Classification tree analysis of second neoplasms after childhood cancer

Jazbec Janez, MD, PhD, Division of oncology and hematology, Department of Pediatrics, Medical Centre Ljubljana, Slovenia

Ljupčo Todorovski, PhD, Institute Josef Stefan, Ljubljana, Slovenia

Jereb Berta, MD, PhD, Institute of Oncology, Ljubljana, Slovenia

Corresponding author: Janez Jazbec, Medical Centre ljubljana, Department of Pediatrics, Division of oncology and hematology, Vrazov trg 1, 1000 Ljubljana. Tel: + 386 1 522 8360, Fax: + 386 1 522 9360, e-mail: Janez.Jazbec@mf.uni-lj.si
Abstract

Reports on childhood cancer survivors estimated cumulative probability of developing secondary neoplasms vary from 3.3% to 25% at 25 years from diagnosis, and the risk of developing another cancer to several times greater than in the general population. In our retrospective study, we have used the classification tree multivariate method on a group of 849 first cancer survivors, to identify childhood cancer patients with the greatest risk for development of secondary neoplasms. In observed group of patients, 34 develop secondary neoplasm after treatment of primary cancer. Analysis of parameters present at the treatment of first cancer, exposed two groups of patients at the special risk for secondary neoplasm. First are female patients treated for Hodgkin’s disease at the age between 10 and 15 years, whose treatment included radiotherapy. Second group at special risk were male patients with acute lymphoblastic leukemia who were treated at the age between 4.6 and 6.6 years of age. The risk groups identified in our study are similar to the results of studies that used more conventional approaches. Usefulness of our approach in study of occurrence of second neoplasms should be confirmed in larger sample study, but user friendly presentation of results makes it attractive for further studies.
Introduction

As the number of childhood cancer survivors grows and the time of their observation gets longer, increasing attention is directed towards adverse effects of treatment. Late effects of treatment on many organ systems, such as cardiovascular, skeletal, endocrinologic, dental, hepatic, pulmonary and renal as well as psychosocial, educational and neuropsychological, have been described. Among the most serious of them are second neoplasms (SN) and the better the treatment results become for the primary malignancy, the more may long-term results expected to be compromised by secondary cancers (1). Reports on childhood cancer survivors estimated cumulative probability of developing SN vary from 3.3% to 25% at 25 years from diagnosis, and the risk of developing another cancer to be up to 35 times greater than in the general population (2). Occurrence of SN is the result of interaction of many independent factors to which the patient is exposed before, during and after treatment of the first malignancy. Some of those factors may have synergistic effect on development of SN. Design of prospective studies to identify those risk factors is difficult, due to the long follow-up period needed for development of SN. In our retrospective study, we have used the decision tree multivariate method to identify the group of childhood cancer patient with the greatest risk for development of SN.
Patients and methods

Patients

The study included 1577 patients registered at the Cancer Registry of Slovenia in the period from 1. 1. 1961 to 12. 10. 2000, who were younger than 16 years at diagnosis of the first malignancy. The decision tree analysis was performed on a group of 849 first cancer survivors, among whom 34 developed SN. In all patients who died, the cause of death was checked in original medical documentation. SNs were defined as neoplasms on a new location, that were not direct spread or metastases of the primary neoplasm, or neoplasms on the same location as the primary ones but of different histological type (3). Primary neoplasms were categorized according to histology as: leukemia, Hodgkin’s disease, non-Hodgkin lymphoma, Ewing sarcoma, osteogenic tumors, nephroblastoma, neuroblastoma, hepatoblastoma, rhabdomyosarcoma, retinoblastoma, thyroid cancer, germ-cell tumors, tumors of central nervous system (CNS) and others. The group of “others” consisted of carcinomas of different organ systems in 41 cases and two melanomas. They were grouped together because each particular group would be too small for further analysis. Data in the database included patient’s name, sex, date of birth, clinical diagnosis, histological type of the neoplasm, date of the diagnosis, treatment modality, date and status at the last follow-up. Detailed information on chemotherapy and radiotherapy was not included in the database. Table 1 presents the independent and dependent variables used for multivariate analysis.
Table 1:

All data were collected through the childhood cancer follow-up program in Slovenia. One pediatric-oncology center at Department of Pediatrics, University Medical Center Ljubljana, serves as a national referral center for all pediatric patients with malignant diseases. It covers the population of approximately 2 million. After the end of treatment all children are followed in the center until the end of schooling or at least four years after the end of treatment, after which they are followed at the outpatient clinic for late effects at the Institute of Oncology, where a team, headed by an oncologist known to the patient as a member of pediatric follow-up team, continues follow-up for life (4). Less than 5% of patients were lost to follow-up, because of permanent migration outside territory of Republic Slovenia. All of them were treated before 1990.
Classification tree analysis

Classification tree is a method for multivariate analysis, where a series of independent variables can be studied simultaneously in relation to one dependent variable. This is done by a successive division of the original group of cases into pairs of subgroups based on the value of a single independent variable. The variable that produces most pure subgroups of cases is chosen for division. A pure group of cases is a group where all the cases have the same outcome. Each of the pairs of subgroups then becomes a parent group and is further divided in the same way. The division stops when the group of cases is pure or when it contains less than operator-defined minimal number of cases. In our study, the C4.5 (5) program for building classification trees was used. C4.5 allows the setting of several parameters that influence the branching and quality of final classification tree. The optimal values of these parameters were determined using a standard cross-validation method (6, 7). Since the SN has been observed in only a minority of patients (about 4%), the classification tree tend to classify all the patients as non-SN cases. To avoid this, we used another C4.5 setting – the misclassification cost, which is used to vary the penalty for different misclassification types. In the analysis presented here, we have two types of misclassification: (1) a non-SN case classified as a SN case and (2) a SN case classified as a non-SN case. By default the penalties for both are the same, i.e., the misclassification cost setting is 1:1. Since by default, C4.5 trees tends to mis-classify all the SN cases into non-SN (misclassification type 2), we increase the cost of this misclassification type using different settings, starting with the default 1:1, through 1:5, 1:10, up to 1:25.
Results
The relatively low number of SN cases in the whole group resulted in a highly branched classification tree. In the analysis of the entire group of 1577 childhood cancer patients, the number of SN cases is below 3% and all the SN cases were classified as non-SN cases. Therefore we reduced our analysis on the group of children who survived their first cancer. There were 849 patients in this group and 34 developed SN. We have build several classification tree models with different misclassification costs in the algorithm. We considered misclassification of SN case in the group without SN as more severe mistake as vice versa. In the extreme case, with misclassification cost 5:1, we have built a tree where all SN cases were allocated in the group without SN. On the other side, if misclassification cost was set too high, there were too many cases without SN classified as patients with SN. In table 2. the classification results with different misclassification cost settings are presented.

On the basis of these results we were able to choose the setting that gave the lowest number of misclassified cases. Classification three that is build with regression tree analysis of our cohort is presented in figure 1.

Figure 1.

Despite optimal setting the branching of the tree is still considerable. There are many grains with individual SN cases and some clusters in which misclassified non-SN cases predominate. In the graphic presentation of the pruned tree the first factor that divides our cohort is radiotherapy. In the group of patients treated without radiotherapy, only 1.4% patients developed SN, which is considerable less than in the group of irradiated patients (5.8%). From these point we can follow two paths. The
first one encompasses patients with Hodgkin's disease and at the end, a group of female patients, aged between 10 and 15 years at the diagnosis and treated with chemotherapy, is exposed as a group in which risk for SN reaches 45%. The other path reveals a group of male patients with acute leukemia, who were old between 4,6 and 6,6 years of diagnosis of leukemia. In these group the risk for SN reaches 40%. 

Discussion

In our study of 849 childhood cancer survivors we have used multivariate analysis with classification trees to identify groups with special risk for development of SN. Girls with Hodgkin's disease, who were old between 10 and 16 years at the diagnosis and were treated with combination of chemotherapy and radiotherapy were identified as the group at highest risk. In all of the cases SN was carcinoma, with the time to development of SN ranging from 3 to 16.5 years after treatment of Hodgkin's disease. These results are similar to the observation of Beaty et. al. (9), who found statistically significant higher risk for SN in the group of adolescent patients with Hodgkin's disease. Bhatia and coworkers found 6.7 fold higher risk for SN in patients treated for Hodgkin's disease between 10 and 16 years of age (10). They also found the risk for secondary solid tumor after a combination of chemotherapy and radiotherapy twice as great as after chemotherapy without radiotherapy. It is possible that some tissues are particularly vulnerable for carcinogenic effect of chemotherapy and radiotherapy during intensive pubertal development. The challenge is to maintain the high rate of cure in Hodgkin's disease and at the same time reduce the risk for second malignancies. Some modern protocols of treatment of Hodkin's disease have already reduced or completely omitted radiotherapy for patients with low stages of disease. Löning et. al found radiation therapy as significant risk factor for SN after treatment of childhood acute lymphoblastic leukemia (11). This is in contrast with the results of Dalton et. al. (12). Löning also states that particularly young children are at increased risk when irradiation has been used. Intensive chemotherapy regimens do not predict a higher risk as reported in several studies (13, 14). In Childhood Cancer Survivor Study the diagnosis of leukemia was independently associated with the occurrence of
a malignant tumor of central nervous system, as was the younger age at the diagnosis (15).

While the improvement in survival rate of childhood cancer should not be overshadowed by the incidence of SN, the patients and health care providers should be aware of the population at the greatest risk and focus their efforts in primary and secondary prevention to groups at the highest risk. Using C 4.5 algorithm for building classification tree, we were able to construct subgroups at different risk, by logical combination of patients characteristics. The risk groups identified in our study are similar to the results of studies that used more conventional approaches. In contrast to traditional regression methods (e.g. Cox proportional hazard regression) which compute prognostic index as a weighted average of patient’s characteristics, in the classification tree model the subgroups are based directly on the patient’s characteristics. The model shows the correlation between the various independent variables and its influence on the end result (16). Another advantage of the method is in its simple and intuitive nature (i.e. find the best split by examining all possible splits in all available variables, form subgroups based on this split, repeat in all subgroups) (17). Classification trees are used in medical and health care applications for more than 20 years and have shown to be a powerful tool for classification support on different areas (18). In oncology the method has been used for developing tumor classification (19, 20), evaluation of biomarkers (21, 22, 23), Although the sample size represents a limitation of our study, the method used is potentially powerful tool for investigating multilevel interactions (24). As occurrence of secondary neoplasms may well be the result of complex interactions of several independent factors such as genetic predisposition, treatment related factors and environmental exposures, study of larger sample might prove our approach correct.
Literature:


**Table 1.** Description and values of the independent variables and the dependent variables (last row) used for multivariate analysis.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>male (485), female (346)</td>
</tr>
<tr>
<td>age_at_diagnosis</td>
<td>numeric</td>
</tr>
<tr>
<td>histology_type (categories)</td>
<td>leukemia, Hodgkin’s disease, non-Hodgkin lymphoma, Ewing sarcoma, osteogenic tumors, nephroblastoma, neuroblastoma, hepatoblastoma, rhabdomyosarcoma, retinoblastoma, thyroid cancer, germ-cell tumors, tumors of central nervous system, others</td>
</tr>
<tr>
<td>surgery</td>
<td>yes (481), no (368)</td>
</tr>
<tr>
<td>radiotherapy</td>
<td>yes (500), no (349)</td>
</tr>
<tr>
<td>chemotherapy</td>
<td>yes (598), no (251)</td>
</tr>
<tr>
<td>secondary_neoplasm</td>
<td>yes (34), no (815)</td>
</tr>
</tbody>
</table>
Table 2. Classification results in different misclassification cost settings.

a) **misclassification cost setting** 5:1, three pruning 0.01

<table>
<thead>
<tr>
<th>Classification</th>
<th>SN cases</th>
<th>Cases with no SN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>Without SN</td>
<td>0</td>
<td>815</td>
</tr>
</tbody>
</table>

b) **misclassification cost setting** 10:1, three pruning 0.01

<table>
<thead>
<tr>
<th>Classification</th>
<th>SN cases</th>
<th>Cases with no SN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>Without SN</td>
<td>66</td>
<td>749</td>
</tr>
</tbody>
</table>

c) **misclassification cost setting** 25:1, three pruning 0.5

<table>
<thead>
<tr>
<th>Classification</th>
<th>SN cases</th>
<th>Cases with no SN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>31</td>
<td>3</td>
</tr>
<tr>
<td>Without SN</td>
<td>18</td>
<td>797</td>
</tr>
</tbody>
</table>
Figure 1. Classification tree for risk of secondary neoplasm after treatment of cancer in childhood. Analysis of 849 first cancer survivors from Cancer Registry of Slovenia.

Legend: ST secondary tumor, HD Hodgkin's disease, KT, Chemotherapy, RT radiotherapy, SRG surgery, ALL acute lymphoblastic leukemia
Additional files provided with this submission:

Additional file 2: Table 2.doc: 21Kb  
http://www.biomedcentral.com/imedia/132023861066634/sup2.DOC

Additional file 1: Table 1.doc: 20Kb  
http://www.biomedcentral.com/imedia/957250851066634/sup1.DOC