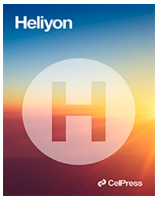
Heliyon 10 (2024) e38563

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

# Knowledge discovery of patients reviews on breast cancer drugs: Segmentation of side effects using machine learning techniques

Mehrbakhsh Nilashi [a,g,*], Hossein Ahmadi [b], Rabab Ali Abumalloh [c], Mesfer Alrizq [d,f], Abdullah Alghamdi [d,f], Sultan Alyami [e,f]

[a] UCSI Graduate Business School, UCSI University, 56000, Cheras, Kuala Lumpur, Malaysia
[b] Centre for Health Technology, Faculty of Health, University of Plymouth, Plymouth, PL4 8AA, UK
[c] Department of Computer Science and Engineering, Qatar University, Doha, 2713, Qatar
[d] Information Systems Dept., College of Computer Science and Information Systems, Najran University, Najran, Saudi Arabia
[e] Computer Science Dept., College of Computer Science and Information Systems, Najran University, Najran, Saudi Arabia
[f] AI Lab, Scientific and Engineering Research Center (SERC), Najran University, Najran, Saudi Arabia
[g] Centre for Business Informatics and Industrial Management, UCSI Graduate Business School, UCSI University, Malaysia

ARTICLE INFO

*Keywords:*
Knowledge discovery
Drugs
Online reviews
Breast cancer
Text mining
Machine learning
Public health

ABSTRACT

Breast cancer stands as the most frequently diagnosed life-threatening cancer among women worldwide. Understanding patients' drug experiences is essential to improving treatment strategies and outcomes. In this research, we conduct knowledge discovery on breast cancer drugs using patients' reviews. A new machine learning approach is developed by employing clustering, text mining and regression techniques. We first use Latent Dirichlet Allocation (LDA) technique to discover the main aspects of patients' experiences from the patients' reviews on breast cancer drugs. We also use Expectation-Maximization (EM) algorithm to segment the data based on patients' overall satisfaction. We then use the Forward Entry Regression technique to find the relationship between aspects of patients' experiences and drug's effectiveness in each segment. The textual reviews analysis on breast cancer drugs found 8 main side effects: Musculoskeletal Effects, Menopausal Effects, Dermatological Effects, Metabolic Effects, Gastrointestinal Effects, Neurological and Cognitive Effects, Respiratory Effects and Cardiovascular. The results are provided and discussed. The findings of this study are expected to offer valuable insights and practical guidance for prospective patients, aiding them in making informed decisions regarding breast cancer drug consumption.

## 1. Introduction

In the United States, breast cancer is both widespread and the second most common cause of cancer-related deaths among women, with lung cancer being the leading cause [1]. Breast cancer represents a substantial risk to their well-being and longevity. There are different types of breast cancer, such as invasive ductal carcinoma, invasive lobular carcinoma, and rarer subtypes like inflammatory breast cancer and triple-negative breast cancer [2,3]. The prevalence of breast cancer is on the rise [4,5], with more than 2 million new cases reported annually. In 2018, it constituted 11.6 % of all cancer instances and 24.2 % of cases among women, making it the most

* Corresponding author. UCSI Graduate Business School, UCSI University, 56000, Cheras, Kuala Lumpur, Malaysia.
*E-mail address:* mehrbakhsh@ucsiuniversity.edu.my (M. Nilashi).

https://doi.org/10.1016/j.heliyon.2024.e38563
Received 10 April 2024; Received in revised form 30 August 2024; Accepted 26 September 2024
Available online 26 September 2024

commonly occurring cancer and the leading cause of female mortality worldwide at 6.6 % [6]. In addition, in 2024, it is estimated that 42,250 women and 530 men will die of breast cancer [7]. Additionally, according to Ref. [7], despite having a 4 % lower incidence rate, Black women experience a 41 % higher mortality rate from breast cancer compared to White women. Risk factors associated with breast cancer comprise family history of the illness, age, genetic mutations like BRCA1 and BRCA2, lifestyle elements including alcohol intake, late menopause, and hormone replacement therapy, hormonal influences such as early onset of menstruation, lack of physical activity, and obesity, as well as exposure to ionizing radiation [3].

Breast cancer has several subtypes, with triple-negative breast cancer posing a particularly difficult prognosis due to the lack of effective treatment targets. Breast cancer treatment strategies and outcomes are primarily determined by its stage and subtype [8]. The prognosis for breast cancer is primarily determined by early tumor detection and prompt treatment [9]. Understanding breast cancer's development from preclinical biomarkers to mortality may help identify key markers for prevention and mortality reduction. Early detection of breast cancer can greatly improve outcomes [10,11]. There have been several methods such as screening mammography, clinical breast exams, and self-examinations for detection of breast cancer. When diagnosed, it is crucial to receive timely treatment. The specific interventions will be customized based on the type and stage of cancer to effectively manage the disease [3].

Estrogen receptor (ER) positivity is seen in approximately 80 % of breast cancers, suggesting that these tumors grow in response to estrogen. Hormone therapy, utilizing medications such as selective ER downregulators (SERDs) [12] and aromatase inhibitors (AIs) [13], is commonly prescribed for this cancer subtype. It is primarily administered to postmenopausal women and, to a lesser degree, to premenopausal women when coupled with a central blocking agent. These treatments work by either reducing estrogen production or promoting the degradation of its receptor [14,15]. Research is ongoing to find effective breast cancer treatments and develop new cancer prevention drugs, although drugs design and development can be quite expensive, time-consuming, and demanding [3,16].

Breast cancer treatment involves a range of individualized therapies and medications. Chemo, hormone, targeted, immuno, and preventative treatments are all part of the package [17,18]. Chemotherapy is the process of killing cancer cells through various mechanisms by administering a variety of drugs, such as capecitabine [19], cyclophosphamide [20], docetaxel [21], and others. Treatment with hormones inhibits either the body's natural hormone production or their ability to kill cancer cells [22]. Hormone therapy may also be used in combination with other treatments like chemotherapy or targeted therapy for optimal results in breast cancer management. Aromatase inhibitors such as anastrozole [23] and letrozole [24], fulvestrant [25], and tamoxifen [26] are common examples. Aromatase inhibitors are often used as adjuvant therapy following surgery or as first-line therapy for metastatic breast cancer. Cancer cells or other targets can be located and attacked with targeted therapies. Examples of targeted therapies include olaparib [27], tyrosine kinase inhibitors lapatinib [28] and neratinib [29], and monoclonal antibodies margetuximab [30] and trastuzumab [31]. Immunotherapy [32], which includes the use of immune checkpoint inhibitors like pembrolizumab, enhances the body's natural defenses against cancer. In high-risk individuals, preventive treatments that interfere with estrogen's effects, such as raloxifene [33] and tamoxifen [26], help reduce the risk of breast cancer. The choice of medication and therapy is ultimately dictated by factors such as cancer type, stage, and individual patient characteristics.

Community pharmacists are in a unique position to provide individualized counseling because of their frequent and direct interactions with patients [34]. The provision of drug-related information to patients and healthcare professionals is seen as crucial, with pharmacists playing a pivotal role in this aspect. In order to effectively carry out their role in patient care, it is expected that they remain informed about new therapies and emerging drug information. Accurate drug information is critical for improving patient outcomes, including symptom relief and avoiding medication errors. Patients rely on pharmacists for accurate medication information as the repercussions of misinformation can be significant. Access to reliable sources ensures the provision of relevant and easily applicable guidance. In addition to pharmacists, patients are also required to have access to a diverse range of dependable and up-to-date sources of medicine information. Web-based resources (e.g., drugs.com) can possess significant value when they provide precise and reliable information [35]. In addition, users can benefit from actual experiences and insights shared by individuals who have used specific medications. A unique perspective on drug effectiveness, side effects, ease of use, and overall satisfaction is offered by online reviews. By reading patient reviews, valuable insights into real-world experiences with medications can be gained, aiding individuals in making more informed decisions about their healthcare.

Patients are increasingly relying on online resources for medication and related health information as a result of the rapid growth of online review sites and discussion boards in medicine [36]. Patients can post unbiased reviews of medications and services they have personally used on review sites, allowing potential patients to make more informed decisions about which medications to take [37]. Online drug reviews typically include two sections: textual comments and ratings [38]. While ratings use a numerical scale to represent a customer's overall assessment, textual comments can provide more detailed information about the medication's specific side effects and usefulness than overall ratings. However, as the number of textual user comments grows by the day, prospective users are finding it increasingly difficult to read through each review before making a decision [39]. As a result, an effective, structured algorithm is required to analyze the reviews and categorize them into relevant features that can be used to inform prospective customers [40]. Given this, the primary goal of this study is to create a novel machine learning technique that utilizes online customer reviews to evaluate the efficacy and identify any potential side effects of prescription breast cancer drugs.

In this research, we propose a machine learning approach for knowledge discovery on breast cancer drugs using patients' reviews collected from drugs.com, which is a free online database powered by four independent leading medical-information suppliers: Cerner Multum, American Society of Health-System Pharmacists, Lexicomp, and Micromedex [41]. Understanding patients' drug experiences is essential to improving treatment strategies and outcomes. The proposed approach employs clustering, text mining, and regression techniques. Latent Dirichlet Allocation (LDA) technique [42] is initially utilized to discover the main topics from the patients' reviews on breast cancer drugs. Additionally, the Expectation-Maximization (EM) algorithm [43] is employed to segment the data based on patients' overall satisfaction. We then use the Forward Entry Regression technique [44,45] to find the relationship between discovered

topics and patients' overall satisfaction in each segment.

## 2. Related works

In [46], the authors examined human breast cancer cell lines and drug sensitivity data from Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE) databases to identify potential drugs. They selected top-ranked biomarkers and tested their radiation resistance using Cleveland database data. Six drugs were identified as effective on breast cancer cell lines. Additionally, five biomarkers showed sensitivity to both the shortlisted drugs and radiation. In Ref. [47], the authors tackled the high costs and risks of 'de novo' drug discovery by focusing on repurposing known drugs. They utilized large-scale data and a network-based integration approach to capture complex relationships among drugs, genes, and diseases. Using a network-based machine learning method, they identified potential therapeutic drugs for breast cancer subtypes, needing further clinical validation. In Ref. [48], the authors analyzed molecular descriptor data of compounds for breast cancer therapy using an ensemble learning algorithm. They selected important features for developing and validating Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) classification models. The process involved data cleaning, splitting, feature selection, and model evaluation. They proposed a Two-Level Stacking Algorithm (TLSA) for the classification of ADMET, reporting various performance measures. Results showed that TLSA with Logistic Regression excelled in Absorption, Distribution, and Excretion, while Support Vector Machine was best for Metabolism and Toxicity.

In [49], the authors developed MEDICASCY, a machine learning method that predicts indications, drug side effects, mode of action, and efficacy using only the chemical structure. MEDICASCY outperforms existing methods, showing 78 % precision and recall for severe side effects and 72 % precision for drug efficacy. It achieved nearly 80 % precision in predicting efficacy for cancer cell lines. MEDICASCY could enhance drug approval success rates and is accessible via MEDICASCY. In Ref. [50], the authors analyzed chemotherapy-related side effects using data from an oncology department in Scotland, focusing on breast cancer patients over three years. They compared several techniques—Markov model, Hidden Markov model, Recurrent Neural Network, and Random Forest —to predict treatment toxicity. In the study by Ref. [51], the authors aimed to detect subtle brain alterations in post-chemotherapy breast cancer patients using machine learning. Nineteen breast cancer patients and twenty healthy controls underwent resting-state functional Magnetic Resonance Imaging (MRI) and Generalized Q-Sampling Imaging (GQI). Machine learning models, including Logistic Regression (LR) with GQI indices, LR with mean regional homogeneity, Classification and Regression Tree (CART) with generalized fractional anisotropy, and XGBoost (XGB) with normalized quantitative anisotropy, effectively classified subjects into chemo-brain or healthy control groups. Leave-one-out cross-validation achieved the highest accuracy of 84 % with LR, CART, and XGB. The study's models show promise for clinically tracking brain changes due to chemotherapy.

In [52], to improve prediction of doxorubicin resistance, the authors evaluated 16 machine learning algorithms across 8 molecular profiles. Among 128 models, only 2 showed substantial predictive power. The best model, a CART, combined 4 miRNA isoforms, achieving a median Matthew Correlation Coefficient (MCC) of 0.56 and Area Under the Curve (AUC) of 0.80. In contrast, HER2 expression had lower predictive accuracy (median MCC of 0.14 and AUC of 0.57). The authors in Ref. [53] explored how variability in breast cancer responses and drug side effects relates to Cell Surface Receptor (CSR) transcriptional profiles. By comparing CSR expression in breast tumors and normal tissues, the research linked drug responses in cell lines to CSR levels. Significant differences in CSR expression were found between tumor subtypes and cell lines, influencing drug response predictions. Clinical data also showed correlations between CSR expression in healthy tissues and adverse drug reactions, aiding in the selection of optimal CSR targets for therapy. In Ref. [54], the authors used NLP (Natural Language Processing) to analyze social media data from breast cancer patients on X (Twitter). They developed a transformer-based classifier to identify relevant posts and a rule-based model to create a side effect lexicon and detect medication usage patterns. Their analysis revealed medication-related side effects and emotional distress, highlighting NLP's effectiveness in studying patient experiences through social media. In Ref. [55], a machine learning method was developed to categorize and forecast cancer drug combinations. The task entailed gathering and annotating data from the O'Neil drug interaction dataset, performing preprocessing on it, and dividing it into separate training and test sets. Classification models were constructed to classify drug combinations into the categories of synergistic, additive, or antagonistic, while regression models were used to forecast combination sensitivity scores. The framework identified successful combinations, such as kinase inhibitors with mTOR inhibitors or DNA damage-inducing drugs, specifically for ovarian, colorectal cancers, melanoma, lung, and prostate. Significant drugs that exhibited synergistic effects were Gemcitabine, MK-8776, and AZD1775. In Ref. [56], the authors investigated the effects of N-acetyl-D-glucosamine (D-GlcNAc) on breast cancer using cell assays and machine learning. Researchers cultured MCF-7 and 4T1 cell lines with D-GlcNAc and found that higher concentrations significantly reduced cell proliferation and increased apoptosis. They also used a xenograft mouse model, where D-GlcNAc administration led to reduced tumor size, mitosis, and angiogenesis. Molecular docking revealed D-GlcNAc's strong binding to HER2, suggesting its potential mechanism of action. These findings revealed that D-GlcNAc could be a promising candidate for breast cancer treatment.

## 3. Methods

**LDA for Text Mining:** This study used LDA for text mining [42]. This approach is widely used for topic modeling in previous studies. LDA uses probabilistic generative models which are used to model documents and words [57–59]. Assume that the indexes of the document are $d \in \{1, 2, \ldots, K\}$ for each topic, and $n \in \{1, 2, \ldots, N\}$ is the word indices in a document. According to Ref. [42], in LDA to generate the topics from the documents we have:

i. $N_d$ (the number of words) is distributed according to a Poisson distribution.
ii. $\theta_d$ (the model parameter) can be used to describe the proportions of the topic in a document per document $d$. The probability distribution over topics is represented by the letter $d$. It is assumed that the random variable $\theta_d$ comes from a Dirichlet distribution with prior parameter $\alpha$ and it is chosen on a random basis.
iii. Word proportions per topic $k$ can be identified by $\beta_k$. Each topic $\beta_k$ in the corpus is a probability distribution over a set of vocabulary that has been constructed. We assume that $\beta_k$ is a random variable. By considering a prior parameter $\eta$, this variable is selected from a Dirichlet distribution.
iv. In the document $z_{d,n}$, the topic of each word is randomly selected from a multinomial $\theta_d$ distribution, and from another multinomial distribution, the words $w_{d,n}$ are randomly selected to define $P(w_{d,n}|z_{d,n},\beta_k)$ terms.

The latent topical scheme was determined using two significant measures in LDA: $\theta_d$ and $\beta_k$. Researchers can determine the likelihood of a topic appearing in a particular text based on the value of $\theta_d$. Scholars can infer the most important terms that define a topic from the value of $\beta_k$. $\theta_d$ and $\beta_k$ are unknown variables in Bayesian statistics, and they are allocated to theoretically acceptable distributions, also called prior distributions. $\theta_d$ and $\beta_k$ are prior distributions in the LDA model, and they are defined using a probability distribution that is considered in the Dirichlet family of distributions. A prior parameter governs each of the two Dirichlet distributions, $\alpha$ for $\theta_d$ and $\eta$ for $\beta_k$.

**EM Clustering Using Gaussian Mixture Model (GMM):** GMM [60] is used for data clustering as a probabilistic technique. It utilizes the EM or Expectation and Maximization [61] to fit data points with the GMM model's parameters, entailing the mixing proportions of the Gaussian components, means, and covariance [62–64]. A parametric Gaussian distribution can be used to determine the boundaries of each segment [65]. Eq. (1) is a linear superposition of Gaussian components that can be used to outline the distribution of the entire dataset $f(x)$:

$$f(x)=\sum_{i=1}^{J} p_j f\left(x|\mu_j,\sum_j\right)=\sum_{i=1}^{J} p_j \frac{exp\left(-\frac{1}{2}\left(x-\mu_j\right)^T \Sigma_j^{-1}\left(x-\mu_j\right)\right)}{(2\pi)^{\frac{d}{2}}\left|\Sigma_j\right|^{\frac{1}{2}}} \tag{1}$$

where $\int_x f\left(x|\mu_j,\sum_j\right)dx=1$ and $\sum_{j=1}^{J} p_j = 1$. $J$ distributions from each point $x$ in $f(x)$ can be drawn from the $J$ model distributions by a probability of $p_j$ where $j=1$ , 2, $J$. To calculate unknown parameters the maximum-likelihood approach can be utilized. Through Eq. (2) and training samples $\{x_k\}_{k=1}^{N}$, $\theta=\left\{\mu_j, Z_j, p_j\right\}_{j=1}^{J}$ is calculated to maximize the log-likelihood function :

$$L=\sum_{k=1}^{N} log\left(\sum_{j1}^{J} p_j f\left(x_k\middle|\mu_j,\sum_j(0)\right)\right) \tag{2}$$

Still, in the data set that is used in the training process, there is a need for the information of labels, particularly with an incomplete dataset. The EM algorithm can address this issue effectively. To estimate the GMM parameters, the following steps are performed which are presented in Fig. 1 [64].

1. Initialization: Select the initial estimates $p_j(0)$, $\mu_j(0)$, $\Sigma_j(0)$, $j=1,2,$ $J$ and compute the initial log-likelihood by Eq. (3) as:

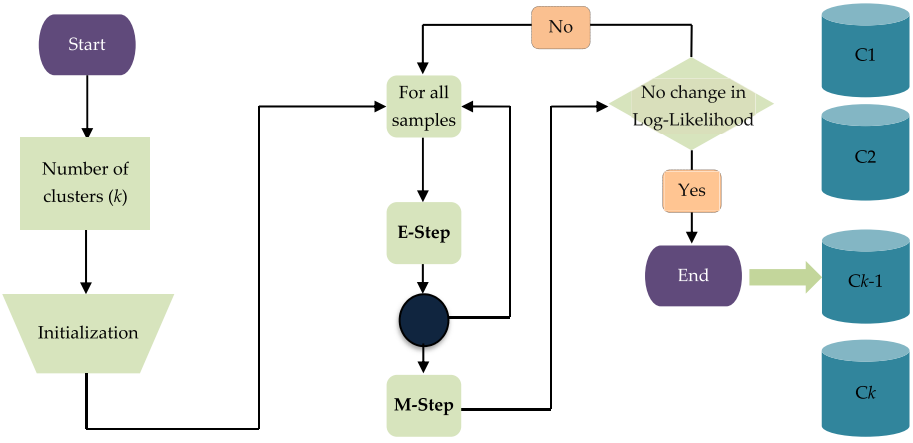



**Fig. 1.** Em clustering.

$$L(0) = \sum_{k=1}^{N} \log \left( \sum_{j=1}^{J} p_j(0) f\left(\mathrm{x}_k | \mu_j(0), \Sigma_j(0)\right) \right) \tag{3}$$

2. As shown in Eq. (4), in E-Step we have:

$$p(j|x_k, \theta(t)) = \frac{p_i(t) f\left(x_k | \mu_j(t), \Sigma_j(t)\right)}{\sum_{l=1}^{j} p_l(t) f\left(x_k | \mu_l(t), \sum_l(t)\right)}, j = 1, 2, \ldots, J, k = 1, 2, \ldots, N.. \tag{4}$$

3. As shown Eq. (5), in M-Step we have:

$$p_j(t+1) = \frac{1}{N} \sum_{k=1}^{N} p(j|x_k, \theta(t)) \tag{5}$$

## 4. Data analysis and results

A general framework of the proposed method is presented in Fig. 2. According to Fig. 2, patients' online reviews on breast cancer drugs are collected from drugs.com. The data is cleaned to keep high quality reviews to be analyzed by the LDA technique. Aspects of patients' experiences are discovered by the LDA technique and combined with the numerical ratings to perform data segmentation using EM. Drug effectiveness prediction is performed using Forward Entry Regression technique. Finally, the method is evaluated using R-squared ($R^2$). To determine the optimal number of topics for our Latent Dirichlet Allocation (LDA) model, we assessed the goodness-of-fit by comparing models with different topic counts. This evaluation involved calculating the perplexity of a separate set of documents, which gauges how effectively the model captures the content of the documents. A lower perplexity score indicates a better model fit. By analyzing perplexity scores across different topic numbers, we identified the number of topics (8 topics) that provided the best fit for our data. According to the LDA analysis, the topics were categorized in Musculoskeletal Effects, Menopausal Effects, Dermatological Effects, Metabolic Effects, Gastrointestinal Effects, Neurological/Cognitive Effects, Respiratory Effects and Cardiovascular Effects (see Table 1).

Musculoskeletal effects encompass a range of symptoms affecting the muscles, bones, and joints in individuals undergoing breast cancer drug therapy. These may include joint pain, stiffness, muscle weakness, and reduced range of motion. Some patients may experience conditions like osteoporosis or osteopenia, leading to increased fracture risk. Furthermore, Menopausal symptoms refer to the physiological changes experienced by individuals undergoing breast cancer drug treatment, often due to hormonal alterations. Common symptoms include hot flashes, night sweats, vaginal dryness, and mood swings. These symptoms can significantly impact

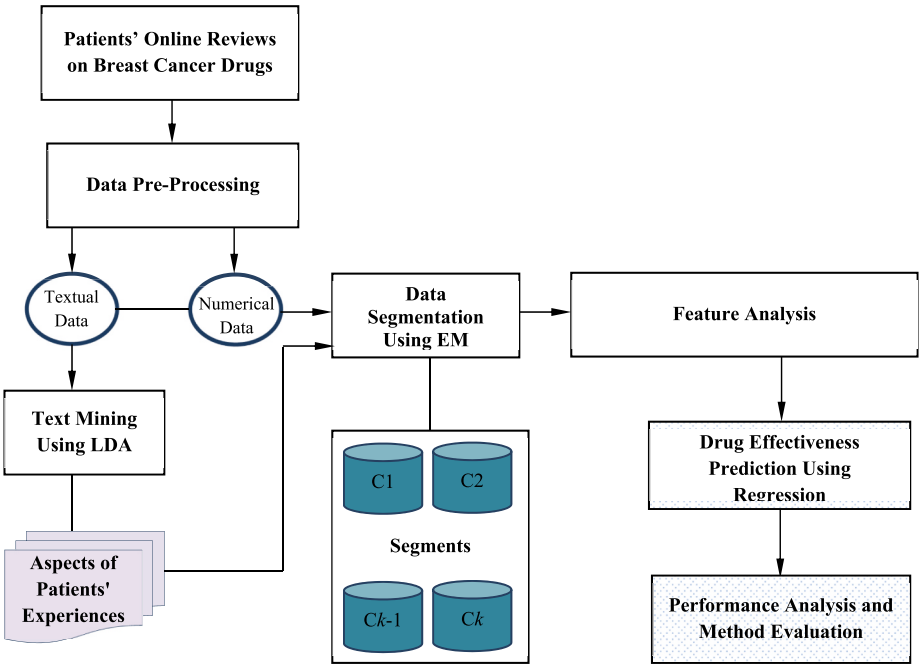



**Fig. 2.** Procedure of patients' reviews analysis for breast cancer drugs.

**Table 1**
An example of patient reviews on breast cancer drugs in drugs.com, side effects and group of side effects.

| Patient Review | Side Effect | Group |
|---|---|---|
| I was prescribed Letrozole by the Oncologist after breast cancer surgery. Exactly one year on this medication I started having arthritis, which progressed to full blown Rheumatoid Arthritis. The Oncologist refused to acknowledge a link, and insisted I stay on the medication for five years. Common sense told me that the medication interfered with my immune system. I will be on medication for the rest of my life, to control the RA, taking injections every three weeks. I am positive it is connected to the Letrozole. | Arthritis, Rheumatoid Arthritis | Musculoskeletal Effects |
| Stage 3 breast cancer. Double mastectomy, chemo, Radiation. Started letrozole about 5 months ago. Severe joint and muscle pain especially in shoulders, arms and hands. 20 lb weight gain. Seriously hate this medication but afraid to go off it because supposedly it reduces my recurrence rate. I'm trying acupuncture for the joint pains. | Severe joint and muscle pain, especially in shoulders, arms, and hands. Weight gain of 20 lbs | Musculoskeletal and Metabolic Effects |
| Started taking letrozole in June. Side effects started almost immediately. Fatique, not flashes, joint and muscle aches. Hair thinning. Have cut my hair very short to hopefully make the thinning less noticeable. I just realized that I cut my hair 5–6 weeks ago and it does not seem to be growing at all. Has anyone else noticed this side effect. I would recommend yoga and acupuncture to help with the side effects. Without them, I would not be able to tolerate this medicine! Good luck to you all who are taking this drug! | Fatigue, Hot flashes<br>Joint and muscle aches, Hair thinning, Delayed hair growth | Musculoskeletal and Dermatological Effects |

quality of life and may require management strategies such as hormone replacement therapy or lifestyle modifications. Dermatological effects (e.g., skin rashes, dryness, or changes in hair and nail health) pertain to the skin-related side effects associated with breast cancer drug therapy. Additionally, Metabolic effects involve alterations in metabolic processes and parameters observed in individuals receiving breast cancer drug therapy. They may involve weight gain, changes in cholesterol levels, or alterations in blood sugar regulation. Gastrointestinal effects (e.g., nausea, vomiting, diarrhea, or constipation) encompass a spectrum of digestive system-related symptoms experienced by individuals undergoing breast cancer drug treatment. Neurological and cognitive effects refer to changes in brain function and cognitive abilities observed in individuals undergoing breast cancer drug therapy. They may present memory problems, difficulty concentrating, or cognitive fog. Moreover, Respiratory effects involve complications affecting the respiratory system in individuals receiving breast cancer drug therapy. These may include dyspnea, cough, pulmonary fibrosis, and susceptibility to respiratory infections. Finally, Cardiovascular effects encompass changes in cardiovascular health and function observed in individuals undergoing breast cancer drug therapy. These effects may include hypertension, arrhythmias, thromboembolic events, and cardiomyopathy.

EM parameters, Clusters, Max Iteration, Trials, and Ridge were respectively set to 3, 50 and 0.001. Totally, 3 clusters were generated by EM clustering. 1-way ANOVA (Clusters vs. Input Attributes) is shown in Table 2.

BSS stands for Between-Cluster Sum of Squares, WSS stands for Within-Cluster Sum of Squares, and TSS stands for Total Sum of Squares. These metrics play a vital role in cluster analysis, particularly in evaluating the quality of clustering algorithms. BSS measures the total variance between clusters, indicating the extent to which the clusters are distinct and well-separated. In contrast, WSS quantifies the variance within clusters, reflecting how tightly grouped the data points are within each cluster. TSS represents the total variance in the dataset, encompassing both between-cluster and within-cluster variability.

In addition, in Table 3, the cluster centroids are presented. The cluster centroids represent the average values for each attribute across three segments. The attributes include musculoskeletal effects, menopausal symptoms, dermatological effects, metabolic effects, gastrointestinal effects, neurological and cognitive effects, respiratory effects, and cardiovascular effects. Each segment is characterized by its level of intensity for these attributes, ranging from low to moderate to high, based on the values ranging from 0 to 1.

Table 4 presents the results of a forward entry regression analysis in Segment 1, revealing coefficients along with their corresponding p-values in parentheses. Notably, negative coefficients signify an inverse relationship between predictor variables and outcomes which is drug effectiveness. Specifically, the Musculoskeletal Effects ($-9.102735$, $p = 0.000000$), Metabolic Effects ($-1.464585$, $p = 0.000000$), and Gastrointestinal Effects ($-0.959759$, $p = 0.000000$) display negative coefficients, indicating that higher values of these factors are associated with lower outcomes in Segment 1. Hence, Musculoskeletal Effects appear to exert the most significant influence on outcomes, followed by Metabolic and Gastrointestinal Effects.

In Table 5, the outcomes of a forward entry regression analysis within Segment 2 are outlined. Particularly, Respiratory Effects ($-4.033943$, $p = 0.000020$), Cardiovascular Effects ($-1.348199$, $p = 0.000009$), Musculoskeletal Effects ($-2.012207$, $p = 0.000000$), and Metabolic Effects ($-3.036903$, $p = 0.000676$) display negative coefficients which suggest that higher values of these attributes correspond to diminished outcomes in Segment 2. Among the effects in this table, Respiratory Effects exert the most substantial influence, followed by Metabolic, Musculoskeletal, and Cardiovascular Effects, respectively.

In Table 6, we provide the outcomes of a forward entry regression analysis performed within Segment 3. It is noteworthy that all coefficients display negativity which are: menopausal effects ($-4.046992$, $p = 0.000000$), dermatological effects ($-4.560264$, $p = 0.000000$), and neurological/cognitive effects ($-2.309966$, $p = 0.000000$). notably, dermatological effects stand out as the most influential factor, followed by menopausal and neurological/cognitive effects, respectively.

**Table 2**
1-way ANOVA results for EM clustering.

| Attribute_Y | Attribute_X | Description | | | | Statistical test | | |
|---|---|---|---|---|---|---|---|---|
| Musculoskeletal Effects | Clusters | | | | | Variance decomposition | | |
| | | Value | Examples | Average | Std-dev | Source | Sum of square | d.f. |
| | | **C1** | 155 | 0.5924 | 0.0999 | BSS | 32.8916 | 2 |
| | | **C2** | 183 | 0.1993 | 0.1277 | WSS | 4.8798 | 430 |
| | | **C3** | 95 | 0.8980 | 0.0633 | TSS | 37.7714 | 432 |
| | | **All** | 433 | 0.4933 | 0.2957 | **Significance level** | | |
| | | | | | | **Statistics** | **Value** | **Proba** |
| | | | | | | Fisher's F | 1449.171035 | 0.000000 |
| Menopausal Symptoms | Clusters | | | | | Variance decomposition | | |
| | | Value | Examples | Average | Std-dev | Source | Sum of square | d.f. |
| | | **C1** | 155 | 0.5600 | 0.2222 | BSS | 22.1906 | 2 |
| | | **C2** | 183 | 0.3073 | 0.1865 | WSS | 14.1843 | 430 |
| | | **C3** | 95 | 0.8996 | 0.0523 | TSS | 36.3748 | 432 |
| | | **All** | 433 | 0.5277 | 0.2902 | **Significance level** | | |
| | | | | | | **Statistics** | **Value** | **Proba** |
| | | | | | | Fisher's F | 336.356121 | 0.000000 |
| Dermatological Effects | Clusters | | | | | Variance decomposition | | |
| | | Value | Examples | Average | Std-dev | Source | Sum of square | d.f. |
| | | **C1** | 155 | 0.5704 | 0.1000 | BSS | 27.6861 | 2 |
| | | **C2** | 183 | 0.2239 | 0.1091 | WSS | 4.1991 | 430 |
| | | **C3** | 95 | 0.8701 | 0.0724 | TSS | 31.8851 | 432 |
| | | **All** | 433 | 0.4897 | 0.2717 | **Significance level** | | |
| | | | | | | **Value** | **Proba** | **Statistics** |
| | | | | | | 1417.575282 | 0.000000 | Fisher's F |
| Metabolic Effects | Clusters | | | | | Variance decomposition | | |
| | | Value | Examples | Average | Std-dev | Source | Sum of square | d.f. |
| | | **C1** | 155 | 0.5899 | 0.1328 | BSS | 32.2423 | 2 |
| | | **C2** | 183 | 0.2097 | 0.1396 | WSS | 6.6518 | 430 |
| | | **C3** | 95 | 0.9049 | 0.0644 | TSS | 38.8941 | 432 |
| | | **All** | 433 | 0.4983 | 0.3001 | **Significance level** | | |
| | | | | | | **Statistics** | **Value** | **Proba** |
| | | | | | | Fisher's F | 1042.141956 | 0.000000 |
| Gastrointestinal Effects | Clusters | | | | | Variance decomposition | | |
| | | Value | Examples | Average | Std-dev | Source | Sum of square | d.f. |
| | | **C1** | 155 | 0.6247 | 0.1088 | BSS | 5.1322 | 2 |
| | | **C2** | 183 | 0.3924 | 0.3235 | WSS | 30.2595 | 430 |
| | | **C3** | 95 | 0.5891 | 0.3160 | TSS | 35.3918 | 432 |
| | | **All** | 433 | 0.5187 | 0.2862 | **Significance level** | | |
| | | | | | | **Statistics** | **Value** | **Proba** |
| | | | | | | Fisher's F | 36.465658 | 0.000000 |
| Neurological/Cognitive Effects | Clusters | | | | | Variance decomposition | | |
| | | Value | Examples | Average | Std-dev | Source | Sum of square | d.f. |
| | | **C1** | 155 | 0.6325 | 0.1119 | BSS | 31.4769 | 2 |
| | | **C2** | 183 | 0.2206 | 0.1256 | WSS | 5.1531 | 430 |
| | | **C3** | 95 | 0.8917 | 0.0614 | TSS | 36.6300 | 432 |
| | | **All** | 433 | 0.5153 | 0.2912 | **Significance level** | | |
| | | | | | | **Statistics** | **Value** | **Proba** |
| | | | | | | Fisher's F | 1313.291942 | 0.000000 |
| Respiratory Effects | Clusters | | | | | Variance decomposition | | |
| | | Value | Examples | Average | Std-dev | Source | Sum of square | d.f. |
| | | **C1** | 155 | 0.5706 | 0.1276 | BSS | 28.8625 | 2 |
| | | **C2** | 183 | 0.2148 | 0.1512 | WSS | 7.1852 | 430 |
| | | **C3** | 95 | 0.8739 | 0.0744 | TSS | 36.0477 | 432 |
| | | **All** | 433 | 0.4868 | 0.2889 | **Significance level** | | |
| | | | | | | **Statistics** | **Value** | **Proba** |
| | | | | | | Fisher's F | 863.635522 | 0.000000 |
| Cardiovascular Effects | Clusters | | | | | Variance decomposition | | |
| | | Value | Examples | Average | Std-dev | Source | Sum of square | d.f. |
| | | **C1** | 155 | 0.4533 | 0.2009 | BSS | 21.6078 | 2 |
| | | **C2** | 183 | 0.3224 | 0.2431 | WSS | 17.3180 | 430 |
| | | **C3** | 95 | 0.9040 | 0.0607 | TSS | 38.9258 | 432 |
| | | **All** | 433 | 0.4969 | 0.3002 | **Significance level** | | |
| | | | | | | **Statistics** | **Value** | **Proba** |
| | | | | | | Fisher's F | 268.258031 | 0.000000 |

**Table 3**
Cluster centroids.

| Attribute | Segment 1 | Segment 2 | Segment 3 |
|---|---|---|---|
| Musculoskeletal Effects | 0.592380 | 0.199338 | 0.897961 |
| Menopausal Effects | 0.559954 | 0.307336 | 0.899648 |
| Dermatological Effects | 0.570424 | 0.223898 | 0.870100 |
| Metabolic Effects | 0.589905 | 0.209742 | 0.904878 |
| Gastrointestinal Effects | 0.624739 | 0.392423 | 0.589133 |
| Neurological/Cognitive Effects | 0.632523 | 0.220635 | 0.891676 |
| Respiratory Effects | 0.570615 | 0.214781 | 0.873862 |
| Cardiovascular Effects | 0.453346 | 0.322423 | 0.903981 |

**Table 4**
Selected effects by forward entry regression in Segment 1 (Sig. Level = 0.05; $R^2 = 0.9421$).

| Attribute | Coef. | std | p-value |
|---|---|---|---|
| Intercept | 10.904579 | 0.092831 | 0.000000 |
| Musculoskeletal Effects | −9.102735 | 0.200672 | 0.000000 |
| Metabolic Effects | −1.464585 | 0.187843 | 0.000000 |
| Gastrointestinal Effects | −0.959759 | 0.142127 | 0.000000 |
| *Drug Effectiveness* = 10.904579 − 9.102735 × *Musculoskeletal Effects* − 1.464585 × *Metabolic Effects* − 0.959759 × *Gastrointestinal Effects* | | | |

**Table 5**
Selected effects by forward entry regression in Segment 2 (Sig. Level = 0.05; $R^2 = 0.9464$).

| Attribute | Coef. | std | p-value |
|---|---|---|---|
| Intercept | 10.732682 | 0.058900 | 0.000000 |
| Respiratory Effects | −4.033943 | 0.911123 | 0.000020 |
| Cardiovascular Effects | −1.348199 | 0.292599 | 0.000009 |
| Musculoskeletal Effects | −2.012207 | 0.316542 | 0.000000 |
| Metabolic Effects | −3.036903 | 0.872232 | 0.000676 |
| *Drug Effectiveness* = 10.732682 − 4.033943 × *Respiratory Effects* − 1.348199 × *Cardiovascular Effects* − 2.012207 × *Musculoskeletal Effects* − 3.036903 × *Metabolic Effects* | | | |

**Table 6**
Selected effects by forward entry regression in Segment 3 (Sig. Level = 0.05; $R^2 = 0.9447$).

| Attribute | Coef. | std | p-value |
|---|---|---|---|
| Intercept | 10.595677 | 0.054707 | 0.000000 |
| Menopausal Effects | −4.046992 | 0.638233 | 0.000000 |
| Dermatological Effects | −4.560264 | 0.494864 | 0.000000 |
| Neurological/Cognitive Effects | −2.309966 | 0.363459 | 0.000000 |
| *Drug Effectiveness* = 10.595677 − 4.046992 × *Menopausal Effects* − 4.560264 × *Dermatological Effects* − 2.309966 × *Neurological/Cognitive Effects* | | | |

## 5. Findings and discussions

Our findings from analysis of reviews on breast cancer drugs suggest that breast cancer patients in Segment 1 are particularly concerned about musculoskeletal, metabolic, and gastrointestinal effects of treatment. In Segment 2, respiratory effects, cardiovascular effects, musculoskeletal effects and metabolic effects and in Segment 3, menopausal effects, dermatological effects, neurological and cognitive effects are the most concern in consuming breast cancer drugs. These potential side effects underscore the need for further investigation to elucidate connections and develop targeted interventions to manage adverse effects in breast cancer patients.

Previous research has investigated osteoporosis and musculoskeletal complications related to breast cancer therapy [66]. The authors highlighted that AIs cause musculoskeletal issues, including muscle pain (myalgias) and joint pain (arthralgias). The authors in Ref. [67] found that a substantial portion of single-agent Trastuzumab administrations resulted in gastrointestinal toxicities, such as vomiting, nausea, abdominal pain, and diarrhea. In Ref. [68], the authors investigated the effects of neoadjuvant chemotherapy on respiratory function in breast cancer patients. They found a significant increase in maximal ventilatory volume, with most patients experiencing dyspnea and fatigue. The authors in Ref. [69] carried out a mixed-methods study to analyze content related to AI-associated side effects posted on twelve message boards between 2002 and 2010. Out of a total of 25,256 posts concerning AIs, 4589 posts, accounting for 18.2 %, mentioned at least one potential side effect. The most cited side effects were joint/musculoskeletal pain, osteoporosis, hot flashes, and weight gain. The authors in Ref. [70] investigated central nervous system toxicity and cardiovascular caused by anticancer drugs in breast cancer patients. They highlighted that, despite improvements in breast cancer treatment,

anticancer drugs are associated with significant side effects, such as cardiotoxicity, which can lead to heart failure. In Ref. [71], the authors revealed that chemotherapy has the potential to induce cognitive impairments, which may be long-lasting or permanent. These cognitive deficits tend to be diffuse, affecting various domains of functioning, including attention and concentration, verbal and visual memory, and executive functioning. In Ref. [72], the authors conducted an analysis on the occurrence of skin toxicity resulting from drugs used in adjuvant and neoadjuvant chemotherapy in women with breast cancer. An evaluation was conducted on the medical records of 72 women who received therapy from 2003 to 2006. Out of the 558 cycles of chemotherapy, a total of 152 adverse events were recorded. 37 cases of dermatological toxicity were reported, with 20 instances of extravasations that affected 17 women. Throughout the course of neoadjuvant therapy, there were nine recorded instances of localized injury, fibrosis, pain, and hyperemia. Throughout the course of adjuvant therapy, there were 11 recorded cases of extravasations, which resulted in reports of fibrosis, hardened local injury, and local pain.

## 6. Conclusion

Drug reviews have served as a critical source of medical information for healthcare professionals and consumers alike. With the rising trend of online platforms, such as review sites, discussion boards, and forums, individuals are increasingly sharing their experiences and opinions about various drugs. These platforms have offered valuable insights and feedback, aiding both healthcare providers and patients in making informed decisions regarding treatment options. In this research, we accordingly developed a new machine learning approach employing clustering, text mining, and regression techniques. Initially, we utilized LDA to identify the main aspects of patients' experiences from their reviews on breast cancer drugs. Subsequently, we employed the EM algorithm to segment the data based on patients' overall satisfaction (drug effectiveness). Finally, we applied the Forward Entry Regression technique to ascertain the relationship between aspects of patients' experiences and drug effectiveness within each segment.

### *6.1. Managerial implications*

Breast cancer is a global health concern, it is the most common cancer among women in many countries [73,74]. While female gender is the primary risk factor, other factors like age, obesity, and alcohol use also contribute. Treatment modalities, including surgery, radiation therapy, and medications, offer hope, but their effectiveness hinges on timely intervention and adherence to full courses. Efforts to reduce breast cancer mortality are underway through initiatives like the WHO Global Breast Cancer Initiative, emphasizing health promotion, early detection, and comprehensive management. Strengthening health systems and fostering awareness are vital steps towards achieving this goal, ensuring that women worldwide receive timely care and support. The objective of the WHO Global Breast Cancer Initiative is to decrease worldwide cancer mortality by 2.5 % annually, thereby preventing 2.5 million death rates caused by breast cancer in women under the age of 70 from 2020 to 2040 [75]. The Global Breast Cancer Initiative is a result of the enduring dedication of breast cancer advocates on a global scale. It currently involves international collaborators in order to effectively coordinate endeavors aimed at promoting the advancement of breast cancer control.

When using medications for breast cancer, it is important to investigate any potential adverse effects as they have the potential to significantly influence the quality of life of patients as well as their adherence to treatment. The ability of healthcare providers to mitigate the impact of side effects and tailor treatment plans accordingly, which ultimately results in improved patient well-being and outcomes, is made possible by the understanding of the nature and severity of side effects as well. In addition, the investigation of adverse effects yields valuable information regarding the safety profiles of drugs. This information enables the identification of potential adverse reactions and the development of strategies for risk management.

Utilizing online patient reviews can prove to be a valuable and cost-effective method for identifying potential side effects linked to medications. Examining these reviews provides immediate insights into patients' experiences, allowing healthcare providers and pharmaceutical companies to quickly identify and resolve common side effects. This approach can offer significant advantages in enhancing drug safety profiles and fine-tuning treatment strategies. Through careful analysis of online feedback, managers and administrators in healthcare can improve decision-making, customize patient communications, and take proactive steps to address reported side effects. This will result in improved patient outcomes and more effective resource allocation. Thus, a reliable data collection platform is crucial for effectively utilizing patient reviews to identify drug side effects. This platform should enable the systematic collection of extensive data from various sources. The system should have sophisticated analytics capabilities to perform comprehensive text mining and sentiment analysis. This will enable the extraction of actionable insights from patient feedback. Moreover, the ability to personalize survey options is essential for capturing precise facets of patients' experiences with medications. The inclusion of real-time monitoring capabilities will guarantee that emerging trends and new reviews are promptly attended to. Seamless integration with pre-existing healthcare databases is crucial in order to offer a comprehensive perspective on patients' experiences.

### *6.2. Limitations and future work*

While this study illuminates critical insights from online reviews regarding breast cancer drugs, its limitations prompt avenues for future research. We acknowledge that patient reviews may contain unverified and subjective information. Although, we applied several pre-processing steps to filter out irrelevant or clearly erroneous data, future work may involve applying more advanced pre-processing steps to filter out irrelevant or clearly erroneous data. In fact, while our findings provide valuable insights, they should be interpreted with caution. Future research could enhance reliability by integrating patient reviews with clinical data and conducting

longitudinal studies to validate patient-reported outcomes. In addition, this study only used the patients reviews in drugs.com, thus, ensuring the generalizability of findings across diverse patient populations and drug regimens remains paramount which underscores the need for further patients' reviews. Future research could explore ways to address these limitations by incorporating diverse data sources, such as clinical records or patient surveys, to provide a more comprehensive understanding of patient experiences with breast cancer drugs. Furthermore, refining the machine learning techniques to enhance the interpretability and robustness of the results could further advance the field. Moreover, longitudinal studies could be conducted to assess the long-term effectiveness and side effects of breast cancer treatments, providing valuable insights into their real-world impact on patient outcomes. Finally, our method for knowledge discovery in detecting the side effects of breast cancer drugs using patients' reviews can be further developed using deep and machine learning-based algorithms [76] and generative artificial intelligence-based diagnostic algorithms [77,78].

## Data availability statement

Sharing research data helps other researchers evaluate your findings, build on your work and to increase trust in your article. We encourage all our authors to make as much of their data publicly available as reasonably possible. Please note that your response to the following questions regarding the public data availability and the reasons for potentially not making data available will be available alongside your article upon publication.

Has data associated with your study been deposited into a publicly available repository? No.

Please select why. Please note that this statement will be available alongside your article upon publication.

Data will be made available on request.

## CRediT authorship contribution statement

**Mehrbakhsh Nilashi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Hossein Ahmadi:** Writing – review & editing, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **Rabab Ali Abumalloh:** Writing – review & editing, Writing – original draft, Validation, Software, Formal analysis, Data curation, Conceptualization. **Mesfer Alrizq:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Abdullah Alghamdi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization, Funding acquisition. **Sultan Alyami:** Writing – review & editing, Writing – original draft, Visualization, Validation, Formal analysis, Data curation, Conceptualization, Investigation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

Hossein Ahmadi was supported by the Baily Thomas Charitable Fund.

## References

[1] M.M. Patel, B.E. Adrada, Hereditary Breast Cancer: BRCA Mutations and beyond, Radiologic Clinics, 2024.
[2] B. Kinteh, S.L. Kinteh, A. Jammeh, E. Touray, A. Barrow, Breast cancer screening: knowledge, attitudes, and practices among female university students in the Gambia, BioMed Res. Int. 2023 (2023).
[3] S. Meharban, A. Ullah, S. Zaman, A. Hamraz, A. Razaq, Molecular structural modeling and physical characteristics of anti-breast cancer drugs via some novel topological descriptors and regression models, Current Research in Structural Biology (2024) 100134.
[4] S. Fu, H. Ke, H. Yuan, H. Xu, W. Chen, L. Zhao, Dual role of pregnancy in breast cancer risk, Gen. Comp. Endocrinol. (2024) 114501.
[5] M. Nilashi, O. Ibrahim, H. Ahmadi, L. Shahmoradi, A knowledge-based system for breast cancer classification using fuzzy logic method, Telematics Inf. 34 (2017) 133–144.
[6] C.H. Barrios, Global challenges in breast cancer detection and treatment, Breast 62 (2022) S3–S6.
[7] R.L. Siegel, A.N. Giaquinto, A. Jemal, Cancer statistics, CA A Cancer J. Clin. 2024 (2024) 74.
[8] A.G. Waks, E.P. Winer, Breast cancer treatment: a review, JAMA 321 (2019) 288–300.
[9] L. Wang, Early diagnosis of breast cancer, Sensors 17 (2017) 1572.
[10] M. Milosevic, D. Jankovic, A. Milenkovic, D. Stojanov, Early diagnosis and detection of breast cancer, Technol. Health Care 26 (2018) 729–759.
[11] S.Y. Loke, A.S.G. Lee, The future of blood-based biomarkers for the early detection of breast cancer, Eur. J. Cancer 92 (2018) 54–68.
[12] M.M. Boisen, C.L. Andersen, S. Sreekumar, A.M. Stern, S. Oesterreich, Treating gynecologic malignancies with selective estrogen receptor downregulators (SERDs): promise and challenges, Mol. Cell. Endocrinol. 418 (2015) 322–333.
[13] S. Chumsri, T. Howes, T. Bao, G. Sabnis, A. Brodie, Aromatase, aromatase inhibitors, and breast cancer, J. Steroid Biochem. Mol. Biol. 125 (2011) 13–22.
[14] B.C. Özdemir, G. Sflomos, C. Brisken, The challenges of modeling hormone receptor-positive breast cancer in mice, Endocr. Relat. Cancer 25 (2018) R319–R330.
[15] B. Mansour, C. Ngo, D. Schlemmer, P. Robidou, J. Blondel, C. Marin, G. Noé, A. Procureur, M. Jamelot, J. Gligorov, Simultaneous quantification of four hormone therapy drugs by LC-MS/MS: clinical applications in breast cancer patients, J. Pharmaceut. Biomed. Anal. (2024) 116032.
[16] T.T. Odunitan, O.A. Saibu, B.T. Apanisile, D.A. Omoboyowa, T.A. Balogun, A.V. Awe, T.M. Ajayi, G.V. Olagunju, F.M. Muhammad, M. Akinboade, Integrating biocomputational techniques for Breast cancer drug discovery via the HER-2, BCRA, VEGF and ER protein targets, Comput. Biol. Med. (2023) 107737.
[17] G. Kroemer, L. Senovilla, L. Galluzzi, F. André, L. Zitvogel, Natural and therapy-induced immunosurveillance in breast cancer, Nat. Med. 21 (2015) 1128–1138.
[18] M. García-Aranda, M. Redondo, Immunotherapy: a challenge of breast cancer treatment, Cancers 11 (2019) 1822.

[19] N. Masuda, S.-J. Lee, S. Ohtani, Y.-H. Im, E.-S. Lee, I. Yokota, K. Kuroi, S.-A. Im, B.-W. Park, S.-B. Kim, Adjuvant capecitabine for breast cancer after preoperative chemotherapy, N. Engl. J. Med. 376 (2017) 2147–2159.
[20] S.E. Jones, B.G. Durie, S.E. Salmon, Combination chemotherapy with adriamycin and cyclophosphamide for advanced breast cancer, Cancer 36 (1975) 90–97.
[21] P. Ellis, P. Barrett-Lee, L. Johnson, D. Cameron, A. Wardley, S. O'Reilly, M. Verrill, I. Smith, J. Yarnold, R. Coleman, Sequential docetaxel as adjuvant chemotherapy for early breast cancer (TACT): an open-label, phase III, randomised controlled trial, Lancet 373 (2009) 1681–1692.
[22] G. Kledzik, C. Bradley, J. Meites, Reduction of carcinogen-induced mammary cancer incidence in rats by early treatment with hormones or drugs, Cancer Res. 34 (1974) 2953–2956.
[23] J. Cuzick, I. Sestak, M. Baum, A. Buzdar, A. Howell, M. Dowsett, J.F. Forbes, Effect of anastrozole and tamoxifen as adjuvant treatment for early-stage breast cancer: 10-year analysis of the ATAC trial, Lancet Oncol. 11 (2010) 1135–1141.
[24] H.M. Lamb, J.C. Adkins, Letrozole: a review of its use in postmenopausal women with advanced breast cancer, Drugs 56 (1998) 1125–1140.
[25] M.R. Nathan, P. Schmid, A Review of Fulvestrant in Breast Cancer, Oncology and Therapy, vol. 5, 2017, pp. 17–29.
[26] C.K. Osborne, Tamoxifen in the treatment of breast cancer, N. Engl. J. Med. 339 (1998) 1609–1618.
[27] G. Griguolo, M.V. Dieci, V. Guarneri, P. Conte, Olaparib for the treatment of breast cancer, Expet Rev. Anticancer Ther. 18 (2018) 519–530.
[28] J.-C. Xuhong, X.-W. Qi, Y. Zhang, J. Jiang, Mechanism, safety and efficacy of three tyrosine kinase inhibitors lapatinib, neratinib and pyrotinib in HER2-positive breast cancer, Am. J. Cancer Res. 9 (2019) 2103.
[29] K. Feldinger, A. Kong, Profile of neratinib and its potential in the treatment of breast cancer, Breast Cancer (2015) 147–162.
[30] P. Tarantino, S. Morganti, J. Uliano, F. Giugliano, E. Crimini, G. Curigliano, Margetuximab for the treatment of HER2-positive metastatic breast cancer, Expet Opin. Biol. Ther. 21 (2021) 127–133.
[31] S. Maximiano, P. Magalhaes, M.P. Guerreiro, M. Morgado, Trastuzumab in the treatment of breast cancer, BioDrugs 30 (2016) 75–86.
[32] J. Zhou, Y. Zhong, Breast cancer immunotherapy, Cell. Mol. Immunol. 1 (2004) 247–255.
[33] E. Barrett-Connor, L. Mosca, P. Collins, M.J. Geiger, D. Grady, M. Kornitzer, M.A. McNabb, N.K. Wenger, Effects of raloxifene on cardiovascular events and breast cancer in postmenopausal women, N. Engl. J. Med. 355 (2006) 125–137.
[34] C.A. Pedersen, P.J. Schneider, M.C. Ganio, D.J. Scheckelhoff, ASHP national survey of pharmacy practice in hospital settings: prescribing and transcribing—2019, Am. J. Health Syst. Pharm. 77 (2020) 1026–1050.
[35] A.L. Plumb, Drugs. com: drug information online 2004, Ref. Rev. 18 (2004) 41, 41.
[36] E. Sillence, P. Briggs, P.R. Harris, L. Fishwick, How do patients evaluate and make use of online health information? Soc. Sci. Med. 64 (2007) 1853–1862.
[37] M. Hardey, Consuming professions: user-review websites and health services, J. Consum. Cult. 10 (2010) 129–149.
[38] E. Saad, S. Din, R. Jamil, F. Rustam, A. Mehmood, I. Ashraf, G.S. Choi, Determining the efficiency of drugs under special conditions from users' reviews on healthcare web forums, IEEE Access 9 (2021) 85721–85737.
[39] H. Hassani, C. Beneki, S. Unger, M.T. Mazinani, M.R. Yeganegi, Text mining in big data analytics, Big Data and Cognitive Computing 4 (2020) 1.
[40] M. Nilashi, S. Samad, A. Alghamdi, M.Y. Ismail, O. Alghamdi, S.S. Mehmood, S. Mohd, W.A. Zogaan, A. Alhargan, A new method for analysis of customers' online review in medical tourism using fuzzy logic and text mining approaches, Int. J. Inf. Technol. Decis. Making 21 (2022) 1797–1820.
[41] P. Vivithanaporn, T. Kongratanapasert, B. Suriyapakorn, P. Songkunlertchai, P. Mongkonariyawong, P.K. Limpikirati, P. Khemawoot, Potential drug-drug interactions of antiretrovirals and antimicrobials detected by three databases, Sci. Rep. 11 (2021) 6089.
[42] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[43] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. B 39 (1977) 1–22.
[44] L. Wilkinson, G.E. Dallal, Tests of significance in forward selection regression with an F-to-enter stopping rule, Technometrics 23 (1981) 377–380.
[45] W.A. Carr, S.A. Ball, Predictors and treatment outcomes of perceived ward atmosphere among therapeutic community residents, J. Subst. Abuse Treat. 46 (2014) 567–573.
[46] A. Mehmood, S. Nawab, Y. Jin, H. Hassan, A.C. Kaushik, D.-Q. Wei, Ranking breast cancer drugs and biomarkers identification using machine learning and pharmacogenomics, ACS Pharmacol. Transl. Sci. 6 (2023) 399–409.
[47] F. Firoozbakht, I. Rezaeian, L. Rueda, A. Ngom, Computationally repurposing drugs for breast cancer subtypes using a network-based approach, BMC Bioinf. 23 (2022) 143.
[48] L. Shi, F. Yan, H. Liu, Screening model of candidate drugs for breast cancer based on ensemble learning algorithm and molecular descriptor, Expert Syst. Appl. 213 (2023) 119185.
[49] H. Zhou, H. Cao, L. Matyunina, M. Shelby, L. Cassels, J.F. McDonald, J. Skolnick, MEDICASCY: a machine learning approach for predicting small-molecule drug side effects, indications, efficacy, and modes of action, Mol. Pharm. 17 (2020) 1558–1574.
[50] A. Silvina, J. Bowles, P. Hall, On Predicting the Outcomes of Chemotherapy Treatments in Breast Cancer, Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Springer, 2019, pp. 180–190. Proceedings 17.
[51] V.C.H. Chen, T.Y. Lin, D.C. Yeh, J.W. Chai, J.C. Weng, Predicting chemo-brain in breast cancer survivors using multiple MRI features and machine-learning, Magn. Reson. Med. 81 (2019) 3304–3313.
[52] A.Z. Ogunleye, C. Piyawajanusorn, A. Gonçalves, G. Ghislat, P.J. Ballester, Interpretable machine learning models to predict the resistance of breast cancer patients to doxorubicin from their microRNA profiles, Adv. Sci. 9 (2022) 2201501.
[53] M. Sinkala, K. Naran, D. Ramamurthy, N. Mungra, K. Dzobo, D. Martin, S. Barth, Machine learning and bioinformatic analyses link the cell surface receptor transcript levels to the drug response of breast cancer cells and drug off-target effects, PLoS One 19 (2024) e0296511.
[54] S. Kobara, A. Rafiei, M. Nateghi, S. Bozkurt, R. Kamaleswaran, A. Sarker, Social Media as a Sensor: Analyzing Twitter Data for Breast Cancer Medication Effects Using Natural Language Processing (2024) arXiv preprint arXiv:2403.00821.
[55] T. Abd El-Hafeez, M.Y. Shams, Y.A. Elshaier, H.M. Farghaly, A.E. Hassanien, Harnessing machine learning to find synergistic combinations for FDA-approved cancer drugs, Sci. Rep. 14 (2024) 2428.
[56] Ö. Baysal, D. Genç, R.S. Silme, K.K. Kırboga, D. Çoban, N.A. Ghafoor, L. Tekin, O. Bulut, Targeting breast cancer with N-Acetyl-D-Glucosamine: integrating machine learning and cellular assays for promising results, Anti Cancer Agents Med. Chem. 24 (2024) 334–347.
[57] M. Nilashi, B. Minaei-Bidgoli, A. Alghamdi, M. Alrizq, O. Alghamdi, F.K. Nayer, N.O. Aljehane, A. Khosravi, S. Mohd, Knowledge discovery for course choice decision in Massive Open Online Courses using machine learning approaches, Expert Syst. Appl. 199 (2022) 117092.
[58] M. Nilashi, R.A. Abumalloh, H. Ahmadi, S. Samad, M. Alrizq, H. Abosaq, A. Alghamdi, The nexus between quality of customer relationship management systems and customers' satisfaction: evidence from online customers' reviews, Heliyon 9 (2023) e21828.
[59] M. Nilashi, R. Ali Abumalloh, H. Ahmadi, M. Alrizq, A. Alghamdi, O.A. Alghamdi, S. Alyami, A proposed method for quality evaluation of COVID-19 reusable face mask, Measurement and Control 57 (2024) 828–840.
[60] D.A. Reynolds, Gaussian Mixture Models, Encyclopedia of biometrics, 2009, p. 741.
[61] T.K. Moon, The expectation-maximization algorithm, IEEE Signal Process. Mag. 13 (1996) 47–60.
[62] M. Nilashi, R.A. Abumalloh, S.Y.M. Yusuf, H.H. Thi, M. Alsulami, H. Abosaq, S. Alyami, A. Alghamdi, Early diagnosis of Parkinson's disease: a combined method using deep learning and neuro-fuzzy techniques, Comput. Biol. Chem. 102 (2023) 107788.
[63] M. Nilashi, R.A. Abumalloh, M. Alrizq, A. Almulihi, O. Alghamdi, M. Farooque, S. Samad, S. Mohd, H. Ahmadi, A hybrid method to solve data sparsity in travel recommendation agents using fuzzy logic approach, Math. Probl Eng. 2022 (2022) 7372849.
[64] M. Nilashi, R.A. Abumalloh, A. Alghamdi, B. Minaei-Bidgoli, A.A. Alsulami, M. Thanoon, S. Asadi, S. Samad, What is the impact of service quality on customers' satisfaction during COVID-19 outbreak? New findings from online reviews analysis, Telematics Inf. 64 (2021) 101693.
[65] M. Nilashi, B. Minaei-Bidgoli, M. Alrizq, A. Alghamdi, A.A. Alsulami, S. Samad, S. Mohd, An analytical approach for big social data analysis for customer decision-making in eco-friendly hotels, Expert Syst. Appl. 186 (2021) 115722.
[66] J. Suskin, C.L. Shapiro, Osteoporosis and musculoskeletal complications related to therapy of breast cancer, Gland Surg. 7 (2018) 411.

[67] N. Al-Dasooqi, J.M. Bowen, R.J. Gibson, T. Sullivan, J. Lees, D.M. Keefe, Trastuzumab induces gastrointestinal side effects in HER2-overexpressing breast cancer patients, Invest. N. Drugs 27 (2009) 173–178.
[68] L. Ding, L. Wang, J. Yin, Z. Fan, Z. He, Effects of neoadjuvant chemotherapy on respiratory function in patients with breast cancer, Chin. J. Cancer Res. 32 (2020) 36.
[69] J.J. Mao, A. Chung, A. Benton, S. Hill, L. Ungar, C.E. Leonard, S. Hennessy, J.H. Holmes, Online discussion of drug side effects and discontinuation among breast cancer survivors, Pharmacoepidemiol. Drug Saf. 22 (2013) 256–262.
[70] G. Natale, G. Bocci, Cardiovascular and central nervous system toxicity by anticancer drugs in breast cancer patients, Brain and Heart Dynamics (2020) 765–789.
[71] T.A. Ahles, A.J. Saykin, Breast cancer chemotherapy-related cognitive dysfunction, Clin. Breast Cancer 3 (2002) S84–S90.
[72] T.d.O. Gozzo, M.S. Panobianco, M.J. Clapis, A.M.d. Almeida, Dermatological toxicity in women with breast cancer undergoing chemotherapy treatment, Rev. Latino-Am. Enferm. 18 (2010) 681–687.
[73] S.S. Coughlin, D.U. Ekwueme, Breast cancer as a global health concern, Cancer epidemiology 33 (2009) 315–318.
[74] R.T. Simo, A.P. Nyemb, E.M. Baiguerel, A.H.N. Kamdje, A. Mohamadou, C. Nangue, P.B. Telefo, Assessment of breast cancer awareness and detection of asymptomatic cases in Ngaoundere, Adamawa region of Cameroon, Heliyon 10 (2024) e32995.
[75] WHO, Breast Cancer (2024).
[76] G. Lăzăroiu, T. Gedeon, E. Rogalska, M. Andronie, K.F. Michalikova, Z. Musova, M. Iatagan, C. Uță, L. Michalkova, M. Kovacova, The economics of deep and machine learning-based algorithms for COVID-19 prediction, detection, and diagnosis shaping the organizational management of hospitals, Oeconomia Copernicana 15 (2024) 27–58.
[77] M. Bugaj, T. Kliestik, G. Lăzăroiu, Generative artificial intelligence-based diagnostic algorithms in disease risk detection, in: Personalized and Targeted Healthcare Procedures, and in Patient Care Safety and Quality, vol. 15, Contemporary Readings in Law and Social Justice, 2023, pp. 9–26.
[78] M. Grupac, A. Zauskova, E. Nica, Generative artificial intelligence-based treatment planning in clinical decision-making, in precision medicine, and in personalized healthcare, Contemp. Read. Law Soc. Justice 15 (2023).