Original article

Determining prognostic factors for gastric cancer using the regression tree method

YOSHITAKA YAMAMURA¹, TOSHIFUSA NAKAJIMA², KEIICHIRO OHTA², ATSUSHI NASHIMOTO³, KUNIYOSHI ARAI⁴,
MASAHIRO HIRATSUKA⁵, MITSURU SASAKO⁶, YASUHIRO KODERA¹, and MASASHI GOTO⁷

¹Department of Gastroenterological Surgery, Aichi Cancer Center Hospital, 1-1 Kanokoden, Chikusa-ku, Nagoya 464-8681, Japan

²Department of Surgery, Cancer Institute Hospital, Tokyo, Japan

³Department of Surgery, Niigata Cancer Center, Niigata, Japan

⁴Department of Surgery, Tokyo Metropolitan Komagome Hospital, Tokyo, Japan

⁵Department of Surgery, Center for Adult Disease, Osaka, Japan

⁶Department of Surgery, National Cancer Center, Tokyo, Japan

⁷Department of Informatics and Mathematical Science, Osaka University, Osaka, Japan

Abstract

Background. The regression tree method is a useful statistical technique that has been little used in the analysis of prognosis.

Methods. The prognostic factors of gastric cancer were investigated, using the regression tree method, in 555 patients who had undergone curative resection for serosa-negative gastric cancer and who were enrolled in a randomized controlled trial of postoperative adjuvant chemotherapy (JCOG [Japan Clinical Oncology Group] 8801 study).

Results. By the regression tree method, the first divided prognostic factor (the most important factor) was lymph node metastasis; in particular, extent of lymphatic spread had the greatest impact on prognosis. In addition, age, tumor size, depth of invasion, and individual dose intensity were found to be significant prognostic factors, whereas sex, tumor location, macroscopic tumor type, and extent of lymph node dissection were not. The resulting tree structure consisted of nine terminal nodes with different prognostic factors, and four clusters were obtained by the merging of terminal nodes that showed a similar prognosis. The cluster which showed the best survival rate (5-year survival rate, 0.986) consisted of two terminal nodes: node 12, which contained N0T1 patients who were younger than 62 years and had a tumor size of less than 7.5 cm, and node 14, which contained N1 patients who were younger than 46 years.

Conclusion. In serosa-negative gastric cancer, lymph node metastasis was the most important prognostic factor. Utilization of the regression tree method enabled visual interpretation of the results of statistical analyses through the graphic representation of prognostic factors. It allowed the identification of the optimal combination of these prognostic factors that defined several groups of patients with distinct prognoses and may serve as a useful reference for the individualization of treatment strategy.

Key words Regression tree method · Prognostic factor · Gastric cancer

Introduction

Various analytical techniques have been developed to evaluate patient survival for different diseases, and one of the important objectives for survival analyses is to identify factors that affect survival and mortality. In the past, multivariate analyses have been used for survival assessment, and proportional hazards methods [1] have been widely applied. These proportional hazards methods have been quoted as being “dream statistical techniques” that allow the analysis of treatment effects and prognostic factors. However, these methods are not without disadvantages: because they are based on proportionality of hazards, they cannot be applied to cases lacking proportionality.

The regression tree method was developed by Friedman [2] and Breiman et al. [3], and was subsequently applied to the analysis of survival by Matsubara et al. [4]. It consists of an analytical process by which the relative significance of each of the prognostic factors is evaluated, and an integral process whereby the optimal combination of these prognostic factors for identifying a group of patients with a similar prognosis can be decided. This method has received more attention in recent years as one of the computer-oriented approximation methods. Its algorithm is very clear and, unlike other regression analysis methods, there are few restrictions and assumptions. A further advantage of this method is that the graphical representation of analytic data is easily understandable by individuals who are not familiar with statistical analyses. As a personal computer is now sufficient to compute the vast amounts of data required by the regression tree

method, greater application is envisaged in the near future.

Despite the obvious advantages that the regression tree model offers, few reports have used it in clinical settings [5,6]. The present study was conducted with the aim of disseminating knowledge about this technique. Statistical analyses using the regression tree model were performed with a personal computer (STATSURV; Fujitsu, Tokyo, Japan) to investigate the prognostic factors for gastric cancer in patients enrolled in a randomized controlled trial (JCOG 8801 study) [7] conducted by the Gastric Cancer Surgical Study Group of the Japan Clinical Oncology Group (JCOG).

Subjects and methods

Subjects

The following institutions participated in the JCOG 8801 study: the Cancer Institute Hospital, Aichi Cancer Center Hospital, Niigata Cancer Center, Tokyo Metropolitan Komagome hospital, Center for Adult Disease, and the National Cancer Center. The subjects had undergone curative resection at one of the above institutions at some point between 1988 and 1992 and were diagnosed as having (1) T1 cancer with lymph node metastasis or (2) T2 cancer regardless of nodal metastasis. The randomized controlled trial aimed to prove the survival benefit of adjuvant chemotherapy with three-drug combination chemotherapy consisting of mitomycin (i.v.), fluorouracil (i.v.), and uracil plus futrafur (UFT; p.o.) in postoperative gastric cancer patients, but the final report disclosed no statistically significant survival advantage with adjuvant chemotherapy [7]. At the end of 1997, all 579 patients who were enrolled in the JCOG 8801 study were still available for follow-up. After the exclusion of patients who died of unrelated causes, 555 patients served as subjects in the present study.

Candidate prognostic factors

Table 1 shows the categories and types of potential prognostic factors for serosa-negative gastric cancer in the previous study [7]. The terminology follows the *Japanese classification of gastric carcinoma (first English edition)* [8]. Only data available from the case report forms used during the JCOG 8801 study were used in this study. Host factors included sex and age, and tumor factors included tumor location, depth of invasion, lymph node metastasis, macroscopic tumor type, and tumor size. Therapeutic factors included lymph node dissection and individual dose intensity, which was calculated by dividing the amount of a drug planned to be administered by the amount of the drug actually admin-

Table 1. Prognostic factors and categories

Factors and categories	No. of cases
Sex	
Male	350
Female	205
Tumor location	
A	207
M	223
C	114
AMC	11
Depth of invasion	
T ₁	425
T ₂	130
Lymph node metastases	
N ₀	230
N ₁	279
N ₂	43
N ₃	3
Macroscopic types	
0	101
1	38
2	127
3	114
4	16
5	159
Lymph node dissection	
D ₀	0
D ₁	18
D ₂	477
D ₃	60
Age	555
Tumor size	555
IDI	555

IDI, Individual dose intensity of chemotherapeutic agents

istered during a given period. In other words, if only half of the planned dose was administered to a patient, then the individual dose intensity for this patient was 0.5 [9,10]. Optimal separation of continuous variables such as age, tumor size, and individual dose intensity measurements was arbitrarily determined after repeated analysis at different cutoff values.

Regression tree method

A summary of the regression tree method is discussed below.

Division of the patient group. The patient group was randomly divided into two subgroups with respect to one of the categories shown in Table 1, and then a log-rank test was performed to compare survival distribution of the two subgroups. This process was repeated for all possible combinations in order to determine the combination with the smallest *P* value. This represents the optimal point at which to divide the patient group into two subgroups or nodes. Next, each node was sub-

jected to the same process so that the node was divided into two smaller nodes. In this manner, nodes were divided into smaller and smaller nodes until the size of the node became smaller than the predetermined sample size ($n = 10$) or P values remained above 0.05. Nodes at which further division is not possible are called terminal nodes.

When a regression tree is constructed in this manner, factors that are divided earlier (upper nodes) are statistically more relevant prognostic factors.

Merging of terminal nodes. When a regression tree is constructed in the above manner, due to the characteristics of the algorithm employed, the survival distribution of terminal nodes, derived from the same node, could be different. However, there is no proof that the survival distribution of all terminal nodes differs. Therefore, in order to group patients with the same survival distribution, it is necessary to reverse the division process and merge terminal nodes.

In other words, by calculating the log-rank statistic for all possible combinations of terminal nodes, two terminal nodes with the largest P value (>0.05) were merged. Then, in the same manner, merging was repeated between two terminal nodes, between a terminal node and a merged node, or between two merged nodes until the P value was below 0.05 for all possible combinations. Terminal nodes were subsequently grouped into several clusters (patient groups) having different survival distributions.

Results

Of the 555 patients, 102 died before the end of 1997, and their 5-year survival rate was 84.8%. However, all of these patients underwent surgery at least 5 years before the end of 1997, and some deceased patients had survived for more than 5 years after surgery.

As shown in Fig.1, the factor that had the greatest impact on the prognosis for the entire patient population (node 1) was the extent of lymph node metastasis: node 2 contained N0 or N1 patients (509 patients, 83 deceased patients; 5-year survival rate, 84.6%), and node 3 contained N2 or N3 patients (46 patients, 19 deceased patients; 5-year survival rate, 67.3%; $P = 0.000003$). Node 3 could not be divided into two groups that were significantly different for any categories, and was a terminal node.

Node 2 was further divided into node 4, which contained patients younger than 74 years (496 patients, 77 deceased patients; 5-year survival rate, 87.0%) and node 5, which contained patients 74 years and older (13 patients, 6 deceased patients; 5-year survival rate, 61.5%; $P = 0.000366$). Node 5 was a terminal node.

Node 4 was further divided into node 6, containing N0 patients (226 patients, 22 deceased patients; 5-year survival rate, 92.7%) and node 7, containing N1 patients (270 patients, 55 deceased patients; 5-year survival rate, 82.3%; $P = 0.000894$). Furthermore, node 6 was divided into node 8, which contained patients with a tumor size of less than 7.5 cm (210 patients, 17 deceased patients; 5-year survival rate, 93.9%) and node 9, which contained patients with a tumor size of 7.5 cm and greater (16 patients, 5 deceased patients; 5-year survival rate, 80.8%; $P = 0.001865$). Node 7 was itself divided into node 14, which comprised patients younger than 46 years (35 patients, 1 deceased patient; 5-year survival rate, 100.0%) and node 15, which comprised patients 46 years and older (235 patients, 54 deceased patients; 5-year survival rate, 79.7%; $P = 0.008745$). Nodes 9 and 14 were terminal nodes.

Node 8 was further divided into node 10, which contained T1 patients (173 patients, 10 deceased patients; 5-year survival rate, 96.5%) and node 11, which contained T2 patients (37 patients, 7 deceased patients; 5-year survival rate, 80.3%; $P = 0.006329$). Node 10 was further divided into node 12, comprising patients younger than 62 years (109 patients, 3 deceased patients; 5-year survival rate, 98.1%) and node 13, comprising patients 62 years and older (64 patients, 7 deceased patients; 5-year survival rate, 93.8%; $P = 0.008275$). Nodes 11, 12, and 13 were terminal nodes.

Node 15 was divided with respect to the level of individual dose intensity, as follows: node 16, which contained patients with an individual dose intensity of less than 0.98 (201 patients, 51 deceased patients; 5-year survival rate, 77.6%) and node 17, which contained patients with an individual dose intensity of 0.98 and above (34 patients, 3 deceased patients; 5-year survival rate, 91.2%; $P = 0.036587$). Both nodes 16 and 17 were terminal nodes.

These findings show that the extent of lymph node metastasis had the greatest impact on the prognosis of patients with serosa-negative gastric cancer, and whether these patients had N0/N1 status or N2/N3 status was particularly relevant. Although age was an important prognostic factor, its statistical significance varied according to the presence or absence of other prognostic factors. Tumor size was a significant prognostic factor for N0 patients younger than 74 years, and was stronger as a prognostic determinant than the depth of invasion. In addition, of the two therapy-related prognostic factors, individual dose intensity was found to be significant in a limited number of patients (N1 patients 46 years and older but younger than 74 years).

Thus, lymph node metastasis (N), age, tumor size, depth of invasion, and individual dose intensity appeared as significant prognostic factors for serosa-negative gastric cancer, whereas sex, tumor location,

Fig. 1. Regression tree for serosa-negative gastric cancer. *I.D.I.*, Individual dose intensity of chemotherapeutic agents

macroscopic tumor type, and lymph node dissection (D) were not significant prognostic factors in this analysis.

The resulting tree structure had nine terminal nodes (nodes 3, 5, 9, 11, 12, 13, 14, 16, and 17). Node 3 contained every N2 or N3 patient; node 5 contained N1/N0 patients who were 74 years and older; node 9 contained N0 patients who were younger than 74 years and had a tumor size of 7.5 cm or greater; node 11 contained N0, T2 patients who were younger than 74 years and had a tumor size of less than 7.5 cm; node 12 contained N0, T1 patients who were younger than 62 years and had a tumor size of less than 7.5 cm; node 13 contained N0, T1 patients who were 62 years and older but younger than 74 years and had a tumor size of less than 7.5 cm; node 14 contained N1 patients who were younger than 46 years; node 16 contained N1 patients who were 46 years

and older but younger than 74 years and had an individual dose intensity of less than 0.98; and node 17 contained N1 patients who were 46 years and older but younger than 74 years and had an individual dose intensity of 0.98 or above.

A comparison of the survival curves for these nodes showed many overlaps, suggesting that the survival distribution of some nodes was similar (Fig. 2).

Therefore, based on log-rank statistics, these terminal nodes were merged. The results showed that nodes 14 and 12 ($P = 0.90539$), nodes 3 and 5 ($P = 0.7317$), nodes 9 and 16 ($P = 0.7085$), and nodes 13 and 17 ($P = 0.5703$) were merged to create nodes 18, 19, 20, and 21, respectively. Furthermore, nodes 20 and 11 ($P = 0.3386$) were merged to form node 22 (Fig. 3).

Because nodes 18, 19, 21, and 22, which were formed by merging terminal nodes, could not be further

Fig. 2. Survival curves of terminal nodes

Fig. 3. Process of merging of terminal nodes

merged, these nodes were named clusters 18, 19, 21, and 22, respectively. The 5-year survival rates were 98.6% in cluster 18, 92.3% in cluster 21, 78.2% in cluster 22, and 66.0% in cluster 19. The survival distribution varied significantly among these clusters (Fig. 4).

Discussion

The prognostic factors for gastric cancer vary for different patient groups. In the present study, through the use of the regression tree method, the type and statistical significance of factors that influence the prognosis for serosa-negative gastric cancer were investigated in patients who had undergone curative resection (these patients served as subjects in the JCOG 8801 study). The present study identified the following prognostic factors for serosa-negative gastric cancer as significant: lymph node metastasis, age, tumor size, depth of invasion, and individual dose intensity. When the regression tree

method is used, the tendency is that the earlier the division (higher in the tree), the greater the effects on prognosis. We found that the variety and weight of prognostic factors varied for different patient groups. Further, although age was one prognostic factor, its statistical significance varied according to the presence or absence of other factors.

Even when patients had different factors, they were classified as belonging to the same prognostic group. For example, cluster 19 suggested that the prognosis for N2 patients (node 3) was comparable to that for N1 or N0 patients of 74 years or older (node 5). Similarly, cluster 21 suggested that the prognosis for N0, T1 patients who were 62 years or older but younger than 74 years and had a tumor size of less than 7.5 cm (node 13) was comparable to that for N1 patients who were 46 years and older but younger than 74 years and had an individual dose intensity of 0.98 or above (node 17). In this manner, the use of the regression tree method enables the type and statistical significance of prognostic

Fig. 4. Survival curves of merged clusters

factors to be determined for each patient, as well as showing the type of prognostic group to which each patient belongs.

Okajima [11] explored independent prognostic factors for gastric cancer, using Cox's proportional hazards model, in 6540 patients with gastric cancer operated on at the National Cancer Center, Tokyo. The most powerful prognostic factor identified in that study was the depth of invasion, followed by the extent of node metastasis, patient age, hepatic/peritoneal metastases, and the extent of lymphadenectomy. These prognostic factors were not identical with what has been found in the current study, but this was attributable to the difference in the patient population (serosa-negative versus all cases) rather than the difference in the statistical methodology.

The target of the current study was a group of patients with relatively early-stage cancers without involvement of the serosa. These patients have less possibility of peritoneal dissemination, and are therefore expected to have a relatively high possibility of being cured by treatment with surgery alone. Despite this, the extent of lymphadenectomy was not identified as one of the significant prognostic factors in this group. This is probably because the vast majority of the patients (85.9%; 477 of 555) underwent D2 resection. On the other hand, the individual dose intensity was found to be a significant prognostic factor among a subset of patients who were between 46 and 73 years of age and had metastases to the N1 lymph nodes. Although no difference in survival between the two groups had been observed in this randomized trial overall, further testing is warranted to confirm the promising role of chemotherapy in this subset identified by the regression tree method.

Finally, though, there is a limitation of this technique that has to be mentioned. Regression analysis requires the availability of data for all cases to be analyzed. Lack of any data in a particular case necessarily results in

deletion of the case, and this limits the applicability of the technique. In this study, the technique was successfully employed because no data were missing among several essential clinicopathologic variables. This was attributable to the fact that the subjects of the analysis were patients who took part in a clinical trial performed by highly qualified centers. Weakness of the technique will emerge whenever additional analyses, using other potential prognostic factors that were not included in the case report form of a trial, and, therefore, are not fully available, are deemed necessary.

The most advantageous aspect of using the regression tree method is that, despite the complexity of the results, these are presented graphically in an easily understandable manner. Using this technique, we have been able to identify subgroups within a randomized trial that would benefit from further study. Although the regression tree method is not without its problems, it may have a role in survival analyses, especially in the setting of clinical trials.

References

1. Cox DR. Regression models and life tables. *J Royal Stat Soc Series B* 1972;34:187–220.
2. Friedman JH. A tree-structured approach to nonparametric multiple regression. In: Gasser T, Rosenblatt M, editors. *Smoothing techniques for curve estimation*. Tokyo Berlin Heidelberg New York: Springer-Verlag; 1979. pp. 5–22.
3. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. Belmont: Wadsworth; 1984.
4. Matsubara Y, Teranishi T, Goto M. Tree-structured approach for evaluating the prognosis of cancer (in Japanese with English abstract). *Proc Soc for Survival Time Studies on Cancer Patients* 1990;10:15–20.
5. Yamamura Y, Nakajima T, Okabayashi K, Iwanaga T, Nakazato H, Kito T, et al. A tree-structured approach for evaluating prognostic factors of gastric cancer (in Japanese with English abstract). *Oncologia* 1993;26:351–8.
6. Furukawa H, Iwanaga T, Nakajima T, Okabayashi K, Nakazato H, Hiratsuka M, et al. Randomized study with mitomycin C + 5-fluorouracil + cytosine arabinoside (MFC) + 5-fluorouracil,

- MFC + tegafur and uracil (UFT), and MF + UFT in advanced gastric cancer: institutional differences in a multicenter study in Japan. *J Surg Oncol* 1995;60:59–64.
7. Nakajima T, Nashimoto A, Kitamura M, Kito T, Iwanaga T, Okabayashi K, et al. Adjuvant mitomycin and fluorouracil followed by oral uracil plus tegafur in serosa-negative gastric cancer: a randomised trial. *Lancet* 1999;354:273–7.
 8. Japanese Research Society for Gastric Cancer. Japanese classification of gastric carcinoma. 1st English ed. Tokyo: Kanehara; 1995.
 9. Coppin CML. The description of chemotherapy delivery: options and pitfalls. *Semin Oncol* 1987;14(Suppl 4):34–42.
 10. Nakajima T, Okabayashi K, Nakazato H, Iwanaga T, Ohta K, Kito T, et al. Individual dose intensity and relative performance as indices for assessing the quality of clinical cancer chemotherapy and their effect on therapeutic results (in Japanese with English abstract). *Oncologia* 1989;22:100–8.
 11. Okajima K. Prognostic factors of gastric cancer patients. A study by univariate and multivariate analysis (in Japanese with English abstract). *Jpn J Gastroenterol Surg* 1997;30:700–11.