Evaluation for Surrogacy of End Points by Using Data from Observational Studies: Tumor Downstaging for Evaluating Neoadjuvant Chemotherapy in Invasive Bladder Cancer

Satoshi Teramukai,1 Hiroyuki Nishiyama,2 Yoshiyuki Matsui,2 Osamu Ogawa,2 and Masanori Fukushima1

Abstract

Purpose: In clinical cancer trials for evaluating neoadjuvant chemotherapy, tumor downstaging is frequently used as a surrogate end point for overall survival. We evaluated the surrogacy of tumor downstaging using data from a follow-up observational study in bladder cancer.

Experimental Design: A total of 586 patients (from 32 Japanese hospitals) who underwent radical cystectomy for invasive bladder cancer (clinical T2 to T4) between 1990 and 2000 were analyzed. We considered changes over time in clinical stage at diagnosis and pathologic stage at cystectomy as a surrogate end point, and survival time after cystectomy as a true end point. First, we developed a new criterion for tumor downstaging. Second, we statistically evaluated surrogacy for the criterion using Prentice's criteria.

Results: To develop the criterion of end points based on tumor downstaging, we selected the best classification among all possible classifications in an attempt to separate prognosis for patients. The hazard ratios after adjustment for prognostic factors in the intermediate effect patients and the poor effect patients were 1.9 (95% confidence interval, 1.0-3.7) and 5.0 (95% confidence interval, 2.6-9.8), respectively, compared with that in the good effect patients. The conditions for correlation and conditional independency of Prentice's criteria were satisfied approximately. Neoadjuvant chemotherapy has a statistically significant tumor downstaging effect, whereas there was no difference on survival between treatment groups.

Conclusions: The tumor downstaging effect could be an appropriate intermediate end point for screening novel neoadjuvant chemotherapy for invasive bladder cancer. The dataset from follow-up studies were useful for evaluating the surrogacy of end points.

Appropriate surrogate end points are critical for developing new therapies through evaluation of biological activity. The surrogate end point is a test, measurement, score, or some other similar variable that is used in place of a clinical event in the design of a trial, or in summarizing results from it. Used because the variable is believed to be correlated with the clinical event of interest and because of its perceived utility in yielding detectable treatment differences (1). In clinical cancer trials, overall survival is considered to be the most reliable and definitive true end point. However, surrogate end points such as tumor burden outcomes including objective tumor effect, disease-free survival, and progression-free survival, or biomarkers including prostate-specific antigen have been widely used because trials with the true clinical outcome are often longer and larger. In a recent analysis for oncologic drugs in the U.S., 68% (39 of 57) of the regular approvals and all of the 14 accelerated approvals were based on end points other than overall survival in the last 13 years (2). To use a valid and reliable surrogate end point in cancer clinical trials, we should evaluate the surrogacy of end points on a case-by-case basis because the adequacy as a surrogate end point is highly dependent upon the type and/or stage of cancer, and other available therapies.

For statistical validation of surrogate end points, Prentice (3) proposed the validity criterion that a valid between-group analysis of the surrogate end point also constitutes a valid analysis of the true clinical end point. Freedman et al. (4) showed that these criteria were not straightforward to verify by hypothesis testing. Recently, Buyse et al. (5) have proposed two new measures, termed “relative effect” and “adjusted association.” However, to explore the validity of a surrogate end point by these measures, we have to combine information from several randomized clinical trials testing the effect of a treatment on both the surrogate and the true end points (6). In practice, we rarely have information about both end points from even single randomized clinical trials before designing a feature clinical trial for new agents. Such situations have motivated us to assess the surrogacy of end points using
available information other than randomized studies. In clinical trials for evaluating neoadjuvant chemotherapy in bladder cancer, “tumor downstaging” is frequently used as a surrogate end point for overall survival. Clinical staging with transurethral resection (TUR) is very important in treatment planning and prognosis. However, the reliability of TUR staging is a problem. The disparity between clinical and pathologic staging may be caused by repeat TUR, i.e., TUR effect, and measurement error. We developed a new criterion of tumor downstaging effect and evaluated the surrogacy of tumor downstaging using data from a follow-up observational study in invasive bladder cancer.

### Patients and Methods

A total of 1,131 patients who underwent radical cystectomy for invasive bladder cancer between 1990 and 2000 at 32 Japanese hospitals were retrospectively registered (8). The information that was collected from the medical records included age, gender, histology, clinical staging, and pathologic staging according to the tumor-node-metastasis classification (9), and the presence of perioperative systemic chemotherapy. In the present study, 586 patients who have clinical stage T2 to T4, N0, M0, transitional cell carcinoma, and who were less than 80 years old were included.

Figure 1 shows a schema of treatment group comparison. The patients were divided into two treatment groups, i.e., neoadjuvant chemotherapy (NAC) group and no neoadjuvant chemotherapy (non-NAC) group. After the clinical staging was done based on diagnostic TUR, chemotherapy followed by radical cystectomy was done in the NAC group, and only cystectomy was done in the non-NAC group. More precise pathologic staging was done at the time of cystectomy.

### Statistical analysis

Prentice’s criterion for evaluating the surrogacy of end points is a set of four conditions as follows (3, 5, 10):

- **PC1:** $f(T|Z) \neq f(T)$ so the treatment affects the distribution of $T$.
- **PC2:** $f(S|Z) \neq f(S)$ so the treatment affects the distribution of $S$.
- **PC3:** $f(T|S) \neq f(T)$ so the surrogate affects the distribution of $T$.
- **PC4:** $f(T|S, Z) \neq f(T|S)$ so that conditionally on $S$, $T$ is independent of $Z$.

where, for example, $f(T|Z)$ is the conditional distribution of the true end point $T$ given the treatment assignment $Z$, and $S$ is the surrogate end point. In the present study, the treatment $Z$ is set to 0 for non-NAC group and 1 for NAC group. The candidate surrogate end point $S$ is a tumor downstaging effect based on the difference between clinical stage and pathologic stage and the true end point $T$ is overall survival after cystectomy. Therefore, in this setting, the PC1 means that neoadjuvant chemotherapy must affect overall survival, PC2 means that neoadjuvant chemotherapy must affect tumor downstaging, PC3 means that tumor downstaging must be correlated with overall survival, and PC4 means that tumor downstaging must fully capture the net effect of neoadjuvant chemotherapy on overall survival.

The survival curves were estimated with the Kaplan-Meier method. The Cox proportional hazards model was used to estimate hazard ratios (HR) after adjustment for covariates. All statistical analyses were done by using SAS version 8.02 (SAS Institute, Inc., Cary, NC).

### Results

A total of 586 patients [481 men (82%) and 105 women (18%)], with a mean age of 65.2 years (range, 33-80 years), were treated with radical cystectomy with bilateral lymph node dissection. Out of 586 patients, 183 patients (31%) were treated with neoadjuvant chemotherapy. As the neoadjuvant chemotherapy, methotrexate, vinblastine, doxorubicin, and cisplatin, was used in 43% of patients and used for 1.5 cycles on average. The other patients were treated with the modified cisplatin-based regimens including methotrexate, epirubicin and cisplatin; and cisplatin, cyclophosphamide, and doxorubicin; and cisplatin, adriamycin, and methotrexate, as well as other miscellaneous regimens (11-15). The distributions of prognostic factors in treatment groups were as follows: mean patient age was 65.8 years (SD, 8.8) and 63.7 years (SD, 8.6)

### Table 1. Hazard ratios by clinical stage and pathologic stage

<table>
<thead>
<tr>
<th>Clinical stage</th>
<th>Pathologic stage (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P0/1</td>
</tr>
<tr>
<td>All cases</td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>1</td>
</tr>
<tr>
<td>T3/4</td>
<td>1.5 (0.6-3.6)</td>
</tr>
<tr>
<td>Non-NAC group</td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>1, n = 59</td>
</tr>
<tr>
<td>T3/4</td>
<td>2.6 (0.8-8.5), n = 26</td>
</tr>
<tr>
<td>NAC group</td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>1, n = 27</td>
</tr>
<tr>
<td>T3/4</td>
<td>1.0 (0.3-3.5), n = 40</td>
</tr>
</tbody>
</table>

Fig. 1. Schema of treatment group comparison.

Clin Cancer Res 2006;12(1) January 1, 2006  140  www.aacrjournals.org
Downloaded from clinometeres.aacrjournals.org on January 6, 2018. © 2006 American Association for Cancer Research.
in the non-NAC and NAC groups, respectively. The patient proportion of positive lymph node involvement was slightly higher in the non-NAC group (17.4%) than in the NAC group (14.2%), but that of clinical T3 or T4 was much higher in the NAC group (70.5%) than in non-NAC group (49.6%). Proportions of receiving postoperative chemotherapy were similar in both groups, i.e., 23.1% in the non-NAC group, 23.0% in the NAC group.

Development of tumor downstaging effect criterion. We estimated HRs on the overall survival after cystectomy by 10 combinations of clinical and pathologic stage after adjustment for age, lymph node involvement, and adjuvant chemotherapy (Table 1). The estimated HRs by treatment group were similar to that in all cases. First, the 10 combinations were ordered according to the size of HR [1, T2 to P0/1 (HR, 1); 2, T3/4 to P0/1 (HR, 1.5); 3, T2 to P2a (HR, 1.9); 4, T3/4 to P2a (HR, 2.2); 5, T2 to P2b (HR, 2.4); 6, T2 to P3 (HR, 4.3); 7, T3/4 to P2b (HR, 4.6); 8, T3/4 to P3 (HR, 5.3); 9, T3/4 to P4 (HR, 5.3); 10, T2 to P4 (HR, 11.1)] in all cases. Second, we selected the best classification among all possible classifications in an attempt to separate the prognosis of patients with respect to the Akaike’s information criteria. The total number of examined classifications was 45—9 for two categories (good/poor) and 36 for three categories (good/intermediate/poor). For example, the examined classifications were 1 (good) versus 2 to 10 (poor), 1 to 2 versus 3 to 10, ..., 1 to 9 versus 10 for two categories, and 1 (poor) versus 2 (intermediate) versus 3 to 10 (poor), 1 versus 2 to 3 versus 4 to 10, 1 versus 2 to 4 versus 5 to 10, ..., 1 to 8 versus 9 versus 10 for three categories.

As a result, patients were classified into three categories, i.e., good effect (1, T2 to P0/1), intermediate effect (2-5, T2 to P2a/2b or T3/4 to P0/1/2a), and poor effect (6-10, T2 to P3/4 or T3/4 to P2b/3/4). Survival curves according to the tumor downstaging effect were shown in Fig. 2A. The HRs in the intermediate effect patients and the poor effect patients were 1.9 [95% confidence interval (CI), 1.0-3.7] and 5.0 (95% CI, 2.6-9.8), respectively, compared with that in the good effect patients after adjustment for age, lymph node involvement, and adjuvant chemotherapy. The risks by tumor downstaging effect were similar between treatment groups (Fig. 2B and C).

Statistical evaluation for surrogacy of the end point. It is obvious that to fulfill the PC3 condition, tumor downstaging must be correlated with overall survival because we selected the tumor downstaging in such a way that the patients can be classified based on their overall survival. To verify the PC4 condition that tumor downstaging must fully capture the net effect of neoadjuvant chemotherapy on overall survival, it is usually stated that the coefficient corresponding to treatment effect corrected for tumor downstaging is required to be equal to zero. The HRs between treatment groups by tumor downstaging effect, pooled HR and their 95% CIs were estimated after adjustment for age, lymph node involvement, and adjuvant chemotherapy (Table 2). The estimated pooled HR was 1.06 (95% CI, 0.77-1.47) when stratifying by tumor downstaging effect. Although the nonsignificance of the test in which HR = 1 does not prove the PC4 condition, it was suggested that PC4 might be plausible in this study because the pooled HR was close to 1.

As the data is not from randomized trials, strictly speaking, the inference for treatment comparison is not valid and thus the PC1 and PC2 conditions cannot be evaluated. However, we attempted to verify the PC1 and PC2 conditions after adjustment for the confounding factors. For evaluating PC2, we used the Cochran-Mantel-Haenszel statistic with rank score, i.e., the stratum-adjusted Wilcoxon test, because of imbalance of clinical stage distribution among treatment groups. The effect of neoadjuvant chemotherapy on tumor downstaging effect was statistically significant ($\chi^2 = 16.1, P = 0.001$; Table 3).
To evaluate the PC1 condition, we compared the overall survival between treatment groups by clinical stage. In clinical stage T2, the treatment effect was not statistically significant (HR, 0.87; 95% CI, 0.44-1.70) after adjustment for age, lymph node involvement, and adjuvant chemotherapy. Similarly, in clinical stage T3 or T4, the treatment effect was not statistically significant (HR, 0.98; 95% CI, 0.67-1.43).

Discussion

In this study, we proposed a new tumor downstaging criterion based on prognosis in invasive bladder cancer patients for evaluating neoadjuvant chemotherapy. Objective tumor response has been a widely accepted measure of cancer chemotherapy activity. According to international standards, including WHO criteria (16) and Response Evaluation Criteria in Solid Tumors (17), patients were usually classified into either responders (complete response or partial response) or non-responders (no change or progressive disease). The objective tumor response can be assessed even in single-arm studies, however, in the NAC group of the present study, overall survival had no difference between responders and non-responders for neoadjuvant chemotherapy (adjusted HR, 1.09; 95% CI, 0.59-2.03; Fig. 3). Therefore, objective tumor response might not be a valid surrogate end point for evaluating neoadjuvant chemotherapy in invasive bladder cancer.

Some investigators defined the criterion for tumor downstaging (7, 18). However, few data were available with regard to clinical staging and pathologic staging for patients who were treated with or without neoadjuvant chemotherapy, and no definite criterion has been developed based on the prognosis of patients. In the present study, the HRs among clinical stages were different even on the same pathologic stage, especially on P0/1 and P2b in the non-NAC group (Table 1). This suggests that unmeasurable components, including the clinician’s subjective judgment on clinical stage, might reflect different prognoses. With regard to tumor downstaging in invasive bladder cancer, it is questionable to generalize the findings to other cancers because downstaging can occur without chemotherapy when the tumor is removed by the diagnostic TUR (7).

We statistically evaluated the surrogacy of the end point using data from a follow-up observational study. Prentice’s criterion was useful for that purpose, especially for the evaluation of PC3 (correlation) and PC4 (conditional independency). In the present study, the PC3 and PC4 conditions were satisfied approximately. Although the study is not a randomized trial, it is suggested that the neoadjuvant chemotherapy affects tumor downstaging, i.e., PC2 (tumor downstaging benefit) is acceptable, but the treatment does not affect overall survival, i.e., PC1 (survival benefit) is unacceptable. We gave an actual example of hypothetical situations from other articles (5, 10), which showed that the PC2 does not imply the PC1. As another actual case, a randomized trial for locally advanced bladder cancer concluded that the survival benefit of neoadjuvant chemotherapy was of borderline statistical significance (P = 0.06), whereas the tumor downstaging effect was statistically significant (P = 0.001; ref. 7). Do the inconsistent results between PC1 and PC2 depend on the differences of statistical power for evaluating these conditions? We calculated the power of two kinds of statistical tests, i.e., Wilcoxon rank-sum test for tumor downstaging effect and log-rank test for overall survival, based on our data. If the expected proportions of downstaging effect are 0.50 (good), 0.39 (intermediate), and 0.11 (poor) in the NAC group and 0.29 (good), 0.55 (intermediate), and 0.16 (poor) in the non-NAC group from the data in clinical stage T2, a sample size of 96 in each group will have 80% power to reject the null hypothesis using a Wilcoxon rank-sum test with a 0.05 two-sided significance level (19). On the other hand, if the expected 5-year survival probability in the non-NAC group is 0.5, 0.6, and 0.7 and HR is 0.87, a corresponding sample size in each group will be 1,595, 2,004, and 2,683, respectively, using a 0.05 level two-sided log-rank test for equality of survival curves (20). The difference of statistical power is critical for evaluating the PC1 and PC2 conditions. In two recently published studies, the survival curves for patients treated with neoadjuvant

<table>
<thead>
<tr>
<th>Tumor downstaging effect</th>
<th>No. of patients</th>
<th>HR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>27 (15%)</td>
<td>59 (15%)</td>
</tr>
<tr>
<td>Intermediate</td>
<td>74 (40%)</td>
<td>161 (40%)</td>
</tr>
<tr>
<td>Poor</td>
<td>82 (45%)</td>
<td>183 (45%)</td>
</tr>
<tr>
<td>Pooled, stratified by tumor downstaging effect</td>
<td>183</td>
<td>403</td>
</tr>
</tbody>
</table>

Table 3. Tumor downstaging effect of treatment according to clinical stage

<table>
<thead>
<tr>
<th>Clinical stage</th>
<th>Treatment</th>
<th>Tumor downstaging effect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Good</td>
<td>Intermediate</td>
</tr>
<tr>
<td>T2</td>
<td>NAC</td>
<td>27 (50%)</td>
<td>21 (39%)</td>
</tr>
<tr>
<td>non-NAC</td>
<td></td>
<td>59 (29%)</td>
<td>111 (55%)</td>
</tr>
<tr>
<td>T3/4</td>
<td>NAC</td>
<td>0</td>
<td>53 (41%)</td>
</tr>
<tr>
<td>non-NAC</td>
<td></td>
<td>0</td>
<td>50 (25%)</td>
</tr>
</tbody>
</table>
methotrexate, vinblastine, doxorubicin, and cisplatin was superior for patients treated with cystectomy alone, with a HR of 0.74 (95% CI, 0.55-0.99) in a randomized trial (7, 21), and platinum-based combination chemotherapy showed a survival benefit with a HR of 0.87 (95% CI, 0.78-0.97) in a meta-analysis of individual patient data (22). The HR which we assumed to calculate the power might be plausible from these results. However, an important question for implementing neoadjuvant chemotherapy for patients with invasive bladder cancer remains, i.e., how do we select the appropriate patients for combination therapy (23).

Buyse et al. (5, 6) have emphasized that we have to combine information from several randomized clinical trials testing the effects of treatment on both surrogate and true end points to explore the validity of a surrogate end point. In practice, we must assess the surrogate of a candidate end point without data from a randomized trial because the primary objective of a randomized trial will often be to evaluate survival benefit, hence, if the survival benefit were known to be true, then one would have to question the value of conducting such a study. Nonetheless, the purpose of the evaluation of surrogacy should be restricted to find out "appropriate intermediate end points" (10). Fleming et al. (24) also pointed out that surrogate end points can be useful in phase 2 screening trials for identifying whether a new intervention is biologically active and for guiding decisions about whether the intervention is promising enough to justify a large definitive trial with clinically meaningful outcomes. The basic premise is that we cannot predict a treatment effect on the true end point from the effect on the surrogate end point. In conclusion, the tumor downstaging effect could be an appropriate intermediate end point in phase 2 trials for screening novel neoadjuvant chemotherapy in invasive bladder cancer. The dataset from follow-up studies were useful for evaluating the surrogacy of end points.

Acknowledgments

We thank the hospitals associated with Kyoto University, Nara Medical University, and Nagoya University for providing clinical data.

References

Evaluation for Surrogacy of End Points by Using Data from Observational Studies: Tumor Downstaging for Evaluating Neoadjuvant Chemotherapy in Invasive Bladder Cancer

Satoshi Teramukai, Hiroyuki Nishiyama, Yoshiyuki Matsui, et al.