Received: 3 November 2020 | Revised: 10 May 2021 | Accepted: 27 May 2021

DOI: 10.1002/nop2.963

RESEARCH ARTICLE

NursingOpen Open Access WILEY

# Using data mining technology to predict medication-taking behaviour in women with breast cancer: A retrospective study

Chen-Chen Kuo[1,2] | Hsiu-Hung Wang[2] | Li-Ping Tseng[3,4]

[1]The Cancer Prevention and Treatment Center, St. Martin De Porres Hospital, Chiayi, Taiwan

[2]School of Nursing, Kaohsiung Medical University, Kaohsiung, Taiwan

[3]Management Center, St. Martin De Porres Hospital, Chiayi, Taiwan

[4]Department of Industrial Engineering and Management, National Yunlin University of Science and Technology, Douliu, Taiwan

**Correspondence**
Hsiu-Hung Wang, School of Nursing, Kaohsiung Medical University, 100, Shih-Chuan 1st Road, San-Ming District, Kaohsiung City 80708, Taiwan.
Email: hhwang@kmu.edu.tw.

**Funding information**
St. Martin De Porres Hospital, Grant/Award Number: 17B-029

## Abstract

**Aims:** Medication-taking behaviours of breast cancer survivors undergoing adjuvant hormone therapy have received considerable attention. This study aimed to determine factors affecting medication-taking behaviours in people with breast cancer using data mining.

**Design:** A longitudinal observational retrospective cohort study with a hospital-based survey.

**Methods:** A total of 385 subjects were surveyed, analysing existing data from January 2010 to December 2017 in Taiwan. Three data mining approaches—multiple logistic regression, decision tree and artificial neural network—were used to build the prediction models and rank the importance of influencing factors. Accuracy, specificity and sensitivity were used as assessment indicators for the prediction models.

**Results:** Multiple logistic regression was the most effective approach, achieving an accuracy of 96.37%, specificity of 96.75% and sensitivity of 96.12%. The duration of adjuvant hormone therapy discontinuation, duration of adjuvant hormone therapy use and age at diagnosis by data mining were the three most critical factors influencing the medication-taking behaviours of people with breast cancer.

KEYWORDS

adherence, breast cancer, data mining, medication-taking behaviours, persistence

## 1 | INTRODUCTION

Breast cancer is one of the most common cancers among women worldwide (World Health Organization, 2019). Approximately, 2.1 million new cases of breast cancer were diagnosed in 2018, and nearly 627,000 deaths were attributed to the disease (WHO, 2019). Approximately, 75% of breast cancers are oestrogen receptor-positive (ER+) (Moon et al., 2017). Most ER + people with breast cancer are prescribed adjuvant hormonal therapy (AHT), which blocks the effects of oestrogen to inhibit breast cancer cell growth (Moon et al., 2017). AHT choices of tamoxifen or aromatase inhibitors (AIs) depend on the menopausal status of the patient (Pourcelot et al., 2018). AHT for a period of 5–10 years is a standard therapy for obtaining the maximum benefits in breast cancer treatment (Davies et al., 2013; Pourcelot et al., 2018; Robinson et al., 2018). It has been proven that AHT is effective in reducing the risk of recurrence and the mortality rate of women with breast cancer (Davies et al., 2013; Makubate et al., 2013; Moon et al., 2017; Pourcelot et al., 2018). Despite the proven clinical benefits (Cahir et al., 2015; Moon et al., 2017), patient non-adherence and non-persistence with AHT through the full course of treatment are common (Robinson et al., 2018).

## 2 | BACKGROUND

In the last two decades, increasing importance has been placed on research on medication-taking behaviour (MTB). WHO (2003) formalized the concept of a person taking medication, that is, medication adherence. This concept has been classified into three common terms: compliance, adherence and persistence (American Society on Aging [ASA] and the American Society of Consultant Pharmacists [ASCP] Foundation, 2006). Compliance is defined as the patient passively following a doctor's orders; adherence is defined as the patient agreeing with receiving treatment and actively following a doctor's orders; and persistence is defined as the patient's ability to continue taking medication for the intended course of therapy (ASA and ASCP Foundation, 2006; Brown & Bussell, 2011).

Adherence is a multidimensional phenomenon (ASA and ASCP Foundation, 2006). In most quantitative studies, common multifactorial reasons for non-adherence were influenced by key factors, including (a) patient factors: a lack of health literacy, comorbidities and treatment side effects; (b) provider factors: patient–provider relationships, office visits and follow-up care with the oncologist; (c) environmental factors: characteristics of society and social isolation; (d) socioeconomic factors: race and medication costs of long-term treatment; (e) healthcare system factors: out-of-pocket expenditure and access to care; (f) condition-type factors: acute or chronic; (g) factors of daily dose number: single or multiple; and (h) ease of use factors: fewer drug interactions and drug-food interactions (Cahir et al., 2015; Jacobs et al., 2018; Mohan et al., 2019). A systematic review published in 2017 reported that most studies concentrating on clinical and demographic factors yielded inconsistent results. It reported that a good patient–physician relationship and self-efficacy for taking medication increased persistence (Moon et al., 2017); however, a younger age and frequent hospitalizations were associated with non-adherence (Moon et al., 2017).

In clinical practice, adherence is defined according to the prescribed time, dosage and frequency of AHT (Cramer et al., 2008; Osterberg & Blaschke, 2005). Persistence is defined as continuing the AHT for a prescribed duration of time (Osterberg & Blaschke, 2005). However, according to WHO, the measurement of medication adherence involves two main methods: subjective and objective measures (WHO, 2003). Subjective measures include healthcare professional assessments and self-report questionnaires, which are the most common tools, such as the Morisky Medication Adherence Scale and modified Siegal scale (Brown & Bussell, 2011; Mohan et al., 2019; Pourcelot et al., 2018). Objective measures, including pill counts, electronic medication monitors, biochemical measurements obtained by detecting serum drug levels in the patient's blood or urine (Brown & Bussell, 2011), and rate of prescription refill (Makubate et al., 2013; Osterberg & Blaschke, 2005). Measuring the adherence to AHT in women with breast cancer by estimating the rate of prescription refills is relatively objective and the data can be easily obtained from medical records (Murphy et al., 2012). Common methods of measuring adherence to AHT are most often defined as a medication possession ratio (MPR) > 80% (Hsieh et al., 2014; Mohan et al., 2019; Murphy et al., 2012); while an MPR of <80% is defined as non-adherence (treatment interruption) (Hsieh et al., 2014; Partridge et al., 2008). Persistence is defined as the discontinuation of the AHT drug after exceeding a permissible gap, which ranged from 60–180 days, depending on various factors (Hsieh et al., 2014; Murphy et al., 2012; Nekhlyudov et al., 2011). Each measurement method has both advantages and disadvantages and no method is regarded as the gold standard (Lam & Fresco, 2015; Osterberg & Blaschke, 2005). Therefore, researchers should use a multi-measure approach to overcome the limitations in clinical practice to improve patients' medication adherence.

About clinical effectiveness, previous studies' evidence has proven that 5 years of AHT significantly reduces breast cancer recurrence and mortality (Moon et al., 2017). Many women with ER + breast cancer do not regularly take AHT as prescribed, resulting in early discontinuation, low adherence and significantly poorer survival (Davies et al., 2013; Makubate et al., 2013; Moon et al., 2017). A systematic review of 61 articles concluded that adherence to tamoxifen or AIs ranged from 47%–97% and fell from an average of 79% in 1 year of AHT use to 56% by the fourth or fifth year (Moon et al., 2017). Discontinuation ranged from 9%–63% and rose from an average of 21% in the year of AHT use to 48% by the fourth or fifth year (Moon et al., 2017). Additionally, a previous cohort study with a large sample showed that women with breast cancer are initially highly adherent to AHT and that as time progresses, patients decrease their adherence. The annual adherence by women with ER + breast cancer changed from 90% in year 1 to 82% in year 2, 77% in year 3, 59% in year 4 and 51% in year 5 (Makubate et al., 2013). That cohort study showed a high discontinuation rate ranging from 30%–51% (Makubate et al., 2013). Therefore, AHT medication adherence and persistence play critical roles in recurrence and mortality.

Previous quantitative studies, systematic reviews and meta-analyses have focussed on objective and subjective factors that influence medical adherence and persistence in AHT for breast cancer. These studies have highlighted the complexity of factors that influence MTB in AHT, leading to different conclusions. This makes factors that affect adherence and persistence to AHT in breast cancer, difficult to understand. Since receiving AHT is long-term and expensive, AHT compliance prediction is important for women with breast cancer. Previous studies used logistic regression or Cox proportional hazard models to determine the explicit factors associated with AHT compliance in women with breast cancer (Bhatta et al., 2013; Makubate et al., 2013). However, it was not clear if these included factors that influenced both AHT adherence and persistence in patients. Therefore, we attempted to solve this problem using other analytical techniques, such as data mining.

With the rapid development of information technology, data mining has been widely applied to health condition assessment, disease detection and diagnoses, treatment effect analysis, survival prediction, and quality improvement in healthcare (Yu et al., 2015). Data mining is an innovative method that is more flexible and powerful than traditional methods (e.g. log-normal, logistic regression, Cox regression and Kaplan–Meier) (Çığşar & Ünal, 2019; Delen et al., 2005;

Jajroudi et al., 2014). The data mining process includes data cleaning, data amalgamation, data selection, data modification, data mining, pattern study and knowledge perception (Sondhi, 2017). Data mining, together with computational technology, can reduce the binding and limitations of data analysis methods (Yu et al., 2015). Data mining can be used to analyse large-size data sets and complex variables, find hidden relationships between different attributes and extract information that is understandable and valuable for data users (Oskouei et al., 2017; Pourhoseingholi et al., 2017). In other words, the use of data mining techniques for various clinical data sets can be helpful in recognizing and obtaining valuable information or knowledge and can aid in accurate prediction (Ghorbani & Ghousi, 2019; Srinivas et al., 2010).

The use of prediction methods to understand the outcome of a disease is one of the most interesting and challenging tasks in clinical practice (Delen et al., 2005). Kourou et al. (2014) demonstrated the use of data mining in the prediction and prognosis of cancer. Similarly, Oskouei et al. (2017) used data mining techniques for a comprehensive survey of breast cancer diagnosis, treatment and prognosis. They reviewed 125 articles and compared various algorithms and techniques to measure the accuracy rate of prediction using data mining. Additionally, Pourhoseingholi et al. (2017) compared different data mining methods for predicting 5-year survival in colorectal cancer cases. The results showed that an ensemble performed better than the basic classifier methods with an AUC of 0.96 (Pourhoseingholi et al., 2017). Ghorbani and Ghousi (2019) selected 23 studies conducted between 1997–2018 on breast cancer by data mining and compared various classification and evaluation methods used in breast cancer diagnosis. They found that decision trees (DTs) and artificial neural networks (ANNs) were the best classification methods and prediction accuracy was the best evaluation method. However, Delen et al. (2005) compared three data mining methods for predicting survivability in breast cancer cases. These authors claimed that the use of a large SEER data set and 72 variables provided insight into the prediction model based on different data mining methods. They used three classification models: logistic regression (LR), DT (C5.0), and ANN. A limitation of most of the previous studies is that neither the experience of physicians nor the use of current techniques was adequate in determining the influencing factors. Until now, minimal research has been published on data mining in nursing. Therefore, the aim of this study was to use a data mining approach to predict the impact factors of long-term AHT adherence and persistence in females with breast cancer in Taiwan.

## 3 | METHODS

### 3.1 | Study design and data source

We conducted a longitudinal observational retrospective cohort study and a single-hospital survey based on the breast cancer population using secondary data. Study cohorts were defined from the Taiwan Cancer Registry (TCR) database for breast cancer from January 2010 to December 2017, and their medical records covered 5 years of AHT or until death. Data sources were extracted from TCR and the patients' medical records. Data were retrieved for all women newly diagnosed with ER + breast cancer on the TCR. We used longitudinal electronic health records (i.e. drug prescriptions, medication possession ratio, medication profile and physician order entry system), and TCR database to estimate adherence and persistence to initial AHT among women with ER + breast cancer before December 31, 2018. Patients were assigned to two groups based on the actual situation of taking medication (treatment adherence and persistence) rather than according to the secondary data based on the sampling at the initial study. In addition, we applied an innovative method of data mining to identify complex and multivariable factors influencing adherence and persistence of MTB in people with breast cancer. Three data mining approaches using multiple logistic regression (MLR), decision tree and artificial neural network were used to build the prediction models. Each model produced a sorted result of the affecting factors. Then, by applying the Borda count, the results were sorted to generate a final ranking of all the factors affecting MTB in AHT (Tseng et al., 2017).

### 3.2 | Participants

The electronic records of all participants with International Statistical Classification of Disease and Related Health Problems, Tenth Revision, Clinical Modification (ICD-10-CM) principal diagnosis of breast cancer (C50. 011-C50. 922) were extracted from the TCR database. The inclusion criteria were as follows: (a) newly diagnosed with ER + breast cancer, as documented in medical charts from January 2010 to December 2017, and (b) individuals 18 years old or above, with breast cancer who had received initial AHT (including tamoxifen, anastrozole, letrozole and exemestane). A sample of 421 female patients was potentially eligible for the study; 35 (8.3%) were excluded due to the following reasons: death in one year ($N = 1$), contraindication owing to patient risk factors (e.g. comorbid conditions, advanced age and tumour progression prior to administration, ($N = 14$)), transfer to another hospital ($N = 12$) and patient or family refusal of AHT ($N = 8$). Finally, 385 patients were retained for analysis in the integrated TCR and HER databases.

### 3.3 | Data collection

Data were collected from May to December 2018 at a regional hospital in Taiwan. Based on factors related to medication adherence behaviour and on a review of related literature, the variables for analysis were selected; these included socio-demographic variables (Mohan et al., 2019; Moon et al., 2017; Pourcelot et al., 2018; Sella & Chodick, 2020), clinical variables (Cahir et al., 2015; Guedes et al., 2017; Moon et al., 2017; Sella & Chodick, 2020), diagnostic variables (Cahir et al., 2015; Murphy et al., 2012), treatment variables (Cahir et al., 2015; Guedes et al., 2017; Mohan et al., 2019; Moon et al., 2017; Pourcelot et al., 2018), healthcare variables (Mohan et al., 2019; Moon et al., 2017; Pourcelot et al., 2018) and health system variables (Cahir et al., 2015; Mohan et al., 2019) and measures involving secondary database analysis (Lam & Fresco, 2015). Sixty-five independent variables were extracted using medical records and TCR data as research data sources (Table 1). In our study, persistence and adherence to MTB were selected as outcome variables. Each ER + breast cancer patient's AHT prescription duration was followed from the first AHT prescription date through the last date of AHT prescription coverage. Persistence was defined as the continuous use of AHT without gaps in prescription refills or treatment. (Cahir et al., 2015; Osterberg & Blaschke, 2005). Persistency was defined as the percentage of patients who continued to fill AHT prescriptions over time (Partridge et al., 2008). Persistence was measured as the AHT-covered period between two consecutive AHT prescriptions for less than 90 days (Hsieh et al., 2014; Murphy et al., 2012; Nekhlyudov et al., 2011). Adherence was defined as the degree of compliance with the prescribed dosage and daily frequency of AHT (Cahir et al., 2015; Osterberg & Blaschke, 2005). In other words, adherence was defined as the proportion of days that patients had oral AHT available over the observation period (i.e. medication possession ratio [MPR]) (Partridge et al., 2008). A conventional MPR cut-off point of more than 80% was used to define adherence (Hsieh et al., 2014; Mohan et al., 2019; Murphy et al., 2012).

### 3.4 | Analysis

SPSS version 19.0 was used for statistical data analysis to calculate the descriptive statistics. Participants were characterized using descriptive statistics, including means and standard deviations for continuous variables, such as age, follow-up time and AHT utilization time, and percentages for categorical variables, such as sex, age ranks, Charlson comorbidity index (CCI) score and AHT utilization pattern. Data were analysed using data mining techniques and algorithms. A data analysis algorithm based on the MATLAB R2013a software was used for data integration, algorithm development and modelling (Kumar & Lenina, 2016). The software is widely used in engineering, statistics, finance, signal and image processing, testing, measuring and various other research applications (Kumar & Lenina, 2016). Sixty-five variables were defined as input data. The output variables were identified as "1" for AHT adherence and persistence and "0" for non-adherence or non-persistence. To build a model for predicting AHT adherence trends using the MATLAB R2013a software, three predictive data mining methods were employed: MLR by SPSS Clementine, DT (C5.0) by SPSS Clementine 12.0, and ANN by a multilayer perception. These methods were used to construct an ensemble-learning scheme. The ANN structure had 65 input variables with one node accounting for bias, 34 hidden neurons with one node accounting for bias, and two output variables of

**TABLE 1** Classification of the variables

| Category | Variables |
|---|---|
| Demographics | Date of birth, education, marital status, occupation, caregiver, income |
| Cancer diagnosis | Age at diagnosis, sequence number, class of case, class of diagnosis status, class of treatment status, date of first contact, date of initial diagnosis, primary site, laterality, histology, behaviour code, grade/differentiation, tumour size, regional lymph nodes examined, regional lymph nodes positive |
| Cancer staging | Clinical stage group, pathologic stage group |
| Type of treatment methods | Type of first course of treatment, date of most definite surgical resection of the primary site, surgical procedure of primary site at this facility, surgical margins of the primary site, scope of regional lymph node surgery at this facility, reason for no surgery of primary site, receipt of radiotherapy, sequence of radiotherapy (RT) and surgery, sequence of locoregional therapy and systemic therapy, reason for no RT, dose to high risk clinical target volume (CTV_H ) (cGy), number of fractions to CTV_H, chemotherapy at this facility, hormone/steroid therapy at this facility, target therapy at this facility, palliative care at this facility |
| Treatment results | Date of first recurrence, type of first recurrence, vital status, cancer status, cause of death |
| Site-specific factors | Height, weight, body mass index (BMI), smoking behaviour, betel nut chewing behaviour, drinking behaviour |
| Site-risk factors | Oestrogen receptor assay, progesterone receptor assay, response to neoadjuvant therapy, number of sentinel lymph nodes examined, number of sentinel lymph nodes positive, Nottingham or Bloom–Richardson score/grade, HER2 IHC test lab value, Paget disease, lymph vessels or vascular invasion |
| Medical records | Charlson comorbidity index (CCI), Eastern Cooperative Oncology Group (ECOG) performance status, type of hormone therapy drugs, duration of AHT discontinuation, duration of AHT use, duration between the first course of treatment date and the initial diagnosis date |

the output layer. A flowchart of the proposed scheme, which comprises five stages, is presented in Figure 1.

### 3.5 | Validity, reliability, and rigour

The Surveillance, Epidemiology and End Results (SEER) programme of the National Cancer Institute (NCI) is regarded as the gold standard for data quality in the US and global cancer registries (Duggan et al., 2016). The TCR is similar to that of the SEER (Huang et al., 2019). Thus, TCR provides reliable cancer statistics for the Taiwanese population. The three predictive performance variables noted in the above paragraph were evaluated and compared in terms of accuracy, specificity, sensitivity and area under the curve (AUC) (Chuang et al., 2016; Jajroudi et al., 2014; Sakr et al., 2017; Tseng et al., 2017; Wang et al., 2017; Zou et al.,2007). Accuracy was defined as the percentage of correctly classified samples from all samples. AUC represented the area under the receiver operating characteristic curve (ROC curve; Chuang et al., 2016; Jajroudi et al., 2014; Sakr et al., 2017; Tseng et al., 2017; Wang et al., 2017; Zou et al., 2007). The AUC classification performance was defined as very poor (AUC $< 0.6$), poor ($0.7 >$ AUC $\geq 0.6$), fair ($0.8 >$ AUC $\geq 0.7$), good ($0.9 >$ AUC $\geq 0.8$) and excellent (AUC $\geq 0.9$; Hosmer et al., 2013). The AUC ranged from 0 (incorrect) to 1 (100% correct; Zou et al., 2007). The following metrics were calculated in each iteration of the measurement process: Accuracy $= (TP + TN)/(TP + TN + FP + FN)$, sensitivity $= TP/(TP + FN)$ and specificity $= TN/(TN + FP)$, where TP, TN, FP and FN denote true positive, true negative, false positive and false negative, respectively (Jajroudi et al., 2014; Sakr et al., 2017).

### 3.6 | ETHICAL CONSIDERATIONS

Research Ethics Committee approval was granted by the Institutional Review Board (IRB) of St. Martin De Porres Hospital, Chiayi City, Taiwan (IRB No: 17B-029). This study did not require patients' informed consent due to the retrospective analysis of routinely collected clinical data. Safe data management was conducted according to the National Health and Medical Research Council Guidelines for Human Research (National Health & Medical Research Council, 2007).

## 4 | RESULTS

We analysed a total of 385 patients' medical records. The overall mean age at diagnosis was 55.1 years (range: 29–90 years), and 100% of the subjects were female. Half of the women were between 50–69 years of age. Fifty-three per cent of the subjects had a senior high school education, 79.7% were married, and 63.1% were unemployed. The AHT follow-up time was $3.6 \pm 2.3$ years. The AHT course was $3.1 \pm 1.9$ years. The CCI score of 0 was 67.5%. The socio-demographic characteristics of the participants are summarized in Table 2.

Adherence information was available for 272 (70.6%) while persistence information was available for 292 (75.8%) of the 385 subjects (Table 2). Moreover, complete information on both medication adherence and persistence was available for only 227 subjects. The tamoxifen-only adherence and persistence rate was 31.6% and 30.8%, respectively, and AIs only were 53.7% and 51.7%, respectively. Most medical records did not include adverse effects of AHT experienced by the patient and their impact on medication adherence and persistence. The findings indicated that the mean adherence decreased each year from 85.5% in year 1 to 75.5% in year 2, 75.1% in year 3, 71.1% in year 4 and 63.9% in year 5. Moreover, medication adherence and persistence were lower in the younger group (age $< 50$ years; 37.9% vs. 38.7%) and older group ($>70$ years old; 14.0% vs. 11.6%) than that among the subjects ranging from 50–69 years old (48.2% vs. 49.7%).

Evaluation results of the three classification techniques for medication adherence and persistence are shown in Table 3. The metrics, namely, overall accuracy, specificity, sensitivity and AUC, were used to compare the evaluation performance in predicting the MTB of patients on AHT. Comparing the three classification models in Table 3, it is observed that the classification accuracies of the MLR, DT and ANN models were 96.37%, 93.52% and 92.29%, respectively. Furthermore, the three classification models showed greater accuracy in predicting the MTB of patients with AHT. Among all the models, the MLR model was the most accurate, followed by the DT model, and the ANN model had the worst performance. The AUC value ranges from 0.90 (ANN) to 1.00 (MLR). It was clear that the MLR model was the optimum approach for MTB prediction in female patients with breast cancer receiving AHT. The importance levels of the predictor variables using the three algorithms are presented in Table 4. For each algorithm, a ranking result was given for all the predictors. From Table 4, the top five influencing factors determined on the basis of the Borda count can also be observed: (a) duration of AHT discontinuation, (b) duration of AHT use, (c) age at diagnosis, (d) body mass index (BMI) and (e) receipt of radiotherapy.

## 5 | DISCUSSION

Three main findings were obtained in this study. First, it is evident that medication adherence and persistence with MTB are influenced by demographic factors. In particular, low adherence and persistence increase cancer recurrence and mortality (Cramer et al., 2008; Makubate et al., 2013). Second, data mining methods can be used to develop a model with a high degree of predictive accuracy. Such a model can help in understanding the factors that influence MTB. Third, the top five influencing factors were identified in the study by employing data mining.

For breast cancer, AHT is recommended for 5–10 years (National Comprehensive Cancer Network, 2020). The Early Breast Cancer Trialists' Collaborative Group (2005) reported that tamoxifen could reduce the risk of recurrence and death rate by 31% in women with breast cancer. A large population-based study reported that

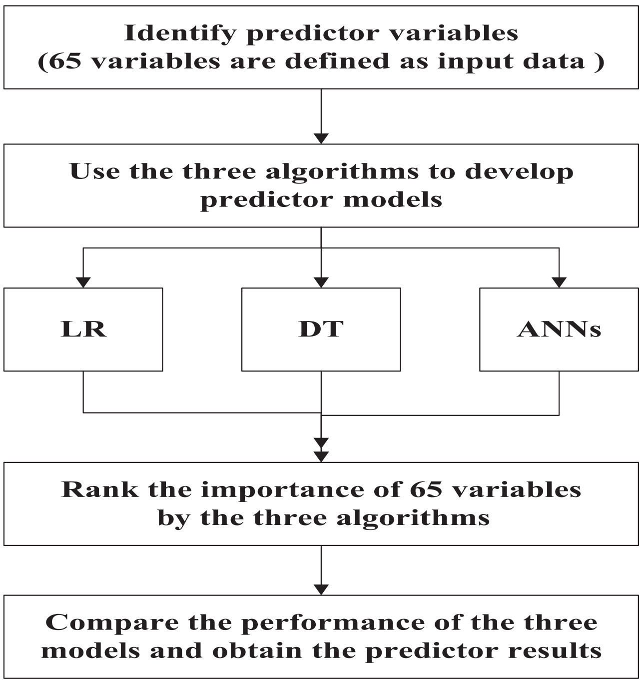


**FIGURE 1** Flowchart of the proposed predictive modeling scheme (Tseng et al., 2017)

the early discontinuation rate for breast cancer was 23% (Sella & Chodick, 2020). These results are in agreement with our findings. Furthermore, a 79% AHT (mean) adherence rate in the first year and a 56% AHT adherence rate in the fourth or fifth year were reported previously (Moon et al., 2017). Similar findings were obtained in a population-based study in Taiwan ($N = 30,573$), which reported that 77.3% and 85.1% of AHT adherence and persistence rates were higher than those in our study (Hsieh et al., 2014). Compared to a previous study in Taiwan about the age at diagnosis, we found that the percentage of patients younger than 50 years was 47.0%, which is higher than that in our study (Hsieh et al., 2014). Of the non-persistent patients in our study, more than half of the patients were younger than 50 years old (35.5%) and older than 70 years (21.5%) at the age of diagnosis.

With respect to the age at diagnosis, the younger group (age < 50 years) and older group (age > 70 years) were associated with lower AHT persistence and adherence, which is consistent with previous studies on women with breast cancer (Hsieh et al., 2014/Hsieh et al., 2015; Makubate et al., 2013; Moon et al., 2017; Pourcelot et al., 2018; Robinson et al., 2018). Moreover, the percentage of patients older than 70 years (14%) in our study was higher than that in a previous study in Taiwan (9%), which led to poorer medication intake for adherence and persistence (Hsieh et al., 2014). Therefore, AHT medication adherence and persistence in breast cancer among special groups (younger and older groups) are important management issues for the healthcare system and its professionals.

In our study, three popular classifiers, namely MLR, DT and ANN, were used to construct the MTB prediction model (Kumar & Lenina, 2016). The differences in predictive performance between the models were measured by comparing the AUCs using MATLAB R2013a software. Hosmer et al., (2013) reported that an AUC ≥ 0.8, which is indicative of a good model. The AUC of all models in our study exceeded 0.9 and the classification accuracy rates reached 92.29% and higher (MLR 96.37%). The results showed that all three models achieved high classification performance. A similar study used breast cancer databases and applied various data-mining techniques (Kharya, 2012). The best classification accuracy obtained in previous study was 93.62%, which was less than that of our study (Kharya, 2012).

Among the three models, the MLR model was the most accurate in predicting MTB in people with breast cancer receiving AHT. This is in complete agreement with the results of Wang et al. (2013) and Chang and Liou (2008). Although the TCR database used in our study was formalized and standardized, the medical records contained some heterogeneous data and were incomplete, imprecise and lacked certain values, all of which could have impacted the results of the data-mining tools. Therefore, some data-related tasks remain to be addressed in future research, including data collection, data mining, and the development of predictive models. Based on the data mining technique and measures by the three different classifiers and the Borda count methods, we observed that the duration of AHT discontinuation and the duration of AHT use were the two most critical predictor variables for the three models. The critical predictor variables were consistent with those reported by Cramer et al. (2008). The extent and duration of the patient's medication strongly affects the effectiveness of the treatment and the survival rate (Cramer et al., 2008). Therefore, adherence and persistence with AHT, at least for the full 5-year course, should be encouraged among patients.

In all three models based on the Borda count method, the Borda count was ranked the most important risk factor twice and was among the first five important risk factors. The top five influencing factors were as follows: duration of the AHT discontinuation period, duration of AHT use, age at diagnosis, BMI and receipt of radiotherapy. First, the duration of AHT discontinuation can be defined as the non-persistence time (total number of days that exceeded the 90-day prescription refill period) for the 5-year treatment period. Longer durations of AHT discontinuation have been associated with a higher risk of mortality (Murphy et al., 2012; Paranjpe et al., 2019). In contrast, the duration of AHT use can be defined as adherence and persistence time, including AHT adherence (MPR > 80%) and persistence (90-day prescription refill period) from initiation to discontinuation. Longer durations of AHT use have been linked to reduced mortality (Clancy et al., 2020; Wassermann & Rosenberg, 2017).

Although the use of tamoxifen for AHT is highly effective, medication adherence and persistence in the case of tamoxifen were lower than that in the case of AIs in a younger group of patients (Hsieh et al., 2014/Hsieh et al., 2015; Makubate et al., 2013; Moon et al., 2017; Sella & Chodick, 2020). Similar to a previous large population-based study ($N = 4,178$), we found that young age (<45 years) and low BMI increased the risk of non-adherence and early discontinuation (Sella & Chodick, 2020). Additionally, previous studies showed that increasing BMI is a statistically significant predictor (OR: 1.35, 95%

**TABLE 2** Subjects' characteristics

| Characteristics | Total (%) | | Adherence (%) | | Non-adherence (%) | | Persistence (%) | | Non-persistence (%) | |
|---|---|---|---|---|---|---|---|---|---|---|
| Number of patients (%) | 385 | 100.0% | 272 | 70.6% | 113 | 29.4% | 292 | 75.8% | 93 | 24.2% |
| Sex | | | | | | | | | | |
| Female | 385 | 100.0% | 272 | 100.0% | 113 | 29.4% | 292 | 100.0% | 93 | 24.2% |
| Follow-up time (year) | | | | | | | | | | |
| Total (patient-year) | 1,374.8 | | 875.6 | | | | 1,128.5 | | | |
| Mean (*SD*) | 3.6 (2.3) | | 3.2 (2.0) | | | | 3.9 (2.4) | | | |
| Age of diagnosis | | | | | | | | | | |
| Mean (*SD*) | 55.1 (12.1) | | 55.1 (12.0) | | | | 54.5 (11.7) | | | |
| Age ranks (%) | | | | | | | | | | |
| <50 years old | 146 | 37.9% | 103 | 37.9% | 43 | 38.1% | 113 | 38.7% | 33 | 35.5% |
| 50–69 years old | 185 | 48.1% | 131 | 48.2% | 54 | 47.8% | 145 | 49.7% | 40 | 43.0% |
| ≧70 years old | 54 | 14.0% | 38 | 14.0% | 16 | 14.2% | 34 | 11.6% | 20 | 21.5% |
| CCI score (%) | | | | | | | | | | |
| 0 | 260 | 67.5% | 186 | 68.4% | | | 199 | 68.2% | | |
| 1 | 73 | 19.0% | 49 | 18.0% | | | 55 | 18.8% | | |
| 2 | 52 | 13.5% | 37 | 13.6% | | | 38 | 13.0% | | |
| HT utilization pattern | | | | | | | | | | |
| Tamoxifen only | 131 | 34.0% | 86 | 31.6% | | | 90 | 30.8% | | |
| Tamoxifen to AIs | 3 | 0.8% | 2 | 0.7% | | | 2 | 0.7% | | |
| AIs only | 190 | 49.4% | 146 | 53.7% | | | 151 | 51.7% | | |
| AIs to tamoxifen | 56 | 14.5% | 33 | 12.1% | | | 46 | 15.8% | | |
| Multiple switches | 5 | 1.3% | 5 | 1.8% | | | 3 | 1.0% | | |
| Mean (*SD*) | 3.1 (1.9) | | 3.1 (0.1) | | | | 3.4 (0.3) | | | |

**TABLE 3** Adherence and persistence results of the three models on the prediction set

| Model | Accuracy (%) | Specificity (%) | Sensitivity (%) | AUC |
|---|---|---|---|---|
| MLR | 96.37 | 96.75 | 96.12 | 1.00 |
| DT | 93.52 | 95.24 | 92.47 | 0.93 |
| ANN | 92.29 | 91.33 | 93.01 | 0.90 |

CI: 1.08-0.1.80 $p = .042$; Kann et al., 2014) and is positively associated with adherence and persistence to AHT (Nyrop et al.,2016). However, patients at the extremes of age—with a risk of cancer recurrence, side effects from drugs, and a lower CCI score—have the highest risk of treatment interruptions and non-adherence (Clancy et al., 2020; Hsieh et al., 2014; Sella & Chodick, 2020). Xu et al. (2020) pointed out that when younger women suffered severe side effects and did not have enough family support or information, they opted to discontinue AHT or seriously considered stopping it. Our results are consistent with other studies showing that medication adherence and persistence in the younger group were initially high, and patients became less adherent or stopped medication as time progressed and severe adverse effects occurred (Cahir et al., 2015; Moon et al., 2017; Robinson et al., 2018; Sella & Chodick, 2020). Thus, switching the use of tamoxifen and AI is safer than the refusal or premature discontinuation of AHT (Pan et al., 2017).

Vergbrugghe et al. (2015) reported that experience with previous treatment (such as surgery, chemotherapy, targeted therapy and radiotherapy) impacted AHT adherence and persistence in women with breast cancer. Generally, most participants felt satisfied when physicians discussed their experience of symptoms at length, provided social support and helped them about the impact of the AHT (Verbrugghe et al., 2015). These results of clinical and demographic variables using data mining are consistent with those of Moon et al. (2017), who reported clinical variables (e.g. receipt of radiotherapy, AHT withdrawal period, actual AHT use) and demographic variables (e.g. younger than 50 years, BMI). For these clinical variables with the AHT use method, the side effects showed a negative relationship with adherence and persistence. The demographic variables with age (younger group under the age of 40 or 45/50 years and older group aged more than 65/75 years) had low odds of adherence and persistence. Therefore, data mining provides a logical and actionable method for extracting useful knowledge from medical databases. However, these new factors must be verified and certified in future research.

**TABLE 4** Ranking results of importance of the predictive variables

| Rank/Methods | MLR | DT | ANN | Borda count (Overall) |
|---|---|---|---|---|
| 1 | Duration of AHT discontinuation | Duration of AHT discontinuation | Duration of AHT discontinuation | Duration of AHT discontinuation |
| 2 | Duration of AHT use | Duration of AHT use | Duration of AHT use | Duration of AHT use |
| 3 | Age at diagnosis | Vitality status | Regional lymph nodes positive | Age at diagnosis |
| 4 | CCI | Age at diagnosis | Cause of death | BMI |
| 5 | Type of first recurrence | BMI | Tumour size | Receipt of radiotherapy |
| 6 | Sequence of locoregional therapy and systemic therapy | Receipt of radiotherapy | Number of sentinel lymph nodes positive | Vitality status |
| 7 | Surgical margins of the primary site | Laterality | Number of sentinel lymph nodes examined | Regional lymph nodes positive |
| 8 | Receipt of radiotherapy | Date of surgical diagnostic and staging procedure | BMI | CCI |
| 9 | Reason for no surgery of the primary site | Smoking behaviour | Duration between the first course of treatment date and the initial diagnosis date | Cause of death |
| 10 | Reason for no RT | | Regional lymph nodes examined | Type of first recurrence |

### 5.1 | Study limitations

Despite the number of participants (385) involved in our study and the accuracy of the three adopted predictive models, it had some limitations. First, we found that information on adverse effects or severe side effects were not collected or were incomplete in the medical records. Second, owing to the relatively small size of the database employed in our study, ANN did not obtain better predictions than the traditional statistical method (MLR). ANN is more appropriate for using a larger database to identify data attributes, and it has a better ability to manipulate data with a time sequence than DT (Delen et al., 2005; Liu et al., 2018; Sakr et al., 2017). Third, prescription refilling was the most common objective measurement of AHT adherence and persistence. However, the use of this measure remains limited because it cannot be known with certainty that the patient actually took the AHT medication. Therefore, it is strongly suggested to use a smart pillbox to remind patients about taking medicines and monitoring medicine intake in further research (Minaam & Abd-ELfattah, 2018). Finally, as we used secondary data analysis and a retrospective study design with monitoring of patients through five years of AHT, we could not collect complete information about the patients. In particular, the unavailability of psychosocial variables (such as motivation, knowledge and patient-healthcare provider communication) prevented a more advanced analysis.

### 5.2 | Clinical implications

The present study revealed that the implications of the five identified influencing factors can be drawn from data mining. In addition, MTB is a complex and dynamic process in clinical practice. The four main types of MTBs include acceptance/persistence, bearing/suffering, hesitation/adjustment and refusing/abandoning. In the switch from one type to another with a change in the patient's situation, healthcare providers should be conscious of individual preferences, the patient's level of motivation and knowledge and the implemented interventions to help the patient accomplish AHT (Vlasnik et al., 2005; Xu et al., 2020). Therefore, health professionals should develop a set of effective strategies to improve MTB among younger and older patients. To build effective MTB strategies, we suggest applying the case management adherence guidelines and combining information technology products (e.g. reminder app or LINE, phone calls and smart pill boxes) to provide accurate and objective measurements of AHT adherence and persistence in women with breast cancer for the full 5-year treatment period (Robinson et al., 2018).

## 6 | CONCLUSION

In our study, the MLR was shown to be the best among the three classifier models used in the small database. By applying a data mining technique to extract relevant knowledge, the study identified new factors, such as duration of AHT discontinuation, AHT use duration, age at diagnosis, BMI and receipt of radiotherapy as factors affecting MTB. Effective management of AHT use duration and discontinuation issues can help patients receive AHT for the full 5-year period. In terms of age at diagnosis, greater attention should be given to younger (<50 years old) and older (>70 years old) people with breast cancer because of their low rates of AHT adherence and persistence, respectively. Moreover, we should consider implementing therapeutic education programmes for people with breast cancer to strengthen their AHT MTB and possibly improve their adherence and persistence for the full 5-year treatment period.

## ACKNOWLEDGEMENTS

This research was supported by St. Martin De Porres Hospital: IRB No: 17B-029. We would like to thank professor Tung-Hsu Hou for providing MATLAB R2013a software with the data-mining techniques and algorithms employed in this study.

## ETHICAL STATEMENT

This study was approved by the ethics committee of St. Martin De Porres Hospital.

## CONFLICT OF INTERESTS

The authors have no conflicts of interest to disclose.

## AUTHOR CONTRIBUTIONS

Kuo, Wang, and Tseng were involved in study concept and design. Kuo was involved in the acquisition of data, analysis, interpretation of data and drafting of the manuscript. Wang was involved in revising it critically for important intellectual content and gave final approval of the version to be published. Tseng was involved in the analysis, interpretation of data and drafting of the manuscript. Hou was involved in analysis and provided MATLAB R2013a software for the data-mining techniques and algorithms. Each author participated sufficiently in the work to take public responsibility for appropriate portions of the content.

## DATA AVAILABILITY STATEMENT

Datasets that are restricted and not publicly available. Due to confidentiality agreements, the data and the source of the data can only be made available to bona fide researcher's subject to a non-disclosure agreement.

## ORCID

*Chen-Chen Kuo* https://orcid.org/0000-0002-0709-9141
*Hsiu-Hung Wang* https://orcid.org/0000-0001-6055-5401
*Li-Ping Tseng* https://orcid.org/0000-0001-5162-4364

## REFERENCES

American Society on Aging (ASA) and the American Society of Consultant Pharmacists (ASCP) Foundation (2006). *Overview Medication Adherence– Where Are We Today?*. http://adultmeducation.com/OverviewofMedicationAdherence.html

Bhatta, S. S., Hou, N., Moton, Z. N., Polite, B. N., Fleming, G. F., Olopade, O. I., Huo, D., & Hong, S. (2013). Factors associated with compliance to adjuvant hormone therapy in black and white women with breast cancer. *SpringerPlus*, *2*, 356. https://doi.org/10.1186/2193-1801-2-356

Brown, M. T., & Bussell, J. K. (2011). Medication adherence: WHO cares? *Mayo Clinic Proceedings*, *86*(4), 304–314.

Cahir, C., Guinan, E., Dombrowski, S. U., Sharp, L., & Bennett, K. (2015). Identifying the determinants of adjuvant hormonal therapy medication taking behaviour in women with stages I-III breast cancer: A systematic review and meta-analysis. *Patient Education and Counseling*, *98*(12), 1524–1539. https://doi.org/10.1016/j.pec.2015.05.013

Chang, W. P., & Liou, D. M. (2008). Comparison of three data mining techniques with genetic algorithm in analysis of breast cancer data. *Journal of Telemedicine and Telecare*, *9*(1), 26. https://scholar.google.com/scholar?hl=zh-TW&as_sdt=0%2C5&q=Comparison+of+three+Data+Mining+techniques+with+Genetic+Algorithm+in+analysis+of+Breast+Cancer+data&btnG

Chuang, M. T., Hu, Y. H., & Lo, C. L. (2016). Predicting the prolonged length of stay of general surgery patients: A supervised learning approach. *International Transactions in Operational Research*, *25*(1), 1–16. https://doi.org/10.1111/itor.12298

Çığşar, B., & Ünal, D. (2019). Comparison of data mining classification algorithms determining the default risk. *Scientific Programming*, *2019*, 1–8. https://doi.org/10.1155/2019/8706505

Clancy, C., Lynch, J., OConnor, P., & Dowling, M. (2020). Breast cancer patients' experiences of adherence and persistence to oral endocrine therapy: A qualitative evidence synthesis. *European Journal of Oncology Nursing*, *44*, 101706. https://doi.org/10.1016/j.ejon.2019.101706

Cramer, J. A., Roy, A., Burrell, A., Fairchild, C. J., Fuldeore, M. J., Ollendorf, D. A., & Wong, P. K. (2008). Medication compliance and persistence: Terminology and definitions. *Value in Health*, *11*(1), 44–47. https://doi.org/10.1111/j.1524-4733.2007.00213.x

Davies, C., Pan, H., Godwin, J., Gray, R., Arriagada, R., Raina, V., Abraham, M., Alencar, V. H. M., Badran, A., Bonfill, X., Bradbury, J., Clarke, M., Collins, R., Davis, S. R., Delmestri, A., Forbes, J. F., Haddad, P., Hou, M. F., Inbar, M., … Adjuvant Tamoxifen: Longer Against Shorter (ATLAS) Collaborative Group (2013). Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of estrogen receptor-positive breast cancer: ATLAS, a randomised trial. *Lancet (London, England)*, *381*(9869), 805–816. https://doi.org/10.1016/S0140-6736(12)61963-1

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127. https://doi.org/10.1016/j.artmed.2004.07.002

Duggan, M. A., Anderson, W. F., Altekruse, S., Penberthy, L., & Sherman, M. E. (2016). The Surveillance, Epidemiology and End Results (SEER) program and pathology: Towards strengthening the critical relationship. *The American Journal of Surgical Pathology*, *40*(12), e94–e102. https://doi.org/10.1097/pas.0000000000000749

Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An overview of the randomised trials. *Lancet (London, England)*, *365*(9472), 1687–1717. https://doi.org/10.1016/S0140-6736(05)66544-0

Ghorbani, R., & Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. *International Journal of Data and Network Science*, *3*(2019), 47–70. https://doi.org/10.5267/j.ijdns.2019.1.003

Guedes, J. B. R., Guerra, M. R., Alvim, M. M., & Leite, I. C. G. (2017). Factors associated with adherence and persistence to hormonal therapy in women with breast cancer. *Revista Brasileira De Epidemiologia*, *20*(4), 636–649. https://doi.org/10.1590/1980-5497201700040007

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Hsieh, K. P., Chen, L. C., Cheung, K. L., Chang, C. S., & Yang, Y. H. (2014). Interruption and non-adherence to long-term adjuvant hormone therapy is associated with adverse survival outcome of breast cancer women—an Asian population-based study. *PLoS One*, *9*(2), e87027. https://doi.org/10.1371/journal.pone.0087027

Hsieh, K. P., Chen, L. C., Cheung, K. L., & Yang, Y. H. (2015). A competing risk analysis of hormone therapy interruption in Asian women with breast cancer. *Pharmacoepidemiology and Drug Safety*, *24*(3), 301–309. https://doi.org/10.1002/pds.3733

Huang, C. C., Chan, S. Y., Lee, W. C., Chiang, C. J., Lu, T. P., & Cheng, S. H. (2019). Development of a prediction model for breast cancer based on the national cancer registry in Taiwan. *Breast Cancer Research*, *21*, 92. https://doi.org/10.1186/s13058-019-1172-6

Jacobs, M. S., Schouten, J. F., de Boer, P. T., Hoffmann, M., Levin, L. Å., & Postma, M. J. (2018). Secondary adherence to non-vitamin-K antagonist oral anticoagulants in patients with atrial fibrillation in Sweden and the Netherlands. *Current Medical Research and Opinion*, *34*(10), 1839–1847. https://doi.org/10.1080/03007995.2018.1459528

Jajroudi, M., Baniasadi, T., Kamkar, L., Arbabi, F., Sanei, M., & Ahmadzade, M. (2014). Prediction of survival in thyroid cancer using data mining technique. *Technology in Cancer Research & Treatment*, *13*(4), 353–359. https://doi.org/10.7785/tcrt.2012.500384

Kann, S., Schmid, S. M., Eichholzer, M., Huang, D. J., Amann, E., & Güth, U. (2014). The impact of overweight and obesity on breast cancer: Data from Switzerland, so far a country little affected by the current global obesity epidemic. *Gland Surgery*, *3*(3), 181–197. https://doi.org/10.3978/j.issn.2227-684x.2013.12.01

Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. *International Journal of Computer Science, Engineering and Information Technology*, *2*(2), 55–66. https://doi.org/10.5121/ijcseit.2012.2206

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Kumar, S. S., & **Lenina**, S. V. B. (2016). *MATLAB: Easy way of learning*. PHI Learning.

Lam, W. Y., & Fresco, P. (2015). Medication adherence measures: An overview. *BioMed Research International*, *2015*, 217047. https://doi.org/10.1155/2015/217047

Liu, M. M., Wen, L., Liu, Y. J., Cai, Q., Li, L. T., & Cai, Y. M. (2018). Application of data mining methods to improve screening for the risk of early gastric cancer. *BMC Medical Informatics and Decision Making*, *18*(Suppl 5), 121. https://doi.org/10.1186/s12911-018-0689-4

Makubate, B., Donnan, P. T., Dewar, J. A., Thompson, A. M., & McCowan, C. (2013). Cohort study of adherence to adjuvant endocrine therapy, breast cancer recurrence and mortality. *British Journal of Cancer*, *108*, 1515–1524. https://doi.org/10.1038/bjc.2013.116

Minaam, D. S. A., & Abd-ELfattah, M. (2018). Smart drugs: Improving healthcare using Smart Pill Box for medicine reminder and monitoring system. *Future Computing and Informatics Journal*, *3*(2), 443–456. https://doi.org/10.1016/j.fcij.2018.11.008

Mohan, A., Wanat, M. A., & Abughosh, S. M. (2019). Medication taking behaviors in patients taking warfarin versus direct oral anticoagulants: A systematic review. *Expert Review of Cardiovascular Therapy*, *17*(6), 427–434. https://doi.org/10.1080/14779072.2019.1620600

Moon, Z., Moss-Morris, R., Hunter, M. S., Carlisle, S., & Hughes, L. D. (2017). Barriers and facilitators of adjuvant hormone therapy adherence and persistence in women with breast cancer: A systematic review. *Patient Prefer Adherence*, *11*, 305–322. https://doi.org/10.2147/PPA.S126651

Murphy, C. C., Bartholomew, L. K., Carpentier, M. Y., Bluethmann, S. M., & Vernon, S. W. (2012). Adherence to adjuvant hormonal therapy among breast cancer survivors in clinical practice: A systematic review. *Breast Cancer Research and Treatment*, *134*(2), 459–478. https://doi.org/10.1007/s10549-012-2114-5

National Comprehensive Cancer Network (2020). *NCCN guidelines for treatment of cancer by site: Breast cancer*. https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1419

National Health and Medical Research Council. (2007). (Updated 2018). *National statement of ethical conduct in human research* [Online]. Canberra: Commonwealth of Australia. https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018

Nekhlyudov, L., Li, L., Ross-Degnan, D., & Wagner, A. K. (2011). Five-year patterns of adjuvant hormonal therapy use, persistence, and adherence among insured women with early-stage breast cancer. *Breast Cancer Research and Treatment*, *130*, 681–689. https://doi.org/10.1007/s10549-011-1703-z

Nyrop, K. A., Williams, G. R., Muss, H. B., & Shachar, S. S. (2016). Weight gain during adjuvant endocrine treatment for early-stage breast cancer: What is the evidence? *Breast Cancer Research and Treatment*, *158*(2), 203–217. https://doi.org/10.1007/s10549-016-3874-0

Oskouei, R. J., Kor, N. M., & Maleki, S. A. (2017). Data mining and medical world: Breast cancers' diagnosis, treatment, prognosis and challenges. *American Journal of Cancer Research*, *7*(3), 610–627.

Osterberg, L., & Blaschke, T. (2005). Adherence to medication. *The New England Journal of Medicine*, *353*, 487–497. https://doi.org/10.1056/nejmra050100

Pan, H., Gray, R., Braybrooke, J., Davies, C., Taylor, C., McGale, P., Peto, R., Pritchard, K. I., Bergh, J., Dowsett, M., Hayes, D. F., & EBCTCG. (2017). 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *New England Journal of Medicine*, *377*(19), 1836–1846. https://doi.org/10.1056/nejmoa1701830

Paranjpe, R., John, G., Trivedi, M., & Abughosh, S. (2019). Identifying adherence barriers to oral endocrine therapy among breast cancer survivors. *Breast Cancer Research and Treatment*, *174*(2), 297–305. https://doi.org/10.1007/s10549-018-05073-z

Partridge, A. H., LaFountain, A., Mayer, E., Taylor, B. S., Winer, E., & Asnis-Alibozek, A. (2008). Adherence to initial adjuvant anastrozole therapy among women with early-stage breast cancer. *Journal of Clinical Oncology*, *26*(4), 556–562. https://doi.org/10.1200/jco.2007.11.5451

Pourcelot, C., Orillard, E., Nallet, G., Dirand, C., Billion-Rey, F., Barbier, G., Chouk, S., Limat, S., Montcuquet, P., Henriques, J., Paget-Bailly, S., Anota, A., Chaigneau, L., & Nerich, V. (2018). Adjuvant hormonal therapy for early breast cancer: An epidemiologic study of medication adherence. *Breast Cancer Research and Treatment*, *169*(1), 153–162. https://doi.org/10.1007/s10549-018-4676-3

Pourhoseingholi, M. A., Kheirian, S., & Zali, M. R. (2017). Comparison of basic and ensemble data mining methods in predicting 5-year survival of colorectal cancer patients. *Acta Informatica Medica*, *25*(4), 254–258. https://doi.org/10.5455/aim.2017.25.254-258

Robinson, B., Dijkstra, B., Davey, V., Tomlinson, S., & Frampton, C. (2018). Adherence to adjuvant endocrine therapy in Christchurch women with early breast cancer. *Clinical Oncology (The Royal College of Radiologists)*, *30*(1), e9–e15. https://doi.org/10.1016/j.clon.2017.10.015

Sakr, S., Elshawi, R., Ahmed, A. M., Qureshi, W. T., Brawner, C. A., Keteyian, S. J., Blaha, M. J., & Al-Mallah, M. H. (2017). Comparison of machine learning techniques to predict all-cause mortality using fitness data: Henry Ford exercise testing (FIT) project. *BMC Medical Informatics and Decision Making*, *17*(1), 174. https://doi.org/10.1186/s12911-017-0566-6

Sella, T., & Chodick, G. (2020). Adherence and persistence to adjuvant hormonal therapy in early-stage breast cancer patients: A population-based retrospective cohort study in Israel. *Breast Care*, *15*, 45–53. https://doi.org/10.1159/000500318

Sondhi, M. D. (2017). Application of data mining in census data analysis using Weka. *International Journal of Engineering Trends and Technology*, *52*(3), 157–161. https://doi.org/10.14445/22315381/IJETT-V52P224

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, *2*(2), 250–255. https://www.semanticscholar.org/paper/Applications-of-Data-Mining-Techniques-in-and-of-Kasikumar-MohamedNajumuddeen/309bc44f32a24044e4ccac1722f96abd2db87ec9?p2df

Tseng, C. J., Lu, C. J., Chang, C. C., Chena, G. D., & Cheewakriangkrai, C. (2017). Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. *Artificial Intelligence in Medicine*, *78*, 47–54. https://doi.org/10.1016/j.artmed.2017.06.003

Vergbrugghe, M., Verhaeghe, S., Decoene, E., De Baere, S., Vandendorpe, B., & Van Hecke, A. (2015). Factors influencing the process of medication (non-) adherence and (non-) persistence in breast cancer patients with adjuvant antihormonal therapy: A qualitative study. *European Journal of Cancer Care*, *26*(2), e12339. https://doi.org/10.1111/ecc.12339

Vlasnik, J. J., Aliotta, S. L., & DeLor, B. (2005). Using case management guidelines to enhance adherence to long-term therapy. *Case Manager*, *16*(3), 83–85. https://doi.org/10.1016/j.casemgr.2005.02.001

Wang, K. M., Makond, B. J., Wu, W. L., Wang, K. J., & Lin, Y. S. (2013). Optimal data mining method for predicting breast cancer survivability. *International Journal of Innovative Management, Information & Production*, *4*(2), 28–33. https://www.semanticscholar.org/paper/OPTIMAL-DATA-MINING-METHOD-FOR-PREDICTING-BREAST-Wang-Makond/043e84748065c9038e29faaff173130ed3cc95e4

Wang, Z., Feng, F., Zhou, X., Duan, L., Wang, J., Wu, Y., & Wang, N. (2017). Development of diagnostic model of lung cancer based on multiple tumor markers and data mining. *Oncotarget*, *8*(55), 94793–94804. https://doi.org/10.18632/oncotarget.21935

Wassermann, J., & Rosenberg, S. M. (2017). Treatment decisions and adherence to adjuvant endocrine therapy in breast cancer. *Current Breast Cancer Reports*, *9*(2), 100–110. https://doi.org/10.1007/s12609-017-0248-5

World Health Organization (2003). *Adherence to long-term therapies: Evidence for action / [edited by Eduardo Sabaté]*. World Health Organization. https://apps.who.int/iris/handle/10665/42682

World Health Organization (2019). *Breast cancer*. https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

Xu, H., Zhang, X. J., Wang, D. Q., Xu, L., & Wang, A. P. (2020). Factors influencing medication-taking behaviour with adjuvant endocrine therapy in women with breast cancer: A qualitative systematic review. *Journal of Advanced Nursing*, *76*(2), 445–458. https://doi.org/10.1111/jan.14253

Yu, T., He, Z., Zhou, Q., Ma, J., & Wei, L. (2015). Analysis of the factors influencing lung cancer hospitalization expenses using data mining. *Thoracic Cancer*, *6*(3), 338–345. https://doi.org/10.1111/1759-7714.12147

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *115*(5), 654–657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929

**How to cite this article:** Kuo, C.-C., Wang, H.-H., & Tseng, L.-P. (2022). Using data mining technology to predict medication-taking behaviour in women with breast cancer: A retrospective study. *Nursing Open*, 9, 2646–2656. https://doi.org/10.1002/nop2.963