Kim *et al. BMC Medical Informatics and Decision Making* 2016, **16**(Suppl 1):63
DOI 10.1186/s12911-016-0297-0

BMC Medical Informatics and Decision Making

RESEARCH Open Access

Discovery of prostate specific antigen pattern to predict castration resistant prostate cancer of androgen deprivation therapy

Yejin Kim[1], Yong Hyun Park[2], Ji Youl Lee[2], In Young Choi[3] and Hwanjo Yu[1*]

From The ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics
Melbourne, Australia. 23 October 2015

Abstract

Background: Prostate specific antigen (PSA) is an important biomarker to monitor the response to the treatment, but has not been fully utilized as a whole sequence. We used a longitudinal biomarker PSA to discover a new prognostic pattern that predicts castration-resistant prostate cancer (CRPC) after androgen deprivation therapy.

Methods: We transformed the longitudinal PSA into a discrete sequence, used frequent sequential pattern mining to find candidate patterns from the sequences, and selected the most predictive and informative pattern among the candidates.

Results: Patients were less likely to be CRPC if, after PSA values reach nadir, the PSA decreases more than 0.048 ng/ml during a month, and the decrease occurs again. This pattern significantly increased the accuracy of predicting CRPC by supplementing information provided by existing PSA patterns such as pretreatment PSA.

Conclusions: This result can help clinicians to stratify men by the risk of CRPC and to determine the patient that needs intensive follow-up.

Keywords: Prostate specific antigen, Longitudinal biomarker, Frequent sequential pattern mining, Prediction

Background

Prostate cancer has been the most common cancer in men worldwide [1–3]; it accounted for 27 % of new cancer cases in 2014 [3]. Androgen deprivation therapy (ADT) is the primary treatment of metastatic prostate cancer. ADT is conducted by suppressing androgens by castration, inhibiting the action of androgen using competing compounds known as anti-androgens, or by combining these treatments. Unfortunately, some patients proceed to castration-resistant prostate cancer (CRPC), whereas others retain hormone-sensitive prostate cancer (HSPC). For those who will be possibly endangered to CRPC, intensive follow-up and additional systematic therapies are required. Thus clinicians must assess the risk of progression to CRPC.

Prostate specific antigen (PSA) has been an important biomarker for diagnosis and prognosis of ADT. PSA level is measured during follow-up to monitor the response to the treatment. Generally, PSA level decreases after ADT begins, reaches the lowest level (nadir), then stabilizes for some period. If the cancer develops, PSA level increases. PSA has been used in therapeutic decision making by stratifying the risk of development to CRPC [4]. Characteristics of PSA variation are summarized as patterns such as pretreatment PSA level, nadir, time to nadir, and doubling time; these patterns have clinical significances as prognostic factors to predict CRPC [5–10]. However, the accuracy of these patterns as predictors is still unclear. They are computed based on only one or two PSA values before or around nadir even though PSA accumulates consistently after the treatment.

Patterns generated from a fully-utilized PSA sequence may increase the accuracy of predicting CRPC. Although

*Correspondence: hwanjoyu@postech.ac.kr
[1]Department of Creative IT Engineering, POSTECH, Pohang, South Korea
Full list of author information is available at the end of the article

the latter parts of PSA accumulation reflect the progression to CRPC [4], they have been discarded for three reasons: (1) Collection of PSA data from electronic medical records has been limited [11]; (2) Computing the characteristic patterns with the whole PSA sequence is more complicated than with one or two representative values; and (3) The relationship between PSA after nadir and CRPC has not been fully quantified. However, the patterns of PSA after nadir can provide insight into this relationship.

Thus we aim to exploit longitudinal PSA data to discover a new prognostic pattern that predicts CRPC after ADT, and to demonstrate clinical significance of the new pattern. We will compare this pattern with existing patterns.

Methods

We described a framework that discovers the prognostic pattern from the longitudinal PSA. This framework consists of three parts: transformation, pattern mining, and pattern selection.

Materials

Patients. We exploited data in electronic medical records (EMRs) that include longitudinal PSA level and other clinical variables. The EMR data were from an observational longitudinal database at Seoul, St. Mary Hospital, Korea; the database has been described in detail previously [12]. Among 1068 men diagnosed from January 2006 to June 2012 at our institution, 458 were treated as ADT. We only included 370 men who had not received any other treatment such as radical prostatectomy or radiation therapy and for whom PSA level data were available.

Longitudinal PSA. Each patients had a set of longitudinal PSA values that were recorded during follow-up every one to six months. The time to nadir has been investigated as the primary time point at which the kinetics of PSA level changes [6]; thus we separated the longitudinal PSA sequences into before and after nadir. Among the total of 2883 PSA values, 1238 were before nadir and 1645 were after nadir. We excluded PSA after CRPC because we should predict CRPC before it occurs. Some patients had only before or after nadir. After separation, 261 men had PSA after nadir, and 306 men had PSA before nadir.

Other variables. Patients had 14 demographic and clinical features: Age, laboratory results of Alb, Plt, Hb, Ca; medication information on intermittent treatment, drug order; bone metastasis, clinical stage; Gleason score, MRI prostate volume, pretreatment PSA, nadir, time to nadir. Patients also had two outcome variables: dichotomous factor for CRPC occurrence, and the time to CRPC. The CRPC variables were determined after being reviewed by the single urologist (Y.H.P.). Missing values were imputed using random survival forest [13]. To provide clinical background profiles of patients, the representative characteristics of the patients that have PSA after nadir were evaluated. Existence of bone metastasis can be cause of CRPC [14], and CRPC patients are more likely to have high Gleason score [15]. The *p*-value of the most representative features to predict CRPC by univariate Cox regression was assessed (Table 1).

Transformation

PSA velocity

We first converted PSA level to PSA velocity (PSAV) [ng/(ml · mo)]:

$$PSAV = \frac{PSA_{t_2} - PSA_{t_1}}{t_2 - t_1}, \quad (1)$$

where PSA_{t_i} is PSA [ng/ml] at time t_i [mo] [16]; so PSA sequence was converted into PSAV sequence, which can capture directions and the amount of PSA change. PSAV $\geq$ 0 and $\leq$ 0 referred to increasing and decreasing PSA level per month, respectively. PSAV is sometimes expressed as logarithm [17], but we did not log-transform PSA because $\log(PSA)$ change means relative change (i.e. multiple of previous PSA) rather than absolute change. We discriminated small change from large change when the ratio of two PSA values was the same. For example, PSA changes from 0.003 to 0.001 and from 30 to 10 have the same $\log(PSA)$ changes, but different PSAV.

Table 1 Characteristics and *p*-value of patients that have PSA after nadir

	Total	CRPC	*p*-value
Number of patients	261	96	-
Mean follow-up ± s.d.	38.7 ± 3.5	15.8 ± 2.4	-
Mean age ± s.d.	74 ± 0.9	75.2 ± 1.7	0.011
Mean time to nadir ± s.d.	10.3 ± 1.3	8.9 ± 1.9	0.001
Mean pretreatment PSA ± s.d.	119.4 ± 39.4	151.5 ± 80.3	0.054
Mean PSA nadir ± s.d.	12.8 ± 11.7	6.6 ± 4.2	0.142
Bone metastasis			
Yes	41	15	0.490
No	220	79	
Gleason score			0.117
≤ 6	59 (22.6 %)	20 (21.2 %)	
7	72 (27.5 %)	21 (22.3 %)	
≥ 8	130 (49.8 %)	53 (56.3 %)	
ADT type			
Leuprin only	154	44	0.012
Zoladex only	36	10	
Leuprin → Zoladex	9	6	
Zoladex → Leuprin	49	27	
Anti-androgen only	13	7	

Discretization

PSAV was abstracted using discretization methods because continuous PSAV contains noise that might reduce its generalizability. We denoted PSAV state as the discretized PSAV; so PSAV sequence was converted to PSAV state sequence. Two discretization methods were used: equal-frequency binning and entropy-based discretization. We used both methods and compared them to avoid biased discretization split-points.

- Equal-frequency binning is an unsupervised discretization technique that splits continuous variables into a specified number of bins to have equal frequency. We set bin size to five. These PSAV states were labeled as Low (L_q), Medium low (ML_q), Medium (M_q), Medium high (MH_q), and High (H_q). This method did not distinguish PSAV of CRPC from PSAV of HSPC.
- Entropy-based discretization is a supervised discretization technique that finds split-points with minimum entropy, and recursively partitions the intervals until a stopping criterion is met [18]. PSAV values were separated into three PSAV states, which were labeled as Low (L_e), Medium (M_e), and High (H_e). Because entropy is increased when PSAV of CRPC and PSAV of HSPC are mixed, this method generated PSAV states so that PSAV values from CRPC and HSPC belonged to distinct PSAV state as much as possible.

As an example of the transformation using after-nadir longitudinal PSA by entropy-based discretization we consider the PSA sequence (Figs. 1 and 2) from a patient treated with two medications (Zoladex and Leuprin) intermittently. PSA level reached nadir at 15 months after treatment began (Fig. 1). The PSA values were converted into PSAVs (Fig. 2), and then PSAVs were assigned to one of the PSAV states (i.e. L_e, M_e, H_e) that separate PSAVs by the dashed lines (Fig. 2). The interval of PSAV states is computed by entropy-based discretization. PSAV values at $(\infty, -1), [-1, 6)$, and $[6, \infty)$ are discretized as L_e, M_e, and H_e, respectively. Consequently, the PSA sequence becomes the PSAV state sequence $M_e \rightarrow M_e \rightarrow M_e \rightarrow L_e \rightarrow M_e \rightarrow M_e \rightarrow M_e \rightarrow M_e \rightarrow M_e \rightarrow M_e \rightarrow L_e$.

Pattern mining

To fully utilize longitudinal PSA data, we split PSAV state sequence into before- and after-nadir PSAV state sequences, and investigated the whole PSAV state sequence at the same time. We then used the frequent sequential pattern mining (FSPM) method to find new prognostic patterns. This method is the most widely used method for a set of discrete sequences. This method is also more computationally advantageous for a set of short and single sequences than are other methods that were mostly devised to analyze heterogeneous and large-scale data [19–22]. The PSAV state sequences were also short due to the short follow-up periods. We used FSPM to find candidate prognostic PSA patterns.

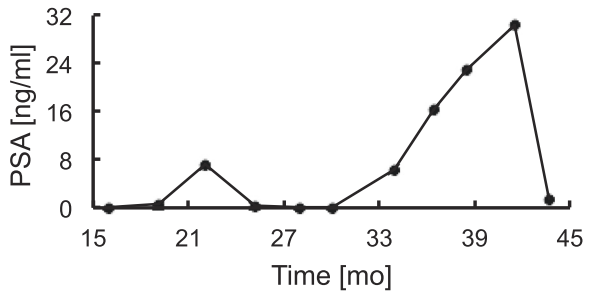

Fig. 1 PSA sequence. Example of transformation based on after-nadir longitudinal PSA by entropy-bases discretization

Particularly we used PrefixSpan algorithm for FSPM, which is a pattern-growth approach that builds prefix patterns that concatenate with suffix patterns to find frequent patterns [18, 23]. For examples, let assume that we have PSAV state sequences in Table 2 and minimal frequency of 0.1. We began with length-one prefix. The number of instance of length-one sequential patterns is L: 5, ML: 4, M: 6, MH: 1, H: 2. We discarded MH with frequency (=1/18) < 0.1. We then divided the search space with each prefix and searched sequential patterns starting with the prefix. We listed up the subset of PSAV state sequences starting with the prefix and discarded PSAV state sequence with frequency < 0.1. We repeated these process until the prefix becomes the whole PSA state sequence.

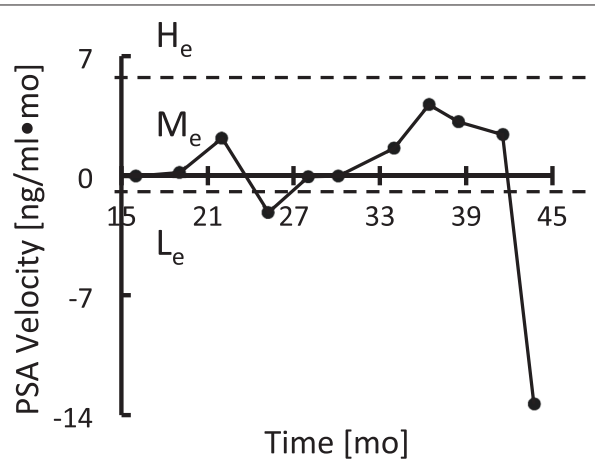

Fig. 2 PSAV sequence and PSAV state. Example of transformation based on after-nadir longitudinal PSA by entropy-bases discretization. PSA sequence (Fig. 1) was transformed to PSAV sequence (Fig. 2), then PSAV state (L_e, M_e, H_e) sequence. Dashed lines: borders between PSAV states (labeled)

Table 2 Example of PrefixSpan

	PSAV state sequences
1	$L \rightarrow L \rightarrow L \rightarrow MH \rightarrow M$
2	$L \rightarrow M \rightarrow ML \rightarrow L$
3	$H \rightarrow ML \rightarrow M \rightarrow M \rightarrow ML$
4	$H \rightarrow M \rightarrow L \rightarrow M \rightarrow ML \rightarrow M$

We restricted the patterns to have a support ≥ 0.3 to ensure the enough frequency, and a length ≤ 3 to avoid over-fitting. The discovered set of candidate patterns were the subset of PSAV state sequences. The patterns were time-ordered but not always consecutive.

Selection

Predictive pattern selection

To select the predictive patterns among the candidates set that were generated from FSPM, we evaluated the accuracy by measuring the area under the receiver-operating characteristic curve (AUC) and Harrell's concordance index (C-index) [24]. Each candidate pattern p was used as the predictor of CRPC because, by contraposition, if HSPC patients have a pattern p, a patient without the pattern p would be CRPC, and vice versa. We used baseline set containing the 14 demographic and clinical features that were extracted from the EMRs. We added each candidate pattern p to the baseline B (i.e. $B \cup p$) to evaluate the pattern p. We evaluated the discriminative power of the pattern p using

$$\triangle \text{AUC} = \text{AUC}(B \cup p) - \text{AUC}(B), \tag{2}$$

and

$$\triangle \text{C-index} = \text{C-index}(B \cup p) - \text{C-index}(B) \tag{3}$$

where AUC(X) denotes AUC of a logistic regression to predict CRPC using dataset X, and C-index(X) denotes the C-index of a Cox regression to predict time to CRPC using dataset X. They represent the net increase in AUC and C-index compared to the baseline. Ten-fold cross validation was used. A paired t-test with 95 % confidence level was conducted to identify patterns that increase the AUC and C-index significantly; patterns that had p-value $\leq$ 0.05 were excluded. The remaining significantly predictive patterns among the candidate set became the final candidate set from which the last pattern was selected.

Informative pattern selection

Among the significantly predictive patterns, we chose the most informative pattern. We preferred specific and rare patterns to broad and prevalent ones if the patterns have similar accuracy, because the former pattern provides relatively more information than the latter one. We formulated the relative information as follows:

Lemma 1. *Let p_1, p_2 denote patterns that the prediction accuracy are not significantly different, and let $I(p)$ denote the relative amount of information expressed by pattern p. Then $I(p_2) \leq I(p_1)$ if*

1. *p_2 is a sub-pattern of p_1 or*
2. *The interval of p_1 is a subset of interval of p_2 or*
3. *The frequency of p_1 is smaller than frequency of p_2.*

Cases 1 and 2 indicate that all patients that have p_1 also have p_2; thus p_1 is more specific than p_2. For example, p_2 is the sub-pattern of p_1 if $p_1 = L \rightarrow L$, $p_2 = L$ because L is a sub-pattern of $L \rightarrow L$ (Case 1). When $p_1 = L_e$, $p_2 = L_q$ where L_e has the interval of PSAV ≤ -0.048, and L_q has the interval PSAV ≤ -0.005, then p_1 is more specific than p_2 (Case 2). Case 3 implies that p_1 occurs more rarely than p_2. If p_1 is rare than p_2 although p_1 and p_2 have similar prediction accuracy, it means that the amount of information that p_1 carries per instance. For example, if $p_1 = L_e$, $p_2 = M_q$ where the frequency of p_1 is 14.4 %, and the frequency of p_2 is 59.3 %, then p_1 is more informative than p_2 because p_1 is more rare than p_2 in spite of the same prediction accuracy. We compared the amount of information using Lemma 1, and selected from the candidate set the final prognostic pattern that has the largest amount of information.

Comparison

We compared the progression to CRPC of the final pattern with that of pretreatment PSA, nadir, and time to nadir, which are known as the prognosis factors of CRPC. We computed the log-rank statistics of Kaplan-Meier analysis to test survival difference between patients with and without the pattern. The thresholds of pretreatment, nadir, and time to nadir were 100 ng/ml, 0.2 ng/ml, and 12 months, respectively [6].

Software

We used R3.0.3 [25] with two packages: `survival` [26, 27] for the significance test, Cox regression and log-rank test; and `randomSurvivalForest` [13, 27] for the imputation. We also used JAVA API, `SPMF` [28] to implement FSPM.

Results

We separated the longitudinal PSA data into before and after nadir. For the after-nadir dataset, we had 261 patients (HSPC: 167, CRPC: 94), and the mean follow-up time was 38.7 $\pm$ 3.5 months; the mean time to CRPC was 15.8 $\pm$ 2.4 months. Median PSAV was 0.022 ng/(ml $\cdot$ mo) (from -1521 to 2091). (Analysis of the before-nadir data did not reveal any prognostic patterns (Appendix A)).

We discretized the continuous PSAV values into five and three PSAV states by equal-frequency binning (Table 3) or

Table 3 Discretization to PSAV state after nadir by equal-frequency binning

PSAV state	Interval	
	Frequency [%]	Value [ng/(ml·mo)]
L_q	(0, 20]	(, −0.005]
ML_q	(20, 40]	(−0.005, 0.005]
M_q	(40, 60]	(0.005, 0.068]
MH_q	(60, 80]	(0.068, 0.454]
H_q	(80, 100]	(0.454,)

L_q = Low, ML_q = Medium low, M_q = Medium, MH_q = Medium high, H_q = High

entropy-based discretization (Table 4). Equal-frequency binning generated PSAV states of which frequency was evenly distributed over the five states. Entropy-based method generated three PSAV states; 85 % were M_e, 14 % were L_e, and 1 % were H_e. M_e occurred in both CRPC and HSPC, L_e occurred only in HSPC, and H_e occurred only in CRPC.

We found the candidate patterns from the set of after-nadir PSAV state sequences. Among the after-nadir 261 patients' sequences, we found 13 HSPC and 3 CRPC frequent patterns by equal-frequency binning (Table 5); and 6 HSPC and 2 CRPC frequent patterns by entropy-based discretization (Table 6). The PSAV state sequences from CRPC were rather uncommon among them; thus we could not find many frequent patterns from CRPC. In contrast, the PSAV state sequences from HSPC contained patterns that occurred repeatedly.

We computed the △AUC and △C-index when each candidate pattern was added to the baseline, and checked the significances of the △AUC and △C-index by calculating the p-values of paired t-tests. Five patterns from HSPC and the after-nadir were predictive in that they increased the AUC and C-index significantly (Table 7), but none of the patterns from CRPC or the before-nadir were predictive. The two equal-frequency binning patterns were observed: (1) L_q: PSA decline $\geq$ 0.005 ng/ml per month after nadir, (2) $L_q \rightarrow L_q$: two PSA declines $\geq$ 0.005 ng/ml per month after nadir; they showed the AUC of 0.81 and C-index of 0.78 – 0.79. The two entropy-based discretization patterns with L_e were observed: (1) L_e: PSA decline $\geq$ 0.048 ng/ml per month after nadir, (2) $L_e \rightarrow L_e$: two PSA declines $\geq$ 0.048 ng/ml per month after nadir; they showed the AUC of 0.81 – 0.82 and C-index of 0.77 – 0.81. One entropy-based discretization pattern with M_e before L_e was observed; this was $M_e \rightarrow L_e$: PSA decline from 0.048 to 5.43 ng/ml per month followed by PSA decline $\geq$ 0.048 ng/ml per month after nadir; this pattern showed the AUC of 0.84 and C-index of 0.81.

Table 4 Discretization to PSAV state after nadir by entropy-based discretization

PSAV state	Interval	
	Frequency [%]	Value [ng/(ml·mo)]
L_e	(0, 14.1]	(, −0.048]
M_e	(14.1, 99.2]	(−0.048, 5.43]
H_e	(99.2, 100]	(5.43,)

L_e = Low, M_e = Medium, H_e = High

Table 5 After-nadir candidate patterns by equal-frequency binning

Pattern (support)					
HSPC				CRPC	
L_q	(0.51),	$L_q \rightarrow L_q$	(0.34)	M_q	(0.45)
ML_q	(0.54),	$M_q \rightarrow L_q$	(0.31)	MH_q	(0.43)
M_q	(0.54),	$M_q \rightarrow M_q$	(0.35)	H_q	(0.36)
MH_q	(0.54),	$MH_q \rightarrow L_q$	(0.40)		
H_q	(0.48),	$MH_q \rightarrow H_q$	(0.31)		
		$MH_q \rightarrow MH_q$	(0.36)		
		$H_q \rightarrow L_q$	(0.34)		
		$H_q \rightarrow H_q$	(0.32)		

$L_q, ML_q, M_q, MH_q, H_q$ = PSAV state (Table 3)

The most informative pattern among the five predictive patterns was $L_e \rightarrow L_e$. Because the AUC and C-index among the predictive patterns were not significantly different, we compared the relative amount of information using Lemma 1 regardless of AUC and C-index. By Lemma 1.1,

$$\begin{aligned} &I(L_e) \leq I(L_e \rightarrow L_e), \\ &I(L_e) \leq I(M_e \rightarrow L_e), \\ &I(L_q) \leq I(L_q \rightarrow L_q). \end{aligned} \tag{4}$$

We then had $I(L_e \rightarrow L_e), I(M_e \rightarrow L_e)$, and $I(L_q \rightarrow L_q)$ after excluding patterns with small amounts of information. By Lemma 1.2,

$$I(L_q \rightarrow L_q) \leq I(L_e \rightarrow L_e). \tag{5}$$

Table 6 After-nadir candidate patterns by entropy-based discretization

Pattern (support)			
HSPC		CRPC	
L_e	(0.49)	M_e	(0.88)
M_e	(0.95)	$M_e \rightarrow M_e$	(0.55)
$L_e \rightarrow L_e$	(0.33)		
$L_e \rightarrow M_e$	(0.38)		
$M_e \rightarrow L_e$	(0.49)		
$M_e \rightarrow M_e$	(0.81)		

L_e, M_e = PSAV state (Table 4)

Table 7 Mean and s.d. of logistic regression and Cox regression when each candidate pattern is added to baseline

Pattern	Logistic regression		Cox regression	
	Mean AUC	s.d.	Mean C-index	s.d.
Baseline	0.6951	0.0686	0.6898	0.0429
L_q	0.8102	0.0549	0.7938	0.0323
$L_q \rightarrow L_q$	0.8110	0.0500	0.7815	0.0384
L_e	0.8103	0.0665	0.8090	0.0442
$L_e \rightarrow L_e$	0.8245	0.0352	0.7733	0.0623
$M_e \rightarrow L_e$	0.8446	0.0459	0.8174	0.0411

Finally, we compared $I(L_e \rightarrow L_e)$ and $I(M_e \rightarrow L_e)$ by Lemma 1.3,

$$I(M_e \rightarrow L_e) \leq I(L_e \rightarrow L_e). \quad (6)$$

Thus the final pattern was $L_e \rightarrow L_e$.

We conducted a Kaplan-Meier analysis of the pattern $L_e \rightarrow L_e$ and the other PSA patterns, and found that patients with $L_e \rightarrow L_e$ showed slow progressions to CRPC, and that patients without $L_e \rightarrow L_e$ showed fast progressions to CRPC (Figs. 3, 4, 5 and 6). The log-rank statistics of all PSA patterns had p-values ≤ 0.05. When compared with other PSA patterns, this pattern $L_e \rightarrow L_e$ had comparable prognostic power.

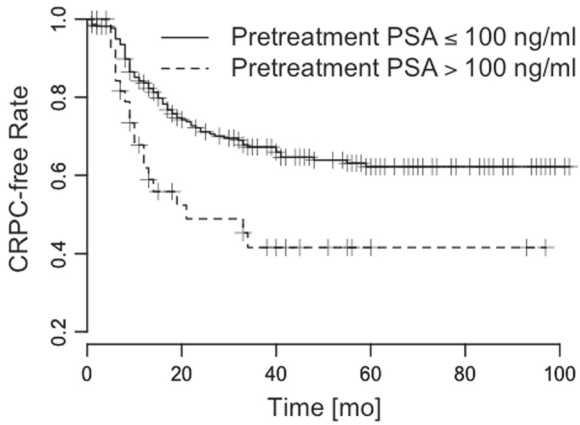

Fig. 4 Pretreatment PSA. Time vs. CRPC-free rate to measure survival difference. $p = 0.003$

Discussion

The objective of this study was to exploit the longitudinal measurements of PSA to discover a new prognostic pattern that predicts CRPC. The results of this study demonstrated that ADT patient is more likely to retain HSPC if, after PSA values reach nadir, PSA level decreases more than 0.048 ng/ml during a month, then the decrease occurs again; thus two PSA declines ≥ 0.048 ng/ml per month after nadir could be the prognostic pattern. This pattern was significantly related to the survival time to CRPC.

This finding has not been described in previous research on how different forms of PSA kinetics are associated with prognosis [5–10]. The most representative PSA prognostic patterns are pretreatment PSA, nadir, time to nadir or doubling time. Previous studies did not investigate all available PSA values, but instead used only one or two PSA values before or around nadir; thus the predictive value of PSA change during ADT was not clearly evaluated. In contrast, our pattern was computed from the whole PSA sequence, including even the latter parts. Although the initial response to ADT has important clinical significance, the subsequent response that is inferred from PSA after nadir might also contain information that can be used to predict CRPC. Incorporating PSA level as the sequence might enable us to understand the response to ADT more specifically.

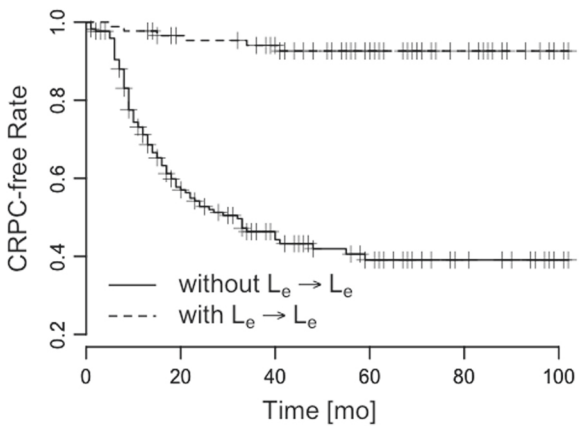

Fig. 3 Pattern $L_e \rightarrow L_e$. Time vs. CRPC-free rate to measure survival difference. $p \leq 0.001$

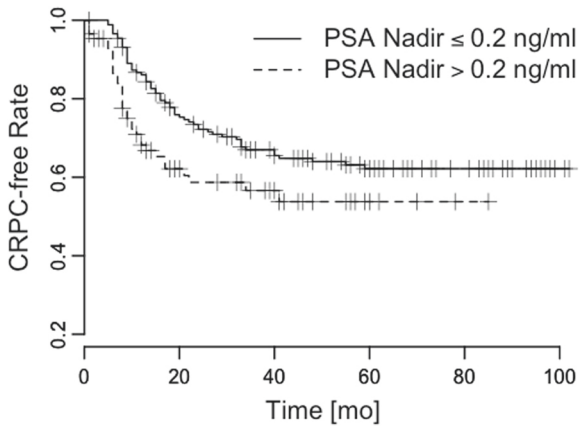

Fig. 5 PSA nadir. Time vs. CRPC-free rate to measure survival difference. $p = 0.031$

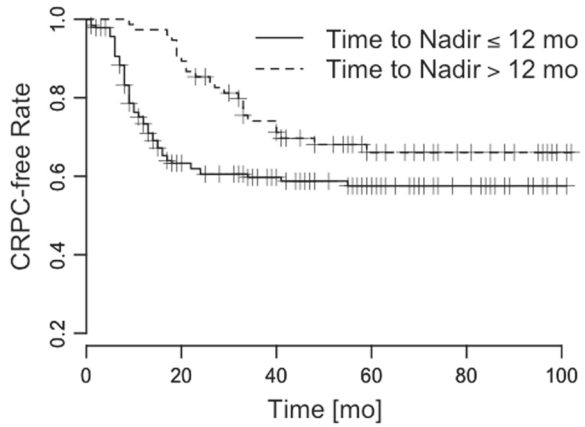

Fig. 6 Time to PSA nadir. Time vs. CRPC-free rate to measure survival difference. $p = 0.021$

We found that the two substantial declines after PSA nadir ensured sensitivity to ADT. The concept of the two substantial declines after PSA nadir can be confusing for clinicians, because nadir PSA means the lowest PSA value around a given point of observation. The occurrence of two substantial declines after PSA nadir implies that PSA fluctuates after nadir. PSA might increase due to some reasons such as intermittent treatment, and decline as sensitively reacting to ADT. We can further infer that the main difference between HSPC and CRPC is on whether PSA level fluctuates as the response to ADT.

The importance of this study is that the PSA decline after nadir helps to stratify men by the risk of CRPC and to determine the patient population that needs intensive follow-up. Risk assessment of the disease progression to CRPC has been based on the early PSA values, (i.e. before or around nadir) and has been limited to measuring the initial response. PSA after nadir has been neglected due to the complicated nature of computation, but we demonstrated that considering the after-nadir PSA pattern significantly increased the accuracy of the risk assessment by supplementing the early risk assessment obtained using the before-nadir PSA. Thus we can easily identify high-risk men who need in-depth follow-up. Therapeutic decision-making based on appropriate risk stratification enables clinicians to use clinical resource effectively.

This study has two main limitations. The first limitation is that the PSA decline pattern may occur at any time after nadir, so clinicians must wait until the pattern occurs, which must occur after the nadir. This means that clinicians must wait a long time to check whether PSA level declines; this delay is a disadvantage because rapid risk assessment is preferable when designing therapies for high-risk patients. However, FSPM with time constraints can solve this problem [18]. The time-gap between the PSAV states in the discovered pattern can be restricted. The PSAV states that occur within a specified gap reduce the time required to detect the occurrence than PSAV states without the gap. The second limitation is that analysis of the PSA decline pattern was focused on predicting HSPC. The median time to CRPC was only 15.8 months, whereas the median follow-up of all population was 38.7 months. The PSAV sequence from CRPC was not long enough to discover meaningful patterns, so most frequent patters were from HSPC. We predicted CRPC indirectly by predicting HSPC using the PSA decline pattern. A prognostic pattern that occurs frequently in CRPC can help detect CRPC directly, and this prognostic pattern from CRPC can be discovered if the quantity of data is increased and the follow-up time is extended.

Conclusions

This study discovered a prognostic PSA pattern that predicts CRPC for ADT using FSPM, and demonstrated the clinical significance of the pattern. A patient in which PSA declined twice by $\geq$ 0.048 ng/ml per month after nadir was predicted to retain HSPC, and a patient in which these declines did not occur was predicted to develop CRPC; the prediction had the AUC of 0.82 if the pattern was combined with pretreatment PSA, nadir, and time to nadir. These results can help risk stratification of ADT patients.

Appendix A: Results of before-nadir PSA values

For the before-nadir dataset, we had 306 patients (HSPC: 233, CRPC: 73), and the mean follow-up time was 37.7 ± 3.5 months; the mean time to CRPC was 17.5 ± 0.3 months. Median PSAV was -0.12 ng/(ml · mo) (from -1917 to 1686). After discretization, equal-frequency binning gave five discrete PSAV states (Table 8); but the entropy-based discretization could not be applied because the PSAV distributions of CRPC and HSPC were too similar. Among the 306 patients' PSAV state sequences, we discovered 6 HSPC and 5 CRPC frequent patterns from equal-frequency binning (Table 9). We computed the AUC and C-index when each frequent pattern is added to

Table 8 Discretization to PSAV state before nadir by equal-frequency binning

PSAV state	Interval	
	Frequency [%]	Value [ng/(ml·mo)]
L_{bq}	(0, 20]	(, −5.567]
ML_{bq}	(20, 40]	(−5.567, −0.611]
M_{bq}	(40, 60]	(−0.611, −0.026]
MH_{bq}	(60, 80]	(−0.026, 0.005]
H_{bq}	(80, 100]	(0.005,)

L_{bq} = Low, ML_{bq} = Medium low, M_{bq} = Medium, MH_{bq} = Medium high, H_{bq} = High

Table 9 Before-nadir candidate patterns by equal-frequency binning

Pattern (support)			
HSPC		CRPC	
L_{bq}	(0.78)	L_{bq}	(0.64)
ML_{bq}	(0.52)	ML_{bq}	(0.61)
M_{bq}	(0.41)	M_{bq}	(0.52)
MH_{bq}	(0.36)	$L_{bq} \rightarrow M_{bq}$	(0.31)
H_{bq}	(0.37)	$ML_{bq} \rightarrow M_{bq}$	(0.33)
$L_{bq} \rightarrow ML_{bq}$	(0.33)		

$L_{bq}, ML_{bq}, M_{bq}, MH_{bq}, H_{bq}$ = PSAV state (Table 8)

the baseline, but we could not find patterns that increase AUC and C-index significantly.

Availability of data and materials
Not available.

Authors' contributions
The first author YK contributed the majority of the writing and conducted major parts of the experiments. YHK contributed writing on discussion and provided important insights and clinical suggestions. JYL and IYC provided the motivation of this work and helpful comments on both method and discussion. HY provided helpful comments on method and presentation with detailed edits. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
This study was approved and performed in accordance with the approved guidelines by the Institutional Review Board at the Catholic University of Korea, Seoul St. Mary's Hospital (IRB number: KC15EISI0103).

Declarations
Publication charges for this work was partly supported by the Ministry of Science, ICT and Future Planning NIPA-2014-H0201-14-1001 "IT Consilience Creative Program", the ICT R&D program of MSIP/IITP B0101-15-0307 "Basic Software Research in Human-level Lifelong Machine Learning", and the Ministry of Education, Science and Technology No. 2012M3C4A7033344 "Next-Generation Information Computing Development Program through the National Research Foundation of Korea". This article has been published as part of BMC Medical Informatics and Decision Making Volume 16 Supplement 1, 2016: Proceedings of the ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics. The full contents of the supplement are available online at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-16-supplement-1.

Author details
[1]Department of Creative IT Engineering, POSTECH, Pohang, South Korea. [2]Department of Urology, Seoul St. Mary's Hospital, Seoul, South Korea. [3]Department of Medical Informatics, The Catholic University of Korea College of Medicine, Seoul, South Korea.

Published: 18 July 2016

References

1. Boyle P, Ferlay J. Cancer incidence and mortality in europe, 2004. Ann Oncol. 2005;16(3):481–8.
2. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, Thun MJ. Cancer statistics, 2008. CA: Cancer J Clinic. 2008;58(2):71–96.
3. Siegel R, Ma JM, Zou ZH, Jemal A. Cancer statistics, 2014. Ca-a Cancer J Clinic. 2014;64(1):9–29. doi:10.3322/Caac.21208.
4. Bubley GJ, Carducci M, Dahut W, Dawson N, Daliani D, Eisenberger M, Figg WD, Freidlin B, Halabi S, Hudes G. Eligibility and response guidelines for phase ii clinical trials in androgen-independent prostate cancer: recommendations from the prostate-specific antigen working group. J Clin Oncol. 1999;17(11):3461–7.
5. Park YH, Hwang IS, Jeong CW, Kim HH, Lee SE, Kwak C. Prostate specific antigen half-time and prostate specific antigen doubling time as predictors of response to androgen deprivation therapy for metastatic prostate cancer. J Urol. 2009;181(6):2520–5.
6. Kwak C, Jeong SJ, Park MS, Lee E, Lee SE. Prognostic significance of the nadir prostate specific antigen level after hormone therapy for prostate cancer. J Urol. 2002;168(3):995–1000.
7. Morote J, Trilla E, Esquena S, Abascal JM, Reventos J. Nadir prostate-specific antigen best predicts the progression to androgen-independent prostate cancer. Int J Cancer. 2004;108(6):877–81.
8. Hussain M, Tangen CM, Higano C, Schelhammer PF, Faulkner J, Crawford ED, Wilding G, Akdas A, Small EJ, Donnelly B. Absolute prostate-specific antigen value after androgen deprivation is a strong independent predictor of survival in new metastatic prostate cancer: data from southwest oncology group trial 9346 (int-0162). J Clinical Oncol. 2006;24(24):3984–90.
9. D'Amico AV, McLeod DG, Carroll PR, Cullen J, Chen MH. Time to an undetectable prostate-specific antigen (psa) after androgen suppression therapy for postoperative or postradiation psa recurrence and prostate cancer-specific mortality. Cancer. 2007;109(7):1290–5.
10. Choueiri TK, Xie W, D'Amico AV, Ross RW, Hu JC, Pomerantz M, Regan MM, Taplin ME, Kantoff PW, Sartor O. Time to prostate-specific antigen nadir independently predicts overall survival in patients who have metastatic hormone-sensitive prostate cancer treated with androgen-deprivation therapy. Cancer. 2009;115(5):981–7.
11. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, Taylor R. Can electronic medical record systems transform health care? potential health benefits, savings, and costs. Health Affairs. 2005;24(5):1103–17.
12. Choi I, Choi R, Lee J, Choi BG. Implementation of single source based hospital information system for the catholic medical center affiliated hospitals. Healthcare Inform Res. 2010;16(2):133–9.
13. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2(3):841–60. doi:10.1214/08-aoas169.
14. Montgomery RB, Mostaghel EA, Vessella R, Hess DL, Kalhorn TF, Higano CS, True LD, Nelson PS. Maintenance of intratumoral androgens in metastatic prostate cancer: a mechanism for castration-resistant tumor growth. Cancer Res. 2008;68(11):4447–54.
15. Gu Z, Thomas G, Yamashiro J, Shintaku I, Dorey F, Raitano A, Witte O, Said J, Loda M, Reiter R. Prostate stem cell antigen (psca) expression increases with high gleason score, advanced stage and bone metastasis in prostate cancer. Oncogene. 2000;19(10):1288–96.
16. Partin AW, Pound CR, Pearson JD, Clemens JQ, Landis PK, Epstein JI, Carter HB, Walsh PC. Evaluation of serum prostate-specific antigen velocity after radical prostatectomy to distinguish local recurrence from distant metastases. Urology. 1994;43(5):649–59.
17. Patel A, Dorey F, Franklin J. Recurrence patterns after radical retropubic prostatectomy: clinical usefulness of prostate specific antigen doubling times and log slope prostate specific antigen. J Urol. 1997;158(4):1441–5.
18. Han J, Kamber M, Pei J. Data mining: concepts and techniques: Elsevier; 2011.
19. Moskovitch R, Shahar Y. Medical temporal-knowledge discovery via temporal abstraction. In: AMIA Annual Symposium Proceedings, vol. 2009. San Francisco: American Medical Informatics Association; 2009. p. 452.
20. Batal I, Sacchi L, Bellazzi R, Hauskrecht M. A temporal abstraction framework for classifying clinical temporal data. In: AMIA Annual Symposium Proceedings, vol. 2009. San Francisco: American Medical Informatics Association; 2009. p. 29.
21. Hripcsak G, Albers DJ, Perotte A. Exploiting time in electronic health record correlations. J Am Med Inform Assoc. 2011;18(Supplement 1):109–15.
22. Hanauer DA, Ramakrishnan N. Modeling temporal relationships in large scale clinical associations. J Am Med Informatics Assoc. 2013;20(2):332–41.

23. Pei J, Han J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, Dayal U, Hsu MC. Mining sequential patterns by pattern-growth: The prefixspan approach. Knowl Data Eng IEEE Trans. 2004;16(11):1424–40.
24. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247(18):2543–6.
25. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing. R Foundation for Statistical Computing. 2014. https://www.r-project.org/.
26. T T. A Package for Survival Analysis in S. 2014. http://CRAN.R-project.org/package=survival. Accessed 15 Jan 2015.
27. Terry M. Therneau PMG. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.
28. Viger PF, Gomariz A, Gueniche T, Soltani A, Wu CW, Tseng VS. Spmf: A java open-source pattern mining library. J Mach Learn Res. 2014;15:3389–93.