Chapter 7

External validity of a prediction model for residual mass histology in testicular germ cell cancer: an evaluation for good prognosis patients

Abstract

We assessed the external validity of a prediction model for nonseminomatous testicular germ cell cancer patients. The model was developed to predict the probability of retroperitoneal metastases being benign (only necrosis/fibrosis) after chemotherapy treatment. Patients with a high probability of benign tissue may be offered observation as opposed to patients with a low probability, which indicates a substantial risk of residual tumour. These patients should undergo retroperitoneal lymph node dissection (RPLND). We compared the observed histology with the predicted probability in 105 patients with good prognosis germ cell cancer who underwent RPLND between 1995 and 1998. We found that predicted probabilities higher than 5% were in good agreement with the observed frequencies of benign tissue. The concordance (*c*)-statistic was 0.76, suggesting that the model could reasonably discriminate between benign tissue and tumour. However, nearly all predicted probabilities (n=101) were lower than 70%, which might be considered as the lowest value at which observation offers a reasonable alternative to RPLND. Further, 35% of patients currently under observation (84/241) had predicted probabilities lower than 70%. In conclusion, the clinical relevance of the prediction model was limited for the patients who underwent RPLND; use of the model would change the policy from RPLND to observation in only a few. On the other hand, the model might support selection of patients for RPLND, who currently are under observation.

Introduction

Computer tomography (CT) often shows small remnants of retroperitoneal masses after chemotherapy for metastatic nonseminomatous testicular germ cell cancer.[1] The histology of the residual masses may be benign (entirely necrotic/fibrotic), or may contain tumour elements (mature teratoma or viable cancer cells). Resection of a totally benign mass has no therapeutic value and should preferably not be performed. Most resection policies consider only one prognostic factor to predict the histology of residual masses, i.e. mass size after chemotherapy.[2,3] Masses smaller than or equal to 10 mm are generally not resected, although more aggressive approaches have been proposed.[4,5]

Mass size as a single prognostic factor has limited predictive power to discriminate benign tissue from tumour. Some small masses containing tumour are left unresected and larger benign masses are unnecessarily resected. A distinction based on several prognostic factors has the potential to classify masses more accurately as benign or tumour.[5,6] Therefore, a clinical prediction model has been developed that incorporates six well-known predictors.[7] It estimates the probability that a residual mass is completely benign. The predicted probability may support the treating physician in deciding whether a residual mass should be resected or not.

Before any wide use of a prediction model can be encouraged, its ability to produce accurate predictions for patients from different but plausibly related populations ('transportability') needs to be assessed.[8] The prediction model was developed for good, intermediate, and poor prognosis patients according to the International Germ Cell Consensus Classification[9] on the basis of data from patients from six European and US study groups (development data), who were predominantly treated in the eighties with cis-platin based chemotherapy. Patients with a good prognosis (56% of all nonseminomas) have an expected 5-year progression free survival probability of 89%.[9] In this group, particularly, it is important to minimise the therapeutic burden; any unnecessary treatment such as resection should be avoided. We therefore studied the transportability of the prediction model in good prognosis patients treated in the nineties. We were particularly interested in the clinical relevance of the prediction model, i.e. its ability to support decision making for patients after chemotherapy.

Patients and methods

Patients participated in an EORTC/MRC trial of the genito-urinary group (EORTC-30941/ MRC-TE20), which compared three cycles of Bleomycin, Etoposide, Cis-platin (3BEP) with four cycles (3BEP-1EP) and the administration of BEP over five days with three days.[10] 812 good prognosis patients were enrolled between March 1995 and April 1998 (Figure 7.1). The present analysis included only nonseminomas (n = 682), which are defined as good prognosis disease when the site of the primary tumour is not mediastinal; no non-pulmonary visceral metastases are present; and the marker levels are good, i.e. AFP and HCG below 1000 ng/ml and LDH below 1.5 x upper limit of normal value.[9] Patients with an extragonadal primary site (n = 21); patients having no retroperitoneal metastasis (n = 182); and patients with elevated markers after chemotherapy (n = 68) were excluded from the analysis. Out of the remaining 411 patients, 306 patients with a prechemotherapy retroperitoneal metastasis did not undergo retroperitoneal lymph node dissection (RPLND), either because the CT was considered to be normal following chemotherapy (n = 241) or for other reasons (n = 65, e.g. uncompleted chemotherapy). This meant that 105 patients were analysed for the relation between the observed histologies and the predicted probabilities (validation data); 241 patients were analysed for the predicted probabilities only.

Histologic findings at RPLND were classified as benign or tumour. Lesions classified as benign contained only necrotic or fibrotic elements while tumour contained mature teratoma or viable cancer cells.

The prediction model was developed in 544 patients and was described in detail before.[7] The following patient characteristics are needed to calculate the probability of benign tissue: the absence/presence of teratoma elements in the primary tumour, determined as teratoma differentiated (TD) or malignant teratoma intermediate (MTI); prechemotherapy levels of the serum tumour markers AFP, HCG and LDH; maximal transversal mass size measured on CT before and after chemotherapy. The exact formula is:

$$\begin{aligned}\text{linear predictor} = &- 0.98 + 0.86(\text{teratoma-negative}) + 0.87(\text{AFPnormal}) + 0.76(\text{HCGnormal}) \\ &+ 0.97(\ln[\text{LDH}_{\text{st}}]) - 0.28(\text{sqrt}[\text{postsize}]) + 0.15(\text{change})\end{aligned}$$

The variables teratoma-negative, AFPnormal, and HCGnormal are 1 if true and 0 if false. Ln[LDH_{st}] is the natural logarithm of LDH/upper limit of the normal value range, sqrt[postsize] is the square root of postchemotherapy transverse diameter expressed in millimetres, and change is the change (per 10%) in mass size during chemotherapy: ((presize-postsize)/presize)*10. The probability of benign tissue is calculated with the formula: probability = $1/(1 + e^{-\text{linear predictor}})$.

This complex formula has been transformed into a score chart[7] for easy estimation of the predicted probability. The value of each variable corresponds to a number of points and the total number of points corresponds directly to the predicted probability. The formula is also implemented in a spreadsheet, which is available in the public domain (`http://www.eur.nl/fgg/mgz/software.html`).

Missing predictor values (2% of all required values) were imputed based on the correlation with the other predictor variables.[11] The statistical performance of the prediction model was studied with respect to calibration and discrimination. Calibration refers to the agreement between the observed frequencies and the predicted probabilities. Calibration was studied graphically[12] and tested with the Hosmer-Lemeshow test for external validation.[13] Discriminative ability, i.e. whether the relative ranking of individual predictions is in the correct order, was determined with the concordance (*c*)-statistic, which is identical to the area under the ROC curve (ROC area) for binary outcomes.[14] The *c*-statistic represents the likelihood that a patient with benign tissue has a higher predicted probability of benign tissue than a patient with tumour for a random pair of patients with different histological masses.

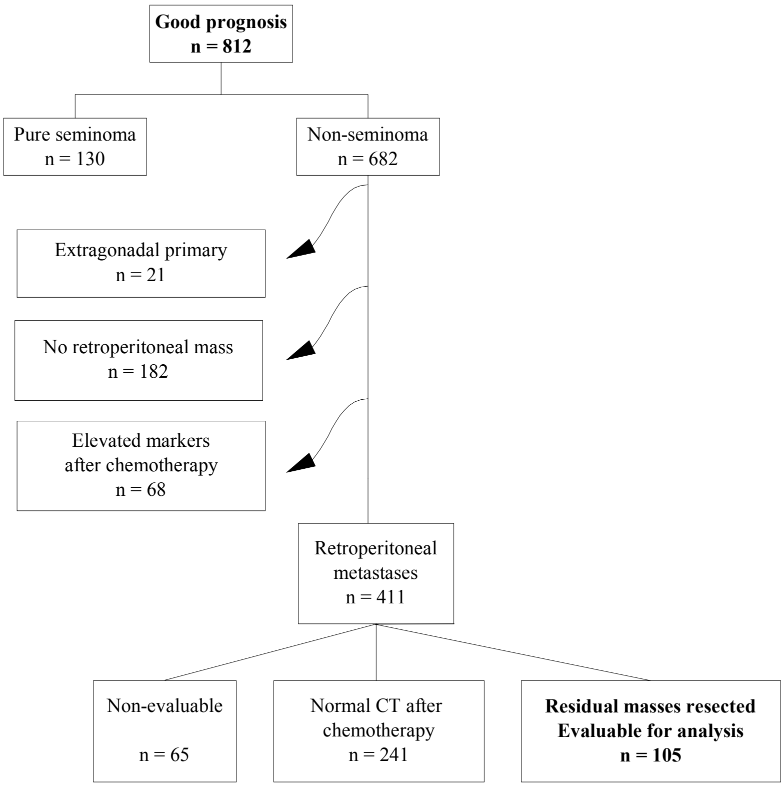

Figure 7.1 Selection of 105 patients from the EORTC-30941/MRC-TE20 study, for whom the prediction model is applied.

In order to classify patients as candidates for observation rather than resection using the prediction model, we applied a threshold value of 70%.[15] Patients with predicted probabilities higher than 70% were selected for observation; patients with probabilities lower than 70% were selected for resection. Using the threshold value, we could study the clinical relevance of the prediction model for the current data. Clinical relevance was expressed as the proportion of patients, who would receive an alternative treatment, if the prediction model was applied (i.e. observation instead of RPLND).

Calculations were performed with SAS version 6.12 and S-plus version 4.5 software, using the *Hmisc* and *Design* library.[12]

Results

Table 7.1 shows the distributions of patient characteristics for the development data and validation data. Merely 26% (27/105) of the patients in the validation data, which only contained patients with good prognosis disease, had benign tissue. The distributions of the prechemotherapy levels of AFP and HCG and of the histology of the primary tumour were similar across the data sets. The validation data contained a far greater number of patients in whom LDH level was normal (72% versus 28%), which follows from the definition of good

Table 7.1 Distribution of the characteristics of nonseminomatous testicular germ cell cancer patients undergoing resection; n (%)

Patient characteristics	*Development data*		*Validation data*	
	n=544		*n=105*	
Primary tumour histology				
Teratoma-negative	252	(46)	56	(53)
Prechemotherapy AFP level				
Normal	186	(34)	33	(31)
Prechemotherapy HCG level				
Normal	205	(38)	50	(48)
Prechemotherapy LDH level				
Normal	151	(28)	76	(72)
Postchemotherapy mass size				
0 - 10 mm	165	(30)	8	(8)
11 - 20 mm	124	(23)	40	(38)
21 - 50 mm	139	(26)	40	(38)
> 50 mm	116	(21)	17	(16)
Change in mass size during chemotherapy				
≥ 70% reduction	161	(30)	6	(6)
0 - 69% reduction	341	(63)	67	(64)
1 - 24% progression	10	(2)	13	(12)
≥ 25% progression	32	(6)	19	(18)
Residual histology				
benign	245	(45)	27	(26)
tumour	299	(55)	78	(74)

prognosis; LDH level should be less than 1.5 times the upper normal value. The post-chemotherapy mass size was larger than 10 mm in 92% of all patients. A very large reduction in mass size during chemotherapy (≥ 70%) was seen in only 6% of the validation data.

Figure 7.2 shows the calibration of the prediction model. The ideal curve represents equality of observed frequencies and predicted probabilities. More than 80% of all patients had predicted probabilities for benign tissue smaller than 50%, which is in agreement with the low proportion of patients who actually had benign tissue (26%). The Hosmer-Lemeshow test for external validation indicated a poor fit (p = 0.001). This was caused by 3 out of 22 patients with benign tissue, while predicted probabilities were below 5%. The fit was satisfactory, when these 3 patients were excluded. The *c*-statistic was 0.76 (95% confidence interval: 0.65 - 0.88), which indicates reasonable discrimination.

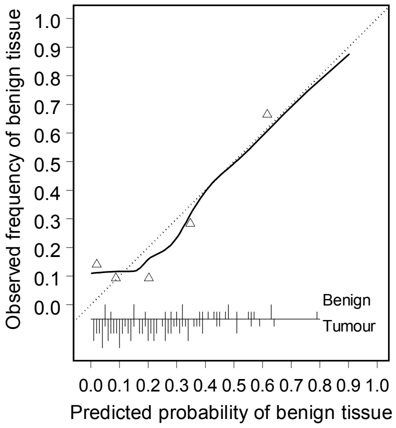

Figure 7.2 Calibration curve of the prediction model in patients of the EORTC-30941/MRC-TE20 study. Triangles indicate the frequency of benign tissue grouped per quintile of predicted probabilities. The solid line shows the relation between observed frequencies and predicted probabilities. Ideally, this line equals the dotted line. Vertical lines at the bottom indicate the distribution of the predicted probabilities; lines upwards represent patients with benign tissue, lines downwards patients with tumour.

At the threshold value of 70%, only four patients (4%) were selected for observation, had the model been applied. Thus the clinical relevance of the model was limited. Three of the four masses were indeed benign. 77 of the 101 patients (76%) selected for resection had masses containing tumour. Some 84 of the 241 masses (35%), which were not resected had predicted probabilities of benign tissue under 70% and would be considered for resection.

Discussion

This study shows a reasonable statistical performance of a prediction model for residual mass histology of nonseminomatous testicular germ cell cancer in 105 recently treated patients with good prognosis disease. However, the clinical relevance of the model was disappointing for these patients.

The prevalence of benign tissue was low, i.e. 26% in contrast to 45% in the development data. This may seem surprising, since we only considered patients with good prognosis disease. However, the studied patients were a selection of all good prognosis patients. Predominantly, patients with residual masses larger than 10 mm were candidates for resection and included in the validation data (92%). It is well known that small masses are more often benign. If more good prognosis patients with very small masses had undergone resection, the proportion of benign tissue would have been higher.

In total, 30% of all patients (32 out of 105) had larger masses after chemotherapy than before, compared with 8% of the patients in the development data. Ignoring the 13 masses that were enlarged by less than 25% (which may simply reflect measurement error) reduces the proportion of enlarged masses to 18% (19 out of 105).

Low predicted probabilities showed disagreement with the observed frequencies, while higher predicted probabilities were well calibrated (Figure 7.2). Since a physician will choose observation over resection only if the predicted probability for benign tissue is relatively high, the model can still be valuable in that decision making process. A larger sample size would, however, be required to provide solid evidence of adequate calibration.

Discriminative ability depends, apart from the studied model, also on the patients to whom the model is applied. If the predictor values of the patients show little variability (homogeneous population), it is difficult to distinguish between patients with different outcomes. Therefore, a *c*-statistic of 0.76 is considered reasonable for our more homogeneous validation data set containing only patients with good prognosis disease. A model with the same six predictor variables developed with the validation data resulted in a slightly larger *c*-statistic (0.78). This confirms the finding that the original model was statistically valid for the good prognosis patients, even though the small sample size and the large confidence interval of the *c*-statistic leave some room for doubt.

If a threshold value of 70% were used for the present patients, only four patients (4%) would be selected for observation over resection. Therefore, application of the model would have little clinical relevance for the present candidates of resection.

We also studied the clinical relevance of simpler models. If all patients with masses ≤ 10 mm were to be closely observed, eight patients would have been denied resection of whom five had tumour. Considering mass size (≤ 10 mm) together with the primary tumour histology (mature teratoma elements absent) would have resulted in only two patients on observation of whom one still had tumour. This suggests that simpler models are not to constitute good alternatives in good prognosis patients. Better discriminating selection models are required, to reduce the morbidity of treatment in these patients.

One third of the patients who did not undergo resection because of small residual masses had predicted probabilities of benign tissue under 70%, which indicates a substantial risk for residual tumour. A number of these patients should have been candidates for resection, particularly since the risks of short-term morbidities associated with resection are probably low given the size of the residual masses.[16] The patients mainly had mature teratoma-positive primaries, elevated prechemotherapy levels of AFP or HCG or a low LDH level. Thus, the prediction model could be particularly relevant in identifying small masses containing tumour. Future studies are required among patients currently on observation to evaluate the role of the prediction model.

Table 7.2 Studies performed to validate the prediction model for residual mass histology

Data	*Hospitals / groups*	*Years*	*n*	*Prognosis*	*Percentage benign tissue*	*Calibration*	*Discrimination*	*Clinical relevance*
Development	six study groups from Europe and US	1979 - 92	544	good/int/poor	45%	OK	*c*-stat = 0.83	142/544 (26%) observation 116/142 (82%) correct
Validation 1	five study groups from Europe	1980 – 96	172	good/int/poor	45%	OK	*c*-stat = 0.82	52/172 (30%) observation 38/ 52 (73%) correct
Validation 2 [a]	IUMC	1985 – 99	276	good/int/poor	28%	recalibration necessary	*c*-stat = 0.79	24/276 (9%) observation 17/ 24 (71%) correct
Validation 3 (present study)	EORTC/MRC	1995 – 98	105	good	26%	OK for predictions > 5%	*c*-stat = 0.76	4/105 (4%) observation 3/ 4 (75%) correct

Calibration: Agreement between observed frequencies and predicted probabilities
Discrimination: Ability to distinguish benign tissue from tumour
c-stat: *c*-statistic
observation: patients selected for observation (predicted probability > 70%)
correct : patients correctly selected for observation (predicted probability > 70% and tissue benign)
[a] an alternative prediction model was validated

To select patients for observation or resection using the prediction model, we applied a threshold value of 70%. The assessment of a sensible threshold value is often difficult. We previously found that the policy to resect all masses larger than 10 mm had an implicit threshold value of 62%.[17] A more stringent policy such as resection in all patients, except in those with masses smaller or equal to 20 mm, having a teratoma-negative primary tumour, and normal prechemotherapy levels of AFP and HCG implied a threshold value of 85%.[5] A threshold value of 70% or 80% therefore seems reasonable.

Like any scientific hypothesis, the transportability of a prediction model is established by being tested and being found valid across increasingly diverse settings.[8] The more numerous and diverse the setting in which the model is tested and found valid, the more likely it is that it will be transportable to an untested setting. Previously, we demonstrated the statistical performance of the prediction model in data of the late eighties (Table 7.2), which was rather similar to the development data.[18] The model predicted slightly too high probabilities, for patients treated between 1985 and 1999 at Indiana University Medical Center (IUMC).[19] For these patients, a simple adjustment of the prediction model may result in better calibrated probabilities.

The model was mainly clinically relevant for the patients from the development and first validation data sets. Around 30% of the patients might have been closely observed. The clinical relevance was poor for the good prognosis patients from the present study (4% would have been closely observed).

In conclusion, the prediction model for residual mass histology is statistically valid in diverse settings. Given the small number of patients in the current study, the validity in good prognosis patients is still not fully certain. Although the clinical relevance was low for the resected patients, the model may be valuable to identify candidates for resection among these with masses smaller than 10 mm containing tumour.

References

1. Peckham M: Testicular cancer. Rev Oncol 1: 439-453, 1988

2. Jansen RL, Sylvester R, Sleyfer DT, et al.: Long-term follow-up of non-seminomatous testicular cancer patients with mature teratoma or carcinoma at postchemotherpy surgery. EORTC Genitourinary tract Cancer Cooperative Group (EORTC GU Group). Eur J Cancer 27: 695-698, 1991

3. Mead GM, Stenning SP, Parkinson MC, et al: The second medical research council study of prognostic factors in nonseminomatous germ cell tumours. J Clin Oncol 10: 85-94, 1992

4. Gelderman WA, Koops HS, Sleijfer DT, et al.: Treatment of retroperitoneal residual tumour after PVB chemotherapy of nonseminomatous testicular tumours. Cancer 58: 1418-1421, 1986

5. Fosså SD, Qvist H, Stenwig AE, et al.: Is postchemotherapy retroperitoneal surgery necessary in patients with nonseminomatous testicular cancer and minimal residual tumour masses? J Clin Oncol 10: 569-573, 1992

6. Donohue JP, Rowland RG, Kopecky K, et al.: Correlation of computerised tomographic changes and histological findings in 80 patients having radical retroperitoneal lymph node dissection after chemotherapy for testis cancer. J Urol 137: 1176-1179, 1987

7. Steyerberg EW, Keizer HJ, Fosså SD, et al: Prediction of residual retroperitoneal mass histology following chemotherapy for metastatic nonseminomatous germ cell tumour: multivariate analysis of individual patient data from six study groups. J Clin Oncol 13: 1177-1187, 1995

8. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. Ann Intern Med 130: 515-524, 1999

9. IGCCG: International germ cell consensus classification: a prognostic factor-based staging system for metastatic germ cell cancers. J Clin Oncol 15: 594-603, 1997

10. de Wit R, Roberts JT, Wilkinson P, et al.: Final analysis demonstrating the equivalence of 3 BEP vs. 4 cycles and the 5 day schedule vs. 3 days per cycle in good prognosis germ cell cancer. An EORTC/MRC phase III study. J Clin Oncol 19: 1629-1640, 2001

11. He X, Shen L: Linear regression after spline transformation. Biometrika 84: 474-481, 1997

12. Harrell FE, Jr.: Design: S-plus functions for biosstatistical/epidemiological modelling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit. Programs available at internet. `http://lib.stat.cmu.edu/DOS/S/Harrell/`. 1997

13. Hosmer DW, Lemeshow S: Applied logistic regression, John Wiley & Sons Inc: New York, 1989

14. Harrell FE, Jr., Califf RM, Pryor DB, et al.: Evaluating the yield of medical tests. JAMA 247: 2543-2546, 1982

15. Steyerberg EW, Marshall PB, Keizer JH,et al.: Resection of small, residual retroperitoneal masses after chemotherapy for nonseminomatous testicular cancer: a decision analysis. Cancer 85: 1331-1341, 1999

16. Gels ME, Nijboer AP, Hoekstra HJ, et al.: Complications of the post-chemotherapy resection of retroperitoneal residual tumour mass in patients with non-seminomatous testicular germ cell tumours. Br J Urol 79: 263-268, 1997

17. Steyerberg EW, Keizer HJ, Habbema JDF: Prediction models for the histology of residual masses after chemotherapy for metastatic testicular cancer. Int J Cancer 83: 856-859, 1999

18. Steyerberg EW, Gerl A, Fosså SD, et al.: Validity of predictions of residual retroperitoneal mass histology in nonseminomatous testicular cancer. J Clin Oncol 16: 269-274, 1998

19. Vergouwe Y, Steyerberg EW, Foster RS, et al.: Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. J Urol 165: 84-88, 2001