A Review of Statistical and Machine Learning Methods for Modeling Cancer Risk using Structured Clinical Data

Aaron N. Richter[a,b,*], Taghi M. Khoshgoftaar[a]

[a] *Florida Atlantic University*
[b] *Modernizing Medicine, Inc.*

Abstract

Advancements are constantly being made in oncology, improving prevention and treatment of cancers. To help reduce the impact and deadliness of cancers, they must be detected early. Additionally, there is a risk of cancers recurring after potentially-curative treatments are performed. Predictive models can be built using historical patient data to model the characteristics of patients that developed cancer or relapsed. These models can then be deployed into clinical settings to determine if new patients are at high risk for cancer development or recurrence. For large-scale predictive models to be built, structured data must be captured for a wide range of diverse patients. This paper explores current methods for building cancer risk models using structured clinical patient data. Trends in statistical and machine learning techniques are explored, and gaps are identified for future research. The field of cancer risk prediction is a high-impact one, and research must continue for these models to be embraced for clinical decision support of both practitioners and patients.

Keywords: cancer prediction, cancer recurrence, cancer relapse, data mining, machine learning, Electronic Health Records

1. Introduction

This paper aims to inform practitioners, namely oncology researchers, statisticians, and data scientists, of the current methods used for performing cancer risk and recurrence prediction. Additionally, this formal review identifies gaps in current research and paths for advancing the field.

The goal of cancer risk prediction is to determine if a given patient will develop cancer (or recur) at some point in the future [1]. The problem is distinct from patient identification (also called phenotyping [2]), as the goal is not to determine if a patient has a certain disease at the present moment, but to determine if the patient will develop it in the future. This task can be formulated as a supervised learning problem, where the input data are certain demographic and clinical elements (e.g. age, sex, and treatment history), and the output

[*] Corresponding author

Email addresses: `arichter@fau.edu` (Aaron N. Richter), `khoshgof@fau.edu` (Taghi M. Khoshgoftaar)

variable is the probability that the patient will develop the cancer at some point in the future. This probability can be tracked over time, assigning risk as time increases. The problem can also be formulated as a binary classification task, attempting to ascertain whether or not a patient will develop cancer at a specified point in time (i.e. developing the cancer within the next five years). A prediction model is built by supplying historical data from patients that did, or did not, develop the cancer in question. Statistical and machine learning techniques are used to fit a model to this historical data (i.e. training data). Then, to prove the model will be generalizable to different patient populations, a validation set (or multiple validation sets) is used to determine the performance of the model. When the performance of the model is adequate, based on several metrics, it can be deployed into clinical settings to help inform patients and providers. For more information about predictive modeling for medicine in general, see [1][3].

In this review, a distinction is made between models that attempt to predict if a patient will develop a cancer in the future (risk prediction), and those that predict whether or not a patient will relapse after a potentially-curative treatment (recurrence prediction). These problems are distinct in that they often have different types of input data. For example, a risk prediction model will not have any variables about cancer in the patient, as the patient has not yet developed cancer (although family histories of cancer would be relevant). For recurrence models, as will be seen in the papers studied, information about the tumor and treatments for the cancer are often chosen for inclusion in the models [4]. While the problem scenarios are distinct, the approaches to solve them can be very similar; in this paper, methods for both cancer risk and recurrence prediction are reviewed.

Accurate models are clinically relevant, as they can provide personalized treatment plans for patients at risk for a new cancer or recurrence of cancer in remission. There are various types of cancers, many of which have a very low incidence rate. It is not economically feasible to screen all patients visiting a doctor for a wide range of different diseases [5][6]. Thus, a model that can predict future development of cancer based on regularly captured clinical biomarkers, demographic, and lifestyle information is of high value to a healthcare system. As the model is built and tested, it can be used to flag high-risk patients for enrollment in a surveillance program, catered towards each patients' individual risk and clinical profile [7]. Therefore, a model must be applicable to large populations of patients, given that cancer is still a relatively rare disease but one of high importance to humanity.

To build high-impact models that can be generalized to a diverse array of patients, structured clinical data is required. As we discuss in Section 3, this review focuses on studies that utilize structured clinical information, not free-text or genetic data. Section 2 outlines the methodology used for our literature review. Rather than provide a summary of each related article, this paper highlights certain patterns about predictive model usage, sources of data (Section 3), statistical and machine learning methods (Section 4), and necessary future work (Section 5). Relevant papers will be mentioned throughout the text, and a summary of the papers profiled can be found in Appendix A.

Table 1: Cancer Types and Risk/Recurrence Prediction

	Prediction Problem		
Cancer Type	Risk	Recurrence	Total
Any	1	0	1
Bladder	0	1	1
Breast	0	8	8
Cervical	0	1	1
Colon	1	2	3
Hepatocellular Carcinoma	2	1	3
Lung	1	0	1
Pancreatic	1	0	1
Sarcoma	0	1	1
Gastric	1	1	2

2. Methodology

We conducted a comprehensive review of literature related to data mining for healthcare applications, and filtered the list of works to those relevant for this review. Therefore, works focusing on other diseases besides cancer, and those using non-clinical data (such as genomic or proteomic data) or primarily free-text clinical notes were excluded.

Papers were first identified by browsing through related journals, followed by a breadth-first search of articles using Pubmed[1] and Google Scholar[2]. Keywords used included but were not limited to: "cancer risk", "cancer recurrence", "cancer prediction", "machine learning", "data mining", and permutations of these keywords. Then, each paper identified was reviewed for relevance and a decision to keep or remove the paper was made. For each paper that was kept, related articles and articles citing the paper (utilizing search features available in both Pubmed and Google Scholar) were reviewed for relevance. This process was repeated until no new papers could be identified, resulting in 22 papers analyzed.

There are many different types of cancers, with different risk factors and treatment options, resulting in researchers with specific and invaluable knowledge of a specific type of cancer. Therefore, each paper focuses on a particular type of cancer for modeling, with the exception of Bayati et al., who attempted to predict cancer in general [8]. Table 1 outlines the type of cancer and prediction problem for the 22 papers reviewed.

3. Cancer Risk Models

3.1. Data Sources and Features

Patient data is collected from a variety of sources, and the availability of each varies based on the ease of collection, cost, and data storage methods [9]. This paper focuses on

[1] http://www.ncbi.nlm.nih.gov/pubmed

[2] https://scholar.google.com/

studies that utilize structured (non-free text) clinical information, as this data is widely collected and has the greatest value for efficient modeling of cancer risk and recurrence.

3.1.1. Molecular Data

Collection of molecular data, such as genomic or proteomic information, is still inhibited by cost and availability of facilities to handle sequencing a large number of patients. While molecular data has been shown to be highly valuable in many cancer research settings [10], it is not yet captured for the majority of patients, so there would be a small impact in the area of population-level cancer risk modeling. Therefore, papers using molecular data are excluded from this review.

3.1.2. Clinical and Practice Data

There is a large amount of information collected about routine clinical encounters in hospitals and private practices. Billing data, such as insurance claims for procedures and medications, have mature data sharing standards due to their financial impact and need for consistency. Coding standards include Current Procedural Terminology (CPT) [11] for procedures performed by a physician, and International Classification of Diseases (ICD) for specifying which diagnoses warrant the procedure being billed for [12]. While these codes provide a standard for data collection, there is more clinically-relevant information that is not captured through routine billing data. For example, the ICD-10 code C50.111 represents “malignant neoplasm of central portion of right female breast”, but the tumor information, progression of the patient’s health, and the patient’s medical and social history are all unknown. Several papers reviewed utilize ICD codes to determine if a patient has a certain condition.

Electronic Health Record (EHR) systems have the potential to capture large databases of clinical patient data relating to office and hospital visits, medical history, lab and pathology results, prescriptions, and social and demographic information. The biggest promise of EHR systems is being able to collect structured data at the point of care by physicians themselves, preventing the “garbage in-garbage out” problem of big data. This information is more advantageous for cancer risk and recurrence prediction, because the clinical information is often more valuable than the financial information (procedures and billing). For example the number of adenomatous polyps, or family history determines the risk profile for colon cancer. With melanoma, family history, proximity to the equator, number of sunburns, and the number of clinically atypical nevi are all factors that lead to developing the cancer. With the increasing adoption of these systems (due to governmental regulations such as the Affordable Care Act [13]) comes greater possibilities for utilizing this data to both improve patient outcomes and reduce healthcare costs. However, there are barriers to fully unlocking the potential of this data. EHR systems are developed independently and often maintain proprietary standards for data collection and storage. Furthermore, many EHRs capture clinical information via free-text notes, making it difficult to extract structured information for use in automated decision support algorithms. While there is a great deal of research involving Natural Language Processing (NLP) techniques to extract structured elements from free-text data [14], papers using these techniques on free-text notes are excluded from this review.

Though not mentioned in the articles reviewed, other standards exist for capturing clinical data that is transferred between multiple parties to efficiently care for patients. ePrescriptions, prescriptions that are sent electronically from the doctor's office to a pharmacy, use standards such as National Drug Code (NDC) numbers and RxNorm [15] to ensure the correct medications are given to the patient. Logical Observation Identifiers Names and Codes (LOINC) are used to maintain consistency in the ordering and reporting of lab results and other clinical observations. The Systematized Nomenclature of Medicine (SNOMED) maintains coding standards for clinical information such as diagnoses, family history, allergies, social information, and others. The adoption of these standards is not consistent across medical providers, but when used, they provide valuable structured information that can be used to further population health research.

3.1.3. Social and Lifestyle Data

Social and lifestyle data can be important to modeling the risk for certain cancers. Smoking has been shown to be associated with lung cancer [16], alcohol consumption with liver cancer [17], and UV light exposure with skin cancer [18]. This data can be captured through routine clinical encounters using EHR systems, or through surveys and questionnaires given to patients. Several studies from the National Cancer Center of the Republic of Korea use data collected from health exams that include lifestyle information such as alcohol use and smoking status [6].

3.1.4. Clinical Registries

Clinical registries help solve research problems by maintaining a centralized database of clinical information specific to certain patient populations. The data points captured are often based on expert knowledge of the disease being studied, and can be submitted through electronic connections with digital record systems or manual input. Therefore, the data stored in these registries can be from multiple different sources, such as demographic, billing, pathologic, and tumor information. Registries are common for high-profile diseases, such as cancer, and many governments require that all cancers be recorded in a local or national cancer registry [19]. The same studies mentioned in Section 3.1.3, from the Korean National Cancer Center [5], and one study from Linköping University in Sweden [20], link data from a national cancer registry to determine when a patient developed cancer, and a national death registry to determine when and why a patient died. Many articles in this review build models from data stored in clinical registries.

3.1.5. Feature Types

Figure 1 outlines the different types of features for papers profiled in this review, based on the prediction problem (risk or recurrence). Note that the feature types used refer to those features that remained in the final model, not all features available to the researchers. The feature categories are as follows:

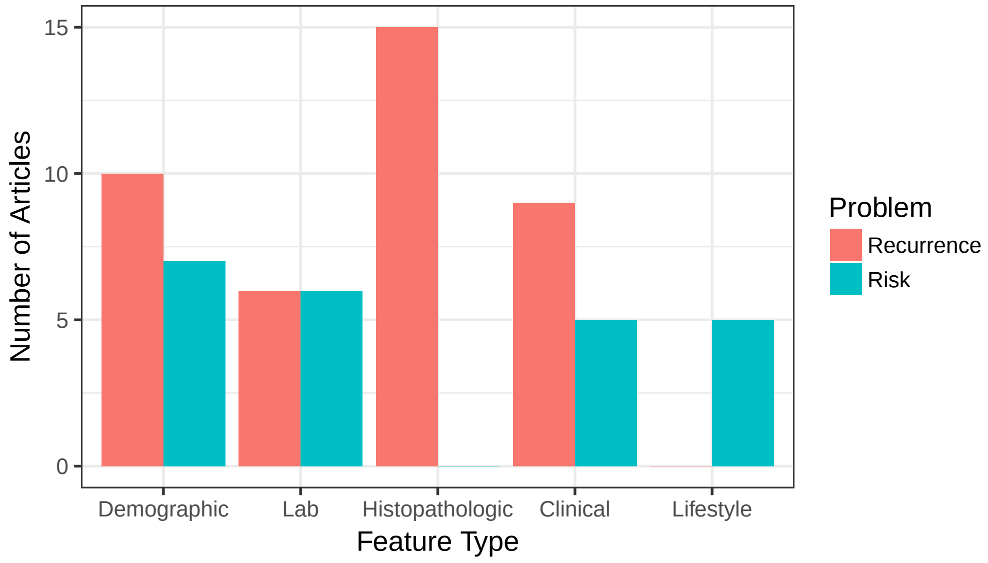


Figure 1: Feature Types by Prediction Problem (22 total articles). Many articles use more than one feature type.

- **Demographic**: Patient demographic information. The papers profiled only used age and/or sex, while one paper used race [7].
- **Lab**: Laboratory test results, such as white blood cell count, hemoglobin, glucose, triglycerides, etc.
- **Histopathologic**: Cancer and tumor-related information, such as the location, tumor size, metastasis, stage, margins, etc.
- **Clinical**: Treatments, family history, vitals, and other routinely captured clinical information that does not fit into any of the other categories.
- **Lifestyle**: Social history information such as smoking status and alcohol use.

Understandably, papers predicting disease risk do not use histopathologic data, since the patient has not yet developed a cancer, while papers predicting recurrence have found that the tumor information is valuable for their models. There is the possibility of using histopathologic data for risk prediction, however, as one cancer may predict another type of cancer. For example, pancreatic cancer patients are at a higher risk for melanoma, while one breast cancer or melanoma puts a patient at risk for another. Four of the risk prediction papers that used lifestyle information were all created from the same biennial health examination program run by the Korean Health Insurance Corporation from 1996-1999 [21]. Additionally, no papers used all five categories of data, while most papers used two or three.

3.2. Models in Practice

Several prognostic and predictive models are used, or available for use, in clinical practice. Some of these are not based on statistical or machine learning models, but rule-based methods or clinical guidelines.

3.2.1. Cancer Staging

The TNM Classification of Malignant Tumors is an international standard developed and maintained by the American Joint Committee on Cancer (AJCC) and Union for International Cancer Control (UICC) to describe the stage of a cancer tumor when it is diagnosed. This standard measures the size of the primary tumor (T), spread to regional lymph nodes (N), and the presence of distant metastasis (M) [22]. The staging is used to bucket patients into mutually exclusive groups based on their tumor characteristics, providing a means to determine prognosis of the disease, including the risk of recurrence [23][24]. Several papers, namely Cahlon et al. [24], Weiser et al. [25], Bochner et al. [26], and Marelli et al. [27], built models to predict the risk of cancer recurrence, and found that their models were more accurate than using TNM staging alone.

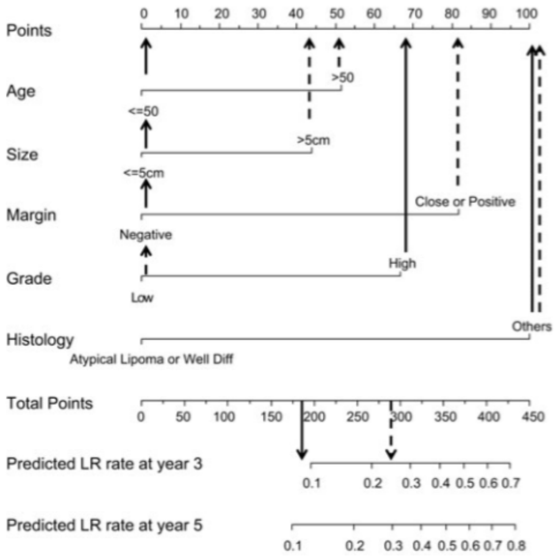
3.2.2. Nomograms

A nomogram is a graphical calculating device that allows a mathematical equation to be answered by aligning a straight-edge across values of different inputs, with the end of the straight-edge pointing to the result of the equation (see Figure 2a). Nomograms for oncology, widely studied by researchers at the Memoral Sloan Kettering Cancer Center (MSKCC), can produce succinct formulas that determine a patient's risk for certain clinical events, including the development or recurrence of a cancer [28]. Rather than utilize the archaic means of aligning a ruler to a page, MSKCC publishes these nomograms as online forms to be used by both physicians and patients[3] (see Figure 2b). These nomograms were built using regression techniques, such as Cox Proportional Hazards or competing risk survivial analysis, with the aim to use the minimum number of variables necessary to produce accurate results. Nomograms specific to cancer recurrence prediction were developed for: Sarcoma (Cahlon et al. [24]), Colon Cancer (Weiser et al. [25]), Breast Cancer (Rudloff et al. [29]), and Bladder Cancer (Bochner et al. [26]).

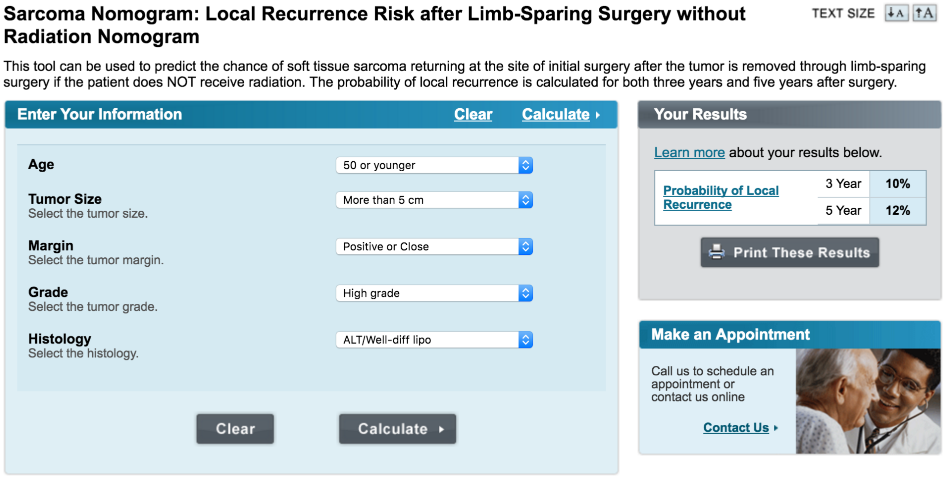
3.2.3. Breast Cancer Recurrence Models

Kim et al. [4] built a model for predicting breast cancer recurrence, and compared it to several other established guidelines: St. Gallen, Nottingham Prognostic Index (NPI), and Adjuvant! Online. The St. Gallen International Expert Consensus, in 2009, published several factors that contribute to a low-risk of recurrence, thus informing the use of adjuvant therapies post-surgery [30]. Researchers at the Nottingham City Hospital, in 1982, conducted retrospective multivariate analysis of breast cancer patients to build a prognostic model for survival, resulting in the NPI [31]. Kim et al. used this score to group patients into risk groups for recurrence. Cirkovic et al. also used the NPI index as an input to their breast

[3] https://www.mskcc.org/nomograms



(a) Manual nomogram. Lines are drawn from each feature to a particular score at the top line depending on the value of that feature. These points are then added up to reveal the predicted recurrence probability at either three or five years. The two styles of arrows indicate two different predictions made using the nomogram.



Sarcoma Nomogram: Local Recurrence Risk after Limb-Sparing Surgery without Radiation Nomogram

This tool can be used to predict the chance of soft tissue sarcoma returning at the site of initial surgery after the tumor is removed through limb-sparing surgery if the patient does NOT receive radiation. The probability of local recurrence is calculated for both three years and five years after surgery.

Enter Your Information

Age	50 or younger
Tumor Size Select the tumor size.	More than 5 cm
Margin Select the tumor margin.	Positive or Close
Grade Select the tumor grade.	High grade
Histology Select the histology.	ALT/Well-diff lipo

Your Results

Learn more about your results below.

Probability of Local Recurrence	3 Year	10%
	5 Year	12%

Make an Appointment

Call us to schedule an appointment or contact us online

(b) Online version of the nomogram.

Figure 2: Example Nomogram [24].

cancer relapse prediction model [32]. Adjuvant! Online is a web-based tool for determining survival and recurrence rates based on several factors[4] [33]. Kim et al. found their Support Vector Machine (SVM) model to be superior to the three established models, indicating that there is more research to be done to build clinically effective recurrence predictors.

[4] https://www.adjuvantonline.com/

4. Statistical and Machine Learning Methods

All articles in this review build predictive models to determine if a patient will develop a cancer, or recur, in the future. The techniques used, however, differ between studies. Generally, a study used either classical statistical methods, such as regression and survival analysis, or machine learning methods, such as Artificial Neural Networks (ANN), Support Vector Machine, or tree models. A few studies used hybrid approaches or compared statistical and machine learning methods. Studies produced by the same institution tend to use the same methods. For instance, four studies from MSKCC all used survival analysis techniques, and four studies from the National Cancer Center in Korea also used survival analysis techniques.

The goal of our analysis is to provide a snapshot of the current techniques used in the literature and discuss gaps in research, but not to extensively describe the theory and implementation of these models. For more details on models specifics, the references cited in each section should be explored.

4.1. Statistical Models

Modeling of disease risk or recurrence is easily framed as a survival analysis problem, and many studies utilize survival analysis techniques to construct their predictive models. Cox Proportional Hazards [34] is typically the model of choice, as it allows for time censoring and multivariate analysis. It is a regression model that creates a function of time, from baseline covariate values, that model the probability of an event occurring at any future time. In risk prediction studies, the event is the diagnosis of cancer, and time zero is either the enrollment in a study, or start of the observation period. In recurrence prediction studies, the event is the recurrence of cancer, and time zero is the date of a potentially curative treatment (often the surgical removal of a tumor). A patient is censored when follow-up is lost before the event occurs, which is typically the end of the follow-up period, but may be other scenarios such as a patient dropping out of the study or death. To handle a large number of patient deaths not due to the recurrence of Sarcoma, Cahlon et al. used a competing risk survival analysis model [35], treating non-recurrent death as a competing risk [24]. This study is the only one profiled performing survival analysis with a model different than the Cox Proportional Hazards model.

To visualize and intepret the results of survival models, a Kaplan-Meier curve is often generated. A Kaplan-Meier curve estimates the survival function of different cohorts of patients and plots the probability of survival along a time axis. Traditionally, this allows for comparison of patient cohorts with different characteristics of treatment regimens to determine which treatment to select for a new patient. A Kaplan-Meier analysis is not limited to predictions made from a statistical survival model, as a machine learning algorithm can also output whether or not a patient will survive. An example is shown in Figure 3, where Kim et al. use a Kaplan-Meier curve to compare the survival rates of high-risk and low-risk patients as determined by a machine learning model [4].

Logistic regression (LR) is another widely used statistical model. This technique allows for multivariate analysis and modeling of a binary dependent variable [36]. Essentially, a

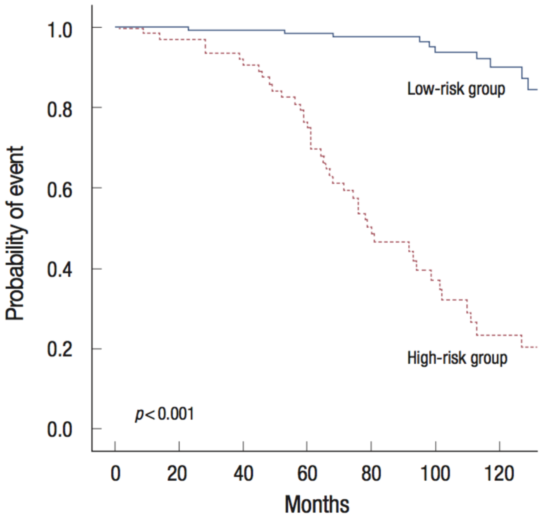


Figure 3: Example Kaplan-Meier curve [4]. The event is recurrence-free survival. The low-risk and high-risk labels are assigned to patients based on the output of the predictive model. This graph shows that the model does indeed discriminate between high-risk and low-risk patients.

linear regression model is built on the covariates, and then a logistic function is applied to discriminate between the two classes of output. El-Serag et al. used logistic regression models to predict the development of Hepatocellular Carcinoma (HCC), a form of liver cancer, within 6 months of an α-fetoprotein (AFP) test [37]. Among other models, Cirkovic et al. built a logistic regression model to predict recurrence after surgery for breast cancer [32]. Bayati et al. compare a traditional LR model to their own improved LR models based on multi-task learning, as their model attempts to predict risk of multiple different diseases (of which cancer is one) [8]. Regression models are widely used and understood in medical literature due to the simplicity of the fitted model. It is easy for a practitioner to see which features contribute toward the prediction [38]. However, regression models are not ideal for problems that do not easily fit to a linear model (or using non-linear techniques such as restricted cubic splines or fractional polynomials). Additionally, interpretability is limited for models that utilize a large number of variables.

4.2. Machine Learning Models

Decision trees are fairly interpretable ML models that can be used for regression or classification. They produce an output similar to a flow chart, allowing a path to be traversed based on the value of the instance in question, resulting in a predicted value. The model is trained by selecting a feature that best discriminates between the different outcomes, splitting the tree on this feature (node), and recursively performing this split on each new node that is generated. This produces a tree-like graph, and new instances can be scored by traversing the path created based on the instance's feature values. Various parameters of the model will determine when this splitting stops (number of iterations, number of nodes, etc.). Since the model is selecting features to split the tree on at each node, there is an inherent feature reduction that occurs, resulting in the most informative features being included in

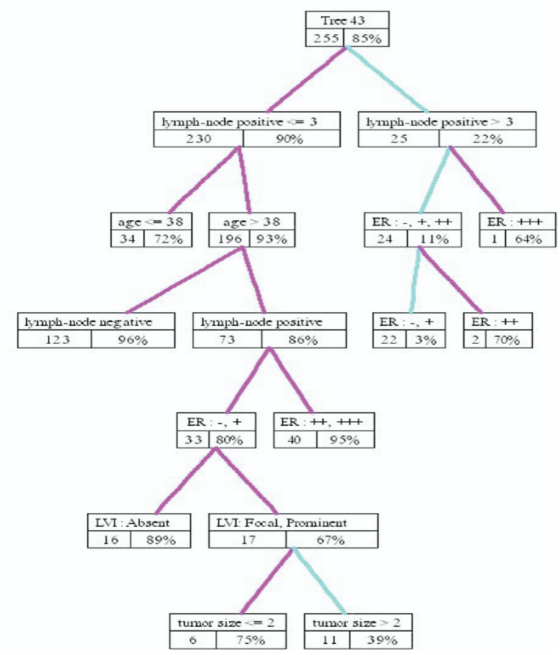


Figure 4: Example Decision Tree [40]. Each node presents the number of training instances in that node, and the percentage of those that did not recur within 5 years.

the model. Several studies use trees for feature selection, as will be seen in Section 4.3. Common algorithms for decision trees are C4.5 [39], C5.0 (an improved, commercialized version of C4.5), and Bayesian trees [40]. An example tree is shown in Figure 4, with the percentage at each node representing the probability of not having local recurrence of breast cancer in 5 years [40]. Tseng et al. find that their C5.0 model performed best when selecting two features to model the risk of recurrent cervical cancer [41]. Radespiel-Tröger et al. construct decision trees to model the recurrence of colon cancer within five years of curative resection [42]. Cirkovic et al. and Ahmad et al. both utilize a C4.5 model (among others) to predict recurrent breast cancer [32][43].

An adaptation to the decision tree model is called Random Forest. In a Random Forest, multiple trees are built and predictions are decided by majority voting. Bagging is used to construct the trees so that a random subset (with replacement) of features and a random subset of data are selected to build each tree. While building the trees, a random subset of features are considered at each decision node. After all trees are built, classification takes place by evaluating the instance with respect to all trees and the decision is the one agreed by the majority of the trees. Singal et al. utilize a Random Forest to predict the development of HCC in patients with cirrhosis [7]. While Random Forests can often be superior to single decision trees, the multiplicity of trees in the model makes it difficult to interpret and present to those not familiar with the technique.

Another widely used model in healthcare analytics is the Support Vector Machine (SVM). An SVM creates a set of hyperplanes for each feature in an infinite dimensional space, and

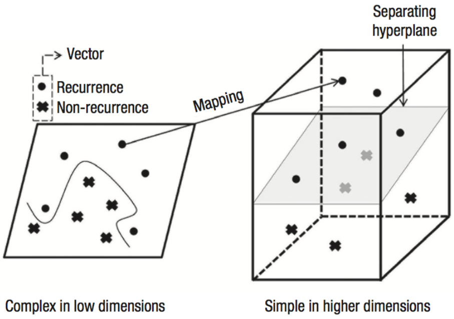


Figure 5: Basic framework of an SVM [4].

fits linear or nonlinear models that most effectively discriminate between the values of a binary output variable [4]. Kim et al. (Figure 5) provide a basic description of an SVM model in their paper that discriminates between recurrence and non-recurrence in breast cancer patients [4]. Tseng et al. [41], Cirkovic et al. [32], Liang et al. [44], and Ahmad et al. [43] all use SVMs to predict cancer recurrence.

A popular model in the machine learning community is the Artificial Neural Network (ANN). Variations of ANNs have been shown to be highly effective in unsupervised learning tasks such as image recognition [45]. ANNs, however, are very useful for supervised learning tasks such as disease prediction [46]. A neural network is roughly modeled after the way the human brain works, by creating nodes (neurons) that give weights to certain inputs and produce an ouput value. Multiple layers of nodes are tied together with an input layer taking in the value of the independent variables, and an output layer with nodes repsenting each of the possible outcome values. The weights at each layer in the network are modified as the model learns through back-propagation. When one node in the output layer is positive, the value at the node is taken as the prediction. When there is a large number of intermediate layers, this is often called “deep learning”, and has shown impressive results for very complex modeling problems [46]. Ahmad et al. provide an illustration of a basic network in Figure 6 [43]. Jerez-Aragonés et al. construct neural networks to predict the recurrence of breast cancer after surgery [47]. They construct multiple models with different network toplogies based on different time intervals, with the theory that recurrence risk is dependent on the amount of time after surgery, and not all features will have the same weight at different follow-up times [47]. Tseng et al. used a modification of an ANN, called Extreme Learning Machine (ELM), that randomly assigns the input weights while modeling the output weights of the network [41]. This makes the ELM model much faster to train than a typical ANN. Razavi et al. [20], Kim et al. [4], Cirkovic et al. [32], and Ahmad et al. [43] also use an ANN to model disease risk.

While machine learning methods can improve prediction accuracy over traditional regression techniques, there are several considerations when exploring different types of models. Van der Ploeg et al. showed that machine learning models can be “data hungry”, meaning that they require more samples than classical techniques to achieve stable results [48]. Since modern machine learning techniques have been shown to have very good predictive

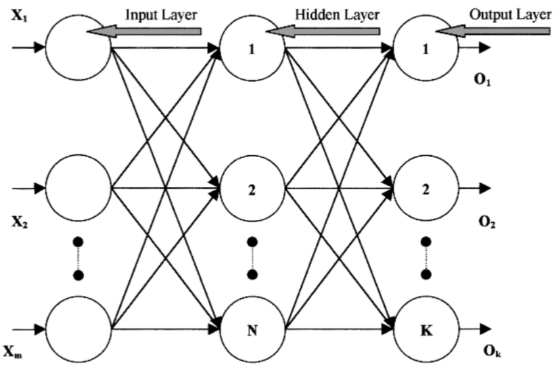


Figure 6: Basic framework of an ANN [43].

accuracy in other domains, it may be tempting to take a large medical dataset and blindly apply a robust model (such as Random Forest) to the data. This may not produce ideal results, because there are likely only a small number of variables that have clinical meaning for the particular problem [38]. Machine learning researchers should work closely with domain experts to determine which types of models work best for a particular problem. For more considerations when choosing between classical and machine learning methods, see Steyerberg et al. [38].

4.3. Feature Reduction

The first step in training a predictive model is determining what input features to provide to the model. Often, this is limited by the availability of variables in the database that is used to collect and store the clinical data (see Section 3.1). Additionally, computational and inferential complexity is a factor that can limit the number of features that can be used by the model. In most statistical models, a small number of features is necessary to interpret the significance of features and their combinations. For both statistical and machine learning models, there can be too many iterations or computations necessary, resulting in models that require too much time or computational resources to be built. Feature reduction can also be important in the context of deploying models for use in the clinical setting. If users of the prediction model need to manually enter the variables before a prediction can be made (i.e. in an online nomogram), the least number of variables should be included to improve the user experience of the model.

In most papers profiled, the authors have access to a dataset with a certain number of attributes, and these attributes are examined in the context of the research problem. In nearly every case, a domain expert, such as a physician or oncology researcher, will inform the analysts about features he or she believes will be important to the model. These studies then only focus on these features.

Most studies perform a univariate analysis to find which covariates have a statistically significant correlation with the output variable (see Figure 7). Then, only these features are used for the subsequent model. Methods include the Pearson correlation coefficient, mutual information, or distance correlation. This generally results in less than 10 features input to the model, which is desirable to allow interpretation of regression models.

4.3.1. Feature Selection Techniques

One approach to feature reduction is the use of feature selection algorithms, commonly used for ML models that require large feature vectors [49]. Many of these algorithms utilize univariate analysis methods, such as information gain or mutual information. They can, however, produce more than a single output for each feature (i.e. significant or not significant). Feature rankers order the features according to a certain statistic, leaving the practitioner to determining how many of the top features he or she wants to include in the model. Cirkovic et al. combined three different feature rankers from the Weka ML toolkit [50] (mRMR, ReliefF, and Information Gain), to select the top 20 most relevant features for use in their ML models.

Feature subset selection techniques evaluate features in groups to determine which subset is most informative for the predictive model. Razavi et al. apply Canonical Correlation Analysis (CCA) to reduce their feature set in the context of breast cancer recurrence prediction. CCA is a subset selection technique that finds the subset of features that most correlate with an output set of features. In CCA, the output must be a set of features, rather than a single variable, so the authors broke down the recurrence variable into different types of recurrence (loco-regional recurrence or distant metastasis) for the feature selection step. Once the most informative features were selected, they included those features in a neural network to predict the binary outcome of recurrence. Liang et al. utilized two feature subset selection techniques, namely Genetic Algorithm (GA) [51] and Simulated Annealing (SA) [52] to reduce the feature space provided to their SVM model.

Several predictive models, such as decision trees, effectively perform feature selection as part of the model building process. The p-values from a statistical model can also be used as a form of feature selection, by only selecting those features that have significant p-values (often <0.05). Jerez-Aragonés et al. use a decision tree model to first select important features, then build ANNs to predict recurrence of breast cancer [47]. Tseng et al. [41], Radespiel-Tröger et al. [42], Cheng et al. [40], and Singal et al. [7] built models with trees or forests, limiting the features used to those in the resultant trees. Li et al. built a logistic regression model using features that were found to be statistically significant from a Cox survival model [53]. In addition to feature subset selection techniques alone, Liang et al. combined both the GA and SA algorithms with Random Forest to create a hybrid model and subset-based feature selection approach [44].

4.4. Hybrid Models and Comparisons

In complex modeling problems, there is often not a one-size-fits-all approach that can be used. Researchers must explore various model options and determine which one works best in the context of the research problem. Additionally, different techniques can be combined to produce the best results.

Several studies profiled in this paper compare different ML models to each other, or compare ML models to a statistical model. Jerez-Aragonés et al., Kim et al., and Singal et

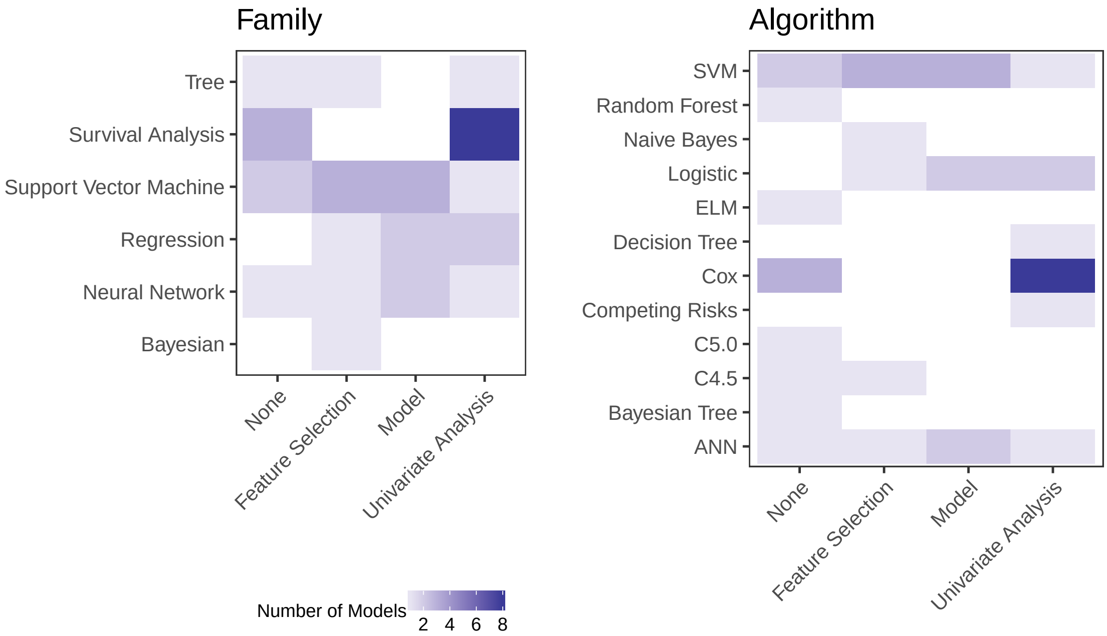


Figure 7: Feature Selection and Model Algorithm Methods. Studies with more than one method are counted multiple times. Feature Selection indicates use of a feature ranker or feature subset-selector. Left: Model Algorithms are grouped by their algorithm family. Right: Each model algorithm is outlined.

al. all find an ML model to have better performance than a Cox survival analysis model. Bayati et al. compare their enhanced multi-task regression models to a typical logistic model and find their method to be superior [8]. Tseng et al. find that a C5.0 decision tree performs better than an SVM and ELM model [41], while Cirkovic et al. find that an ANN model performs better than a C4.5 tree, SVM, logistic, and Naïve Bayes models [32]. Ahmad et al. find that an SVM outperforms both a C4.5 and ANN model [43].

Jerez-Aragonés et al. use a hybrid model by building a decision tree for feature selection, then an ANN for prediction [47]. Li et al. use Cox regression for finding important covariates then use those features as inputs for a logistic predictive model [53].

Figure 7 illustrates the different statistical and machine learning models, and feature selection methods used in the articles reviewed. The most widely used model combination is Cox regression, and most of those models utilized univariate analysis to select important features. SVM models tended to use feature rankers, subset selectors, or model-based feature selection.

4.5. Model Evaluation

Performance evaluation is an important step when creating a classification model, as the model must be proven to be accurate before using it to inform clinical decision-making. The most basic form of performance evaluation is predictive accuracy, which gives the percentage of instances that the model correctly labeled. This can be a biased measure, especially in cancer settings, as the class labels can be imbalanced, meaning there are many more patients that do not develop the disease than those that do. For example, if 10 out of 1,000 patients in a dataset develop the disease, the model can simply label all 1,000 patients as negative (not developing the disease), and still achieve an accuracy of 99%. Therefore, other metrics based on a confusion matrix (see Table 2) are calculated:

- True Positive Rate (TPR, sensitivity, recall): $TP/(TP + FN)$
- True Negative Rate (TNR, specificity): $TN/(TN + FP)$
- False Negative Rate (FNR): $FN/(TP + FN)$
- False Positive Rate (FPR, 1 - specificity): $FP/(FP + TN)$
- Positive Predictive Value (PPV, precision): $TP/(TP + FP)$

Confusion matrix-based metrics can also be biased, as many models produce a score, or a probability as the output rather than a concrete class label. A discrimination threshold must be set to determine at which point the score results in a positive or negative class value. It is important to evaluate these metrics in the context of the research problem. Many papers in the healthcare space report metrics using sensitivity and specificity, and since these metrics are inversely proportional to each other, the importance of each of will vary depending on what is being modeled. For example, if the model is trying to flag patients to screen for a particular cancer, a high sensitivity is desired to make sure that patients at risk for the disease are not missed. Setting the discrimination threshold for a maximum sensitivity, however, will drastically decrease the specificity and increase the false positive rate. This results in many low-risk patients being advised for screening, increasing patient mental burden and overall healthcare costs.

To handle multiple different discrimination thresholds, a Receiver Operating Characteristic curve (ROC) is generated [54]. The ROC curve plots the TPR (or sensitivity) against the FPR (or 1 - specificity) against a range of discrimination thresholds. An example ROC curve is shown in Figure 8, taken from Kim et al. [4]. By taking the area under the curve (AUC), a single metric is produced that is not dependent on the discrimination threshold. This metric is the probability that the model will rank an arbitrary positive instance higher than an arbitrary negative instance (in terms of the probability of an instance being positive). In papers that use statistical methods, the AUC is also denoted as the Concordance Index, or C-index. Most papers profiled report confusion matrix-based metrics, as well as an AUC score, for their models.

In addition to reporting performance measures on the training dataset, some sort of validation set must be used to prove that the prediction model can accurately predict on new instances, and is not overfit to the training data. This can be accomplished by splitting the dataset into training and testing sets, using an independent validation set, or by performing boostrapping or cross-validation. Validation using bootstrapping resamples the training data with replacement to create a training set, and uses the rest of the instances as a test set. This is repeated n number of times and the results combined to produce the final

Table 2: Confusion Matrix.

		Predicted Values	
		Positive	*Negative*
Actual Values	*Positive*	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

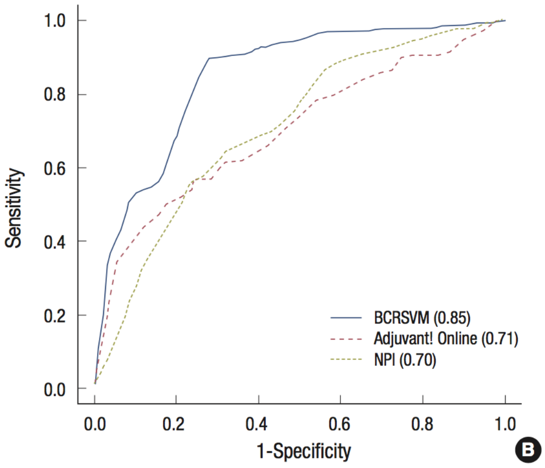


Figure 8: Example ROC Curve [4]. A larger area under the curve indicates better model performance.

performance score. Cross-validation is similar to bootstrapping, but divides the dataset into n folds, using $n-1$ folds for training and the final fold for testing. This is then repeated for the rest of the folds and the results are combined to produce the final performance score.

Of the papers profiled, only two papers did not use a form of validation when reporting performance of their models. Cheng et al. [40] did not report any model metrics, and Li et al. [53] only reported performance metrics on their training data. Yu et al. [6], Eom et al. [5], Shin et al. [55], and Park et al. [21] all utilize the same dataset collected through biennial health examinations conducted by the Korean Health Insurance Corporation. These four papers use patients examined from 1996-1997 as their training dataset, and patients examined from 1998-1999 as their validation dataset. Radespiel-Tröger et al. perform a similar split, using patients from 1984-1998 for training, and patients from 1978-1983 for testing [42]. Bochner et al. utilize an international multicenter prospective database from twelve different institutions [26]. To evaluate model performance, they treat each institution's data as a fold, effectively creating a 12-fold institution-based cross-validation. The rest of the papers analyzed use a typical train/test split, cross-validation, or bootstrapping methods.

5. Discussion

There are several patterns that can be seen from the studies reviewed in this paper. Many of the studies achieved relatively low performance compared to predictive results seen in other domains [56]. These may be improved through certain statistical and machine learning techniques, while other improvements may be more systemic to the healthcare system.

5.1. Advanced Methods

Performance of cancer risk and recurrence models can be improved by employing advanced data mining techniques. While many articles in this field apply statistical survival analysis models, it has been shown that machine learning techniques can outperform the statistical techniques [7]. There are many statistical models with acceptable AUC values (see Table A.4); but as the goal of this field is to provide the most accurate predictions, potential gains in AUC are worth the exploration of advanced methods.

The papers employing machine learning models tend to use decision tree, neural network, or SVM models. Decision tree models, similar to regression models, are easy to interpret, but they can lack in predictive performance. SVMs and ANNs are difficult to interpret, but can achieve good classification results. Other models, such as Naïve Bayes and Random Forest, were only used once in the papers studied. Random Forest has been shown to be particularly useful in the field of genetics [57]. Future work is needed to compare the performance of different machine learning algorithms in the context of cancer risk prediction.

While algorithm choice can improve model performance, there can be a bottleneck related to the quality of the input data and how it is structured. The following sections explore these ideas.

5.2. Temporal Data

Most studies, especially those employing Cox models, take a snapshot of patient data at the beginning of the follow-up period (time 0), and use this data to make predictions months or years into the future. While some variables do not change (such as tumor staging information), clinical data about a patient (such as vitals, lab tests, treatment, or prescription history) can change as time progresses. Models that only use data from one point in time can lose this information. Bayati et al. build predictive models using lab results taken during one month; these results may be captured multiple times during the month, in which case they average the results of the lab tests [8]. The four papers built using physical examination data from the Korean Health Insurance Corporation only use a snapshot of data taken from a certain point in time. Liang et al. only used data that is closest to the treatment time [44]. All other articles studying recurrence prediction use data from the time of the potentially curative surgery. Predictive models should exploit the value of longitudinal data, rather than suppress it. Changes in patient characteristics could potentially result in biomarkers that are valuable to a predictive model.

5.3. Missing Data

Missing data is an important problem in all modeling efforts, especially in the healthcare domain. If certain patient data is missing, such as tumor information or treatment history, the results can be significantly skewed. In addition to dropping patients, some studies will merely ignore variables that are available because there is not enough information filled in. El-Serag et al. chose to ignore lab results because they were too sparsely recorded in the input data [37]. It is important that all clinical variables are present, as to not bias the model, but it is also important to have a large sample size to make the model more generalizable for future instances. While most studies drop patients with missing clinical variables, there are several techniques that can help keep as many patients as possible in the model.

The benefit of using a Cox Proportional Hazards model, as opposed to simpler survival models, is that Cox models allow for censoring of patients that drop out of the study without experiencing the event in question. This may be due to death not related to the cancer, or simply not following up at the clinic.

Prospective databases and clinical registries can help produce the most integrous data, as they can make certain fields mandatory for practitioners to populate as they see patients. There is a trade-off however, if too many fields are required, the participating investigators may simply not submit data as it takes away from their time seeing patients. Additionally, there is an overhead of regulation and management in dealing with prospective studies, as compared to retrospective studies from EHR systems that are used regularly in practice [58]. Cahlon et al. [24], Tseng et al. [41], Singal et al. [7], and Bochner et al. [26] all use prospective databases or registries and do not mention the problem of missing data. This does not mean they did not encounter missing data, however, as they could have been filtered from their cohort counts beforehand. Weiser et al. were able to fill in some missing tumor information by having pathologists review the original slides [25].

Algorithmic techniques can be used to fill in missing values, such as mean imputation, or the Expectation-Maximization (EM) method. In mean imputation, a certain variable with missing data is filled in by taking the mean of the other instance's values [59]. Bayati et al. utilize mean imputation to substitute values for missing lab tests [8]. Naturally, this technique can only be used for continuous variables, and may not be desirable as it may bias those instances that do not have the value recorded by reducing the variance between values. EM is another method to impute missing values, and involves iteratively maximizing the log-likelihood of certain parameter values [60]. Radespiel-Tröger et al. dropped patients with more than one missing variable, but imputed values for one variable with EM as to not drop too many patients [42]. Ahmad et al. dropped patients with certain missing values, but used EM to impute other values [43]. While no studies reviewed used multiple imputation methods [61], techniques such as predictive mean matching can provide additional options for handling missing clinical data.

5.4. Feature Reduction

While feature reduction is useful for making models more interpretable, it can negatively impact the performance of the model. Nearly all studies reviewed used features that were deemed useful by a domain expert. A machine learning model, however, can often pick up on hidden patterns in data that humans cannot. Therefore, it can be advantageous to at least try building a model with all available features, or use a feature selection algorithm to reduce the feature set. Additionally, many studies only included covariates that were found to be statistically significant through univariate analysis. This can also decrease model performance, as a variable can still provide value to a model even if the p-value is not significant. Liang et al. found that univariate analysis resulted in only one significant variable, but their feature selection techniques selected four features to be included in their predictive models [44]. The models that utilized feature selection had better results than those that used the single significant variable.

5.5. Necessity of Structured Clinical Data

The field of cancer risk modeling can benefit most by increasing the amount of data that is available to researchers and machine learning experts. This advancement is hindered by the lack of structured clinical data available in EHR systems, as many still record free-text clinical notes. Medical providers must also utilize all the functionalities available in an EHR system to capture the most complete and valuable data. Paré et al. studied family practice physicians in Canada, and found that the majority of them did not utilize all available features in their EHR systems, which included e-prescribing, electronic lab ordering, secure data transmissions, and more [62]. Additionally, data privacy concerns often result in institutions or cloud-based EHR systems keeping terabytes of data locked away in private servers, especially if the data is free-text, as it is especially difficult to de-identify clinical notes [63]. Research in anonymization techniques must continue to help alleviate these concerns [64], as well as policies put in place to allow for more data sharing without breaching patient privacy.

While the adoption of EHRs has increased due to governmental requirements (such as the Affordable Care Act), the EHR industry is fragmented and data sharing is difficult. Standards need to be developed and enhanced to allow sharing of detailed clinical information. Through a study of mental health patients in Massachusetts, Madden et al. found that over half of the incidents of outpatient care were not captured in the patients' EHR system, as they occurred outside of the medical practice [65]. These data points were still covered, however, by insurance claims data. Ahmadian et al. specifically studied the data standards used in Clinical Decision Support Systems (CDSS), and found that many users of these systems were limited by incomplete data sharing standards and capabilities [66].

All papers studied only have age or sex as available demographic information (except one that uses race [7]). Due to practical necessity and Meaningful Use (MU) requirements from the Affordable Care Act, many other variables such as geographic information, smoking status, and alcohol use, are collected. These can provide valuable insights for modeling cancer risk, as there may be hidden biomarkers that contribute to cancer development. Additionally, EHR systems record real-world clinical data at the point-of-care, making models built from these datasets more generalizable to the public. Clinical trials and prospective observational studies may have small cohort sizes and can be biased towards the patients in the study.

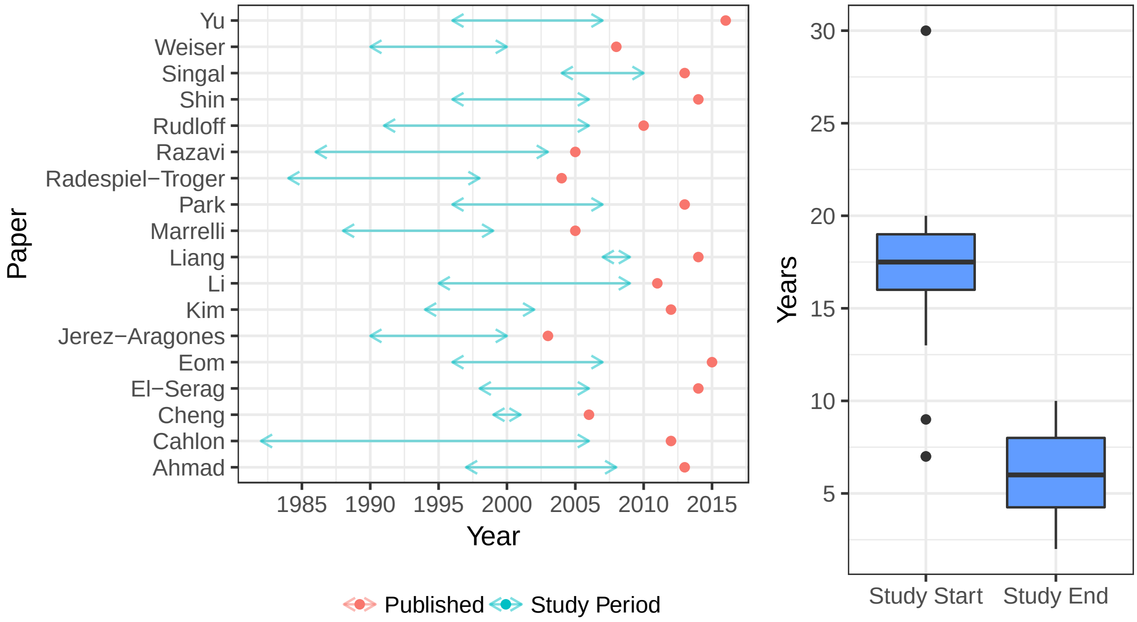


Figure 9: Year Published vs. Study Period. Four papers did not disclose study period and are excluded from this figure. Left: The line with arrows indicates the duration of the study period, and the dot indicates the year of publication. Right: Boxplot summarizing the distributions of the time between the year published and year of study start or end.

Data must also be shared between clinical and non-clinical settings. For example, four papers studied data from Korea that were linked from a physicial health examination, the national cancer registry, and the national death registry. This allowed for large-scale population health analysis, and they where able to build personalized predictive models for many different types of cancers. Razavi et al. were able to used linked data from the breast cancer registry, tumor registry, and death registry from Sweden [20].

5.6. Old Data

Due to the overhead of prospective data collection, privacy and legal issues, and modeling difficulty, studies often analyze data from many years in the past. This is not desirable, as clinical guidance is constantly changing based on medical breakthroughs and clinical trial results. A model built from data that is ten years old will be biased towards the treatments used and knowledge from that era, and may not be as accurate for current patients. Rudloff et al. created a covariate indicating if a patient was treated from either 1991-1998 or 1999-2006, as several key articles were published, resulting in significant changes in treatment patterns of breast cancer [29]. They found this covariate to be a statistically significant predictor in their model. Figure 9 illustrates the time between the study period start or end and the date of publication of the study. It can be seen that most studies are published at least five years after the end of the study period. Operational, policy, and data management efforts must be made to enhance the speed at which models can be built from current data. Additionally, online models can be built to utilize real-time data coming from EHR systems. While this requires major enhancements in infrastructure and data management, it will provide the most valuable models for predicting the risk and recurrence of different types of cancers.

6. Conclusion

This paper presents a comprehensive review of literature utilizing data mining techniques to perform cancer risk and recurrence prediction. This field is important, as these models can inform patient screening and treatment patterns, potentially improving patient outcomes and reducing overall healthcare costs. The key impact of these models is reducing costs. Governments spend billions of dollars on chronic conditions and acute end of life care. These models can determine who to spend those resources on, and more importantly, who not to spend those resources on. This both improves patient care and reduces operating costs, allowing funds to be spent advancing cutting-edge developments in cancer care.

The data provided to these models must be structured, frequently captured, and clinically relevant as to apply to large populations of patients. Coding standards must be enhanced to allow many different clinics and hospitals to exchange structured clinical data. While many standards exist for financial, laboratory, and prescription data, there are gaps in the transfer of point-of-care data such as outcomes and treatment plans.

Trends in statistical and machine learning techniques are presented, and analysis is performed to provide several valuable avenues for future work. Many studies utilize statistical survival analysis techniques, such as the Cox Proportional Hazards Model. Those that do not use survival analysis build predictive models using machine learning techniques such as Decision Trees, Neural Networks, and Support Vector Machines. To propel research in this area, advanced modeling methods using state of the art machine learning techniques must be employed, including time-series analysis, missing data imputation, and feature selection.

Acknowledgements

The authors would like to thank the editor and the anonymous reviewers for their insightful comments. They would also like to thank various members of the Data Mining and Machine Learning Laboratory, Florida Atlantic University, Boca Raton, for their assistance reviewing this manuscript.

Appendix A. Appendix

Tables A.3 and A.4 summarize all articles reviewed in this survey. Both tables have a row for each paper, while the columns display different information.

Table A.3: Summary of Articles

Paper	Data Source(s)	Cancer Type	Prediction Problem	Published	Study Period	# Instances	# Features
Park et al. [21]	Korean Health Insurance Corporation, Korean Central Cancer Registry, National Statistical Office (Korea)	Lung	Risk	2013	1996-2007	1,309,144	Unknown
Singal et al. [7]	University of Michigan Hepatology and Transplant Hepatology clinics (prospective database), Hepatitis C Antiviral Long- term Treatment against Cirrhosis (HALT-C) Clinical Trial	Hepatocellular Carcinoma	Risk	2013	2004-2010	1,492	Unknown
El-Serag et al. [37]	Department of Veterans Affairs Hepatitis C Virus Clinical Case Registry	Hepatocellular Carcinoma	Risk	2014	1998-2006	11,721	Unknown
Shin et al. [55]	Korean Health Insurance Corporation, Korean Central Cancer Registry, National Statistical Office (Korea)	Colon	Risk	2014	1996-2007	1,326,008	
Bayati et al. [8]	Kaggle Practice Fusion (KPF) & Stanford Hospitals and Clinics (ST)	Any	Risk	2015	Unknown	75,619 (ST) 1,096 (KPF)	1,313 (ST) 285 (KPF)
Eom et al. [5]	Korean Health Insurance Corporation, Korean Central Cancer Registry, National Statistical Office (Korea)	Gastric	Risk	2015	1996-2007	2,176,501	
Yu et al. [6]	Korean Health Insurance Corporation, Korean Central Cancer Registry, National Statistical Office (Korea)	Pancreatic	Risk	2016	1996-2007	2,975,369	Unknown
Jerez-Aragonés et al. [47]	Medical Oncology Service of the Hospital Clinico Universitario of Málaga, Spain	Breast	Recurrence	2003	1990-2000	1,035	85
Radespiel-Tröger et al. [42]	Erlangen Registry of Colorectal Carcinoma	Colon	Recurrence	2004	1984-1998	641	16
Razavi et al. [20]	Breast cancer registry, Tumor registry, Death registry (Sweden)	Breast	Recurrence	2005	1986-2003	5,787	150
Marrelli et al. [27]	Italian Research Group for Gastric Cancer Prospective Database	Gastric	Recurrence	2005	1988-1999	536	
Bochner et al. [26]	International Bladder Cancer Nomogram Consortium Post-RC Database	Bladder	Recurrence	2006	Unknown	9,064	14
Cheng et al. [40]	Koo Foundation, Sun Yat-Sen Cancer Center, Taipei, Taiwan	Breast	Recurrence	2006	1999-2001	1010 (Cox) 255 (tree)	
Weiser et al. [25]	Memorial Sloan-Kettering Cancer Center	Colon	Recurrence	2008	1990-2000	1,320	Unknown
Rudloff et al. [29]	Memoral Sloan-Kettering Cancer Center Prospective DCIS Database	Breast	Recurrence	2010	1991-2006	1,681	14
Li et al. [53]	Shanghai Cancer Hospital of Fudan University Prospective Breast Malignancy Database	Breast	Recurrence	2011	1995-2009	454	
Cahlon et al. [24]	Memoral Sloan-Kettering Cancer Center Prospective Sarcoma Database	Sarcoma	Recurrence	2012	1982-2006	684	Unknown
Kim et al. [4]	Korean tertiary teaching hospital	Breast	Recurrence	2012	1994-2002	679	193
Ahmad et al. [43]	Iranian Center for Breast Cancer	Breast	Recurrence	2013	1997-2008	547	22
Tseng et al. [41]	Chung Shan Medical University Hospital Tumor Registry (Taiwan)	Cervical	Recurrence	2014	Unknown	168	12
Liang et al. [44]	National Taiwan University Hospital	Hepatocellular Carcinoma	Recurrence	2014	2007-2009	83	16
Cirkovic et al. [32]	Clinical Center of Kragujevac, Serbia prospective database	Breast	Recurrence	2015			58

Table A.4: Summary of Articles (continued)

Paper	Feature Reduction	# Selected Features	Feature Types	Handle Missing Data?	Models	Best Results
Park et al. [21]	Univariate Analysis	7	Demographic, Lab, Clinical, Lifestyle	Dropped patients	Cox	0.864 (AUC)
Singal et al. [7]	Univariate Analysis	2	Demographic, Lab, Clinical, Lifestyle		Cox, Random Forest	0.64 (AUC)
El-Serag et al. [37]	Univariate Analysis	4	Demographic, Lab	Dropped features	Logistic Regression	0.815 (AUC)
Shin et al. [55]	Univariate Analysis	10 (men), 8 (women)	Demographic, Lab, Clinical, Lifestyle	Dropped patients	Cox	0.77 (men), 0.72 (women)
Bayati et al. [8]		30	Demographic, Lab	Imputed missing lab values	Single task learning, Multi-task learning, OLR-M	0.87 (AUC, Stanford), 0.73 (AUC, Kaggle)
Eom et al. [5]	Univariate Analysis	7 (men), 5 (women)	Demographic, Lab, Clinical, Lifestyle	Imputed values from nearest time point	Cox	0.782 (AUC, men), 0.705 (AUC, women)
Yu et al. [6]	Univariate Analysis	7	Demographic, Lab, Clinical, Lifestyle	Dropped patients	Cox	0.80 (AUC)
Jerez-Aragonés et al. [47]	Decision Tree	4-7 (multiple models)	Demographic, Lab, Histopathologic, Clinical	Dropped patients	ANN, Cox	0.948 (accuracy)
Radespiel-Tröger et al. [42]	Univariate Analysis	6	Histopathologic, Clinical	EM imputation for one missing feature, else dropped patients	Decision Tree	0.23 (Brier score)
Razavi et al. [20]	Canonical Correlation Analysis	12	Lab, Histopathologic	Dropped patients	ANN	0.71 (accuracy)
Marrelli et al. [27]	Maximum Likelihood	5	Demographic, Histopathologic, Clinical	Dropped patients	Logistic Regression	0.861 (accuracy)
Bochner et al. [26]		7	Demographic, Histopathologic		Cox	0.75 (AUC)
Cheng et al. [40]	Univariate Analysis	5	Demographic, Histopathologic, Clinical	Dropped patients	Cox, Bayesian Tree	Unknown
Weiser et al. [25]		11	Demographic, Lab, Histopathologic, Clinical	Reviewed pathology specimen for missing variables	Cox	0.77 (AUC)
Rudloff et al. [29]	Univariate Analysis	10	Histopathologic, Clinical	Dropped patients	Cox	0.68 (AUC)
Li et al. [53]	Univariate Analysis	3	Histopathologic	Dropped patients	Logistic Regression	0.70 (AUC)
Cahlon et al. [24]	Univariate Analysis	5	Demographic, Histopathologic		Competing Risk Survival Analysis	0.74 (AUC)
Kim et al. [4]	Univariate Analysis	7	Histopathologic	Dropped patients	Cox, SVM, ANN	0.85 (AUC)
Ahmad et al. [43]			Demographic, Lab, Histopathologic, Clinical	EM imputation, dropped patients	C4.5, ANN, SVM	0.957 (accuracy)
Tseng et al. [41]		2	Demographic, Histopathologic, Clinical		C5.0, SVM, Extreme Learning Machine (ELM)	0.924 (accuracy)
Liang et al. [44]	GA, SA, RF	Unknown	Demographic, Lab, Histopathologic		SVM	0.69 (AUC)
Cirkovic et al. [32]	mRMR, ReliefF, InfoGain	20	Demographic, Lab, Histopathologic, Clinical		Naïve Bayes, C4.5, SVM, Logistic Regression, ANN	0.96 (AUC)

References

[1] E. W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 1st Edition, Statistics for Biology and Health, Springer-Verlag New York, 2009.

[2] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, A. M. Lai, A review of approaches to identifying patient phenotype cohorts using electronic health records, Journal of the American Medical Informatics Association 21 (2) (2014) 221–230. doi:10.1136/amiajnl-2013-001935.
URL `http://jamia.oxfordjournals.org/lookup/doi/10.1136/amiajnl-2013-001935`

[3] R. Bellazzi, B. Zupan, Predictive data mining in clinical medicine: Current issues and guidelines, International Journal of Medical Informatics 77 (2) (2008) 81–97. doi:10.1016/j.ijmedinf.2006.11.006.
URL `http://linkinghub.elsevier.com/retrieve/pii/S1386505606002747`

[4] B. J. Kim, S.-w. Chung, others, Development of novel breast cancer recurrence prediction model using support vector machine, Journal of breast cancer.

[5] B. W. Eom, J. Joo, S. Kim, A. Shin, H.-R. Yang, J. Park, I. J. Choi, Y.-W. Kim, J. Kim, B.-H. Nam, Prediction Model for Gastric Cancer Incidence in Korean Population, PLOS ONE 10 (7) (2015) e0132613. doi:10.1371/journal.pone.0132613.
URL `http://dx.plos.org/10.1371/journal.pone.0132613`

[6] A. Yu, S. M. Woo, J. Joo, H.-R. Yang, W. J. Lee, S.-J. Park, B.-H. Nam, Development and Validation of a Prediction Model to Estimate Individual Risk of Pancreatic Cancer, PLOS ONE 11 (1) (2016) e0146473. doi:10.1371/journal.pone.0146473.
URL `http://dx.plos.org/10.1371/journal.pone.0146473`

[7] A. G. Singal, A. Mukherjee, B. J. Elmunzer, P. D. Higgins, A. S. Lok, J. Zhu, J. A. Marrero, A. K. Waljee, Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma, The American journal of gastroenterology 108 (11) (2013) 1723–1730.
URL `http://www.nature.com/ajg/journal/v108/n11/abs/ajg2013332a.html`

[8] M. Bayati, S. Bhaskar, A. Montanari, A Low-Cost Method for Multiple Disease Prediction, in: AMIA Annual Symposium Proceedings, Vol. 2015, American Medical Informatics Association, 2015, p. 329.
URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765607/`

[9] M. Herland, T. M. Khoshgoftaar, R. Wald, A review of data mining using big data in health informatics, Journal of Big Data 1 (1) (2014) 1–35.
URL `http://link.springer.com/article/10.1186/2196-1115-1-2`

[10] B. Heredia, T. M. Khoshgoftaar, A. Fazelpour, D. J. Dittman, Building an Effective Classification Model for Breast Cancer Patient Response Data, in: Information Re-use and Integration, IEEE, 2015, pp. 229–235. doi:10.1109/IRI.2015.46.
URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7300982`

[11] A. M. Association, Current procedural terminology: CPT, American Medical Association, 2007.

[12] W. H. Organization, others, International classification of diseases (ICD) (2012).

[13] P. Protection, A. C. Act, Patient protection and affordable care act, Public Law 111 (2010) 48.

[14] S. Doan, M. Conway, T. M. Phuong, L. Ohno-Machado, Natural language processing in biomedicine: a unified system architecture overview, Clinical Bioinformatics (2014) 275–294.
URL `http://link.springer.com/protocol/10.1007/978-1-4939-0847-9_16`

[15] S. J. Nelson, K. Zeng, J. Kilbourne, T. Powell, R. Moore, Normalized names for clinical drugs: RxNorm at 6 years, Journal of the American Medical Informatics Association 18 (4) (2011) 441–448. doi:10.1136/amiajnl-2011-000116.
URL `http://jamia.oxfordjournals.org/lookup/doi/10.1136/amiajnl-2011-000116`

[16] S. D. Stellman, T. Takezaki, L. Wang, Y. Chen, M. L. Citron, M. V. Djordjevic, S. Harlap, J. E. Muscat, A. I. Neugut, E. L. Wynder, others, Smoking and lung cancer risk in American and Japanese men: an international case-control study, Cancer Epidemiology Biomarkers & Prevention 10 (11) (2001) 1193–1199.
URL `http://cebp.aacrjournals.org/content/10/11/1193.short`

[17] F. Turati, C. Galeone, M. Rota, C. Pelucchi, E. Negri, V. Bagnardi, G. Corrao, P. Boffetta, C. La Vecchia, Alcohol and liver cancer: a systematic review and meta-analysis of prospective studies, Annals of

Oncology 25 (8) (2014) 1526–1535. doi:10.1093/annonc/mdu020.
URL http://annonc.oxfordjournals.org/cgi/doi/10.1093/annonc/mdu020

[18] C. Watts, M. Dieng, R. Morton, G. Mann, S. Menzies, A. Cust, Clinical practice guidelines for identification, screening and follow-up of individuals at high risk of primary cutaneous melanoma: a systematic review, British Journal of Dermatology 172 (1) (2015) 33–47. doi:10.1111/bjd.13403.
URL http://doi.wiley.com/10.1111/bjd.13403

[19] CDC - national program of cancer registries (NPCR).
URL http://www.cdc.gov/cancer/npcr/

[20] A. R. Razavi, H. Gill, H. Ahlfeldt, N. Shahsavar, , Studies in health technology and informatics 116 (2005) 175.
URL https://books.google.com/books?hl=en&lr=&id=HXTk5rWOdG4C&oi=fnd&pg=PA175&dq=%22documents.+In+medicine,+information+is+saved+in+different+forms+such+as%22+%22for+a+long+period+of+time.+Before+data+are+analysed+by+a+data+mining%22+&ots=ZThCVyrLa9&sig=Pt7Rcl44Js2CWFcBNdqOxq53Kac

[21] S. Park, B.-H. Nam, H.-R. Yang, J. A. Lee, H. Lim, J. T. Han, I. S. Park, H.-R. Shin, J. S. Lee, Individualized Risk Prediction Model for Lung Cancer in Korean Men, PLoS ONE 8 (2) (2013) e54823. doi:10.1371/journal.pone.0054823.
URL http://dx.plos.org/10.1371/journal.pone.0054823

[22] S. Edge, D. R. Byrd, C. C. Compton, A. G. Fritz, F. L. Greene, A. Trotti, AJCC Cancer Staging Manual, 7th Edition, Springer-Verlag New York, 2010.

[23] H. B. Burke, Outcome Prediction and the Future of the TNM Staging System, JNCI Journal of the National Cancer Institute 96 (19) (2004) 1408–1409. doi:10.1093/jnci/djh293.
URL http://jnci.oxfordjournals.org/cgi/doi/10.1093/jnci/djh293

[24] O. Cahlon, M. F. Brennan, X. Jia, L.-X. Qin, S. Singer, K. M. Alektiar, A Postoperative Nomogram for Local Recurrence Risk in Extremity Soft Tissue Sarcomas After Limb-Sparing Surgery Without Adjuvant Radiation, Annals of Surgery 255 (2) (2012) 343–347. doi:10.1097/SLA.0b013e3182367aa7.
URL http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00000658-201202000-00023

[25] M. R. Weiser, R. G. Landmann, M. W. Kattan, M. Gonen, J. Shia, J. Chou, P. B. Paty, J. G. Guillem, L. K. Temple, D. Schrag, L. B. Saltz, W. D. Wong, Individualized Prediction of Colon Cancer Recurrence Using a Nomogram, Journal of Clinical Oncology 26 (3) (2008) 380–385. doi:10.1200/JCO.2007.14.1291.
URL http://jco.ascopubs.org/cgi/doi/10.1200/JCO.2007.14.1291

[26] International Bladder Cancer Nomogram Consortium, Postoperative Nomogram Predicting Risk of Recurrence After Radical Cystectomy for Bladder Cancer, Journal of Clinical Oncology 24 (24) (2006) 3967–3972. doi:10.1200/JCO.2005.05.3884.
URL http://www.jco.org/cgi/doi/10.1200/JCO.2005.05.3884

[27] D. Marrelli, A. De Stefano, G. de Manzoni, P. Morgagni, A. Di Leo, F. Roviello, Prediction of Recurrence After Radical Surgery for Gastric Cancer: A Scoring System Obtained From a Prospective Multicenter Study, Annals of Surgery 241 (2) (2005) 247–255. doi:10.1097/01.sla.0000152019.14741.97.
URL http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00000658-200502000-00009

[28] V. P. Balachandran, M. Gonen, J. J. Smith, R. P. DeMatteo, Nomograms in oncology: more than meets the eye, The Lancet Oncology 16 (4) (2015) e173–e180.
URL http://www.sciencedirect.com/science/article/pii/S1470204514711167

[29] U. Rudloff, L. M. Jacks, J. I. Goldberg, C. A. Wynveen, E. Brogi, S. Patil, K. J. Van Zee, Nomogram for Predicting the Risk of Local Recurrence After Breast-Conserving Surgery for Ductal Carcinoma In Situ, Journal of Clinical Oncology 28 (23) (2010) 3762–3769. doi:10.1200/JCO.2009.26.8847.
URL http://jco.ascopubs.org/cgi/doi/10.1200/JCO.2009.26.8847

[30] A. Goldhirsch, J. N. Ingle, R. D. Gelber, A. S. Coates, B. Thurlimann, H.-J. Senn, Panel members, Thresholds for therapies: highlights of the St Gallen International Expert Consensus on

the Primary Therapy of Early Breast Cancer 2009, Annals of Oncology 20 (8) (2009) 1319–1329. doi:10.1093/annonc/mdp322.
URL `http://annonc.oxfordjournals.org/cgi/doi/10.1093/annonc/mdp322`

[31] M. H. Galea, R. W. Blamey, C. E. Elston, I. O. Ellis, The Nottingham prognostic index in primary breast cancer, Breast Cancer Research and Treatment 22 (3) (1992) 207–219. doi:10.1007/BF01840834.
URL `http://dx.doi.org/10.1007/BF01840834`

[32] B. R. A. Cirkovic, A. M. Cvetkovic, S. M. Ninkovic, N. D. Filipovic, Prediction models for estimation of survival rate and relapse for breast cancer patients, in: Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on, IEEE, 2015, pp. 1–6.
URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7367658`

[33] National Collaborating Centre for Cancer (Great Britain), Early and locally advanced breast cancer diagnosis and treatment: full guideline, National Collaborating Centre for Cancer, Cardiff, 2009.

[34] D. R. Cox, D. Oakes, Analysis of survival data, Vol. 21, CRC Press, 1984.

[35] D. G. Kleinbaum, M. Klein, Competing risks survival analysis, Survival Analysis: A self-learning text (2005) 391–461.

[36] T. M. Khoshgoftaar, E. B. Allen, Logistic regression modeling of software quality, International Journal of Reliability, Quality and Safety Engineering 6 (04) (1999) 303–317.

[37] H. B. El-Serag, F. Kanwal, J. A. Davila, J. Kramer, P. Richardson, A New Laboratory-Based Algorithm to Predict Development of Hepatocellular Carcinoma in Patients With Hepatitis C and Cirrhosis, Gastroenterology 146 (5) (2014) 1249–1255.e1. doi:10.1053/j.gastro.2014.01.045.
URL `http://linkinghub.elsevier.com/retrieve/pii/S0016508514001048`

[38] E. W. Steyerberg, T. van der Ploeg, B. Van Calster, Risk prediction with machine learning and regression methods: Risk prediction with machine learning and regression methods, Biometrical Journal 56 (4) (2014) 601–606. doi:10.1002/bimj.201300297.
URL `http://doi.wiley.com/10.1002/bimj.201300297`

[39] J. R. Quinlan, Improved use of continuous attributes in C4. 5, Journal of artificial intelligence research (1996) 77–90.
URL `http://www.jair.org/papers/paper279.html`

[40] S. H. Cheng, C.-F. Horng, J. L. Clarke, M.-H. Tsou, S. Y. Tsai, C.-M. Chen, J. J. Jian, M.-C. Liu, M. West, A. T. Huang, L. R. Prosnitz, Prognostic index score and clinical prediction model of local regional recurrence after mastectomy in breast cancer patients, International Journal of Radiation Oncology*Biology*Physics 64 (5) (2006) 1401–1409. doi:10.1016/j.ijrobp.2005.11.015.
URL `http://linkinghub.elsevier.com/retrieve/pii/S0360301605029688`

[41] C.-J. Tseng, C.-J. Lu, C.-C. Chang, G.-D. Chen, Application of machine learning to predict the recurrence-proneness for cervical cancer, Neural Computing and Applications 24 (6) (2014) 1311–1316. doi:10.1007/s00521-013-1359-1.
URL `http://link.springer.com/10.1007/s00521-013-1359-1`

[42] M. Radespiel-Tröger, W. Hohenberger, B. Reingruber, Improved prediction of recurrence after curative resection of colon carcinoma using tree-based risk stratification: Recurrence Prediction in Colon Ca, Cancer 100 (5) (2004) 958–967. doi:10.1002/cncr.20065.
URL `http://doi.wiley.com/10.1002/cncr.20065`

[43] L. Ahmad, A. Eshlaghy, M. Ebrahimi, A. Razavi, Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence, Journal of Health & Medical Informatics 04 (02). doi:10.4172/2157-7420.1000124.
URL `http://www.omicsonline.org/using-three-machine-learning-techniques-for-predicting-breast-can`
`1000124.php?aid=13087`

[44] J.-D. Liang, X.-O. Ping, Y.-J. Tseng, G.-T. Huang, F. Lai, P.-M. Yang, Recurrence predictive models for patients with hepatocellular carcinoma after radiofrequency ablation using support vector machines with feature selection methods, Computer Methods and Programs in Biomedicine 117 (3) (2014) 425–434. doi:10.1016/j.cmpb.2014.09.001.
URL `http://linkinghub.elsevier.com/retrieve/pii/S016926071400323X`

[45] I. Maglogiannis, C. Doukas, Overview of Advanced Computer Vision Systems for Skin Lesions Characterization, IEEE Transactions on Information Technology in Biomedicine 13 (5) (2009) 721–733. doi:10.1109/TITB.2009.2017529.
URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4801738`

[46] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, E. Muharemagic, Deep learning applications and challenges in big data analytics, Journal of Big Data 2 (1). doi:10.1186/s40537-014-0007-7.
URL `http://www.journalofbigdata.com/content/2/1/1`

[47] J. M. Jerez-Aragonés, J. A. Gómez-Ruiz, G. Ramos-Jiménez, J. Muñoz-Pérez, E. Alba-Conejo, A combined neural network and decision trees model for prognosis of breast cancer relapse, Artificial intelligence in medicine 27 (1) (2003) 45–63.
URL `http://www.sciencedirect.com/science/article/pii/S0933365702000866`

[48] T. van der Ploeg, P. C. Austin, E. W. Steyerberg, Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints, BMC medical research methodology 14 (1) (2014) 137.
URL `https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-14-137`

[49] K. Gao, T. M. Khoshgoftaar, H. Wang, N. Seliya, Choosing software metrics for defect prediction: an investigation on feature selection techniques, Software: Practice and Experience 41 (5) (2011) 579–606. doi:10.1002/spe.1043.
URL `http://doi.wiley.com/10.1002/spe.1043`

[50] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA data mining software: an update, ACM SIGKDD explorations newsletter 11 (1) (2009) 10–18.
URL `http://dl.acm.org/citation.cfm?id=1656278`

[51] J. Yang, V. Honavar, Feature subset selection using a genetic algorithm, in: Feature extraction, construction and selection, Springer, 1998, pp. 117–136.

[52] E. Aarts, J. Korst, Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing, John Wiley & Sons, Inc., New York, NY, USA, 1989.

[53] S. Li, K.-D. Yu, L. Fan, Y.-F. Hou, Z.-M. Shao, Predicting Breast Cancer Recurrence Following Breast-Conserving Therapy: A Single-Institution Analysis Consisting of 764 Chinese Breast Cancer Cases, Annals of Surgical Oncology 18 (9) (2011) 2492–2499. doi:10.1245/s10434-011-1626-2.
URL `http://www.springerlink.com/index/10.1245/s10434-011-1626-2`

[54] M. H. Zweig, G. Campbell, Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine., Clinical chemistry 39 (4) (1993) 561–577.
URL `http://www.clinchem.org/content/39/4/561.short`

[55] A. Shin, J. Joo, H.-R. Yang, J. Bak, Y. Park, J. Kim, J. H. Oh, B.-H. Nam, Risk Prediction Model for Colorectal Cancer: National Health Insurance Corporation Study, Korea, PLoS ONE 9 (2) (2014) e88079. doi:10.1371/journal.pone.0088079.
URL `http://dx.plos.org/10.1371/journal.pone.0088079`

[56] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, A comparative evaluation of feature ranking methods for high dimensional bioinformatics data, in: Information Reuse and Integration (IRI), 2011 IEEE International Conference on, IEEE, 2011, pp. 315–320.
URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6009566`

[57] D. Dittman, T. M. Khoshgoftaar, R. Wald, A. Napolitano, Random forest: A reliable tool for patient response prediction, in: Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on, IEEE, 2011, pp. 289–296.
URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6112389`

[58] W. W. LaMorte, Prospective versus Retrospective Studies (May 2016).
URL `http://sphweb.bumc.bu.edu/otlt/MPH-Modules/EP/EP713_CohortStudies/EP713_CohortStudies2.html`

[59] A. R. T. Donders, G. J. van der Heijden, T. Stijnen, K. G. Moons, Review: A gentle introduction to imputation of missing values, Journal of Clinical Epidemiology 59 (10) (2006) 1087–1091.

doi:10.1016/j.jclinepi.2006.01.014.
URL http://linkinghub.elsevier.com/retrieve/pii/S0895435606001971

[60] T. K. Moon, The expectation-maximization algorithm, Signal processing magazine, IEEE 13 (6) (1996) 47–60.

[61] S. Van Buuren, Flexible imputation of missing data, CRC press, 2012.

[62] G. Paré, L. Raymond, A. O. d. Guinea, P. Poba-Nzaou, M.-C. Trudel, J. Marsan, T. Micheneau, Electronic health record usage behaviors in primary care medical practices: A survey of family physicians in Canada, International Journal of Medical Informatics 84 (10) (2015) 857–867. doi:10.1016/j.ijmedinf.2015.07.005.
URL http://linkinghub.elsevier.com/retrieve/pii/S1386505615300228

[63] I. Spasić, J. Livsey, J. A. Keane, G. Nenadić, Text mining of cancer-related information: Review of current status and future directions, International Journal of Medical Informatics 83 (9) (2014) 605–623. doi:10.1016/j.ijmedinf.2014.06.009.
URL http://linkinghub.elsevier.com/retrieve/pii/S1386505614001105

[64] C. A. Kushida, D. A. Nichols, R. Jadrnicek, R. Miller, J. K. Walsh, K. Griffin, Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies, Medical care 50 (2012) S82–S101.

[65] J. M. Madden, M. D. Lakoma, D. Rusinak, C. Y. Lu, S. B. Soumerai, Missing clinical and behavioral health data in a large electronic health record (EHR) system, Journal of the American Medical Informatics Association (2016) ocw021doi:10.1093/jamia/ocw021.
URL http://jamia.oxfordjournals.org/lookup/doi/10.1093/jamia/ocw021

[66] L. Ahmadian, M. van Engen-Verheul, F. Bakhshi-Raiez, N. Peek, R. Cornet, N. F. de Keizer, The role of standardized data and terminological systems in computerized clinical decision support systems: Literature review and survey, International Journal of Medical Informatics 80 (2) (2011) 81–93. doi:10.1016/j.ijmedinf.2010.11.006.
URL http://linkinghub.elsevier.com/retrieve/pii/S1386505610002261