Int. J. Radiation Oncology Biol. Phys., Vol. 75, No. 5, pp. 1357–1363, 2009
Copyright © 2009 Elsevier Inc.
Printed in the USA. All rights reserved
0360-3016/09/$–see front matter

ELSEVIER

doi:10.1016/j.ijrobp.2009.01.013

A STATISTICAL EVALUATION OF RULES FOR BIOCHEMICAL FAILURE AFTER RADIOTHERAPY IN MEN TREATED FOR PROSTATE CANCER

Carine A. Bellera, Ph.D.,[*†] James A. Hanley, Ph.D.,[†] Lawrence Joseph, Ph.D.,[†]
and Peter C. Albertsen, M.D.[‡]

[*] Department of Clinical Epidemiology and Clinical Research, Institut Bergonié, Regional Comprehensive Cancer Center, Bordeaux, France; [†] Department of Epidemiology and Biostatistics, McGill University, Montreal, PQ, Canada; and [‡] Division of Urology, University of Connecticut Health Center, Farmington, CT

Purpose: The "PSA nadir + 2 rule," defined as any rise of 2 ng/ml above the current prostate-specific antigen (PSA) nadir, has replaced the American Society for Therapeutic Radiology and Oncology (ASTRO) rule, defined as three consecutive PSA rises, to indicate biochemical failure (BF) after radiotherapy in patients treated for prostate cancer. We propose an original approach to evaluate BF rules based on the PSAdt as the gold standard rule and on a simulation process allowing us to evaluate the BF rules under multiple settings (different frequency, duration of follow-up, PSA doubling time [PSAdt]).

Methods and Materials: We relied on a retrospective, population-based cohort of individuals identified by the Connecticut Tumor Registry and treated for localized prostate cancer with radiotherapy. We estimated the 470 underlying true PSA trajectories, including the PSAdt, using a Bayesian hierarchical changepoint model. Next, we simulated realistic, sophisticated data sets that accurately reflect the systematic and random variations observed in PSA series. We estimated the sensitivity and specificity by comparing the simulated PSA series to the underlying true PSAdt.

Results: For follow-up of more than 3 years, the specificity of the PSA nadir + 2 rule was systematically greater than that of the ASTRO criterion. In few settings, the nadir + 2 rule had a lower sensitivity than the ASTRO. The PSA nadir + 2 rule appeared less dependent on the frequency and duration of follow-up than the ASTRO.

Conclusions: Our results provide some refinements to earlier findings as the BF rules were evaluated according to various parameters. In most settings, the PSA nadir + 2 rule outperforms the ASTRO criterion. © 2009 Elsevier Inc.

Prostate cancer, PSA failure, Radiotherapy, Sensitivity, Specificity.

INTRODUCTION

Prostate specific antigen (PSA) is used to monitor patients after treatment for prostate cancer. After prostatectomy, the PSA concentration decreases immediately to undetectable levels because of the removal of the prostate. Any subsequent production of PSA indicates biochemical failure, and by extension treatment failure. The situation is more complex for patients treated with external beam radiation therapy, since, as the prostate is not removed, the concentration of PSA not only depends on the tumour eradication, but also on the cellular effects of radiation on the prostate. Postradiotherapy PSA levels decrease more or less rapidly over the first 2 years, and then start to rise at rates which vary between individuals. As a result, establishing a precise biochemical failure (BF) rule based on PSA becomes complex; and up to a decade ago, different rules were used in the literature. In an effort to standardize the reporting of treatment outcomes and to facilitate their comparison, in 1996 the American Society for Therapeutic Radiology and Oncology (ASTRO) consensus panel proposed guidelines to unify the scientific community on the use of a single definition. The panel considered three consecutive PSA rises as an appropriate definition of biochemical failure after radiation therapy, with the date of failure as the time midway between the posttreatment PSA nadir and the first of the three consecutive increases (1).

After the publication of the guidelines, the performance of the ASTRO and other rules have been extensively studied, including the so-called ‘‘PSA nadir + 2’’ criterion, defined as

Reprint requests to: Carine A. Bellera, Ph.D., Department of Clinical Epidemiology and Clinical Research, Institut Bergonié, Regional Comprehensive Cancer Center, 229 Cours de l’Argonne, 33076 Bordeaux, France. Tel: +33 (0)5 56 33 04 95; Fax: +33 (0)5 56 33 04 85; E-mail: bellera@bergonie.org

Conflict of interest: none.

Acknowledgments—This work was supported by a doctoral training award from Le Fonds de la Recherche en Santé du Québec (to C.B.), a studentship from the McGill University Health Center (to C.B.), as well as grants from The Natural Sciences and Engineering Research Council of Canada (to J.H.) and Le Fonds Québécois de la Recherche sur la Nature et les Technologies (to J.H.). Support was also provided through a. is supported by a Senior Scientist Award from Le Fonds de la Recherche en Santé du Québec (to L.J.).

Received Sept 8, 2008, and in revised form Dec 9, 2008. Accepted for publication Jan 8, 2009.

any increase of 2 ng/ml above the current PSA nadir. The objective of these studies was to assess the classification performance of the BF rules in terms of sensitivity and specificity. The sensitivity was defined as the probability that the BF criteria were met among men experiencing clinical failure; and conversely, the specificity was defined as the probability of not satisfying the BF criteria among men without clinical failure. Different estimates of the ASTRO sensitivity and specificity have been reported, varying respectively from 55% to 92% and from 68% to 80% (2–6). This large variability can be explained by several factors, including the choice of the gold standard rule (local failure, distant failure, clinical failure, PSA above some threshold value, or a combination of these events), the population under study, or the intensity of PSA surveillance (frequency of measurements and length of follow-up). However, most studies have suggested that in general the PSA nadir + 2 rule has better classification properties than the ASTRO rule. Thus, in 2006 a second ASTRO consensus panel, in Phoenix, AZ, led to the adoption of the PSA nadir + 2 rule, now referred as the ‘‘Phoenix definition’’ and replacing the criterion based on three consecutive rises (7).

To date, studies have evaluated rules for PSA failure by estimating their sensitivity and specificity with respect to the incidence of particular clinical events (clinical, local, or distant failure) (2–6, 8, 9). It is however recognized that a rising PSA precedes clinical failure by several years and reflects the true ability to cure patients (10). Moreover, a fundamental point, before one considers how well even a perfectly measured PSA trajectory correlates with clinical outcomes, is how good a BF rule is at correctly identifying a PSA trajectory that is truly rising, and how often it can recognize series that are truly stable or rising only slowly. We thus propose an original numerical assessment of PSA rules using the underlying PSA trend, or similarly the PSA doubling time (PSAdt) as the gold standard rule. Indeed, a property for PSA-based rules should be the ability to identify rising from non rising PSA series. Our approach is based on an original statistical procedure that relies on the simulation of realistic postradiotherapy PSA series in which we know the ‘‘true’’ trajectories. Our flexible simulation process enables us to evaluate the performance of BF rules as a function of the PSA doubling time, the duration of follow-up, and the frequency of PSA measurements, all known to affect the sensitivity and specificity. To our knowledge, such comparison of BF rules under multiple settings has not yet been performed.

We first describe the dataset of postradiotherapy PSA series used to generate realistic PSA profiles. We next describe the statistical approach and the simulation process. Finally, we present estimates of the sensitivity and specificity of both the ASTRO and the PSA nadir + 2 rules under various settings.

METHODS AND MATERIALS

We based our evaluation procedure and the inputs to our simulations on a PSA dataset that was assembled retrospectively, on a population-based cohort identified from the Connecticut Tumor Registry. The men were 75 years or less of age and were residents of Connecticut when diagnosed with localized cancer between 1990 and 1992. More details are available in a report by Albertsen *et al.* (11). We based our study on men diagnosed with a localized cancer of the prostate and treated with radiotherapy without any hormonal pretreatment. Men with advanced disease or an initial PSA greater than 50 ng/ml were excluded. In addition, we required each PSA series to have at least one baseline PSA measurement as well as two subsequent PSA measurements. In some instances, the men could receive a subsequent hormonal treatment; however, PSA measurements taken while patients were receiving hormone therapy are not considered here.

Previous studies have been conducted to evaluate whether BF rules predict clinical failure. On the other hand, our aim was to perform a numerical evaluation of these rules, and specifically to assess whether BF rules adequately identify PSA trajectories that are truly rising and conversely series that are truly stable or rising only slowly. Because a property of PSA-based rules should be the ability to identify rising from nonrising PSA series, we thus relied on the PSAdt as the gold standard rule. The general idea of our approach was the following. First, based on a real PSA dataset, we simulated several realistic postradiotherapy PSA series. Because the schedule of PSA measurements can affect the performance of BF rules, these series were simulated assuming various length of follow-up (up to 10 years) and intervals between PSA measurements (every 3 or 6 months). Next, for each generated series, we estimated the underlying PSA doubling-time that we categorized (PSAdt <1 year, 1–2 years, 2–5 years, 5–10 years, >10 years). Finally, we evaluated whether the simulated series satisfied the BF rules according to the schedule of PSA measurements and PSAdt; these evaluations provided the estimates of sensitivity and specificity for each PSA-based rule.

We now describe our procedure in details. Our evaluation of PSA failure rules relied on statistical modeling and sophisticated simulations. Although we provide only a general presentation of our procedure, complete statistical details can be found in reports by Bellera *et al.* (12, 13). First, for every man, we estimated the post-radiotherapy PSA profiles, including the PSA doubling time after the PSA nadir, using a Bayesian hierarchical change point model (Appendix A). Second, we used these estimated individual PSA profiles to simulate thousands of PSA series; more specifically, for each real PSA series, we generated 150 PSA profiles. Because we based our simulations on estimates derived from real PSA profiles, the generated series accurately reflected the systematic random variations observed in real PSA trajectories. Series were simulated assuming various follow-up schedules, including PSA measurements taken every 3 or 6 months after treatment and follow-up durations of up to 10 years. We then sorted these simulated series according to their underlying estimated PSA doubling time (<1 year, 1–2 years, 2–5 years, 5–10 years, >10 years). We estimated the sensitivity of the BF rules based on the simulated PSA series with a PSAdt of less than 10 years. Specifically, we estimated the sensitivity of the ASTRO criterion as the proportion of simulated realistic series with three consecutive PSA increases. Conversely, we estimated the specificity in men with a close-to-flat post-nadir PSA curve, that is, in the subgroup of simulated series with an estimated true doubling time of more than 10 years. The specificity of the ASTRO criterion was estimated as the proportion of series with two or fewer consecutive PSA rises. The same procedure was used to assess the PSA nadir + 2 definition: that is, the sensitivity of the Phoenix rule was estimated within series with a short PSAdt as the proportion of series satisfying the BF rule. The specificity was estimated from close-to-flat post-nadir PSA curves as the proportion of series not satisfying the rule.

Table 1. Baseline characteristics of the 470 men treated with radiotherapy for prostate cancer

	All men (*N* = 470)	Men with subsequent hormotherapy (*n* = 139)	Men without subsequent hormonotherapy (*n* = 331)
Age at diagnosis (years)			
Average	70.1	69.4	70.4
Range	(49.2–76.0)	(53.2–75.9)	(49.2–76.0)
Pre-treatment PSA level in ng/ml*			
0–3.9	7% (33)	2% (3)	9% (30)
4–4.9	39% (183)	24% (34)	45% (149)
10–19.9	32% (149)	34% (47)	31% (102)
20–50	22% (105)	40% (55)	15% (50)
Pretreatment Gleason score*			
2–4	2% (10)	2% (3)	2% (7)
5	6% (26)	4% (5)	6% (21)
6	45% (210)	30% (42)	51% (168)
7	27% (126)	35% (48)	24% (78)
8–10	19% (91)	25% (35)	17% (56)
Missing	1% (7)	4% (6)	0% (1)

* Percentages (counts).

RESULTS

A total of 470 men satisfied our inclusion criteria and were included as the inputs to our numerical evaluation in our analysis. Of these, 139 men subsequently received hormonal therapy and 331 did not. The shortest and longest original PSA series had three and 36 measurements, respectively; there were nine PSA measurements on average, and the mean follow-up time was 5.7 years. Baseline characteristics of the 470 patients, such as age at diagnostic, initial T-stage, PSA level and Gleason score, are provided in Table 1.

From the 470 men, 377 had an estimated PSAdt of less than 10 years, 52 had an estimated PSAdt of more than 10 years, and for 41 men, the PSAdt was estimated to be infinite.

We generated 150 randomly distributed trajectories per man, that is, a total of 70,500 different PSA profiles (470 × 150). We estimated the sensitivity from the 56,550 (377 × 150) simulated PSA series with a PSAdt of less than ten years. Similarly, we estimated the specificity relying on the 7,800 (52 × 150) PSA trajectories with a long PSAdt, that is, a close-to-flat PSA profile. The estimated parameters are presented in Tables 2 and 3 for various follow-up durations, frequencies of measurements, and PSAdt. For illustration, when PSA levels are measured every 3 months over a 3-year period, 80.8% of the series generated from the men with an estimated true PSAdt of less than 1 year had three consecutive PSA rises. Thus, under this schedule of measurements, and for PSAdt of less than 1 year, the rule of three rises had an 80.8% sensitivity.

Our simulation process allowed us to estimate the sensitivity and specificity as a function of the PSAdt. For example, depending on the PSAdt, and when PSA levels are measured every 3 months over a 3-year period, the ASTRO sensitivity ranges from 20.6% (PSAdt between 5 and 10 years) to 80.8% (PSAdt <1 year), whereas the sensitivity of the PSA nadir +2 rule varies between 21.1% and 87.1%. In the same settings, the ASTRO and PSA nadir + 2 rules provided specificity estimates of 78.2% and 82%, respectively.

The PSA concentrations are known to decrease over the first 2 years after radiotherapy, achieving their lowest level around the second year (14, 15). This explains the very low sensitivities observed for the BF rules during the first 2 years, and similarly their high specificities. The ASTRO criterion requires three consecutive PSA rises and thus a minimum of four observations. Therefore, because we tested 3- and 6-month intervals between PSA measurements, we could not evaluate the ASTRO criterion when the follow-up duration was 1 year, and the frequency of measurements was 6 months only.

Overall, and ignoring the first 2 years, the sensitivity of the rules improved as the follow-up duration increased, and intervals between measurements were shortened. These results are intuitively reasonable. Indeed, given a fixed true doubling time, and a fixed follow-up duration, a 3-month interval between measurements provides twice as many PSA readings as a 6-month interval between measurements. Therefore, the probability of observing three consecutive PSA rises is higher, leading to a larger sensitivity. Similarly, given a fixed doubling time and a fixed interval between measurements, an increased length of follow-up provides more PSA observations, and thus increases the chance of observing three consecutive PSA rises. Conversely, the specificity decreases with longer follow-up time and increases with longer intervals between measurements. In general, this observation holds for diagnostic tests that are used as surveillance tools and that are thus repeated over time, such as mammograms in the context of breast cancer surveillance after treatment.

In Table 3, estimates are shaded if the PSA nadir + 2 rule has a better sensitivity or specificity. After 3 years of follow-up, we observe that the specificity of the PSA nadir +2 rule is systematically higher than that of the ASTRO criterion. With respect to the sensitivity, the PSA nadir +2 provides better results than the ASTRO criterion, provided that the PSAdt is less than 5 years. For PSAdt of more than 5 years, the sensitivity the ASTRO criterion is greater than that of ASTRO. It should be noted, however, that when the interval between

Table 2. Sensitivity and specificity of the ASTRO rule

Sensitivity when PSA levels are measured every 3 months										
	Duration of follow-up (years)									
PSA doubling time	1	2	3	4	5	6	7	8	9	10
0–1 year	7.9%	50.8%	80.8%	93.2%	97.5%	98.8%	99.1%	99.4%	99.5%	99.6%
1–2 years	1.1%	19.0%	45.2%	66.1%	80.3%	89.6%	94.4%	97.0%	98.3%	98.9%
2–5 years	0.8%	11.2%	27.1%	42.9%	56.0%	66.7%	75.3%	81.7%	86.6%	89.9%
5–10 years	0.7%	8.0%	20.6%	32.8%	43.7%	52.6%	61.2%	67.6%	73.9%	78.6%
Sensitivity when PSA levels are measured every 6 months										
	Duration of follow-up (years)									
PSA doubling time	1	2	3	4	5	6	7	8	9	10
0–1 year	N/A	24.0%	75.2%	90.5%	96.7%	98.2%	98.8%	99.1%	99.2%	99.2%
1–2 years	N/A	6.7%	34.0%	58.9%	76.0%	86.9%	92.9%	95.8%	97.4%	98.2%
2–5 years	N/A	2.2%	14.2%	28.1%	41.4%	53.6%	63.4%	71.0%	77.1%	81.5%
5–10 years	N/A	1.1%	8.0%	15.6%	23.9%	32.1%	39.3%	46.0%	52.6%	58.4%
Specificity										
Interval between	Duration of follow-up (years)									
PSA measurements	1	2	3	4	5	6	7	8	9	10
Every 3 months	99.1%	90.6%	78.2%	67.1%	57.2%	48.3%	41.2%	34.7%	28.9%	24.2%
Every 6 months	N/A	98.0%	91.4%	84.1%	76.3%	69.3%	63.2%	57.6%	52.1%	47.1%

Abbreviations: ASTRO = American Society for Therapeutic Radiology and Oncology; N/A = not applicable; PSA = prostate-specific antigen.

PSA measurements is 6 months, the sensitivity of the ASTRO rule is only slightly greater than that of ASTRO (<4% difference). This difference in sensitivity between the two BF rules is greater when the interval between PSA levels is reduced to 3 months (up to 15% difference).

Overall, the PSA nadir + 2 rule is less dependent than the ASTRO criterion on the frequency and duration of follow-up. For a fixed PSAdt and a fixed interval between PSA measurements, we observe that for two consecutive values of follow-up duration, the difference in the sensitivity of the nadir + 2 rule is less than the difference in the sensitivity of the ASTRO criterion. Similarly, fixing the duration of follow-up, the difference in the sensitivity of the nadir + 2 rule between a 3-month and a six-month frequency is less than the difference in the sensitivity of ASTRO between these two frequencies of measurements. The same observation holds for the specificity of the nadir + 2 rule, which appears to be less affected by the schedule of measurements than is the ASTRO criterion.

DISCUSSION

Compared with the ASTRO criterion, the PSA nadir + 2 rule has been shown to have several advantages. First, because of the backdating process, the ASTRO definition overestimates the PSA failure early, and underestimate it afterwards (16). While this bias may be theoretically reduced with longer follow-up, the rule then becomes less practical to apply as one would not want to wait for three consecutive rises. On the other hand, the PSA nadir + 2 rule, by definition, does not backdate the failure time as it is considered to be the timing of the nadir. Moreover, given the important within-individual PSA variability (17, 18), patients do not show a steadily rising or stable PSA pattern, but rather experience a succession of downs, ups and plateaus, which become overly complex to interpret. The use of consecutive rises is thus in essence more sensitive to random fluctuations or bounces, contrary to a rule based on trend. For illustration, Horwitz *et al.* showed that patients who experienced a PSA bounce (there defined as a minimal rise of 0.4 ng/mL over a 6-month follow-up followed by a drop of PSA of any magnitude) had an increasing risk of depicting three consecutive PSA rises, but with longer follow-up, this did not translate into a difference in clinical failure (5).

The most comprehensive studies of PSA failure rules following radiation were provided by Thames et al. in 2003 (4), and Horwitz et al. in 2005 (6). These two publications were based on the same large institutional database and reported the sensitivity and specificity for more than 100 definitions of BF using either clinical or distant failures as gold-standard rules. Several definitions were identified as being both more sensitive and specific than ASTRO. With respect to distant failure only, approximately 20 definitions, including the

Table 3. Sensitivity and specificity of the rule based on the "nadir + 2 ng/ml" rule

Sensitivity when PSA levels are measured every 3 months

	Duration of follow-up (years)									
PSA doubling time	1	2	3	4	5	6	7	8	9	10
0–1 year	24.7%	66.1%	87.1%	94.0%	98.8%	99.6%	99.6%	99.7%	99.7%	99.6%
1–2 years	11.3%	37.2%	61.8%	78.3%	88.1%	92.8%	96.5%	98.5%	99.2%	99.5%
2–5 years	4.0%	15.4%	29.6%	43.6%	56.6%	67.4%	75.3%	81.4%	85.3%	87.8%
5–10 years	5.0%	13.3%	21.1%	28.2%	35.4%	41.5%	47.2%	53.5%	58.5%	62.7%

Sensitivity when PSA levels are measured every 6 months

	Duration of follow-up (years)									
PSA doubling time	1	2	3	4	5	6	7	8	9	10
0–1 year	18.9%	61.1%	85.5%	93.3%	98.6%	99.5%	99.6%	99.6%	99.7%	99.6%
1–2 years	5.7%	27.2%	53.2%	73.6%	85.9%	91.6%	95.8%	98.0%	98.9%	99.3%
2–5 years	1.6%	8.5%	19.2%	32.6%	46.4%	59.3%	69.0%	76.3%	81.6%	84.8%
5–10 years	2.2%	6.9%	12.3%	18.2%	25.0%	32.0%	38.3%	44.8%	50.0%	54.7%

Specificity

Interval between	Duration of follow-up (years)									
PSA measurements	1	2	3	4	5	6	7	8	9	10
Every 3 months	93.5%	87.2%	82.0%	77.1%	71.9%	68.1%	63.6%	59.4%	55.2%	51.3%
Every 6 months	96.4%	91.9%	88.3%	84.9%	80.5%	76.8%	72.8%	68.4%	64.1%	60.3%

Abbreviation: PSA = prostate-specific antigen.
Shading of estimates indicates that the "nadir + 2" rule has a better sensitivity (or specificity) for the schedule of follow-up under study.

PSA nadir + 2 rule, performed better than ASTRO (6). Three rules had better sensitivity and specificity than ASTRO when relying on clinical failure as the gold standard rule. While in this case the PSA nadir + 2 rule was not one of the rules shown to perform better (*i.e.,* both its sensitivity and specificity were not simultaneously greater), it did remain a potential candidate when other rules were considered, as only its specificity was slightly lower than ASTRO (71.4% vs. 72.1% for ASTRO).

Our study confirmed that in general, the PSA nadir + 2 rule outperforms the ASTRO criterion and is thus in accordance with previous findings (7). Moreover, our simulation procedure allowed us to refine these conclusions and to assess how these two BF rules behave under various settings. The flexible simulation process allowed us to control for characteristics (*PSAdt*, schedule of measurements) known to affect the sensitivity and specificity, and we thus evaluated the BF rules accordingly. Moreover, contrary to previous similar evaluations, we focused on the ability of the BF rules to adequately distinguish PSA series with underlying rising PSA trend from non rising PSA series.

Up to now, when BF rules were investigated, one sensitivity estimate and one specificity estimate were reported per study. Given these estimates were based on data usually not collected in the context of controlled trials, it is usually complex to compare them between studies. The study populations might be different with respect to baseline characteristics (variations in the distribution of baseline PSA or Gleason scores), PSA surveillance procedures can vary across studies and centers. Our objective was thus to evaluate BF rules while controlling for parameters that can affect sensitivity and specificity estimates. As our results suggest, sensitivity and specificity estimates of both BF rules are greatly affected by the underlying PSAdt. For example, from Table 3, assuming PSA are measured every 6 months over 5 years, a PSAdt of less than 1 year leads to a sensitivity estimate of 98.6%, whereas a PSAdt between 5 and 10 years corresponds to a 25% sensitivity estimate. Collapsing these results into one single number would lead to a loss of valuable information.

Whatever the frequency of measurements, as long as the duration of follow-up was more than 3 years, the specificity of the PSA nadir + 2 rule was greater than that of ASTRO. For PSAdt of less than 5 years, the sensitivity of the PSA nadir + 2 criterion was greater than that of ASTRO. In some few occasions, specifically, when the PSAdt was more than 5 years, the ASTRO rule had a better sensitivity than the PSA nadir + 2 rule. However, it should be noted that when the interval between PSA measurements was 6 months, which in practice is more realistic than a 3-month interval, there was only a slight improvement in the sensitivity provided by the ASTRO rule. However, because the specificity of the PSA nadir + 2 was much greater under these circumstances, this BF rule appeared still to be valid compared with the ASTRO rule. Finally, our results also suggested that the sensitivity and specificity of the PSA nadir + 2 rule were less affected by the schedule of PSA measurements than those of the ASTRO rule.

This analysis could easily be extended to the evaluation of other BF rules and to other schedules of measurements (*e.g.*, frequent PSA assessments early on, followed by a less intensive PSA surveillance).

REFERENCES

1. American Society for Therapeutic Radiology and Oncology Consensus Panel. Consensus statement: Guidelines for PSA following radiation therapy. *Int J Radiat Oncol Biol Phys* 1997;37: 1035–1041.
2. Buyyounouski MK, Hanlon AL, Eisenberg DF, *et al.* Defining biochemical failure after radiotherapy with and without androgen deprivation for prostate cancer. *Int J Radiat Oncol Biol Phys* 2005;63:1455–1462.
3. Kuban DA, Thames HD, Shipley WU. Defining recurrence after radiation for prostate cancer. *J Urol* 2005;173:1871–1878.
4. Thames H, Kuban D, Levy L, *et al*. Comparison of alternative biochemical failure definitions based on clinical outcome in 4839 prostate cancer patients treated by external beam radiotherapy between 1986 and 1995. *Int J Radiat Oncol Biol Phys* 2003;57:929–943.
5. Horwitz EM, Levy LB, Thames HD, *et al*. Biochemical and clinical significance of the posttreatment prostate-specific antigen bounce for prostate cancer patients treated with external beam radiation therapy alone: A multiinstitutional pooled analysis. *Cancer* 2006;107:1496–1502.
6. Horwitz EM, Thames HD, Kuban DA, *et al*. Definitions of biochemical failure that best predict clinical failure in patients with prostate cancer treated with external beam radiation alone: A multi-institutional pooled analysis. *J Urol* 2005;173: 797–802.
7. Roach M III, Hanks G, Thames H Jr., *et al*. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: Recommendations of the RTOG-ASTRO Phoenix Consensus Conference. *Int J Radiat Oncol Biol Phys* 2006;65:965–974.
8. Horwitz EM, Uzzo RG, Hanlon AL, *et al*. Modifying the American Society for Therapeutic Radiology and Oncology definition of biochemical failure to minimize the influence of backdating in patients with prostate cancer treated with 3-dimensional conformal radiation therapy alone. *J Urol* 2003; 169:2153–2157.
9. Critz FA. A standard definition of disease freedom is needed for prostate cancer: Undetectable prostate specific antigen compared with the American Society of Therapeutic Radiology and Oncology consensus definition. *J Urol* 2002;167: 1310–1313.
10. Shipley WU, Thames HD, Sandler HM, *et al*. Radiation therapy for clinically localized prostate cancer: A multi-institutional pooled analysis. *J Am Med Assoc* 1999;281:1598–1604.
11. Albertsen PC, Hanley JA, Penson DF, *et al*. 13-Year outcomes following treatment for clinically localized prostate cancer in a population based cohort. *J Urol* 2007;177:932–936.
12. Bellera CA, Hanley JA, Joseph L, Albertsen PC. Detecting Trends in Noisy Data Series: Application to Biomarker Series. *Am J Epidemiol* 2008;167:1130–1139.
13. Bellera CA, Hanley JA, Joseph L, Albertsen PC. Hierarchical changepoint models for biochemical markers illustrated by tracking postradiotherapy prostate-specific antigen series in men with prostate cancer. *Ann Epidemiol* 2008;18:270–282.
14. Hanlon AL, Diratzouian H, Hanks GE. Posttreatment prostate-specific antigen nadir highly predictive of distant failure and death from prostate cancer. *Int J Radiat Oncol Biol Phys* 2002;53:297–303.
15. Kestin LL, Vicini FA, Ziaja EL, *et al*. Defining biochemical cure for prostate carcinoma patients treated with external beam radiation therapy. *Cancer* 1999;86:1557–1566.
16. Coen JJ, Chung CS, Shipley WU, Zietman AL. Influence of follow–up bias on PSA failure after external beam radiotherapy for localized prostate cancer: Results from a 10–year cohort analysis. *Int J Radiat Oncol Biol Phys* 2003;57:621–628.
17. Eastham JA, Riedel E, Scardino PT, *et al*. Variation of serum prostate-specific antigen levels: An evaluation of year-to-year fluctuations. *J Am Med Assoc* 2003;289:2695–2700.
18. Carter HB, Pearson JD, Metter EJ, *et al*. Longitudinal evaluation of prostate-specific antigen levels in men with and without prostate disease. *J Am Med Assoc* 1992;267: 2215–2220.

APPENDIX A: HIERARCHICAL CHANGEPOINT MODEL USED FOR THE ESTIMATION OF THE PROSTATE-SPECIFIC ANTIGEN TRAJECTORIES

After radiotherapy, prostate-specific antigen (PSA) levels decrease and then start to rise again at various rates across individuals, thus applying a $\log_2$ transformation allows one to obtain a piecewise linear pattern. Figure 1 illustrates a prototypic PSA trajectory defined by its four parameters: α_i, τ_i, β_{1i}, and β_{2i} corresponding respectively to the $\log_2$PSA nadir, its timing, also called changepoint, the $\log_2$PSA decline rate prior to the PSA nadir, and the post-nadir $\log_2$PSA growth rate for the ith man.

For every man, we estimated the post-radiotherapy PSA profile using a Bayesian hierarchical changepoint model with three hierarchical levels to account for the presence of a random changepoint, as well as the wide between-subjects variations in PSA trajectories (for more details, see (13)).

At the first level, each individual $\log_2$PSA profile was modelled as follows. Let $\log_2\text{PSA}_{ij}$ be the PSA concentration on the $\log_2$ scale for the jth measurement for the ith man. We assumed that the $\log_2\text{PSA}_{ij}$ were normally distributed, with expected value μ_{ij}, and variance σ^2_{ij} : $\log_2\text{PSA}_{ij} \sim \text{N}(\mu_{ij}, \sigma^2_{ij})$.

The expected $\log_2$PSA value, μ_{ij}, was related to the timing of the measurement, t_{ij}, through linear regression functions before and after the unknown changepoint τ_i:

$$\begin{aligned}\mu_{ij} &= \alpha_i + \beta_{1i}(t_{ij} - \tau_i),\ t_{ij} < \tau_i\\ &= \alpha_i + \beta_{2i}(t_{ij} - \tau_i),\ t_{ij} \geq \tau_i.\end{aligned}$$

We expressed the PSA variability σ^2_{ij} as a function of the PSA concentration because interassay coefficients of variation tend to be larger at lower PSA levels. We thus modeled the logarithm of the precision as a linear function of the $\log_2$-PSA level: $\log(1/\sigma^2_{ij}) = \theta_1 + \theta_2 \log_2\text{PSA}_{ij}$.

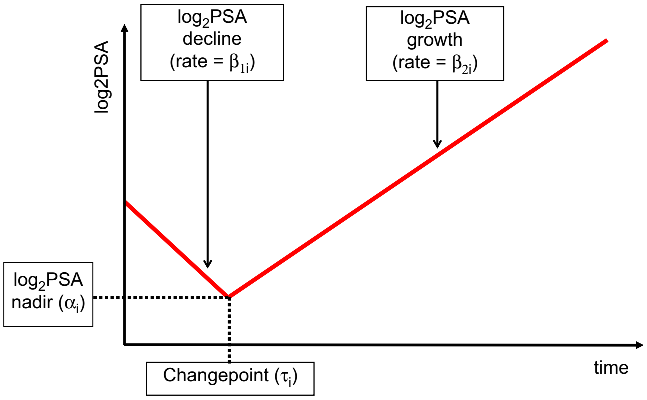

Fig. 1. Individual prototypic prostate-specific antigen (PSA) profile.

At the second level, we assumed that the individual parameters, α_i, β_{1i}, β_{2i}, and τ_i, were a priori uncorrelated both within and between subjects, although they are related through the likelihood function. The complete model assumes the following distributions:

$$\alpha_i \sim \mathrm{N}\left(\mu_\alpha, \sigma_\alpha^2\right),\ \mu_\alpha \sim \mathrm{N}(0,100),\ \sigma_\alpha^2 \sim \mathrm{U}(0,4),$$
$$\beta_{1i} \sim \mathrm{N}\left(\mu_{\beta 1}, \sigma_{\beta 1}^2\right),\ \mu_{\beta 1} \sim \mathrm{N}(0,100),\ \sigma_{\beta 1}^2 \sim \mathrm{U}(0,4),$$
$$\beta_{2i} \sim \mathrm{N}\left(\mu_{\beta 2}, \sigma_{\beta 2}^2\right),\ \mu_{\beta 2} \sim \mathrm{N}(0,100),\ \sigma_{\beta 2}^2 \sim \mathrm{U}(0,4),$$
$$\theta_1 \sim \mathrm{N}(0,100), \theta_2 \sim \mathrm{N}(0,100).$$

The prior distribution of the changepoint was a continuous uniform distribution; the range was selected according to previous biologic knowledge and depending on the subgroup of men. Secondary treatment usually is initiated when it is suspected that radiotherapy has failed, indicated by a rising PSA pattern starting within the first 2 to 3 years after radiotherapy; we thus selected a range of 5 years for this subgroup, $\tau_j \sim$ U(0, 5). Most men who do not receive a secondary treatment generally are those for whom radiotherapy is successful. In such cases, PSA are still produced by the remaining healthy prostate cells, although in very small quantities. Thus, the PSA concentrations for these men will start to rise at a later time, and at a very slow rate. For this reason, we selected a uniform distribution with a ten-year range for this subgroup: $\tau_j \sim$ (0, 10).