Computers in Biology and Medicine

Computers in Biology and Medicine 36 (2006) 634–655

www.intl.elsevierhealth.com/journals/cobm

Patient-recognition data-mining model for BCG-plus interferon immunotherapy bladder cancer treatment

Shital C. Shah[a], Andrew Kusiak[a,*], Michael A. O'Donnell[b]

[a]*Intelligent Systems Laboratory, MIE, 3131 Seamans Center, The University of Iowa, Iowa City, IA 52242-1527, USA*
[b]*Department of Urology, University of Iowa Hospital and Clinics, IA, USA*

Received 28 January 2005; accepted 4 March 2005

Abstract

Bladder cancer is the fifth most common malignant disease in the United States with an annual incidence of around 63,210 new cases and 13,180 deaths. The cost for providing care for patients with bladder cancer disease is high. Bladder cancer treatment options such as immunotherapy, chemotherapy, radiation therapy, transurethral resection, and cystectomy, are used with varying success rates. In this research, data from a nationwide bacillus Calmette-Guérin (BCG) plus interferon-alpha (IFN-α) immunotherapy clinical trial was considered. Data mining algorithms were used to analyze the effectiveness of immunotherapy treatment and to understand the prominent parameters and their interactions. The extracted knowledge was used to build a patient recognition model for prediction of treatment outcomes. The data was analyzed to understand the impact of various parameters on the treatment outcome. A list of significant parameters such as cumulative tumor size, presence of residual disease, stages of prior bladder cancer, current state of bladder cancer, and the presence of current bladder cancer (T1) is provided. The decision-making approach outlined in the paper supplemented with additional knowledge bases will lead to a comprehensive analytical road map of the BCG/IFN-α immunotherapy treatment. It will provide individualized guidelines for each stage of the treatment as well as measure the success of the treatment.

Keywords: Data mining; BCG/IFN-α immunotherapy; Bladder cancer; Patient recognition model

* Corresponding author. Tel.: +1 319 335 5934; fax: +1 319 335 5669.
E-mail address: andrew-kusiak@uiowa.edu (A. Kusiak)
URL: http://www.icaen.uiowa.edu/~ankusiak.

0010-4825/$ - see front matter
doi:10.1016/j.compbiomed.2005.03.007

1. Introduction

1.1. Relevance of bladder cancer

Bladder cancer is the fifth most common malignant disease in the United States with an annual occurrence of around 63,219 new cases and 13,180 deaths [1]. In most industrialized western countries the situation is similar, with incidence rates of 18 to 30 new cases per 100,000 men. Bladder cancer is among the top 5 cancers in men while women are three times less likely to acquire this disease than men. The early superficial disease stage in 75% of patients is diagnosed due to gross or microscopic blood in the urine [2]. The overall recurrence rate of 65% and progression rate of 20% necessitates lifelong medical observation involving periodic internal inspections of bladders for even these patients.

Though the exact cause of bladder cancer is unclear, various factors such as exposure to certain aromatic chemicals, notably aniline dyes and benzidine compounds, cigarette smoking, and prior family bladder cancer history have a strong association with the disease [2,3]. Smoking causes the release of α-napthalene and β-napthalene into the urine, which is responsible for over 50% of the bladder cancers found in men and 33% of those found in women [4]. Over 90% of bladder cancers found in such settings are of the transitional cell type, with squamous cell carcinoma usually occurring in patients with chronic infections such as bilhariasis [2]. The most significant factors in accurate prognosis of bladder cancer are the depth of penetration (stage) and degree of cellular anaplasia (grade) among the transitional cell carcinomas. Carcinoma-in-situ, a mucosal high-grade lesion is not surgically accessible due to its diffuse surface-spreading behavior and if left untreated will progress to a muscle-invasive disease in up to 80% of cases within 5 years [5].

The intracavitary treatment of superficial bladder cancer is performed to prevent tumor recurrence after successful local surgical resection and to eradicate residual disease such as carcinoma-in-situ. Various intravesical chemotherapeutic agents such as thiotepa, doxorubicin, and mitomycin were the mainstay of therapy for years. They achieved short remissions with a net durable benefit for only 7–14% of patients [6,7]. More unconventional forms of treatment, such as immunotherapy were introduced due to unsatisfactory chemotherapy results [2]. One of the most successful therapies for superficial bladder cancer and carcinoma in situ of the bladder is intravesical administration of bacillus Calmette-Guérin (BCG) [8,9]. However, this therapy is unsuccessful in 25–40% of patients that never respond to BCG. Long-term remission (> 5 yr) is achieved in only 50% of patients [10–13]. Furthermore, toxicity associated with BCG therapy is frequent with incidence of severe adverse effects in 5% of patients and life-threatening symptoms in 0.5% of patients [14,15].

Recent scientific investigations have revealed BCG anti-tumor activity involves activation of a specific arm of the immune system referred to as T Helper Type 1 or TH1 [8]. Importantly, the addition of interferon-alpha (IFN-α) to the BCG mixture has been shown to significantly augment TH1 responses with the suggestion of improved efficacy in small clinical trials. On this basis a large (> 1100 patient) clinical trial of BCG/IFN-α therapy was begun both for previous BCG failures (prior BCG treated) and for those not previously treated with this agent (BCG naïve). An extensive data set of historical and prospective information was generated for this treatment group.

Means to accurately predict the clinical response to BCG therapy are currently lacking. If the latter were possible, patients at high risk for failure could be spared the morbidity of the treatment, offered alternatives, or proceed with radical therapy without incurring the risk of unfruitful delay. Thus, there is a need for measuring the effectiveness of BCG/IFN-α immunotherapy treatment and understanding the prominent parameters and their interactions. Currently, poor clinical outcomes from BCG therapy are thought to be due to a combination of either inadequate/inappropriate immune responses or intrinsic tumor resistance both of which may have indirect or direct associations with environmental or clinical parameters. The aim of this work is to develop a comprehensive analytical road map of the treatment involving various aspects of the therapy at various stages of the treatment. The hope is that this will provide individualized guidelines for each stage of the treatment as well as measure the progress and success of the treatment.

1.2. Hypotheses

The success of BCG/IFN-α immunotherapy treatment is measured by various factors such as prior history, symptomatic response (measurable by a Quantitative Symptom Score (QSS)), toxicity analysis, recurrence analysis, and long term follow up for later, less frequent events such as progression and survival. A goodness metric of the overall treatment can be formulated based on these factors. This metric should evolve over time with predefined milestone measurements.

The first phase of the metric building is the development of a patient recognition model (PRM), which is based on prior history before the treatment initiation. In the second phase (after the treatment initiation) status of the treatment response (measured every 3–6 months in accordance with the clinical surveillance pattern) can be devised based on prior history, QSS analysis, and toxicity analysis. In the final phase the recurrence analysis can be performed based on all the collected parameters as well as identified prominent parameters from other phases.

The main hypothesis for the development of the PRM is as follows:

{Main H_o: Applying data mining algorithms to the patient's disease history data can effectively predict the treatment outcome, and the rules/knowledge generated can form the basis for decision-making}

The treatment success can be evaluated based on various aspects of the patient's history such as prior treatments, toxic exposures, daily supplements (dietary and over-the counter drugs), other medical conditions, and so on. Thus various lower level hypotheses are formulated to support the main hypothesis in the development of PRM. The lower level hypotheses were developed based on prior literature and parameters of interest for domain experts. These hypotheses may involve a single parameter as well as the interactions among various factors.

Hypothesis 1: {H_o: Daily supplements increase the treatment success}
Hypothesis 2: {H_o: Prior unsuccessful BCG treatment predisposes to treatment failure}
Hypothesis 3: {H_o: A higher grade of stage T1 leads to treatment failure}
Hypothesis 4: {H_o: A higher cumulative tumor size predisposes toward treatment failure}
Hypothesis 5: {H_o: Cardiac medication is a detrimental factor for treatment success}
Hypothesis 6: {H_o: Smoking has no effect on the treatment success}
Hypothesis 7: {H_o: Skin eczema allergy (a Th1 antagonistic process) predisposes to treatment failure}
Hypothesis 8: {H_o: Vitamin E (antioxidant) along with vitamin C results in treatment success}

1.3. Computational intelligence models for BCG/IFN-α immunotherapy treatment

The first phase of PRM is of major significance. The PRM metric provides the information for admitting a patient into the treatment protocol based on the likelihood of treatment success. Thus patients will be well informed about various treatment options while the physicians will be provided with guidelines for the treatment success in the form of IF-THEN rules. This knowledge will allow customizing a treatment protocol to each individual patient.

The customization of the treatment is achieved by the use of computational intelligence tools. It provides algorithms and tools for identifying valid, novel, potentially useful, and ultimately understandable patterns from data and constructs high confidence predictions for individuals [16]. Discovering hidden patterns in the data may represent valuable knowledge that might lead to discoveries, i.e., control setting, treatment selection, and so on. As an emerging science, data mining draws from the existing theories, e.g., statistics, as well as contributes its own developments, e.g., the rough set theory [17]. The most widely used data mining algorithms are decision tree algorithms [18,19] decision rule algorithms [17], support vector machines [20–22], neural networks, and so on [23].

Data mining has two facets: knowledge discovery and decision-making. As a knowledge discovery tool, some data mining algorithms used in this research produce explicit knowledge (IF ... THEN rules) that can be analyzed by a user. The users may learn new knowledge and at the same time may pose questions to be addressed by a targeted research. The decision-making facet of data mining overlays with decision making and prediction theories, and it produces outcomes of three different types: high confidence decision (type I), low confidence decision (type II) and no-decision (type III). The outcomes of type I patients (subgroups corresponding to the discovered patterns) are expressed with high confidence based on the collected data. The subgroups of type II patients require more clinical attention, as their outcomes cannot be accurately predicted. This is due to conflicts in matching the extracted patterns caused by the lack of relevant data presented to the data mining algorithms. Type III patients will be of particular interest to the medical research community as there is no sufficient knowledge to make evidence-based decisions for these patients. The selected data mining algorithms will emphasize what is unique about a patient subpopulation (in particular a patient) rather than what is common about the patient population. Three classes of data mining algorithms were used to extract knowledge, namely the decision tree algorithm (C4.5) [18,19], a decision-rule algorithm (based on rough set theory) [17], and support vector machines (SVM) [20–22]. Employing bagging [24,25] and boosting [25,26] techniques further enhance the above algorithms.

2. Data preparation

2.1. Data set

The data used in this research has been collected at multiple locations, including the University of Iowa Hospitals and Clinics (UIHC) [27]. Data collection was based on known as well as unknown indicators of bladder cancer treatment responses. The information from 13 forms was categorized into two major sections, patient-provided (8 forms) and physician-provided (5 forms) information (see Table 1). The patient-provided information covered demographics, exposure risk factors, bladder cancer history, prior BCG history, and medical conditions for various physiological systems. The physician-provided information included current bladder cancer information, prior history, individual on-site physician's assessments, and periodic assessments. The parameters (nominal, ordinal, binary, and integer types) were of relevance to the success of current treatment as indicated in the literature as well as domain experts. All the parameters except periodic assessment parameters were collected before the BCG/IFN-α

Table 1
Information content

Patient-provided information			
Demographics	*Bladder cancer history*	*Other medical conditions*	
	Initial signs	Cancer activities	Respiratory conditions
	Prior treatments	Cardiovascular conditions	Skin conditions
	Reason for stopping	Immune system and infections	Bones and joint conditions
		Glandular - Endocrine disorders	Blood disorders
BCG history	*Exposures risk factors*	Gastrointestinal problems	Nutritional state
Side effects	Smoking	Neurological disorders	Urological conditions
BCG tolerances	Family history		
	Toxic exposure		
Daily supplements			
Aspirin	Aspirin like drugs	Vitamin B complex	Coumadin
Selenium	Tylenol	Vitamin A	Vitamin C
Herbal	Multivitamins	Vitamin E	None
On site physician-provided bladder cancer information			
Current information	*History*	*Global assessment*	*Periodic assessment*
Grade	Clinical parameters	Summary cancer status	Every 3 months
Multicentricity	Prior treatments	Tumor recurrence estimates	3 to 24 months
Tumor architecture	Prior BCG history	Tumor progression estimates	
Tumor size			

immunotherapy treatment initiation. The periodic assessments were performed every 3 months starting at month three after the treatment starting date and ending at month 24.

The data sets were preprocessed to remove unwanted characters such as dashes, spaces, slash, and so on. The inconsistencies (such as None or none are equivalent) and similar parameter values (not applicable, not available, and none) in the data set were resolved with the help of domain experts. The "not applicable" parameter value signifies applicability of the parameter, for example not applicable for prior chemotherapy indicates absence of any prior chemotherapy treatment. Some of the categories were recoded so as to assist in data mining and the subsequent analysis. Some of the parameters such as prior treatment were in semi-colon separated format (Table 2). To analyze this information, the coding scheme was modified to form a column for each parameter value. For example, in Table 2 the prior treatment parameter was modified to form five additional columns with binary parameter values. The coding scheme provided information regarding the importance of the presence as well as absence of the parameter. Thus, a rule can provide information stating that prior treatment of chemotherapy is not a factor for the success of the BCG/IFN-α immunotherapy, and so on.

Data cleaning was performed on all the individual data sets and these data sets were merged to form a single data set. Data set was pruned to consider only patients who had completed 24 months and above of the BCG/IFN-α immunotherapy. The dimension of pruned data set was 668×221 (i.e., 688 patients and 221 parameters).

Table 2
Examples of creating additional informative parameters

Prior coding scheme		Present coding scheme				
Patient	Prior treatment	BCG	Int_alpha	other	chemo	none
1	BCG	1	0	0	0	0
2	Chemo; BCG	1	0	0	1	0
3	BCG; Int-alpha	1	1	0	0	0

The data set was further partitioned into two classes, namely the BCG naïve subpopulation and the prior BCG treated subpopulation. This partitioning was performed to analyze the response of each subpopulation to the treatment separately. All the previously treated patients that have failed BCG were likely candidates for poor immune response and intrinsic tumor resistance. The majority of the BCG naïve patients had either never experienced intravesical therapy or received chemotherapy only, a situation previously shown not to significantly influence a subsequent response to BCG. The parameter representing the prior BCG treatment column was utilized to perform the partitioning of the data set. The information provided by the patients as well as the physicians regarding the prior BCG related factors was removed from the analysis for the BCG naïve subpopulation. The dimensions of the BCG naïve and prior BCG treated subpopulation data sets were 366×165 and 302×201, respectively.

2.2. Decision formulation

The first 3 months of the BCG/IFN-α immunotherapy treatment was an induction regiment where six out of total 15 doses, separated by a week each, were administered. The Kaplan Meier analysis [27] indicated that approximately half of the total positive events occur at 3 months and a similar number of positive events occur over the next 15 months at a relatively constant rate (linear slope). The total recurrence curve thus displayed properties of at least second order kinetics and steadied out at and after 24 months. The total bladder cancer-free percentages at 24 months for the BCG naïve and prior BCG treated subpopulation are 57.3% and 42%, respectively. Thus, the assessment at the end of 24 months is indicative of the patient's long-term response to the treatment. The treatment response for the patients was assessed using standard-care bladder cancer monitoring every 3 months with cytoscopy, urinary cytology, and biopsies (when clinically indicated). As the patient recognition model is designed to measure the applicability of the BCG/IFN-α immunotherapy treatment for a patient, the ideal decision parameter is the assessment (positive, i.e., bladder cancer present, and negative i.e., bladder cancer not present) at 24 months.

2.3. Relevance-based data subset formulation

A micro-level (i.e., specific aspects of the hypotheses) analysis required formulation of a relevance-based data subset (Fig. 1). After initial partitioning of the data set into BCG naïve and prior BCG treated, each subpopulation provided parameters for the formation of eight relevance-based data subsets. The first data subset consisted of all relevant parameters for the BCG naïve subpopulation. The second data subset contained parameters corresponding to the daily supplements (Table 1) with the decision parameter.

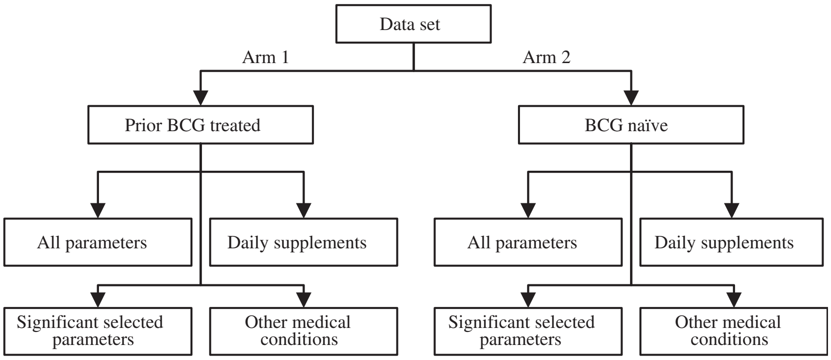

Fig. 1. Relevance based data subset formulation.

It included binary parameters for aspirin, multivitamin, tylenol, herbal vitamins A and E, and so on. The third data subset included the medical conditions (Table 1) of the patients such as cardiovascular, gastrointestinal, neurological, skin, blood, and so on. The last data subset for the BCG naïve subpopulation was comprised of significant parameters selected through the genetic algorithm and correlation-based heuristics (Table 3). Similarly, four data subsets were formulated for the prior BCG treated subpopulation (Tables 1 and 3).

2.4. Identification of significant parameter subsets

Parameter selection with a genetic algorithm (GA) [28] can be performed using two approaches, filter and wrapper search [29]. The wrapper search uses the machine-learning algorithm (e.g., decision tree) to evaluate the GA solutions [30,31]. The filter approach evaluates the parameters on heuristic-based general characteristics (for example, correlation) of the data. The correlation based parameter selection (CFS) filter is an effective way for parameter selection [29]. It selects a parameter if it correlates with the decision outcome but not with any other parameter that has already been selected. Thus, GA provides the global search framework for decision tree wrapper and the CFS filter, which in turn use their built-in functionality to optimize the parameters selected.

The GA and CFS were applied to both the BCG naïve and prior BCG treated subpopulation. The algorithm was replicated 10 times to ensure repeatability as well as exploration of the complete set of the possible significant parameter set. The higher value of the occurrence rate (i.e., number of times the parameter was selected) indicates the better quality of the selected parameter. A threshold on the number of parameters selected as well as the threshold occurrence rate can be set for inclusion in the final parameter set. The threshold for this research was set to parameters with an occurrence rate $\geqslant 6$ (i.e., above the 0.5 probability of inclusion by chance). Forty-one of the original 165 parameters for the BCG naïve subpopulation were significant (Table 3). Thirty-eight of the original 201 parameters for the prior BCG treated parameters were determined to be significant, i.e., a reduction of 81.1% (Table 3). In addition to data mining, an analysis of the individual parameter's effect on the outcome (positive or negative) was performed.

Table 3
Significant parameters

Significant parameters: BCG naïve		Significant parameters: Prior BCG treated	
1	Ethnicity	1	Recent_tobacco_use
2	Family_history_bladder_cancer_sibling	2	Family_history_bladder_cancer_grandparent
3	Family_history_bladder_cancer_grandparent	3	Family_history_bladder_cancer_aunt_uncle_cousin
4	Number_bladder_tumors_removed_TURBT	4	Family_history_bladder_cancer_children
5	Second_hand_smoke_exposure	5	Number_bladder_tumors_removed_TURBT
6	Length_time_first_sign_diagnosis_of_bladder_cancer	6	Toxic_exposure_to_cytoxan_chemotherapy
7	Primary_reason_for_avoiding_bladder_removal	7	Length_time_first_sign_diagnosis_of_bladder_cancer
8	First_other_non_skin_cancer	8	Number_of_individual_maintenance_treatments
9	First_other_non_skin_cancer_activity	9	Number_of_separate_courses_of_BCG
10	Second_other_non_skin_cancer	10	Primary_reason_for_avoiding_bladder_removal
11	Second_other_non_skin_cancer_activity	11	Feel_about_your_progress
12	Cardiovascular_history_of_artificial_heart_valve	12	Given_anti_tuberculosis_antibiotics_before
13	Cardiovascular_history_of_impaired_circulation	13	Permanent_effects_from_BCG_urination_pain
14	Cardiovascular_history_of_other_heart_circulatory _problems	14	Pattern_of_BCG_intolerance_uncontrolled_bladder _instability
15	General_medical_condition	15	First_other_non_skin_cancer
16	Respiratory_history_of_emphysema_COPD	16	Second_other_non_skin_cancer
17	Respiratory_history_of_other	17	Cardiovascular_history_of_high_blood_pressure
18	Glandular_history_of_diabetes_insulin_dependent	18	Glandular_history_of_diabetes_insulin_dependent
19	Gastrointestinal_history_of_cirrhosis	19	Neurological_surgery
20	Gastrointestinal_history_of_hepatitis_active	20	Skin_history_of_other
21	Gastrointestinal_history_of_other	21	Blood_medications
22	Gastrointestinal_history_of_pancreatitis	22	Nutrition_medications
23	Urological_history_of_prostate_disease	23	Nutrition_surgery
24	Urological_medications	24	Daily_supplements_vitamin_E
25	Neurological_history_of_degenerative_disease	25	Daily_supplements_aspirin_like_drugs
26	Skin_history_of_non_melanoma_skin_cancer	26	Daily_supplements_vitamin_C
27	Bone_history_of_chronic_injury_defect	27	Patients_performance_status
28	Bone_history_of_rheumatoid_arthritis	28	Highest_grade_if_current_bladder_cancer__is_T1
29	Nutritional_state	29	Cumulative_tumor_size
30	Daily_supplements_multivitamin	30	Cytology_result_prior_most_recent_TURBT_bx_cysto
31	State_current_bladder_cancer	31	Cytology_result_after_most_recent_TURBT_bx_cysto
32	Cumulative_tumor_size	32	Prior_chemotherapy
33	Cytology_result_after_most_recent_TURBT_bx_cysto	33	Stages_prior_bladder_cancer
34	Prior_chemotherapy	34	Permanent_effects_from_BCG_bladder_spasms
35	Administered_chemotherapy_after_last_TURBT	35	Estimated_success_of_study
36	Stages_prior_bladder_cancer	36	Intolerant_standard_intravesical_therapy
37	Prior_number_of_courses_with_intravesical_interferon	37	Appropriate_based_strictly_on_disease
38	Summary_cancer_status_at_expected_treatment_failed _prior_intravesical_therapy	38	Worst_side_effects_during_prior_BCG_treatment _frequency_of_urination
39	Residual_disease_presence		
40	Failed_prior_intravesical_therapy		
41	Intolerant_standard_intravesical_therapy		

3. Analysis and results

The support vector machines (SVM) [20–22] provide information about the optimal hyperplanes separating the two classes. The decision tree (DT) [18,19] and decision rule (RSA) [17] algorithms used in this research have produced rules in the following format, respectively:

DT_Rule1: IF Prior_chemotherapy = Not_performed AND Ethnicity = White AND First_other_non_skin_cancer_activity = Not_applicable AND Cummulative_tumor_size = 5 THEN Decision_24 = Positive [Strength = 12.97/1.91]

RS_Rule 1: IF Worst_side_effects_during_prior_BCG_treatment_frequency_of_urination in {Moderate, severe} AND Nutrition_medications = Not_applicable AND Daily_supplement_vitamin_C = Not administered THEN Decision_24 = Positive [6 (Support), 6 (Strength), 3.37% (Relative strength), 100.00% (Confidence)]

To evaluate the accuracy of the knowledge generated by data mining algorithms, a 10-fold cross-validation [32] was used, where a random 10% of the records were removed and remaining 90% were utilized to generate the rules. The 10% removed were then tested on the generated rule set. This process was repeated 10 times to ensure the generality of the rule sets for future predictions.

Classification accuracy (CA) is the ratio of correctly predicted over the total number of predictions made [33]. This can be applied to each decision outcome as well as for the overall data set. For example, the DT classification accuracy for the negative decision in Table 4 is $(131/(131 + 83)) * 100 = 61.21\%$. The confusion matrix provides information regarding the actual decision outcomes (row) and the predicted decision outcomes (column) for patients [33]. Ideally the actual and predicted decision outcome should both be the same (NN, i.e., negative predicted as negative and PP i.e., positive predicted as positive). However, due to misclassification, the actual and predicted decision outcome may not agree, which are represented by the NP (i.e., negative predicted as positive) and PN (i.e., positive predicted as negative).

3.1. All parameters analysis

The Kaplan Meier analysis for the BCG naïve and prior BCG treated subpopulation has a 57.3% and 42% cancer free population at the end of 24 months, respectively, which are represented by the zero classifiers (true population distribution) [27]. Data mining of the BCG naïve subpopulation with all parameters produced a classification accuracy of 55.19% for RS, 52.46% for DT, and 52.73% for SVM (Table 4). The bagging with DT as base classifier and boosting with DT as base classifier resulted in a classification accuracy of 53.55% and 52.73%, respectively. Though the above approaches (RS, DT, SVM, bagging with DT, and boosting with DT) provided better classification accuracy (sensitivity) for a negative outcome (∼ 60%), the performance was worse than that of true population distribution. RS, DT, and SVM data mining algorithms produced classification accuracy of 55.25%, 55.96%, and 52.98%, respectively for the prior BCG treated subpopulation (Table 4). Similar to BCG naïve subpopulation, the prior BCG treated subpopulation provides better classification accuracy for positive outcome (specificity), but the overall classification accuracy is worse than that of the true population distribution. However, the bagging approach with DT as the base classifier had a classification accuracy of 61.26%, a 2.32% absolute increase over the true population distribution (zero classifier) for the prior BCG treated population. This classifier has a high specificity (71.91%), indicating that positive outcomes are better predicted.

Table 4
Results: all parameters and significant parameters

Classifier	CA (%)	Absolute change CA (%)	% change CA (%)	Sensitivity (%)	Specificity (%)	NN	PP	NP	PN
BCG naïve: All parameters									
Zero	58.47	—	—	100.00	0.00	214	0	0	152
DT	52.46	−6.01	−10.28	61.21	40.13	131	61	83	91
SVM	52.73	−5.74	−9.81	60.28	42.11	129	64	85	88
RS	55.19	−3.28	−5.61	56.07	53.95	120	82	94	70
Bagging_DT	53.55	−4.92	−8.41	67.76	33.55	145	51	69	101
Boosting_DT	52.73	−5.74	−9.81	62.62	38.82	134	59	80	93
Prior BCG treated: All parameters									
Zero	58.94	—	—	0.00	100.00	0	178	124	0
DT	55.96	−2.98	−5.06	45.97	62.92	57	112	67	66
SVM	52.98	−5.96	−10.11	43.55	59.55	54	106	70	72
RS	55.25	−3.69	−6.26	80.65	37.64	100	67	24	111
Bagging_DT	61.26	2.32	3.93	45.97	71.91	57	128	67	50
Boosting_DT	56.95	−1.99	−3.37	40.32	68.54	50	122	74	56
BCG naïve: Significant parameters									
Zero	58.47	—	—	100.00	0.00	214	0	0	152
DT	54.64	−3.83	−6.54	63.55	42.11	136	64	78	88
SVM	58.74	0.27	0.47	70.09	42.76	150	65	64	87
RS	61.24	2.77	4.74	78.97	36.18	169	55	45	97
Bagging_DT	58.20	−0.27	−0.47	70.56	40.79	151	62	63	90
Boosting_DT	55.46	−3.01	−5.14	61.68	46.71	132	71	82	81
Bagging_SVM	60.38	1.91	3.27	74.30	40.79	159	62	55	90
Prior BCG treated: Significant parameters									
Zero	58.94	—	—	0.00	100.00	0	178	124	0
DT	56.95	−1.99	−3.37	40.32	68.54	50	122	74	56
SVM	63.25	4.30	7.30	52.42	70.79	65	126	59	52
RS	65.87	6.93	11.76	70.16	62.92	87	112	37	66
Bagging_DT	65.23	6.29	10.67	46.77	78.09	58	139	66	39
Boosting_DT	55.30	−3.64	−6.18	41.13	65.17	51	116	73	62
Bagging_SVM	63.91	4.97	8.43	56.45	69.10	70	123	54	55

The poor performance of most of the algorithms may have been due to three reasons namely, the disproportionate data set, the lack of a sufficient knowledge base, and a noisy data set. The disproportionate data set can be handled by re-balancing (e.g. one-sided selection) the data sets [34], cost-based sensitivity functions [35], and so on. As the subpopulation distribution for BCG naïve and prior BCG treated is roughly 57.3% (negative) to 42.7% (positive) and 42% (negative) to 58% (positive), respectively, the data sets can be considered proportionate [27]. The noise in the data set can be due to the presence of irrelevant parameters or an incorrect decision assignment. The decisions are based on a definitive clinical test, thus eliminating the possibility of an incorrect decision assignment. The presence of irrelevant parameters is handled by applying significant parameter selection algorithms on the subpopulation data sets. The selected significant parameters list retains only the relevant parameters with respect to the decision values. The lack of a sufficient knowledge base can be handled by adding new relevant parameters and/or through data transformation.

3.2. Significant parameters analysis

Mining the BCG naïve data set with significant parameters (derived by application of GA with a CFS objective function) (Table 3) produced rules with a classification accuracy of 61.24% for RS, a 4.74% relative change from the population distribution (Table 4). The classification accuracy significantly increased (by 6.05%) for the RS algorithm as compared to the all parameter BCG naïve data set. The significant parameters were just a small fraction (i.e., $\sim$ 24.85%) of all available parameters. The SVM, bagging with DT as a base classifier and bagging with SVM as a base classifier, performed either equivalent to or better than the true population distribution, indicating that the significant parameter list was able to remove irrelevant noisy parameters and increase classification accuracy. The sensitivity for the BCG naïve subpopulation was above 70%, while the specificity was 36–42%. Thus the negative patients were more likely to be identified correctly in the BCG naïve subpopulation.

A significant parameter data subset for prior BCG treated (similar to the BCG naïve significant parameter data subset) was created. This data subset consisted of 38 parameters and produced a classification accuracy of 63–66% for various algorithms (Table 4). The classification accuracy significantly increased (by 4–7%) over the true population distribution, indicating that the selected significant parameters were superior. The specificity for the prior BCG treated subpopulation was about 62–78%. Thus the positive patients were more likely to be correctly predicted in this subpopulation. The rules produced by the algorithms provided access into the prominent parameters, interactions, and individual parameter contributions for predicting the patient's treatment outcome. These rule sets provide the information to test the sub-hypotheses stated in Section 1.2. The prominent rules extracted by the data mining algorithms and their medical interpretation (provided by the physician) for both BCG naïve and prior BCG treated subpopulations are provided next. The rules should be interpreted with respect to the subpopulations (i.e., the initial conditions for the prior BCG treated subpopulation is that the patients should have had prior BCG treatments) and other rules.

BCG naïve

DT Rule 1: IF Ethnicity = White AND Prior_chemotherapy = Not_applicable AND Cardiovascular_history_of_impaired_circulation = 0 AND Second_other_non_skin_cancer_activity = Not_applicable AND Gastrointestinal_history_of_other = 1 AND Cardiovascular_history_of_other_heart_circulatory_problems = 0 THEN Decision_24 = Negative [24.84/1.0]

Interpretation: *No prior chemotherapy*, *i.e.*, *no history of prior chemotherapy failure*, *with no heart related conditions*, *with no second non-skin cancers*, *some other gastrointestinal history*, *i.e.*, *in all a healthy patient produce negative outcomes after 24 months*.

DT Rule 2: IF Ethnicity = White AND Second_other_non_skin_cancer_activity = Not_applicable AND First_other_non_skin_cancer_activity = Not_applicable AND Stages_prior_bladder_cancer = Ta AND Daily_supplements_multivitamin = 1 AND State_current_bladder_cancer = Ta THEN Decision_24 = Positive [5.53]

Interpretation: *Absence of any other non-skin cancer with prior as well as current bladder cancer as Ta* (*i.e.*, *the bladder cancer was not cured*), *with the intake of multivitamins resulted in unsuccessful treatment*.

DT Rule 3: IF Ethnicity = White AND Second_other_non_skin_cancer_activity = Not_applicable AND First_other_non_skin_cancer = Not_applicable AND Residual_disease_presence = None AND Administered_chemotherapy_after_last_TURBT = No AND Failed_prior_intravesical_therapy = Not_applicable AND State_current_bladder_cancer = T1 AND Cytology_Result_After_most_recent_TURBT_bx_Cysto = Not_done AND Daily_supplements_multivitamin = 0 AND Primary_reason_for_avoiding_bladder_removal = Not_applicable_never_advised THEN Decision_24 = Negative [29.89/5.84]

Interpretation: *Absence of any other non-skin cancer*, *residual disease*, *multivitamin*, *with no history of failed intravesical therapy*, *and never advised to remove bladder suggest the patients are in good health even though the current stage of bladder cancer is T*1. *Thus the possibility of successfully treatment is higher*.

DTRule 4: IF Summary_cancer_status_at_expected_treatment_failed_prior_intravesical_therapy = 0 AND Second_other_non_skin_cancer_activity = Not_applicable AND Residual_disease_presence = visible THEN Decision_24 = Positive [7.19/0.11]

Interpretation: *Absence of second non-skin cancer with a chance for success of intravesical therapy*, *but presence of visible residual disease results in a positive outcome*.

DT Rule 5: IF First_other_non_skin_cancer_activity = Not_applicable AND Ethnicity = White AND Primary_reason_for_avoiding_bladder_removal = Not_applicable_never_advised AND Prior_chemotherapy = Not_applicable AND Residual_disease_presence = None AND Respiratory_history_of_other = 0 AND Respiratory_history_of_emphysema_COPD = 0 AND Cytology_result_after_most_recent_TURBT_bx_cysto = Not_done AND Daily_supplements_multivitamin = 0 AND Stages_prior_bladder_cancer = No_prior_cancer THEN Decision_24 = Negative [38.97/5.54]

Interpretation: *Absence or non-applicability of various parameters suggests healthier individuals and thus the success of immunotherapy.*

DT Rule 6: IF Glandular_history_of_diabetes_insulin_dependent = 0 AND First_other_non_skin_cancer_activity = Not_applicable AND Ethnicity = White AND Neurological_history_of_degenerative_disease = 0 AND Prior_chemotherapy = Not_applicable AND Stages_prior_bladder_cancer = No_prior_cancer AND Second_hand_smoke_exposure = minimal AND Nutritional_state = 30_to_51_lbs_overweight THEN Decision_24 = Positive [13.51]

Interpretation: *Absence or non-applicability of various parameters but overweight individuals lead to treatment failure. Further investigation of nutritional state and second-hand smoking is needed.*

Prior BCG treated

DT Rule 1: IF Permanent_effects_from_BCG_bladder_spasms = 0 AND Cytology_result_prior_most_recent_TURBT_bx_cysto = Positive THEN Decision_24 = Positive [33.24/2.95]

Interpretation: *The positive cytology results prior to the most recent TURBT bx cysto along with no permanent BCG* (*bladder spasms*) *effects to produce non-responsive treatment outcomes.*

DT Rule 2: IF Daily_supplements_vitamin_E = 0 AND Number_of_individual_maintenance_treatments = 6 THEN Decision_24 = Positive [12.0]

Interpretation: *A large number of individual maintenance treatments with no daily supplements of vitamin E results in a positive outcome.*

DT Rule 3: IF First_other_non_skin_cancer = Not_applicable AND Highest_grade_if_current_bladder_cancer_is_T1 = 3_or_4_high THEN Decision_24 = Positive [11.39/2.0]

Interpretation: *The higher grades of T*1 (*T*1 *in general is a worse disease*) *along with no other non-skin cancer results in a positive outcome.*

DT Rule 4: IF Cytology_result_prior_most_recent_TURBT_bx_cysto = Negative AND Number_of_individual_maintenance_treatments = 0 THEN Decision_24 = Negative [13.95/1.0]

Interpretation: *No individual maintenance treatment with cytology results prior to the most recent TURBT Bx cysto is a beneficial condition for treatment success*.

DT Rule 5: IF Appropriate_based_strictly_on_disease = Not_applicable AND Highest_grade_if_current_bladder_cancer_is_T1 = 1_to_2_low_to_int THEN Decision_24 = Negative [6.0]

Interpretation: *The lower grades of $T1$ ($T1$ in general is a worse disease) along with no comments for appropriateness of the treatment results in a negative outcome.*

DT Rule 6: IF Permanent_effects_from_BCG_bladder_spasms = 0 AND Daily_supplements_vitamin_E = 0 AND Appropriate_based_strictly_on_disease = Med_approp_refused THEN Decision_24 = Positive [10.0]

Interpretation: *With medium appropriateness (i.e., recommended for refusal of treatment) of the treatment with no vitamin E and no permanent bladder spasms BCG related effects resulted into positive outcome.*

DT Rule 7: IF Daily_supplements_vitamin_E = 0 AND Blood_medications = yes THEN Decision_24 = Negative [7.03/1.03]

Interpretation: *With no vitamin E but on blood medications can lead to negative outcome. Further analysis of type of blood medication is necessary*.

DT Rule 8: IF Second_other_non_skin_cancer = Not_applicable AND First_other_non_skin_cancer = Not_applicable AND Highest_grade_if_current_bladder_cancer_is_T1 = Not_applicable AND Appropriate_based_strictly_on_disease = Not_applicable AND Number_bladder_tumors_removed_TURBT = 1 AND Feel_about_your_progress = Tumors_in_check AND Cardiovascular_history_of_high_blood_pressure = 1 THEN Decision_24 = Negative [6.0]

Interpretation: *Absence of any other non-skin cancer and $T1$ type bladder cancer, low number of bladder tumors removed, positive attitude regarding the health, and high blood pressure medications results in treatment success.*

DT Rule 9: IF Second_other_non_skin_cancer = Not_applicable AND First_other_non_skin_cancer = Not_applicable AND Cardiovascular_history_of_high_blood_pressure = 1 AND Current_ness_tobacco_use = 4 THEN Decision_24 = Positive [13.33]

Interpretation: *Absence of any other non-skin cancer, high blood pressure medications, with high use of tobacco is detrimental for the treatment success*.

DT Rule 10: IF Primary_reason_for_avoiding_bladder_removal = Not_applicable_never_advised THEN Decision_24 = Negative [4.0]

Interpretation: *As the removal of bladder is not applicable or never advised the prior and current condition of bladder cancer are suggested. Potentially the patients are in good conditions in-spite of prior BCG treatment and thus succeed with the BCG/IFN-α immunotherapy*.

DT Rule 11: IF Second_other_non_skin_cancer = Not_applicable AND Highest_grade_if_current_bladder_cancer_is_T1 = 1_to_2_low_to_int THEN Decision_24 = Negative [6.0]

Interpretation: *Absence of second non-skin cancer with smaller $T1$ cancer is beneficial for the treatment*.

DT Rule 12: IF Given_anti_tuberculosis_antibiotics_before = no AND Intolerant_standard_intravesical_therapy = Not_applicable AND Number_of_individual_maintenance_treatments = 0 THEN Decision_24 = Negative [42.56/12.56]

Interpretation: *Absence of prior anti-tuberculosis antibiotics, with no history of prior failed intravesical therapy, with no prior individual maintenance treatment is a good sign for BCG/IFN-α immunotherapy success*.

DT Rule 13: IF Appropriate_based_strictly_on_disease = Inapprop_high_risk THEN Decision_24 = Positive [7.0]

Interpretation: *High-risk patients with probable known history of BCG treatment failures are likely candidates for the current BCG/IFN-α immunotherapy failures*.

Table 5
Trends of significant parameter: BCG naïve

Parameter	Positive	Negative	Remarks
First_other_non_skin_cancer_activity	Active_surgery, Active	Inactive_chemo_surgery, Inactive_surgery, Surgery, Radiation	Isuff.
Residual_disease_presence	Visible, visible_CIS	No_CIS	Isuff.
Stages_prior_bladder_cancer	CIS	Ta_CIS	Isuff.
State_current_bladder_cancer	Tl_ CIS	Ta_CIS, Not_available	
Number_bladder_tumors_removed_TURBT	7	2,3	Isuff.
Administered_chemotherapy_after_last_TURBT	Yes_immed		Isuff.
Summary_cancer_status_at_expected_treatment	1 (Present)		Isuff.
failed_prior_intravesical_therapy			
Prior_chemotherapy	Others	Mil, Thi	
Respiratory_history_of_emphysema_COPD	1 (Present)		0 is 50–50
Cumulative_tumor_size	0, 5	2, 4	Isuff.
Skin_History_of_non_melanoma_skin_cancer		1 (Present)	Isuff.
Second_other_non_skin_cancer_activity			Isuff.
First_other_non_skin_cancer	Prostate		Isuff.
Cytology_result_after_most_recent_TURBT_bx_cysto			Isuff.
Daily_supplements_multivitamins			Isuff.
Glandular_history_of_diabetes_insulin_dependent	1 (Present)		Isuff.
Nutritional_state			Isuff.
Second_other_non_skin_cancer			Isuff.
Primary_reason_for_avoiding_bladder_removal			Isuff.
Bone_history_of_rheumatoid_arthritis	1 (Present)		Isuff.
Second_hand_smoke_exposure			Isuff.
Length_time_first_sign_diagnosis of bladder_cancer		3, 4.5	Isuff.
Bone_history_of_chronic injury_defect		1 (Present)	Isuff.
Urological_history_of_prostate_disease		1 (Present)	Isuff.
Cardiovascular_history_of_impaired_circulation		1 (Present)	Isuff.
Cardiovascular_history_of_other_heart_circulatory problems		1 (Present)	Isuff.
Respiratory_history_of_other			Isuff.
Gastrointestinal_history_of_other		1 (Present)	Isuff.
Urological_medications			Isuff.
Family_history_bladder_cancer_sibling		1 (Present)	Isuff.
Cardiovascular_history of_artificial_heart_valve			Isuff.
General_medical_condition			Isuff.
Failed_prior_intravesical_therapy	BCG	Not_applicable	
Neurological_history_of_degenerative_disease	0 (Absent)		
Ethnicity			

The rules generated by the algorithms had reduced complexity. The general trend of response for each prominent parameter per subpopulation was formulated based on the multiple rule sets (from bagging), occurrence of parameter values in various rules with respect to the decision class, and the analysis of individual patients (rule strength). No prior chemotherapy for both subpopulations (Tables 5 and 6) indicates that the patients have had no history of prior failures to chemotherapy, which makes the patients more likely to have negative outcomes after 24 months. The cardiac problems (Tables 5 and 6) such

Table 6
Trends of significant parameter: Prior BCG treated

Parameter	Positive	Negative	Remarks
Appropriate_based_strictly_on_disease	lnapprop_high_risk, Med_approp_refused		Positive outcome trend parameter
Number_of_individual_maintenance _treatments	> 2	Not applicable (< − 1 or n_a)	> = 2 is positive outcome
Highest_grade_if_current_bladder _cancer_is_Tl	Intermediate (2) and above	Low (1) to intermediate (2)	Intermediate (2) and above is positive
Prior_chemotherapy		None	No prior chemotherapy is beneficial
Cytology_Result_After_most _recent_TURBT_bx_cysto	Atypical, Not_done, Positive	Negative	Clearer relationship
Stages_prior_bladder_cancer	CIS, Tl.Ta_Tl		Ta is insufficient
Feel_about_your_progress		tumors_in_check	Clearer relationship
Daily_Supplements_aspirin_like_drugs	0 (Absent)	1 (Present)	Take aspirin and negative
Cytology_Result_Prior_most _recent_TURBT_bx_cysto	Not_done, Positive	negative	Clearer relationship
Number_of_separate_courses_of_BCG	2		0, 1 is insufficient
Worst_side_effects_during_prior _BCG_treatment frequency_of_urination		None	None means BCG can be in the bladder for longer
Blood_medications	Not_applicable, No	Yes	Taking blood medications beneficial
Length_time_first_sign_diagnosis _of_bladder_cancer	1, 2	4, 5	Not clear
Estimated_success_of_study	S1, S2, S3		Clearer relationship
Currentness_of_tobacco_use	4	3	Not clear
Given_anti_tuberculosis_antibiotics_before	Not_sure		No results in 50–50 in rules
Permanent_effects_from_ BCG_bladder_spasms		1 (Present)	Presence of spams may be beneficial
Nutrition_medications	No		Not taking nutrition medications is not beneficial
Intolerant_standard_intravesical_therapy		BCG	Not applicable is insufficient
Family_history_bladder_cancer _aunt_uncle_cousin		1 (Present)	Other values insufficient
Number_Bladder_Tumors _Removed_TURBT		4	Not clear
Daily_supplements_vitamin_C			Not clear
Patients_performance_status	2		Other values insufficient
Skin_history_of_other	0 (Absent)	1 (Present)	No other skin history is positive
Neurological_surgery	no		
Cardiovascular_history_of_high _blood_pressure	0 (Absent)		High bp tends to be positive
Cumulative_tumor_size	0, 2	3	
Primary_reason_for_avoiding _bladder_removal	Still_other_options	N_a_never_advised	
Nutrition_surgery	No	Not_applicable	

Table 6 (*continued*)

Parameter	Positive	Negative	Remarks
Family_history_bladder_cancer _grandparent		0	
Toxic_exposure_to_cytoxan_chemotherapy		0	
Glandular_history_of_diabetes_insulin_dependent		0	
Second_other_non_skin_cancer			Not clear
Daily_supplements_vitamin_E	0 (Absent)	1 (Present)	Taking vitamin E beneficial
First_other_non_skin_cancer	Colon, Prostate		Not clear

as other heart circulatory problems, impaired circulation, high blood pressure may require the intake of some aspirin, which may strengthen the immune system and thus result in a successful BCG treatment. Positive feeling regarding the progress of the bladder cancer (indicates less concerned attitude) may help in BCG/IFN-α immunotherapy treatment (Table 6) for prior BCG treated patients. More than two (i.e., higher) individual maintenance treatments (Table 6) for the prior BCG treated subpopulation indicate the failure to respond to the treatments, suggesting a strong BCG resistance. The visible or visible with CIS-type of residual disease presence, or the presence of activate or active with surgery, of first other non-skin cancer activity may be indicative of BCG/IFN-α immunotherapy failure (Table 5). High-risk and medium risk prior BCG treated patients are likely to fail the current BCG/IFN-α immunotherapy treatment (Table 6). Patients with prior as well as current bladder cancer stages as T1_CIS, Ta_T1 or T1 or CIS will probably not respond to the treatment, as T1 in general is a worse disease. Aspirin, aspirin-like drugs and vitamin E and vitamin C are beneficial for the success of BCG/IFN-α immunotherapy treatments. Vitamin E is a potent antioxidant that can react with damaging oxygen free radicals. Vitamin E also protects white blood cells, which play a major role in the immune system's defense against disease, thus resulting in successful treatment for some individuals. The lower or no frequency of urination side effects due to prior BCG treatments indicates that the BCG + IFN-α can stay longer in the bladder and thus create a higher immune response resulting in treatment success. Other parameter trends are presented in Tables 5 and 6. The parameter trend analysis along with higher strength rules provides information regarding a small subset of population, which can be predicted with minimum errors. This provides the physicians an on-site, convenient, and easy parameter-based prognosis tool.

3.3. *Other medical condition related parameters analysis*

The other medical condition data subset included medical conditions (Table 1) of the patients such as cardiovascular, gastrointestinal, neurological, skin, blood, and so on. The nature of these parameters was to provide information regarding the presence/absence of the parameter. For example, the cardiovascular history of impaired circulation parameters has only binary values with 0 being absent and 1 being present. This parameter failed to provide information regarding the severity of the impaired circulation, which may have a likely effect on the treatment. The binary medication and surgery history parameters did not provide information regarding the kind of medications, dosages, number of such medications, complexity of the surgery, number of such surgeries, and so on. Thus the analysis of the other medical history subset was performed to identify prominent general parameters that can be further expanded with additional

Table 7
Results: other medical conditions and daily supplements

Classifier	CA (%)	Absolute change CA (%)	% change CA (%)	Sensitivity (%)	Specificity (%)	NN	PP	NP	PN
BCG naïve: Other medical conditions									
Zero	58.47	—	—	100.00	0.00	214	0	0	152
DT	50.55	−7.92	−13.55	59.81	37.50	128	57	86	95
SVM	54.10	−4.37	−7.48	63.55	40.79	136	62	78	90
RS	52.47	−6.00	−10.26	49.07	57.24	105	87	109	65
Bagging_DT	50.55	−7.92	−13.55	63.55	32.24	136	49	78	103
Boosting_DT	50.82	−7.65	−13.08	62.62	34.21	134	52	80	100
Prior BCG treated: Other medical conditions									
Zero	58.94	—	—	0.00	100.00	0	178	124	0
DT	54.64	−4.30	−7.30	45.16	61.24	56	109	68	69
SVM	59.60	0.66	1.12	41.94	71.91	52	128	72	50
RS	51.30	−7.64	−12.96	35.48	62.36	44	111	80	67
Bagging_DT	65.23	6.29	10.67	47.58	77.53	59	138	65	40
Boosting_DT	57.62	−1.32	−2.25	46.77	65.17	58	116	66	62
BCG naïve: Daily supplements									
Zero	58.47	—	—	100.00	0.00	214	0	0	152
DT	59.02	0.55	0.93	75.23	36.18	161	55	53	97
SVM	54.64	−3.83	−6.54	83.18	14.47	178	22	36	130
RS	57.07	−1.40	−2.39	70.56	37.18	151	58	63	98
Bagging_DT	57.92	−0.55	−0.93	73.83	35.53	158	54	56	98
Boosting_DT	55.74	−2.73	−4.67	74.30	29.61	159	45	55	107
Prior BCG treated: Daily supplements									
Zero	58.94	—	—	0.00	100.00	0	178	124	0
DT	65.89	6.95	11.80	40.32	83.71	50	149	74	29
SVM	66.89	7.95	13.48	46.77	80.90	58	144	66	34
RS	65.92	6.98	11.84	35.48	87.08	44	155	80	23
Bagging_DT	65.56	6.62	11.24	37.90	84.83	47	151	77	27
Boosting_DT	62.91	3.97	6.74	37.10	80.90	46	144	78	34

information in future model refinements. For example, if the skin history of non-melanoma skin cancer is a prominent parameter, then additional information such as stage of the cancer, number of such cancers, tumor analysis, others related test, and so on can be included in the analysis.

The best classification accuracies of the BCG naïve and prior BCG treated subpopulations were 54.10% (SVM) and 65.23% (bagging with DT), respectively (Table 7). Bagging with DT for prior BCG treated subpopulations performed relatively better than the population distribution by 10.67%, while SVM performed the same as the population distribution.

The other medical condition analysis produces classification accuracies (rule sets) that were worse than the true population except for the bagging with DT for the prior BCG treated subpopulation. This clearly indicates the inadequacy of the generated knowledge base and the information content in the data sets. Thus the analysis should be interpreted in conjunction with other information. The majority of other medical condition parameters do not affect the outcome of the immunotherapy. The rules formed have the other medical condition parameter value as zero, indicating healthier individuals, which improves the chances of success for immunotherapy. The presence of seasonal allergies may weaken the immune response and thus may result in the failure of the BCG/IFN-α immunotherapy treatment. Diabetes is associated with immune dysfunction, thus the presence of a glandular history of insulin-dependent diabetes may reduce the probability of treatment success. The sample prominent rules extracted by the data mining algorithms are provided next.

BCG naïve

RS Rule 1. IF Cardiovascular_history_of_high_blood_pressure = 0 AND Respiratory_history_of_emphysema_COPD = 1 AND Bone_medications = Yes THEN Decision_24 = Negative [6, 6, 4.84%, 100.00%]

RS Rule 2. IF General_medical_condition is {Good, fair} AND Gastrointestinal_medications is {Not_applicable, Yes} AND Uroloical_history_of_prostate_disease = 1 AND Bone_history_of_non_rheumatoid_arthritis = 0 AND Immune_medications is {Not_applicable, Yes} AND Blood_Disorders_History = Not_applicable AND Nutrition_medications = No THEN Decision_24 = Positive [15, 15, 8.43%, 100.00%]

RS Rule 3. IF Glandular_history_of_diabetes_non_insulin_dependent = 1 AND Urological_medications = Not_applicable AND Nutritional_state in {Well_nourished, malnourished} THEN Decision_24 = Positive [17, 17, 9.55%, 100.00%]

RS Rule 4. IF Cardiovascular_history_of_impaired_circulation = 0 AND General_medical_condition is {Good, Fair} AND Respiratory_history_of_seasonal_allergies = 1 AND Bone_history_of_non_rheumatoid_arthritis = 1 THEN Decision_24 = Positive [6, 6, 3.37%, 100.00]

Prior BCG treated

RS Rule 1. IF Cardiac_medications = Not_applicable AND Gastrointestinal_surgery is {Not_applicable, No} AND Uroloical_history_of_None = 0 AND Urological_medications = No AND Bone_surgery is {Not_applicable, No} THEN Decision_24 = Negative [19, 19, 8.88%, 100.00%]

RS Rule 2. IF General_Medical_Condition = excellent AND Bone_history_of_non_rheumatoid_arthritis = 0 AND Bone_medications = No AND Bone_surgery = No AND Nutritional_state = Well_nourished THEN Decision_24 = Negative [6, 6, 2.80%, 100.00]

RS Rule 3. IF Cardiovascular_history_of_impaired_circulation = 0 AND Cardiac_surgery = No AND Glandular_history_of_diabetes_insulin_dependent = 1 THEN Decision_24 = Positive [4, 4, 2.63%, 100.00%]

RS Rule 4. IF General_medical_condition is {Fair, Poor} AND Respiratory_history_of_emphysema_COPD = 1 AND Glandular_history_of_diabetes_non_insulin_dependent = 1 THEN Decision_24 = Positive [9, 9, 5.92%, 100.00%]

3.4. Daily supplement related parameters analysis

The role of daily supplements in the success of the treatment is of interest to the physicians. Thus data sets with only daily supplements (binary parameter values) were formulated for BCG naïve and prior BCG treated subpopulations. The binary parameter values lacked the daily supplement-related information such as dosages, which may be crucial in determining their relevance.

The classification accuracies were around 54–59% and 63–67% for the BCG naïve and prior BCG treated subpopulations, respectively (Table 7). Though the accuracies were higher, the number of high confidence rules was smaller. Many approximate rules were generated from the RS algorithms for both subpopulations. The relative rule strengths were small, and they covered only a small number of patients. Thus the daily supplements were indeed significant/beneficial in the bladder cancer treatment for a certain small patient subpopulation in both BCG naïve and prior BCG treated. The results obtained through this analysis also highlighted the fact that the daily supplement knowledge base is not adequate to formulate high confidence predictions for all patients. The generated rules must be interpreted with other knowledge.

Nine prior BCG treated individuals taking daily supplements of vitamin E and B-complex and no aspirin had no recurrence of bladder cancer at 24 months. Aspirin-like drugs contributed to a higher treatment success for the prior BCG treated subpopulation. The treatment was unsuccessful for BCG naïve individuals who took multivitamins and vitamin C but not vitamin E. The sample prominent rules extracted by the data mining algorithms are provided next.

BCG

RS Rule 1.	IF Daily_supplements_multivitamin = 0 AND Daily_supplements_tylenol = 0 AND Daily_supplements_vitamin_E = 1 AND Daily_supplements_aspirinlike = 0 AND Daily_supplements_vitamin_C = 0 AND Daily_supplements_herbal = 0 AND Daily_supplements_selenium = 0 THEN Decision_24 = Negative [10, 10, 4.67%, 100.00%]
DT Rule 2.	IF Daily_supplements_multivitamin = 1 AND Daily_supplements_vitamin_C = 1 AND Daily_supplements_vitamin_E = 0 THEN Decision_24 = Positive [7.12/0.02]
DT Rule 3.	IF Daily_supplements_selenium = 1 AND Daily_supplements_vitamin_C = 1 THEN Decision_24 = Negative [8.13/1.11]

Prior BCG treated

RS Rule 1.	IF Daily_supplements_tylenol = 0 AND Daily_supplements_vitamin_E = 0 AND Daily_supplements_vitamin_C = 1 THEN Decision_24 = Positive [13, 13, 7.30%, 100.00%]
RS Rule 2.	IF Daily_supplements_aspirin = 0 AND Daily_supplements_vitamin_E = 1 AND Daily_supplements_Bcomplex = 1 AND Daily_supplements_coumadin = 0 AND Daily_supplements_selenium = 0 THEN Decision_24 = Negative [9, 9, 7.26%, 100.00%]
DT Rule 3.	IF Daily_supplements_aspirin_like_drugs = 1 THEN Decision_24 = Negative [22.22/6.15]

4. Future research

The addition of other data (i.e., genetic, initial treatment responses, and so on) with the present knowledge base will improve the probability of correct predictions. Due to the inherent nature and dynamics of the data, each data set (daily supplements, other medical conditions, significant parameters, and all parameters) provided treatment success information with varying degrees of confidence. These predictions can be used to create a weighted meta-decision-making algorithm [36]. More refined data mining algorithms can be applied to these data sets to produce reliable knowledge bases. Based on similar approach, PRM models can be built for various other bladder cancer treatment options such as chemotherapy, radiation therapy, transurethral resection, cystectomy, and so on. Another layer of processing will be required to combine the results of each PRM so as to determine the best (or least harmful) treatment option. This will provide patients with individualized risk and/or benefits regarding each treatment option so as to facilitate in making a well-informed decision. Various other models to address other aspects of bladder cancer BCG/IFN-α immunotherapy treatment can be built similar to the PRM models.

5. Conclusion

This research provided valuable knowledge to verify the known bladder cancer predictors as well as identified other unknown predictors. This will lead to a greater understanding of the BCG/IFN-α immunotherapy treatment and its applicability to individual patients. Mining a patient's history successfully predicted the treatment outcome for both BCG naïve and prior BCG treated subpopulations. Various hypotheses were successfully tested via relevance/significant parameter-based data mining, rule-set analysis, and general trend of each parameter. Various interactions and interdependencies were explored to build robust rule sets. The current PRM models based on significant parameters performed (classification accuracies above 60%) better than the true population distributions of both subpopulations. The identified significant parameters and parameter trends considerably reduce the burden of data collection for the next phases of the project/clinical trials as well as the computational burden of data mining algorithms.

The prominent trend parameters for the BCG naïve population were the activity of non-skin cancer, presence of residual disease, stages of prior bladder cancer, current state of bladder cancer, number of bladder tumors removed by TURBT, administered chemotherapy after last TURBT, expected treatment failed prior intravesical therapy, prior chemotherapy, respiratory history of emphysema COPD, cumulative tumor size, and skin history of non-melanoma skin cancer. The parameters for the prior BCG treated population were treatment appropriateness based strictly on disease, the number of individual maintenance treatments, the highest grade presence for current bladder cancer of T1, prior chemotherapy, cytology results after the most recent TURBT bx cysto, stages of prior bladder cancer, feel about your progress, cytology results prior to the most recent TURBT bx cysto, number of separate courses of BCG, frequency of urination (worst side effects during prior BCG treatment), blood medications, length of time from first sign diagnosis of bladder cancer, estimated success of study, aspirin, aspirin-like drugs, vitamin E, and current tobacco use. Other medical-conditions and daily-supplements analysis provides some parameters (such as respiratory emphysema COPD, seasonal allergies, high hypertension, vitamin A, C, and E, aspirin, aspirin-like medications, and so on) for further investigations. The addition of relevant parameters, type and dosages of BCG/IFN-α, additional information for other medical conditions and daily supplement parameters, knowledge base, patients, meta-decision-making, relabeling algorithms, and so on will lead to a robust PRM with enhanced and reliable predictions.

Acknowledgements

Our special thanks to Christina Leopold, Eric Axelson, Josh Hoffman, and Ryan Redington for assisting in this research.

References

[1] A. Jemal, T. Murray, E. Ward, A. Samuels, R.C. Tiwari, A. Ghafoor, E.J. Feuer, M.J. Thun, Cancer statistics, 2005, CA Cancer J. Clin. 55 (1) (2005) 10–30.
[2] A.B. Alexandroff, A.M. Jackson, M.A. O'Donnell, K. James, BCG immunotherapy of bladder cancer: 20 years on, The Lancet 353 (9165) (1999) 1689–1694.
[3] Mayo Clinic staff, Bladder cancer, Accessed on 21 September 2004 Website: http://www.mayoclinic.com/invoke.cfm?id = DS00177.
[4] E.L. Wynder, R. Goldsmith, The epidemiology of bladder cancer: a second look, Cancer 40 (3) (1977) 1246–1268.

[5] D.C. Utz, G.M. Farrow, Management of carcinoma-in-situ of the bladder: the case for surgical management, Urol. Clin. North Am. 7 (3) (1980) 533–541.
[6] A. Pawinski, R. Sylvester, K.H. Kurth, C. Bouffioux, A. Van Der Meijden, M.K.B. Parmar, L. Bijnens, A combined analysis of European organization for research and treatment of cancer, and medical research council randomized clinical trials for the prophylactic treatment of stage TaT1 bladder cancer, J. Urol. 156 (6) (1996) 1934–1941.
[7] D.L. Lamm, D.R. Riggs, C.L. Traynelis, U.O. Nseyo, Apparent failure of current intravesical chemotherapy prophylaxis to influence the long-term course of superficial transitional cell carcinoma of the bladder, J. Urol. 153 (5) (1995) 1444–1450.
[8] Y. Luo, X. Chen, T.M. Downs, W.C. Dewolf, M.A. O'Donnell, IFN-alpha 2B enhances Th1 cytokine responses in bladder cancer patients receiving Mycobacterium bovis bacillus Calmette-Guérin immunotherapy, J. Immunol. 162 (4) (1999) 2399–2405.
[9] M.A. O'Donnell, J. Krohn, W.C. DeWolf, Salvage intravesical therapy with interferon-alpha 2b plus low dose bacillus Calmette-Guérin is effective in patients with superficial bladder cancer in whom bacillus Calmette-Guérin alone previously failed, J. Urol. 166 (4) (2001) 1300–1304.
[10] L.R. Kavoussi, R.J. Torrence, D.P. Gillen, M.A. Hudson, E.O. Haaff, S.M. Dresner, T.L. Ratliff, W.J. Catalona, Results of 6 weekly intravesical bacillus Calmette-Guérin instillations on the treatment of superficial bladder tumors, J. Urol. 139 (5) (1988) 935–940.
[11] R. De Jager, P. Guinan, D. Lamm, O. Khanna, S. Brosman, J. De Kernion, R. Williams, C. Richardson, L. Muenz, D. Reitsma, et al., Long-term complete remission in bladder carcinoma in situ with intravesical TICE bacillus Calmette-Guérin. Overview analysis of six phase II clinical trials, Urology 38 (6) (1991) 507–513.
[12] H.W. Herr, C.M. Pinsky, W.F. Whitmore Jr., P.C. Sogani, H.F. Oettgen, M.R. Melamed, Long-term effect of intravesical bacillus Calmette-Guérin on flat carcinoma in situ of the bladder, J. Urol. 135 (2) (1986) 265–267.
[13] D.L. Lamm, Long-term results of intravesical therapy for superficial bladder cancer, Urol. Clin. North Am. 19 (3) (1992) 573–580.
[14] A. Steg, S. Adjiman, B. Debre, BCG therapy in superficial bladder tumours: complications and precautions, Eur. Urol. 21 (1992) 35–40.
[15] D.L. Lamm, P.M. Van Der Meijden, A. Morales, S.A. Brosman, W.J. Catalona, H.W. Herr, M.S. Soloway, A. Steg, F.M. Debruyne, Incidence and treatment of complications of bacillus Calmette-Guérin intravesical therapy in superficial bladder cancer, J. Urol. 147 (3) (1992) 596–600.
[16] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA, 1996.
[17] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Boston, MA, 1991.
[18] J.R. Quinlan, Induction of decision trees, Machine Learning 1 (1) (1986) 81–106.
[19] E. Frank, I.H. Witten, Generating Accurate Rule Sets Without Global Optimization, in: J. Shavlik (Ed.), Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, CA, 1998, pp. 144–151.
[20] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, NY, 1999.
[21] J. Platt, Fast Training of Support Vector Machines using Sequential Minimal Optimization, in: B. Scholkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods—Support Vector Learning, MIT Press, 1999, pp. 185–208.
[22] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy, Improvements to Platt's SMO Algorithm for SVM Classifier Design, Neural Comp. 13 (3) (2001) 637–649.
[23] A. Kusiak, Feature transformation methods in data mining, IEEE Transactions on Electronics Packaging Manufacturing 24 (3) (2001) 214–221.
[24] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140.
[25] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, CA, 2001.
[26] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: L. Saitta (Ed.), Proceedings of the International Conference on Machine Learning, ICML'96, Morgan Kaufmann, San Francisco, CA, 1996, pp. 148–156.
[27] M.A. O'Donnell, K. Lilli, C. Leopold, and The National Bacillus Calmette-Guérin/Interferon Phase 2 Investigator Group, Interim Results from A National Multicenter Phase II Trial of Combination Bacillus Calmette-Guérin Plus Interferon ALFA-2B for Superficial Bladder Cancer, J. Urol. 172 (3) (2004) 888–893.
[28] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley, Longman, New York, NY, 1989.

[29] M.A. Hall, L.A. Smith, Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper, in: A. Kumar, I. Russell (Eds.), Proceedings of the Florida Artificial Intelligence Research Symposium, Orlando, Florida, AAAI Press, Menlo Park, CA, 1999, pp. 235–239.
[30] H. Vafaie, K. DeJong, Genetic algorithms as a tool for restructuring feature space representations, in: Proceedings of the Seventh International Conference on tools with Artificial Intelligence, IEEE Computer Society Press, Los Alamitos, CA, 1995, pp. 8–11.
[31] L. Zhang, Y. Zhao, Z. Yang, J. Wang, Feature selection in recognition of handwritten Chinese characters, Proceedings of the 2002 International Conference on Machine Learning and Cybernetics 3 (2002) 1158–1162.
[32] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. Royal Stat. Soc. 36 (1974) 111–147.
[33] M. Kantaradzic, Data Mining: Concepts, Models, Methods, and Algorithms, IEEE Press and John Wiley, New York, NY, 2003.
[34] M. Kubat, S. Marvin, Addressing the Curse of Imbalanced Training Sets: One-Sided Selection, in: D.H. Fisher (Ed.), Proceedings of the Fourteenth International Conference on Machine Learning, Morgan Kaufmann, 1997, pp. 179–186.
[35] P.D. Turney, Cost-sensitive classification: empirical evaluation of hybrid genetic decision tree induction algorithm, J. Artificial Intell. Res. 2 (1995) 369–409.
[36] A. Kusiak, B. Dixon, S. Shah, Predicting survival time for kidney dialysis patients: a data mining approach, Comput. Biol. Med. 35 (4) (2005) 311–327.

Shital C. Shah is a Ph.D. student in Industrial Engineering at The University of Iowa, Iowa City. He completed M.S. in IE at The University of Alabama, Tuscaloosa and BE in production engineering at VJTI, Bombay, India. His area of interest is in computational intelligence, data mining, informatics, operations research, simulation, and decision support systems. He has published data mining and informatics related research papers. He is the member of ALPHA-PI-MU and INFORMS. His E-mail address is scshah@engineering.uiowa.edu

Andrew Kusiak is a Professor of Industrial Engineering at the University of Iowa, Iowa City. He is interested in theory and applications of computational intelligence, data mining, and optimization in healthcare, pharma, product development, and manufacturing. He has published research papers in journals sponsored by AAAI, IEEE, IIE, INFORMS, ESOR, IFIP, IFAC, IPE, ISPE, and SME. He speaks frequently on international meetings, conducts professional seminars, and consults for industrial corporations. He serves on the editorial boards of 16 journals, and edits book series. He is the Editor-in-Chief of the Journal of Intelligent Manufacturing. His E-mail address is andrew-kusiak@uiowa.edu

Michael A. O'Donnell is a Professor in Urology, Holden Cancer Center at The University of Iowa. O'Donnell received his M.D. degree from Duke. He was a resident in urology, clinical and research fellow, instructor, and assistant professor in surgery at Harvard and a biomedical research fellow at Beth Israel Deaconess Medical Center in Boston. He is currently Director of Urologic Oncology at the University of Iowa. The author of over 50 peer-reviewed manuscripts, 10 book chapters, and multiple reviews and abstracts, Dr. O'Donnell also serves on the editorial board of the Journal of Urology and CURE. He leads several national clinical trials using immune therapy to treat bladder cancer and holds patents in the genetic manipulation of BCG bacteria. His E-mail address is michael-odonnell@uiowa.edu