes. The nurses were satisfied with e-Tracker as it has made them more efficient and productive. More importantly, the offline feature enabled them to capture data in areas with no Internet coverage. Using e-Tracker has reduced the cost and the physical effort required to collect and process data and has improved data quality. Lack of confidence in using the tablet in front of clients and poor Internet connectivity were among the challenges identified. While inevitable, these need to be addressed as they could influence the usability and continuity of use of the technology.
Conclusions: While the device has the potential to improve routine data collection, some contextual factors might hinder its usability. An important lesson from this study is that the implementation of new technology among healthcare staff, particularly those at the peripheral level, requires continuous training and support. The study contributes to the discourse on digitalization of routine data management processes at the point-of-care.

Keywords: *Usability, e-Tracker, mHealth Technology, Routine health information management, Digitalization, Ghana*

1 Introduction and Background

Manual data collection has traditionally been the mainstay of collecting and managing routine data in the healthcare sector, especially at the peripheral levels in low and middle-income countries (LMIC). However, it has high potential for human error such as incomplete records, poor recording, and underreporting of data. Manual data collection procedures can also be cumbersome and time-consuming, thus increasing the burden on already overloaded staff, and risking the quality of both the data collected and the services provided [1,2]. Poor quality data and its limited use to support decision-making characterize health management information systems (HMIS) in many LMIC, including Ghana, which was the focus of this study.

Ghana is a developing country with a population of 28,102,471 (July 2018 est.) [3]. In the country's healthcare sector, routine data is collected manually, with nurses completing several registers before attending to patients. Due to the high volume of clients, paper-based registers can be cumbersome and time-

*Corresponding author address: florana@uio.no

J Health Inform Afr. 2020;7(2):18-28. DOI: 10.12856/JHIA-2020-v7-i2-285

consuming and could result in late submission of data. In turn, this results in poor monitoring of, and delays in patient follow-up, and inability to take timely action, which could affect the delivery of primary healthcare services. To address this situation, the Ghana Health Service (GHS) decided to digitalize data management related activities by implementing mobile technology (e-Tracker). E-Tracker is used by front-line healthcare providers such as community-based Health Planning and Services (CHPS) for the management of routine data [3] at the point-of-care.

The introduction and implementation of mobile technology in the healthcare sector in general and for managing data collection related activities hold tremendous potential [4]. Features such as GPS navigation, web-browser, instant messaging, and high-speed wireless network [5] have opened up a vast space for mobile interactions in the healthcare sector. Studies have shown that mHealth can improve health systems in areas such as maternal, child, and reproductive health [6]. Mobile applications including the use of mobile devices can improve medical data collection, service delivery, and patient-doctor communication, and facilitate real-time monitoring of patients. In HMIS, mobile technology can be employed to address data collection challenges [7,8], especially in LMIC.

Using mHealth technology for data management has significantly reduced the effort expended. For example, it can be used in HMIS to track and monitor the reporting of health indicators, while the use of Personal Digital Assistant (PDAs) has increased access to data, improved accuracy (timeliness and feedback), reduced time and costs, and improved data quality [9]. In Mozambican health centers, staff who use handheld phones to send routine data to district offices reported up to 50% improvement in data quality [10].

Despite the advantages of mobile technologies, studies conducted in Nigeria and South Africa have cited poor network infrastructure as a predominant barrier to their use, especially in rural areas where there is poor network coverage [11]. The high cost of Internet access in landlocked countries hinders full implementation of mobile technology. For example, a World Bank report showed that on average, the cost of Internet access is US$206.6 per Mbit/s per month in coastal countries in Africa, compared to US$438.82 per Mbit/s per month in landlocked ones. Chad, Cameroon, Equatorial Guinea, Lesotho, Mali, and Niger have some of the highest access costs [12]. Another challenge is the lack of a regulatory framework; there is also a need to review and update existing ones to address emerging issues and new technologies. The lack of or poor implementation of laws on cybersecurity, data protection, and privacy, could slow the momentum of the growth of the African digital economy [13]. A lack of skilled human resources such as ICT professionals and electronic content developers hinders the implementation of mobile technology, especially in rural areas [14]. Other factors that can affect usability include limited Internet connectivity, high levels of power consumption, small screen sizes, limited input modalities [8], low Internet penetration, and poor mobile connectivity [15,16], which, for example, hamper the use of SMS and voice call reminders to take medication [12] and negatively impact perspectives of usability [8].

Organizations implement mobile technology in order to increase efficiency. However, the full potential of the system and thus the benefits can only be exploited if they are used. Nielsen [17] explains that as the prevalence of mobile technology with Internet connectivity increases, so too, does the need to pay attention to the usability of these devices. Usability refers to how easy it is to use a device, or how easily it can be used to accomplish a given task. A system that is difficult to operate is likely to fail; therefore, acceptance and ease of use are factors in the successful implementation of mHealth in LMIC[12]. For example, a major challenge faced by mHealth users is lack of use acceptance of the technologies [18,19] as they are designed to be used while on the move. Therefore, ascertaining mobile technology's usability is critical for continuity of use [12,20].

Furthermore, routine data is generated at the-point-of-care (peripheral level). Studies have revealed that the use of mobile technologies at this level is influenced by multifaceted contextual factors and unanticipated challenges [18]. Ghana has a large rural healthcare sector that is nurse-driven and studies have revealed that frontline care providers lack basic IT skills [13]. The fact that users in rural healthcare settings are reluctant to accept mHealth technologies [17] motivated the study. The main objective was to understand the contextual factors that influence the usability of these technologies among frontline nurses at the point-of-care in primary healthcare facilities. The findings could serve as an important component of the pilot phase of mHealth implementation.

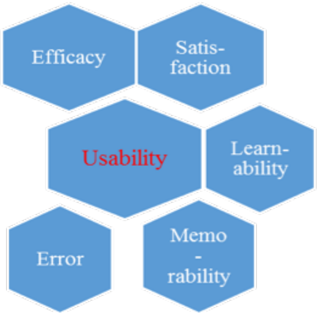


Figure 1 Usability Framework (Lin, 2013)

Definition of Concepts

The concept of usability emanates from the field of human-computer interactions (HCI) and is concerned with the relationship between humans and computers. Usability is a quality attribute that measures the ease-of-use of a user interface. It concerns the quality of a user's experience when interacting with an application, website, or mobile within a specific context [21]. Measuring usability involves a combination of factors. The International Standardization Organization (ISO) 9241-11 defines usability as "the extent to which a device can be used by a specific user to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [21]. In this case, the device is the e-Tracker, the users are community health nurses (CHNs), the tasks involve managing data related activities and the context of use is primary healthcare. Data related activities include data collection, analysis, interpretation, and dissemination. Figure 1 above shows the five constructs that have been used to measure usability [21]. The key constructs which guided this study were: Efficiency – a measure that assesses whether users have learned to use the device and how quickly they can perform tasks that is considered a quantitative measurement; Satisfaction – a qualitative measurement that measures how pleasant it is to use a device; Learnability – a quantitative measurement to assess how easy it is to learn the system and quickly begin to use it for work; Memorability – a measure of how well the user can re-establish proficiency after a period of not using the system; and Error – a measure that assesses how many errors users make when using a mobile device and how easy it is to recover from the error[21]. Our study did not use the error rate as a separate measurement because it can be incorporated into the learnability attributes [22]. The remainder of the article is arranged as follows: Section 2 presents the research setting and methods employed, section 3 discusses the results, section 4 presents the discussion, and section 5 concludes the article.

2 Research Setting and Methods

2.1 Research Setting

The study was conducted in two districts (Central Region Awutu Senya and the Volta Region, Ho Municipality), across 52 health facilities. These districts were selected because at the time of data collection, they had implemented e-Tracker, and nurses had been trained and had started using the mobile device. These health facilities are the first point-of-entry for most people seeking healthcare and data processing is done manually. They are under-resourced in terms of both human resources and technological and other infrastructure. The staff has limited skills, and few opportunities for further training. They also lack access to technology, experience sporadic power cuts, and have limited or no access to the Internet. This is the level where data is generated and the source of many data quality issues. Redman [23] explains that poor quality data at this level affects the entire system, and cleaning data at another level is difficult and expensive.

At CHPS, routine data is gathered from patients who seek different healthcare like outpatient, immunization, and maternal care and recorded on different registers and forms. Next, the data is manually compiled and aggregated weekly, monthly or quarterly, and forwarded to the next level of reporting. At the highest level, the data is analyzed using the District Health Management Information System (DHMIS2),

J Health Inform Afr. 2020;7(1):18-28. DOI: 10.12856/JHIA-2020-v7-i2-285

an open source software platform for reporting, analysis and dissemination of routine data. Each district has a district health directorate, with a district information manager. At the health centers, nurses have poor access to information because the books are out of date, there is no access to journals and the Internet, and the information available is not appropriate for the local situation.

2.2 Methods

This qualitative study adopted an interpretive stance to understand the phenomenon of digitalization of routine data management at the point-of-care through the meaning that people ascribe to it [24]. Behaviors that stem from experiences help to describe realty. The interpretive stance was chosen because the study aimed to descriptively understand the contextual factors that influence frontline healthcare staff's use of mobile technology in delivering healthcare services from their own perspective. Interpretive studies aim to produce "an understanding of the context of IS, and the process whereby the information system influences and is influenced by the context" [24, pp.4-5]. Walsham [24] notes that interpretive studies in information systems (IS) research incorporate thick descriptions of human interactions in the use of IS. Data was collected from 52 health facilities using a combination of data collection techniques including questionnaires, interviews, and document analysis [25].

A questionnaire was the primary source of data collection. This is a well-established tool to acquire information on public knowledge and perceptions. It enables respondents to consider their responses carefully without interference from, for example, an interviewer and it is possible to access a large audience within a short period of time and to compare the data [26]. In addition, questionnaire is particularly useful when participants wish to remain anonymous and more comfortable way for participants to divulge information that would make them uncomfortable in a face-to-face setting. In this case, the purpose is to understand nurses' perception of e-Tracker (mobile health) and to rate how they feel about using the device. In addition, we sought to compare the data from the two districts to establish any differences in experiences of using e-Tracker at Awutu Senya (urban) and Ho municipality (rural).

The team adapted a usability questionnaire to meet the needs of the study. The questions were closed-ended and were scored on a 5 point Likert-type scale ranging from strongly agree to strongly disagree. Space was provided for nurses to add their comments after each question and to express their views on using e-Tracker. For validating and administering of the questionnaire, it was cross-checked by IT students at Lucas College in Accra, under the supervision of the fourth author and senior health information department staff to validate the content. It was pre-tested with five surveillance officers at Ho municipality. Pre-testing a questionnaire helps to determine if the respondents understand the questions and provides the most direct evidence of the validity of the questionnaire data [27]. Feedback from the pre-test led to the modification and clarification of some questions. The questionnaire was paper-based and one questionnaire was administered per facility.

2.3 Data collection

This study focused on the e-Tracker on the tablet, with the nurse in charge of the tablet at each health facility responding to the questionnaire. Data was collected between April and June 2018. The questionnaires were handed to the CHNs in charge at the district office while they were attending the monthly data management review meeting. At the meeting, the district manager introduced the researchers to the CHNs in charge and the researchers were given ten minutes to present the study's aims and purpose. At the end of the meeting, an envelope containing a questionnaire, a covering letter that explained the study's objectives, and a consent form for the CHN in charge to sign was handed to the CHNs in charge. Envelopes for CHNs in charge who were absent from the meeting were later given to the Health Information Officer (HIO) at Awutu Senya and the Public Health Officer at Ho to deliver at the health facility, as they visit facilities in the district on a weekly basis. Sixty-two questionnaires were distributed, 40 to Ho Municipality and 22 in Awutu-Senya. The facilities were given approximately a month and-a-half to complete the questionnaire and two reminders were sent in the space of two weeks to increase the response rate.

While a questionnaire is an effective data collection tool, it is not always comprehensive. Hence, the questionnaire was supplemented with interviews and document analysis. Interviews were conducted after the health facilities had returned the questionnaires. They probed why the nurses held the opinions recorded on the questionnaires. Interviews were conducted with four individuals charged with overseeing the implementation and use of e-Tracker in these districts, namely, the district health director, two information managers, and a public health officer. The interviews were semi-structured with broad and open-ended questions to allow the respondents to explain the issues they encountered in trying to get the nurses to use e-Tracker. Before the start of each interview, time was dedicated to building trust that helped to set the tone for the rest of the discussion and for the respondents to read the information sheet and sign the consent form. The interviews were conducted at the respondents' offices and lasted 30 minutes. They were auto-taped, and transcribed verbatim. Printed and electronic documents were also analyzed. These included strategic plans, project reports, quarterly reviews, and bulletins. They provided contextual information on health care practices and strategies in Ghana. Collecting data from multiple sources increased the study's internal validity [28]. Written informed consent was obtained from those interviewed after they had read the covering letter explaining the study's objectives. Ethical approval was obtained from the Ghana Health Services Review Committee, Reference 017/11/17 and permission was obtained from the relevant authorities.

2.4 Quantitative and Qualitative Data Analysis

Data analysis for this study was twofold. Firstly, the questionnaires were sorted and entered on the Statistical Package for the Social Sciences (SPSS), with descriptive analyses employed to summarize the deductive results. The data was analyzed based on usability concepts. Secondly, the interviews (qualitative analysis) were transcribed and analyzed. Analysis of the data from the interviews was cyclic [27], with the researcher going back and forth from the data to the analysis and from the analysis back to the data to gain a good understanding of what the participants were saying. An inductive approach to thematic analysis was adopted based on Braun and Clarke's [29] process. The transcripts were read and codes that gathered similar data together were developed. Phrases were written in the transcripts that placed similar data together. After identifying the phrases, the researchers reviewed the transcripts and started developing interpretive codes. The coding process was done manually using colored pens to highlight key themes while inserting comments in the margins to record the researchers' thoughts. Thereafter, the themes were charted to link key phrases from the respondents and identify patterns of phrases. As an inductive process of coding and categorizing was adopted, the themes that emerged are rooted in the participants' own words.

3 Results and Analysis

This section presents the results from the qualitative and quantitative analysis. Where appropriate, comments and verbatim quotes are inserted.

3.1 Characteristics of Study Participants

Of the 62 questionnaires administered, Awutu-Senya returned all 22 while Ho municipality returned 30. In total, 52 health centers returned the questionnaires. Of the 52 nurses who participated in the study, 43 (83%) were female and 9 (17%) were male. This affirms that nursing is still regarded as "the quintessential" female profession [30]. Most (96%) of the nurses were CHNs with more than three years' work experience. The majority (63%) of the participants were in the age bracket 31 to 40 years, while 23% were aged 15 to 30. Approximately 75% of the participants had worked at the facility for more than two years. There was no significant difference between nurses' use of e-Tracker in the rural and urban districts.

3.2 e-Tracker Usage

Each health facility has a desktop computer and one tablet and both have e-Tracker installed. The tablet is used by the CHN in charge. E-Tracker is used to collect routine data and to track clients on different health programs such as family planning, immunization, child health, maternal health, and post-natal health services, and to send SMSs to remind mothers of their appointments. It is also used to manage data; i.e., to capture, calculate coverage, analyze trends, compile data, and submit data to the DHMIS2 platform, where facility reports are generated.

Comparing the data collection process before and after the implementation of e-Tracker

The nurses reported that before the implementation of e-Tracker, a nurse, for example, was required to manually complete an average of five registers or forms before attending to a patient. S/he had to juggle between examining the mother and the baby, and completing the forms. A nurse explained the challenges they encountered on a daily basis:

> *"Mothers come very early to the clinic to have their babies vaccinated before starting their day's activities. At the clinic, we have to examine both mother and child, weigh and check the baby's vital signs. If the mother is still breastfeeding the child, we have to verify ...* [that] *the baby is eating properly. All information is written down. Also, one has to fill in about five registers. Mothers are getting angry, complaining that we are wasting their time." (HO_2)*

Another nurse added that at month end, when data has to be submitted, they had to flip through tens of pages of the registers and the tally sheet to verify and validate the data, then capture it on an excel file, and perform the necessary validation, before the data was sent to the next level. This participant described manual data collection as cumbersome and time consuming. Using e-Tracker and the digitalization of registers, including the use of the Unique Patient Identifier (UPI) has reduced duplication of processes, enabling nurses to perform their tasks more quickly. A feature of e-Tracker that facilitates ease of use is the device's interface. Around 75% of the participants agreed that the e-Tracker interface is pleasant to work with. They explained that the organization of information on the device is very clear and easy to grasp. The HIO added that data management was easier and faster as nurses submitted data almost daily, making it easy to manage and validate and provide timeous feedback. This resulted in improved efficiency. Although most of the nurses said that they were happy with using the device, around 30% said that they did not feel comfortable using it in front of patients.

3.3 Satisfaction

More than half (59%) of the nurses expressed satisfaction with using the device and 64% agreed that it made them more productive. The participants noted that digitalization increased the speed with which data was processed, because e-Tracker enables reports to be generated instantly. This function was previously done at the district office, and took a couple of days after the data had been submitted.

One participant compared using e-Tracker on a desk-top computer to using it on the mobile device. It was noted that the mobile device offered mobility and an offline feature enables nurses to capture data without necessarily being connected to the Internet. This feature is not possible with a laptop. The Information Manager concurred:

> "…using e-Tracker on the desktop was difficult as compared to the one on the tablet. This version on the tablet is user-friendly. But an additional advantage is the offline feature which allows nurses to capture, save data and then send it later. One does not need to have access to the Internet during data capturing" (AS_4)

Another nurse added that e-Tracker enables instant generation of clients' schedules and they are able to send text messages from the device as reminders to clients anywhere at any time. The participants expressed satisfaction with the device as they can do everything on it. Similarly, the HIO noted that the introduction of e-Tracker had reduced the cost and the physical effort required to collect and process data. For example, the amount spent on paper, printing reports and phone calls has decreased significantly. The HIO added that data timeliness and submission rates have improved and because data is submitted daily, so too is the quality of data because nurses have more time to look at the data before it is submitted.

However, it was interesting to note that, 40% of the nurses expressed reluctance to use e-Tracker because it was time-consuming. A further reason was that they had to do double data capturing, i.e., on paper-based registers and the device. However, the HIO explained that the MoH requested this as the digitalization process is in its initial stages and that, in order to ensure that no data is missed, CHNs should continue capturing data manually. This will ensure a smooth transition. Once the system is stable, the CHNs will only use the e-Tracker.

3.4 Learnability and Error

Before nurses started using e-Tracker, they attended a two-day training course on how to use the tablet. Fifty-two percent of the nurses reported that it was easy to learn how to use the e-Tracker and 50% added

DOI: 10.12856/JHIA-2020-v7-i2-285

that it was easy to rectify an error. The HIO added that the MoH is aware that the training was brief; however, health centers are visited on a regular basis by HIOs who provide on-site support and supervision to ensure that the device is used properly and functioning well. Furthermore, the fact that the nurses found the interface easy to use and were able to navigate the various features easily implies that the e-Tracker was memorable.

3.5 Factors that Influence Nurses' Use of e-Tracker

Inadequate Skills and Lack of Support

The nurses were of the view that they required more training and support on how to use the device and to ensure that it is functioning. For example, three health facilities in Ho municipality did not complete the questionnaire because the device was not functional. From the deductive results presented above, it would seem that e-Tracker offers exciting new opportunities to nurses to satisfy high demand for healthcare delivery, and to be able to work faster, and be more flexible and effective. However, a lack of adequate skills and support to use the device effectively can cause frustration, with negative impacts on their motivation. Training is an important component when implementing new IS and at the rural health facilities, continuous training and support is even more crucial due to the dearth of skills [1,2].

Lack of Resources

Lack of mobile device - as noted previously, there is only one device per facility. The nurses expressed the need for more devices, as having one tablet per facility with many consulting points does not reduce the workload. There is also a lack of finance for fuel and to purchase data for the device. The HIOs and nurses need fuel for the vehicles and motorcycles they use to travel to health facilities for supervision and to provide on-site support and the device need mobile data. A lack of resources such as finance and equipment, are among the biggest challenges hindering the implementation of a new IS in low and middle-income countries (LMIC) [1].

Lack of Confidence to Use e-Tracker Device

The nurses reported that they lacked confidence to use the device on a daily basis. One explained:

> *"When we stand in front of the clients talking to them with this device in our hand, some of them feel that instead of attending to them instead we are chatting so I do not feel confident"(HO_1).*

Another commented:

> *"...though nurses are happy with the device, most of them are still reluctant to use [it]. ... They prefer to capture data on the paper registers and when the facility is less busy, they transfer the data on the device..." (AS_5).*

Limited Internet Coverage

Furthermore, the nurses reported poor Internet coverage at some health facilities, especially those in deep rural areas. They noted that after capturing routine data they have to drive long distances to obtain Internet coverage to send it, meaning that they spend more on fuel.

4 Discussion

The study aimed to understand the contextual factors that influence the usability of e-Tracker (mHealth technologies) among frontline nurses at the point-of-care. Quantitative and qualitative data collection methods were employed. To measure usability, we used Lin's [21] concepts, namely, efficiency, satisfaction, learnability, error, and memorability. Our quantitative analysis to measure the concepts of usability revealed that four concepts (efficiency, satisfaction, learnability, and error) were evident in this study. Due to lack of data, we were unable to measure the concept memorability.

The study found that the digitalization of numerous paper-based registers and forms has reduced the time spent on data management processes. The user-friendliness of the e-Tracker interface facilitates easy access to different components and data management has become flexible and faster. Using e-Tracker gave nurses the freedom to perform their duties (to collect routine data) easier and faster, hence giving them more time to validate data before it is submitted, consequently improving data quality (data timeliness and submission rates have improved). The mobility of the device and availability of the offline feature facilitate data capture and analysis even when there is no connectivity, which improves performance. While using e-Tracker has made nurses more productive and efficient, from a management perspective, it has reduced the cost and the physical effort required to collect and process data.

Apart from establishing whether e-Tracker is usable or not, the study also aimed to identify the challenges encountered by nurses in using it. This is important as challenges can affect usability. The findings show that nurses lacked confidence to use the device daily to complete their tasks. They also point to the challenge of undermining clients' trust, as some clients assume that nurses are using the e-Tracker device for fun. This finding is consistent with Velez et al.'s [31] study that found that because nurses were not confident about using the device, they reverted to paper-based registers and forms to collect routine data.

Self-confidence is related to uncertainty [32]. Bearden [32] notes that when individuals are confronted with an intricate situation, self-confidence plays a significant role in backing their actions or decisions, and can determine their attitude. There are two types of self-confidence; general self-confidence and specific self-confidence. Specific self-confidence implies that the individual has ample information and knowledge that makes them confident about handling the specific device. General self-confidence involves negative and positive attitudes towards a particular object or individual. Bearden [32] contends that general self-confidence is associated with a person's decisions and behavior, and is often associated with non-users with little or no experience of a particular product. In this study, the nurses lack general self-confidence. Individuals with high levels of general self-confidence are accustomed to using new technologies and willing to take risks [33]. In contrast, those with low levels of general self-confidence feel that they are insignificant and their fallibilities make them uncomfortable, and uncertain that they can successfully manipulate a new technology.

A lack of general self-confidence to use a technology in an organization could have a serious impact on the workforce [34]. As noted above, a lack of self-confidence is the result of a lack of skills. In the IS domain, an important objective of the implementation of a new technology is to ensure that its users are equipped with the skills required to properly use it [35]. However, building staff confidence to use technology, particularly among staff at the peripheral level [36], requires continuous and contextual training because it empowers staff by giving them the confidence they need to keep abreast of the new technology, and pushes them to perform better.

Although the managers acknowledged that the two-day initial training given to nurses was not sufficient and noted the need for on-going on-site support and supervision, performing these services might be challenged due to the lack of financial resources to cover items such as fuel to travel to health facilities to provide such. Lack of finance and poor Internet connectivity hamper the development and sustainability of IS in LMIC. While IS implementation is complex and challenges are inevitable [37], it is important to manage them because they are key to continuity of use of the technology[38].

5 Conclusion

The study found that the implementation of mobile technology has enhanced and improved the processes of data collection and use at the point-of-care. The digitalization of data collection registers and forms significantly reduced the data management processes. These findings are consistent with those of

[9,10,13,19]. However, the implementation of mobile technology was not without challenges. While these are unavoidable, they need to be attended to because if they persist, they might affect usability; that is, nurses might abandon the device. An important lesson from this study is that the implementation of new technology among healthcare staff, particularly those at the peripheral level requires continuous training and support. The traditional two-day HIS training is inadequate because it does not provide staff with the necessary skills to adequately use the technology. In order to develop staff skills and self-confidence to use the new technology properly, the two-day training should be complemented with on-site support and supervision. Overall, the positive effects of e-Tracker outweigh the challenges. However, to gain the full benefits, it is important to focus more resources on building staff capacity.

This study contributes to the discourse on digitalization of routine data management processes at the point-of-care. While it was conducted in two health districts, its findings raise issues that have wider applicability to the implementation of information technology in resource-constrained settings. The primary concerns relating to qualitative research revolve around validity and reliability [27]. To address these concerns, in terms of validity, we employed triangulation; that is, a variety of data sources (questionnaires, interviews, and documents analysis) as opposed to relying solely on one source. We also included verbatim quotations [39] in the analysis section. To ensure reliability, the interviews were recorded and transcribed, and at the end of data collection, a preliminary report was written and presented at the district information meeting [40]. In terms of future research, while there is a rich body of literature on data management in health facilities, very little attention has been paid to nurses at the peripheral level who generate this data. We recommend that further study is conducted to check the validity and quality of the data since data quality problems arise at the point-of-care.

Acknowledgements

We wish to thank all the nurses who participated in this project. We would like to express my sincere gratitude to Madame Patricia Antwi, Messrs Christian Gohoho, Adams Agbeko, Frederick Ofosu-Ababio, Fredie Amponsah, Dominic Atweam, Adu Geegee Kwame, and Vadis Adesi for orienting us towards relevant information on the project. Special gratitude to my department, IFI at the University of Oslo for the financial support.

Statement on Conflicts of interest No conflict of interest

Funding: No funding

References

[1] Lippeveld T, Sauerborn R, Bodart C, World Health Organization. Design and implementation of health information systems.

[2] Lippeveld, T. Achieving Universal Health Coverage: The role of Routine Health Information System. RHINO Forum, January 15, 2019. Available at https://www.rhinonet.org/summary-of-achieving-universal-health-coverage-the-role-of-routine-health-information-systems/

[3] Association of Chartered Certified Accountants. (2013). Key Health Challenges in Ghana. Available athttps://www3.accaglobal.com/content/dam/acca/global/PDF-technical/health-sector/tech-tp- khcg.pdf

[4] ASANGANSI I, MACLEOD B, MEREMIKWU M, ARIKPO I, ROBERGE D, HARTSOCK B, MBOTO I. Improving the routine HMIS in Nigeria through mobile technology for community data collection. Journal of Health Informatics in Developing Countries. 2013 Jun 1;7(1).

[5] Taylor, Katie Headrick; Takeuchi, Lori; Stevens, Reed (2017). Mapping the daily media round: novel methods for understanding families' mobile technology use. Learning, Media and Technology. 43: 1–15.

[6] Wallis L, Blessing P, Dalwai M, Shin SD. Integrating mHealth at point of care in low- and middle-income settings: the system perspective. Global health action. 2017 Jun 30; 10 (sup3):1327686.

[7] Källander K, Tibenderana JK, Akpogheneta OJ, Strachan DL, Hill Z, ten Asbroek AH, Conteh L, Kirkwood BR, Meek SR. Mobile health (mHealth) approaches and lessons for increased performance and retention of community health workers in low-and middle-income countries: a review. Journal of medical Internet research. 2013; 15 (1): e17

[8] Van Biljon J, Kotzé P. Cultural factors in a mobile phone adoption and usage model.

[9] Xiong K. Mobile technology for monitoring and evaluation and health information systems in low-to middle-income countries. Meas Eval Spec Rep. 2015:11-6.

[10] Macanze J. Monitoring and Evaluation Report of PDAs for Malaria Monitoring in Maputo Province, Mozambique - Final Report. 2007

[11] Pankomera R, Van Greunen D. Challenges, benefits, and adoption dynamics of mobile banking at the base of the pyramid (BOP) in Africa: A systematic review. African Journal of Information and Communication. 2018;21: 21-49.

[12] Little A, Medhanyie A, Yebyo H, Spigt M, Dinant GJ, Blanco R. Correction: Meeting Community Health Worker Needs for Maternal Health Care Service Delivery Using Appropriate Mobile Technologies in Ethiopia. PloS one. 2014;9(1).

[13] Kiberu VM, Mars M, Scott RE. Barriers and opportunities for implementation of sustainable e-Healthprogrammes in Uganda: A literature review. African journal of primary health care & family medicine. 2017;9(1):1-0.

[14] Ewusi-Mensah K. Problems of information technology diffusion in sub-Saharan Africa: the case of Ghana. Information Technology for Development. 2012 Jul 1;18(3):247-69.

[15] Mayes J, White A. How smartphone technology is changing healthcare in developing countries. InJ. Glob. Health 2016 Nov 1.

[16] Park YT. Emerging new era of Mobile health technologies. Healthcare informatics research. 2016 Oct 1;22(4):253-4.

[17] Neilsen Mobile (2008) Critical Mass: The Worldwide State of the Mobile Web. [Electronic version]. Retrieved November 2008 from http://www.nielsen.com/us/en/reports/2008/critical-mass-worldwide-state-of-the-mobile-web.html

[18] Banda CK, Gombachika H. mobile phone technology acceptance and usability in the delivery of health services among health surveillance assistants in rural areas of Malawi. In Internationa Conference on e-Infrastructure and e-Services for Developing Countries 2012 Nov 12 (pp. 249-258). Springer, Berlin, Heidelberg.

[19] O'Connor Y, O'Donoghue J. Contextual barriers to mobile health technology in African Countries: A perspective piece. Journal of Mobile Technology in Medicine, JMTM. 2015;4(1):31-4.

[20] Kaplan WA. Can the ubiquitous power of mobile phones be used to improve health outcomes in developing countries? Globalization and health. 2006 Dec 1:2(1):9.

[21] Lin CC. Exploring the relationship between technology acceptance model and usability test. Information Technology and Management. 2013 Sep 1;14(3):243-55.

[22] Adeoye B, Wentling RM. The relationship between national culture and the usability of an e-learning system. International Journal on E-learning. 2007 Jan; 6(1):119-46.

[23] Thomas R Redman. Data. To Improve Data Quality, Start at the Source. February 2020. Available at https://hbr.org/2020/02/to-improve-data-quality-start-at-the-source

[24] Walsham G. (2006). Doing Interpretive Research, European Journal of Information Systems 15(3):320-330.

[25] Lincoln, Y. S., & Guba, E. G. (2002). Judging the quality of case study reports. The qualitative researcher's companion, 6(4), 205-215.

[26] Bird, D. K. (2009). The use of questionnaires for acquiring information on public perception of natural hazards and risk mitigation–a review of current knowledge and practice. Natural Hazards and Earth System Sciences, 9(4), 1307-1325.

[27] Marshall, C., & Rossman, G. B. (2014). Designing qualitative research. Sage publications.

[28] Creswell, J. W. (2003). Qualitative Inquiry and Research Design: Choosing among five traditions. Thousand Oaks CA, Sage Publications, Inc.

[29] Clarke V., & Braun, V. (2014). Thematic analysis. In Encyclopedia of critical psychology (pp. 1947-1952). Springer, New York, NY.

[30] Evans JA. Cautious caregivers: gender stereotypes and the sexualization of men nurses' touch. Journal of advanced nursing. 2002 Nov; 40(4):441-8.

[31] Vélez O, Okyere PB, Kanter AS, Bakken S. A usability study of a mobile health application for rural Ghanaian midwives. Journal of midwifery & women's health. 2014 Mar; 59(2):184-9[1.

[32] Bearden, W.O., Hardesty, D.M., Rose, R.L., 2001. Consumer self-confidence: refinements in conceptualization and measurement. J. Consum. Res. 28 (1), 121–134.

[33] Agarwal, R., Sambamurthy, V., Stair, R.M., 2000. Research report: the evolving relationship between general and specific computer self-efficacy—an empirical assessment. Inf. Syst. Res. 11 (4), 418–430.

[34] Thomas, T. Low self-confidence: Technology and Self-confidence 31 August 2011. Available at https://www.counselling-directory.org.uk/memberarticles/technology-and-self-confidence

[35] Braa, J., Monteiro, E., & Sahay, S. (2004). Networks of action: sustainable health information systems across developing countries. MIS Quarterly, 337-362.

[36] Nutley, T., & Reynolds, H. W. Improving the use of health data for health system strengthening. Glob Health Action. 2013; 6: 2001.

[37] Idler, S., (2011). The Paradox of Technology and 5 Ways to Avoid it, Available at https://usabilla.com/blog/the-paradox-of-technology-and-5-ways-to-avoid-it/

[38] Ketata I, Sofka W, Grimpe C. The role of internal capabilities and firms' environment for sustainable

innovation: evidence for Germany. R&D Management. 2015 Jan; 45(1):60-75.

[39] Johnson, B.: Examining the validity structure of qualitative research. Education 118(2), 282 (1997). Winter 1997, Research Library

[40] Roberts, P., Priest, H., Traynor, M.: Reliability and validity in research. Nurse. Stand. (through 2013) 20(44), 41–45 (2006). Proquest

JHIA - online.org
Journal of Health Informatics in Africa

Reliability of Predictions Using Hybrid Models: The Case of Malaria Incidence Rates in Uganda

Francis Fuller Bbosa[a],[c],[*], Ronald Wesonga[b] , Peter Nabende[c], Josephine Nabukenya[c]

[a]School of Statistics and Planning, Makerere University, Kampala, Uganda
[b]Department of Statistics, College of Science, Sultan Qaboos University, Muscat, Oman
[c]School of Computing and Informatics Technology, Makerere University, Kampala, Uganda

Background and purpose: Reliability of estimates emanating from predictive independent data mining techniques is a complex problem. This could be attributed to cross-cutting weaknesses of individual techniques such as collinearity due to high dimensionality of attributes in a dataset, biasedness due to under fitting and over fitting of data as well as noise accumulation due to outliers and thus affecting the reliability of predictions emanating from these models. This study thus aimed at developing a hybrid data mining technique for predicting reliable malaria incidence rate thresholds.
Methods: The decision tree and naïve Bayes classifiers were used to build a hybrid prediction model. Results of the developed hybrid model were compared with independent data mining models using 10-fold cross-validation on a previously unlearned data set. Accuracy, F-measure and the area under the receiver operating characteristics curve (AUC) were the key performance metrics used to evaluate the generalizability of the hybrid model in comparison to the independent models.
Results: Findings revealed that the hybrid classifier attained an accuracy of 79.3% and an F-measure score of 84.2%, the naïve Bayes classifier achieved accuracy and F-measure value of 69% while the decision tree classifier registered an accuracy of 72.4% and an F-measure score of 80%.
Conclusions: The developed hybrid model outperformed both independent decision tree and naïve Bayes models. Hence merging several independent homogeneous predictive data mining techniques enhances the accuracy of the estimates leading to reliable estimates.

Keywords: Hybrid, Data mining, Prediction, Hybrid, Malaria, Incidence

1 Introduction

Garg and Vishwakarma [1] argue that predicting a reliable estimate based on independent data mining techniques is an intricate obstacle, as each technique has its weaknesses with respect to the data structure, shape, and validity [2] [3]. According to Gidron [4], reliability refers to the degree of consistency in measurement. Easterby-Smith, Thorpe, and Jackson [5] corroborate that reliability can be explored by answering the following three questions: i) Do the predicting methods employed generate similar results on different occasions? ii) Are the same results generated by other researchers? iii) Is the analytical process from raw data to the discovery of new knowledge transparent? Hence reliability is a measure of the consistency of the information [6].

Prediction of estimates in databases is often made based on traditional statistical techniques rather than data mining techniques [7] [8] [9] [10] [11] to mine formerly hidden patterns and information from databases. However, the current proliferation of data that is "big" in nature and unstructured, characterized by its Volume, Velocity, Variety, Veracity, and Value have made it difficult for traditional statistical procedures that are often exclusively accustomed to the investigation of structured and homogeneous data, to process and analyze large and complex data sets [12] [13] [14]. The fact that the data is too big and in different forms as well as from various sources led to several scholars [14] [15] [16] breaking down big data into five characteristics, commonly referred to as 5 V's: Volume relates to the size of data, Variety pertains to the data which appears in different forms, Velocity denotes the high pace at which new data is

[*]Corresponding author address: Department of Planning and Applied Statistics, School of Statistics and Planning, Makerere University, P.o Box 7062, Kampala, Uganda. Email: fullerbbosa@gmail.com. Tel: (+256)-702792202

J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

generated, Veracity measures the authenticity of the data, and Value assesses how good the quality of the data is in reference to the intended results. Therefore, the rise of big data has forced scholars [13] [17] [18] to advance data mining as a plausible solution to extract previously unknown and unseen patterns and information that are challenging to discover with traditional statistical techniques with respect to big data.

Agyapong, Hayfron-Acquah, and Asante [19] assert that data mining techniques are mainly categorized into two categories: predictive and descriptive methods. Predictive approaches also known as classification learn from the training set, where all attributes are already associated with known class labels and build a model which is used to estimate unknown values of new attributes [20] [21] whereas descriptive approaches are also known as clustering usually identify patterns or associations among attributes in datasets by looking for human-interpretable patterns that describe data [19].

Pertinent literature reveals that predictive approaches are the dominant method in the data mining arena [20] [22] [23] possibly due to their strength such as making the computation process easy to understand, generation of inclusive rules for classification, handling both real and discrete data [21] [22] [24]. However, majority of the independent predictive techniques share common weaknesses such as dependence on the nature of the dataset or data type for classifier performance [22], imprecision of estimates in scenarios where various attributes in a dataset are dependent on each other [21] , replication of sub-trees on different paths leading to collinearity [25], information overload due to the large size of input datasets thus increasing the time to mine information, which decelerates the decision-making process [25], collinearity due to high dimensionality of attributes in a dataset [26], biasedness due to under fitting and over fitting of data as well as noise due to outliers [27]. Thus several researchers [26] [27] [28] suggest that different independent data mining models have varying predicting capabilities based on their strengths and weakness; the authors claim there is no universally employable independent data mining model for all prediction scenarios. Hence the practice of employing independent data mining techniques leads to unwanted biases, errors and omissions, noise accumulation and spurious correlations among variables which affects the accuracy and reliability of predictions emanating from these models [29] [30] [31].

Various scholars [32] [33] [34] have applied more than one independent data mining technique to predict estimates on the same dataset but all these techniques generated varying results with dissimilar accuracies. As a result, the above scholars conclude that there is no single data mining model that produces the most reliable result. To address the above gap associated with variances in estimates of predictions using individual classifiers, hybridization of several individual data mining techniques is suggested [25] [35] [36] [37] alluding to the fact that merging several independent data mining techniques improves the accuracy of the estimates leading to reliable estimates [25] [28] [34] [38]. According to Ahlawat and Suri [25], hybrid procedures in data mining are a logical amalgamation of various individual techniques, thereby utilizing the strengths of the individual procedures of the hybrid algorithm to improve the performance of prediction models to generate reliable estimates. Kazienko, Lughofer, and Trawiński [39] suggest that it's imperative to note that both hybrid and ensemble techniques utilize the concept of information amalgamation nonetheless in diverse ways. In case of hybrid classifiers, diverse heterogeneous data mining approaches are combined [39] [40] [41] whereas ensemble classifiers instead merge numerous but homogeneous, feeble techniques [42], characteristically at their individual output level, utilizing several merging methods [43].

1.1 Data mining models employed in the study

Despite the presence of several predictive data mining techniques, scholars are facing the challenge of choosing the best model for a particular data set [44]. In keeping with relevant published literature, the most frequently employed predictive data mining techniques include: Decision trees, Artificial Neural Networks (ANN), KNearest Neighbor (k-NN), Support Vector Machines (SVM), algorithm, logic-based algorithms especially Decision Trees (DT) and bayesian related classifiers [45]. Furthermore, Hamblin et al. [45] reveal that ANN and SVM generate better estimates when dealing with continuous-valued attributes whereas K-NN is biased to noise and hence very sensitive to outliers in datasets. However, given that the researchers' problem under investigation involved discrete data from heterogeneous sources, the above limitations disqualify ANN, SVM, and KNN techniques.

However, logic-based systems such as Bayes and decision trees classifiers tend to perform better when dealing with categorical attributes [45]. As a result, the researchers employed the decision tree and naive Bayes as the predictive data mining techniques for this study on the basis of their greater ability of modelling classification type prediction problems [46]. Above all, the choice of decision trees and Bayesian classifiers takes into consideration decision boundary-based and probability-based approaches to prediction in machine learning respectively [47].

1.2 Malaria disease model

Malaria was chosen as a disease model for this study because the World Health Organization (WHO) [48] recognizes the presence of weak surveillance systems that are unable to reliably predict future malaria incidence rates particularly in Low and Middle-Income Countries (LMICs) particularly Uganda, making it hard to optimize response to malaria outbreaks. Additionally, the latest WHO world malaria report [49] reveals that at a global level, Africa accounted for 213 million cases out of the 228 million cases recorded globally in 2018 with Uganda accounting for 5% of the global burden. Despite increased consideration paid to malaria surveillance systems and their key role for improving health systems in Low and Middle-Income Countries (LMICs) such as Uganda, it is believed that the majority of the existing surveillance systems cannot be used to reliably predict future malaria incidence rates [49]. Hence the need to develop a robust hybrid prediction model in order to enhance early warning leading to effective and timely response to future outbreaks.

The overall motivation underlying this study is that considerable work has been done on boosting the predictive accuracy of individual homogeneous data mining techniques but little work with regards to enhancing the reliability of heterogeneous techniques. To this end, the researchers argue that there is a need for building a hybrid data mining approach, which is an effective amalgamation of numerous independent data mining techniques, to utilize the strengths of each individual technique and compensate for each other's weaknesses.

1.3 Problem statement

In order to address the drawbacks of the traditional statistical methods, data mining techniques have been adopted. However, the conventional independent data mining techniques are not capable of producing reliable predictions. This is mainly attributed to their weaknesses with respect to the data structure, shape, and validity. As a result, the performance of conventional independent data mining techniques is weakened due to noise accumulation emanating from measurement errors, outliers, and missing values as well as spurious correlations which may lead to false scientific conclusions or poor predictions and varying measures of accuracy. Hence, the need to develop a hybrid data mining algorithm in order to improve the predictive accuracy and reliability of individual data mining techniques.

2 Related Literature

In this section, the researchers review recent research on the amalgamation of independent data mining for various real-world predictive problems.

Sumana and Santhanam [44] examined single and hybrid classification techniques as viable tools to achieve enhanced predictions for the presence or absence of heart diseases in a cohort of patients. Their findings reveal that the proposed hybrid model produced more accurate estimates of 99.54% compared to single classifiers and ensemble classifiers.

Ogwoka, Cheruiyot, and Okeyo [50] proposed "A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms". Findings reveal that merging Decision tree and k-means algorithms generated better results after extracting previously unknown features, thus improving the accuracy of prediction.

In 2016, Dubey and Saxena [51] developed a hybrid prediction model for feature selection. They amalgamated correlation and support vector machines to classify big data. This proposed hybrid technique was tested on five big data boolean datasets. The authors attest that the hybrid yielded better accuracies in three out of the five big data datasets with a fewer number of features.

Ahlawat and Suri [25] developed a hybrid algorithm by combining decision trees and clustering to classify data samples. Their results proved that an amalgamation of decision trees and clustering is suitable to improve the accuracy of estimates. They concluded that using hybridization can be used to enhance performance and prediction values to get better results.

In 2016, Raghavendra and Indiramma [52] proposed a "Hybrid data mining model for the classification and prediction of medical datasets". The researchers employed attribute separation selection techniques particularly the forward selection and backward elimination method generate an appropriate subset of attributes to enhance the performance of the model. Findings revealed that the proposed hybrid model outperformed the linear regression and artificial neural networks with fewer number of significant attributes.

Hakizimana et al. [2] proposed "A Hybrid Based Classification and Regression Model for Predicting Diseases Outbreak in Datasets". The authors built a hybrid model for predicting infections occurrence in datasets by merging naïve Bayes, random forest, simple logistic, Bayesian logistic regression, and SMO. The hybrid technique produced the best accuracy with 100%, compared to the naïve Bayes with 90.9%, SMO with 90.9%, and Bayesian logistic regression with 36.4%. Hence the hybrid model is superior to individual models in terms of improved accuracy of estimates.

Ren, Fei, Liang, Ji, and Cheng [53] in their study to predict kidney disease in hypertension patients proposed a hybrid neural network that integrates Bidirectional Long Short-Term Memory (BiLSTM) and Autoencoder networks. Findings from their study revealed that the proposed hybrid model attains 89.7% accuracy and thus the proposed integrated model outperformed traditional stand-alone prediction models with distinct features and neural baseline systems.

In order to predict diseases, [54] developed a hybrid model that amalgamated k-nearest neighbor,case-based reasoning, and fuzzy set classifiers. Findings revealed that the hybrid model enhanced the accuracy of the model compared to the stand-alone classifiers. The authors concluded that the integration of several independent predictive models aids to yield improved estimates from predictions. In 2020, Ju-young, Rang, and Jong-chul [55] proposed "Seasonal forecasting of daily mean air temperatures using a coupled global climate model and machine learning algorithm for field-scale agricultural management". The researchers used a hybrid model that integrated multiple regression and artificial neural networks to forecast mean daily temperature. The hybrid model yielded results with a root mean square error (RMSE) of 1.02-3.35 compared to the standard climate model that achieved a RMSE of 1.61-3.37.

Junliang Fan, Wu, Ma, Zhou, and Zhang [56] suggested three novel hybrid support vector machines (SVM) with bat algorithm (SVM-BAT), whale optimization algorithm (SVM-WOA) and particle swarm optimization algorithm (SVM-PSO) for daily diffuse solar radiation in air-polluted regions using, multivariate adaptive regression spline (MARS), SVM and extreme gradient boosting (XGBoost) models. Their findings showed that hybrid models generate more accurate estimates. The researchers proved that the hybrid SVM-BAT model was a better classifier than the SVM, XGBoost, and MARS models by attaining more accurate daily rates and quicker convergence rates.

A synopsis of the reviewed relevant literature suggests that the application of various independent predictive data mining techniques in the context of classification on similar datasets yields varying estimates, leading to poor, unreliable estimates and consequently insufficient scientific conclusions. The above shortcomings have stimulated the curiosity of the researchers to continuously struggle to improve the algorithms for undertaking classifications and predictions using single data mining techniques related to the realization of reliable estimates. To this end, the researchers argue that there is a need for building a hybrid data mining approach, which is an effective amalgamation of numerous independent data mining techniques, in order to utilize the strengths of each individual technique and compensate for each other's weaknesses.

3 Materials and methods

3.1 Data Pre-processing

3.1.1 Data Collection .

Monthly data for the period January 2012 to December 2019 on confirmed and suspected (clinically diagnosed) cases of malaria were obtained by the researchers from the Ministry of Health through the

District Health Information System 2 (DHIS2)[1]. Temperature (average maximum and average minimum) and rainfall data for a similar period were also obtained from the Uganda National Meteorological Authority (UNMA)[2], whereas demographic data was obtained from the Uganda Bureau of Statistics (UBoS)[3].

3.1.2 Data Cleaning

The researchers verified and validated the raw datasets in order to check for errors, omissions, and outliers in preparation for compiling a complete and merged dataset that was used for building predictive models. R version 3.6.3 (R Core Team, 2020) served as the primary tool for data management. In cases of missing climate data, the researchers imputed the missing data by substituting each missing value with the average of identified values of that attribute using equation (1) adapted from [57];

$$\mathrm{Y}_{\mathrm{i}}^{\mathrm{j}} = \sum \mathrm{k} \in \mathrm{r(complete)} \frac{\mathrm{Y}_{\mathrm{k}}^{\mathrm{j}}}{\mathrm{n}_{|\mathrm{r(complete)}|}} \tag{1}$$

Where Y_i^j denotes the j^{th} the missing attribute of the i^{th} observation
$r(complete)$ denotes non-missing values from Y_i
$r_{|I(complete)|}$ denotes the total number of observations where the j^{th} attribute is not missing.

3.1.3 Data Transformation

***Normalization*.**
The fact that attribute values were measured on different scales .e.g. temperature in degrees Celsius and rainfall in milimetres implied that the attributes couldn't be compared meaningfully [58]. Hence data normalization by standardization (z-scores) was undertaken to adjust for the above discrepancies, thereby ensuring that all continuous attribute values are scaled and belong in similar ranges [58] [59]. The mean and standard deviation of the attributes were used for normalization as illustrated in equation (2);

$$\mathrm{B} = \frac{x_i - \mu}{\sigma} \tag{2}$$

where $\mathrm{B} = normalised\ attribute\ value,\ x_i = original\ attribute\ value$
μ=*mean of attribute* , σ =*attribute standard deviation*

***Discretization*.**
According to [60], data discretization enhances the comprehensibility of the discovered previously unknown knowledge from databases. Hence data discretization was undertaken in order to produce a homogeneous group of antecedent attributes since the dataset comprised both continuous and categorical attributes, thus alleviating outliers and conquering noise accumulation [59] [60] [61].

3.1.4 Training and Testing data

The dataset was split into training and testing datasets. 80% of the data was assigned to the training group for the development of the classifiers. The rest of the data (20% of the total cases) was assigned to the validation groups for the assessment of model performance [62].

3.2 Ethical statement on data access

The datasets were accessed with official permission granted from the Ministry of Health, Uganda National Meteorological Authority and Uganda Bureau of Statistics.

[1] www.health.go.ug
[2] www.unma.go.ug
[3] www.ubos.org

3.3 Analysis

3.3.1 Proposed hybrid data mining model

The researchers' proposed hybrid model was developed in two phases, utilizing two different single techniques. In the first phase, decision trees were used in a cascaded style for important attribute extraction based on the gain ratio to pre-process the data. Then, the output of the first stage was employed to construct the second stage weighted naïve Bayesian classifier as the prediction model. The overall methodological workflow of the hybrid model is illustrated in Figure 1.

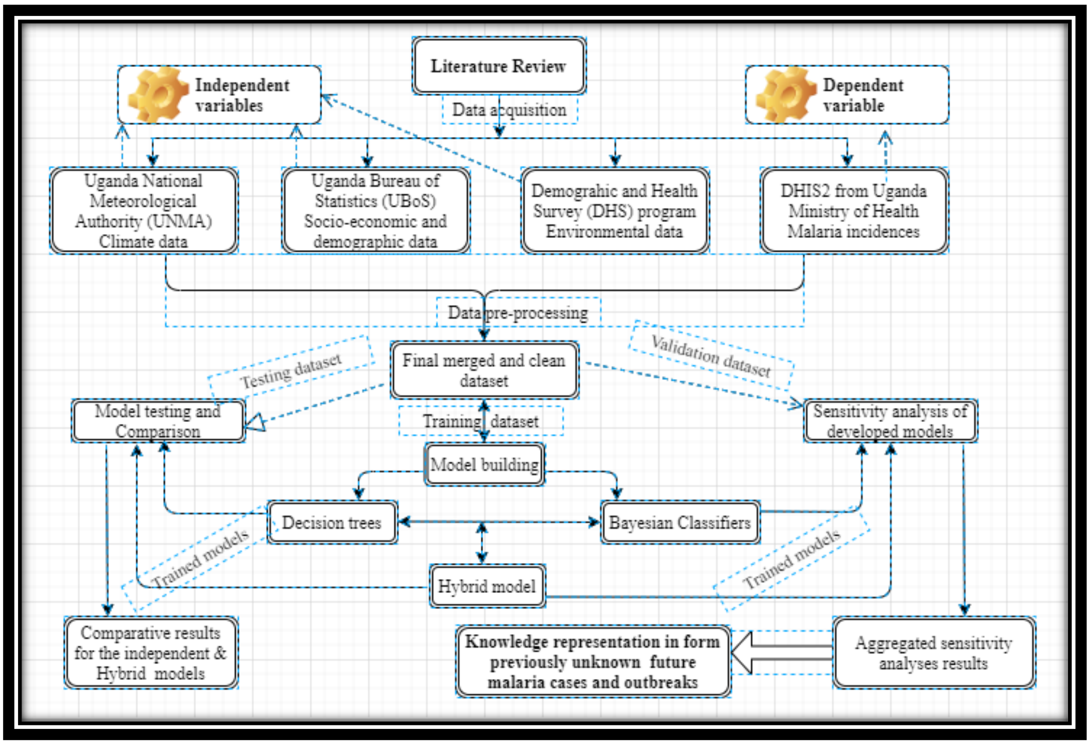

Literature Review
Independent variables
Data acquisition
Dependent variable
Uganda National Meteorological Authority (UNMA) Climate data
Uganda Bureau of Statistics (UBoS) Socio-economic and demographic data
Demographic and Health Survey (DHS) program Environmental data
DHIS2 from Uganda Ministry of Health Malaria incidences
Data pre-processing
Testing dataset
Final merged and clean dataset
Validation dataset
Model testing and Comparison
Training dataset
Sensitivity analysis of developed models
Model building
Decision trees
Bayesian Classifiers
Trained models
Hybrid model
Trained models
Comparative results for the independent & Hybrid models
Knowledge representation in form previously unknown future malaria cases and outbreaks
Aggregated sensitivity analyses results

Figure 2: The overall framework and methodological workflow of the hybrid model

First Phase

Under this phase, the researcher employed the decision tree technique based on the C4.5 algorithm to identify the most significant attributes to improve estimates [63]. Additionally, the researchers assigned gain ratio values as weights for each attribute based on the fact that weighted classification assigns various degrees of significance to different attributes and classes in order to denote the relative importance of each attribute and class [64].

The following procedures were followed in the first phase.

i) Given a training dataset, T with instances y_i where $y_i = \{y_1, y_2, \dots, y_n\}$.The training dataset T is defined by attributes B_i .i.e. $T = \{B_1, B_2, \dots, B_n\}$ and generate an attributes list B which has v possible values. The training data also belongs to a set of classes $Z = \{Z_1, Z_2, \dots, Z_n\}$

ii) Build a decision tree classifier employing the C4.5 formula adapted from [65];

a) Entropy for the root node:

$$\text{Entropy}(\text{T}) = \sum_{\text{i}=1}^{\text{m}}(\text{p}_\text{i})\log_2(\text{p}_\text{i})$$

Where p_i denotes the probability of the target attribute

b) Entropy of Attribute *(B)* for attribute list B with respect to the root note *(T)*

$$\text{Entropy}_\text{B}(\text{T}) = \sum_{\text{j}=1}^{\text{v}}\frac{|\text{T}_\text{j}|}{|\text{T}|} * \text{Entropy}(\text{T}_\text{j})$$

where T_j is a collection of the instances in the dataset T with attribute B having value j

c) Compute the information gain for each attribute

$$\mathrm{Info\_Gain(T)} = \mathrm{Entropy(T)} - \mathrm{Entropy_B(T)}$$

d) Compute the split information *(V) for a set of attributes* (T_i) *and* (T_j)

$$\mathrm{Splitinfo(B)}{=}-\left[\left|\frac{\mathrm{T_i}}{\mathrm{B}}\right| \log_2 \left[\left|\frac{\mathrm{T_i}}{\mathrm{B}}\right|\right] + \left|\frac{\mathrm{T_j}}{\mathrm{B}}\right| \log_2 \left[\left|\frac{\mathrm{T_j}}{\mathrm{B}}\right|\right]\right]$$

e) Compute the gain ratio*(B)* for attribute list *B w*ith respect to the root note *(T*)

$$\mathrm{Gainratio(B)} = \left(\frac{\mathrm{Info\_Gain(B)}}{\mathrm{Splitinfo(B)}}\right) \qquad (3)$$

f) After computing information gain ratio for each attribute on the decision tree classifier, assign and initialize the weight (W_i) for each attribute(B_i), where $B_i \in T$ as the gain ratio value of the respective attribute.

NB: The weight for an attribute is computed as $W_i = Gainratio(B)_i$

g) If the attribute, $B_i \in T$, is not tested in the decision tree, then the weight (W_i) of the attribute, (B_i), was initialized to zero.

h) Thus, the parent node of the tree will have a higher weight value in comparison with those of its child nodes [66].

Second Phase

Under this phase, the researchers employed the naïve Bayes technique. The naïve Bayes technique is a Bayesian classifier grounded on statistical methods and utilizes Bayes Theorem proposed by Thomas Bayes to calculate unknown conditional probabilities [67]. Bayesian classifiers handle real and discrete data and make the computation process easier. The main advantages of naive Bayes classifiers are that they are resilient to noise and outliers, and they handle missing values by ignoring the instance during probability estimate calculations [68]. The naïve Bayes technique is referred to as "naïve" due to the fact that it assumes that the occurrence of a certain attribute is independent of other attributes conditional on a similar consequent target value [67].

On the other hand, the naïve Bayes technique assumes conditional independence of antecedent attributes given the target attribute [69] [70], which hardly ever holds in real-world scenarios [71]; weakening its performance in models with complex attribute dependencies [72].

Hence in order to alleviate the independence hypothesis of the naive Bayes, the researchers applied weight values derived from the first phase to the attribute set based on each attribute's importance in the classification process [45]. According to [45], the Naive Bayes algorithm is derived from Bayes' theorem (equation 4);

$$\mathrm{P(X/Y)} = \frac{\mathrm{P(Y/X)P(X)}}{\mathrm{P(Y)}} \qquad (4)$$

Where $\mathrm{P(X/Y)}$=probability of X given Y has occurred
$\mathrm{P(Y/X)}$=probability of Y given X has occurred
$\mathrm{P(X)}$ and $\mathrm{P(Y)}$ are probabilities of X and Y occurring independently from each other.

However, bases on the assumption of independence among antecedent attributes, equation (4) is transformed to equation (5) for the naïve Bayes formula [67];

$$\mathrm{P(Z_n/y_1, y_2, \ldots, y_n)} = \frac{\mathrm{P(y_1, y_2.y_3, \ldots., y_n, Z_i)}}{\prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P(y_i)}} \qquad (5)$$

Where $\mathrm{P}(Z_n)$= the prior probability of the class that reflects background knowledge due to the chance of Z to be correct.
$\mathrm{P}(y_i)$= the probability of y to be observed
$\mathrm{P(Z_n/y_i)}$= the posterior probability of class (malaria incidence) given predictor (attribute).
$\mathrm{P(y_i/Z_n)}$= the probability of observing y given Z holds

Hence simplifying the numerator on the right hand side in equation (5) leads to equation (6);

DOI: 10.12856/JHIA-2020-v7-i2-289

$$\mathrm{P(Z_n/y_1, y_2, \dots, y_n)} = \frac{\mathrm{P(y_n)} \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P}(y_i/Z_n)}{\prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P(y_i)}} \tag{6}$$

Since the denominator in equation (6) is invariant across various consequent attribute classes, it can be dropped as illustrated in equation (7).

$$\mathrm{x = argmax(P(Z_n/y_i) = \ P(y_n)} \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P}(y_i/Z_n)) \tag{7}$$

Where x is the class with the highest probability given a set of attributes.

The following procedures were followed in the second phase.

i) The researcher will compute the class conditional probabilities utilizing only the significant attributes nominated by decision tree technique in the first phase (i.e. $W_i \neq 0$) and classify each instance $B_i \in T$ based on the gain ratio values.

ii) Assume that there are m classes, $Z_1, Z_2, \dots, Z_m$. Given an object Y, the classifier will predict that B belongs to the class having the highest posterior probability.

That is, the naïve Bayesian classifier predicts that tuple B belongs to the class Z_i if and only if

$$\mathrm{P(Z_i/Y) > P\big(Z_j/Y\big) \text{ for } 1 \leq j \leq m, j \neq i}$$

iii) Thus the $P(Z_i/Y)$ needs to be maximized. The class Z_i for which $P(Z_i/Y)$ is maximized is called the maximum posterior hypothesis.

iv) Compute the class conditional probabilities using the weights of significant attributes as exponential constraints using the formula below [72];

$$\mathrm{P(Z/y_i) = P(Z)} \prod_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{P(y_i|Z)^{Wi}} \tag{8}$$

Where W_i refers to the weight of the attribute, (y_i), which effects on class conditional probability calculation as an exponential parameter.

$P(y/Z) = Probability\ of\ attribute\ y\ given\ Z\ has\ occured$

$P(Z) = Probability\ of\ \ consequent\ attribute(class)$

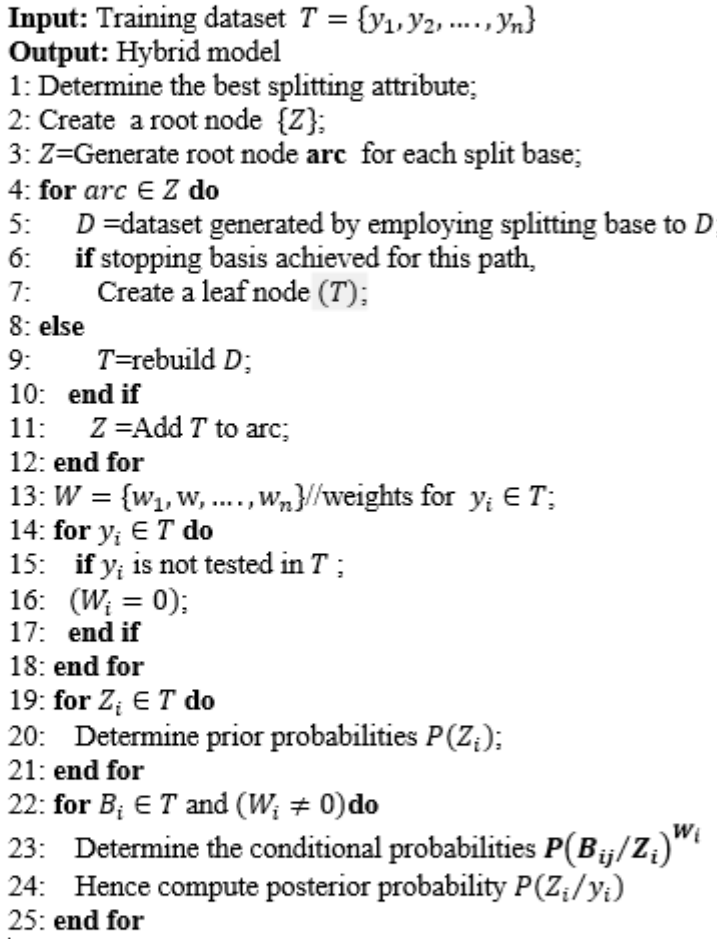
$P(y) = Probability\ of\ \ antecedent\ attribute$

$P(Z/y) = Probability\ of\ event\ Z\ \ given\ y\ has\ occured$

v) The class conditional probabilities of the non-significant attributes ($W_i = 0$) by decision trees will not be employed in the prediction of estimates in the second phase.

vi) The researcher reiterated this process until all the attributes were correctly predicted.

The algorithm in Figure 2 adapted from [63] [73] outlines the proposed hybrid algorithm;



```
Input: Training dataset $T = \{y_1, y_2, \dots, y_n\}$
Output: Hybrid model
Determine the best splitting attribute;
Create a root node $\{Z\}$;
$Z$=Generate root node arc for each split base;
for $arc \in Z$ do
    $D$ =dataset generated by employing splitting base to $D$;
    if stopping basis achieved for this path,
        Create a leaf node $(T)$;
    else
        $T$=rebuild $D$;
    end if
    $Z$ =Add $T$ to arc;
end for
$W = \{w_1, \mathrm{w}, \dots, w_n\}$//weights for $y_i \in T$;
for $y_i \in T$ do
    if $y_i$ is not tested in $T$ ;
    $(W_i = 0)$;
    end if
end for
for $Z_i \in T$ do
    Determine prior probabilities $P(Z_i)$;
end for
for $B_i \in T$ and $(W_i \neq 0)$do
    Determine the conditional probabilities $\boldsymbol{P(B_{ij}/Z_i)}^{w_i}$
    Hence compute posterior probability $P(Z_i/y_i)$
end for
```

Figure 2: Proposed hybrid algorithm

3.4 Goodness of Fit

The researchers used k-fold cross-validation (CV) method and six performance evaluation metrics.

3.4.1 Classifier Validation Method

The K-fold cross-validation method was employed [74]. In k- fold validation, the set of training data was divided into k- groups of equal size. In our experiment, we used the K=10 cross-validation due to the fact that its performance is reliable [60]. Hence under the 10-fold cross-validation process, 90% of the data was used for training and 10% of the data was used for testing purposes.

3.4.2 Performance Evaluation Metrics

The researchers used a confusion matrix in order to evaluate the performance of the classifiers based on various performance evaluation metrics as illustrated in Table 1.

Table 5: Performance metrics computed

Metric	Formula	Description
Accuracy/recognition rate (%)	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$	Number of correctly classified malaria incidence thresholds to total number of incidences
Sensitivity/ true positive rate (%)/Recall	$\frac{TP}{(TP+FN)}$	The proportion of low incidence thresholds that are correctly classified

DOI: 10.12856/JHIA-2020-v7-i2-289

Specificity/ true negative rate (%)	$\frac{TN}{(FP + TN)}$	The proportion of “moderate” incidence thresholds that are correctly classified
Precision (%)	$\frac{TP}{(TP + FP)}$	The proportion of “low” incidences predicted to be “low” that are truly “low” incidences
F-Score/F-measure	$\left(\frac{2 * Precision * Recall}{Precision + Recall}\right)$	The harmonic mean of precision and recall
Area under the Curve (AUC)		The area under the curve (AUC) is a model goodness-of-fit measure that compares it to a baseline 50% measure (the straight line).

Source: Mehdiyev, Enke, Fettke, & Loos [75]

In the context of this study, the entries in the confusion matrix were defined as:

i) True positive (TP): is the number of actual “LOW” instances classified as “LOW”.

ii) False-positive (FP): is the number of actual “MODERATE” instances classified as “LOW”

iii) False Negative (FN): is the number of actual “LOW” instances classified as “LOW”.

iv) True Negative (TN): is the number of actual “MODERATE” instances classified as “MODERATE”.

3.5 Software Tools

The data processing and analysis was undertaken entirely in R, version 3.6.3 [76], by means of R packages “funModeling” version 1.9.3 [77], “dplyr” version 0.8.5 [78], “tidyr” version 1.0.2 [79], “caret” version 6.0.86 [80], “reshape2” version 1.4.4 [81].

4 Results

The overall performance of the classifiers was evaluated based on their prediction accuracy in classifying the instances of the data set into low and moderate malaria incidence thresholds. The researchers utilized 10-fold cross-validation to assess the performance of the three classifiers on previously unlearned data. Figure 3 shows the classification results of the test data.

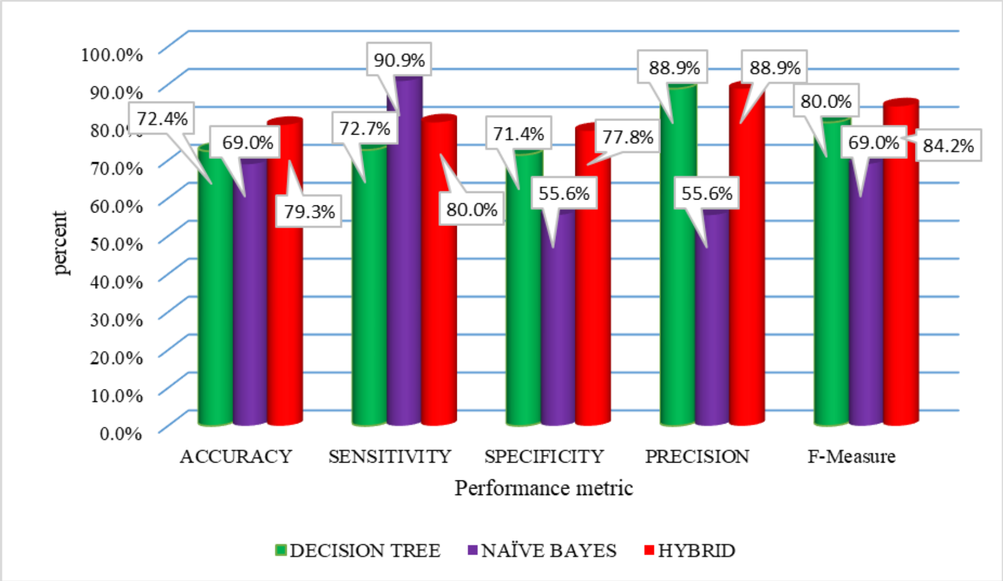


Figure 3: Comparison of classifiers’ performance using 10 fold cross-validation

The performance metrics for all the classifiers were separately identified using trained models that were fit on previously untrained test data. According to figure 2, the hybrid model attained the highest performance with respect to the accuracy, specificity, and F-measure metrics recorded at 79.3%, 77.8%, and 84.2% respectively. On the other hand, the naïve Bayes classifier registered the highest sensitivity at 90.9% compared to 72.7% registered by decision tree and 80% obtained by the hybrid classifier. The achieved accuracy results indicate that the proposed hybrid model outperformed the independent decision tree and naïve Bayes classifiers by 6.9% and 10.3% respectively.

4.1 Receiver Operating Characteristics (ROC) Curve

The ROC curve (Figure 4) is a graphical plot that symbolizes how the performance of the sensitivity and specificity of a classifier varies in relation to one another (Wu, Yang, Huang, He, & Wang, 2018; Zhu, Idemudia, & Feng, 2019). The ROC permitted the researchers to assess the performance of the developed models at various thresholds. Figure 3.0 reveals that the Area under the Curve (AUC) was recorded at 67.17%, 88.38%, and 86.87% for the decision tree, naïve Bayes, and hybrid classifier respectively. A random model would have an AUC of 50% (the straight line), give that it basically dissects the graph (Winters, 2015). Hence the generated the ROC curve for all the classifiers outperform a random model (straight line).

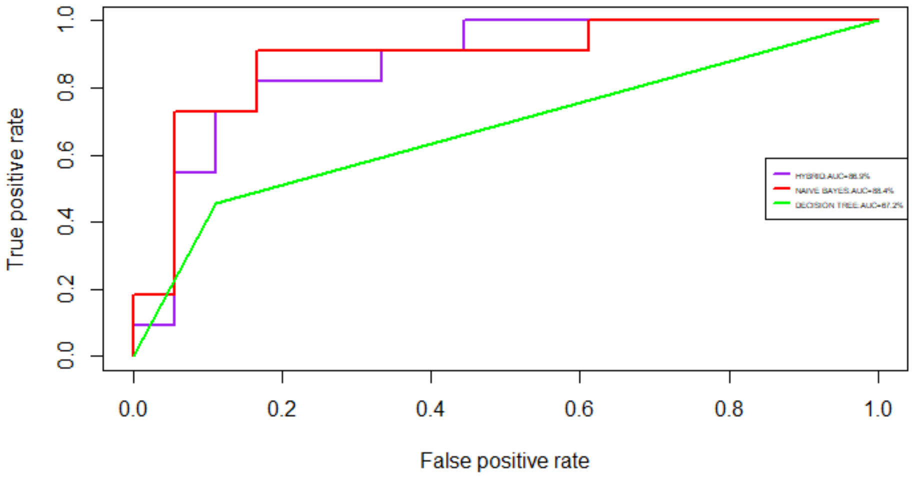


Figure 4: Comparison of the ROCs for the classifiers at various thresholds

Based on Figure 4, the hybrid and naïve Bayes classifiers attained a high sensitivity (True positive rate) of approx.70% at a very low false-positive rate (1-specificity). Nevertheless, the hybrid model denoted by the purple line returned a better cut-off determination threshold since it yielded a higher true positive rate at lower false-positive rates compared to the naïve Bayes classifier. The decision tree was a poor classifier given that it achieved a high True positive rate at the cost of a high false-positive rate.

4.2 Reliability of the proposed hybrid model

To further demonstrate and evaluate the reliability of the developed hybrid model on high dimensional data, the researchers applied the model on three demonstration datasets from various application domains. The datasets were sourced from the UCI machine learning repository[4] [82]; an assembly of databases used by several scholars [83] [84] [85] [86] for experimental investigation of machine learning techniques. Table 2 shows the characteristics of the demonstration datasets.

[4] https://archive.ics.uci.edu/ml/datasets.php

J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

Table 2: Datasets used to test the reliability of the hybrid model

Dataset name	Description	Number of attributes	Number of observations	Target attribute	Data Source
Heart failure	Dataset for predicting mortality caused by Heart Failure	13	299	Death_Event	https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records
Heart_UCI	Data set with attributes to detect the presence of heart disease in the patient	14	303	target	https://archive.ics.uci.edu/ml/datasets/Heart+Disease
Wine quality	Data set with attributes to determine which physiochemical properties make a wine 'good'!	12	1599	Quality (=<6.5=”BAD” >6.5=”GOOD”	https://archive.ics.uci.edu/ml/datasets/wine+quality

The researchers subjected the datasets in table 2 to data pre-processing steps similar to the malaria incidence rate dataset collected from a known population in Kampala. The researchers employed k-fold cross-validation to compare the performance of the developed hybrid model in terms of the F-measure metric with the independent decision tree and naïve Bayes classifiers. The results are shown in Table 3.

Table 3: Performance of the decision tree, naive Bayes and proposed hybrid models on various demonstration datasets

Dataset	Model	Accuracy	Sensitivity	Specificity	Precision	F-measure
heart_failure	Decision tree	81.7%	86.4%	68.8%	88.4%	87.4%
	Naïve Bayes	81.7%	90%	65%	83.7%	86.7%
	Hybrid	86.7%	90.7%	76.5%	90.7%	90.7%
heart	Decision tree	83.6%	89.5%	81%	68%	77.3%
	Naïve Bayes	83.6%	75.9%	90.6%	88%	81.5%
	Hybrid	88.5%	82.1%	93.9%	92%	86.8%
winequality	Decision tree	84.1%	86.8%	52%	95.5%	90.9%
	Naïve Bayes	79.7%	88.9%	39%	86.6%	87.7%
	Hybrid	84.7%	88.2%	54.5%	94.4%	91.2%

Table 3 reveals that the hybrid model improved the performance the independent models across all the datasets. The proposed hybrid model outperformed the independent models by obtaining the highest F-measure score of 90.7%, 86.8% and 91.2% on the “heart_failure”, “heart” and “winequality” datasets respectively. Similarly, the proposed hybrid model outperformed the independent models in terms of predictive accuracy in all the demonstration datasets. The attained results indicate that the proposed hybrid model could help in enhancing the performance of the independent data mining techniques.

5 Discussion

The main purpose of this study is to build a hybrid data mining approach robust to noises, dependence, and data complexity to improve the predicting of malaria incidence rates, leading to early prediction of malaria occurrences and thus dipping the transmission risk in the community. In this work, the researchers

compared a hybrid data mining technique with single data mining techniques in the form of decision trees and naïve Bayes classifiers. The results showed that the hybrid model outperformed the independent models in terms of classification accuracy, specificity, and F-measure. To assess the robustness of the hybrid model, the researchers undertook further experiments using datasets from different application domains. These datasets were obtained from the UCI machine learning repository [82]. Findings from these experimental analyses alluded to the researchers' initial findings with the hybrid model outperforming the individual models. Furthermore, the experimental results showed that employing the hybrid model by weighting the naïve Bayes classifier using gain ratio values emanating from the C4.5 algorithm enhances the naïve Bayes algorithm. This is similar to findings of [72] [87] [88] [89] who alluded that nullifying the conditional independence assumption of the naïve Bayes through weighting can help improve its performance.
Above all, the results are in agreement with the concept that high sensitivity and specificity may not be achievable in real-world scenarios concurrently [90] because they are inversely related, implying that as the specificity increases, the sensitivity decreases and vice versa [91]. Hence there is a trade-off between sensitivity and specificity with the hybrid model recording a lower sensitivity of 80% compared to the naive Bayes model registered at 90.9%. However, a similar trend was observed in terms of specificity with the hybrid model registering the highest specificity of 77.8% compared to 71.4% and 55.6% recorded for the decision tree and naive Bayes models respectively.

The study faced a key challenge of available monthly data being limited to a few predictor attributes for the period under investigation and hence the researchers were unable to subject the developed hybrid model to a higher dimensional dataset from a known population which would return more reliable and robust performance results [92]. Additionally, the researchers did not take into consideration the effect of biasedness associated with predictions emanating from imbalanced data [93].

6 Conclusion

This study aimed to develop a hybrid model for predicting reliable estimates of malaria incidence thresholds. After reviewing relevant literature, the researchers proposed a hybrid model which was an amalgamation of the C4.5 decision tree and naïve Bayes classifiers. The hybrid model was developed in two phases with phase one employing the C4.5 algorithm to generated information gain ratio values for antecedent attributes which were used as attribute weights for the naïve Bayes classier in the second phase. Empirically, the developed hybrid model outperformed both independent decision tree and naïve Bayes models. Notably, the hybrid model outperformed the independent decision tree and naïve Bayes classifiers in terms of accuracy by 6.9% and 10.3% respectively. Hence merging several independent homogeneous predictive data mining techniques enhances the accuracy of the estimates leading to reliable estimates.

Acknowledgements

The authors extend their appreciation to Mr. Douglas Candia and Mr. Frank Namugera who contributed to improving this research. This research was partly funded by Makerere University through the Staff Development, Welfare and Retirement Benefits Committee (SDWRBC).

Conflict of Interest

The authors declare that they have no competing interests.

References

[1] Garg, P., & Vishwakarma, S. K. (2019). An efficient prediction of share price using data mining techniques. International Journal of Engineering and Advanced Technology, 8(6), 3110–3115. https://doi.org/10.35940/ijeat.F9085.088619

[2] Hakizimana, L., Cheruiyot, K., Kimani, S., & Nyararai, M. (2017). A Hybrid Based Classification and Regression Model for Predicting Diseases Outbreak in Datasets. International Journal of Computer (IJC), 27(1), 69–83.

[3] Tan, J., & Wang, F. (2017). A Hybrid Mining Approach to Facilitate Health Insurance Decision : Case Study of Non-Traditional Data Mining Applications in Taiwan NHI Databases. Proceedings of the 50th Hawaii International Conference on System Sciences, 3253–3262.

[4] Gidron, Y. (2013). Reliability and Validity. In M. D. Gellman & J. R. Turner (Eds.), Encyclopedia of Behavioral Medicine. Springer Science+Business Media. https://doi.org/10.1007/978-1-4419-1005-9

[5] Easterby-Smith, M., Thorpe, R., & Jackson, P. R. (2008). Management Research (3rd ed.). Sage.

[6] Curry, A., Flett, P., & Hollingsworth, I. (2006). Managing information and systems:The Business Perspective. Routledge, New York-London.

[7] Appiah, S. ., Otoo, H., & Nabubie, B. (2015). Times series analysis of malaria cases in Ejisu- Juaben Municipality. International Journal of Scientific and Technology Research, 4(06).

[8] Carillo, M., Largo, F., & Ceballos, R. (2018). Principal Component Analysis on the Philippine Health Data. International Journal of Ecological Economics and Statistics, 39(1), 91–96.

[9] Hussien, H. H. (2019). Malaria's association with climatic variables and an epidemic early warning system using historical data from Gezira State , Sudan. Heliyon, 5. https://doi.org/10.1016/j.heliyon.2019.e01375

[10] Muwanika, F., Atuhaire, L., & Ocaya, B. (2017). Journal of Medical Diagnostic Prediction of Monthly Malaria Incidence in Uganda and its Implications for Preventive Interventions. Journal of Medical Diagnostic Methods, 6(2). https://doi.org/10.4172/2168-9784.1000248

[11] Twumasi-Ankrah, S., Wa, P., Nyantakyi, K., & Addo, D. (2019). Comparison of Statistical Techniques for Forecasting Malaria Cases in Ghana. Journal of Biostatistics and Biometric Applications, 4(1), 1–9.

[12] Basco, A., & Senthilkumar, N. C. (2017). Real-time analysis of healthcare using big data analytics. IOP Conference Series: Materials Science and Engineering, 263(4). https://doi.org/10.1088/1757-899X/263/4/042056

[13] Hu, C. H., Lee, H. S., Lara, E., & Gan, S. (2018). The Ensemble and Model Comparison Approaches for Big Data Analytics in Social Sciences. Practical Assessment, Research & Evaluation, 23(17).

[14] Wyber, R., Vaillancourt, S., Perry, W., Mannava, P., Folaranmi, T., & Anthony, L. (2015). Big data in global health : improving health in low- and middle-income countries. Big Data in Health Care, January, 203–208.

[15] Bains, J. K. (2016). Big Data Analytics in Healthcare- Its Benefits, Phases and Challenges. International Journal of Advanced Research in Computer Science and Software Engineering, 6(4). www.ijarcsse.com

[16] Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. Journal of Big Data, 1(2). http://www.journalofbigdata.com/content/1/1/2

[17] Aparna, K., Reddy, C. S., Prabha, S., & Srinivas, V. (2014). Disease prediction in data mining techniques. International Journal of Computer Science and Technology, 5(2), 17–21.

[18] Sahay, S. (2016). Big data and public health: Challenges and opportunities for low and middle income countries. In Communications of the Association for Information Systems (Vol. 39). https://doi.org/10.17705/1CAIS.03920

[19] Agyapong, K. B., Hayfron-Acquah, J., & Asante, M. (2016). An Overview of Data Mining Models (Descriptive and Predicitve). International Journal of Software & Hardware Research in Engineering, 4(5), 53–60. https://doi.org/10.1007/978-3-319-13084-2_59

[20] Patil, T. R., & Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal Of Computer Science And Applications, 6(2).

[21] Krishnaiah, V., Narsimha, G., & Subhash, C. (2013). Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. (IJCSIT) International Journal of Computer Science and Information Technologies, 4(1), 39–45.

[22] Gorade, S., Ankit, D., & Preetesh, P. (2017). A Study Some Data Mining Classification Techniques. International Research Journal of Engineering and Technology, 4(1), 210–215. https://doi.org/10.21884/ijmter.2017.4031.zt9tv

[23] Thorat, S., & Kute, S. (2014). Medical Data Mining Life Cycle and its Role in Medical Domain. International Journal of Computer Science and Information Technologies, 5(4), 5751–5755

[24] Ying-ying, W., Yi-bin*, L., & Xue-wen, R. (2017). Improvement of ID3 Algorithm Based on Simplified Inform ation Entropy and Coordination Degree. IEEE, 1526–1530.

[25] Ahlawat, A., & Suri, B. (2016). Improving Classification in Data mining using Hybrid algorithm. IEEE, 2–5.

[26] Kumar, P., & Wahid, A. (2015). Performance Evaluation of Data Mining Techniques for Predicting Software Reliability. International Journal of Computer and Systems Engineering, 9(8), 1946–1953

[27] Anwar, H., Qamar, U., & Qureshi, A. W. (2014). Global optimization ensemble model for classification methods. Scientific World Journal, Vol. 2014. https://doi.org/10.1155/2014/313164

[28] Abuassba, A. O. M., Zhang, D., Luo, X., Shaheryar, A., & Ali, H. (2017). Improving Classification Performance through an Advanced Ensemble Based Heterogeneous Extreme Learning Machines. Computational Intelligence and Neuroscience, Vol. 3405463. https://doi.org/10.1155/2017/3405463

[29] Fan, Jianqing, Han, F., & Liu, H. (2014). Challenges of Big Data analysis. In National Science Review (Vol. 1). https://doi.org/10.1093/nsr/nwt032

[30] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

[31] La Sorte, F. A., Lepczyk, C. A., Burnett, J. L., Hurlbert, A. H., Tingley, M. W., & Zuckerberg, B. (2018). Opportunities and challenges for big data ornithology. The Condor, 120(2), 414–426. https://doi.org/10.1650/condor-17-206.1

[32] Almarabeh, H., & Amer, E. (2017). A Study of Data Mining Techniques Accuracy for Healthcare. International Journal of Computer Applications, 168(3), 12–17. https://doi.org/10.5120/ijca2017914338

[33] DEEPAK, S. (2016). Knowledge Discovery With Hybrid Data Mining Approach. DAYALBAGH EDUCATIONAL INSTITUTE.

[34] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, Vol. 19, pp. 1–16. https://doi.org/10.1186/s12911-019-1004-8

[35] Lal, A., & Kumar, C. R. . (2017). Hybrid Classifier for Increasing Accuracy of Fitness Data Set. IEEE, 1246–1249.

[36] Nimala, K., & ThamizhArasan, R. (2018). HYBRID DATA MINING APPROACHES FOR ACCURATE PREDICTION OF DIABETES AND HEART DISEASE. International Journal of Pure and Applied Mathematics, 120(6), 2693–2705

[37] Singhal, N., & Ashraf, M. (2015). Performance Enhancement of Classification Scheme in Data Mining using Hybrid Algorithm. International Conference on Computing, Communication & Automation, 138–141. https://doi.org/10.1109/CCAA.2015.7148360

[38] Kaushik, D., & Kaur, K. (2016). Application of Data Mining for High Accuracy Prediction of Breast Tissue Biopsy Results. IEEE Transactions on Knowledge and Data Engineering, 40–45.

[39] Kazienko, P., Lughofer, E., & Trawiński, B. (2011). Hybrid and Ensemble Methods in Machine Learning. New Generation Computing, 29(3), 241–244. https://doi.org/10.1007/s00354-011-0300-3

[40] Castillo, O., Melin, P., & Pedrycz, W. (2007). Hybrid Intelligent Systems: Analysis and Design (Studies in Fuzziness and Soft Computing). Springer, Berlin Heidelberg.

[41] Corchado, E., Abraham, A., & De Carvalho, A. (2010). Hybrid intelligent algorithms and Applications. Information Sciences, 180(14), 2633–2814

[42] Kajdanowicz, T., Kazienko, P., & Kraszewski, J. (2010). Boosting algorithm with sequence-loss cost function for structured prediction. In R. M. Graña, E. Corchado, & S. M. . Garcia (Eds.), Hybrid Artificial Intelligence Systems (pp. 573–580). Berlin, Heidelberg: Springer.

[43] Kuncheva, L. (2004). Combining pattern classifiers: Methods and algorithms. Southern Gate, Chichester, West Sussex, England: John Wiley & Sons.

[44] Sumana, B. V., & Santhanam, T. (2014). An Empirical Comparison of Ensemble and Hybrid Classification. Proc. of Int. Conf. on Recent Trends in Signal Processing, Image Processing and VLSI, 4322–4322. https://doi.org/10.1158/1538-7445.am2015-4322

[45] Hamblin, D., Wang, D., & Chen, G. (2016). Measurement classification using hybrid weighted Naive Bayes. IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, CIVEMSA 2016 - Proceedings. https://doi.org/10.1109/CIVEMSA.2016.7524248

[46] Şen, B., Uçar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Systems with Applications, 39(10), 9468–9476. https://doi.org/10.1016/j.eswa.2012.02.112

[47] Jiang, L., & Li, C. (2011). Scaling up the accuracy of decision-tree classifiers: A naive-bayes combination. Journal of Computers, 6(7), 1325–1331. https://doi.org/10.4304/jcp.6.7.1325-1331

[48] WHO. (2019). Malaria. Retrieved November 26, 2019, from WHO website: https://www.who.int/en/news-room/fact-sheets/detail/malaria

[49] World Health Organization[WHO]. (2019). World Malaria Report 2019. Retrieved from https://www.who.int/publications-detail/world-malaria-report-2019

[50] Ogwoka, T. M., Cheruiyot, W., & Okeyo, G. (2015). A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms. International Journal of Computer Applications Technology and Research, 4(9), 693–697. https://doi.org/10.7753/ijcatr0409.1009

[51] Dubey, V. K., & Saxena, A. K. (2016). Hybrid classification model of correlation-based feature selection and support vector machine. 2016 IEEE International Conference on Current Trends in Advanced Computing, ICCTAC 2016, 1–6. https://doi.org/10.1109/ICCTAC.2016.7567338

[52] Raghavendra, S., & Indiramma, M. (2016). Hybrid data mining model for the classification and prediction of medical datasets. Int. J. Knowledge Engineering and Soft Data Paradigms, 5(3/4), 262–284.

[53] Ren, Y., Fei, H., Liang, X., Ji, D., & Cheng, M. (2019). A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records. BMC Medical Informatics and Decision Making, 19(51). https://doi.org/10.1186/s12911-019-0765-4

[54] Malathi, D., Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., & Sangaiah, K. (2019). Hybrid Reasoning-based Privacy-Aware Disease Prediction Support System. Computers and Electrical Engineering, 73, 114–127. https://doi.org/10.1016/j.compeleceng.2018.11.009

[55] Ju-young, S., Rang, K. K., & Jong-chul, H. (2020). Seasonal forecasting of daily mean air temperatures using a coupled global climate model and machine learning algorithm for field-scale agricultural management. Agricultural and Forest Meteorology, Vol. 281. https://doi.org/10.1016/j.agrformet.2019.107858

[56] Fan, Junliang, Wu, L., Ma, X., Zhou, H., & Zhang, F. (2020). Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. Renewable Energy, 145, 2034–2045. https://doi.org/10.1016/j.renene.2019.07.104

[57] Benhar, H., Idri, A., & Fernandez-Aleman, J. (2020). Data preprocessing for heart disease classification: A systematic literature review. In Computer Methods and Programs in Biomedicine. https://doi.org/10.1016/j.cmpb.2020.105635

[58] Aggarwal, C. (2015). Data mining:The Text book. https://doi.org/10.1007/978-3-319-14142-8 14

[59] Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining : An evaluation of classifier sensitivity in direct marketing. European Journal of Operational Research, 173, 781–800. https://doi.org/10.1016/j.ejor.2005.07.023

[60] Witten, I., Frank, E., & Hall, M. (2011). Data mining: Practical machine learning tools and techniques (3rd ed.). Morgan Kaufmann.

[61] Maslove, D. M., Podchiyska, T., & Lowe, H. J. (2013). Discretization of continuous features in clinical datasets. 544–553. https://doi.org/10.1136/amiajnl-2012-000929

[62] Li, G., Zhou, X., Liu, J., Chen, Y., Zhang, H., Chen, Y., … Nie, S. (2018). Comparison of three data mining models for prediction of advanced schistosomiasis prognosis in the Hubei province. PLoS Neglected Tropical Diseases, 12(2), 1–19. https://doi.org/10.1371/journal.pntd.0006262

[63] Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert Systems with Applications, 41, 1937–1946. https://doi.org/10.1016/j.eswa.2013.08.089

[64] Polo, J. L., Berzal, F., & Cubero, J. C. (2007). Weighted Classification Using Decision Trees for Binary Classification Problems. II Congreso Español de Informática, 333–341.

[65] Prasad, N., & Naidu, M. M. (2013). Gain Ratio as Attribute Selection Measure in Elegant Decision Tree to Predict Precipitation. EUROSIM Congress on Modelling and Simulation, 141–150. https://doi.org/10.1109/EUROSIM.2013.35

[66] Hall, M. (2006). A decision tree-based attribute weighting filter for Naive Bayes. In Research and Development in Intelligent Systems XXIII - Proceedings of AI 2006, the 26th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. https://doi.org/10.1007/978-1-84628-663-6-5

[67] Yildirim, P., & Birant, D. (2014). Naive Bayes Classifier for Continuous Variables using Novel Method (NBC4D) and Distributions. IEEE.

[68] Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal, 16(3), 261–273. https://doi.org/10.1016/j.eij.2015.06.005

[69] Ali, M. F. M., Asklany, S. A., El-wahab, M. A., & Hassan, M. A. (2019). Data Mining Algorithms for Weather Forecast Phenomena : Comparative Study. International Journal of Computer Science and Network Security, 19(9), 76–81.

[70] MAKHTAR, M., NAWANG, H., & SHAMSUDDIN, S. N. W. (2017). Analysis on Students Performance Using Naïve classifier. Journal of Theoretical and Applied Information Technology, 95(16), 3993–4000. Retrieved from www.jatit.org

[71] Zhang, H., Jiang, L., & Yu, L. (2020). Class-specific attribute value weighting for Naive Bayes. Information Sciences, 508, 260–274. https://doi.org/10.1016/j.ins.2019.08.071

[72] Zhang, L., Jiang, L., Li, C., & Kong, G. (2016). Two feature weighting approaches for naive bayes text classifiers. Knowledge-Based Systems, 100, 137–144.

[73] Kharya, S., & Soni, S. (2016). Weighted Naive Bayes Classifier : A Predictive Model for Breast Cancer Detection. International Journal of Computer Applications, 133(9), 32–37.

[74] Raschka, S. (2018). Model Evaluation , Model Selection , and Algorithm Selection in Machine Learning. Wisconsin–Madison.

[75] Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016). Evaluating Forecasting Methods by Considering Different Accuracy Measures. Procedia Computer Science, 95, 264–271. https://doi.org/10.1016/j.procs.2016.09.332

[76] R Core Team. (2020). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org/

[77] Casas, P. (2019). funModeling: Exploratory Data Analysis and Data Preparation Tool-Box. Retrieved from https://cran.r-project.org/package=funModeling

[78] Wickham, H., François, R., Henry, L., & Müller, K. (2020). dplyr: A Grammar of Data Manipulation. Retrieved from https://cran.r-project.org/package=dplyr

[79] Wickham, H., & Henry, L. (2020). *tidyr: Tidy Messy Data*. R Foundation for Statistical Computing

[80] Kuhn, M. (2020). caret: Classification and Regression Training. Retrieved from https://cran.r-project.org/package=caret

[81] Wickham, H. (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/.

[82] Dua, D., & Graff, C. (2017). UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml]

[83] Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(16), 1–16. https://doi.org/10.1186/s12911-020-1023-5

[84] El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). Feature Analysis of Coronary Artery Heart Disease Data Sets. International Conference on Communication, Management and Information Technology, 65, 459–468. https://doi.org/10.1016/j.procs.2015.09.132

[85] Tougui, I., Jilbab, A., & El Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. Health and Technology. https://doi.org/10.1007/s12553-020-00438-1

[86] Zriqat, I. A., Altamimi, A. M., & Azzeh, M. (2016). A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods. International Journal of Computer Science and Information Security (IJCSIS), 14(12), 868–879. Retrieved from http://arxiv.org/abs/1704.02799

[87] Lee, C. H. (2018). An information-theoretic filter approach for value weighted classification learning in naive bayes. Data & Knowledge Engineering, 113, 116–212.

[88] Yu, L., Jiang, L., Wang, D., & Zhang, L. (2018). Toward naive bayes with attribute value weighting. Neural Computing & Applications.

[89] Zaidi, N. A., Cerquides, J., Carman, M. J., & Webb, G. I. (2013). Alleviating naive bayes attribute independence assumption by attribute weighting. Journal of Machine Learning Research, 14, 1947–1988.

[90] Dinov, I. . (2020). Evaluating Model Performance. Data Science and Predictive Analytics. http://www.socr.umich.edu/people/dinov/courses/DSPA_notes/13_ModelEvaluation.html.

[91] Parikh, R. ., Mathai, A., Parikh, S., Sekhar, C. ., & Thomas, R. . (2008). Understanding and using sensitivity, specificity and predictive values. Indian Journal of Opthamology, 56(1), 45–50

[92] Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. International Conference on Advanced Computing, (6). https://doi.org/10.1109/IACC.2016.25

[93] Wang, Z. (2018). Practical tips for class imbalance in binary classification. https://towardsdatascience.com/practical-tips-for-class-imbalance-in-binary-classification-6ee29bcdb8a7.

JHIA online.org Journal of Health Informatics in Africa

Prevalence and Complications Associated with Diabetes Mellitus at the Nairobi Hospital, Nairobi City County, Kenya

Amos Olwendo [1*], George Otieno[1], Kenneth Rucha[1]

Department of Health Management & Informatics, Kenyatta University, P.O. Box. 43884, 00100, Nairobi, Kenya.

Background and Purpose: Diabetes mellitus (DM) is a lifestyle disease and a global health challenge. About 14.2 million people in Africa had the disease in 2015. Kenya is presently experiencing an increase in mortality and morbidity related to diabetes.
Methods: This research employed a retrospective cross-sectional study design that sampled records of confirmed cases of diabetes mellitus collected during routine care between January 2012 and December 2016 at the Nairobi Hospital located in Nairobi city, Kenya. A stratified sample of 652 records of male and female patients were retrieved from the EHR database and analyzed in this research. The dataset was subjected to pre-processing; that involved handling cases of missing values, smoothing for the removal of noise, identification and removal of outliers, and resolving cases of inconsistencies. Data were normalized using the z-score standardization and analyzed based on dimensions of EHR data quality and through cluster analysis using Density-Based Spatial Clustering of Applications with Noise (DBSCAN).
Results: The prevalence of T2DM is at 92% and the most common complications of diabetes include; retinopathy (12%), neuropathy (11%), and cardiovascular (11%). Hypertension was present in 39% of cases of diabetes.
Conclusion: Diabetes is increasingly becoming a health problem in Kenya thus there is need for increased public awareness of the dangers of diabetes mellitus. Members of the public need to sensitized on the usefulness of physical exercise and dietary requirements to slow the development and progression of diabetes. Also, there is need for understanding the causal relationship between T1DM and T2DM and hypertension.

Keywords: Diabetes Mellitus; Complications; Computational Phenotyping; Density-based Clustering; DBSCAN;

1 Introduction

Diabetes refers to a group of diseases that affect how the body uses glucose. Glucose is vital to health because it is an important source of energy for the cells that make up muscles and tissues. DM is a global health challenge and about 14.2 million people in Africa had the disease in 2015. The number of DM cases are expected to rise in Africa especially in countries transitioning from low to middle income economies such as Kenya. A number of DM cases stay undiagnosed for long in Kenya and the prevalence of DM among Kenyans aged 27-79 was 2.2% (approximately 484,000 persons) in 2015 (Mohamed et al., 2018; Mwangi et al., 2017). The principal cause of diabetes varies by type. Nevertheless, all types of diabetes can lead to excess sugar in the blood. Diabetes mellitus is a leading cause of mortality and morbidity that is characterized by insulin deficiency. Diabetes can be classified as type 1, 2, 3, and 4. Type 1 diabetes mellitus (T1DM), characterized by insulin deficiency and develops at any point during an individual's lifetime, is currently on the rise worldwide. On the other hand, T2DM affects millions of people worldwide. T2DM makes about 90% of all diabetes cases in Kenya and is known to develop in a person at about age 40 (Mwangi et al., 2017). Other forms of diabetes include gestational, chemical-induced. Chronic diabetes conditions include T1DM and T2DM. Potentially reversible diabetes conditions include prediabetes which occurs when the blood sugar levels are higher than normal, but not high enough to be classified as diabetes. Also, gestational diabetes, which occurs during pregnancy but may resolve after the baby is delivered (American Diabetes Association, 2016; Conti et al., 2017; Szendroedi et al., 2016; WHO, 2014).

*Corresponding author address: Email: aolwend@gmail.com, Tel: +254712655407

J Health Inform Afr. 2020;7(2):47-57. DOI: 10.12856/JHIA-2020-v7-i2-290

The rising urbanization has resulted in the change in lifestyles especially with regards to nutrition hence the rising cases of diabetes is associated with the change in the lifestyle. Diabetes is the next epidemic in low income countries owing to the changing lifestyles triggered by a number of factors such as unhealthy diets that encompass consumptions of high calories, and sedentary lifestyles due to socioeconomic development. Sub-Saharan Africa is reported to be experiencing the fastest growing rates of diabetes worldwide. The development and progression of diabetes mellitus is characterized by a number of complications which include cardiovascular diseases such as coronary heart disease with chest pain, heart attack, stroke, and atherosclerosis; neuropathy; nephropathy, retinopathy; skin infections; hearing impairment; Alzheimer's disease; preeclampsia, macrosomia (Conti et al., 2017; Kharono et al., 2017; Mwangi et al., 2017; Yadav et al., 2017).

Nutrition is an important factor influencing the risk of developing T2DM and to some extent T1DM. Excess availability of metabolites such as free fatty acids (sources include dark green leafy vegetables, olive oil, whole grain foods, and eggs) and branched-chain amino acids (sources include chicken, fish, eggs, beans, nuts, and soya) induce whole-body insulin resistance hence minimize the development of diabetes. DM is one of the diseases that require biomarker discovery and translation research to determine the clinical characteristics of their sub-phenotypes right from onset to the manifestation of its complications (American Diabetes Association, 2016; Jones, 2013; Kharono et al., 2017; Szendroedi et al., 2016; Tenenbaum & Avillach, 2016; Yadav et al., 2017).

The long-term complications of diabetes develop gradually and the longer a person lives with diabetes with the blood sugar level less controlled the higher the risk of complications. Diabetes complications may be disabling or even life-threatening. Some of the most common complications of diabetes include: neuropathy, nephropathy, retinopathy, the risk of developing cardiovascular diseases, skin conditions, foot damage, hearing impairment, and depression. Neuropathy is the damage that occurs to the nerve damage. Excess sugar in the blood can wound the walls of the capillaries that nourish the nerves, especially on the legs. This may lead to tingling, numbness, burning or pain that usually begins at the tips of the toes and gradually spreads upward (Daga et al., 2015). On the other hand, nephropathy occurs when diabetes damages the glomeruli, the vessels that filter waste from the blood. Severe damage can lead to kidney failure or irreversible end-stage kidney disease, which may require kidney transplant (Daga et al., 2015). In addition, diabetes can damage the blood vessels of the retina that result in diabetic retinopathy leading to blindness. Diabetes may also increase the risk of other serious vision conditions, such as cataracts and glaucoma (Daga et al., 2015; Mwangi et al., 2017). Moreover, diabetes also increases the risk of various cardiovascular problems, including coronary artery disease with chest pain, heart attack, stroke and narrowing of arteries (atherosclerosis). Persons experiencing cardiovascular complications are more likely to have heart disease or stroke. Left untreated, you could lose all sense of feeling in the affected limbs. Damage to the nerves related to digestion can cause problems with nausea, vomiting, diarrhea or constipation. For men, it may lead to erectile dysfunction. Foot damage occurs when the nerves in the feet are damaged resulting in poor blood flow to the feet. Left untreated, cuts and blisters can develop serious infections, which often heal poorly thus may result into amputation of the affected area. Diabetes may also leave the body susceptible to skin problems and hearing problems. Finally, persons with diabetes may experience depression and increase the risk of dementia, such as Alzheimer's disease (Conti et al., 2017; Richesson et al., 2014; Szendroedi et al., 2016).

1.1 Secondary Utility of EHR Data in Clinical Informatics Research

Precision Medicine (PM) is medical care designed to optimize efficiency and/or therapeutic benefit for a group of patients thus an effort for improve healthcare quality. The primary goal of PM is to uncover disease sub-phenotypes defined by distinct molecular mechanisms that underlie various disease manifestations. However, critical disease subtype distinctions may also be impacted by nonmolecular factors such as socioeconomic status (Tenenbaum & Avillach, 2016).

The field of medicine has undoubtedly grown over the years. One of the many appreciated reasons for the advances in medicine is the applications of health information technologies in various departments within hospitals for purposes such as keeping patient records, medical imaging, and health information exchange. Electronic Health Records (EHRs) are increasingly employed for the management of patient data in primary care worldwide due to the fact there are standards for the design and development of an EHR (van der Bij et al., 2017). This has led to an explosion of electronic patient records collected during routine care. Moreover, historical EHR data may be used for secondary purposes such as in conducting research.

The use of EHR data to conduct retrospective study designs would undoubtedly reduce research costs and promote patient-centered research. However, lack of standards for EHR data and the increasing demand of EHR software worldwide has led to the introduction of software products that record data with questionable quality. Presently, EHR data are characterized by noisy data that comes as a result of erroneous inputs and coding inaccuracies. EHR data are normally inaccurate, redundant, incomplete, and/or irrelevant. Moreover, EHR data may also experience fragmentation in records due to inconsistent patient visits to various healthcare providers without data integration plans. Therefore, clinical data with questionable qualities may not be suitable for secondary uses such as understanding the natural history of disease, cohort identification, biomarker discovery and computational phenotyping just to mention a few (Farrell et al., 2017; Richesson et al., 2014; Tenenbaum & Avillach, 2016; N G Weiskopf & Weng, 2013; Yadav et al., 2017).

1.2 Dimensions for Evaluation of the Quality of EHR Data

The domains of EHR data quality are generally categorized as; conformance, completeness, timeliness, and consistency (Feder, 2017; Kahn et al., 2016; N G Weiskopf & Weng, 2013). However, considering its objectives, this research was limited to assessing chose conformance, completeness, and consistency of the EHR data.

1.3 Conformance

Conformance of data is investigated to ascertain whether its value meets syntactic or structural constraints such as the expected format and data values for each data element. Conformance of EHR data element are categorized as; value conformance, relational conformance, and computational conformance. Value conformance determines whether the data element is a true representation of the expected value. For example, age is expressed using positive integer within an acceptable value range. On the other hand, relational conformance determines whether the data value conforms to relational constraints based on external standards. Finally, computational conformance determines whether the computed data values match validation values defined by external standards (Kahn et al., 2016; Zozus et al., 2014).

1.4 Completeness

Data completeness assesses the presence or absence of data at a single moment over time. Data completeness not only checks the absence of data but also the underlying reason for which such data is missing. Data could be missing due to imputation failure or the personnel failing to document and/or enter such data or due to the patient in question having failed to provide the required data (Kahn et al., 2016; Nicole G Weiskopf et al., 2013; Zozus et al., 2014).

1.5 Consistency

The consistency in data is its constancy with regards to the stipulated data validation rules. That is, the absence of a difference when comparing two or more representations of a data element.

Consistency of data is both affected by the training offered to personnel and the relevancy of the data definitions. Furthermore, data consistency may also be affected by the guidelines and procedures that guide its collection. Data consistency is measured through evaluation of the; procedure for measurement, data measures, and the granularity of the data values. Procedure for measurement evaluates the technique employed for data collection and documentation. On the other hand, data measure evaluates consistency of a variable’s unit of measurement and reference range. Finally, data granularity evaluates the degree of detail of the given data element (Kahn et al., 2016; van Engen-Verheul et al., 2016; Nicole G Weiskopf et al., 2013; Zozus et al., 2014).

1.6 Computational Phenotyping

On the other hand, computational phenotyping is the practice of learning latent relationships from raw data without human intervention. Computational phenotyping is achieved through identification of disease phenotypes and sub-phenotypes from healthcare data for adequate management of patient health. The goal of this research was to assess the quality of EHR and determine the suitability of such data to conduct computational phenotyping of diabetes mellitus. Computational phenotyping tasks include; discovering and stratifying new disease sub-phenotypes and;

discovering specific phenotypes for improving classification under existing disease boundaries and definitions. This research was limited to the development of an unsupervised learning model using the DBSCAN algorithm to explore the model's ability to discover and stratify diabetes mellitus cases to help in improving categorization of cases of diabetes under existing complications (Che & Liu, 2017; Denaxas et al., 2017; Ghosh et al., 2016; Richesson et al., 2014; Tenenbaum & Avillach, 2016; Yadav et al., 2017).

Computational phenotyping is achieved through clustering data which involves grouping of objects such that the objects in a group (cluster) are similar (or related) to one another and different from (or unrelated to) the objects in other groups. DBSCAN is a density-based clustering algorithm which determines important properties about the distribution of the dataset. Thereafter, the algorithm constructs a model and a target function that when supplied with the unencountered dataset x, it's upon the target function to determine the cluster(s) to which each record in the dataset is to be assigned. A cluster is a dense region of points which is separated by low-density regions from other regions of high density. Density-based clustering is best applied in the case that clusters are irregular or intertwined, and when noise and outliers are present in the data (Ester et al., 1996; Richesson et al., 2014; Schubert et al., 2019; Yadav et al., 2017; Zozus et al., 2014).

2 Materials and Methods

A retrospective cross-sectional study design that utilized 652 records of confirmed cases of diabetes mellitus that were collected and stored in the EHR database during routine care at The Nairobi Hospital between January 2012 and December 2016. The Nairobi Hospital is located in Nairobi city and it is one of the leading hospitals in Kenya. It has a client-base comprising of persons from middle and upper social classes. Nairobi hospital was chosen because it is one of the few hospitals that have utilized EHR for a period of no less than five years and also embraces research. The data sample size was determined through a stratified sampling of considering both genders. The dataset was subjected to pre-processing focusing on data entry and typing errors. Moreover, missing values were replaced with the arithmetic mean and inconsistencies in data values were also addressed appropriately. Furthermore, outliers were identified and replaced by the arithmetic mean as well. Finally, data were smoothened for the removal of noise in the data and normalized through z-score standardization method. The quality of the dataset was evaluated based on the dimensions of EHR data quality; conformance, completeness, and consistency. Descriptive statistics were measured using SPSS version 21. On the other hand, cluster analysis was conducted using the DBSCAN algorithm and cluster results verified using the International Classification of Diseases version ten (ICD 10) codes assigned to each data record during diagnosis.

3 Results

3.1 Description of the EHR Dataset

A total of 652 records of confirmed cases of diabetes mellitus were extracted from the EHR database. The attributes of the dataset comprised of; Age, Gender, Body Mass Index (BMI), BSA, Pulse, Systolic, Diastolic, Random Blood Sugar (RBS), SPO2 (Oxygen saturation), Temperature, and Respiration as summarized in Table 1.

Table 1. Descriptive statistics of the attributes of the EHR data set.

Attribute	N	Minimum	Maximum	Mean	Std. Deviation
Age	652	21	82	53.4	11.1
Gender	652	0	1	0.5	0.5
BMI	652	16.5	311.5	30.0	17.3
BSA	652	.70	5.6	1.9	0.2
Pulse	652	55	118	82.7	11.2
Systolic	652	52	203	132.9	18.3

Diastolic	652	45	111	80.9	10.7
RBS	652	2.1	22.0	8.2	3.8
SPO2	652	10	100	97.3	3.9
Temperature	652	20.0	37.2	35.9	0.9
Respiration	652	12	181	18.4	6.6

3.2 Evaluation for quality based on domains of EHR data quality

Both Conformance and Consistency of the EHR data elements was evaluated to beyond average except for values for Temperature and BMI which were out of range. Errors with BMI data were as a result in wrong recordings of body Weight and Height or both. However, data incompleteness (as a result of an unexpected value or no value present at all) was the main challenge to the quality of EHR data. Completeness of EHR data was evaluated at 75% and data attributes with cases of Incompleteness were mainly due to data entry errors. The rest of the details are as summarized in Table 2.

Table 2. A summary of the evaluation for quality based on domains of EHR data quality.

Dimension	**Parameter**	**Frequency N= 652**	**Percentage**
Conformance	Value conformance	646	99%
	Relational conformance	652	100%
	Computational conformance	646	99%
Consistency	Procedure for measurement	652	100%
	Data measure	646	99%
	Data Granularity	642	98%
Completeness	Completeness	489	75%

3.3 Distributions of the types of diabetes in the dataset

The distributions of cases of diabetes mellitus in the dataset based on assigned ICD 10 codes included; gestational (1%), Prediabetes (0.4%), T1DM (30%), and T2DM (92%). on the other hand, co-morbidities identified in the dataset were variations of hypertension (67%). Amongst T2DM cases, 60% had hypertension and only 18% of T2DM cases had T1DM. More of the details are as summarized in Table 3.

Table 3. Distribution of the types of diabetes mellitus in the EHR dataset.

Category	Number of Cases	Percentage (%)
Prediabetes	3	0.4
Gestational diabetes	7	1
Hypertension only	25	4
T1DM only	19	3
T1DM with Hypertension	0	0

T2DM only	171	26
T2DM with Hypertension	255	39
T1DM and T2DM without Hypertension	18	3
T1DM and T2DM with Hypertension	155	24

3.4 Complications of diabetes identified from the dataset based on ICD 10 Codes

Retinopathy (12%) was the leading complication associated with diabetes followed by neuropathy (11%) and cardiovascular complications (11%). Other complications in the data set include kidney damage (nephropathy), foot damage, skin conditions, and depression. However, it is interesting to note that 40% of cases of diabetes were not associated with any complications yet only 0.4% of the EHR dataset were prediabetic as summarized in Table 3 and Table 4.

Table 4. A summary of complications of diabetes mellitus based on ICD 10 codes.

ICD 10 Code and Description	Complication	Percentage
F31-Bipolar affective disorder,	Depression	4%
B35.3 - athletes foot B35.3 - Athlete's foot, also known as tinea pedis M21.6 - deformities of foot	Foot damage	4%
H52.4 - Presbyopia-long-sightedness H52.1 - Nearsightedness (myopia) H52.0 - a condition of the eye H35.0 - retinopathy and retinal…	Retinopathy	12%
N08.3-Glomerular disorders in diabetes mellitus I13 - hypertensive heart and kidney diseases	Nephropathy	3%
L30 - unspecified dermatitis – skin L20 - Atopic dermatitis R23 - other skin changes	Skin conditions	2%
I10-neuropathy G63.2 - Diabetic polyneuropathy	Neuropathy	11%
H45 - Transient cerebral ischaemic attacks I64 – Stroke I65 - Occlusion and stenosis of precerebral arteries, … I50 - Heart failure	Cardiovascular	11%

E78.0 - Pure hypercholesterolaemia	High cholesterol	5%
E03.9 – Hypothyroidism E03 - other Hyperthyroidism	Thyroid	1%
E78 - Disorders of lipoprotein metabolism and other lipidaemias	Lipids	7%
Others – no other ICD 10 code except for the diabetic classes	Unclassified	40%

3.5 Clusters identified from the dataset using DBSCAN algorithm

The DBSCAN algorithm categorized 88% of the EHR dataset as noise and the remaining 12% of the dataset were categorized into 23 clusters. The other details are as summarized in Table 5.

Table 5. Clusters identified from cases of diabetes mellitus using DBSCAN algorithm.

Cluster	ICD 10 Codes	Complication
1	E11 – T2DM E10 – T1DM I10 – Essential Hypertension I75 - Atheroembolism H52.4 - Presbyopia-long-sightedness	Eye damage
2	E11 – T2DM	None
3	E11 – T2DM K30 – Functional dyspepsia H40 – Glaucoma – eye condition	Eye damage
4	E11 – T2DM M75.4 – impingement syndrome of shoulder M21.6 – deformities of foot	Neuropathy Foot damage
5	E11 – T2DM E10 – T1DM I10 – Essential hypertension I79 - Disorders of arteries, arterioles and capillaries…	Cardiovascular disease
6	E11 – T2DM E10 – T1DM I10 – Essential hypertension I86 - Varicose veins of other sites E78 - Disorders of lipoprotein metabolism and other lipidaemias	Cardiovascular disease Lipids

7	E11 – T2DM E10 – T1DM I10 – Essential hypertension	None
8	E10 – T1DM	None
9	E11 – T2DM E10 – T1DM I10 – Essential hypertension I74 – I78 - Diseases of capillaries	Cardiovascular disease
10	E11 – T2DM M54.5 - Low back pain	Neuropathy
11	E11 – T2MD F31 - Bipolar affective disorder B35.3 - Athlete's foot, also known as tinea pedis	Depression Foot damage
12	E11 – T2DM I10 – Essential hypertension E78 - Disorders of lipoprotein metabolism and other lipidaemias	Lipids
13	E11 – T2DM I10 – Essential hypertension E78 - Disorders of lipoprotein metabolism and other lipidaemias E66 – Obesity	Obesity Lipids
14	E11 – T2DM E10 – T1DM I10 – Essential hypertension F31- Bipolar affective disorder M13.9 – arthritis – unspecified	Depression Neuropathy
15	E11 – T2DM I10 – Essential hypertension M47 – Spondylosis- Spondylosis-arthritis to the spine	Neuropathy
16	E11 – T2DM	None
17	E11 – T2DM	None
18	E11- T2DM E10 – T1DM H52.4 – Presbyopia	Eye damage
19	E11 – T2MD E10 – T1DM I10 – Essential hypertension	None

20	E11 – T2DM	None
21	E11 – T2DM	None
22	E11 – T2DM E10 – T1DM I10 – Essential hypertension	None
23	E11 – T2DM E10 – T1DM I10 – Essential hypertension	None

4 Discussion

The prevalence of prediabetes (0.4%) in the EHR dataset was quite unbelievable (Mohamed et al., 2018). As a matter of fact, 40% of the cases of diabetes were reported unclassified (based on the assigned ICD 10 codes) meaning that they had not developed into any of the complications of diabetes. Moreover, cases of Prediabetes are believed that usually develop into T2DM thus prevalence of T2DM at 92% was not matching with the prevalence of Prediabetes at 0.4% (Chung et al., 2020). Therefore, the task of diagnosis and assignment of the associated ICD 10 codes must have been compromised. However, effective management of diabetes requires early diagnosis of the disease. Therefore, the few cases of prediabetes show that the actual statistics of persons living with prediabetes maybe incorrect hence the need for urgent measures for the identification of all possible cases of prediabetes cases among the general population.

Moreover, it was interesting to observe that there were no identified cases with both hypertension and T1DM. On the other hand, hypertension was observed in 60% of cases of T2DM. Also, 70% of the cases of T2DM had T1DM and hypertension as co-morbidities. The presence of T1DM, which has genetic predispositions and usually develops at any time right from birth, is increasing among patients that had never until the time of the diagnosis of T2DM. This calls for the need for analysis of the causal relationship between T1DM and T2DM and the causal relationship T2DM and hypertension.

Moreover, the complications arising from cases of diabetes mellitus such as retinopathy; which causes eye damage that results into blindness and complications like neuropathy, and cardiovascular challenges need to be identified early in time since the aftermaths of such complications would not only burden the healthcare system but also lead to a rise in mortality and also affect the productivity of the general population. As a matter of urgency, the government of Kenya should come up with the measures for targeted case finding for not only undiagnosed cases of diabetes but also cases that would develop into retinopathy and cardiovascular complications.

The clustering algorithm determined that 88% of the records had noise hence only 12% of the dataset were utilized in the construction of the target function for the model. This could have been as a result of incompleteness that was evaluated at 25% and the inconsistencies and lack of conformance recorded for data attributes such as Temperature and BMI. Height, Weight, and Temperature are measured values and the challenges experienced with them must have come from the human resource doing the measurements. However, BMI calculations must have been questionable since it is calculated from both Weight and Height whose measurements had already been determined to be questionable.

Moreover, the clustering algorithm seems to agree with the results in Table 3 and takes it further to show the connectedness among the various cases of diabetes. Results in Table 4 show the traditional view of the categories of diabetes mellitus. However, the algorithm showed that not only are the types of diabetes related but also that the complications of diabetes mellitus are intertwined. Despite the wanting quality of the EHR data, clusters identified from 12% of the dataset disclose that computational phenotyping would be realizable from EHR data.

In Table 5, Clusters 7, 19, 22, and 23 show similarities in the disease and associated complications based on the ICD 10 codes. However, the algorithm determined that these clusters are significantly distinct hence form distinct sub-phenotyping groups. Similarly, clusters 2, 16, 17, and 20, based on the characteristics of the dataset look similar. However, the algorithm identified each of them as distinct. This means there is an underlying difference in the data that is not obvious thus only known to the learning algorithm. clusters; 1, 3, and 18 are distinct sub-groups of retinopathy. On the hand, clusters; 5, 6, and 9 are distinct sub-groups of cardiovascular complications. Moreover, clusters; 12 and 13 are distinct sub-groups of lipids. This shows that the task of computational phenotyping from routine healthcare data is achievable since as the algorithm has not only identified the phenotypes but also the sub-types of the given phenotypes. As a result, clinical decision support systems could easily be developed from such backgrounds to assist physicians in the delivery of healthcare services. Moreover, minimal effort to reduce the amount of "noise" from EHR data would probably guarantee much better results.

5 Conclusion

Diabetes mellitus is increasingly infiltrating the health and wellbeing of the Kenyan population hence the need for urgent action by the government to come up with measures for the control of development of the disease. This could be achieved through the combined efforts from both the government and the general population. The Ministry of Health (MoH) should not only increase awareness for diabetes but also conduct targeted case finding exercises for early diagnosis and management of diabetes. According to results in Table 3, cases of prediabetes appear negligible. However, this may not be the true picture given that 40% cases of diabetes were unclassified meaning the diagnosis of associated cases of diabetes were not conclusive.

On the other hand, the MoH should instill measures to ensure all food items undergo clearance by the Kenya Bureau of Standards before they are made available for human consumption as an effort to control their cholesterol contents. Moreover, the MoH should as well increase their efforts of educating the general population on the benefits of eating balanced diets and better nutrition at large.

Funding

No funding was acquired to support this research.

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethical Approval

The proposal was submitted for ethical approval from Kenyatta University Ethical and Research Committee. Thereafter, the study sought a permit to conduct the research from National Commission for Science, Technology and Innovation (NACOSTI). Moreover, the study sought relevant permissions and approvals from The Nairobi Hospital administration. Finally, consent of participants was sought before their involvement in this research.

References

American Diabetes Association, G. of C. N. R. C. (2016). Standards of Medical Care in Diabetes-2016 Abridged for Primary Care Providers. *Clinical Diabetes : A Publication of the American Diabetes Association*, *34*(1), 3–21. https://doi.org/10.2337/diaclin.34.1.3

Che, Z., & Liu, Y. (2017). Deep Learning Solutions to Computational Phenotyping in Health Care. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 1100–1109. https://doi.org/10.1109/ICDMW.2017.156

Chung, W., Erion, K., Florez, J., Hattersley, A., Hivert, M.-F., Lee, C., McCarthy, M., Nolan, J., Norris, J., Pearson, E., Philipson, L., McElvaine, A., Cefalu, W., Rich, S., & Franks, P. (2020). Precision Medicine in Diabetes: A Consensus Report From the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care*, *43*, 1617–1635. https://doi.org/10.2337/dci20-0022

Conti, C., Mennitto, C., Di Francesco, G., Fraticelli, F., Vitacolonna, E., & Fulcheri, M. (2017). Clinical Characteristics of Diabetes Mellitus and Suicide Risk. *Article Mini Review*, *8*(40). https://doi.org/10.3389/fpsyt.2017.00040

Daga, R. A., Naik, S. A., Maqbool, M., Laway, B. A., Shakir, M., & Rafiq, W. (2015). Demographic and Clinical Characteristics of Diabetes Mellitus among Youth Kashmir, India. *Int J Pediatr*, *3*(19), 4–1. http://

Denaxas, S., Direk, K., Gonzalez-Izquierdo, A., Pikoula, M., Cakiroglu, A., Moore, J., Hemingway, H., & Smeeth, L. (2017). Methods for enhancing the reproducibility of biomedical research findings using electronic health records. *BioData Mining*, *10*(1), 31. https://doi.org/10.1186/s13040-017-0151-7

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*. 226--231. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220

Farrell, L. J., Shimeng, D., Steege, L. M., Cartmill, R. S., Wiegmann, D. A., & Wetterneck, T. B. (2017). Understanding Cognitive Requirements for EHR Design for Primary Care Teams. *Proceedings of the 2017 International Symposium on Human Factors and Ergonomics in Health Care The*. https://doi.org/10.1177/2327857917061005

Feder, S. L. (2017). Data Quality in Electronic Health Records Research : Quality Domains and Assessment Methods. *Western Journal of Nursing Research*, 1–14. https://doi.org/10.1177/0193945916689084

Ghosh, S., Cheng, Y., & Sun, Z. (2016). Deep State Space Models for Computational Phenotyping. *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 399–402. https://doi.org/10.1109/ICHI.2016.71

Jones, T. L. E. (2013). Diabetes Mellitus: the increasing burden of disease in Kenya. *South Sudan Medical Journal*, *6*(3). http://www.southsudanmedicaljournal.com/assets/files/Journals/vol_6_iss_3_aug_13/Diabetes_in_Kenya.pdf

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S.-T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N., & Schilling, L. (2016). A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Washington, DC)*, *4*(1), 1244. https://doi.org/10.13063/2327-9214.1244

Kharono, B., Nabisere, R., Kiddu Persis, N., Nakakeeto, J., Openy, A., & Bakeera Kitaka, S. (2017). Knowledge, Attitudes, and Perceived Risks Related to Diabetes Mellitus Among University Students in Uganda: A Cross-Sectional Study. *The East African Health Research Journal*, *1*(2), 105–112. http://www.ncbi.nlm.nih.gov/pubmed/29250612

Mohamed, S. F., Mwangi, M., Mutua, M. K., Kibachio, J., Hussein, A., Ndegwa, Z., Owondo, S., Asiki, G., & Kyobutungi, C. (2018). Prevalence and factors associated with pre-diabetes and diabetes mellitus in Kenya: Results from a national survey. *BMC Public Health*, *18*(S3), 1215. https://doi.org/10.1186/s12889-018-6053-x

Mwangi, N., Macleod, D., Gichuhi, S., Muthami, L., Moorman, C., Bascaran, C., & Foster, A. (2017). Predictors of uptake of eye examination in people living with diabetes mellitus in three counties of Kenya. *Tropical Medicine and Health*. https://doi.org/10.1186/s41182-017-0080-7

Richesson, R. L., Horvath, M. M., & Rusincovitch, S. A. (2014). Clinical Research Informatics and Electronic Health Record Data. *IMIA and Schattauer GmbH*, 215–223.

Schubert, E., Hess, S., & Morik, K. (2019). *The Relationship of DBSCAN to Matrix Factorization and Spectral Clustering*. http://ceur-ws.org/Vol-2191/paper38.pdf

Szendroedi, J., Saxena, A., Weber, K. S., Strassburger, K., Herder, C., Burkart, V., Nowotny, B., Icks, A., Kuss, O., Ziegler, D., Al-Hasani, H., Müssig, K., Roden, M., & GDS Group. (2016). Cohort profile: the German Diabetes Study (GDS). *Cardiovascular Diabetology*, *15*(1), 59. https://doi.org/10.1186/s12933-016-0374-9

Tenenbaum, J., & Avillach, P. (2016). An informatics research agenda to support precision medicine: seven key areas. *Journal of The*. http://jamia.oxfordjournals.org/content/23/4/791.abstract

van der Bij, S., Khan, N., ten Veen, P., de Bakker, D. H., & Verheij, R. A. (2017). Improving the quality of EHR recording in primary care: a data quality feedback tool. *Journal of the American Medical Informatics Association*, *24*(1), 81–87. https://doi.org/10.1093/jamia/ocw054

van Engen-Verheul, M. M., Peute, L. W. P., de Keizer, N. F., Peek, N., & Jaspers, M. W. M. (2016). Optimizing the user interface of a data entry module for an electronic patient record for cardiac rehabilitation: A mixed method usability approach. *International Journal of Medical Informatics*, *87*, 15–26. https://doi.org/10.1016/J.IJMEDINF.2015.12.007

Weiskopf, N G, & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment : enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 144–151. https://doi.org/10.1136/amiajnl-2011-000681

Weiskopf, Nicole G, Hripcsak, G., Swaminathan, S., & Weng, C. (2013). Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*, *46*(5), 830–836. https://doi.org/10.1016/j.jbi.2013.06.010

WHO. (2014). WHO | Kenya faces rising burden of diabetes. *WHO*. http://www.who.int/features/2014/kenya-rising-diabetes/en/

Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2017). *Mining Electronic Health Records: A Survey*. http://arxiv.org/abs/1702.03222

Zozus, M. N., Hammond, W. E., Green, B. B., Kahn, M. G., Richesson, R. L., Rusincovitch, S. A., Simon, G. E., & Smerek, M. M. (2014). *Assessing Data Quality*. https://www.nihcollaboratory.org/Products/Assessing-data-quality_V1 0.pdf

Journal of Health Informatics in Africa

ISSN: 2197-6902
DOI: http://dx.doi.org/10.12856/JHIA-2020-v7-i2

Publisher
Koegni eHealth, Innovation for Development e.V. Germany
Postfach 652166
D-22372 Hamburg, Germany
www.koegni-ehealth.org
E-mail: info@koegni-ehealth.org

HELINA is the Africa Region of the International Medical Informatics Association (IMIA)
www.helina-online.org
www.imia-medinfo.org
www.jhia-online.org