Published in final edited form as:

Stat Med. 2014 March 15; 33(6): 930–939. doi:10.1002/sim.6003.

Modeling Grade Progression In An Active Surveillance Study

Lurdes Y.T. Inoue, PhD, Bruce J. Trock, PhD, Alan W. Partin, MD, H. Ballentine Carter, MD, and Ruth Etzioni, PhD

Fred Hutchinson Cancer Research Center, Seattle, Washington (LI, RE), Johns Hopkins Medical Institutions, Baltimore, MD (BJT, AWP, HBC), University of Washington, Seattle, WA (LI, RE).

Abstract

Prostate cancer grade, assessed with the Gleason score, describes how abnormal the tumor tissue and cells appear and it is an important prognostic indicator of disease progression. Whether prostate tumors change grade is a question that has implications for screening and treatment. Empirical data on tumor grade over time have become available from men biopsied regularly as part of active surveillance (AS). However, biopsy grade is subject to misclassification. In this article we develop a model that allows for estimation of the time of grade change while accounting for the misclassification error from biopsy grade. We use misclassification rates from studies of prostate cancer biopsies followed by radical prostatectomy. Estimation of the transition times from true low-grade to high-grade disease is conducted within a Bayesian framework. We apply our model to serial observations on biopsy grade among 627 cases enrolled in a cohort of AS patients at Johns Hopkins University who were biopsied annually and referred to treatment if there was any evidence of disease progression on biopsy. We consider different prior distributions for the time to true grade progression. The estimated likelihood of grade progression within 10 years of study entry ranges from 12% to 24% depending on the prior. We conclude that knowledge of rates of grade misclassification allows for determination of true grade progression rates among men with serial biopsies on AS. While our results are sensitive to prior specifications they indicate that in a non-trivial fraction of the patient population, tumor grade can progress.

1. Introduction

Whether prostate cancers change grade, that is, dedifferentiate over time, represents a tantalizing and elusive question in prostate cancer research, and one with important implications for early detection and treatment. Until recently, empirical evidence of true grade change over time in prostate cancer tumors was lacking. The observation that older men were more likely to have high-grade prostate cancers was regarded as suggestive, but not conclusive evidence in favor of the dedifferentiation hypothesis.

Several modeling studies have been conducted to explore whether prostate cancers dedifferentiate as they grow [1–4]. One study showed that a natural history model allowing dedifferentiation fit to data from the Rotterdam section of the European Randomized Study of Prostate Cancer Screening (ERSPC) performed considerably better than one that did not [1]. A second study also compared models in which grade was fixed as high or low at disease onset and utilized models that allowed low-grade disease to transition to high grade during the pre-clinical phase of the disease [2]. This study also found that the models incorporating grade progression fit the observed data better than those that did not.

However, these studies were not able to validate their findings against empirical data on grade change, since at the time that the studies were conducted, these data did not exist.

The advent of active surveillance (AS) has provided a unique opportunity to learn about the dedifferentiation issue [5–9]. In AS studies, newly diagnosed, low-risk prostate cancer patients are intensively monitored, typically under a protocol that calls for serial biopsies at regular intervals. Patients are referred to treatment if there are any signs of disease progression, as evidenced by a change in disease grade or volume on repeat biopsy. In prostate cancer disease grade is measured by the Gleason Score (GS) on which a pathologist assesses the degree of loss of normal glandular tissue appearance with respect to shape, size and differentiation of the glands. A GS of 6 or below corresponds to low-grade or well-differentiated disease; a GS of 7 corresponds to moderate grade or moderately differentiated disease and a GS of 8–10 corresponds to poorly differentiated disease.

Several AS cohorts are currently under study for disease progression and outcomes given delayed treatment. The cohorts differ somewhat in their inclusion criteria, surveillance schedules and modalities, and criteria for referral to treatment [5–9]. As an example, through June 2011, the Johns Hopkins AS cohort enrolled 627 men with very-low risk (80%) to low-risk (20%) disease, with very low risk defined as GS 6 or below, stage T1c, PSA density (PSA concentration per unit of gland volume) less than or equal to 0.15 ng/mL, 2 or fewer biopsy cores with cancer and less than or equal to 50% involvement of any core with cancer, and low risk defined as early localized stage T1c-T2a, GS 6 or below and PSA below 10ng/ml [8]. The surveillance protocol calls for annual biopsies with referral to treatment if there is any disease reclassification, i.e., exceeding the above thresholds for volume/core involvement or higher GS at repeat biopsy. As an extreme comparison, the UK-based ProtecT trial does not call for regular biopsies; rather, men are referred to treatment based on their PSA trajectories and clinical symptoms [10,11].

The accumulation of data on serial biopsies in the absence of definitive therapy provides the best opportunity to date to address the dedifferentiation question. In the Johns Hopkins cohort, 106 of 235 men referred to treatment showed GS upgrading at their final pre-treatment biopsy [8]. At face value, this observation would seem to put the question of how frequently tumors dedifferentiate to rest. However, it is important to recognize that biopsy GS does not always accurately reflect the “true” grade within the prostate. In particular, it is not uncommon for a tumor to be classified as low grade based on a biopsy specimen, but high grade based on analysis of tissue from a subsequent prostatectomy. This is because a considerably larger sample of tissue is provided from a surgery specimen and more thorough interrogation of disease grade is therefore possible. A seminal study conducted by investigators on the Prostate Cancer Prevention Trial (PCPT) found that, among tumors that were classified as moderate to high grade following surgery, more than half (52.6%) 53% had been classified as low grade on biopsy [12]. This study showed that it is not uncommon for high-grade tumors to be incorrectly labeled as low grade on biopsy. A recent study [13] confirmed this finding under more contemporary biopsy schemes. In this study, among tumors classified as moderate to high grade at surgery, 45% were classified as low grade on biopsy. Conversely, among low-grade tumors at surgery, 10% were labeled high grade on biopsy. The study investigators acknowledged [13] that although under-grading due to the biopsy missing a higher grade component is the more common problem, overgrading is also a real issue commonly caused when glands are tangentially sectioned and the resulting image mimics a poorly-differentiated focus of tumor.

Given the fact that serial biopsy data is subject to misclassification, the question arises as to whether the likelihood of true grade progression can be extracted from the noisy serial biopsy data. This article addresses precisely this question. We consider serial biopsy data

from an AS program and model the longitudinal grade observations as depending on both the true grade (i.e., the grade that would be observed in a prostatectomy specimen) and the biopsy misclassification probability. We use a separate dataset containing both biopsy and radical prostatectomy grades on a large sample of patients to provide information on the frequency of biopsy grade misclassification. We consider two approaches assuming this error to be (a) known or (b) unknown but modeled with an informative prior. We extract the misclassification error from the observed grade trajectories to reveal an estimate of the frequency of true dedifferentiation or upgrading over time. Our model aims to estimate the time to a change in the true grade, where true grade is defined as the grade that would be determined based on a prostatectomy specimen. We describe the model and our Bayesian estimation approach (Section 2), show that it produces intuitively reasonable results (Section 3), and apply it to data on serial prostate biopsies from the Johns Hopkins AS cohort (Section 4).

2. Statistical Model

We consider each individual as having a time to true progression (a change point) from true low- to true moderate-to-high-grade disease and allow this time to fall either before, during, or after a study window. Our goal is to estimate the distribution of the time to the change point. For simplicity we refer to GS 6 or below as low grade and GS 7 or above as high grade. Consider the observed data on biopsy grade where grade is dichotomized as low or high. Each patient enters the study with low grade disease and his data consists of his serial biopsy grade indicators. Each biopsy grade indicator may be misclassified; our approach uses external estimates of the misclassification errors to extract the distribution of the change point. Furthermore, we assume that the dropout mechanism is non-informative of grade progression.

Let Y_{ij} denote, the observed disease grade in subject i at follow-up j , where $i = 1, \dots, N$ and $j = 1, \dots, n_i$. Let θ denote a vector with all model parameters. We formulate our models of the observed responses conditional on the true underlying latent time to grade change. Let t_{iG} and G_{ij} denote, respectively, the true time to grade change and the underlying true grade. We denote the conditional distribution of the observed grade given the underlying grade as $[Y_{ij} | \theta, G_{ij}]$. The joint likelihood of the model parameters θ given the observed data and the latent data is given by:

$$L(\theta) = \prod_{i=1}^N \prod_{j=1}^{n_i} [Y_{ij} | \theta, G_{ij}] \times [G_{ij} | \theta, t_{iG}] \times [t_{iG} | \theta]$$

The specific model components are described next:

1. Model for observed grade indicator on biopsy, $[Y_{ij} | \theta, G_{ij}]$: An individual's observed grade indicator at time t_{ij} depends on his true grade (dichotomized similarly as low or high) at this time and a misclassification error. Specifically, we assume that at time the t_{ij} observed grade is $Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$ where $\pi_{ij} = FP \times (1 - G_{ij}) + (1 - FN) \times G_{ij}$ and FP and FN are biopsy false positive and false negative rates. We consider two formulations:
 - a. The misclassification error is assumed known;
 - b. The misclassification error is unknown and assumed to have a prior distribution.

Specification 1a) is a special case of 1b) with point mass priors. We use specification 1a) in our simulation studies only to investigate the role of misclassification rates in the inference about grade change under the best-case situation where there is no uncertainty about the misclassification rates.

2. Models for the change point and the true grade, that is, $[t_{iG}|\theta]$ and $[G_{ij}|\theta, t_{iG}]$: The change point is modeled with a discrete distribution with the change potentially occurring before the study, at any year during the study, or after the study. By definition, the true grade indicator is low before the change point and high after the change point. This model allows subjects to upgrade after the study, but the change may not be experienced within any given study window. Specifically, we consider a study window of length n . This can be, for example, the maximum follow-up for any patient in the study. A subject can experience grade change within the study window when t_{iG} takes values within the set $\{1, \dots, n\}$. However, it is possible that a patient experiences a grade change before or after the study window. To accommodate the possible left-and right-censoring of the time to grade change, we expand the time horizon so that an individual can experience a grade change before entering the study or after leaving the study so that $t_{iG} \in \{0^-, 0, 1, \dots, n, n^+\}$. Using the above notation, a grade change at time 0^- means that it occurred before the study, while a grade change at time n^+ means that the change occurred after the study. We model $t_{iG} \sim \text{Multinomial}(1, p)$ where p is a vector with components such that $p_k = 0, k = 1, \dots, (n+3)$ and $\sum_{k=1}^{n+3} p_k = 1$. We note that the time to grade change, t_{iG} , is subject-specific as we expect it to vary across subjects depending on each subject's underlying risk factors for disease progression. However, the variation of the times to grade change in the cohort is modeled with a multinomial distribution with a common probability vector p which describes the probabilities of grade change at the population level. Given time to grade change, the model for true disease grade is deterministic. Specifically,

$$G_{ij} = \begin{cases} 0, & \text{if } t_{iG} = n^+ \\ 1, & \text{if } t_{iG} = 0^- \\ I_{(t_{ij} > t_{iG})}, & \text{if } t_{iG} \in \{0, \dots, n\} \end{cases}.$$

In words, if an individual has a change at time 0^- , he has true underlying high grade throughout the observation time. Analogously, if an individual has a change at time n^+ , then he has true low disease grade throughout the study. Otherwise, the change occurs within the study period.

We estimate model parameters via a Bayesian approach. We assume a priori that $p \sim \text{Dirichlet}(r)$, where r is a vector of length $(n+3)$. Moreover, under model specification 1b) in which error rates are assumed unknown, we assume a priori that $FP \sim \text{Beta}(a_{FP}, b_{FP})$ and $FN \sim \text{Beta}(a_{FN}, b_{FN})$. The models are fit using WinBugs. Code to fit the models is available as Web-based supporting materials for this article.

3. Simulation Study

3.1. Methods

To evaluate the validity of our method we conducted a simulation study and measured the bias of our projections against results that would typically be computed based only on observed biopsy grade information. Serial biopsy data were generated over a period of ten years for a hypothetical cohort of 500 subjects, under a theoretical (uniform) distribution of the time to true grade change with 70% of the cohort upgrading within the 10-year interval

and the remaining 25% and 5%, respectively, upgrading before and after the study window, and assuming FN and FP rates of 40% and 20%, based on the application data (Section 4). We investigated our model under different priors for the changes before, within and after the study window as well as under different priors for the misclassification rates, including point masses. The bias was estimated as the average difference between the individual-specific posterior mean time to upgrading and the theoretical (true) time to upgrading across 100 simulations among those cases whose true times to upgrading were within the study window.

3.2. Results

The simulation study confirmed that our method was considerably more accurate than one based only on the observed biopsy data. The bias was -0.95 when using the observed biopsy data, i.e., the estimated change point was on average at least 0.95 later than the true change point. However, when using our method, the bias was -0.04 on average (Table 1, scenario 1). Over- or under-estimating the biopsy misclassification errors generated changes in the estimated distribution of time to true upgrading that were intuitively reasonable. Underestimating the FN rate, representing greater confidence in observed low-grade biopsy results, increased the estimated likelihood of upgrading after the end of the study; conversely, overestimating the FN rate, representing less confidence in the observed low-grade results, increased the estimated likelihood of upgrading before or within the study window. Results when underestimating or overestimating the FP rate were in the opposite direction than those obtained when similarly misestimating the FN rate. Scenarios 6–7 show results when changing the distribution of upgrades before, within or after the study window via data simulation and/or prior specification. The results show sensitivity to the prior on time to upgrade. This is reasonable given the moderate cohort size used in the simulation. Scenarios 8–10 show results from the model where we specify priors for the misclassification rates and time to upgrade. Results from scenarios 8–9 are comparable to those under scenario 1. Relative to the results under scenario 6, the results under scenario 10 indicate that the added uncertainty in the misclassification rates lead to increased bias in the estimation of time to upgrade.

4. Application

4.1. Data Sources

Johns Hopkins Active Surveillance (JH-AS) cohort [8]—We use serial biopsy results from the JH-AS cohort which enrolled 627 men between January 1995 and June 2011. All men had very-low-risk or low-risk disease and were followed with semi-annual PSA tests and digital rectal exams and annual biopsies that sampled at least 12 tissue cores. Only cases with at least one biopsy after diagnosis were included. Referral to treatment occurred on the basis of biopsy reclassification of disease (higher volume, more cores involved with cancer or higher biopsy Gleason score). Some patients without biopsy changes underwent treatment for personal reasons [8] but no further specifics regarding the reason for off-protocol treatment are provided by study investigators. We assume therefore that loss to follow-up is non-informative. Data used for the analysis consisted of patients' indicators of high or low-grade disease at each biopsy and time of follow-up.

Johns Hopkins Immediate Treatment (JH-IT) Cohort—This cohort of 8,544 newly diagnosed prostate cancer cases underwent radical prostatectomy at Johns Hopkins between 2004 and 2011 inclusive and, in addition to their pathologic grade based on the surgical specimen, also had a grade determination based on their diagnostic biopsy. We use these data to determine the false negative (FN) rate, defined as the fraction of men with moderate- to high-grade on prostatectomy who were classified as low grade on biopsy and the false

positive (FP) rate, defined as the fraction of men with low-grade disease on prostatectomy who were classified as high grade on biopsy. For validation purposes we compare our results against a subcohort of the JH-AS data with 153 subjects who underwent radical prostatectomy after terminating surveillance (Table 2).

4.2. Model specification

We analyzed data from men in the JH-AS cohort who had at least one biopsy while on AS. We used their biopsy results excluding the biopsy at the time of diagnosis (all of these were low-grade biopsies). Results are interpretable as the probability of upgrading in the absence of other-cause death.

Our analysis of the JH-AS data assumes a follow-up window of ten years. We consider several different prior distributions for the time to transition to high grade disease. All place 26% probability before the study based on the JH-AS and JH-IT misclassification probabilities which imply that 26% of men with a low grade biopsy have true high grade disease. They differ in the probability of transitioning within and after the study. The different settings reflect uniform distributions placing probability up to 20, 30, 40 or 50 years after enrollment, with the sum of the components of the vector r defining the Dirichlet prior distributions are equal to 43, 58, 74 and 85, respectively. These settings imply prior probabilities of transition within the study of 25%, 19%, 15% and 13%, respectively. Corresponding prior probabilities of transition after the study are 49%, 55%, 59% and 61%.

Our primary analysis assumes priors for biopsy misclassification error rates derived from the JH-IT cohort. For sensitivity analysis, we also conduct an analysis using priors for misclassification rates derived from the subcohort of the JH-AS data for whom grade information at surgery was available.

4.3. Results

Biopsy misclassification rates—Table 1 shows Gleason score classification on biopsy (BX) and on radical prostatectomy (RP) from the JH-IT cohort and patients in the JH-AS who underwent RP. In the JH-IT cohort, 46.9% of cases had Gleason scores of 7 or above in the surgical specimen. Of these, 36.2% were classified as Gleason 6 or below on biopsy. In the JH-AS cohort, the numbers were similar; 43.1% had Gleason 7 or higher in their surgical specimen, and 37.9% of these had been wrongly graded as Gleason 6 or below on their final biopsy. Thus, the false negative rates in the two cohorts were comparable. However, the false positive rate was higher in the JH-AS cohort (17.2% versus 10.5% in the JH-IT cohort).

Based on the data displayed in Table 1, for our primary analysis we assume a priori that $FP \sim \text{Beta}(479, 4064)$ and $FN \sim \text{Beta}(1452, 2553)$ based on the JH-IT cohort data. For the sensitivity analysis, our prior is based on the JH-AS data and, thus, we assume that a priori $FP \sim \text{Beta}(16, 73)$ and $FN \sim \text{Beta}(26, 42)$.

Data analysis—The dataset comprises 627 men diagnosed with prostate cancer at a mean age of 65.7 years (standard deviation (sd)= 5.8). The average PSA at diagnosis was 5.3 ng/ml (sd=2.7). On average, participants had 9.1 follow-up visits (sd=5.3) with an average time between visits of 5.0 months (sd=1.7). Of the 627 men, 130 were upgraded while on AS.

Table 3 summarizes the calculation for posterior probabilities of grade change among hypothetical individuals with different series of observations. The table considers cases with different lengths of observation and different final outcomes and shows how the prior probabilities of transitioning before, during and after the study are altered by the data and

the estimation algorithm. For example, for an individual with a single negative biopsy, the probability of transitioning before the study is reduced (from 26% to 13%) and the probability of transitioning after the study is increased (from 49% to 59%). Conversely, for an individual with a single positive biopsy, the probability of transitioning before the study is increased (from 26% to 66%) while the probability of transitioning after the study is reduced (from 49% to 20%).

Figure 1 shows our model estimates of the population parameter vector p and at times 0^- or 10^+ these correspond, respectively, to upgrading before (8%) or after the study window of 10 years (71%). The sum of estimates at times 1 to 10 provides an estimate for Gleason upgrade within the study window (22%). If we only use observed data on biopsy grade we can estimate the analogous distribution of time to upgrading based on a Kaplan-Meier curve fit to the observed times to Gleason upgrading. This is interpretable as the distribution of the time to upgrading in the absence of other cause death. Using the Kaplan-Meier curve we estimate that the probability of transitioning within the study is 62.6%. Thus, the estimate based on observed serial biopsy grade indicators considerably overestimates the estimate based on our approach.

Table 4 presents estimates of the likelihood of upgrading before, within and after the study period under several different priors for the misclassification rates and time to upgrading. The results using priors for the error rates based on JH-AS and JH-IT studies are similar when making inferences about changes within the ten-year study window. Our estimate of the likelihood of upgrading before the study window is slightly lower when using priors derived from the JH-IT cohort than when using priors derived from the JH-AS cohort, with the opposite occurring for inferences about upgrades after the study window. The results under different upgrading priors consistently show high probability of upgrading after the study. Moreover, while the probability of upgrades during the study depends on the prior, they all indicate a non-negligible probability. In particular, among those upgrading within the ten-year study window, we estimated that the average number of years from diagnosis to grade change, is 6.35 (sd=0.79) and 6.04 (sd=0.79) years when using, respectively, priors derived from the JH-IT and JH-AS cohorts and assuming that the prior probabilities for upgrade before, within and after the study are (0.26,0.25,0.49). The estimates are fairly robust to different prior specifications and indicate that among those upgrading, the change occurs around six years after disease diagnosis.

5. Discussion

This analysis represents a study of grade change in an AS study that goes beyond the observed information on grade and attempts to infer the natural history of true grade change. We used data from the JH-AS cohort which monitors men intensively, with annual biopsies and a referral to treatment on the basis of any aspect of biopsy reclassification. Given rates of biopsy grade misclassification estimated from two contemporaneous Johns Hopkins cohorts with data on both biopsy and prostatectomy grade, we attempted to correct the observed serial grade. Our methods assume non-informative loss to follow-up even though some patients were treated in the absence of biopsy changes. These cases may have elected treatment because of elevated PSA but prior studies have not found clear associations between PSA and grade progression [14]. It is possible that some patients elected treatment because of anxiety or fatigue with the AS protocol, but we cannot think of a reason why this would affect our estimates of the frequency of grade progression.

Our work is within the scope of misclassification methods for diagnostic medicine [18–20]. Earlier Bayesian work on the analysis of diagnostic data under misclassification and in the absence of a gold-standard [21] uses informative priors in a manner similar to our approach.

Our approach also bears resemblance to the models of Zelen and colleagues [15–17] who used data from serial screening studies to estimate the frequencies of disease onset and transition from a latent to a clinical state. Our problem is analogous in that we are estimating the rate of transition from a low-grade to a high-grade disease state and, as in the case of screening, we are using a diagnostic test (here, prostate biopsy) to determine whether the transition has happened. Indeed, our likelihood is analogous to that of Zelen and colleagues but rather than integrating over the latent transition time we use an MCMC approach which treats the latent transition time as a missing variate and generates it as one of the steps in the iterative sampling process. In a sense our problem is more straightforward than the screening problem in that we know that all of our cases have already had disease onset, so we can focus solely on estimating the rate of transitioning to the high-grade state. Like the screening setting, we are essentially using interval censored data to make our inferences. However, there are some differences. First, we allow for the diagnostic test to have a non-zero false-positive probability; the screening models generally ignore false positive tests. Second, our data are all interval censored; we are not able to observed the exact time of transitioning to high-grade disease in any of the cases in our dataset. In the screening setting this would correspond to only having data on cases detected at the various screening tests and not having any data on cases detected between screening tests (interval cases). In the screening setting, these interval cases are critically important and without them, the tail of the distribution of the time spent in the latent state may be difficult to identify.

The results consistently indicate low probability of transition before the study window of ten years and this probability is robust to the prior distribution. This is expected given the accumulation of negative biopsy results in the data and suggests a self-selection into the AS program. However, the partitioning of the transition probability, into portions representing transitions occurring during and after the study, is sensitive to the prior distribution. This is likely a consequence of the availability of only interval censored data on the timing of the upgrading transition, and the allowance for falsely positive biopsies (biopsies that are erroneously interpreted as high grade). In this setting the prior distribution becomes critically important for identification of the posterior, therefore the results should be interpreted in the context of the specific prior distribution used. Indeed, perhaps the most valuable contribution our approach is that it provides an explicit means of showing how observed data on serial biopsies modify any specific estimate of the likelihood of biopsy upgrading within a specified study interval.

It is noteworthy that none of the posterior distributions place zero (or close to zero) mass within the study interval, indicating that the data do not rule out the possibility of true upgrading during the decade following diagnosis. Under the 30-year prior that places 25% probability in this interval and under misclassification frequencies based on the JH-IT cohort we estimate that approximately one-fifth of men with (true) Gleason score 6 or below would transition to Gleason 7 or above within 10 years. Using misclassification frequencies from the JH-AS cohort yields lower progression rates (approximately one-sixth over 10 years) due to a considerably higher FP rate among the JH-AS cases who underwent surgery. Whether the higher FP rate among these cases represents a true difference in pathology errors between men treated immediately and those treated following a period of AS is unclear.

Two prior modeling studies examined dedifferentiation in prostate tumors. Draisma et al. [1] did not present estimates of the likelihood that prostate tumors would progress from low grade to high grade but concluded, based on observed frequencies of high-grade tumors in a serial screening program, that few tumors dedifferentiate within a four year screening interval. In contrast to our model, theirs has three grade groups: low, moderate and high grade. Pashayan et al. [2], on the other hand, estimated that 51% to 66% of low-grade,

preclinical tumors transition to high grade if they are not otherwise clinically diagnosed. However, our setting of AS represents a somewhat different population of cases than both of these prior studies which were conducted in the context of population early detection programs.

Our analysis indicates that it is important to incorporate accurate estimates of the likelihood that disease grade will be incorrectly classified by prostate biopsy. It is possible to modify our estimation procedure to account for uncertainty in the FN and FP estimates; when we did so in the analysis of the JH-AS data, we obtained similar results but our credible intervals were considerably wider, particularly when we used the misclassification errors from the JH-AS rather than the JH-IT cohort and accounted for the smaller sample used to estimate these errors.

In conclusion, our analysis indicates that dedifferentiation is a real phenomenon; it may occur in practice. Indeed, the only way we were able to produce a posterior estimate of progression within the ten-year study that was close to zero was to set the prior probability of progression during this time to zero. At face value, this is a concern because of the considerably poorer prognosis of high-grade cases. However, most upgrading in the JH-AS cohort was from Gleason 3+3 to Gleason 3+4 and there was a very low incidence of Gleason 8–10 disease at treatment among these cases. The implications for prognosis remain to be seen, but the information provided by our analysis should be useful in the further understanding the natural history of prostate cancer and in optimizing AS protocols for the many men who are being diagnosed with low-risk disease in contemporary practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by Awards U01 CA157224, R01 CA160239 and RC2 CA148086 from the National Cancer Institute and the Prostate Cancer Foundation (PCF.org). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute. No conflict of interest exists for any author regarding the funding source. We thank Alex Tsodikov for comments on a previous draft of this manuscript.

References

1. Draisma G, Postma R, Schroder FH, van der Kwast TH, de Koning HJ. Gleason score, age and screening: Modeling dedifferentiation in prostate cancer. *International Journal of Cancer*. 2006; 119:2366–2371.
2. Pashayan N, Pharoah P, Neal DE, Hamdy F, Donovan J, Martin RM, Greenberg D, Duffy SW. PSA-detected prostate cancer and the potential for dedifferentiation-estimating the proportion capable of progression. *International Journal of Cancer*. 2011; 128:1462–1470.
3. Tseng KS, Landis P, Epstein JI, Trock BJ, Carter HB. Risk Stratification of Men Choosing Surveillance for Low Risk Prostate Cancer. *Journal of Urology*. 2010; 183:1779–1785. [PubMed: 20304433]
4. Porten SP, Whitson JM, Cowan JE, Cooperberg MR, Shinohara K, Perez N, Greene KL, Meng MV, Carroll PR. Changes in Prostate Cancer Grade on Serial Biopsy in Men Undergoing Active Surveillance. *Journal of Clinical Oncology*. 2011; 29:2795–2800. [PubMed: 21632511]
5. Warlick C, Trock B, Landis P, Epstein J, Carter H. Delayed versus immediate surgical intervention and prostate cancer outcome. *Journal of the National Cancer Institute*. 2006; 98:355–357. [PubMed: 16507832]

6. Dall'Era MA, Konety BR, Cowan JE, Shinohara K, Stauf F, Cooperberg MR, Meng MV, Kane CJ, Perez N, Master VA, Carroll PR. Active surveillance for the management of prostate cancer in a contemporary cohort. *Cancer*. 2008; 112:2664–2670. [PubMed: 18433013]
7. Newcomb LF, Brooks JD, Carroll PR, Feng Z, Gleave ME, Nelson PS, Thompson IM, Lin DW. Canary Prostate Active Surveillance Study: Design of a Multi-institutional Active Surveillance Cohort and Biorepository. *Urology*. 2010; 75:407–413. [PubMed: 19758683]
8. Tosoian JJ, Trock BJ, Landis P, Feng Z, Epstein JI, Partin AW, Walsh PC, Carter HB. Active Surveillance Program for Prostate Cancer: An Update of the Johns Hopkins Experience. *Journal of Clinical Oncology*. 2011; 29:2185–2190. [PubMed: 21464416]
9. Klotz L, Zhang L, Lam A, Nam R, Mamedov A, Loblaw A. Clinical results of long-term follow-up of a large active surveillance cohort with localized prostate cancer. *Journal of Clinical Oncology*. 2010; 28:126–131. [PubMed: 19917860]
10. Donovan, J.; Hamdy, F.; Neal, D.; Peters, T.; Oliver, S.; Brindle, L.; Jewell, D.; Powell, P.; Gillatt, D.; Dedman, D.; Mills, N.; Smith, M.; Noble, S.; Lane, A. Health Technology Assessment. Vol. 7. England: Winchester; 2003. Prostate testing for cancer and treatment (ProtecT) feasibility study; p. 1-88.
11. Lane JA, Hamdy FC, Martin RM, Turner EL, Neal DE, Donovan JL. Latest results from the UK trials evaluating prostate cancer screening and treatment: the CAP and ProtecT studies. *Eur J Cancer*. 2010; 46:3095–3101. [PubMed: 21047592]
12. Pinsky P, Parnes H, Ford L. Estimating Rates of True High-Grade Disease in the Prostate Cancer Prevention Trial. *Cancer Prevention Research*. 2008; 1:182–186. [PubMed: 19138954]
13. Epstein J, Feng Z, Trock B, Pierorazio P. Upgrading and Downgrading of Prostate Cancer from Biopsy to Radical Prostatectomy: Incidence and Predictive Factors Using the Modified Gleason Grading System and Factoring in Tertiary Grades. *EUROPEAN UROLOGY*. 2012; 61:1019–1024. [PubMed: 22336380]
14. Ross AE, Loeb S, Landis F, Partin AW, Epstein JI, Kettermann A, Feng Z, Carter HB, Walsh PC. Prostate-specific antigen kinetics during follow-up are an unreliable trigger for intervention in a prostate cancer surveillance program. *Journal of Clinical Oncology*. 2010; 28:2810–2816. [PubMed: 20439642]
15. Shen Y, Zelen M. Screening Sensitivity and Sojourn Time From Breast Cancer Early Detection Clinical Trials: Mammograms and Physical Examinations. *Journal of Clinical Oncology*. 2001; 19:3490–3499. [PubMed: 11481355]
16. Shen Y, Zelen M. Robust modeling in screening studies: estimation of sensitivity and preclinical sojourn time distribution. *Biostatistics*. 2005; 6:604–614. [PubMed: 15860542]
17. Zelen M, Feinleib M. On the theory of screening for chronic diseases. *Biometrika*. 1969; 56:601–614.
18. Pepe, M. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2003.
19. Gustafson, P. Measurement error and misclassification in Statistics and Epidemiology: Impacts and Bayesian adjustments. New York: Chapman and Hall; 2003.
20. Broemelling, L. Bayesian biostatistics and diagnostic medicine. New York: Wiley; 2007.
21. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*. 1995; 141:263–272. [PubMed: 7840100]

Figure 1.

Analysis of JH-AS data. Empirical and posterior probabilities of Gleason grade upgrade (based on estimates of the parameter vector p describing the probability of grade change at the population level) assuming prior distributions of biopsy true- and false-negative rates based on the JH-IT cohort. Crosses denote observed proportions based on the Kaplan-Meier curve, while filled squares or filled circles denote posterior median estimates. Full and dashed lines denote the corresponding limits of the 95% posterior credible intervals for the posterior probabilities of grade change. The numbers in the bottom of the figure represent the number of event upgrades (E), the number of subjects at risk of upgrade (N), the posterior medians (M) and the upper (U) and lower (L) boundaries of the 95% posterior credible intervals. Prior probabilities for transition before, during and after the study are 0.26, 0.25 and 0.49 respectively.

Simulation Study. Scenarios (SC) are specified by the percent of true grade changes before (B), within (W) or after (A) a ten-year study-window, and priors for the grade changes and for the misclassification rates. Scenarios 1–7 specify point masses for FP and (1-FN), and the remaining ones incorporate uncertainty in the misclassification rates. We summarize simulation results in terms of the average percent of changes before, within and after the study window (based on posterior estimates of the parameter vector p), and among those placing the change within the study window, we also evaluate bias and RMSE (based on estimates of the subject-specific time to grade change t_{IG}).

Table 1

SC	True Changes			Prior Change			Prior for FP, FN	Estimates			Bias	RMSE
	%B	%W	%A	%B	%W	%A		%B	%W	%A		
1	25	70	5	25	70	5	0.20, 0.40	26	69	5	-0.04	3.91
2	25	70	5	25	70	5	0.10, 0.40	31	66	3	0.76	3.73
3	25	70	5	25	70	5	0.30, 0.40	17	74	9	-1.09	4.17
4	25	70	5	25	70	5	0.20, 0.30	21	70	9	-0.42	3.81
5	25	70	5	25	70	5	0.20, 0.50	32	65	3	0.64	4.06
6	25	70	5	25	50	25	0.20, 0.40	29	59	12	0.03	3.84
7	25	50	25	25	50	25	0.20, 0.40	25	49	25	0.00	3.90
8	25	70	5	25	70	5	Be(1,1), Be(1,1)	25	70	5	-0.05	4.00
9	25	70	5	25	70	5	Be(1,4), Be(2,3)	26	69	5	0.08	3.92
10	25	70	5	25	50	25	Be(1,4), Be(2,3)	22	58	19	-0.60	4.01

Table 2

Gleason score classification on biopsy (BX) and on prostatectomy (RP) in the JH-IT and JH-AS cohorts. Percentages in parentheses represent fractions of the column totals and thus reflect false negative (FN) and false positive (FP) rates and their complements (true positive and true negative rates). In the JH-IT cohort the FN rate is 36% and the FP rate is 11%. In the JH-AS cohort the FN rate is 37% and the FP rate is 17%.

	Immediate RP patients 2004–2011 n=8544 (all with >10 bx cores) JH-IT cohort		Active Surveillance RP patients n=153 (all with >12 bx cores) JH-AS cohort	
BX GLEASON (at final biopsy before RP)	RP GLEASON		RP GLEASON	
	6	> 6	6	> 6
6	4063 (89%)	1451 (36%)	72 (83%)	25 (38%)
> 6	478 (11%)	2552 (64%)	15 (17%)	41 (62%)

Probabilities of true grade change before, during or after the study window of ten years given serial biopsy results for hypothetical men with different lengths of observation and different biopsy outcomes. Prior probabilities for transition before, during and after the study are 0.26, 0.25 and 0.49 respectively.

Table 3

Grade	# Biopsies prior to final negative					# Biopsies prior to final positive				
	0	1	2	3	4	0	1	2	3	4
Change										
Before	0.13	0.06	0.02	0.01	0.00	0.66	0.43	0.23	0.11	0.05
Within	0.28	0.29	0.27	0.26	0.23	0.15	0.26	0.35	0.39	0.41
After	0.59	0.66	0.70	0.73	0.76	0.20	0.32	0.42	0.50	0.54

Table 4

Case Study and Sensitivity Analysis. Posterior estimates (mean, standard deviation, median, lower and upper limits of the 95% posterior credible intervals) of the probabilities of grade upgrade before, within or after the ten year study window (based on estimates of the parameter vector p describing the probability of grade change at the population level). Priors for FN and FP rates are based on the JH-IT or JH-AS cohorts. Prior distributions for the time to upgrade are such that the probability of upgrade before the study is on average equal to 0.26 and place uniform probabilities for upgrades within 20, 30, 40 or 50 years after enrollment.

Upgrade	Prior Prob	Priors for FP/FN based on JH-IT Data					Priors for FP/FN based on JH-AS data				
		Mean	SD	Median	Lower	Upper	Mean	SD	Median	Lower	Upper
Before	0.26	0.08	0.02	0.08	0.05	0.12	0.11	0.03	0.11	0.06	0.18
Within	0.25	0.22	0.06	0.21	0.12	0.36	0.24	0.06	0.24	0.13	0.37
After	0.49	0.71	0.06	0.71	0.56	0.80	0.65	0.06	0.65	0.52	0.76
Before	0.26	0.09	0.02	0.09	0.06	0.13	0.13	0.03	0.13	0.08	0.19
Within	0.19	0.17	0.05	0.16	0.09	0.28	0.18	0.05	0.18	0.09	0.29
After	0.55	0.74	0.05	0.75	0.63	0.82	0.69	0.05	0.69	0.58	0.78
Before	0.26	0.10	0.02	0.10	0.07	0.14	0.14	0.03	0.14	0.09	0.20
Within	0.15	0.13	0.04	0.13	0.07	0.21	0.15	0.04	0.14	0.08	0.24
After	0.59	0.76	0.04	0.77	0.68	0.84	0.71	0.05	0.71	0.62	0.80
Before	0.26	0.11	0.02	0.11	0.07	0.15	0.15	0.03	0.15	0.10	0.20
Within	0.13	0.12	0.04	0.12	0.06	0.20	0.13	0.04	0.13	0.07	0.20
After	0.61	0.77	0.04	0.77	0.69	0.84	0.72	0.04	0.73	0.64	0.80