Artificial Intelligence In Medicine xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

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, United States
[b] Modernizing Medicine, Inc., United States

ARTICLE INFO

Keywords:
Cancer prediction
Cancer recurrence
Cancer relapse
Data mining
Machine learning
Electronic health records

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.

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 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

* Corresponding author.
E-mail addresses: arichter@fau.edu (A.N. Richter), khoshgof@fau.edu (T.M. Khoshgoftaar).

https://doi.org/10.1016/j.artmed.2018.06.002
Received 7 January 2017; Received in revised form 8 September 2017; Accepted 13 June 2018

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.

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 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.

Table 1
Cancer types and risk/recurrence prediction.

Cancer type	Prediction problem		Total
	Risk	Recurrence	
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

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

[1] http://www.ncbi.nlm.nih.gov/pubmed.
[2] https://scholar.google.com/.