Decision Support Systems 74 (2015) 150–161

Contents lists available at ScienceDirect

Decision Support Systems

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

Predicting overall survivability in comorbidity of cancers: A data mining approach

Hamed Majidi Zolbanin *, Dursun Delen, Amir Hassan Zadeh

Department of Management Science and Information Systems Spears School of Business, Oklahoma State University, Stillwater, OK, United States

ARTICLE INFO

Article history:
Received 7 February 2015
Received in revised form 1 April 2015
Accepted 2 April 2015
Available online 19 April 2015

Keywords:
Medical decision making
Comorbidity
Concurrent diseases
Concomitant diseases
Predictive modeling
Random forest

ABSTRACT

Cancer and other chronic diseases have constituted (and will do so at an increasing pace) a significant portion of healthcare costs in the United States in recent years. Although prior research has shown that diagnostic and treatment recommendations might be altered based on the severity of comorbidities, chronic diseases are still being investigated in isolation from one another in most cases. To illustrate the significance of concurrent chronic diseases in the course of treatment, this study uses SEER's cancer data to create two comorbid data sets: one for breast and female genital cancers and another for prostate and urinal cancers. Several popular machine learning techniques are then applied to the resultant data sets to build predictive models. Comparison of the results shows that having more information about comorbid conditions of patients can improve models' predictive power, which in turn, can help practitioners make better diagnostic and treatment decisions. Therefore, proper identification, recording, and use of patients' comorbidity status can potentially lower treatment costs and ease the healthcare related economic challenges.

1. Introduction

Cancer is the second leading cause of death in the United States.[1] It is also a major cause of death worldwide, especially (and ironically) in high income countries.[2] Research on causes and behavior of cancers has resulted in significant advances in our understanding of the disease over the past four decades. Even though cancer studies have traditionally been clinical and biological in nature, the recent technological advances have made data driven analytic studies a common complement.

The exploration of massive medical databases with the aid of new computational tools has confirmed the existence of coexisting diseases, including certain cancers. However, current medical research has a tendency to follow a reductionist approach to the study of ailments by investigating them in isolation from one another, rather than considering their interactions [57]. Recent findings urge taking a different stance toward comorbid diseases by denoting how coexisting illnesses might affect the diagnosis, treatment, and evaluation of treatment effectiveness, as well as survival of patients [2,20,21,26,27,45,61].

Yet, another equally important reason for the consideration of comorbidities is their impact on treatment costs, which in turn affect economies. According to the National Health Council, 133 million Americans are affected by incurable, ongoing chronic diseases, and this number is expected to grow to 157 million in 2020, with 81 million suffering from multiple conditions.[3] These figures find more salience with chronic conditions accounting for more than 75% of all healthcare costs. While in 2007, $1.3 trillion was reported as the adverse economic impact of chronic diseases, including cancer, it is projected to increase to $4.2 trillion for superfluous treatment costs and lost economic output.

Prior research has shown that cancer treatment recommendations might significantly be altered based on the severity of comorbidities. Specifically, the extent of the tumor spread is not the sole indicator of treatment. Instead, the overall health of the patient might have a greater weight in choosing the treatments [17]. Regarding the serious interplay of coexisting complications with all different phases of cancer treatment, and with the increasing trend in the development of intercurrent illnesses, the present cancer classification system needs to be revised, as it does not account for the severity of comorbid conditions [46]. Even if concurrent health issues are diagnosed and accounted for during the course of treatments, excluding them from general data sets or storing them in disconnected systems hampers prospective statistical analyses that might reveal useful patterns about their interplay. Similarly, elimination of comorbidity information may compromise the effectiveness of clinical decision support systems (CDSS). These systems "apply best-

* Corresponding author.
E-mail address: zolban@okstate.edu (H.M. Zolbanin).

[1] http://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm.

[2] http://www.who.int/mediacentre/factsheets/fs310/en/index1.html.

[3] http://www.nationalhealthcouncil.org/NHC_Files/Pdf_Files/AboutChronicDisease.pdf.

http://dx.doi.org/10.1016/j.dss.2015.04.003
0167-9236/

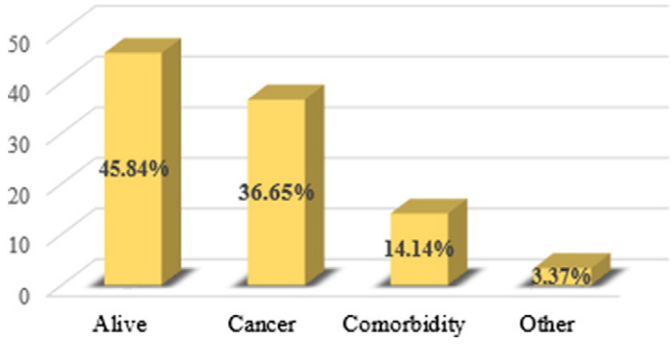

Fig. 1. Final outcomes in the SEER data (1973–2011).

known medical knowledge to patient data for the purpose of generating case-specific decision support" [68], particularly in preventive care services and treatment recommendation [6].

The reductionist approach to the study of diseases ignores parts of the potential interactions among comorbidities; thus, rendering the CDSS less effective. Prior studies on physicians' information needs have shown that in as many as 81% of clinical encounters in ambulatory care, clinicians may be missing critical information. As a result, providers confront serious challenges in accessing relevant information, obtaining a thorough picture of the patient's clinical state and history, and determining the optimal testing or therapeutic actions that should be further taken [54]. It seems, therefore, that collective consideration of concurrent diseases can potentially improve the quality of clinical decision support.

To demonstrate the importance of this issue, the current study seeks to investigate how concurrence of two cancers, namely urinary with male genital and breast with female genital, might affect the predictability of the disease outcomes. Improved predictability not only depreciates the reductionist approach, but can also help build more accurate clinical decision support systems, which in turn, would allow practitioners to make more effective decisions and eventually, lower overall healthcare costs.

2. Motivation

Comorbidity was first defined by Feinstein [22] as "any distinct additional clinical entity that has existed or may occur during the clinical course of a patient who has the index disease under study". Others have limited it to neoplasia, i.e., conditions and diseases that existed before a cancer diagnosis and are not adverse effects of cancer treatment [46], or to illness processes that coexist and are not related to the index disease under study [58]. Some authors have used such terms as intercurrent disease [28,30,55] or coexisting illness [44] interchangeably with comorbidity. Others have been more specific by making a subtle distinction between comorbidity and multimorbidity, which is simply defined as the presence of several diseases in one individual [57,58]. While some researchers refer to comorbidity as coexisting non-cancer medical conditions [18], others have made a distinction between cancer and non-cancer concurrent diseases; implicitly confirming that comorbidity can also be considered for the coexistence of two or more different cancers [19]. Although multimorbidity would be a better term for the coexistence of two cancers in a same patient, for the sake of simplicity, we will use these terms interchangeably.

The impact of comorbid health conditions on patients' overall survival cannot be overstated. In more than seven million unique cancer incidents in the Surveillance, Epidemiology, and End Results (SEER) data set between 1973 and 2011, almost 26% of all deaths were due to non-cancer comorbid causes. Comorbidities accounted for more than 14% of overall outcomes, including survivals. This information is illustrated in Fig. 1.

Investigation of historical cancer incidences reveals that certain types of cancers have higher correlations. Table 1 shows the number of patients who suffered from two different types of cancers during their lives. As it can be seen, urinary and male genital cancers, with 46,204 cases, co-develop the most frequently in a single patient's lifetime. Coexisting breast and female genital cancers stand third with 34,056 cases. Therefore, this study focuses on two of the most common comorbid subsets of SEER cases; i.e., those individuals who were diagnosed with male genitourinary or breast and female genital cancers during their lives.

Development of two types of cancers in one patient may occur at different ages. As stated previously, comorbidity (or more accurately, multimorbidity in a study of concurrent cancers) refers to coexisting ailments in one individual. Consequently, for the two cancers to be considered "comorbid", we limited our sample to those cases whose second cancer was diagnosed within one year after the first cancer's diagnosis. As a result, the new sample of comorbid genitourinary cancers composed of 14,243 cases. For breast and female genital cancers, the comorbid sample included 3664 cases. Figs. 2 and 3 show the distribution of patients' final outcomes in these samples.

Based on these figures, coexisting complications account for a significant proportion of deaths in the study's samples, and their impact is specifically greater in the male genital and urinary cancer patients. Comorbidities, including other cancers, cause more than 47% of deceases among genitourinary, and 37% among breast and female genital patients.

Investigating the impact of comorbid cancers on patients' final outcomes is the main impetus for this study. More specifically, we want to show how two concurrent diseases may change the predictability of disease outcomes. We will develop a set of predicting models for the aforementioned cancers, for each of the cancers alone and for their integration. The models' performances will be compared and the role of comorbidities will be discussed.

3. Background

The primary criterion in the traditional TNM system of cancer classification is the morphology of the carcinoma, and as such, it does not consider patients' overall health and comorbidity [45]. With the sizable body of literature that attests to the prognostic impact of factors other than tumor stage and treatment, comorbidities have found more

Table 1
Co-occurrence and total counts of different types of cancer.

Cancer type	Breast	Colon and rectum	Other digestive	Female genital	Leukemia and lymphoma	Male genital	Other	Respiratory	Urinary
Breast	1,275,422	24,259	11,687	34,056	14,559	809	31,584	22,032	10,478
Colon and rectum		841,426	14,856	12,280	12,501	34,553	21,498	21,531	17,147
Other digestive			619,246	6,251	6,000	18,219	13,395	10,461	8,859
Female genital				631,841	6,452	0	12,889	10,562	5,609
Leukemia and lymphoma					639,434	21,657	22,238	12,614	9,984
Male genital						1,117,558	43,989	33,688	46,204
Other							1,358,973	29,668	19,052
Respiratory								1,058,253	22,166
Urinary									522,376