Scientific Research Publishing

Journal of Software Engineering and Applications, 2022, 15, 275-307
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

A Decision Support Model for Predicting Avoidable Re-Hospitalization of Breast Cancer Patients in Kenyatta National Hospital

Christopher Oyuech Otieno, Oboko Robert Obwocha, Andrew Mwaura Kahonge

Department of Computer Science, University of Nairobi, Nairobi, Kenya
Email: christopheroyuech1@gmail.com

How to cite this paper: Otieno, C.O., Obwocha, O.R. and Kahonge, A.M. (2022) A Decision Support Model for Predicting Avoidable Re-Hospitalization of Breast Cancer Patients in Kenyatta National Hospital. *Journal of Software Engineering and Applications*, **15**, 275-307.
https://doi.org/10.4236/jsea.2022.158017

Received: June 25, 2022
Accepted: July 26, 2022
Published: July 29, 2022

Abstract

This study aimed to develop a clinical Decision Support Model (DSM) which is software that provides physicians and other healthcare stakeholders with patient-specific assessments and recommendation in aiding clinical decision-making while discharging Breast cancer patient since the diagnostics and discharge problem is often overwhelming for a clinician to process at the point of care or in urgent situations. The model incorporates Breast cancer patient-specific data that are well-structured having been attained from a prestudy's administered questionnaires and current evidence-based guidelines. Obtained dataset of the prestudy's questionnaires is processed via data mining techniques to generate an optimal clinical decision tree classifier model which serves physicians in enhancing their decision-making process while discharging a breast cancer patient on basic cognitive processes involved in medical thinking hence new, better-formed, and superior outcomes. The model also improves the quality of assessments by constructing predictive discharging models from code attributes enabling timely detection of deterioration in the quality of health of a breast cancer patient upon discharge. The outcome of implementing this study is a decision support model that bridges the gap occasioned by less informed clinical Breast cancer discharge that is based merely on experts' opinions which is insufficiently reinforced for better treatment outcomes. The reinforced discharge decision for better treatment outcomes is through timely deployment of the decision support model to work hand in hand with the expertise in deriving an integrative discharge decision and has been an agreed strategy to eliminate the foreseeable deteriorating quality of health for a discharged breast cancer patients and surging rates of mortality blamed on mistrusted discharge decisions. In this paper, we will discuss breast cancer clinical knowledge, data mining techniques, the classifying model accuracy, and the Python web-based decision support model that predicts avoidable re-hospitalization of a breast cancer patient through an informed clinical discharging support model.

Keywords

Re-Engineering Processes (RP), Data Mining, Machine Learning, Classification, Decision Tree, Python Web-Based, Decision Support Model (DSM), Clinical Decision Support Systems (CDSSs)

1. Introduction

Generally, avoidable re-hospitalizations result from care failures in the period immediately, before, or after a transition from hospital to the next source of care [1]. These care failures result in clinical deterioration that leads to subsequent hospital utilization, known as re-hospitalization. **Figure 1** below is a structured model illustrating cycles for a Breast cancer patient at any given time that has been obtained after reengineering unstructured pathways.

Dr. Alice Musibi, Medical Oncologist (2008) argues that Breast cancer is the deadliest [2], and most common cancer ailing women all over the world. The United State of America Institute of Medicine has estimated that up to 98,000 Americans die each year as a result of preventable medical errors for example due to mistrusted discharge decisions [3]. In Australia 1 in 13 women will develop breast cancer at some time in her life, while in USA 215,990 women were found to have breast cancer in 2004.

A cancer survey conducted in Nairobi Kenya between the years 2000-2003 showed that breast cancer was leading with 22.9% followed by cervical cancer with 19.3%.

Breast cancer is the most common cancer affecting women in Kenya whose healthcare costs impose a burden on the government while the quality of care provided is arguably inadequate.

Dr. C. Nyogesa-Watt (2007) reported that there are few public and private hospitals in Kenya providing radiotherapy services with only 10 ecologists and as such, patients have to travel across the country some as far as 600 kilometers away to access such scarce medical services [4].

Dr. Ian Hampson, from The University of Manchester's Institute of Cancer Sciences, oversaw cancer research in Kenyatta National Hospital (KNH) and noted that available radiotherapy centers are very few and exclaimed that breast cancer patients referred from other periphery hospitals are either being re-hospitalized or they're new cases that have to wait for months before accessing medical services as radiotherapy centers sometimes leading to preventable death [5]. One of the most promising strategies for addressing the re-hospitalized crisis is the use of clinical decision support systems (CDSSs), which are systems that provide physicians and other healthcare stakeholders with patient-specific assessments or recommendations to aid in clinical decision-making.

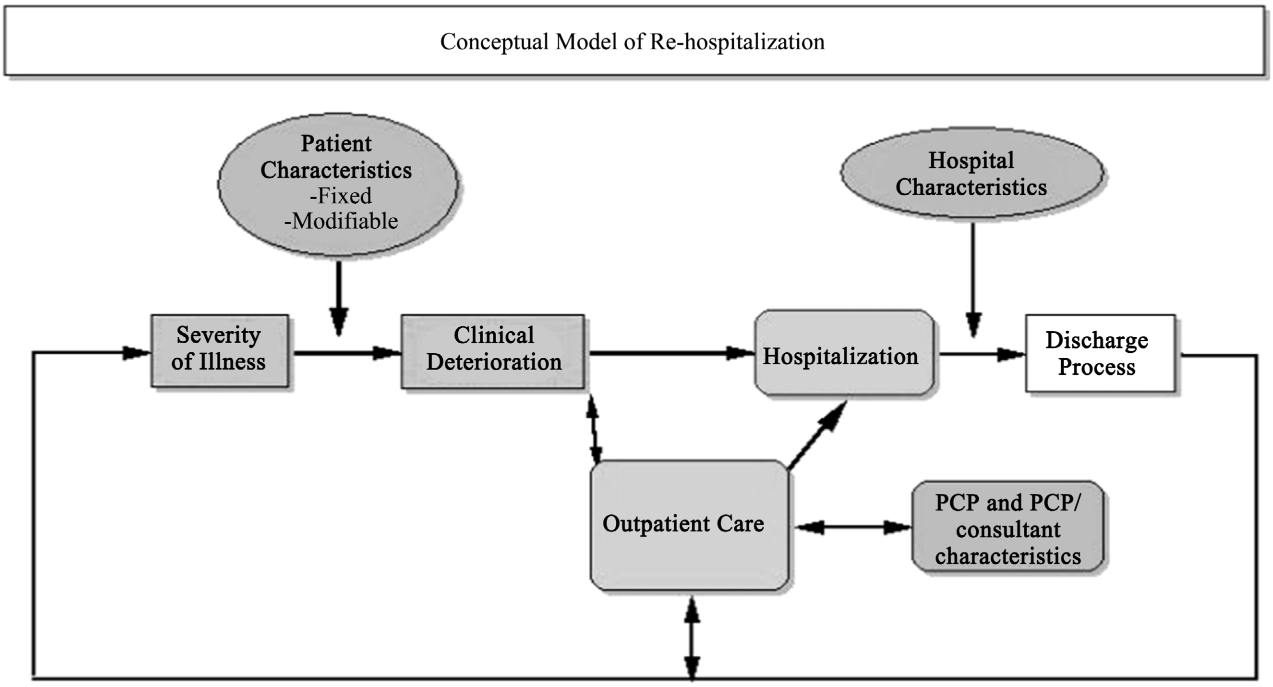

Conceptual Model of Re-hospitalization
Patient Characteristics -Fixed -Modifiable
Hospital Characteristics
Severity of Illness
Clinical Deterioration
Hospitalization
Discharge Process
Outpatient Care
PCP and PCP/ consultant characteristics

Figure 1. Re-engineering the hospital discharge (David Anthony, VK Chetty, *et al.*, 2013).

CDSS/DMS are computerized physician order entry systems that provide patient-specific recommendations as part of the order entry process and alert physicians when critical lab values are detected [6]. CDSSs and DMS are used in the study interchangeably to mean the same thing.

A DSM can be conceptually understood as the custodian of one or more modules of medical knowledge, wherein each DSM knowledge module is capable of utilizing coded patient data to arrive at machine-interpretable conclusions regarding the patient risk attributes under evaluation. In this study, the scope module of the DSM knowledge is on the assessment of a specific breast cancer patient on avoidable re-hospitalization. The knowledge area is made narrow for instance, if a breast cancer patient to be discharged doesn't complete medical dose timely or if a breast cancer patient to be discharged takes unhealthy fat, or if a breast cancer patient to be discharged abuse antibiotic and or if a breast cancer patient to be discharged is a committed smoker and doesn't take physical exercises then this breast cancer patient maybe be re-hospitalized and otherwise if contrary.

There is an urgent need for such CDSS/DMS in the Cancer and Oncology Department, Kenyatta National Hospital (KNH) to classify patients been discharged into their different preventive risk levels based on their vulnerability for re-hospitalization. Identification of likely risk factors predisposing breast cancer patients to avoidable re-hospitalization before discharging is considered a breakthrough in curbing the prevalence of breast cancer menace as this may inform clinicians upfront of avoidable risks categories that a particular patient would assume upon discharge to strategies clinical management targeting the most vulnerable ones as argued by Jaimie Oh, (2012) that, "medicare advantaged patients experience fewer re-hospitalizations."

By implementing this study, we bridge the existing gap in breast cancer discharging and re-hospitalization currently witnessed in Kenyatta National Hospital occasioned by less informed clinical Breast cancer discharge that is merely based on experts' opinions which is insufficiently reinforced for better treatment value outcomes. The rein-forced discharge decision achieved by this study provides better treatment outcomes through a timely decision to work hand in hand with the expertise in deriving an integrative discharge decision having been agreed upon in the medical circles as a strategy that is likely to eliminate the fore-seeable deterioration quality of health for a discharged breast cancer patients thus surging rates of mortality blamed on mistrusted discharge decisions. Reviewed literature didn't reveal existing studies conducted in Kenyatta National Hospital (KNH) in the context of breast cancer discharging and re-hospitalization. This is another compelling force relevant for undertaking this study. This study is also significant as it may be among the transformative practices desired for value-based treatment and management of breast cancer besides. This study is also contributing to theoretical knowledge for discharging and re-hospitalization of Breast cancer patients. The study access relevance of the decision made by Breast cancer experts to discharge a patient thus remains true that if we can predict success based on a certain explanation (*i.e.* C4.5 model or ID3 model), then we have a good reason, and perhaps the best sort of reason, for accepting the explanation. This model is, therefore, a useful tool for assessing the distance between theory (Statistical model) and practice especially when headed to infinity hence assessing the predictive power of a theory to sheds light on the actual performance of an empirical model. The model (*i.e.* C4.5 model or ID3 model), can therefore be used to assess the practical relevance of a theory, (Keil *et al.*). The study through its models (*i.e.* C4.5 model or ID3 model), can also be used to improve existing models since it captures complex underlying patterns and relationships, thereby improving existing explanatory statistical models (Collopy *et al.*).

The remaining part of this paper is organized as follows: Section 2 presents the related work on Breast cancer Re-hospitalization upon discharge. Section 3 describes data mining techniques for modeling the Re-hospitalization prediction classifier while Section 4 presents the conclusion and future work.

2. The Related Work

A joint project of 28 strategic health authorities worked on a risk prediction system that was to be used by PCTs (Primary Care Trusts) to identify patients who are at high risk of hospitalization and was jointly implemented with Essex Strategic Health Authority. In this system, tools such as Ambulatory Cost Groups (ACGs), Diagnostic Cost Groups (DCG), and Hierarchical Coexisting Conditions (HCC) were integrated. ACGs adopted the ICD9-CM coding system based that is based on the assumption that a patient's illness burden better characterizes the patient's need for health services than only the presence of a specific disease such as Breast cancer. The Diagnostic Cost Groups (DCG) tool was also integrated to predict future costs of Medicare for a population-based on the "worst" inpatient diagnosis recorded at a time. DCG/HCC was for predicting total medical expenditure which was an essential feature of the model, (Rosen AK, 2001). The predictive regression model was the main analytic tool in this joint project.

Johns Hopkins University (Bloomberg School of Public Health, 2009) also developed the Adjusted Clinical Group (ACG) model which was based on an aggregation of comorbidities diagnosis as its major methodology. ACG identified patient groups and a population that had a high probability of hospitalization in the future from the aggregation of comorbidities. The model presented the morbidity burden of a population, subgroups, and patients and thus could predict resource use or cost of health care. This model also supported the detection of people with specified diseases, such as HIV. ACG was therefore qualified as a good resource management tool [7].

Health Dialog Analytic Solutions similarly developed Patients at risk of Re-hospitalization (PARR1 and PARR2) algorithms. The algorithms are patient-specific that produce a "risk score" for the probability of future readmission from a patient's past readmission records. The algorithm was used in "real-time" (while the patient is hospitalized) with recent readmission records and diagnostic information. PARR1 and PARR2 algorithms indicated high readmission rates for patients who had experienced readmission before and less for those who have never. In this model, there was no general database to draw inferences from except for specific patient records. The model had shortcomings, for example, it couldn't comprehensively define the risk of readmission that could be assumed by a patient, Schoenmaker & Russo, (1993). It also underestimated the total number of high-risk patients, as it screens patients using a single criterion which may neglect other potentially important risk factors. The model also lacked needed accuracy, for example, individuals who were at risk one year ago may not be at risk in the following next year and vice versa, Dove, Duncan & Robb (2003) [8].

Besides, the Centre for Innovation in Health Management (CIHM) based at the University of Leeds a consulting company gathering expert information from the health sector, public sector, organizational change consultancy, and academics was engaged in a project to develop a model to predict readmission of a discharged patient from a regression model. The project was being developed through decision tree technologies and was to be used by the general practitioners to decide on a patient-oriented intervention while taking into account the clinical knowledge and outputs generated from the Risk classification tool [9]. Additionally, the risk classification tool which is using a regression model whose aim was to stratify patients' risk in terms of their future re-hospitalization. It was therefore meant to act as an intervention by being responsive to the patient's risk categories. Predictive modeling is another tool of risk stratification. General practitioners, nurses, and pharmacists participate in the project modeling as their clinical knowledge is a substantial factor. A predictive model is designed as a statistical model whose output is a risk score for each patient, which is the probability of re-hospitalization in the future [10].

Finally, James Natale and Shengyong Wang of the University of Akron USA, (2013) also develop a model for predicting readmissions of Heart failure via decision tree and Rapid miner as the primary software for the model [11]. A confusion matrix was used in the analysis of specificity and sensitivity. The shortcomings of James Natale and Shengyong Wang's predictive model were that it couldn't focus on the process mapping of the discharging structure. Also, this model predictability was not comprehensive having not factored in avoidable re-admissibility risks such as adverse events, slips, and risk of neglect for its discharging structure processes.

3. Methodology

3.1. The Pre-Study

This is the initial approach adopted in implementing this study. In the pre-study strategy, a retrospective review of clinical non-invasive Breast cancer risks that are likely to cause re-hospitalization of discharged breast cancer patients was identified through questionnaires designed on the Likert scale. These questionnaires were then administered to stakeholders who were the oncology experts, nurse oncologists, clinical oncologists, radiologists, and Breast cancer pathologists. Acquired non-invasive Breast cancer risk variables that were likely to predispose discharged Breast cancer patients to the deteriorating quality of health thus surging rates of mortality were processed to remove outliers and obvious errors in the data set.

3.2. Re-Engineering Current Breast Cancer Discharging Processes in KNH

The re-engineering of discharging Breast cancer processes in KNH was undertaken to provide a conceptual understanding of the causes of preventable breast cancer risk variables to inform safety design concepts aimed at preventing and minimizing them by detecting them upfront before harm occurs. **Figure 2** below highlights fundamental causes of Breast cancer re-hospitalization according to Forster *et al.* (2003) which were adopted to guide the engineering processes of the Kenyatta National Hospital (KNH) discharging framework. Medical errors are an important concern at hospital discharge. Forster and his colleagues identified four areas for improvement before discharging such as assessment and communication of problems that remain unresolved at the time of discharge; patient education regarding medications and other therapies; monitoring of drug therapies after discharge; and monitoring of the overall condition after discharge. Many adverse effects occur during the discharge period which could

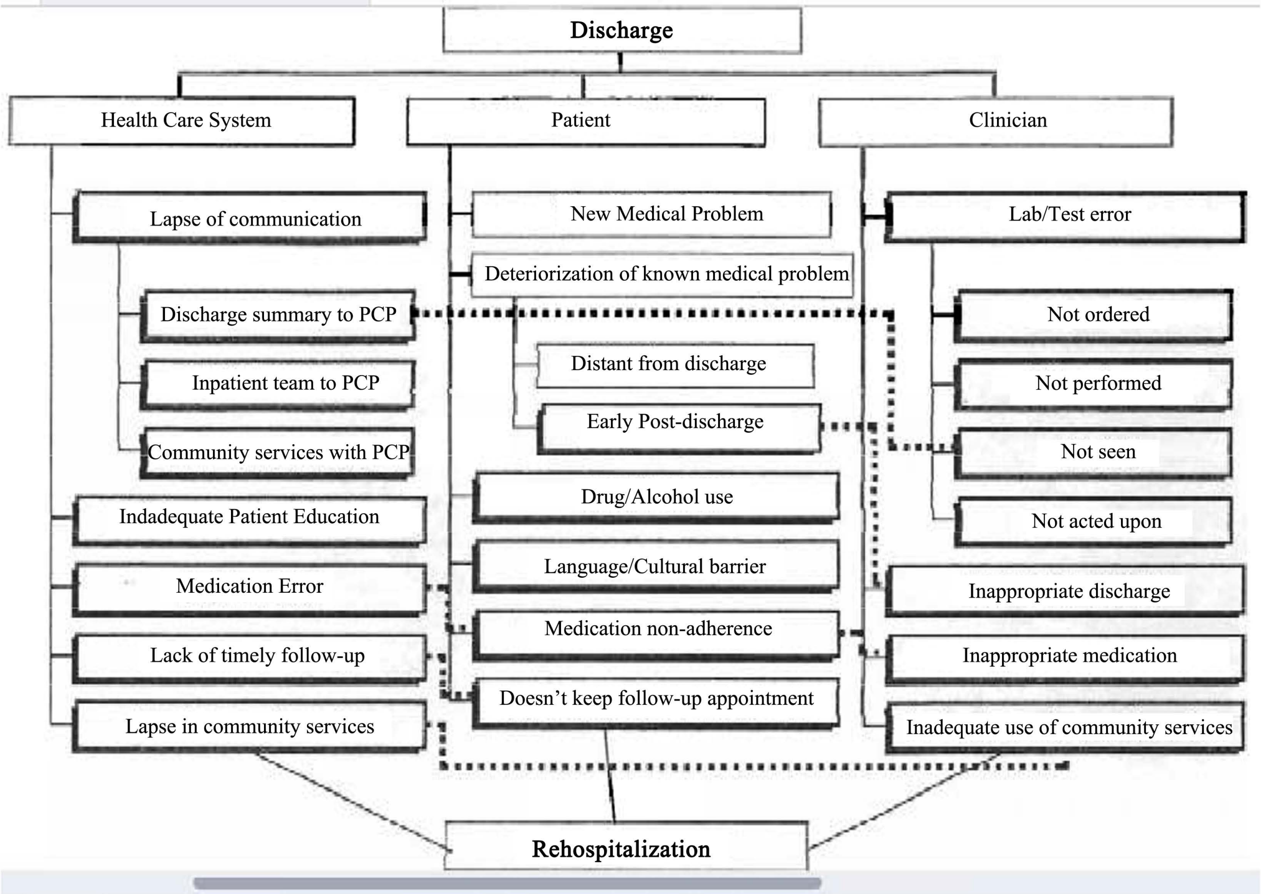


Figure 2. Boxes indicating Breast cancer risks potentially preventable with an intervention. While not detailed here, each type of risk can be further specified, (Forster *et al.*, 2003).

be prevented with relatively simple strategies. Consequently, the re-engineering of the discharging Breast Cancer framework in Kenyatta National Hospital (KNH) considered both active and latent risks occurring at the time of hospital discharge.

Active re-hospitalization risks include those occurring at the time of hospital discharge during knowledge-based decision-making performed at the point of care by Clinicians. Active re-hospitalization risks are hospital characteristics related as shown in the conceptual model (Figure 2) above. Latent risks are observed when there is a system failure. Latent conditions are also clinicians and patients related as shown similarly in the Breast cancer-associated risk in the discharging Model (Figure 2). An example of a latent risk clinician related is when nurses and students are responsible for the discharge process and the harried nature of their work, and competing interests for example new admissions requiring their attention while at the same time, discharging a patient thus may not be considered by them a high priority and can therefore lead to an incomplete or haphazardly discharging resulting to re-hospitalization [12]. Another example of latent risk, is patient-related, for example, their lifestyle and non-compliance to the discharging guidelines or regime for better healing.

Thus taxonomy of avoidable Breast Cancer risk at the point of discharge in Figure 2, demonstrates how latent and active risks inter-relate, and their importance in the rule-based decision-making that the proposed decision support model will action.

With the re-engineered breast cancer discharging processes based on the concept shown in Figure 2 above, tend to promise value-based discharging outcomes needed in value-based healthcare in Breast Cancer as hospital discharge is the main stage that most re-hospitalization risks, lapses, and the adverse event happens. Also to note is that at the point of discharge is when latent conditions (system failures) combine with active failures pointing to the fact that the patient may be discharged with a huge health burden thus guaranteeing re-hospitalization within 30 - 45 days of discharge. As such, Breast cancer discharging processes in Kenyatta National Hospital may be improved through Business Re-engineering Processes (BRP) followed by robust modeling of its best practice processes.

3.3. Improved Breast Cancer Discharging via Business Re-Engineering Processes (BRP) Designed on Best Practices to an Optimal Decision Support Model

Figure 3 displays the Breast Cancer risk elements that pertain to the decision-making of a Breast Cancer patient's readiness for an improved discharge [13]. An optimal decision support model to discharge Breast cancer was coined

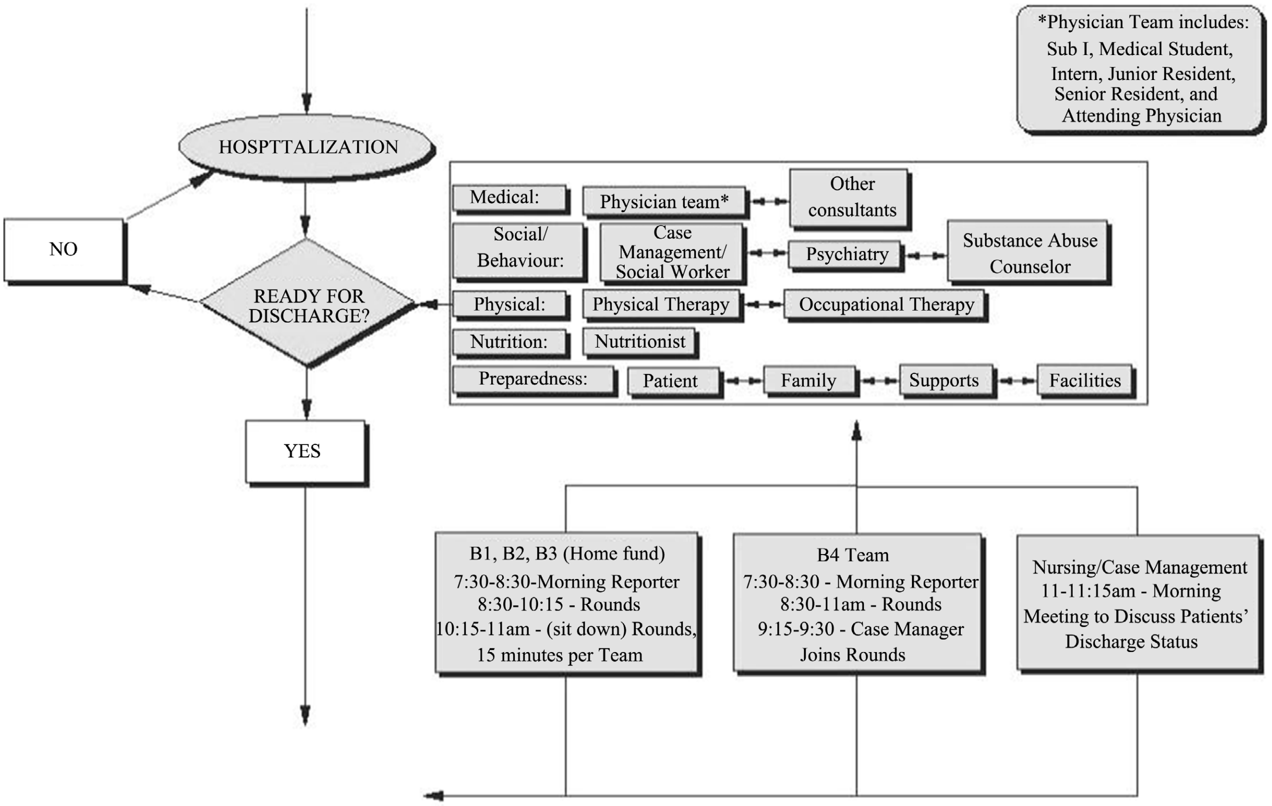


Figure 3. An improved discharging system, David Anthony, VK Chetty, *et al.*, (2013).

around this decision-making structure that required that Breast Cancer risk variables evaluate to a boolean expression, for example, that re-hospitalization shall take place within 30 - 45 days of discharge and or otherwise based on sufficient medical risk evidence of the patients. The decision-making structures that evaluate Breast Cancer risk variables to a boolean expression outcome are *the if… else statement*, *the if… else if… else Statement*, and *the nested if statements* inbuilt in Python programming language.

3.4. The Data Source

A research proposal of the entire study was submitted to Kenyatta National Hospital and the University of Nairobi Ethical Research Committee (KNH/UON-ERC) for review, suggestions, and approval upon satisfaction. The approved proposed study thus authorized access to the data. The source of data for this study was obtained therefore via a pre-study that had to be approved by the KNH/UON-ERC. The prestudy that obtained the dataset was conducted through a retrospective review of clinical non-invasive Breast cancer data risks that are likely to cause re-hospitalization and mistrusted discharges. The prestudy designed questionnaires on the Likert scale. These questionnaires were then administered to stakeholders who were the oncology experts, nurse oncologists, clinical oncologists, radiologists, and Breast cancer pathologists. Acquired data were non-invasive that were likely to predispose discharged Breast cancer patients to a deteriorating quality of health and surging rates of mortality. These data sets were processed to remove outliers and obvious errors. Also, additional first-hand data were obtained from Breast Cancer Past Records which was structured and unstructured and available in files.

3.5. Relevance of Obtained Dataset to Modeling the Clinical Decision Support for Breast Cancer Discharging and Re-Hospitalization Problem

The identified data sets instances can comfortably be represented by attribute-value pairs. For example, in smoking (committed, sneaking, or no smoking), in isolation (psychological, physical, or no Isolation), and data attributes (alcoholism, abuse of drugs, promiscuous) that have discrete output values (yes/no) therefore befitting this clinical decision problem well [14].

3.6. The Research Design

As the study is largely exploratory, the following designs are used;

1) Survey of literature concern, 2) Experience or clinician survey.

The survey of relevant literature e.g. hypothesis specified by earlier researchers is vigorously reviewed and evaluated to form the ground for the design study. Experience survey involves surveying of clinicians on practical experience of problem been studied e.g. surveys of the experienced clinician(s) who are selected randomly. The selected clinicians are given Likert scale questionnaires to respond to on how the design study should be approached. They are also allowed to raise additional issues which the investigator might have not considered within the questionnaires.

3.7. The Study Population

The targeted population included all breast cancer patients re-hospitalized within a period of three months in the years 2007-2013. It's worth noting also that 10,000 breast cancer patients were estimated to have been re-hospitalized upon discharge between these years which translates to 2000 patients yearly on average.

3.8. The Inclusion Criteria for a Patient to Participate in the Study

- A participant should be a patient who has been diagnosed with breast cancer and has experienced re-hospitalized within 30 - 45 days upon discharge.
- A participant may also be a person who survived Breast cancer incidence and experienced re-hospitalization within 30 - 45 days upon discharge.
- A participant should also be a person who met any of the above criteria and has also consented by filling out a consent form to participate in the study.

3.9. The Inclusion Criteria for Clinical Oncologist, Oncology Expert, Nurse Oncologist, Radiologist, and Breast Cancer Pathologist

- For clinical oncologists, oncology experts, nurse oncologists, radiologists, and Breast cancer pathologists to participate in the study, they must have experience obtained through practice in oncology for not less than 1 year. In addition, a participant should have also consented by filling out a consent form to participate in the study.

4. Exclusion Criteria for a Breast Cancer Patient

A breast cancer patient below 8 years of age is not considered for this study.

4.1. Exclusion Criteria for a Clinician

Any clinical oncologist, oncology expert, nurse oncologist, radiologist, and Breast cancer pathologist who doesn't meet the inclusion criteria above.

4.2. The Population Sample Size: Disproportionate Sampling Design

The disproportionate sampling design method is deployed in this study as the data strata differ not only in size but also in variability, for example, the three data strata from the patients, oncologists, and records differ in size and variability. A larger sample is taken from more strata and smaller samples from the less variable strata [15]. The formula for the above statements is as follows:

$$n_1/N_1\sigma_1 = n_2/N_2\sigma_2 = \cdots = n_k/N_k\sigma_k$$

where $\sigma_1, \sigma_2, \cdots$ and σ_k denotes the standard deviations of the k strata,

$N_1, N_2, \cdots, N_k$ denotes the sizes of *k* strata. This is called "optimum allocation" in the context of disproportionate sampling. The allocation in such situations results in the following formula for determining the sample sizes of different strata:

$$n_i = n \cdot N_1\sigma_1 / N_2\sigma_2 + \cdots + N_k\sigma_k$$

for $i = 1, 2, \cdots$ and *k*. For example, in this study, we had strata oncologist *O*, strata patient *P*, and Strata data Record *R* with Standards deviation *O* = 15, *P* = 18, and *R* = 5, projected in the population of:

$$O_p = 5000, P_p = 2000 \text{ and } R_P = 3000 .$$

Thus total sample size of the dataset was *n* = 49 or *n* ≅ 50, apportion differently to sample size strata as follows:

$$O_p = 5000, n_O = n \cdot N_1\sigma_1 / N_2\sigma_2 + \cdots + N_k\sigma_k$$
$$= 49 \times 5000 \times 15 / 5000 \times 15 + 2000 \times 18 + 3000s = 29.16 \cong 30$$

as the sample size of oncologists or pathologists to be included in the study.

$$O_E = 2000, n_O = n \cdot N_1\sigma_1 / N_2\sigma_2 + \cdots + N_k\sigma_k$$
$$= 49 \times 2000 \times 18 / 5000 \times 15 + 2000 \times 18 + 3000s = 14$$

as the sample size of the data record to be included in the study.

$$O_p = 3000, n_O = n \cdot N_1\sigma_1 / N_2\sigma_2 + \cdots + N_k\sigma_k$$
$$= 49 \times 3000 \times 5 / 5000 \times 15 + 2000 \times 18 + 3000s = 5.833 \cong 6$$

as the sample size of patients to be included in the study.

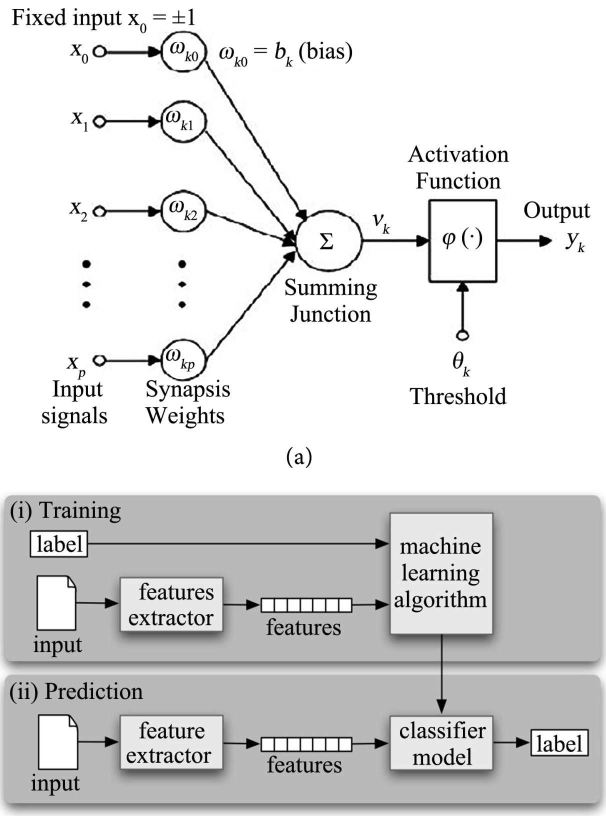
4.3. Data Collection

The dataset in this research study was collected from surveyed questionnaires. The surveyed questionnaires were built from suitable questions that were modified from related studies. In the questionnaire design, the Likert scale technique was used to determine if the respondent agreed or disagreed with a statement. The clinician's, and oncologists' survey questionnaires comprised section A, of 12 questions; section B, of 9 questions; section C, of 11 questions; and section D, of 21 questions on their clinician's perception regarding discharging and re-hospitalization of a breast cancer patient. Patient or Breast cancer survivor's survey questionnaires were made of section A, of 9 questions, and section B, of 20 questions. The clinician's survey questionnaire was distributed to the participating clinicians and oncologists in the department of Cancer Treatment and Oncology within KNH. Other questionnaires were distributed by pathologists who referred their colleagues from other breast cancer treatment and management centers in Kenya such as Aga Khan hospital. The principal investigator also interviewed the Breast cancer patient and Breast cancer survivor against their questionnaires but if the patient or Breast cancer survivor was in a position to write, then he or she was given a questionnaire to answer.

Irrelevant attributes in administered questionnaire such as a patient's residential address, name, application ID, etc. are removed. Obvious outliers and typo errors, for instance, a patient's father's name is irrelevant in predicting future re-hospitalizations. Cleaned "Re-hospitalization" data attributes are then added to hold the predictability result, which can either be "readmitted", "no readmission" or "both". Finally, a sample of the dataset tabular (database) as shown in Table 1 below is constructed from a cleaned dataset obtained from answered clinician's or oncologist's questionnaires, patient interview questions, and past patient records.

4.4. Data Preprocessing

Prediction of the re-hospitalization is obtained from summation threshold of risk variables which forms the basis upon which decision is made using the machine learning perceptron algorithm inbuilt in Rapid miner software. The concept of Breast cancer Risk variables thresholding is better conveyed by the perceptron algorithm shown in Figure 4(a) below. Supervised Classification here is done by perceptron algorithm.



(a)

(b)

Figure 4. (a): Perceptron algorithm inbuilt in Rapid miner software thresholding risk variable for Breast Cancer re-hospitalization; (b): Supervised Classification. (i) During training, the Breast cancer readmission risk feature extractor is used to convert each input value to a feature set. (ii) During prediction, the same feature extractor is used to convert unseen breast cancer risk inputs to pair classification segments. These feature sets are then fed into the model, which generates predicted labels, Steven Bird, Ewan Klein, and Edward Loper (2009).

Table 1. Dataset risk attributes predicting breast cancer re-hospitalization. perceptron algorithm is inbuilt in rapid miner software to learn the patterns.

Patients who were discharged (P) and their re-hospitaliation outcome	Stress Burden	Complete Medical dose	Smoking	Alcoholism	Physical Exercise	Fat Intake	Abuse Antibiotic	Bad rays	Abortion/Miscariage	Follow up appointments	Use Oral Contraceptive	Had Hormon Replacement Therapy	Bearing children	Appropriate Medication	Many Love partners	Overweight (BMI)	Appropriate discharge	Re-hospitalization Predicted Outcome
P1	Psychological isolated	Not Timely	Not a smoker	YES	Not Daily	No	Yes	NO	Yes	Yes	No	No	NO	Yes	No	No	Yes	**Readmitted**
P2	No Stress Burden	Timely	Not a smoker	No	Daily	No	No	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes	Yes	**No readmission**
P3	psychological & Physically isolated	Timely	Not a smoker	No	Daily	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	**Readmitted**
P4	No Stress Burden	Timely	Sneaky Smoker	YES	Not Daily	Yes	No	Yes	No	No	No	No	No	No	Yes	No	No	**Readmitted**
P5	psychological & Physically isolated	Not Timely	Commited Smoker	No	Not Daily	YES	YES	NO	No	Yes	Yes	Yes	YES	Yes	No	Yes	Yes	**Readmitted**
P6	No Stress Burden	Timely	Sneaky Smoker	YES	Daily	NO	YES	NO	Yes	No	No	No	NO	Yes	Yes	Yes	No	**Readmitted**
P7	Physically isolated	Timely	Not a smoker	No	Daily	No	No	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	**No readmission**
P8	No Stress Burden	Timely	Commited Smoker	No	Daily	Yes	No	Yes	No	Yes	Yes	No	Yes	No	No	No	No	**No readmission**
P9	Psychological isolated	Not Timely	Commited Smoker	No	Not Daily	No	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	No	No	**Readmitted**
P10	No Stress Burden	Timely	Not a smoker	YES	Daily	No	No	Yes	No	Yes	No	No	No	Yes	No	No	Yes	**Readmitted**
P11	Physically isolated	Not Timely	Sneaky Smoker	No	Daily	No	Yes	NO	No	Yes	No	Yes	YES	No	Yes	Yes	Yes	**Readmitted**
P12	No Stress Burden	Not Timely	Not a smoker	YES	Daily	Yes	No	NO	No	No	Yes	No	NO	No	Yes	Yes	No	**Readmitted**
P13	psychological isolated	Not Timely	Sneaky Smoker	No	Not Daily	No	No	No	No	No	Yes	No	No	No	No	No	Yes	**No readmission**
P14	No Stress Burden	Not Timely	Commited Smoker	No	Not Daily	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No	**No readmission**

Continued

P15	Physically isolated	Not Timely	Not a smoker	Yes	Daily	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	No	**Readmitted**
P16	psychological isolated	Not Timely	Sneaky Smoker	YES	Daily	Yes	Yes	Yes	No	No	No	Yes	No	No	Yes	Yes	No	**Readmitted**
P17	Physically isolated	Timely	Not a smoker	No	Daily	Yes	No	NO	No	Yes	No	Yes	YES	Yes	Yes	No	Yes	**Readmitted**
P18	psychological & Physically isolated	Timely	Commited Smoker	No	Daily	No	No	NO	Yes	No	Yes	Yes	NO	Yes	No	Yes	Yes	**Readmitted**
P19	Psychological isolated	Not Timely	Sneaky Smoker	No	Daily	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No	Yes	No	**No readmission**
P20	Physically isolated	Not Timely	Commited Smoker	Yes	Daily	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	No	Yes	**No readmission**
P21	No Stress Burden	Not Timely	Not a smoker	Yes	Daily	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	**Readmitted**
P22	psychological & Physically isolated	Timely	Not a smoker	No	Daily	No	No	Yes	Yes	No	No	No	No	Yes	No	No	Yes	**Readmitted**
P23	No Stress Burden	Timely	Not a smoker	No	Daily	No	No	NO	No	Yes	Yes	No	YES	Yes	No	Yes	Yes	**Readmitted**
P24	psychological & Physically isolated	Timely	Sneaky Smoker	YES	Not Daily	Yes	No	NO	Yes	Yes	No	Yes	NO	Yes	Yes	Yes	No	**Readmitted**
P25	psychological & Physically isolated	Not Timely	Commited Smoker	No	Not Daily	Yes	Yes	No	No	No	Yes	No	No	Yes	No	Yes	No	No readmission
P26	No Stress Burden	Timely	Sneaky Smoker	Yes	Daily	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	No	**No readmission**
P27	Physically isolated	Not Timely	Sneaky Smoker	No	Daily	No	Yes	Yes	No	No	No	No	Yes	No	No	No	Yes	Readmitted
P28	No Stress Burden	Not Timely	Commited Smoker	No	Not Daily	No	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Readmitted
P29	Psychological isolated	Not Timely	Not a smoker	YES	Daily	Yes	Yes	NO	No	No	Yes	Yes	YES	Yes	Yes	Yes	No	Readmitted
P30	Physically isolated	Not Timely	Sneaky Smoker	YES	Daily	Yes	Yes	NO	No	Yes	Yes	Yes	NO	No	No	No	Yes	Readmitted
P31	Physically isolated	Timely	Not a smoker	No	Daily	Yes	No	No	No	No	No	Yes	No	No	Yes	Yes	Yes	No readmission
P32	Physically isolated	Timely	Commited Smoker	No	Daily	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	No readmission
P33	Physically and psychological isolated	Not Timely	Sneaky Smoker	No	Not Daily	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Readmitted

Continued

P34	psychological isolated	Not Timely	Commited Smoker	YES	Not Daily	Yes	No	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	Readmitted
P35	psychological isolated	Not Timely	Not a smoker	YES	Not Daily	No	YES	NO	Yes	No	No	No	NO	No	Yes	Yes	No	Readmitted
P36	No Stress Burden	Timely	Not a smoker	No	Daily	No	No	No	No	Yes	Yes	No	No	No	Yes	No	Yes	No readmission
P37	psychological & Physically isolated	Timely	Not a smoker	No	Daily	No	No	Yes	Yes	Yes	No	No	Yes	Yes	No	No	No	No readmission
P38	No Stress Burden	Timely	Sneaky Smoker	Yes	Not Daily	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Readmitted
P39	psychological & Physically isolated	Not Timely	Commited Smoker	No	Not Daily	Yes	Yes	Yes	No	No	Yes	Yes	No	No	Yes	No	No	Readmitted
P40	No Stress Burden	Timely	Sneaky Smoker	YES	Daily	No	Yes	NO	Yes	Yes	No	No	YES	Yes	No	No	Yes	Readmitted
P41	Physically isolated	Timely	Not a smoker	No	Daily	No	No	NO	No	No	Yes	No	NO	No	No	Yes	Yes	Readmitted
P42	No Stress Burden	Timely	Commited Smoker	No	Daily	Yes	No	No	No	Yes	No	No	No	Yes	Yes	No	No	No readmission
P43	psychological isolated	Not Timely	Commited Smoker	No	Not Daily	No	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	No	No readmission
P44	No Stress Burden	Timely	Not a smoker	Yes	Daily	No	No	Yes	No	No	Yes	No	Yes	Yes	No	Yes	Yes	Readmitted
P45	Physically isolated	Not Timely	Sneaky Smoker	No	Daily	No	Yes	Yes	No	No	No	No	No	No	Yes	Yes	Yes	Readmitted
P46	No Stress Burden	Not Timely	Not a smoker	YES	Daily	YES	No	NO	No	Yes	No	Yes	YES	No	Yes	Yes	No	Readmitted
P47	psychological isolated	Not Timely	Sneaky Smoker	No	Not Daily	No	No	NO	No	No	No	Yes	NO	Yes	No	No	No	Readmitted
P48	No Stress Burden	Not Timely	Commited Smoker	No	Not Daily	No	Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No readmission

4.5. Modeling of a Clinical Decision Support Classifier to Predict Avoidable Re-Hospitalization of Discharged Breast Cancer Patient before Actual Discharging

The implementation Clinical Decision Support Classifier is divided into five stages. In the first stage, data attributes that cause re-hospitalization are collected and then uploaded into the Rapidminer software algorithm. In the third stage, data preprocessing and visualization in the rapid miner are executed. In the fourth stage, modeling and generation of predictive decision trees based on ID3 and C4.5 algorithms are actioned.

- **Splitting Re-hospitalization dataset to train and validate generated Breast Cancer Re-hospitalization prediction Model upon discharge**

The Rapidminer software has been designed by default to split the Re-hospitalization dataset attributes into training and validation sets. The hypothetical optimal decision tree that is generated from these datasets, learns the training dataset and gets validated using the validation dataset. The patient's dataset is split in the ratio of 0.7 and 0.3 to train and validate generated Breast cancer Re-hospitalization prediction model respectively. The diagram in Figure 4(b) below shows how the dataset is split in the ratio of 0.7 and 0.3 to train and validate generated Breast cancer Re-hospitalization prediction model. In Figure 4(b), Breast cancer feature sets on re-hospitalization, capture the basic information about each input risk variable to classify it as re-hospitalization related or non-re-hospitalization related. Pairs of these feature sets and labels are fed into the machine learning algorithm *i.e.* rapid miner software to generate a re-hospitalization predictive model.

- **Training of generated re-hospitalization decision tree classifier model**

Rapidminer software algorithm is configured in such a way that it generates a decision tree model classifier when provided with suitable risk datasets. Generated decision tree classifier is then trained to learn the training dataset as shown in Figure 4(b). As a result, a more optimal decision tree classifier is generated which is superior in classifying new unseen breast cancer re-hospitalization risk attributes. The ID3 and C4.5 Algorithms which are instances of decision tree algorithms are at play as the same are inbuilt in the Rapid miner software. The ID3 decision tree classifying model that has been generated from these breast cancer re-hospitalization risk datasets is represented in Figure 5 below.

- **The ID3 Classifier errors due to overfitting of the Breast Cancer re-hospitalization dataset**

The unpruned clinical decision tree model *i.e.* ID3 algorithm has a high error rate as can be demonstrated by the aid of Lemma's theorem illustrated in figure 6 below.

The overfitting of the Breast cancer re-hospitalization dataset takes place when a classifier function is too closely aligned to a limited set of data points. As a result, the model is useful in reference only to its initial training data set, and not to any other data sets. For example, let D be a set of Breast cancer data examples and let H be a hypothesis space. The hypothesis space $h \in H$ is considered to overfit D if an $h' \in H$ with the following property exists:

$$Err(h', D) < Err(h, D) \text{ and } Err^*(h) > Err^*(h'),$$

where $Err^*(h)$ denotes the true misclassification rate of h, while $Err(h', D)$ denotes the error of h' on the example set D. Reasons for overfitting are rooted in the example data set D and are that D is noisy, D is biased, and hence non-representative. D is too small and hence pretends unrealistic data properties. Thus given a hypothesis universal space H, and hypothesis h (or unpruned decision tree classifier), then $h \in H$. $h \in H$ is said to overfit the training

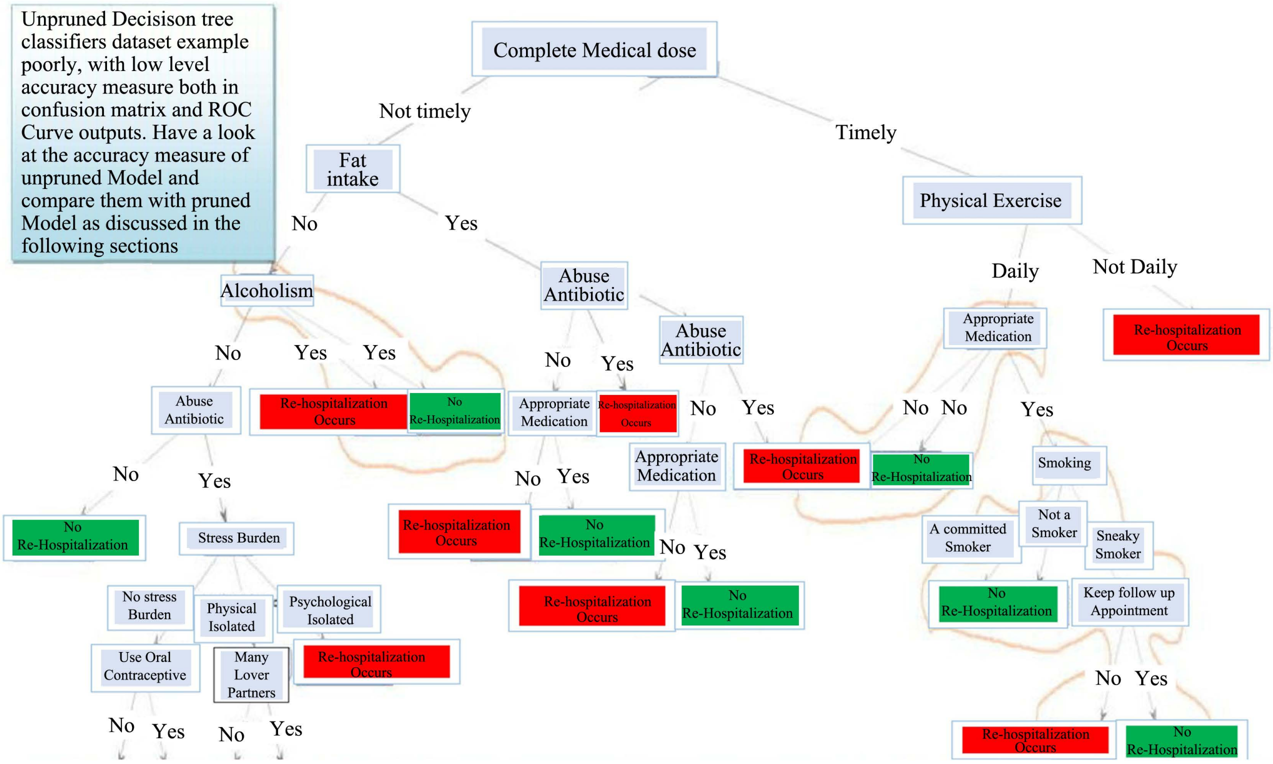

Unpruned Decision tree classifiers dataset example poorly, with low level accuracy measure both in confusion matrix and ROC Curve outputs. Have a look at the accuracy measure of unpruned Model and compare them with pruned Model as discussed in the following sections

Figure 5. The ID3 decision tree breast cancer re-hospitalization classifier prediction model (unpruned). Red implies re-hospitalization while green suggests no re-hospitalization.

dataset if there exists in H, an alternative hypothesis $h' \in H$ such that $h \in H$ has a bigger error than an alternative hypothesis over the training examples as $h' \in H$ which is an alternative has a smaller error than $h \in H$ which is considered to be unpruned. We can therefore conclude that $h \in H$ memorizes the training dataset and thus can be a less classifier beyond memorized training dataset than the $h' \in H$. This has been outlined by the curves in Figure 6.

- **The C4.5 decision tree classifier is an optimal Breast Cancer rehospitalization prediction model improved from the ID3 classifier Algorithm due to pruning**

The C4.5 algorithm is an improved version of the ID3 classifier algorithm with fewer decision nodes compared to the ID3 Algorithm as shown in Figure 7 below. By deployment of C4.5 Algorithm in this study, we ensure that pruning of the resulting decision tree from ID3 classifier algorithm has been done. The model in Figure 7 is the generated C4.5 decision tree classifier for the prediction of re-hospitalization of discharged Breast cancer patients obtained by pruning ID3 decision tree Breast cancer classifier model.

4.6. Evaluation and Analysis of Clinical Decision Support Model (C4.5 Algorithms) for Prediction of Breast Cancer Re-Hospitalization upon Discharge

- **Evaluation and Analysis via Confusion Matrix of Clinical Decision support Model (C4.5 Algorithms)**

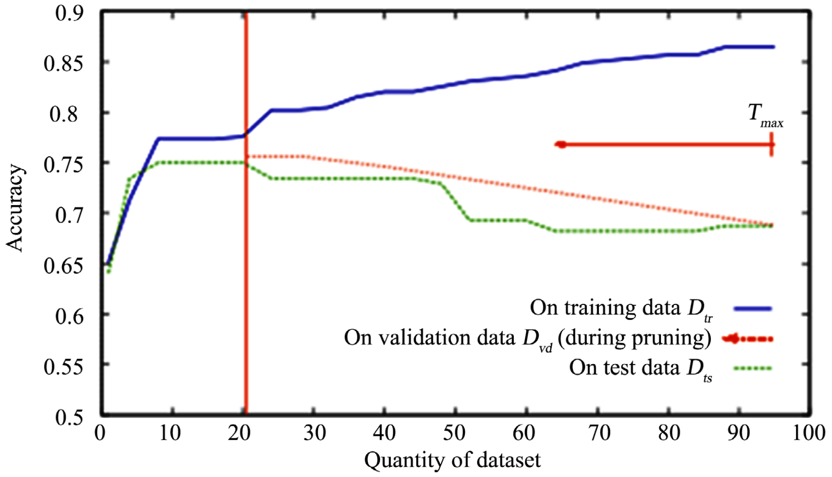


Figure 6. The effect of improving ID3 decision tree algorithm via pruning and cycles of training. This plot shows curves of training and test set accuracy. In addition, it shows the impact of reduced error by pruning and training the ID3 decision tree. Notice increase in accuracy over the test set as nodes are pruned. Here, the validation dataset used for pruning is distinct from both the training and test sets (Mitchel 1997).

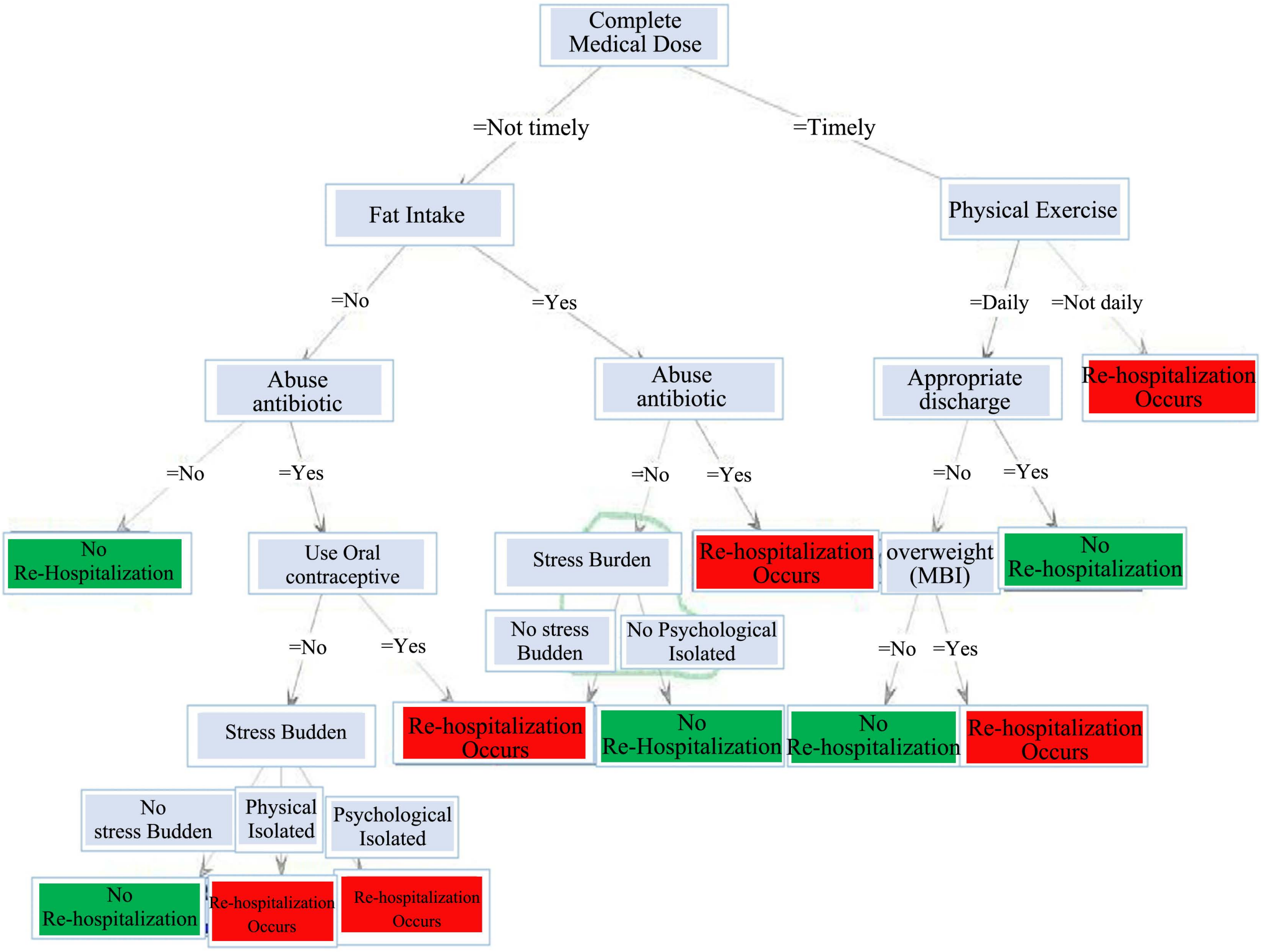


Figure 7. Red implies re-hospitalization while green suggests no re-hospitalization. The pruned clinical decision support model (C4.5 algorithms) for optimal breast cancer re-hospitalization prediction model improved from the ID3 classifier algorithm.

Considering classification problems using only two classes, each instance is mapped to one element of the set {p, n} of positive and negative class labels. A classification model (or classifier) is mapping from instances to predicted classes. The classification model may produce continuous output (e.g., an estimate of an instance I). Also, they may produce discrete class labels indicating only the predicted class of the instance.

In this analysis and evaluation, to distinguish between the actual class and the predicted class we use the labels {P1, N2}. The below **Figure 8(a)** shows a 78.57% measure of the accuracy of the Clinical Decision support Model when built on C4.5 Algorithm *i.e.* pruned ID3 Algorithm. We also obtained the detection of a True negative of 85.71% and a True positive of 71.43% indicating that clinical decisions modeled on C4.5 Algorithm have better performance outcomes as shown in **Figure 8(a)** below.

For precision measures, the Clinical Decision support Modelled on C4.5 Algorithms for the prediction of Breast Cancer re-hospitalization upon discharge recorded 75.00% in identifying True negative and 83.33% level of precision in detection of true positive as shown in the below **Figure 8(b)**.

Recall metrics measure the quantification of the number of positive class predictions made out of all positive examples in the dataset. In this instance, the Clinical Decision support Model built on C4.5 Algorithm has been able to quantify correctly the prediction of Breast Cancer re-hospitalization upon discharge by 71.43% for True positive and true negative by 85.71% showing the strength of the Clinical Decision support Model when built on C4.5 Algorithm. **Figure 8(c)** demonstrates this accomplishment.

These metrics outcomes mentioned above from the Clinical Decision support Model built on C4.5 Algorithm are also realized through the confusion matrix computation as shown in **Table 2** below.

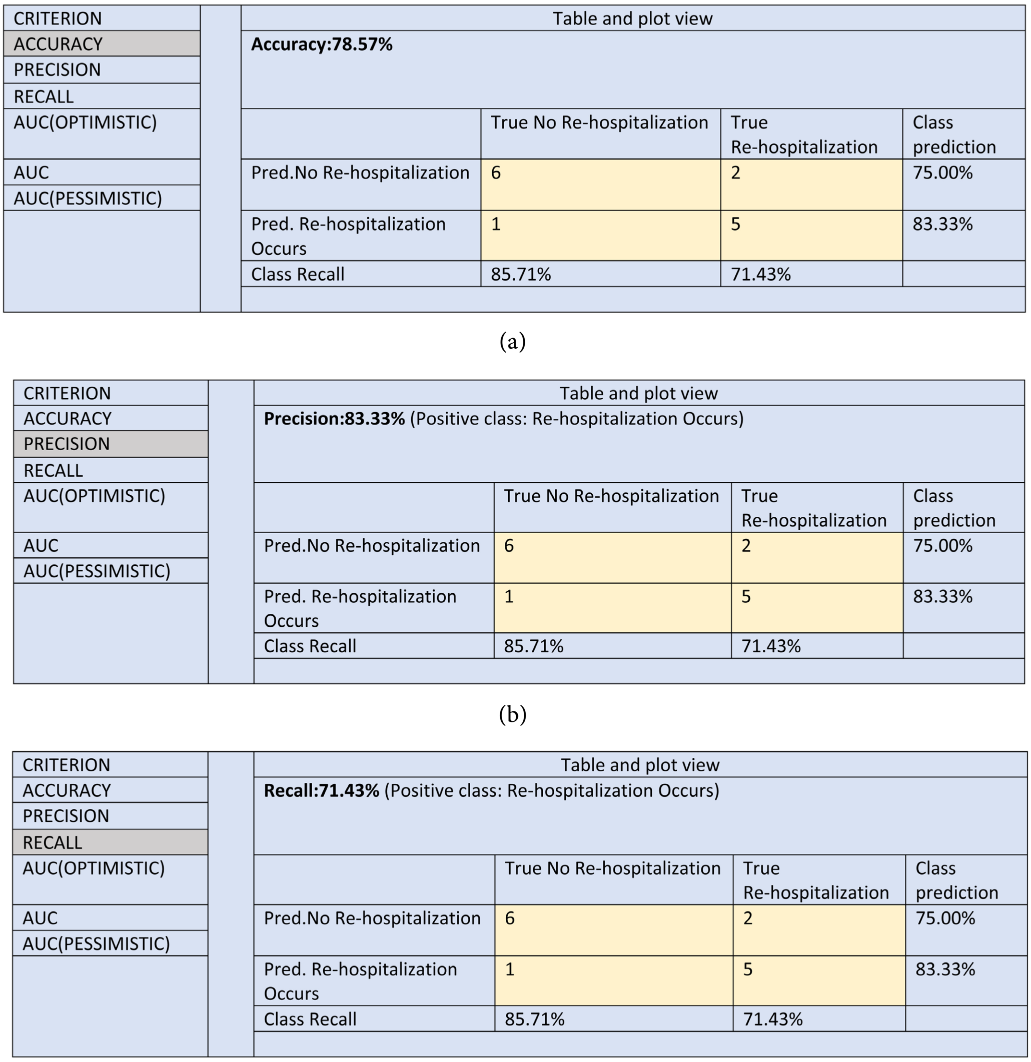
The confusion matrix and common performance metrics were calculated from the model. These were noted as follows:

Precision for True negative = (TP)/(TP) + (FP) = 6/8 = 75%, also

Precision for true positive = (TN)/(TN) + (FN) = 5/6 = 83.33%

Table 2. The confusion matrix of the clinical decision support model (C4.5 Algorithms) for prediction of breast cancer rehospitalization.

Patient class	Predicted by the model to be Re-hospitalized (P_1)	Predicted by the model not to be Re-hospitalized (N_2)	total	Recognition (%)
To be Re-hospitalized (P_1)	6 (TP)	2 (FP)	8	75% (*precision*)
Not Re-hospitalized (N_2)	1 (FN)	5 (TN)	6	71.42% (*specificity*)
Total	7 (P)	7 (N)	14 (All)	78.57% (*accuracy*)



CRITERION
ACCURACY
PRECISION
RECALL
AUC(OPTIMISTIC)
AUC
AUC(PESSIMISTIC)

Table and plot view

Accuracy:78.57%

	True No Re-hospitalization	True Re-hospitalization	Class prediction
Pred.No Re-hospitalization	6	2	75.00%
Pred. Re-hospitalization Occurs	1	5	83.33%
Class Recall	85.71%	71.43%	

(a)

CRITERION
ACCURACY
PRECISION
RECALL
AUC(OPTIMISTIC)
AUC
AUC(PESSIMISTIC)

Table and plot view

Precision:83.33% (Positive class: Re-hospitalization Occurs)

	True No Re-hospitalization	True Re-hospitalization	Class prediction
Pred.No Re-hospitalization	6	2	75.00%
Pred. Re-hospitalization Occurs	1	5	83.33%
Class Recall	85.71%	71.43%	

(b)

CRITERION
ACCURACY
PRECISION
RECALL
AUC(OPTIMISTIC)
AUC
AUC(PESSIMISTIC)

Table and plot view

Recall:71.43% (Positive class: Re-hospitalization Occurs)

	True No Re-hospitalization	True Re-hospitalization	Class prediction
Pred.No Re-hospitalization	6	2	75.00%
Pred. Re-hospitalization Occurs	1	5	83.33%
Class Recall	85.71%	71.43%	

(c)

Figure 8. (a) The overall accuracy of the clinical decision support model (C4.5 Algorithms) for the prediction of Breast Cancer re-hospitalization upon discharge stands at 78.57%; (b) The overall class precision of the Clinical Decision support Model (C4.5 Algorithms) for the prediction of Breast Cancer re-hospitalization upon discharge stands at 75.00% for True negative and true positive at 83.33%; (c) The overall recall for the Clinical Decision support Model (C4.5 Algorithm) for the prediction of Breast Cancer re-hospitalization upon discharge stands at 71.43% for True positive and true negative at 85.71%.

Recall = (TP)/(TP) + (FN) = 6/7 = 85.71% also similar to sensitivity.

Classifier Accuracy, or recognition rate: percentage of test set tuples that are correctly classified. Accuracy = (TP + TN)/All = 6 + 5/14 = 78.57%.

Error rate: 1 − accuracy, or Error rate = (FP + FN)/All = 2 + 1/14 = 0.2142857142857143.

Sensitivity: True Positive recognition rate (Recall positive),

Sensitivity = TP/P = 6/7 = 85.714%.

Specificity: True Negative recognition rate (Recall Negative), Specificity = 5/7 = 71.43%.

By convention, the performance of a classification model such as the Clinical

Decision support Model (C4.5 Algorithms) for the prediction of Breast Cancer re-hospitalization is usually summarized by the following two quantities related to the two types of errors: true-positive rate and false-positive rate. In this context, the true-positive rate is the probability that a patient to be re-hospitalized is correctly classified as such and the false-positive rate is the probability that a patient who shall not be re-hospitalized is incorrectly classified as shall be re-hospitalized. (The true-positive rate is also called sensitivity or recall and one minus the false-Positive rate is also called specificity). For an ideal classification rule, the true-positive rate is one and the false-positive rate is zero. The magnitudes of acceptable false-positive rates and true-positive rates depend on the corresponding costs and perceived benefits for the problem in question. For instance above results outcome for the Clinical Decision support Model (C4.5 Algorithm) for the prediction of Breast Cancer re-hospitalization is fairly acceptable to reinforce experts discharging decisions.

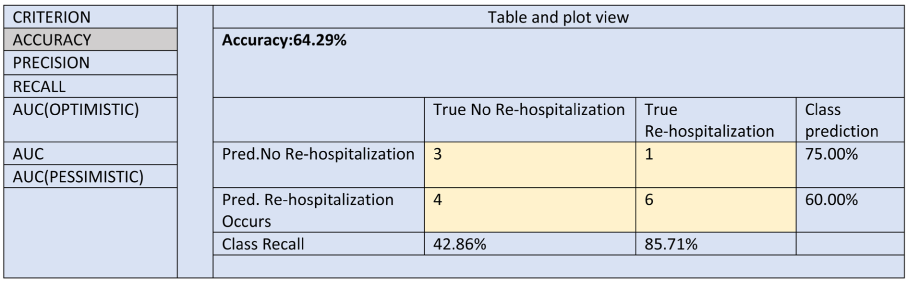
On the flip side, a comparison of C4.5 Algorithms performance to unpruned Clinical Decision support Model (ID3 Algorithms) for the prediction of Breast Cancer re-hospitalization indicates that the latter is a fewer performer in terms of accuracy, precision, and recalling while predicting re-hospitalization for discharged Breast cancer patients.. For example, **Figure 9(a)** below presents an accuracy of 64.29% which is lesser when compared to 78.57% which was obtained by the Clinical Decision support Model modeled on C4.5 Algorithm.

Furthermore, the unpruned Clinical Decision support Modeled on the ID3 Algorithm, provides the precision of 60.00% for True negative and true positive at 75.00% as shown in **Figure 9(b)** below unlike its counterpart implemented on the C4.5 Algorithm that scored precision of 75.00% for True negative and true positive of 83.33%.

It's worth also noting that recall of prediction by the Clinical Decision support Model fashioned on ID3 Algorithm stands at 85.71% for True positive and true negative at 42.86% as shown in the **Figure 9(c)** below which is away lower as compared Clinical Decision support Model produced on the C4.5 Algorithm.

- **The receiver operating characteristic (ROC) evaluation and analysis of the Clinical Decision support Model (built on ID3 and C4.5 Algorithm) for the prediction of Breast Cancer rehospitalization**

ROC graphs have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm rates of a classifier [16]. The medical decision-making community has extensive literature on the use of ROC graphs for diagnostic testing (Zou, 2002). Recent years have seen an increase in the use of ROC graphs in the machine learning community, due in part to the realization that simple classification accuracy is often a poor metric for measuring performance [17], and that they have properties that make them especially useful for domains with skewed class distribution (test set) and unequal classification error costs. These characteristics have become increasingly important as research continues into the areas of cost-sensitive learning and learning in the presence of unbalanced classes (test set).



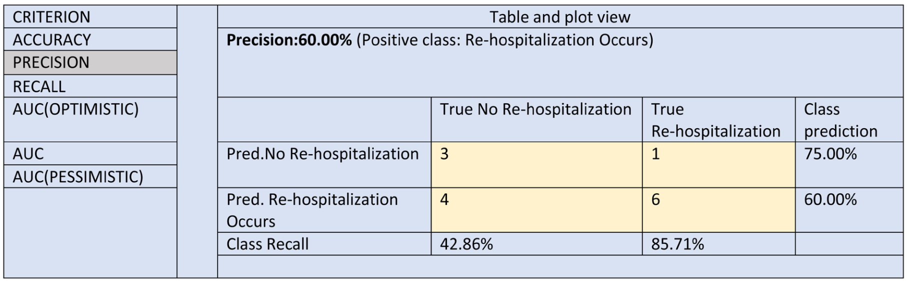
CRITERION
ACCURACY
PRECISION
RECALL
AUC(OPTIMISTIC)
AUC
AUC(PESSIMISTIC)

Table and plot view

Accuracy:64.29%

	True No Re-hospitalization	True Re-hospitalization	Class prediction
Pred.No Re-hospitalization	3	1	75.00%
Pred. Re-hospitalization Occurs	4	6	60.00%
Class Recall	42.86%	85.71%	

(a)



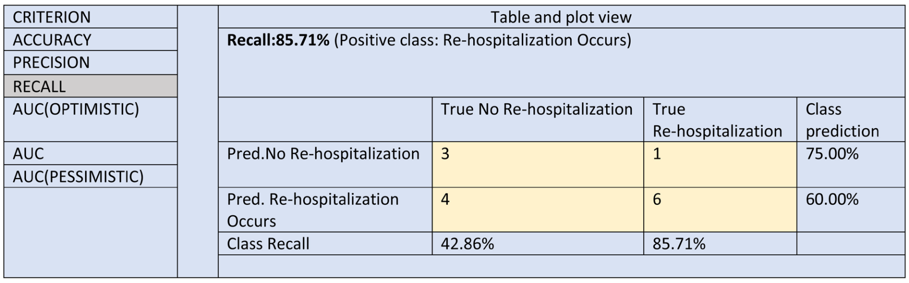
CRITERION
ACCURACY
PRECISION
RECALL
AUC(OPTIMISTIC)
AUC
AUC(PESSIMISTIC)

Table and plot view

Precision:60.00% (Positive class: Re-hospitalization Occurs)

	True No Re-hospitalization	True Re-hospitalization	Class prediction
Pred.No Re-hospitalization	3	1	75.00%
Pred. Re-hospitalization Occurs	4	6	60.00%
Class Recall	42.86%	85.71%	

(b)



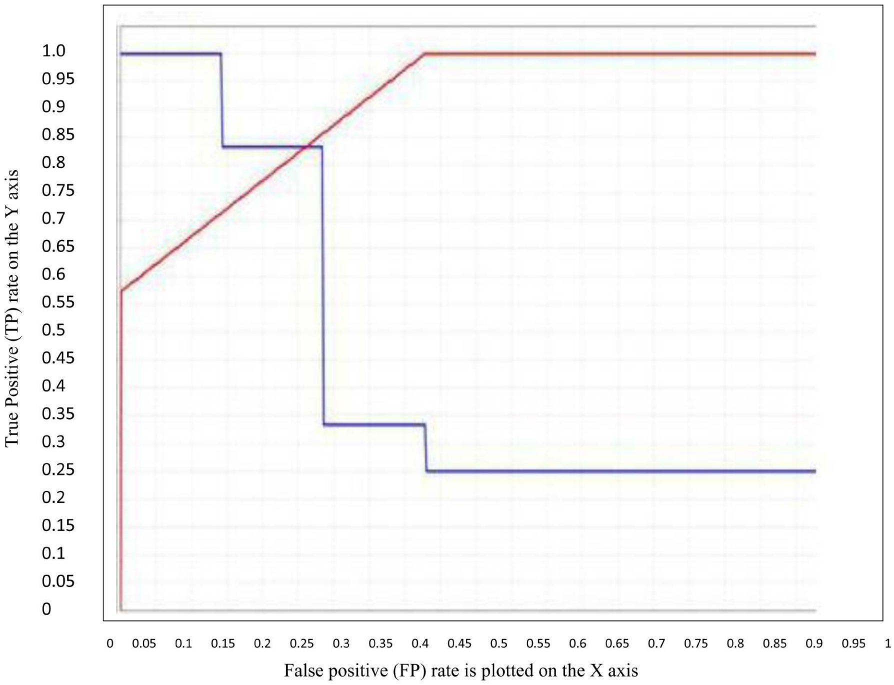
CRITERION
ACCURACY
PRECISION
RECALL
AUC(OPTIMISTIC)
AUC
AUC(PESSIMISTIC)

Table and plot view

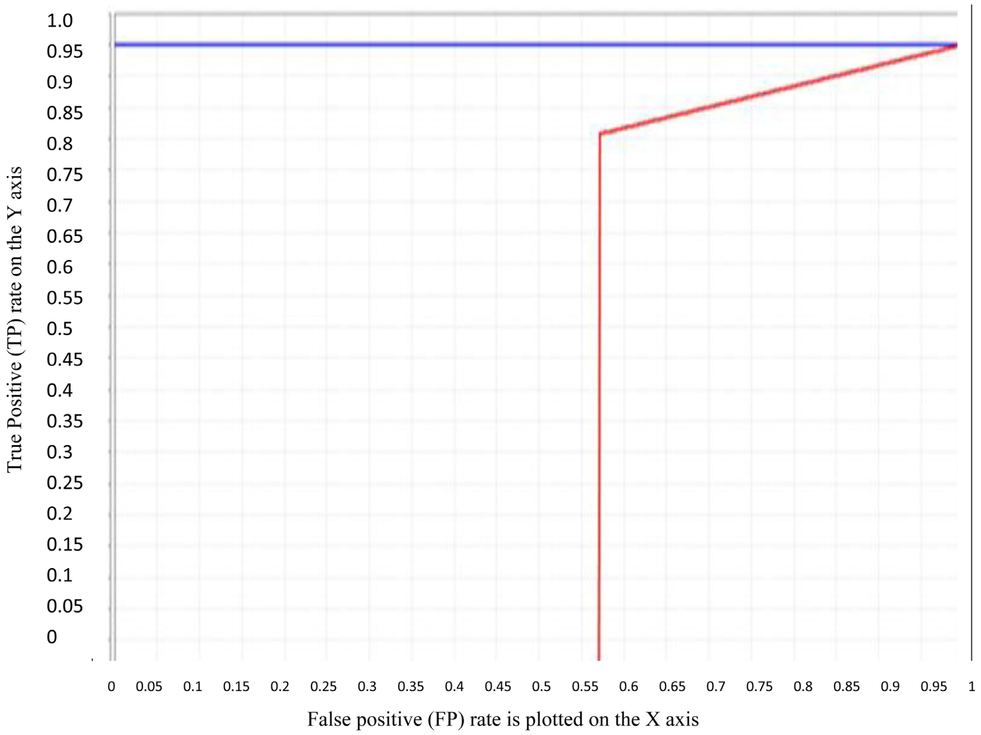
Recall:85.71% (Positive class: Re-hospitalization Occurs)

	True No Re-hospitalization	True Re-hospitalization	Class prediction
Pred.No Re-hospitalization	3	1	75.00%
Pred. Re-hospitalization Occurs	4	6	60.00%
Class Recall	42.86%	85.71%	

(c)



(d)



(e)

Figure 9. (a) The overall accuracy of the clinical decision support model (ID3 Algorithms) for the prediction of breast cancer re-hospitalization upon discharge stands at 64.29%. The detection of True negative is at 42.86% while True positive is at 85.71%; (b) The overall class precision of the clinical decision support model (ID3 Algorithms) for the prediction of breast cancer re-hospitalization upon discharge stands at 60.00% for True negative and true positive at 75.00%; (c) The overall recall for the clinical decision support model (ID3 Algorithm) for the prediction of breast cancer re-hospitalization upon discharge stands at 85.71% for True positive and true negative at 42.86%; (d) The ROC curve analysis for the clinical decision support model (C4.5 Algorithm) for the prediction of breast cancer re-hospitalization show less coverage on the receiver operating characteristic (ROC) graph indicating weak metrics; (e) The ROC curve analysis for the clinical decision support Model (ID3 Algorithm) for the prediction of breast cancer re-hospitalization show less coverage on the receiver operating characteristic (ROC) graph indicating weak metrics.

As seen above, ROC graphs are two-dimensional graphs in which the TP rate is plotted on the Y-axis and the FP rate is plotted on the X-axis. A ROC graph depicts relative tradeoffs between benefits (true positives) and costs (false positives). Figure 9(d) shows that each discrete classifier produces a (TP rate, FP rate) pair corresponding to a single point in the ROC space. Several points in the ROC space are important to note. The lower left point (0, 0) represents the strategy of never issuing a positive classification; such a classifier commits no false-positive errors but also gains no true positives. The opposite strategy, of unconditionally issuing positive classifications, is represented by the upper right point (1, 1). The point (0, 1) represents perfect classification. Informally, one point in ROC space is better than another if it is to the northwest (TP rate is higher, FP rate is lower) appearing on the left-hand side of a ROC graph, near the X-axis, may turn to correspond to one ROC point. Thus, a discrete classifier produces only a single point in the ROC space. The rationale for the optimal ROC curve is that: 1) One wants the highest true-positive rate for a given false-positive rate, and 2) One can specify a rule on the ROC line linking two (false-positive rate, true-positive rate) points by applying the rule for one point with some probability and the rule for the other point with one minus that probability. However, in practice, one would like one of the points on the optimal ROC curve to lie near the target false- and true-positive rates. For the reasons given above, interest is in the part of the ROC curve corresponding to a low false-positive rate when evaluating prediction. The area under (ROC) curve is known as AUC. This area, therefore, should be greater than 0.5 for a model to be acceptable; a model with an AUC of 0.5 or less is worthless. Understandably, this area is a measure of the predictive accuracy of the model. Based on this information, the area under the curve (ROC) for the C4.5 Algorithm is relatively bigger than for the ID3 Algorithm shown in the Figure 9(e) below implying that C.4.5 is a superior classifier.

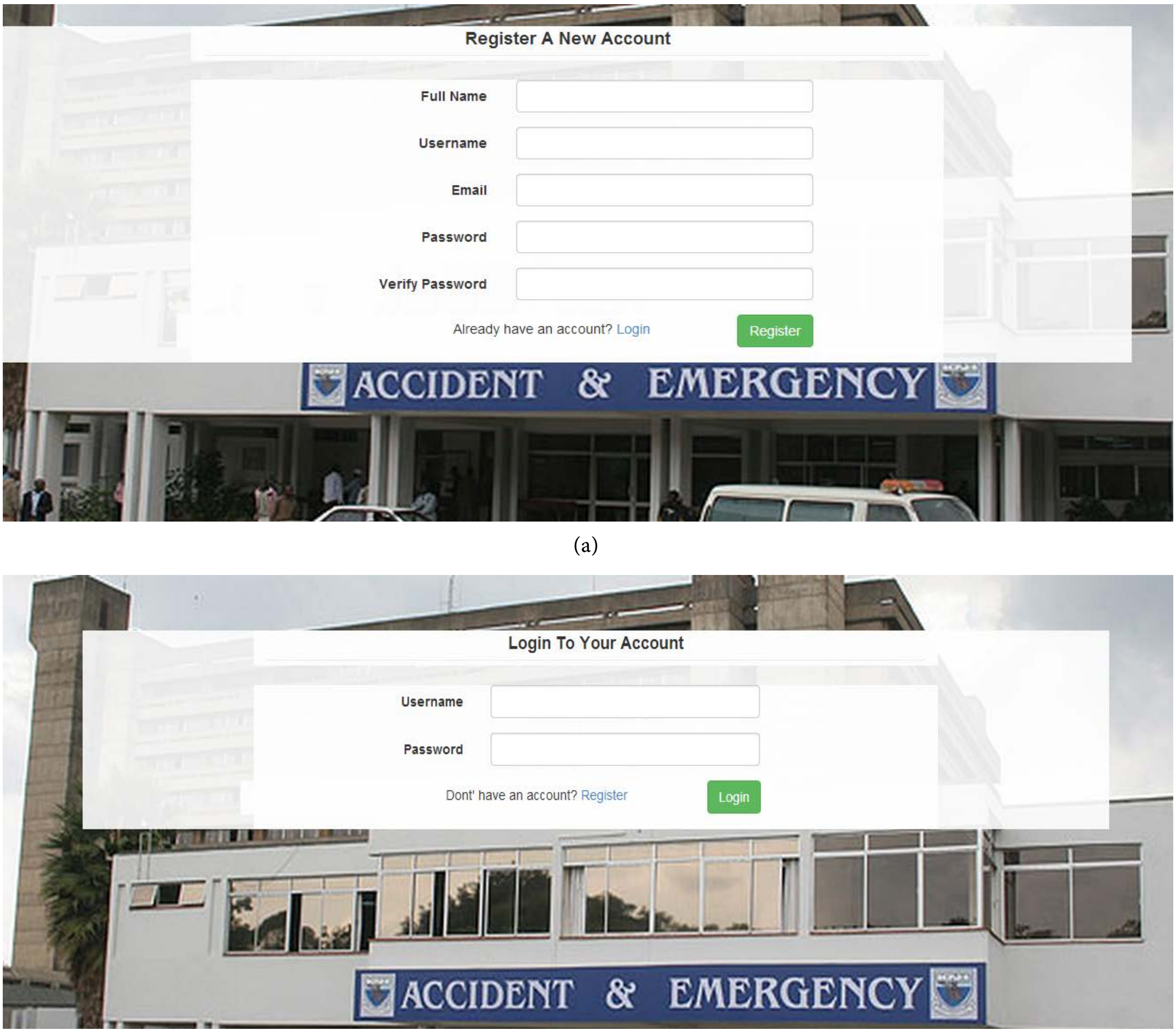
Finally, the confusion matrix and common performance metrics based on the Clinical Decision support Model (ID3 Algorithm) for the prediction of Breast Cancer re-hospitalization have lower metrics performance unlike if the Clinical Decision support Model is modeled on the C4.5 Algorithm for the prediction of Breast Cancer re-hospitalization.

4.7. Mapping of the Clinical Decision Support (Built on the C4.5 Algorithm) Model to Predict Re-Hospitalization of Breast Cancer Patients before Discharging Them to the General Public Using Python Web Application

The predictive patterns generated from data risk variables linked to Breast cancer readmission are coded into class methods of if-else ladders in Python programming language as shown in appendix I. The class method is configured in such a way that they accept only the splitting risk attributes. The class methods return the final result of that particular risk evaluation, indicating whether that patient would be re-hospitalized or will not should he/she be discharged. This understanding is mapped onto a web-based platform to assess and evaluate Breast Cancer patients before being integrated into the general public as, "medicare advantaged patients experience fewer re-hospitalizations", Jaimie Oh, (2012). Figure 10(a) and Figure 10(b)) seen below are the web-based application designed and mapped from this understanding. This web-based application is fashioned from Django-1.6.5 as explained in the succeeding section.

- **Code Igniter-Django-1.6.5.**

As illustrated above, a web application that is mapping Clinical Decision Support (built on the C4.5 Algorithm) model to predict the re-hospitalization of Breast cancer patients before discharging is developed through a Python framework named Django 1.6.5. The web application has provisions for multiple simultaneous clinician registrations and logins thus ensuring independent clinical evaluation. Figure 10(a) and Figure 10(b) below depicts the built Clinical


Register A New Account

Full Name

Username

Email

Password

Verify Password

Already have an account? Login

Register

(a)

Login To Your Account

Username

Password

Dont' have an account? Register

Login

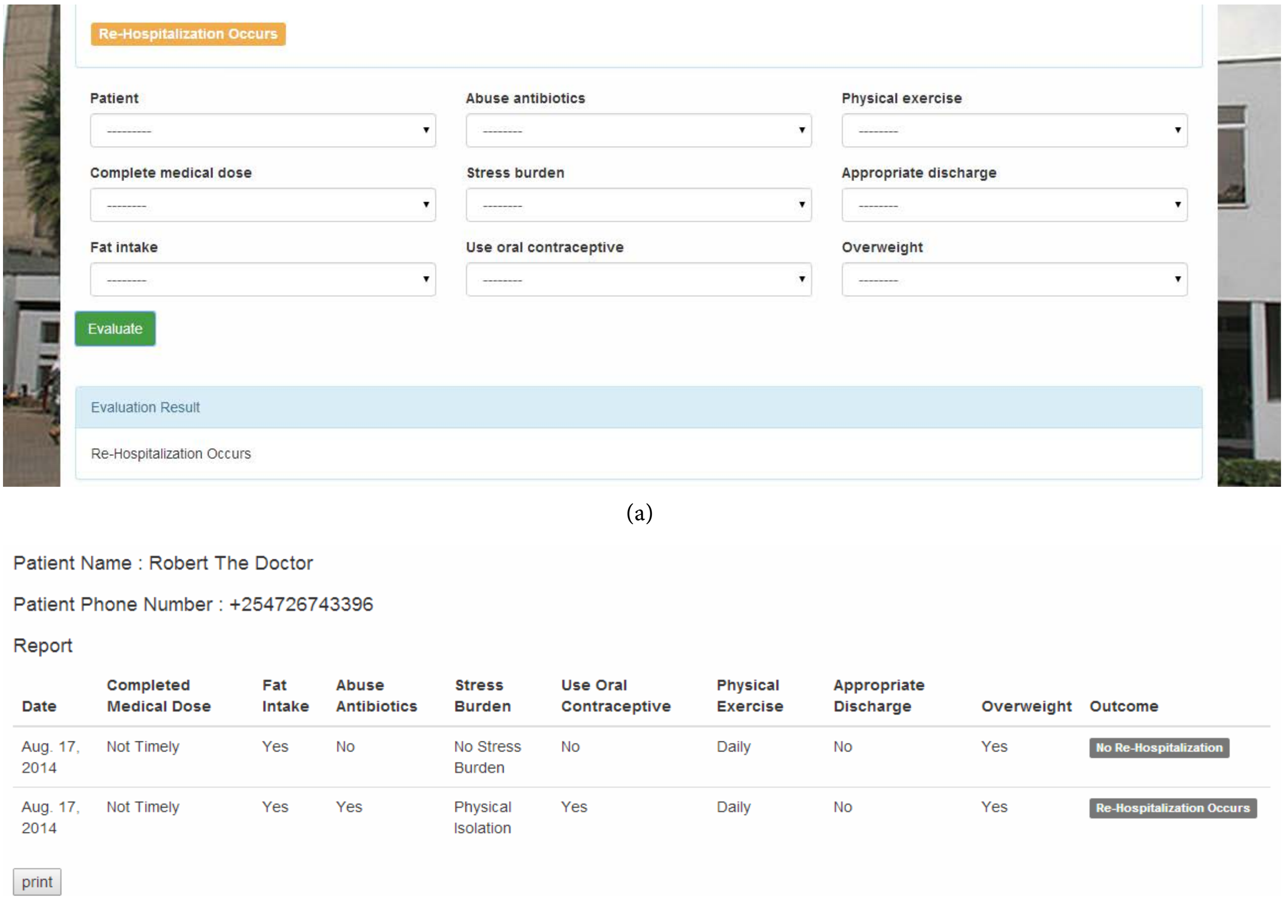
(b)

Figure 10. (a): Registration interface for clinician/oncologist who is discharging the patient; (b): Login interface for registered clinician/oncologist to undertake the prediction for re-hospitalization of Breast cancer patients before discharging.

Decision Support on the C4.5 Algorithm. Figure 10(a) is the registration interface for this clinical decision application while Figure 10(b) is the login interface respectively.

- **The web-based Python application Predicting to Re-hospitalization of Breast cancer patients before discharging**

Once the web-based Python application for Prediction of Re-hospitalization to Breast cancer patients is mapped in class python methods, a web-based application is designed for clinicians to do entries and evaluations on the splitting risk values attribute of a patient as conveyed below in Figure 11(a) and Figure 11(b) respectively. Breast cancer Risk values for re-hospitalization are used to evaluate avoidable re-hospitalization of a patient as either "Re-hospitalization occurs", or "no re-hospitalization".



Evaluation Result

Re-Hospitalization Occurs

(a)

Patient Name : Robert The Doctor

Patient Phone Number : +254726743396

Report

Date	Completed Medical Dose	Fat Intake	Abuse Antibiotics	Stress Burden	Use Oral Contraceptive	Physical Exercise	Appropriate Discharge	Overweight	Outcome
Aug. 17, 2014	Not Timely	Yes	No	No Stress Burden	No	Daily	No	Yes	No Re-Hospitalization
Aug. 17, 2014	Not Timely	Yes	Yes	Physical Isolation	Yes	Daily	No	Yes	Re-Hospitalization Occurs

(b)

Figure 11. (a): Web interface for breast cancer risk values attributes entry for results outcome of the re-hospitalization; (b): Printable result outcome of evaluated risk value attributes entered on re-hospitalization assessment.

5. Discussion and Conclusions

- **Discussions of the main Results**

The results of the Clinical Decision Support (built on the C4.5 or ID3 Algorithms) model to predict the re-hospitalization of Breast cancer patients before discharging are illustrated in Figure 8(a) which illustrated that the overall accuracy of the Clinical Decision support Model (C4.5 Algorithms) for the prediction of Breast Cancer re-hospitalization upon discharge as 78.57%. The detection of the true negative was noted as 85.71% while the true positive was at 71.43%. Also in Figure 8(b), the overall class precision of the Clinical Decision support Model (C4.5 Algorithms) for the prediction of Breast Cancer re-hospitalization upon discharge stood at 75.00% for true negative and true positive at 83.33%. Figure 8(c) noted that the overall recall for the Clinical Decision support Model (C4.5 Algorithms) for the prediction of Breast Cancer re-hospitalization upon discharge was 71.43% for true positive and true negative at 85.71%. Table 2, which is a confusion matrix manually generated and meant to validate presented metrics of the Clinical Decision support Model (C4.5 Algorithms) to Predict Breast Cancer re-hospitalization was noted generally to be congruent to presented metrics observed in **Figures 8(a)-(c)** and the ROC curve analysis of same Clinical Decision support Model (C4.5 Algorithm) predicting Breast Cancer re-hospitalization in **Figure 9(d)**. It is detailed also that the Clinical Decision support Model (on C4.5 Algorithms) has a small error rate of approximately 0.21.

In **Figure 9(a)**, the overall accuracy of the Clinical Decision support Model (ID3 Algorithms) for the prediction of Breast Cancer re-hospitalization upon discharge stood at 64.29% while the detection of true negative samples was at 42.86% as true positive stayed at 85.71%. **Figure 9(b)** also illustrates that the overall class precision of the Clinical Decision support Model (ID3 Algorithms) for the prediction of Breast Cancer re-hospitalization upon discharge is 60.00% for true negative and true positive remained at 75.00% as **Figure 9(c)** reports that the overall recall for the same Clinical Decision support Model (ID3 Algorithm) for the prediction of Breast Cancer re-hospitalization upon discharge stood at 85.71% for true positive and true negative at 42.86%. The juxtaposition of these metrics on the ROC curve analysis shown in **Figure 9(d)** for the Clinical Decision support Model (ID3 Algorithm), disclosed less coverage on the receiver operating characteristic (ROC) graph indicating weak metrics that are harmonious to its metrics observed for same the Clinical Decision support Model (ID3 Algorithms) to Predict Breast Cancer re-hospitalization upon discharge.

To sum up, the face value of metrics from either the C4.5 model or ID3 model, confidently says that the Clinical Decision Support built on the C4.5 model to predict the re-hospitalization of Breast cancer patients before discharging is more acceptable evidenced by its accuracy, recall, and precision metrics unlike the Clinical Decision Support built on the ID3 algorithm. Observed also is that the Clinical Decision Support model built on the ID3 algorithm is a poorer classifier may be because of its unpruned nature, noise, insufficient data, and or skewed data distribution.

Poorer classification at the initial stage by the Clinical Decision support Model (C4.5 Algorithms) depicted by its ROC curve could be due to the wastage of "heat" needed to overcome the data "friction" having been observed also by Collins Mowel in simulation and modeling.

- **Value of the study**

By implementing this study, we bridge the existing gap in breast cancer discharging and re-hospitalization currently witnessed in Kenyatta National Hospital occasioned by less informed clinical Breast cancer discharge that is merely based on experts' opinions which is insufficiently reinforced for better treatment value outcomes. The rein-forced discharge decision achieved by this study provides better treatment outcomes through a timely decision to work hand in hand with the expertise in deriving an integrative discharge decision having been agreed upon in the medical circles as a strategy that is likely to eliminate the fore-seeable deterioration quality of health for a discharged breast cancer patients thus surging rates of mortality blamed on mistrusted discharge decisions. Reviewed literature didn't reveal existing studies conducted in Kenyatta National

Hospital (KNH) in the context of breast cancer discharging and re-hospitalization. This is another compelling force relevant for undertaking this study. This study is also significant as it may be among the transformative practices desired for value-based treatment and management of breast cancer besides. This study is also contributing to theoretical knowledge for discharging and re-hospitalization of Breast cancer patients. The study generates new theories hence valuable for theory building. The new types of breast cancer risk data sets available are rich in detail and combine information of multiple types (e.g., temporal, cross-sectional, geographical, and textual) with a high level of granularity. Such data often contain complex relationships and patterns that are hard to hypothesize, especially given theories that exclude many newly measurable concepts. The model is designed to operate in such environments and detects new patterns and behaviors and helps uncover potential new causal mechanisms, in turn leading to the development of new theoretical models.

This study is solving the poorly structured problem of uncertainty in clinical decision-making by minimizing avoidable mistakes, adverse events, and the problem of thinking hard at the point of decision making hence reduction of subsequent source-related death. Through this study, an unstructured or poorly structured problem is transformed into a structured problem by stating the problem's initial state, solution state, and target state. When the problem is in these states, the curse of dimensionality in the decision-making is lowered significantly because there are some rules and directives on how to reach the target solution.

The study access relevance of the decision made by Breast cancer experts to discharge a patient thus remains true that if we can predict success based on a certain explanation (*i.e.* C4.5 model or ID3 model), then we have a good reason, and perhaps the best sort of reason, for accepting the explanation. This model is, therefore, a useful tool for assessing the distance between theory (Statistical model) and practice especially when headed to infinity hence assessing the predictive power of a theory to sheds light on the actual performance of an empirical model. The model (*i.e.* C4.5 model or ID3 model), can therefore be used to assess the practical relevance of a theory (Keil *et al.*).

The study through the Model (*i.e.* C4.5 model or ID3 model), can also be used to improve existing models since it captures complex underlying patterns and relationships, thereby improving existing explanatory statistical models (Collopy *et al.*).

- **Limitations of the research**

In this study, one cannot make progress without adequate size and quality of Breast cancer dataset risk for training the model that is to predict Breast cancer patients' rehospitalization upon discharge. Getting these Breast cancer risk variables dataset is tedious and cumbersome. This study also requires that you have a clear definition of the concept to be predicted and concept examples.

Among the difficulties in implementing the Clinical Decision Support (built on the C4.5 or ID3 Algorithms) model to predict the re-hospitalization of Breast cancer patients before discharging in everyday clinical practice come mainly from programmers' insufficient understanding of medical reasoning and decision analyses to accurately inform the model.

For this predictive model to be successful in predicting re-hospitalization before discharging Breast cancer patients, the training risk data variables must be representative of the test data. Typically, the training data come from the past, while the test data arise in the future. If the re-hospitalization to be predicted is not stable over time as in the case of the Breast case which is made of obscurity and keeps on changing, then predictions are likely not to be useful. For example, changes in the general economy, lifestyle, and social attitudes towards breast cancer are all likely to change the behavior of patients in the future. The model, therefore, needs constant updates with time which has not been implemented in this study.

Also, the predictive model can mislead clinicians and Breast cancer experts to an ever-increased focus on optimizing predictive power at the expense of understanding the broader situation of theory building and richer content of attributes on avoidable re-hospitalization. Clinicians should be aware of this model's temptation which can shift away from their attention to the real problem of concept building. Clinicians should also be aware that the model doesn't read their minds but work on the "sword of data" and that the model is supportive but they make the actual decision.

- **Future Research Works**

Clinicians and oncologists expect the predictive model to read their minds and deliver the exact verdict on the re-hospitalization problem. However, up to date, artificial intelligence and Machine learning don't read minds. They simply give causal relations and underlying patterns. Thus there is a need to research and provide an autonomous "mind-reading" model to predict the re-hospitalization of Breast cancer patients before discharging.

As noted also for the predictive model to be successful, the training data must be representative of the test data. Typically, the training data come from the past, while the test data arise in the future. Since the re-hospitalization to be predicted is not stable over time, then predictions are likely not to be useful. Changes in the general economy, lifestyle, and social attitudes towards breast cancer are all likely to change the behavior of patients in the future. It is therefore recommended for research on a predictive model which will automatically update itself with the constant lifestyle changes and other Breast cancer re-hospitalization risks.

In **Figure 5**, we observed a single decision tree (ID3) algorithm weakness for prediction as it's shallow and thus unable to predict deeper variables risk thus rendering an incomplete prediction. This study recommends replicating a similar study but now on a random forest algorithm or the Gradient Boosting algorithm. Implementation of this study on the decision tree (ID3) algorithm or C4.5

algorithm was easy for prediction of Breast Cancer re-hospitalization by following the path nodes and finding the result. The random forests algorithm is the most accurate learning and classifier algorithm due to its nature of reducing variance through using different samples for training the model and also building and combining small (shallow) trees. When the Random Forest algorithm is compared to the Gradient Boosting algorithm, it is noted that just like random forests, gradient boosting uses a set of single decision trees though the random Forest algorithm builds each tree independently while gradient boosting builds one tree at a time. The Random Forest algorithm is an additive model (ensemble) that works in a forward stage-wise manner, by introducing a weak single decision tree learner to improve the shortcomings of existing single weak decision tree learners thus covering each other black spots. Gradient boosting combines results along the way. It shall be thought-provoking to see this study replicated through either the Random Forest algorithm or the Gradient Boosting algorithm instead of a single decision tree (*i.e.* ID3 Algorithm or C4.5 algorithm).

- **The Conclusion**

The Clinical Decision Support (built on the C4.5 Algorithms) model to predict the re-hospitalization of Breast cancer patients before discharging that has been implemented in this study doesn't work independently of a clinician or an oncologist. The goal of the model is to make clinicians and oncologists less wrong than they were and not for them to assume that this model will make them 100 percent right in their clinical discharging decision. The model has a high level of performance metrics though slightly lower prediction at the initial stages. We recommend this model for side-by-side Breast cancer clinical practice before discharging Breast cancer patients who may seem to have improved in their quality of health after prolonged medication in the hospital ward.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Anthony, D., Chetty, V.K., Kartha, A., McKenna, K., Rizzo DePaoli, M. and Jack, B. (2005) Re-Engineering the Hospital Discharge. In: Henriksen, K., Battles, J.B., Marks, E.S., *et al.*, Eds., *Advances in Patient Safety*: *From Research to Implementation* (*Volume* 2: *Concepts and Methodology*), Agency for Healthcare Research and Quality, Rockville, 379-394.

[2] Musimbi, A. (2013) Strategic Management of Breast Cancer. https://slideplayer.com/slide/5754092/

[3] Kohn, L.T., Corrigan, J.M., Donaldson, M.S., (Eds.) (1999) To Err Is Human: Building a Safer Health System. National Academy Press, Washington DC.

[4] Nyongessa, C. (2014) Kenya Society of Haematology and Oncology (KESHO). https://www.slideshare.net/drnyongesa1/history-of-kesho-by-dr-catherine-nyongesa

[5] Kelly, C. (2014, February 18) Cervical Cancer Treatment Breakthrough Using Anti-HIV Drugs Discovered at Manchester University/Kenyatta National Hospital in

Nairobi. https://www.mancunianmatters.co.uk/news/18022014-cervical-cancer-treatment-breakthrough-using-anti-hiv-drugs-discovered-at-manchester-university/

[6] Healthcare Services Specification Project (2005) Service Functional Model Specification, Decision Support Service (DSS), Version 0.432.

[7] The Johns Hopkins ACG System (2009) Excerpt from Technical Reference Guide Version 9.0. The Johns Hopkins University, Baltimore.

[8] Billings, J., Mijanovich, T., Dixon, J., Curry, N., Wennberg, D., Darin, B., *et al.* (2006) Case Finding Algorithms for Patients at Risk of Re-Hospitalization PARR1 and PARR2. NYU Center for Health and Public Service Research, New York.

[9] Yi, C. (2012) Developing Decision Trees to Classify Patients Suited for Similar Interventions by Combining Clinical Judgments with Leeds Risk Stratification Tool. The Johns Hopkins University. https://www.hopkinsacg.org/document/developing-decision-trees-to-classify-patients-suited-for-similar-interventions-by-combining-clinical-judgments-with-leeds-risk-stratification-tool/

[10] Curry, N., Billings, J., Darin, B., Dixon, J., Williams, M. and Wennberg, D. (2005) Predictive Risk Project Literature Review. https://www.kingsfund.org.uk/sites/default/files/field/field_document/predictive-risk-literature-review-june2005.pdf

[11] Natale, J. and Wang, S. (2006) Prediction of Readmission of Heart Failure via Decision Tree Model. University of Akron, Akron.

[12] Pantilat, S.Z., Lindenauer, P.K., Katz, P.P. and Wachter, R.M. (2002) Primary Care Physician Attitudes Regarding Communication with Hospitalists. *Disease-a-Month*, **48**, 218-229. https://doi.org/10.1016/S0011-5029(02)90029-5

[13] http://www.bu.edu/fammed/projectred/

[14] https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm

[15] Kothari, C.R. (2004) Research Methodology: Methods and Techniques. New Age International Publishers, New Delhi, 78.

[16] Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010 https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X#

[17] Provost, F. and Fawcett, T. (2001) Robust Classification for Imprecise Environments. *Machine Learning*, **42**, 203-231. https://doi.org/10.1023/A:1007601015854 https://www.researchgate.net/publication/225910756_Robust_Classification_for_Imprecise_Environments

Appendix

The Clinical Decision Support (built on the C4.5 Algorithms) model to predict the re-hospitalization of Breast cancer patients before discharging code Prototype

```
import json
NODE_NAMES = ['Timely Medical Dose', 'Fat Intake', 'Abuse Antibiotics',
              'Use Oral Contraceptive', 'Stress Burden', 'Physical Exercise',
              'Appropriate Discharge', 'Overweight',
              'No Re-Hospitalization', 'Re-Hospitalization Occurs']
class Node(object):
               def __init__(self, node_value):
        '''Initialize the node
        '''
        self.key = node_value
        self.positive_node = None # When the response is positive
        self.negative_node = None # When the response is negative
    def __repr__(self):
        return self.key
    def get_node_value(self):
        '''
        Sets the node' name
        '''
        return self.key
    def set_node_value(self, value):
        '''
        Returns the node's value
        '''
        self.key = value
    def set_positive(self, value):
        '''
        Sets the node returned when given a positive value
        '''
        if not self.positive_node:
            self.positive_node = Node(value)
        else:
            node = Node(value)
            current = self.positive_node
            node.positive_node = current
            self.positive_node = node
    def set_negative(self, value):
        '''Sets the node returned when given a positive value
        '''
        if not self.negative_node:
            self.negative_node = Node(value)
        else:
            node = Node(value)
            current = self.negative_node
            node.negative_node = current
            self.negative_node = node
    def get_positive_node(self):
        return self.positive_node
    def get_negative_node(self):
        return self.negative_node
    def is_tree(self):
        if self.negative_node and self.positive_node:
            return True
        return False
def addtree(data, tree=None):
    '''    Build the tree recursively
    '''if not tree:
        return
    positive_value = data[tree.key_dict][1]
    if positive_value:
        tree.set_positive(data[positive_value][0])
        tree.get_positive_node().key_dict = positive_value
    addtree(data, tree.get_positive_node())
    negative_value = data[tree.key_dict][2]
    if negative_value:
        tree.set_negative(data[negative_value][0])
        tree.get_negative_node().key_dict = negative_value
    addtree(data, tree.get_negative_node())
    return tree
def buildtree(path, root_value="Timely Medical Dose"):
    f = open(path)
    data = json.load(f)
    root = Node(root_value)
    root.key_dict = "Timely Medical Dose"
    return addtree(data, tree=root)
def evaluate(data, node):
    if not node.is_tree():
        return node
    value = data[node.get_node_value()]
    if value:
        leave = evaluate(data, node.get_positive_node())
    else:leave = evaluate(data, node.get_negative_node())
return leave
```