Prognostic Modeling for Clinical Decision Making

Theory and Applications

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Steyerberg, Ewout Willem

Prognostic modeling for clinical decision making: theory and applications /
Ewout Willem Steyerberg. - [S.l.:s.n.]. - Ill.
Proefschrift Erasmus Universiteit Rotterdam. - Met lit. opg.
- Met samenvatting in het Nederlands.
ISBN 90-9009431-8
NUGI 743
Trefw.: voorspellen; klinische besluitvorming.

Cover: 'concentrische schillen', M.C. Escher, 1953.

Printed by Elinkwijk BV, Utrecht, The Netherlands

Address for correspondence: E.W. Steyerberg, Department of Public Health,
Ee2091, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam

Prognostic Modeling for Clinical Decision Making

Theory and Applications

Prognostische Modellering ten behoeve van Klinische Besluitvorming

Theorie en Toepassingen

Proefschrift

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR
AAN DE ERASMUS UNIVERSITEIT ROTTERDAM
OP GEZAG VAN DE RECTOR MAGNIFICUS
PROF. DR P.W.C. AKKERMANS M.A.
EN VOLGENS HET BESLUIT VAN HET COLLEGE VOOR PROMOTIES.

DE OPENBARE VERDEDIGING ZAL PLAATSVINDEN OP
WOENSDAG 29 MEI 1996 OM 15.45 UUR.

door

Ewout Willem Steyerberg

geboren te Delft

Promotiecommissie

Promotor:	Prof. Dr J.D.F. Habbema
Co-promotor:	Dr H.J. Keizer
Overige leden:	Prof. Dr E. Bos Prof. Dr D.E. Grobbee Prof. Dr H. van Urk

Financial support by the Netherlands Heart Foundation for the publication of this thesis is gratefully acknowledged.

Additional financial support for the publication of this thesis was provided by Pharmachemie BV, Janssen-Cilag BV and ASTA-Medica BV.

"Data! Data! Data! he cried impatiently,
I can't make bricks without clay" (*Sherlock Holmes*)

Ter nagedachtenis aan mijn moeder
Voor Aleida

Contents

Introduction

1 Clinical decision making, prognosis and modeling

Clinical decision making is concerned with choices made for individual patients, with a focus on diagnosis, therapy and prognosis. These choices can be supported by empirical research. Diagnostic research includes the assessment of test characteristics, such as sensitivity and specificity, while therapeutic research is preferably performed in randomized clinical trials. Both diagnostic and therapeutic decisions aim to improve the prognosis for the patient. Prognosis is therefore at the heart of clinical decision making[1].

1.1 Prognosis

Prognosis refers to all medical outcomes that may occur during the patient's disease process, for example mortality, complications of therapy, reoperations after surgery, or complete recovery from disease. These outcomes have to be viewed in a time-perspective. The time-perspective may be short-term, for example the occurrence of surgical mortality, or long-term, like long-term survival. With a short-term perspective, prognostic estimates can readily be interpreted, for example 'this patient has a 5% risk of surgical mortality'. With a long-term perspective, prognostic estimates have to be more carefully described. For example, long-term survival is for any patient 0% if follow-up is sufficiently long. A commonly used time perspective in oncology is 5 years, although longer perspectives may be required for cancers that may recur during later years. To assess prognosis quantitatively, we may think of the patient's life-expectancy as a summary measure. The life-expectancy is defined as the area under the survival curve. Apart from duration, we may also want to incorporate the quality of the patient's health status in a prognostic qualification. This may be attempted with the construction of a quality corrected life-expectancy, expressed as quality adjusted life years ('QALYs').

Estimates or predictions of prognosis can be made in several ways. A treating physician may rely on his pathophysiologic knowledge and previous experience with more or less similar patients, either in an informal way ('intuitively' or using 'expert opinions') or in a formal way ('5 of my 48 patients died after surgery'). The treating physician may also use published patient series from the medical literature to estimate prognosis. From these series, general knowledge on the disease course may be derived. For example, for patients with metastatic testicular cancer, it can be stated that the prognosis is generally good (long term cure rate of 80%). In case of individual patients, more specific patient and disease characteristics should preferably be taken into account. For example, in the treatment of metastatic testicular cancer, the choice between standard or more intensive chemotherapy is based on the prognostic classification of patients. The distinction is often made between 'good prognosis' and 'poor prognosis', based on multiple disease characteristics, such as tumor marker levels and the extent of disease[2]. Such a classification is often used in the selection of candidates for clinical trials.

Prognostic estimates can also be based on statistical models, which may combine multiple patient and disease characteristics to estimate prognosis quantitatively. For example, such models may estimate the surgical mortality of a 70-year-old patient with an aortic aneurysm and no major comorbidity as 2%, which may be interpreted as that on average 2 out of every 100 of this type of patients will not survive surgery. Statistical models can most support clinical decision making in situations where the benefits of treatment do not clearly outweigh the risks. For example, surgery may be contemplated for an 80-year-old patient with an abdominal aortic aneurysm of 5 cm, or for a 60-year-old patient with an artificial heart valve with a small risk of mechanical failure.

Statistical models make quantitative estimates available for patients with quite diverse patterns of characteristics. This feature makes statistical models a powerful tool for the physician to benefit from the experience with previous patients. The use of statistical models in clinical practice has, however, several risks[3]. The prognostic estimates may for example not be reliable or may not incorporate important prognostic characteristics of individual patients. The role of prognostic models in clinical decision making can therefore only be supportive, with the aim to assist rather than to take over the responsibility of the treating clinician.

1.2 Prognostic modeling: clinical applications

This thesis contains several prognostic models, which were developed for application in three clinical decision problems. These three problems illustrate the practical side of prognostic modeling. The first problem concerns decision making on patients with testicular cancer, the second on patients with an aortic aneurysm, and the third on patients with a mechanical heart valve.

1.2.1 Residual masses in testicular cancer

Patients with a non-seminomatous tumor in one of the testicles are usually first treated with orchidectomy to remove the testicle with the primary tumor. The primary tumor may have caused spread of the disease through the body. Metastases most commonly arise in the abdomen, in the retroperitoneal lymph nodes. Other sites include the lung, mediastinal or supraclavicular lymph nodes, the liver, the brain and the skeleton. If metastases are present, cis-platin-based chemotherapy is administered, usually in four cycles of about three weeks each. The success of this treatment can be monitored by the shrinkage of the metastases and the normalization of tumor markers in the blood. After completion of the chemotherapy courses, remnants of the initial metastases may remain, while tumor markers have normalized. These remnants are called residual masses and can be detected by radiographic methods, especially computer tomography (CT) scanning. These residual masses may contain one of three histologies:

- totally benign tissue (necrosis/fibrosis)
- potentially malignant tissue (mature teratoma)
- residual malignancy of the metastases (viable cancer cells)

The decision problem is whether residual masses have to be surgically removed

(under the condition that surgical resection is considered technically feasible). Alternatively, the mass may be observed regularly during follow-up, without active treatment. In principle, we may think of a third possibility, i.e., additional chemotherapy to treat the residual mass. This strategy is not considered seriously nowadays in patients with normalized tumor markers; one argument against it is that mature teratoma does not respond to further cytostatic treatment.

The two alternatives are thus resection and observation. For example, we may consider a male patient, 25 years of age with a retroperitoneal residual mass of 10 mm (Figure 1). The resection is a laparotomy in this case, which is a major procedure.

If resection is performed, the patient risks surgical mortality and morbidity. A total lymphadenectomy may frequently cause ejaculation problems. Hospital admission for several days is required, with a subsequent period of further recovery. In case of necrosis or mature teratoma in the residual mass, the patient is considered cured and no additional therapy is given. If residual malignancy is found, additional chemotherapy is administered to improve the patient's prognosis, since small amounts of residual malignancy may still be present at other sites as well.

Observation may look interesting, since the probability of totally benign tissue (necrosis) is high in small residual masses, e.g. over 70% in masses ≤ 10 mm. In the case of necrosis, resection has no therapeutic value. On the other hand, if residual mature teratoma or malignancy is still present, the prognosis for the patient is jeopardized by a higher risk of relapse, which is associated with a poor survival.

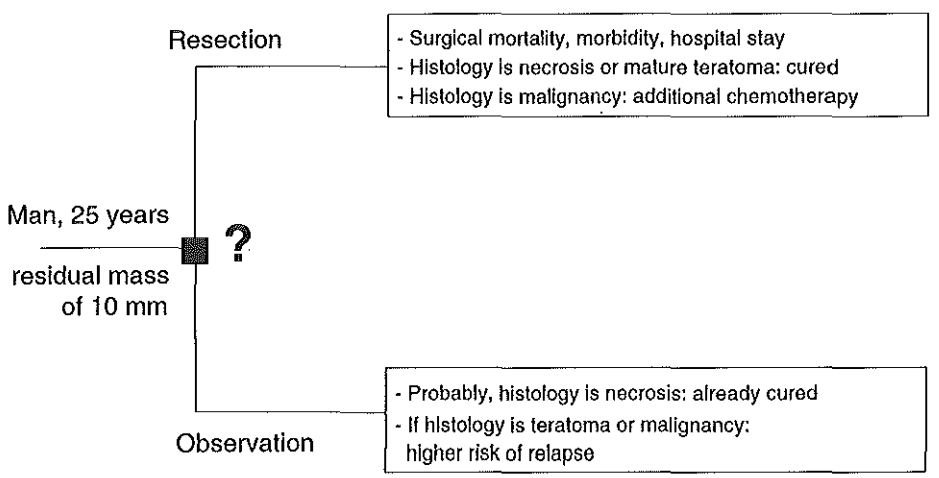

Figure 1 Schematic representation of the decision problem in a testicular cancer patient, 25 years of age, with a residual mass of 10 mm after chemotherapy treatment for metastatic disease.

A crucial factor in this decision making process is the probability of each of the three histologies. Several chapters in this thesis address the prediction of the residual histology, which may, strictly speaking, be considered a diagnostic problem with prognostic implications rather than a pure prognostic problem.

1.2.2 Elective aortic aneurysm surgery

The second clinical decision problem concerns the choice between elective surgery and observation in patients with an abdominal aortic aneurysm. As an example, a female patient is considered, 80 years of age, with an aneurysm of 5 cm (Figure 2). Such an aneurysm may for example have been detected by the patient herself, or during a diagnostic evaluation for another disease, or during specific screening for aneurysms. If elective surgery is performed, the patient risks surgical mortality and morbidity, needs several days of hospital admission and requires a period of reconvalescence. If the peri-operative period is survived, the patient is cured from the aneurysm problem. Alternatively, observation of the aneurysm may be considered, which carries the risk of rupture of the aneurysm. If rupture occurs, the risk involved with acute surgery is much higher than that with elective surgery. Moreover, many patients do not reach the hospital in time for acute surgery. On the other hand, the patient may die of other causes before the rupture occurs, which is not unlikely at the age of 80.

Essentially, the risk of elective surgery has to be weighed against the cumulative risk of rupture with its consequences. Estimation of surgical risk thus is only one of the factors in this decision problem; a prognostic model to estimate this risk is presented.

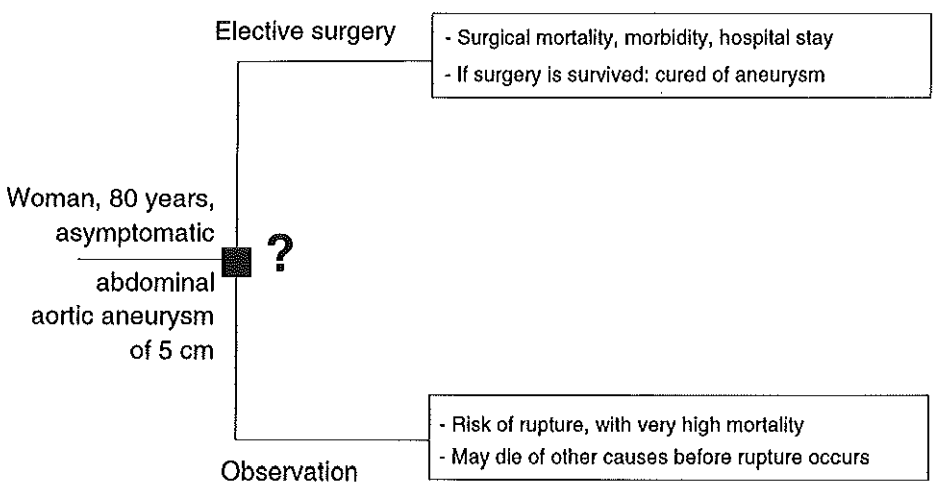

Figure 2 Schematic representation of the decision problem in a female patient, 80 years of age, with an asymptomatic abdominal aortic aneurysm of 5 cm.

1.2.3 Replacement of mechanical heart valves

The third clinical decision problem concerns patients with artificial heart valves with an increased risk of mechanical failure (Björk-Shiley convexo-concave heart valve). For example, we may consider a male patient, 40 years of age with a specific type of such a valve (Figure 3). The valve may be replaced by another valve electively. The elective surgery is associated with a risk of surgical mortality, a risk of morbidity, and with a hospital stay and a period of reconvalescence. If surgery is survived, the patient will live with another artificial heart valve, which may have different hemodynamic characteristics than the original valve. Alternatively, the patient may be observed during follow-up. This choice implies that the patient may suffer the consequences of mechanical failure of the valve. On the other hand, the patient may die from other causes before mechanical failure occurs, since the yearly failure risk is small. Finally, future studies may revise the estimates of the failure risk, thus refining the indication for surgery in this patient.

Prognostic estimates are required for four factors in this problem:

- the risk of elective surgical mortality
- the survival with another artificial heart valve
- the risk of mechanical failure
- the mortality associated with acute failure

Prognostic models are developed to provide these estimates, which are used in a formal decision analysis to indicate which patients may benefit from elective replacement.

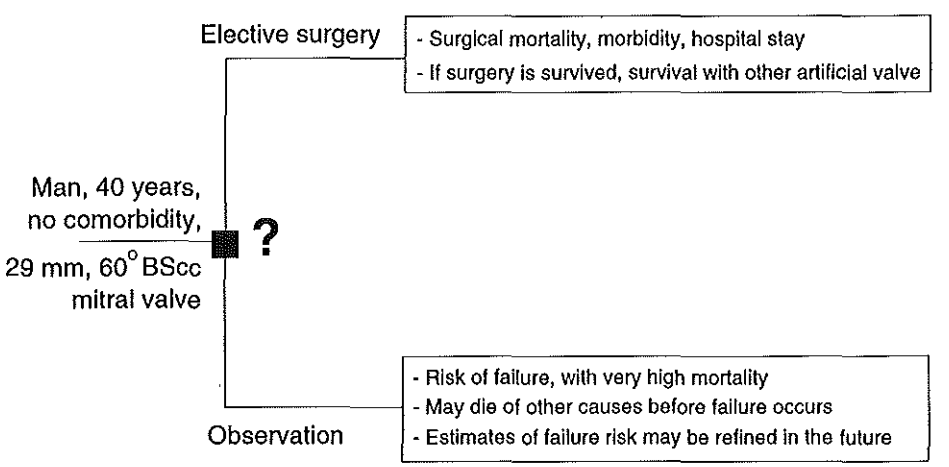

Figure 3 Schematic representation of the decision problem in a male patient, 40 years of age, with no major comorbidity and a Björk-Shiley convexo-concave valve in the mitral position, 29 mm in size and an opening angle of 60°.

1.3 Outline of this thesis

This thesis addresses the development of prognostic models for application in clinical practice. The first part considers theoretical aspects of predictive modeling. These aspects deserve explicit consideration, since it has been noted that the "aim of prognostic modeling requires a change from traditional biostatistical 'explanatory' enterprises of estimation and hypothesis testing"[4]. Chapter 2 focuses on different sources of overoptimism on the performance of prognostic models. Overoptimism refers to the phenomenon that prognostic models perform better on the patients used to derive the prognostic models than on new patients. It is shown that commonly used statistical methods such as stepwise selection greatly contribute to this problem. Suggestions for improvement are given. Chapter 3 introduces a new method for prognostic modeling, which explicitly considers literature data in the model-building process.

Part two of this thesis describes several applications of prognostic models for clinical practice. Chapters 4 to 8 relate to the prognostic aspects in the treatment of metastatic testicular cancer. Chapter 4 is a study of prognosis (especially 5-year relapse-free percentage) after surgery for residual masses. Chapters 5 to 8 relate to the histological content of residual masses, which may, broadly speaking, be benign or malign in a 50:50 ratio. A meta-analysis of 19 studies indicated that several characteristics are related to the histology of residual masses (Chapter 5). A subsequent analysis used these characteristics to predict the histology of residual masses with multivariate logistic regression analysis. Cooperation with several study groups was sought, which resulted in over 500 patients for the analysis of abdominal masses (Chapter 6) and over 200 patients for lung masses (Chapter 8). It appeared that the use of the predictive model for abdominal masses could substantially improve the selection of patients for surgery, which means that more patients with benign histology would be spared surgery while at the same time more patients with residual malignancy would undergo resection (Chapter 7).

Chapter 9 describes the development of a prognostic score chart to estimate elective surgical mortality of abdominal aneurysm surgery, which may be highly relevant for decision making when the risks and benefits of surgery are not obvious. The analysis is based on the theory described in Chapter 3. Chapters 10 and 11 address the replacement of risky artificial heart valves. Chapter 10 is a decision analysis, where prognostic models are used to estimate survival, the risk of mechanical failure and surgical mortality. Chapter 11 shows that the key results of the decision model can adequately be shown graphically or described with a 'meta-model' formed with linear regression analysis. This thesis ends with a discussion of the theoretical and practical results.

References

1. Hilden J, Habbema JDF. Prognosis in medicine: an analysis of its meaning and roles. Theor Med 1987; 8: 349-365
2. Bajorin DF, Geller NL, Bosl GJ. Assessment of risk in metastatic testis carcinoma: impact on treatment. Urol Int 1991; 46: 298-303
3. Concato JC, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Int Med 1993; 118: 201-210
4. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986; 5:421-433

Theory

2 Theoretical aspects of prognostic modeling: a critical review

This chapter addresses theoretical aspects of prognostic modeling for medical decision making. Prognostic models may predict outcomes of various types, such as continuous outcomes (e.g. blood pressure), discrete outcomes with or without an ordening (e.g. benign histology, potentially malign, cancer) or dichotomous (e.g. alive/dead). Regression analysis is the most frequently used statistical method to relate prognostic characteristics of patients to these outcomes. The issues in this chapter focus on logistic regression analysis for dichotomous outcomes[1,2], but many concepts apply to other types of regression modeling which are frequently used in the medical field (ordinary least square regression, Cox proportional hazards regression, Poisson regression[3]).

The regression analyses considered here relate the outcome or transformations of the outcome to a linear combination of predictors x_1 .... x_i. In the case of linear regression the outcome is simply $\hat{y}$, the expected value of the outcome y:

$$\hat{y} = \beta_0 + \beta_1 \bullet x_1 + ... + \beta_i \bullet x_i.$$

In the case of logistic regression analysis, the outcome is the logit of the dichotomous outcome:

$$\widehat{\text{logit}}(y) = \beta_0 + \beta_1 \bullet x_1 + ... + \beta_i \bullet x_i.$$

The term β_0 is the intercept of the model and the terms β_1 .... β_i are named regression coefficients. Regression analyses thus estimate coefficients for each variable (or predictor) in relation to the outcome. The outcomes are in this context labeled 'dependent variables' and the predictors are often labeled 'independent variables'. The regression coefficient of an independent variable has a direct interpretation: with one unit increase in its value, the dependent variable increases with the value of the regression coefficient. This ease of interpretation has certainly contributed to the popularity of regression models. In the case of logistic regression analysis with a single predictor, the odds ratio (OR) of the predictor can be calculated as the exponential of the regression coefficient and is equivalent to the OR as calculated from the cross-table of the predictor with the outcome. The OR is often interpreted as a relative risk. Mathematically, the OR is similar in magnitude to the relative risk if the risk is small (e.g. << 10%) or if the OR is near to one.

Regression analyses may contain a single predictor ('univariate' analysis) or multiple predictors ('multivariate' or 'multivariable' analysis). In a multivariate analysis, the regression coefficients of the predictors are corrected for the prognostic contribution of each other. Multivariate coefficients are therefore often referred to as 'adjusted' coefficients. The multivariate coefficient of a predictor generally differs from the univariate coefficient, if this predictor is correlated with one or more other predictors that are also correlated with the outcome. Most often, the predictors are positively

correlated with each other and with the outcome, and the multivariate coefficients are smaller than the univariate coefficients.

We consider the construction of multivariate logistic regression models on the basis of patient data with the aim to predict a dichotomous outcome in future patients. The modeling process should result in quantitative predictions, based on a set of predictors. This aim of predictive modeling is in contrast to the epidemiological aim of identifying (potentially) etiologic variables. In epidemiology, the prognostic value of one or more individual variables is mainly of interest, similarly to the treatment effect in a randomized clinical trial. Correction for confounding variables is important in this context. In contrast, prognostic models for medical decision making should aim to maximize the prognostic value of a combination of variables, while the specific variables with their coefficients are of lesser importance.

A common prognostic modeling strategy is currently as follows. A clinician has gathered a data set with patient characteristics that may be related to an outcome, for example surgical mortality. In cooperation with a methologically skilled investigator, e.g. a biostatistician, a prognostic model will be developed which uses the patient characteristics in the data set. There are no strong pre-specified hypotheses about the prognostic importance of these characteristics, which are candidate variables for inclusion in the model. Some characteristics may have been reported in the medical literature as relevant, some may be plausible predictors because of pathophysiologic mechanisms, others may have a special interest of the clinician. The first step in the modeling process will be to obtain an impression of the data set under study, for example by simple frequency tables and cross-tabulations of the predictors and the outcome. Further, continuous variables may be classified with different cut-off values and categories of nominal variables may be collapsed to define 'optimal' arrangements of the predictors. Next, a set of variables is selected for the prognostic model, frequently based on an automatic stepwise selection procedure (either forward, backward or in a combined forward/backward way). In this procedure, variables are selected based on the prognostic importance of a variable in addition to a set of other variables in the model. The resulting model after stepwise selection may be modified to some extent based on typical or implausible findings, and the effect of specific combinations of variables may be evaluated (interaction terms). Model performance may be examined with goodness-of-fit tests and measures of discriminative ability. The final model is usually presented as a table showing the selected predictors, the regression coefficients, and the corresponding confidence intervals and p-values. For application in clinical practice, the regression formula is usually presented.

In this illustration of prognostic modeling, the data set is used for far more than the estimation of the regression coefficients. This common modeling strategy is shown schematically in Table 1.

Table 1 Schematic overview of a common prognostic modeling strategy.

Modeling phase	Method
Selection of variables	
Classification of variables	Univariate analysis; optimal classification
Inclusion of variables	Stepwise selection, p<.05
Inclusion of interaction terms	Multivariate analysis
Estimation of regression coefficients	Multivariate analysis
Evaluation of model performance	Discriminative ability and goodness-of-fit
Presentation of model	Regression formula

Decisions that are based on the findings in the data set include the classification of variables, the inclusion of variables and the inclusion of interaction terms. The model developed in this way may perform well on this data set, but much poorer for future patients[4]. In the following, several issues related to this overoptimism are critically discussed. The issues discussed are:

- assumptions of regression modeling
- the selection of variables
- the estimation of the regression coefficients
- the evaluation of prognostic models
- validation and re-sampling methods
- the presentation of prognostic models for application in clinical practice

This chapter ends with a number of practical considerations and conclusions.

2.1 Assumptions of regression modeling

Regression models are valid under a number of assumptions, which will briefly be mentioned here. First, it is assumed that the patients in the data set is a random sample. Technically speaking, the sample should be random conditionally on the values of the predictors. In most regression models it is further assumed that the patients may be considered as independent observations.

Second, the distribution of the regression residuals needs to have certain properties. In the case of ordinary linear regression, the residuals are assumed to have a normal distribution. In the case of logistic regression, the assumption on the residuals might be formulated as that they are binomial distributed. This assumption is fairly natural, as the data are binary (0/1).

Third, the relation of a continuous predictor to the outcome variable has a certain shape. In the case of linear regression, this shape is a straight line. In the case of logistic regression, the 0 to 1 outcome scale is transformed by a the logit transformation (or ln(Odds)) to a scale ranging from minus infinity to plus infinity. Linearity of continuous predictors is assumed on this scale. Shown as a probability, the shape is a characteristic sigmoid curve, which approximates zero with lower values of the predictor, rises

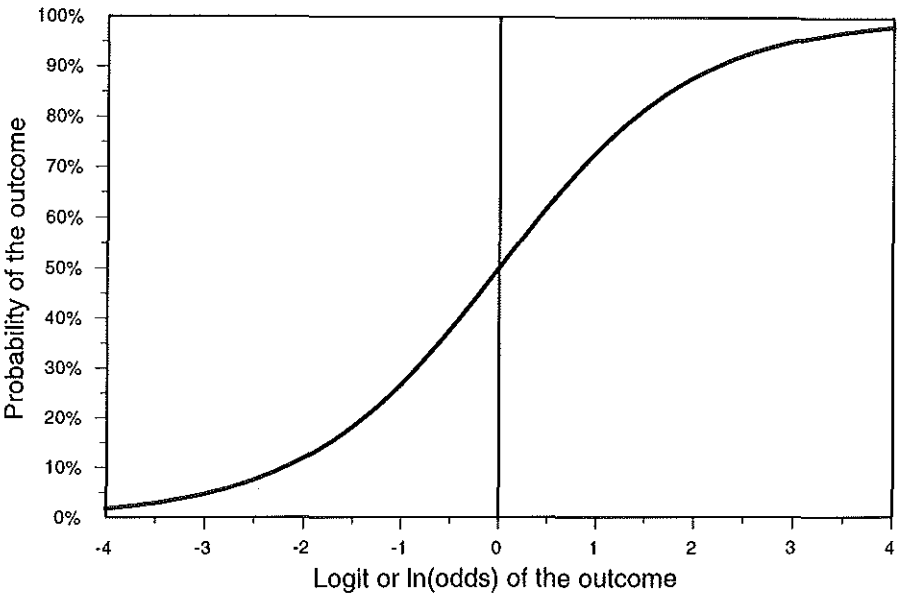

Figure 1 The relation between the logit transformation and the probability of the outcome (coded 0/1).

through the value 0.5 and then flattens to the value 1 with high values of the predictor (Figure 1).

This linearity assumption implies that a change of one unit on the continuous predictor has identical effects over the whole range of values of the predictor. This assumption can statistically be tested. A general method is the addition to the regression model of transformations of the continuous predictor. Simple transformations for a predictor like the age of a patient (in years), may include age^2, $\sqrt{}$(age), 1/age, ln(age), e^{age}. More general transformations can for example be formed with restricted cubic splines[5,6]. These functions describe the relation between a continuous predictor and the outcome with a flexible and smooth curve, while estimates of statistical significance of the non-linearity can readily be obtained with all standard computer software packages[7,9]. The statistical power of a such test for non-linearity will be strong, since the continuous character of the predictor is maintained. Another alternative is to add a categorized version of a continuous predictor to a model with the original continuous predictor already included to indicate non-linearity of the predictor.

Fourth, regression models make assumptions on the combination of predictors (additive or multiplicative). The regression models considered here are additive in their linear form[8]. In the case of logistic regression, the linear form is the logit of the outcome (Figure 1). When combining predictors, a prognostic index can be calculated. In additive models, the prognostic index is the summation of the coefficients multiplied by the values of the predictors. Confusingly, calculation with the Odds Ratios ($e^{coefficient}$)

thus involves a multiplication. For example, a patient with two independent prognostic characteristics, both with an OR of 3, has an OR of 9 ($3 \cdot 3 = e^{(\ln(3)+\ln(3))}$) compared to a patient without these characteristics. The additivity of predictors can statistically be tested by the evaluation of interaction terms between predictors[9].

2.2 Selection of variables

The selection of variables for prognostic models is a complex issue. Many potentially predictive patient characteristics may be available for a prognostic model (possibly 50-200), and it may seem both impractical and unnecessary to use all these available characteristics. Selection of a limited number of patient characteristics is related to the general scientific principle of parsimony: theories with simpler or easier explanations are considered more plausible than more complex theories. In statistical models this translates to the use of a limited number of variables. Further, there may be a concern that a model with many predictors leads to overfitting of the data and hence a poor performance of the model in future patients. Also, the interpretation of the regression coefficients in a model with a limited number of strong predictors may be easier than the interpretation of a model with many variables, in which some regression coefficients have a counterintuitive sign. Finally, selection of a limited number of variables leads to a higher precision of the predictions.

Strategies for selection of variables are discussed below. Special attention is given to stepwise selection.

2.2.1 Stepwise selection

Stepwise selection of variables is probably the most widely used selection strategy nowadays. Stepwise selection may be applied in a forward, a backward or combined backward-forward way. The usual significance level for selection of a variable in the model is 5%, which is identical to the significance level commonly used for hypothesis testing. The significance is usually calculated from the amount of variance (or related measures like the log-likelihood) explained by the variable, although measures based on the posterior probabilities may be more appropriate[10,11,12]. An extension of the stepwise selection strategies (forward, backward, combined backward-forward) is 'all possible subsets regression'. With this method, every possible combination of predictors is examined to find a best fitting model. The advantage of this method is that it may identify combinations of predictors not found by stepwise selection strategies, since all combinations are considered. This advantage holds especially against forward stepwise selection, where correlated variables may only appear prognostically important when considered together. In the following, 'all possible subsets' regression will be considered as a specific form of stepwise selection methods.

Advantages of stepwise selection methods are that they lead to a limited number of variables in a prognostic model, and that they are widely available in most standard statistical computer packages. The methods also nicely correspond to the concept that

once a limited number of predictors is included in the model, the remaining variables add nearly no additional prognostic information.

The fundamental problem of stepwise selection is however that the regression coefficients are biased to more extreme values. This bias is caused by the fact that coefficients that are -by chance- more extreme, are more likely to be selected than coefficients that are less extreme, since the more extreme coefficients are associated with lower p-values. This is illustrated in Figure 2 and 3. Figure 2 shows the distribution of the regression coefficient of a random, noise variable with mean zero and standard error 0.5. The value of 0.5 is the asymptotically calculated standard error in a sample with size 64, where the distributions of both the outcome variable and a dichotomous variable are optimal (50%:50%). The use of the standard significance level of 5% results by definition in a risk of 5% of falsely selecting the random variable as a predictor (*shaded areas*, alpha error).

Figure 3 shows the distribution of the regression coefficient of a predictor with mean 1 and the same standard error as the random variable in Figure 2 (0.5). The probability of selecting this predictor is 52% (*shaded area*). This probability is known as the power (or 1-beta error) in statistical test theory. In the context of clinical trials, statistical power is nowadays explicitly considered in sample size calculations before the start of the trial. A common requirement is that the power to detect an important treatment effect should exceed 80%. In studies of prognosis, sample size calculations are unfortunately rarely performed.

Figure 3 also shows the consequence of the selection of only about half of the regression coefficients of the predictor. The mean of the regression coefficients is 1.39 in the samples where the predictor was selected. This is equivalent to a bias of +0.39 in the estimated coefficient.

Table 2 shows the power (1-beta) of logistic regression analyses with several sample sizes and coefficients of a single predictor. A coefficient of 0.5 corresponds to an Odds Ratio of 1.65, 1.0 to 2.72, and the extreme coefficient of 2.0 to an OR of 7.4. The distributions of both the outcome variable and the dichotomous predictor were optimized (50%:50%). It appears that the power is low for small regression coefficients, even with a sample size of 256 patients. In contrast, a very large regression coefficient will almost always be found significant. Correspondingly, the bias in the coefficients of the selected predictors is large for a small coefficient and minimal with large coefficients. The power and the bias are thus favorably influenced by larger sample sizes. Note that medical studies may rarely have the optimal distribution used in Table 2 and that sample sizes are often less than 250 patients. The regression coefficients of interest may however be in the range of 0.4 to 1.1 (Odds Ratio 1.5 to 3). Therefore, the estimated regression coefficients will currently be biased substantially when the standard p-value of 5% is applied in stepwise selection procedures.

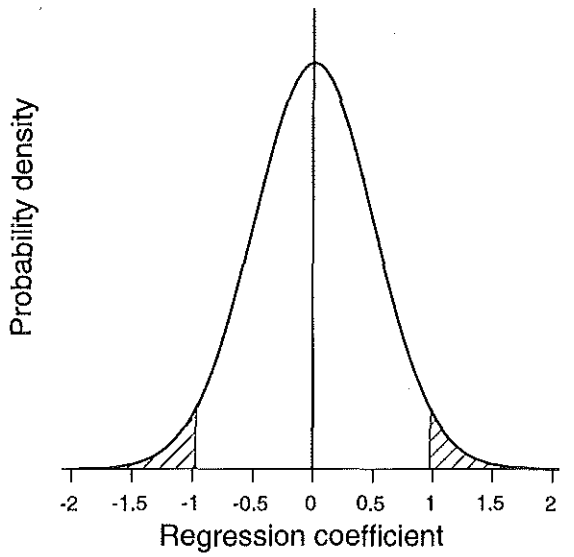

Figure 2 Effect of selection with a significance level of 5% on a random, non-predictive variable with a true regression coefficient of 0 and a standard error of 0.5. *Shaded area*: selected coefficients.

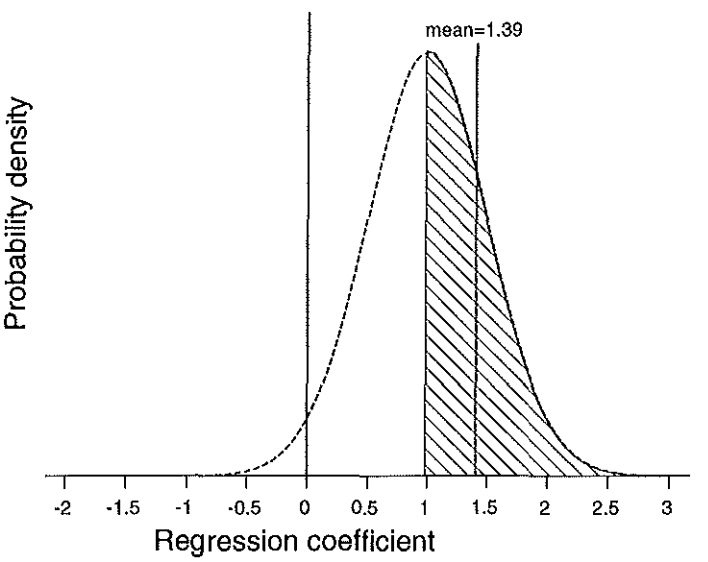

Figure 3 Effect of selection with a significance level of 5% on a predictor with a true regression coefficient of 1 and a standard error of 0.5. *Shaded area*: selected coefficients

In practice, the magnitude of the two errors alpha and beta will depend on the sample size and the distribution of both the predictor and the outcome variables, but also on the exact selection strategy applied (forward/backward, levels of p-values for entry/removal), and the number of important and unimportant predictors in the data set. It has been found that if many prognostically unimportant variables are present, the likelihood of selecting one of these is much higher than when only a minority of the variables is randomly associated with the outcome[13]. This finding is explained by the fact that multiple comparisons are made, which increases the overall risk of falsely selecting at least one unimportant variable to considerably higher levels than the nominal p-value[9].

Table 2 Power for selection of a predictor and resulting bias in the coefficients of the selected predictors. Logistic regression coefficients were 0.5, 1 and 2 and alfa was 5%. The standard errors of the coefficient were 0.50, 0.35, and 0.25, corresponding to sample sizes of N=64, N=128, and N=256.

		True regression coefficient		
	N =	0.5	1.0	2.0
Power	64	17%	52%	98%
	128	29%	81%	99.9%
	256	48%	98%	100%
Mean of selected coefficients	64	1.25	1.39	2.03
	128	0.92	1.12	2.00
	256	0.69	1.01	2.00
Bias	64	0.75	0.39	0.03
	128	0.42	0.12	0.00
	256	0.19	0.01	0.00

The above described phenomena are well-known in the theory of interim analyses of clinical trials[14]. At each interim analysis, a statistically significant effect may be observed in favor of a treatment, while this effect may have occurred by chance only. It would not have been observed at a later moment in time of the trial progress. Moreover, the treatment effect estimated at an interim analysis is biased to more extreme values[15].

Stepwise selection may use other significance levels than the conventional 5% level, since this 5% level is arbitrary and has no direct relation to clinical importance[16]. Alternative criteria might for example be 20% or 50%. With these more liberal significance levels, more variables will be included in the prognostic models. Among these additional variables will be some prognostically important variables, and relatively more noise variables which are randomly associated with the outcome[17]. The inclusion of noise variables causes a loss of precision in model predictions. On the other hand, the bias decreases with more liberal selection criteria. The choice of a particular significance level (.01/.05/.10/.20/.50/1.0) therefore poses a type of bias/precision dilemma[18]. In small studies, the use of the standard significance level of 5% leads to inclusion of a limited number of predictors, which will predict the outcome with a reasonable

precision. On the other hand, the bias will be relatively large and many important predictors will be excluded from the model because of a lack of power (Table 2). The balance between bias and precision might positively be influenced by a considerably more liberal significance level in smaller studies, for example 50%.

In summary, drawbacks of standard stepwise selection methods include the frequent inclusion of random, non-predictive variables, the exclusion of prognostically important variables and bias in the estimated regression coefficients. Moreover, the selection is unstable, in the sense that the addition of deletion of a few patients may change the set of predictors selected[19,20,21]. Therefore, one may doubt the validity of the resulting model: the variables selected, the order of selection, nor the p-values associated with the regression coefficients are reliable[16,21].

2.2.2 Alternative strategies

Alternative selection strategies have been proposed in the literature to diminish the above mentioned problems of automatic selection of variables. Generally these strategies aim at variable reduction with as limited as possible use of the regression results in the data set under study. An obvious step is the limitation of the number of potential predictors to be assessed[13]. This can for example be achieved by critically reviewing the plausibility of the prognostic importance of the candidate variables using clinical knowledge of the disease or related diseases. Other criteria may include the cost or reliability of the potential predictor and the number of missing values.

Variable reduction can also be based on a review of the findings in other studies. Such a review may take the formal statistical form of a meta-analysis ('analysis of analyses'). All predictors which are significant in the meta-analysis might then be selected for the prognostic model. This strategy has the advantage that the number of variables in the prognostic model will be limited, and that the estimates of the regression coefficients will not be biased by selection from the data under study. However, selection based on the literature may lead to inclusion of highly (positively) correlated variables, some of which will have no prognostic value as such. In the statistical literature it is well known that so called multi-collinearity of variables causes unstable estimates of the regression coefficients[22]. In predictive prognostic models we are however not primarily interested in the regression coefficients of individual variables, but in the prognostic performance of the whole model. Therefore, selection based on univariate literature data is only problematic if variables are really very strongly correlated, since in this case no additional prognostic information comes available once one of the positively correlated variables is included.

In contrast to positive correlation, some predictors may be correlated negatively. If two predictors, both with a positive association with the outcome, are correlated negatively, both predictors may appear as unimportant predictors in the literature, and therefore not be selected from a meta-analysis. Fortunately, such a negative correlation between predictors is less common than a positive correlation in medical data sets.

Further variable reduction may be achieved by clustering of variables in related groups[20,23]. Formal clustering techniques may be used to define the clusters, in

combination with medical knowledge. The prognostic weight of the variables should ideally be determined without using the data under study. Physicians may rate the prognostic importance in a few categories (e.g. 1, 2, or 3 points), or, even simpler, all variables in each cluster might be given equal prognostic weight. This relatively rough weighting has the advantage that the relations of each characteristic with the outcome are not used at this stage. On the other hand, this rough weighting may lead to underestimation of the weight of important variables and to overestimation of the weight of unimportant variables. It may therefore be tested if the weight (or regression coefficient) has to be markedly different from the initial weight for one or more of the variables in the cluster. In the regression analysis, the sumscores of each cluster are subsequently used as predictors.

A variation of this clustering strategy is to use principal components analysis to derive scores[24]. This technique summarizes the variance in the predictors, without taking into account their relation with the outcome. Empirical results with this technique have been favorable[21]. A disadvantage of principal components analysis is however that no clear relation can be identified between the patient characteristics and the outcome. Further, a hypothetical example can be formulated where the first principal component would fail to distinguish patients with and without the outcome.

2.2.3 Other aspects

Thus far, this discussion of selection strategies has focused on potentially predictive characteristics. If some of these characteristics are continuous variables, linearity has also to be assessed. Adjustment of the shape of a continuous variable so that it fits the data well again harbors the risk of a too favorable performance of the prognostic model on the data under study. Choosing an 'optimal' cut-off point for continuous variables may even be more dangerous. It has been shown that this practice can lead to an actual rate of falsely declaring a random variable significant of 40%, using the standard 5% significance level[25]. This again illustrates that multiple testing may lead to overoptimistic results. Further, interactions between predictors have to be considered. Assessment of such model assumptions is necessary, although it may contribute to the overoptimism of prognostic models. Some methods automatically incorporate interaction in the modeling strategy, such as classification and regression trees (CART)[26,27] or variations on this method[28,29]. These approaches also typically include the optimization of cut-off values for continuous variables.

2.3 Estimation of regression coefficients

Extensive and complicated theoretical statistical research has shown that regression models tend to overestimate the regression coefficients, even when a fixed set of predictors is used without selection on the same data[30,31]. This overestimation may be interpreted as related to regression to the mean[32]. The bias increases when the number of variables is larger or when the number of patients (or number of events) is relatively

small. It can even theoretically be shown that model performance decreases when many variables are added to a model, even if all variables have prognostic value[31,33].

For logistic regression analysis, the bias in the coefficients can be corrected by multiplication with the following heuristic shrinkage factor[31]:

$$c_{heur}=\frac{\chi^2_m-(k-1)}{\chi^2_m} \quad (1)$$

, where χ^2_m is the log likelihood of the model and k the number of variables in the model. This formula helps to correct the regression coefficients for overoptimism by shrinking them towards zero.

The correction of coefficients by equation 1 is applicable to a predetermined set of variables, which is fitted on the data without selection of variables based on the data. Interestingly, it seems also possible to correct the regression coefficients with the above shown formulas in the case of stepwise selection. The number of variables (k) is then based on the number of candidate variables for the regression equation rather than the number of finally selected variables[30,31]. It must be feared that many prognostic models would have a shrinkage factor near zero if this rule were applied, leaving no prognostic value for the model. Indeed, in situations without strong predictors, it has been suggested that a model without any predictors and only an intercept may perform better than a model with a few predictors selected from a large number of candidate variables[30].

Other methods have been developed to estimate the logistic regression coefficients in such a way that they lead to better predictions. For example, the principle of ridge regression has been applied to obtain biased but more stable parameters in the situation that the number of variables is relatively large compared to the number of patients[34]. Penalized likelihood methods have been applied in the modeling of ordinal categorical predictors[35]. Another method is outlined in Chapter 3, which combines data from the literature and from the data set under study to obtain more stable estimates of the multivariate regression coefficients.

2.4. Evaluation of prognostic models

Two aspects of the prognostic performance of logistic regression models for dichotomous outcomes are usually distinguished: reliability and discriminative ability. These aspects of model performance are discussed below.

2.4.1 Reliability

Reliability or calibration refers to the characteristic that the predicted probabilities correspond to the actual probabilities[36]. For example, if the predicted mortality is 10%, on average 10 out of every 100 patients are expected to die. Reliability may be assessed by goodness-of-fit tests. For logistic regression, the Hosmer-Lemeshow test is frequently used[37]. This test compares the observed and expected frequencies of the outcome in

groups, which are formed by pooling according to the predicted probabilities. Commonly, 10 groups are formed by deciles of the predicted probability. A chi-square test is then performed which gives a p-value that may indicate a good or poor fit of the model. Several graphical methods have also been proposed to examine the fit of a prognostic model. Kernel methods have been used[38] as well as purely non-parametric methods[36]. A drawback of these methods is that they are insensitive to differences within the pooled groups. To overcome this drawback, a goodness of fit test has been developed based on smoothing of the standardized residuals [39].

Simple overall checks on the goodness-of-fit can also be thought of for logistic regression models[40]. First, the average of the predicted probabilities should be equal to the observed average ('calibration in the large'). Also, the prognostic index can be used as a predictor, in which case the corresponding regression coefficient should be equal to 1. Logistic regression models fulfill both requirements by definition when evaluated on the same data set as the regression analysis was performed upon.

2.4.2 Discriminative ability

Discrimination refers to the ability to distinguish patients with and without the outcome from each other. Discriminative ability may generally be measured by an index of concordance[41]. In the case of logistic regression, the index of concordance corresponds to the area under the receiver operating characteristic (ROC) curve[42, 43,44]. The ROC methodology stems from psychophysics and has been applied in signal processing for radar detection. The ROC curve gained interest in the 1980's as summary measure of diagnostic test performance. In a ROC curve, the true-positive rate (or sensitivity) is plotted against the false-positive rate (or 1 minus specificity). The area under the ROC curve for sensible models varies between 0.5 (a useless model) and 1.0 (a perfect model). The statistical interpretation of the area is that it indicates the probability that for a randomly chosen pair of patients, one with the outcome and one without, the patient with the higher probability is the one with the outcome. This interpretation is not directly applicable to clinical situations, since it is unlikely that clinicians are ever asked to classify two patients in this way. The valuation of the magnitude of the area under the ROC curve depends on the clinical situation. In the comparison of different prognostic models, misleading conclusions on the superiority of a model can be drawn if the shapes of the ROC curves differ[45].

An interesting observation is that well discriminating models may be constructed in a simple and naive way, namely with prognostic models assuming conditional independence between the predictors[23,46]. In these independence models, the univariate Odds Ratios are simply multiplied with each other for multivariate prediction. The method may thus be seen as a simple application of Bayes theorema to calculate posterior probabilities. Because of the naiveness it is sometimes labeled "Idiot's Bayes". It has however been shown that the conditional independence is a sufficient, but not a necessary condition for validity of the independence model, which may explain its sometimes respectable performance[47]. The method may be useful in situations where only literature data are available to construct a discriminative model. An example of such

an application is the selection of ankle trauma patients for X-ray evaluation[48] or the detection of normal pressure hydrocephalus in demented patients[49]. An improvement of the method is the use of an 'overall association factor' to correct for dependence between the predictors[50]. Of course, this factor is less refined than the correction for interdependence obtained with logistic regression analysis.

2.4.3 Summary measures

Summary measures have been developed for predictive performance, such as the mean squared error: $n^{-1} \Sigma (x_i - p_i)^2$, which can be decomposed into aspects of reliability and discrimination[51]. An alternative error measure is the mean minus log-likelihood error: $n^{-1} \Sigma (Y_i \log(p_i) + (1-Y_i)\log(1-p_i))$[31]. This measure is closely related to the likelihood function used to estimate the logistic regression coefficients. Although these measures have theoretical value, especially when comparing alternative models developed on the same data set, they have not been used often in the medical field.

2.4.4 Illustration: peptic ulcer

The effects of overoptimism in the selection of variables and the estimation of the regression coefficients on model performance have been illustrated on medical data. In an example of the prediction of peptic ulcer, four modeling strategies were followed in a random sample of 117 patients with dyspepsia, of whom 41 had a peptic ulcer[46]. Thirteen variables were available which were judged reasonable predictors from a clinical point of view. The strategies were stepwise selection with the traditional regression coefficients, selection of all potential predictors with either the univariate regression coefficients (conditional independence model), with the traditional multivariate logistic regression coefficients, or with shrunk multivariate logistic regression coefficients using the formula of Copas[30]. These four strategies were evaluated in a test set of 993 patients (411 with ulcers). It appeared that the discriminative ability of a model with stepwise selected predictors discriminated worse than the models with all 13 potential predictors (area under the ROC 0.71 versus 0.79-0.81). Calibration was best for the model with shrunk coefficients, less for the traditional multivariate coefficients in the full model, and worst for the model with univariate regression coefficients. Hence, inclusion of all potential predictors and shrinkage of the regression coefficients was the best strategy in this study.

2.4.5 Illustration: residual mass histology

We further investigated these prognostic modeling strategies in a database of testicular cancer patients, where six binary predictors were related to the histology found at resection (necrosis vs other histology). These six predictors were found as significant predictors in a meta-analysis of the literature (Chapter 5). As a selection strategy, we compared stepwise forward selection with a p-value less than 5% to selection of all six predictors without consideration of the statistical significance of the individual predictors. For the regression coefficients, we compared the traditional regression coefficients to coefficients shrinked with equation 1 and the simple univariate

coefficients. Shrinkage of the coefficients in the stepwise models was performed with the number of selected variables in the equation (k = # selected) and with the number of candidate variables (k = 6). These strategies were followed in each of five research groups which participated in this study (1 from New York, USA, 1 from Oslo, Norway and 3 from The Netherlands). Patients from a sixth study group (Indiana, USA) were excluded because of a large number of missing values in one predictor (LDH elevated) and zero cells for another predictor (residual mass size < 20mm).

The results obtained in each study were evaluated in the complete data set (including all studies). The column 'total' is included in Table 3 as a reference for the model performance with availability of the total data set. The other columns (1-5) indicate the results when each study would be the only source of data available. Stepwise selection of variables in the complete data set led to inclusion of all six predictors. In the individual studies, 2 to 4 variables were selected.

The overall goodness-of-fit was evaluated with the regression coefficient of the prognostic index in the total data set of 502 patients. In the models based on the complete data set this coefficient was by definition equal to 1 for the traditional coefficients. A value lower than 1 indicates overoptimism of the regression coefficients estimated in the studies. This overoptimism may be caused by overestimation of the regression coefficients and by selection of variables with a relatively large effect, if selection of variables was performed. The traditional multivariate coefficients, as estimated in the studies, appear to be overoptimistic to a considerable extent. The coefficients of the prognostic index vary between 0.37 and 0.91, which corresponds to an average overestimation of 10% to 270%. Remarkably, the overoptimism is similar or even larger when stepwise selection is followed, compared to when the fixed set of 6 predictors is selected. If the coefficients are shrunk with formula 1, the overoptimism decreases. Study 4, which contained only 33 patients, still suffers from considerable overoptimism after shrinkage of the coefficients. For the stepwise models, shrinkage with the number of candidate variables as k in equation 1 performed generally better than the use of the number of selected variables in the equation. These findings thus support the suggestion that the number of candidate variables should be used in the shrinkage formula (equation 1)[31]. Finally, we found that the overoptimism in the coefficients is the largest if the univariate coefficients are used in the model.

We also evaluated the discriminative ability of the fixed and the stepwise models. The area under the ROC curve was lower for all stepwise models in the individual studies. The inclusion of more prognostically important variables greatly improved the discriminative performance of the models. These results confirm the conclusion of the previously mentioned study[46] (*section 2.4.4*), i.e. that the performance of stepwise models was less than the performance of models including all potential predictors with shrinkage of the regression coefficients.

Table 3 Performance of different selection strategies and estimation procedures for logistic regression models in testicular cancer prediction study.

			Study				
		Total N=502	1 N=121	2 N=127	3 N=137	4 N=33	5 N=84
Number of variables							
Six predictors		6	6	6	6	6	6
Stepwise		6	4	3	4	2	2
Shrinkage factor (equation 1)							
Six predictors:	k=6	.97	.88	.89	.90	.76	.75
Stepwise:	k=# selected	.97	.92	.95	.94	.94	.91
	k=6	.97	.87	.88	.90	.69	.55
Reliability: coefficient of PI in total data set							
Six predictors:	unshrunk	1.0	.91	.82	.74	.37	.80
	shrunk k=6	1.03	1.04	.92	.82	.48	1.07
	univariate	.72	.71	.58	.78	.48	.73
Stepwise:	unshrunk	1.0	.83	.76	.76	.38	.89
	shrunk k=#sel	1.03	.90	.80	.81	.41	.98
	shrunk k=6	1.03	.96	.86	.85	.55	1.61
Discriminative ability: ROC area in total data set							
Six predictors:	unshrunk/shrunk	.80	.80	.79	.78	.75	.80
	univariate	.80	.80	.78	.80	.79	.78
Stepwise	unshrunk/shrunk	.80	.76	.75	.76	.69	.72

2.5 Validation and re-sampling methods

As indicated before, prognostic modeling involves several phases where overoptimistic results are obtained if the same data set is used. These phases include the selection of variables, the estimation of the regression coefficients and the evaluation of model performance. The biases in each phase may be labeled selection bias, estimation bias, and evaluation bias, respectively. Note that both selection bias and estimation bias lead to overestimated regression coefficients.

2.5.1 Independent data sets

To eliminate these biases, it has been suggested that for each of the three phases an independent data set might be used[31]. These three data sets might be obtained by random selection from the original data set ('split-sample approach'). The first data set might be labeled the selection sample, the second the estimation sample and the third the validation sample.

A simplification of this division in three samples is quite common nowadays, and includes separation in two samples: a development or training sample and a validation or test sample[52,53,54]. Both selection of variables and estimation of the coefficients are performed on the development sample. The validation sample can only be kept apart if the original data set was relatively large. Another drawback is that the randomly selected validation sample may show poor results of the model by chance only ('bad luck'). The analyst may in such a case be tempted to repeat the random selection of the

validation sample until more favorable results can be shown. Further, the final prognostic model is usually not based on all information available. The prognostic model is merely checked on the validation data set, without using these data to improve the final model.

2.5.2 *Cross-validation*

A more efficient use of the data can be made with cross-validation methods. Part of the data is used for analysis and the results are evaluated on the other part. Examples of cross-validation are the split-half or the split-quarter method. With the split-quarter method, 25% of the data set is used to evaluate a modeling phase performed on 75% of the data set. This is repeated four times and gives an impression of the validity of that modeling phase. The most extreme cross-validation method is to leave out a single patient at a time for evaluation of a modeling phase which was performed on the remaining N-1 patients ('jack-knife method'). This procedure is repeated for all patients (N times, resulting in N models).

2.5.3 *Bootstrap re-sampling*

An increasingly popular model evaluation method is the bootstrap re-sampling procedure[55]. This procedure was developed by Efron and was originally presented as an extension of the jack-knife method[56,57]. The principle of bootstrapping is that random samples are drawn with replacement from the total data set. These samples are labeled bootstrap samples and may contain each patient 0, 1, 2, 3, ..., N times. These bootstrap samples have a structure similar to the original data set. If many bootstrap samples are drawn, the underlying structure of the population where the data were drawn from is revealed, without using new data from this population. The more bootstrap samples are drawn, the more stable become the estimates based on them. A minimum amount is 100 replications for most applications[57]. Computer time may, even nowadays, be a limiting factor for very high numbers of replications. The bootstrap is generally slightly more efficient than the jack-knife method, as each bootstrap sample contains N patients, compared to N-1 in the jack-knife samples. On the other hand, in small data sets ($N<100$), the jack-knife requires less computer time to obtain stable estimates.

The bootstrap re-sampling method can be used for selection of variables. As noted before, stepwise selection procedures produce unstable results, in the sense that the addition or deletion of a few patients may change the set of predictors selected. This variability can excellently be illustrated with the bootstrapping technique[32]. The most important prognostic variables should however be included in most bootstrap samples, and the frequency of inclusion can thus be a criterion for selection[19,20,58]. In this way, selection of variables can be made less dependent on idiosyncracies of the original data set. On the other hand, the selection strategy still is stepwise and still uses information from the data set under study.

The bootstrap has also been proposed to correct the regression coefficients for overoptimism. The procedure is essentially a calibration procedure[31]. Bootstrap samples are drawn from the original data set. A prognostic model is fit on each sample, and the

prognostic index for each patient in the sample is calculated. This prognostic index is then evaluated as the only independent variable in the original data set, resulting in a regression coefficient which is the shrinkage factor for that particular bootstrap sample. Averaging over a large number of replications yields an estimate of the shrinkage factor for the original model. This estimate will in most instances be very similar to the heuristic estimate shown in equation 1.

Finally, the bootstrap can be used to indicate the model performance in future patients. Especially, discriminative ability can be evaluated easily, as discrimination only depends on the ordering of the patients according to the prognostic index, and is thus insensitive to values of the shrinkage factor. After fitting a fixed model in a bootstrap sample, evaluation bias may be quantified as the difference between the performance of that model (with the regression coefficients estimated from the bootstrap sample) in the original data set and in the bootstrap sample[32]. If the prognostic model was developed with selection of variables using the same data set, selection bias should also be assessed. This can be achieved by applying the selection strategy in each bootstrap sample.

In conclusion, the bootstrap method provides a useful tool for model development and validation. The method should become a standard procedure in prognostic modeling, especially since modern computer facilities allow for an acceptable calculation time. The bootstrap may especially help to reduce the bias in the estimated regression coefficients, and to give an impression of the discriminative ability in future patients. The bootstrap evaluation may however still underestimate the total overoptimism, if data driven decisions were taken by the analyst in a way that cannot be simulated.

2.5.4 Internal and external validation

It should be realized that all validation methods that are based on (part of) the original data set may only give an impression of *internal* validity. Internal validity of a prognostic model refers to the prognostic value in the same type of patients as in the data set analyzed. Internal validity may be contrasted to external validity or generalizability, which refers to the prognostic value in patients that may be slightly different than the patients studied. For example, patients in other centers may have prognostically relevant differences that are not taken into account in the model, or the definition of prognostic variables may be different. The results of a model based on data from several centers may therefore have more widespread validity than a singlecenter study. Also, a multicenter study provides the possibility to assess external validity, by leaving out one of the centers and evaluating the performance of the model built on the other centers on the data from this center (see e.g. Chapter 6). More commonly, an external validation study is performed totally independent from the modeling stage. Practical guidelines for statistical evaluation of such studies have been described [59]. External validity of existing prognostic models should gain more emphasis as a research goal, as the prognostic performance of a model when applied by an outsider in a different clinical environment is the ultimate yardstick for a prognostic model.

2.6 Presentation of prognostic models

A prognostic model may aim to support decision making in clinical practice directly. Clinical prediction rules, for example, may aim to help physicians to identify patients who require diagnostic tests, treatment, or hospitalization[60]. In this situation, the applicability of the prognostic model to the clinician's specific patient population must be evaluable. This means that general information needs to be available on the setting and the patient population where the model was developed, and that outcome and predictive characteristics are clearly defined. Further, an indication of model performance is required and a description of the mathematical technique used to develop the model[60]. The presentation of the model itself is another point of interest.

2.6.1 Model presentation

The results of prognostic models are often presented just as the regression equations obtained from the statistical package used to estimate the equation. This presentation may not be suitable for practical application, especially if variables were transformed (quadratic, logarithmic, inverse transformation, spline transformations[9]) or if the predicted probability can only be calculated with transformations, like exponentials as in logistic regression. A regression equation should therefore not be the only presentation of the prognostic model. When presented in addition to a more practical representation, the regression formula may aid those clinicians favoring the use of computers over paper and pencil, since a regression formula can easily be implemented in a simple computer application, like a spreadsheet, or a pocket calculator.

More practical presentations of the prognostic model include the construction of a table with the predicted probabilities for all combinations of the predictors (see for example Chapter 8). This presentation can however only be realized if the number of predictors is limited and if no continuous predictors are involved. If the latter is the case, an alternative presentation is as a prognostic score chart (see for example Chapter 6 and 9). A score chart lists the prognostic variables, their possible values, and their corresponding scores in the prognostic model (rounded to whole integers). The relevant scores are added in a sum score (the prognostic index) and the corresponding probability may be read from a table or graph. This two step process is considerably easier to perform in clinical practice than the application of a formula. Moreover, this calculation of the prognostic index gives a lucid insight in the quantitative weight of the predictors involved. The prognostic index thus provides a attractive summary statement on the predicted outcome[61]. To facilitate the interpretation of the scores in a score chart, it has been proposed to use the ^{2}log scale ($10 \bullet {}^2\log$ of the regression coefficients) rather than the ^{e}log (=ln) form[62]. The advantage of this scale is that a doubling of the odds is indicated with a score of 10 points.

2.6.2 Risk groups

Frequently, the results of prognostic models are grouped into risk groups (e.g. good, intermediate, and poor risk). This grouping may facilitate the practical use of the model, especially when the grouping is linked to the type of therapy. Further, the average

prediction of a larger prognostic group will be more precise than the predictions in the original, smaller, subgroups. Two objections can be formulated against this grouping. First, patients at the borders of the risk groups may be candidates for treatment as in the adjacent group because of center-specific or individual circumstances. These 'threshold' patients can be identified with a continuous predicted probability, where the threshold for treatment is not superimposed by the analyst. Second, the aim to develop simple risk groups based on only a few predictors may negatively influence the modeling process. Prognostically important variables may be left out of the model for the sake of simplicity. For example, the selection criterium for variables has in some applications been lowered to $p<0.01$ to limit the number of predictors in the model[63]. This must be considered a waste of prognostic information.

2.6.3 Prior probability

Another aspect of prognostic models is that they may require an estimate of the average probability of an event, before the prognostic characteristics are considered ('prior probability')[64]. In some clinical problems, prognosis may differ considerably by patient-independent characteristics. For example, mortality of elective abdominal aneurysm surgery will be affected by the surgeon's skill and the availability of technical facilities (see e.g. Chapter 9). Such influences need to be incorporated in prognostic models in addition to patient-dependent characteristics. To estimate the patient-independent prognostic component in an average probability, the observed higher or lower prior probability has to be corrected for the prevalence of risk factors. When the prevalence of risk factors is similar between the model development and validation environment, a simple adaptation of the average probability may improve the calibration of the predictions[65,66,67].

Assessment of the patient-independent component may even be the aim of the analysis, such as in the comparison of the quality of different institutions. It is clear that this quality can only be properly ranked after correction for prognostic variables [52,53,68], although this correction may be difficult in practice[69].

2.7 Practical considerations and conclusions

As discussed in this chapter, prognostic models may show overoptimistic results, when all three phases of modeling are performed on the same data set. The regression coefficients will be too extreme because of the selection strategies applied, and because of bias in the estimation method. Model validation will show overoptimistic results if the same data are used again.

2.7.1 Overoptimism

The combined effect of selection and estimation bias leads to too large regression coefficients and hence a overoptimistic expectation on model performance. It has been noted that this overoptimism or 'overfitting' can be detected by a large variability in a regression coefficient[70]. It is however evident that precisely estimated regression

coefficients may also be overfitted, especially if the variables were selected from a large number of candidate variables. On the other hand, if no selection of variables was performed at all, the regression coefficients may be imprecise, but will show only estimation bias which can largely be corrected with a shrinkage procedure.

In current practice, selection bias is probably the most important cause of overfitting. As shown before, the common attempts to limit overfitting by stepwise selection of predictors may merely aggreviate the problem instead of solving it.

The magnitude of the overoptimism in logistic regression models will depend on the statistical strength of the data set relative to the number of predictors. For statistical strength, the variance around the average of the outcome might be considered a proper measure. As a practical rule it has been suggested that the number of candidate variables should be less than 1/10 of the number of events[9].

2.7.2 *Large data sets*

If a limited number of predictors (e.g. <20) is used in a very large data set is (e.g. N>5000) with a considerable number of events (e.g. >1000), the overoptimism will be very limited (see for example[52,71]). If stepwise selection is applied, it leads to the inclusion of all variables with a substantial prognostic effect. Probably, these variables will have small p-values (<0.001). Most variables with larger p-values will have a modest prognostic importance (small Odds Ratios). The estimation bias in the regression coefficients will be very small (resulting in a shrinkage factor somewhere between 0.990 and 0.999). In such large data sets, a validation sample is often kept apart to evaluate the results of the modeling procedure. This validation sample will however show practically identical results as evaluation on the development sample, since the selection bias and estimation bias are limited. As noted before, this validation sample only indicates internal validity. External validity in slightly different patients outside the sample is not yet assessed in this way.

It might be concluded that the best way to develop a valid prognostic model is to gather an enormous amount of high-quality data. The same statement has been formulated in the context of clinical trials, where 'mega-trials' are set up to answer simple questions in a simple manner, but with a very large number of patients[72,73]. In the medical setting, data acquisition on the individual patient level with sufficient quality control is however expensive. Also, prognostic models for rare diseases cannot be developed in thousands of patients. A promising approach is to perform prognostic meta-analyses with individual patient data from several centers to obtain a sufficiently high number of patients[16]. Chapters 6 and 8 are illustrations of this approach.

2.7.3 *Small data sets*

The main difficulties for the development of prognostic models exist in relatively small data sets. Selection of variables is the key problem here, as estimation bias and evaluation bias can be corrected for relatively easily, for example with bootstrapping techniques. It should be avoided as much as possible that selection is based on the same data as the regression coefficients are estimated upon.

A relatively small data set may be analyzed while literature data on the relevant prognostic variables can be found. A proposed modeling strategy is shown in Table 4. In this situation, both selection of variables and estimation of the regression coefficients may benefit from explicit analysis of the literature data. Chapter 3 discusses this method in detail. The literature may further provide previously developed prognostic models, which may be tested on the data set under study (external validation). Modification of the regression coefficients and inclusion of additional variables might be considered as a next step.

Table 4 Schematic overview of a proposed modeling strategy in the presence of literature data.

Modeling phase	Method	Data set
Selection of variables		
Classification of variables	Published classifications	Literature
Inclusion of variables	Meta-analysis	Literature
Inclusion of interaction terms	Multivariate analysis	Study
Estimation of regression coefficients	Adaptation of coefficients	Lit. + study
Evaluation of model performance	Discriminative ability and goodness-of-fit	Study *
Presentation of model	Table or prognostic score chart	-

* The study data can be used to evaluate a model with regression coefficients based on the same data; if the regression coefficients were based on the *combination* of the study data and the literature data, the study data are less suitable for evaluation.

Another situation occurs when no supportive literature data are available. The first option in this situation is to refrain from prognostic modeling. The data set may then be used for exploratory analyses about the relations between predictors and the outcome, focusing on univariate analyses. If a prognostic model is desired, Table 5 shows a possible strategy in a small data set, without empirical evidence from other studies.

Table 5 Schematic overview of proposed modeling strategies in the absence of literature data.

Modeling phase	Method
Selection of variables	
Classification of variables	Univariate analysis; conservative attitude
Inclusion of variables	Clinical knowledge; clustering of related variables; backward stepwise selection with $p<.50$ for inclusion
Inclusion of interaction terms	Multivariate analysis
Estimation of regression coefficients	Multivariate analysis with shrinkage
Evaluation of model performance	Discriminative ability and goodness-of-fit
Presentation of model	Table or prognostic score chart

Selection of variables should not be based on automatic stepwise procedures with low p-values for inclusion of variables. This would result in severe overoptimism. If a higher p-value for inclusion is chosen, the selection bias decreases. The extreme is to use a p-value of 1, which means that all variables are included and selection bias is minimal. This extreme is only reasonable once the number of potential predictors is limited, either by selection with clinical knowledge or by the formation of clusters of related variables. If the number of candidate variables remains too large to be all included in the prognostic model, backward stepwise selection with a p-value of 50% may be sensible. This p-value is arbitrary, but may be a good compromise between bias reduction (higher p-values) and precision (lower p-values). Bootstrapping techniques should be applied in all three modeling phases (selection of variables, estimation of regression coefficients, evaluation of model performance).

2.7.4 Final model and presentation

In data sets of any size, the final prognostic model should be based on all information available. This means that if a model is tested in a validation sample, the final model should be re-estimated on the total of the development and validation sample. The final regression coefficients should be shrunk towards zero to correct for overoptimism of the estimation procedure. Attention should be given to a practical presentation of the prognostic model. With these guidelines, prognostic models may become less overoptimistic and may more validly support clinical decision making.

References

1. Cornfield J. Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. Feder Proc 1962; 21: 58-61
2. Cox DR. The analysis of binary data. London: Methuen, 1970
3. Frome EL, Checkoway H. Use of Poisson regression models in estimating incidence rates and ratios. Am J Epidem 1985; 121: 309-323
4. Faraway JJ. On the cost of data analysis. Am Stat Ass 1992; 1: 213-229
5. Smith PL. Splines as a useful and convenient statistical tool. Am Stat 1979; 33: 61-83
6. Stone CJ, Koo CY. Additive splines in statistics. Proc Statist Computing Sect ASA 1985: 45-48
7. Devlin TF, Weeks BJ. Spline functions for logistic regression modeling. In: Proceedings of th e11th annual SAS Users Group International Conference. Cary, NC: SAS Institute, Inc., 1986: 646-651
8. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quatitative methods. Lifetime learning publications, London, 1982: pp 403-418
9. Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep 1985; 69: 1071-1077
10. Habbema JDF, Hermans J. Selection of variables in discriminant analysis by F-statistic and error rate. Technometrics 1977; 19: 487-493
11. Hilden J, Habbema JDF, Bjerregard B. The measurement of performance in probabilistic diagnosis. III. Methods based on continuous functions of the diagnostic probabilities. Meth Inform Med 1978; 17: 238-246

12. Habbema JDF, Gelpke H. A computer program for selection of variables in diagnostic and prognostic problems. Comp Progr Biomed 1981; 13: 251-270
13. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset algorithms: frequency of obtaining authentic and noise variables. Br J Math Stat Psych 1992: 42, 265-282
14. Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: a review and commentary. Stat Science 1990; 5: 299-317
15. Simon R. Some practical aspects of the interim monitoring of clinical trials. Stat Med 1994; 13: 1401-1409
16. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994; 69: 979-985
17. Flack VF, Chang PC. Frequency of selecting noise variables in subset regression analysis: a simulation study. Am Stat 1987; 41: 84-86
18. Hand DJ. Statistical methods in diagnosis. Stat Meth Med Res 1992; 1: 49-67
19. Chen CH, George SL. The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. Stat Med 1985; 4: 39-46
20. Altman DG, Andersen PK. Booststrap investigation of the stability of the Cox regression model. Stat Med 1989; 8: 771-783
21. Harrell FE, Lee K, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984; 3: 143-152
22. Kleinbaum DG, Kupper LL, Muller KE. Applied regression analysis and other multivariable methods. PWS-Kent Publishing Company, second edition, 1988: pp 206-218
23. Titterington DM, Murray GD, Murray LS, Spiegelhalter DJ, Skene AM, Habbema JDF, Gelpke GJ. Comparison of discrimination techniques applied to a complex data set of head injured patients. J Roy Stat Soc, Series A, 1981: 144: 145-175
24. Marquardt DW, Snee RD. Ridge regression in practice. Am Stat 1975; 29: 3
25. Altman DG, Lausen B, Sauerbrei W, Schumacher M. The dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. Comment in: J Natl Cancer Inst 1994; 86: 1798-1799
26. Breiman L, Friedmand L, Olshen R, Stone CJ. Classification and regression trees. Belmont, CA: Wadsworth, 1984
27. Goldman L, Weinberg M, Weisberg M, et al. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. N Engl J Med 1982; 307: 588-596
28. Knottnerus JA. Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables. Med Decis Making 1992; 12: 93-108
29. Aitchison TC, Sirel JM, Watt DC, MacKie RM. Prognostic trees to aid prognosis in patients with cutaneous malignant melanoma. BMJ 1995; 311: 1536-1539
30. Copas JB. Regression, prediction and shrinkage (with discussion). J Roy Stat Soc, Ser B, 1983; 45: 311-354
31. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990; 9: 1303-1325
32. Harrell F, Lee,K, Mark D. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996; 15: 361-387
33. Breiman L, Freedman D. How many variables should be entered in a regression equation? J Am Stat Assoc 1983; 78: 131-136
34. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. Appl Statistics 1992; 41: 191-201.

35. Verweij PJM, Van Houwelingen JC. Penalized likelihood in Cox regression. Stat Med 1994; 13: 2427-2436

36. Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. Meth Inform Med, 1978; 17: 227-237

37. Hosmer DW, Lemeshow S. Applied logistic regression. New York, NY: John Wiley & Sons Inc, 1989, pp 140-145

38. Copas JB. Plotting p against x. Applied Statistics 1983; 32: 25-31

39. Le Cessie S, Van Houwelingen JC. Building logistic models by means of a non parametric goodness of fit test: a case study. Statistica Neerlandica 1993; 47: 97-109

40. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Stat Med 1991; 10: 1213-1226

41. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA, 1982; 247: 2543-2546

42. Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol 1979; 14: 109-121

43. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8: 283-298

44. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology 1982; 143: 29-36

45. Hilden J. The area under the ROC curve and its competitors. MDM 1991; 11: 95-101

46. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986; 5: 421-433

47. Hilden J. Statistical diagnosis based on conditional independence does not require it. Comput Biol Med 1984; 14: 429-435

48. Zwetsloot-Schonk JH, Verhoef W, Kievit J, Van Dam W. On the use of a hospital information system in evaluating clinical care: a case report. Med Inf 1993; 18: 243-254

49. Dippel DWJ, Habbema JDF. Probabilistic diagnosis of normal pressure hydrocephalus and other treatable cerebral lesions in dementia. J Neurol Sci 1995; 119: 123-133

50. Hilden J, Bjerregaard B. Computer aided diagnosis and the atypical case. In: FT de Dombal and F Grémy, eds: Decision making and medical care: can information science help? North Holland, Amsterdam, 1976

51. Yates JF. External correspondence: decomposition of the mean probability score. Organ Bahav Human Perf 1982; 30: 132-156

52. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993; 270: 2478-2486

53. O'Connor GT, Plume SK, Olmstead EL, et al. Multivariate prediction of in-hospital mortality associated with coronary artery bypass graft surgery. Circulation 1992; 85: 2110-2118

54. Marcantonio ER, Goldman L, Mangione CM. A clinical prediction rule for delirium after elective noncardiac surgery. JAMA 1994; 271: 134-139

55. Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman & Hall Inc., London, 1993

56. Efron B. Bootstrap methods: another look at the jackknife. Ann Stat 1979; 7: 1-26

57. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 1983; 78: 316-331

58. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. Stat Med 1992; 11: 2093-2109

59. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. MDM 1993; 13: 49-58

60. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules: applications and methodological standards. N Engl J Med 1985; 313: 793-799

61. Fielding LP, Fenoglio-Preiser CM, Freedman LS. The future of prognostic factors in outcome prediction for patients with cancer. Cancer 1992; 70; 2367-2377

62. Habbema JDF. Interperation and calibration of predictive clinical scoring rules. Commenatary in: Theor Surg 1990; 6: 14-18

63. De Wit R, Stoter G, Sylvester R. Prognostic factors in disseminated non-seminomatous testicular cancer. Advances in the Biosciences: germ cell tumours III, Jones WG, Harden P, Appleyard I (eds), 1994: pp 237-238

64. Sox HC, Hickam DH, Marton KI, et al. Using the patient's history to estimate the probability of coronary artery disease: a comparison of primary care and referral practices. Am J Med 1990; 89: 7-14.

65. Poses RM, Cebul RD, Collins M, Fager SS. The importance of disease prevalence in transporting clinical prediction rules. Ann Int Med 1986; 105: 586-591

66. Wigton RS, Connor JL, Centor RM. Transportability of a decision rule for the diagnosis of streptococcal pharyngitis. Arch Intern Med 1986; 146: 81-83

67. Segaar RW, Wilson JHP, Habbema JDF, Malchow-Møller A, Hilden J, Maas PJ van der. Transferring a diagnostic decision aid for jaundice. Neth J Med 1988; 33: 5-15

68. Brand R, Van Hemel DJ, Elferink-Stinkens PM, Verloove-Vanhorick SP. Comparing mortality and mortality in hospitals: theory and practice of quality assessment in peer review. Methods Inf Med 1994; 33: 196-204

69. Iezzoni LI, Ash AS, Schwartz M, Daley J, Hughes JS, Mackiernan YD. Predicting who dies depends on how severity is measured: implications for evaluating patient outcomes. Ann Intern Med 1995; 123: 763-770

70. Concato JC, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Int Med 1993; 118: 201-210

71. Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study. JAMA 1993; 270: 2957-2963

72. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomised trials? Stat Med 1984; 3: 409-420

73. Woods KL. Mega-trials and management of acute myocardial infarction. Lancet 1995; 346: 611-614

3 Prognostic models based on individual patient data and literature data in logistic regression analysis

E.W. Steyerberg, M.J.C. Eijkemans, J.D.F. Habbema.
Partly presented in Theor Surg 1994; 9: 192 (abstract)

3.1 Introduction

Quantitative approaches to clinical decision making are often based on prognostic estimates for different patient profiles, which are defined by two or more crucial clinical characteristics. The prognostic estimates may be obtained by regression analysis of a data set with individual patient data. For dichotomous outcomes (yes/no), logistic regression models are often used to provide probability estimates of the outcome in relation to a number of patient characteristics. Such a multivariate logistic regression analysis can be applied to a data set from a single center, but also to a larger but possibly more heterogeneous data set containing patient data from several centers. The larger number of patients is advantageous for prognostic modeling: the statistical power for the selection of predictive patient characteristics is higher since the variability of the regression coefficients is smaller. Also, the generalizability may be better for a prognostic model based on data from several centers.

In practice, several papers may have been published in the literature on the relation between patient characteristics and the outcome of interest for a particular clinical problem. If the published papers describe comparable patient series, a meta-analysis may combine the available evidence quantitatively. The information in these papers is usually only sufficient to calculate a univariate regression coefficient for each of the patient characteristics. Multivariate coefficients can only be estimated if individual patient data are available from the published series. This may not be feasible especially for papers published several years ago. Furthermore, this requires a substantial research effort.

In this chapter, we aim to estimate the multivariate logistic regression coefficients as would be obtained in the literature data, in case no individual patient data are available for the published studies. These coefficients are traditionally estimated by the multivariate coefficients in the individual patient data, while the univariate data from the literature are ignored. We propose to estimate the multivariate coefficients by adapted univariate coefficients from the literature with an adaptation factor. This adaptation factor is calculated by comparing the univariate and multivariate regression analysis in the individual patient data. Figure 1 schematically shows the data sources and analysis techniques involved.

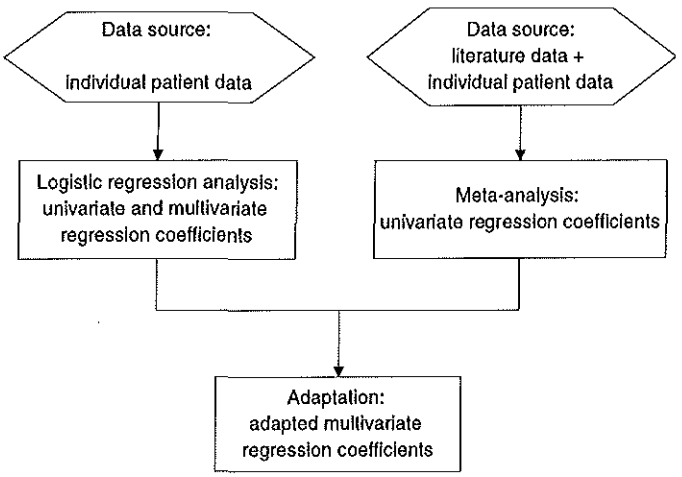

Figure 1 Schematic overview of the adaptation method to combine the regression results from a data set with individual patient data with those from a meta-analysis. The meta-analysis includes data from published series and from the data set with individual patient data.

3.2 Methods

We address the situation that a number of patient characteristics is considered as a set of covariates in a logistic regression analysis with a dichotomous outcome. Two sources of data are available: individual patient data and literature data. The univariate and multivariate regression coefficients are denoted as ßUNI and ßMULT. Note that the logistic regression coefficient is equivalent to the natural logarithm of the Odds Ratio (ln(OR)). For the individual patient data these coefficients can both be calculated and are denoted as ßUNI|IND and ßMULT|IND. For the literature data, only the univariate coefficients can directly be estimated (ßUNI|LIT).

We assume that the individual patient data and the studies forming the literature data are both random samples from a common underlying patient population. Therefore, the individual patient data may also be considered as a random subsample from the literature data. Under this assumption, the multivariate coefficient in the individual patient data (ßMULT|IND) is an asymptotically unbiased estimator of the multivariate coefficient in the literature data (ßMULT|LIT). This is the traditional calculation. The coefficient may however be estimated more efficiently (i.e. with less variance) by combining the evidence from the individual patient data and the literature data.

3.2.1 Adaptation method

Our adaptation method is an extension of a method developed by Greenland to adapt an unadjusted Odds Ratio for confounding[1]: if one study is available which corrected the logistic regression coefficient of an exposure variable (for example coffee

consumption) for a confounder (for example alcohol consumption), the change from unadjusted to adjusted logistic regression coefficient can be used to adapt the unadjusted coefficients in an other study. This procedure was labeled 'external adjustment of coefficients'[1]. In our case of regression analysis on literature data and individual patient data, the formula reads like:

$$\beta_{MULT|LIT} = \beta_{UNI|LIT} + (\beta_{MULT|IND} - \beta_{UNI|IND}). \quad (1)$$

The term ($\beta_{MULT|IND}$ - $\beta_{UNI|IND}$) is interpreted as an adaptation factor (or 'external adjustment factor'). In case no other data are available than the individual patient data, the literature data is equal to the individual patient data and the formula is self-evident.

In the following we will explore the conditions under which the variance of the $\beta_{MULT|LIT}$ is smaller than the variance of the $\beta_{MULT|IND}$. If these conditions are fulfilled, use of the univariate results from other studies ($\beta_{UNI|LIT}$) will be helpful in the estimation of the multivariate coefficients ($\beta_{MULT|LIT}$).

We approximate the variance of the $\beta_{MULT|LIT}$ as follows:

$$\begin{aligned} \text{var}(\beta_{MULT|LIT}) = {} & \text{var}(\beta_{UNI|LIT}) + \text{var}(\beta_{MULT|IND}) + \text{var}(\beta_{UNI|IND}) \\ & +2 \cdot \text{covariance}(\beta_{UNI|LIT}, \beta_{MULT|IND}) - 2 \cdot \text{covariance}(\beta_{UNI|LIT}, \beta_{UNI|IND}) \\ & -2 \cdot \text{covariance}(\beta_{MULT|IND}, \beta_{UNI|IND}). \end{aligned} \quad (2)$$

Since the individual patient data are assumed to form a random subsample from the literature, $\beta_{UNI|LIT}$ and $\beta_{UNI|IND}$ are positively correlated. Moreover, it may safely be assumed that

$$\text{covariance}(\beta_{UNI|LIT}, \beta_{MULT|IND}) < \text{covariance}(\beta_{UNI|LIT}, \beta_{UNI|IND}). \quad (3)$$

Assumption (3) says that univariate coefficients are stronger correlated between the two data sources than univariate and multivariate coefficients. Leaving out both of these covariances may lead to a slight overestimation of the variance of the var($\beta_{MULT|LIT}$), since the larger term has a minus sign in (2):

$$\begin{aligned} \text{var}(\beta_{MULT|LIT}) = {} & \text{var}(\beta_{UNI|LIT}) + \text{var}(\beta_{MULT|IND}) + \text{var}(\beta_{UNI|IND}) \\ & -2 \cdot \text{covariance}(\beta_{MULT|IND}, \beta_{UNI|IND}). \end{aligned} \quad (4)$$

We investigate the conditions under which var($\beta_{MULT|LIT}$) is smaller than var($\beta_{MULT|IND}$), which is equivalent with that

$$\text{covariance}(\beta_{MULT|IND}, \beta_{UNI|IND}) > 1/2 \cdot \text{var}(\beta_{UNI|LIT}) + 1/2 \cdot \text{var}(\beta_{UNI|IND}). \quad (5)$$

Instead of with the covariance($\beta_{MULT|IND}$, $\beta_{UNI|IND}$), we may prefer to define this condition with the correlation coefficient ρ. Since $\rho(\beta_{MULT|IND}, \beta_{UNI|IND}) = \text{covariance}(\beta_{MULT|IND}, \beta_{UNI|IND}) / [SE(\beta_{MULT|IND}) \cdot SE(\beta_{UNI|IND})]$, it follows that

$$\begin{aligned} \rho(\beta_{MULT|IND}, \beta_{UNI|IND}) > {} & 1/2 \cdot \text{var}(\beta_{UNI|LIT}) / [SE(\beta_{MULT|IND}) \cdot SE(\beta_{UNI|IND})] \\ & + 1/2 \cdot [SE(\beta_{MULT|IND}) / SE(\beta_{UNI|IND})]. \end{aligned} \quad (6)$$

The correlation coefficient ρ(ßMULT|IND, ßUNI|IND) is not directly estimated with logistic regression analysis, in contrast to all other terms. Re-sampling methods such as the jack-knife or bootstrap method may however be used to estimate this correlation.

Further, the literature data set will often contain many more patients than the data set with individual patient data. In this case, the var(ßUNI|LIT) will be small, and the term 1/2 • var(ßUNI|LIT) / [SE(ßMULT|IND) • SE(ßUNI|IND)] may be negligible. The ratio SE(ßUNI|IND) / SE(ßMULT|IND) is influenced by the correlation between the covariates and the strength of the multivariate relation of the covariates with the outcome. A positive correlation between covariates increases the SE(ßMULT|IND), while a strong multivariate association between the covariates and the outcome decreases the SE(ßMULT|IND). Condition (6) indicates that the adaptation method will lead to a substantial improvement of the regression coefficients if the ρ(ßMULT|IND, ßUNI|IND) is strong (close to 1), the var(ßUNI|LIT) or the SE(ßUNI|IND) small or the SE(ßMULT|IND) relatively large. In epidemiological terms, the adaptation method will work, if the confounding of risk factors is not too strong.

3.2.2 Simulation

The adaptation method was evaluated by simulation to obtain an impression of the magnitude of the improvement in the estimation of the multivariate coefficients ßMULT|LIT. Four databases were constructed with two covariates and one database with three covariates. All covariates and the dichotomous outcome had a 50%:50% ratio of 0 and 1 values. The associations between covariates and the outcome were varied, such that the multivariate logistic regression coefficients varied between 1.0 and 2.3. The correlation between covariates then determined the univariate coefficients. Without correlation, the univariate and multivariate coefficients were identical. With a positive correlation, the univariate coefficients were larger than the multivariate coefficients. This is illustrated in Table 1, which shows the structure of the first and second database. In both databases, two covariates had multivariate coefficients (ßMULT) of 1.0 and 1.5. This means that the probability of the outcome was identical in both databases for a patient with given values of covariate 1 and 2. For example, a patient with var1=0 and var2=0 had a probability of 22.3%. Since no correlation was present between var1 and var2 in the first database, the univariate coefficients (ßUNI) were 1.0 and 1.5 as well. The positive correlation in the second database resulted in a ßUNI of 1.4 and 1.8 respectively.

Table 1 Illustration of the structure of the first and second hypothetical database for the simulation study.

Var1	Var2	First database*	Second database*	Probability of the outcome**
0	0	25%	33.3%	22.3%
0	1	25%	16.7%	56.2%
1	0	25%	16.7%	43.8%
1	1	25%	33.3%	77.7%

* percentage of patients with each combination of values for var1 and var2 (total 100%).
** percentage of patients with the outcome, given the values of var1 and var2.

These databases may be interpreted as representing very large hypothetical patient series from published papers. Random Monte Carlo samples of 100 patients were drawn from the databases to represent the data set with individual patient data. Random sampling was repeated 500 times, which sufficiently limited random noise.

The regression coefficients ßUNI|IND and ßMULT|IND were calculated in each Monte Carlo sample with logistic regression analysis. Subsequently, the coefficient ßMULT|LIT was estimated with the adaptation method, using the univariate coefficients of the complete database (ßUNI). These estimates could be compared with the multivariate regression coefficients in the complete database (ßMULT). As noted before, these coefficients are usually not available in practice. In the evaluation they served as the gold standard.

Regression analyses tend to overestimate the true regression coefficients slightly[2]. This bias will however be small compared to the variance in the estimates of the coefficients. Our evaluation therefore focused on the improvement in the variance of the estimates achieved by the adaptation method. We calculated the percentage reduction of the standard error (SE) of the estimates with the adaptation method relative to the SE of the traditional logistic regression estimates. The SE was calculated as the standard deviation in the 500 replications of the estimates. Moreover, we calculated the fraction of simulations where the adaptation method led to an improvement of the multivariate regression coefficients. Improvement was defined as 'positive' if the adapted regression coefficient was more than 0.10 (in absolute terms) closer to the true coefficient in the meta-analysis database than the traditional coefficient, 'negative' if the adaptation resulted in a coefficient more than 0.10 away from the true multivariate coefficient, and equivalent in between. The values +0.10 and −0.10 were chosen arbitrarily, but are intended in a way similar to the quantification of the presence of confounding in etiological research[3].

3.3 Results

3.3.1 Performance of the adaptation method

The adaptation method reduced the standard deviation by over 40% in all five databases (Table 2). The improvement percentages varied between 50% and 70% compared to the traditional method. The risk of adaptation of the regression coefficients in the wrong direction (further from the true coefficient compared to the traditional estimate) was small in most simulations, and always less than 20%. In around 25% of the simulations, the adapted coefficients were similar to the traditional estimates (improvement ±). Within simulations, the percentage of improvement increased with magnitude of the coefficients. The fifth database contained three covariates, resulting in similar improvements as observed in the first four databases with two covariates. In other exploratory simulations, the relative improvement appeared independent of the sample size (N=100, N=200 or N=400).

Table 2 Improvement by using the adaptation method instead of the traditional method in estimating multivariate regression coefficients. Simulation study with samples of 100 patients.*

		Meta		Traditional	Adaptation method		Improvement		
#	Var	ßUNI	ßMULT	ßMULTIIND	ßMULTILIT	%SE	+	±	-
1	Var1	1.0	1.0	1.03±.45	1.02±.17	62%	66%	25%	9%
	Var2	1.5	1.5	1.57±.47	1.52±.12	74%	72%	24%	4%
2	Var1	1.4	1.0	1.04±.49	0.99±.21	57%	64%	26%	9%
	Var2	1.8	1.5	1.56±.50	1.51±.16	68%	68%	27%	6%
3	Var1	1.5	1.0	1.03±.50	1.00±.25	50%	56%	29%	15%
	Var2	2.3	2.0	2.10±.54	2.02±.15	71%	68%	28%	4%
4	Var1	1.5	1.5	1.58±.57	1.53±.33	43%	60%	24%	16%
	Var2	2.3	2.3	2.40±.60	2.34±.25	58%	68%	26%	6%
5	Var1	1.3	1.0	1.03±.55	1.01±.31	43%	58%	26%	16%
	Var2	1.7	1.5	1.57±.57	1.53±.29	49%	61%	27%	12%
	Var3	2.2	2.0	2.11±.60	2.04±.25	59%	66%	23%	11%

* For the five databases (#) with two or three covariates (Var), logistic regression coefficients are shown (ßUNI and ßMULT; mean ± SE), the percentage reduction in standard error (%SE = 1 − [SE ADAPTATION METHOD /SE TRADITIONAL]), and the percentage of simulations that the adaptation method resulted in an improvement (+/±/-).

3.3.2 Performance on real medical data

The adaptation method was further evaluated on a database of testicular cancer patients, where six binary predictors were related to the histology found at resection (necrosis vs other histology). We combined the data of five research groups which participated in this study (1 from New York, 1 from Oslo, Norway and 3 from The Netherlands). The results obtained with the complete data set were compared with the results in each study individually, with or without the adaptation method (Table 3). It appears that the adaptation method results in improved estimates of the multivariate regression coefficients in most instances (15 of 30 adaptations = 50%), but worsened the estimates in 17% (5/30). This performance is somewhat less favourable than in the simulation study. This may be caused by heterogeneity between the studies forming the meta-analysis, in contrast to the strictly random selection of samples in the simulation study.

Table 3 Evaluation of the adaptation method in testicular cancer prediction study. Study sizes were N=121, N=127, N=137, N=33, N=84 (Meta N=502).

	Meta		Study #					Improvement		
	ßUNI	ßMULT	1	2	3	4	5	+	±	-
Teratoma elements	1.14	1.02	+	-	±	+	-	2	1	2
AFP normal	1.04	0.88	+	±	+	+	+	4	1	0
HCG normal	0.71	0.70	±	±	-	+	±	1	3	1
LDH elevated	0.51	1.12	+	+	+	-	+	4	0	1
Postsize ≤ 20mm	1.57	1.32	±	+	±	-	±	1	3	1
Change >70%	1.56	0.87	+	±	±	+	+	3	2	0
								15	10	5

3.3.3 Model performance and influence of selection

In addition to the estimation of the individual multivariate regression coefficients, we studied the performance of the predictive model as a whole. Calibration and discriminative ability were determined for models developed in each participating center in the testicular cancer study when evaluated in the total data set consisting of all patients from the five centers.

We compared the performance of models developed with a fixed set of covariates with models developed with a common selection strategy for the covariates, i.e. forward stepwise selection with the standard significance level of 5% for entry of variables. Thus three methods could be evaluated:

- fixed set of variables, traditional estimation of the coefficients
- fixed set of variables, use of the adaptation method to estimate the coefficients
- stepwise selected set of variables, traditional estimation of the coefficients

The evaluation on the total data set will be slightly biased in favour of the adaptation method, since information of the total data set (the univariate regression results) is used in the estimation of the coefficients. This bias will however be very small, since the number of coefficients is limited (N=6) compared to the total number of patients (N=502).

Overall calibration was studied by calculating a prognostic index (PI) for all patients in the total data set with the regression coefficients of each of the three methods[4]. A logistic regression coefficient of the PI smaller than 1 indicates that the applied method leads to overoptimistic coefficients (too large in the study data set). Table 4 shows that overoptimistic coefficients are estimated with all three methods. The coefficient of the PI is however closest to one with application of the adaptation method on a fixed set of predictors for most studies. A considerable improvement is achieved relative to the traditional calculation of regression coefficients, especially in study 1 and 2. Study 4 however showed a large overestimation of the regression coefficients, even with adaptation of the coefficients. This may be explained by the small sample size (N=33). Stepwise selection led to inclusion of 2 to 4 variables in the study samples. The regression coefficients are in most studies more overoptimistic than with the traditional estimation of coefficients in the fixed set of 6 predictors. Forward stepwise selection thus appears not to improve the calibration of the models.

Discriminative ability of the models was measured by the area of the ROC curve or *c* statistic[5]. The absolute discriminative ability achieved with the models containing 6 predictors was around 0.8, and only the smallest study had a lower area under the ROC curve (#4, N=33, c=.75). The adaptation method led to a slightly larger area in the first three studies, and was more or less equivalent in the last two compared to the traditional method. The area was much smaller when stepwise selection of variables was applied as a modeling strategy.

Table 4 Performance of different logistic regression models in testicular cancer prediction study. Study sizes were N=121, N=127, N=137, N=33, N=84.

		Study				
		1	2	3	4	5
Reliability: coefficient of PI						
Six predictors:	traditional	.91	.82	.74	.37	.80
	adapted	1.00	.99	.77	.48	.73
Stepwise:	traditional	.83	.76	.67	.38	.80
Discriminative ability: ROC area						
Six predictors:	traditional	.795	.790	.780	.752	.798
	adapted	.800	.805	.796	.748	.800
Stepwise:	traditional	.763	.753	.762	.691	.719

3.4 Discussion

In this study we developed a new method to estimate logistic regression coefficients in the presence of quantitative literature data. This method combines the results of a univariate meta-analysis with the results of a univariate and multivariate logistic regression analysis on individual patient data. It appears that this combination can result in a substantial improvement of the estimates as compared to the traditional way of estimating the regression coefficients (without explicit consideration of the literature).

The development of a prognostic logistic regression model would ideally take place in a very large data base with individual patient data of high quality. In practice, this ideal is seldom achieved and the data set of individual patient data is usually relatively small. This has several disadvantages. First, it may be difficult to select predictors for the prognostic model. Standard stepwise variable selection methods lead to overestimation of the coefficients. Secondly, regression models inherently tend to estimate the coefficients too extreme[2], and this problem occurs especially in relatively small data sets with a large number of predictors. Thirdly, the estimates of the regression coefficients will be imprecise with relatively large confidence intervals. Prognostic models are thus often unreliable when developed in small data sets, since the regression coefficients are both biased to more extreme values and imprecise.

This study shows that the reliability of a prognostic model can be improved in such situations, if literature data are available that can be summarized quantitatively in a meta-analysis. First, the selection of variables may be based on the meta-analysis. The higher number of patients will indicate prognostically important variables more clearly than one single study. Next, the univariate literature data may be used to improve the estimates of the multivariate regression coefficients. In this study, we considered a simple adaptation method. We defined the situations where the adaptation method would be more efficient than the traditional method to estimate the multivariate regression coefficients. In a simulation study, we found that the variance of the regression coefficients was substantially smaller with this adaptation method. Also, the systematic overestimation of the coefficients appears smaller compared to the traditional method[2].

The better estimates of the regression coefficients are expected to result in improved performance of a prognostic model, as distinguished in calibration and discrimination. In a real medical data set, it was shown that calibration improved clearly. Discriminative ability improved to a much lesser extent. This is in concordance with the finding that the estimates of the regression coefficients are not of major importance for discrimination. In the context of diagnostic tests, the application of Bayes rule assuming conditional independence ("Idiot's Bayes") has sometimes resulted in good discriminative performance[6]. Discriminative ability is however strongly influenced by the selection of variables. We found that discrimination was much worse if a limited number of variables was selected with a standard forward stepwise selection method, compared to a fixed set of predictors, which were all highly significant in the univariate meta-analysis.

The proposed modeling method comprises a central role for the meta-analysis of published literature data. A potential problem of meta-analyses is that publication bias may have led to overestimation of the regression coefficients. The presence of publication bias may however be detected by examination of the relation between study size --, or more specific, the variance of the Odds Ratio,-- and the magnitude of the Odds Ratio. If larger ORs are observed in smaller studies, this indicates publication bias. Publication bias may partly be corrected by a regression model with the observed OR as dependent variable and the variance of the OR as independent variable, using a modification of a previously described method[7].

Further, the meta-analysis will only provide univariate statistics, which are influenced by the correlations between predictors. In case of predictors that are positively related with the outcome (OR>1), a positive correlation of a predictor with other predictors will lead to a large univariate Odds Ratio and thus to inclusion in the prognostic model. On the other hand, a predictor may seem unimportant in the meta-analysis because of a negative correlation with other predictors. This predictor would falsely be excluded from the prognostic model. Fortunately, clinical characteristics are more often positively correlated than negatively in most medical data sets.

The central assumption in the adaptation method is that the data set under study is a random subsample from the literature data. This implies that the relations between predictors and the outcome are similar, and that the correlations between predictors are similar in the individual patient data and in the literature data. Similarity of the relation between predictors and the outcome can statistically be assessed by tests for homogeneity. Unfortunately, similarity of the correlations between predictors cannot be examined in most instances, since correlations between predictors are infrequently published. The application of the here proposed adaptation method might be validated better if publications included a small table with correlations between the predictors to allow comparison between studies.

A final aspect of the adaptation method is that the estimation of the constant, or intercept, in the prognostic model may be difficult. The constant might be re-calculated in the data set with individual patient data, using the adapted regression coefficients in a prognostic index and the average frequency of the outcome in this data set. An

alternative approach is to calculate the intercept by assuming that the average frequency of the outcome corresponds to a patient with average frequencies on all predictors. This approach is mathematically incorrect, because of the non-linearity in the logistic transformation of the prognostic index. A mathematically more correct calculation might include likelihood ratios, which are by definition related to the average frequency of the outcome. An example of this approach is given in the estimation of mortality in elective aortic aneurysm surgery (Chapter 9). In most situations, the difference between the latter methods will be small.

We conclude that literature data should be considered explicitly for prognostic modeling. The literature data may guide the selection of variables, as well as improve the estimates of the regression coefficients. This modeling strategy will result in more reliable logistic regression models than obtained with a strategy that considers a data set with individual patient data as the sole basis for prognostic modeling.

We would like to thank Professor Hans van Houwelingen, PhD, and Ronald Brand, PhD, Dept of Medical Statistics, University of Leiden, for many helpful comments on a previous version of this chapter.

References

1. Greenland S. Quantitative methods in the review of epidemiologic literature. Epidemiologic Rev 1987, 9: 1-30
2. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990; 9: 1303-1325
3. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quatitative methods. Lifetime learning publications, London, 1982
4. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Stat Med 1991; 10: 1213-1226
5. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982; 247: 2543-2546
6. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986; 5: 421-433
7. Brand R, Kragt H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. Stat Med 1992; 11: 2077-2082

Applications

4 Prognosis after resection of residual masses following chemotherapy for metastatic nonseminomatous testicular cancer: a multivariate analysis

E.W. Steyerberg, H.J. Keizer, J. Zwartendijk, G.L. Van Rijk, C.J. Van Groeningen, J.D.F. Habbema & G. Stoter.
Br J Cancer 68: 195-200, 1993

Abstract

Following chemotherapy for metastatic nonseminomatous testicular cancer, 86 patients with normal serum markers AFP and HCG underwent resection of residual tumor masses (63 laparotomy, 11 thoracotomy, 12 both). Prognostic factors for relapse and survival were analyzed with Kaplan-Meier curves and Cox regression analysis. Putative prognostic factors included age, the primary histology, prechemotherapy level of the tumor markers AFP and HCG, the extent of disease (lymph nodes, lung and hepatic metastases) before and after chemotherapy, the histology of the resected material and the completeness of the surgical procedure. Eleven patients relapsed during follow-up (median 47 months), accounting for a 5-year relapse free percentage of 87.4%. Adverse prognostic factors were:

- prechemotherapy level of HCG (≥ 10,000 IU/l)
- incomplete resection
- the extent of disease, especially of lung metastases (prechemotherapy number ≤3, 4-19, ≥20; or size after chemotherapy >1 cm; or presence of any residual lung metastasis after chemotherapy without residual abdominal metastases)

The histology found at resection was not associated with the risk of relapse, which might be explained by the effectiveness of postresection chemotherapy, which in the majority of these patients was a salvage regimen rather than two further cycles of the initial cytostatics. A good and a poor risk group were formed, based on HCG level and completeness of resection. The effect of salvage chemotherapy after resection of viable cancer cells needs further investigation.

4.1 Introduction

Cisplatin combination chemotherapy yields a 60-80% cure rate in metastatic nonseminomatous germ cell tumors (NSGCT) of the testis[1,2,3]. If residual masses are detected after chemotherapy, surgical resection is usually performed[4], although no general agreement exists whether all patients should be operated on[5,6,7]. Additional chemotherapy is usually given if viable cancer cells are present in the resected specimens, to kill remaining microscopic disease[4]. It has been suggested that the type of additional chemotherapy should preferably be a salvage regimen, rather than two further cycles of the initial chemotherapy[8].

The goal of this study was to analyze the prognosis of patients after resection of residual masses detected on CT scan, while tumor markers were normal. Study parameters were relapse of tumor and survival. Putative prognostic factors included the patient's age, the primary histology, prechemotherapy level of the tumor markers AFP and HCG, the extent of disease (lymph nodes, lung and hepatic metastases) before and after chemotherapy, the histology of the resected material and the completeness of the surgical procedure. First, we investigated which factors univariately affected prognosis. Further, we analyzed multivariately whether information obtained at resection (completeness and histology) influenced the prognosis of the patient. Finally, we tried to identify which factors were most important in predicting relapse, combining all factors known after resection.

4.2 Patients and Methods

4.2.1 Patients studied

We reviewed the charts of 210 consecutive patients with first presentation of metastatic nonseminomatous testicular cancer or seminoma with elevated tumor markers, referred to three Dutch cancer centers. The patients were treated between July 1980 and June 1991, most of them in randomized trials of the EORTC. Treatment consisted of cisplatin combination chemotherapy (Peckham 1988; Einhorn, 1990). After completion of induction chemotherapy, the size of metastases was determined on CT scan. If residual masses were detected (≥ 1 cm), resection was planned, provided that tumor markers were normal. If tumor markers remained elevated, additional chemotherapy was usually given until normalization of tumor markers, and subsequent resection was performed of residual masses (n=5). In 99 patients residual masses were resected. Excluded were the following patients to prevent prognostic inhomogeneity: 7 patients not treated according to standard protocol (e.g. treated with radiotherapy before induction chemotherapy); 3 patients with extragonadal tumors; 3 patients who were operated while tumor markers were above normal. After this selection 86 patients were included in the analysis.

4.2.2 Patient data

All patient data were updated until October 1991. The histological diagnosis of the original testicular cancer was made in the participating hospitals and was reviewed for patients in EORTC trials. In the analysis, the British classification[9] was used. The disease was staged according to the Royal Marsden Hospital classification[1]. Further, the maximum transverse diameter of abdominal masses, the maximum transverse diameter of pulmonary tumor nodules, and the number of lung metastases were determined on computed tomographic (CT) scan before and after chemotherapy. The highest serum levels of AFP (ng/ml) and HCG (IU/l) prior to chemotherapy were recorded. The type and number of chemotherapy regimens were registered, before and after resection, as well as the completeness of resection as noted by the surgeon, and the type of histology in the resected material. The data were divided in 3 groups of potential prognostic factors: factors known at the start of cytostatic treatment ('prechemotherapy factors'), factors known after chemotherapy but before resection ('postchemotherapy factors') and factors known only after resection of the residual mass ('resection factors').

4.2.3 Patient characteristics

Table 1 gives the characteristics of the patient population. For each participating center the number of patients in this study and the period of accrual is presented. Median age of the patients was 26.5 years. The primary histology of the testicular cancer was predominantly MTI (teratocarcinoma, 48%) and MTU (embryonal carcinoma, 42%). Abdominal metastases were present in 80 patients (93%), mediastinal metastases in 4 (5%), supraclavicular metastases in 8 patients (9%), and lung metastases in 43 (50%), of whom 10 patients had a largest diameter ≥ 3 cm and 7 had ≥ 20 lung metastases. Hepatic metastases were observed in 7 patients (8%), and a metastatic inguinal node in one patient. AFP serum values were elevated (> 16 ng/ml) in 63 patients, and higher than 1000 ng/ml in 17 patients. HCG serum values were elevated (> 4 IU/l) in 60 patients with 10 over 10,000 IU/l. Standard chemotherapy changed during the last decade from PVB to BEP or EP, alternating PVB/BEP and more recently alternating to BOP/VIP regimens. After chemotherapy, a laparotomy was performed in the majority of the patients in the study group (63 laparotomy only, 12 laparotomy and thoracotomy). The procedure was a radical retroperitoneal lymph node dissection (RPLND) at the University Hospital Leiden (39 patients) and was limited to resection of all pathological masses in the other centers (36 patients). The surgeons reported incomplete resections in 8 patients (9%). Two of these patients had viable cancer cells. Overall, the histology of the resected material was necrosis/fibrosis in 38 patients (44%), mature teratoma in 32 (37%) and viable cancer cells in 16 (19%). The malignant cells most often resembled the primary histology.

Table 1 Characteristics of 86 patients resected for a residual mass after chemotherapy.

Factor	Classification
Hospital*: n, period	RCI: 24, 1983-1990 AZVU: 18, 1980-1990 AZL: 44, 1981-1991
Age: median, range	26.5 year, 18-43
Primary histology	41 MTI (teratocarcinoma) 36 MTU (embryonal carcinoma) 5 MTT 2 TD (teratoma differentiated) 2 Seminoma (HCG 200 and 19.000 IU/l)
Tumor markers: n elevated, median	AFP: 63, 117 ng/ml BHCG: 60, 38 IU/l
Stage II** Abdominal lymph node metastases	6 none 10 A 33 B 37 C
Stage III** Mediastinal or supraclavicular metastases	74 none 3 M1 1 M2 3 N1 5 N2
Stage IV** Lung metastases	43 none 24 L1 5 L2 14 L3
Stage IV** Other metastases	78 none 7 H+ 1 soft tissue
Chemotherapy ***: n, type, period	15 PVB, 1980-1982 14 EP, 1983-1984 33 BEP, 1983-1991 8 PVB/BEP, 1983-1987 6 VIP, 1987-1989 10 BOP/VIP, 1987-1990
Type of surgery	63 laparotomy 11 thoracotomy 12 both
Complete resection	78 yes, 8 no
Histology at resection	38 necrosis 32 mature teratoma 16 viable cancer cells

* RCI : Rotterdam Cancer Institute; AZVU: Free University Hospital Amsterdam; AZL: University Hospital Leiden; ** Royal Marsden Classification; *** B = Bleomycin; E = Etoposide; I = Ifosfamide; O = Vincristine (Oncovin); P = Cisplatinum; V = Vinblastine (in PVB regimen) / Etoposide (VP-16 in VIP regimen)

4.2.4 Statistical analysis

The main endpoint in this study was the diagnosis of relapse of tumor. Relapse was defined as a rise of AFP or HCG serum levels above normal levels, or, in the absence of elevated markers, histological proof of malignancy. Growing mature teratoma without viable cancer cells was not considered as a relapse, because the patient's prognosis is not directly jeopardized by this event. The relapse free period was calculated from the date of resection and ended by relapse in eleven patients. In the censored patients the relapse free period ended by death due to surgery (2 patients), death to unrelated causes (1 patient after 60 months), or the most recent visit to the hospital (72 patients; median follow-up 47.1 months; range: 5.2 - 127). Overall survival used the endpoint death. Kaplan-Meier curves were used to describe the relationship of single variables (Table 2) and the endpoint[10] and groups were compared by the log-rank test[11]. Cox regression[12] was applied to model the simultaneous effect of several variables. Significance for entry of variables was calculated from a Likelihood Ratio (LR) statistic. The additional prognostic influence of resection factors (histology, completeness) was assessed by including these factors in Cox regression models which already contained prechemotherapy and postchemotherapy factors. To identify the variables with the most important effect on relapse, a forward stepwise selection method was used, with $p<.05$ as an entry criterion. The Hazard Ratios (HR) provided by these models may be interpreted as relative risks. Continuous data from completely and incompletely resected patients were compared by the Mann-Whitney test, which is the non-parametric equivalent of the classical t-test.

Table 2 Potential prognostic factors for relapse.

Factor	coding and categorization
Prechemotherapy	
Age at orchidectomy	(continuous)
Primary histology	main diagnosis (MTI yes/no; MTU yes/no)
	presence of elements (Seminoma yes/no; trophoblastic yes/no)
Extent of disease	lymph node metastases size (≤ 2, 2-5, ≥ 5 cm; < 5, 5-10, ≥ 10 cm)
	lung metastases size (none, ≤ 3, > 3 cm) and
	number (none or ≤ 3, 4-19, ≥ 20; < 20; ≥ 20)
	hepatic metastases (presence)
	number of sites of metastases (0, 1, ≥ 2)
Tumor marker levels	AFP and HCG (elevated, continuous, discrete)
Postchemotherapy	
Extent of disease	lymph node size (< 1 and lung ≥ 1, ≥ 1) decrease (continuous; yes/no)
	lung metastases size (≤ 1, > 1 cm) and decrease in size/number
Resection	
Completeness	(complete/incomplete)
Histology at resection	(necrosis/fibrosis, mature teratoma, viable cancer cells)

4.3 Results

4.3.1 Relapse and survival

Overall survival is shown in Figure 1. Ten patients died during follow-up: two shortly after resection, seven after relapse and one at 60 months, due to unrelated causes. The two patients who died shortly after operation account for a 2.3 % operative mortality. One patient suffered from bleomycin toxicity, the other had postoperative cardiac problems, and both had mature teratoma resected. Two years after resection 91.4% of the patients were still alive (86.9% alive without relapse, 4.5% alive after relapse). After five years 87.2% were alive: 85.4% and 1.8% without and after relapse, respectively.

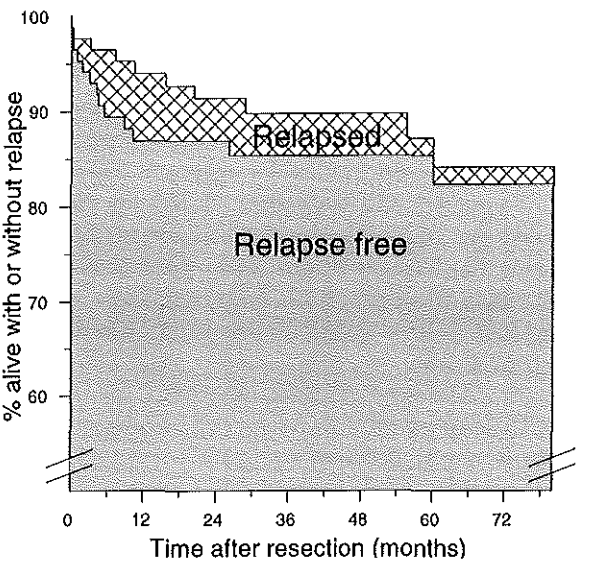

Figure 1 Survival after resection. Kaplan-Meier plot showing the percentage alive (upper line) and the percentage alive and relapse free (lower line).

Survival after relapse of the 11 patients who relapsed after resection is depicted in Figure 2. Seven patients died (median 14.6 months after relapse). If the relapse occurred early (within 2 months), subsequent survival appeared poor (p=0.042, log-rank test).

Figure 3 shows the relapse free Kaplan-Meier plot of all 86 patients. The 5-year relapse free percentage (5y-RF%) was 87.4 %, with a 95% confidence interval (95%-CI) ranging from 78% to 93%. Most relapses (9/11) occurred within 12 months. One patient relapsed after 26 months. One late relapse occurred after 123 months, while only 2 patients were still at risk at that time (not displayed in Figure 1 and 3, but used in statistical analyses). This patient was incompletely resected in 1980 (histology: viable cancer cells and mature teratoma) and he relapsed in 1990 with extensive masses in the abdomen, liver and lung. Salvage chemotherapy was successful and the patient had no evidence of disease ten months after the relapse.

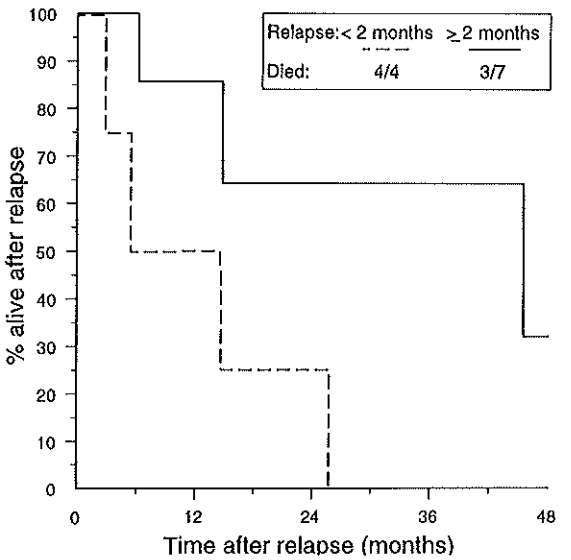

Figure 2 Survival after relapse (n=11), stratified for relapse free period (< or ≥ 2 months, p=.042, log-rank test)

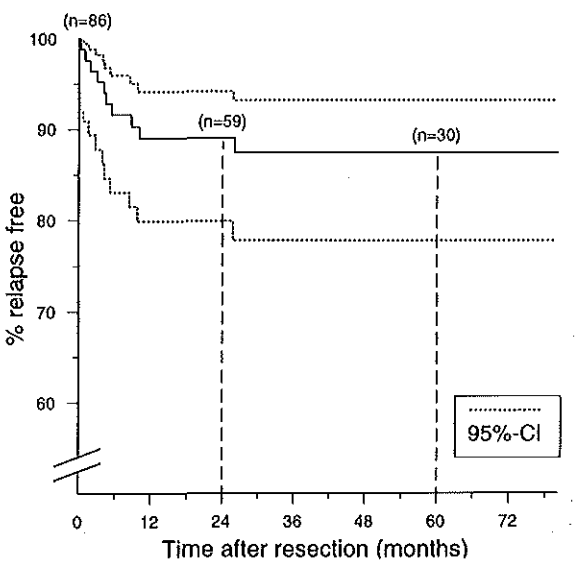

Figure 3 Relapse after resection, censoring deaths without prior relapse (two postoperative deaths, one at 60 months).

4.3.2 Univariate relations with relapse

Univariate analyses revealed significant associations with relapse after resection ($p < 0.05$, log-rank test) for several of the potential prognostic factors of Table 2. These are shown in Table 3. Of the prechemotherapy factors, age or primary histology were not significantly related with relapse. The extent of lung metastases influenced the relapse rate: size (> 3cm, p=.047) and number, coded in two groups (< 20, ≥ 20: p=.007) or, more significantly, coded in three groups (≤ 3, 4-19, ≥ 20: p=.003). The extent of lymph node metastases, the presence of hepatic metastases or the number of sites had p-values > 0.10. For example, of 15 patients with abdominal lymph nodes > 10 cm, 3 relapsed (5y-RF%: 80%). Of the 7 patients with hepatic metastases, only 1 relapsed (5y-RF%: 86%). Differences in relapse rate were observed according to the prechemotherapy serum HCG values (0-999, 1000-9999, ≥10000, p=.014; 0-9999, ≥10000, p=.001), contrary to the prechemotherapy level of AFP (p > 0.10). It is of note that of the ten patients with HCG ≥ 10.000 IU/l, three of four who relapsed, relapsed with brain metastases.

Postchemotherapy lung metastases were prognostically important. Adverse characteristics were a postchemotherapy lung metastasis size > 1 cm (p=.003) or the presence of any lung metastasis without detectable residual abdominal metastases (p=.001). No difference in relapse rate was observed according to the decrease in size or decrease in number of metastases.

The most significant factor for relapse was the completeness of resection (Table 3). The 5y-RF% was only 50% in incompletely resected patients, compared to 92% in completely resected patients (p=.0004). The histology of the resected material had no significant relationship with relapse: 5y-RF% [95%-CI] was 89% [73%-96%] (4 relapsed of 38), 85% [64%-94%] (4 of 32) and 88% [59%-97%] (3 of 16) for necrosis, mature teratoma and cancer respectively (p=.89).

4.3.3 Prognostic influence of resection

The extent of disease was significantly correlated with the completeness of resection: incomplete abdominal resections occurred more frequently in large lymph nodes (before and after chemotherapy, p=.023 and p=.020, respectively, Mann-Whitney test) and incomplete lung resections occurred more frequently if more residual nodules had to be resected (p=.010, Mann-Whitney test). Because of this correlation, the additional prognostic effect of the completeness of resection was explicitly studied while taking into account the extent of disease. Also, correction was made for the prechemotherapy HCG level and the center (Leiden or other) where the patient was resected, as the technique of abdominal lymph node resection varied between the centers. It then appeared that incompletely resected patients had a much poorer prognosis than completely resected patients (Hazard Ratio > 5, p<.02). The histology at resection provided no additional prognostic information (p > .20, Likelihood Ratio test).

Table 3 Significant prognostic factors for relapse (from Table 2). N indicates the number of patients in each category, with the observed 5 year relapse free percentage in the column 5y-RF%. P-values are calculated with the log-rank test.

Factor	Categorization	N	5y-RF%	p-value
Prechemotherapy				
Number of lung metastases	0 - 3	67	94%	.003
	4 - 19	12	63%	
	≥ 20	7	57%	
Highest HCG serum level	0-9999 IU/l	76	91%	.001
	≥ 10000 IU/l	10	58%	
Postchemotherapy				
Residual lung metastases without abdominal metastases	lymph nodes ≥1cm, lymph nodes <1cm and lung ≥1cm	77 9	91% 56%	.001
Size of lung metastases	none or ≤ 1 cm	71	92%	.003
	> 1 cm	15	67%	
Resection				
Completeness of resection	complete	78	92%	.0004
	incomplete	8	50%	

4.3.4 Multivariate prediction of relapse

Following a forward stepwise selection procedure (Table 4), the completeness of resection (p=.004) and prechemotherapy HCG level (p=.006) appeared to be the most important predictors of relapse. The Hazard Ratios were 8.8 and 7.9 respectively. The third variable that entered the model was the presence of residual lung metastases without abdominal metastases (HR=6.8, p=.02). At step 3, neither the postchemotherapy size of lung metastases (≤ 1 cm or > 1 cm) nor the number of prechemotherapy lung metastases (≤ 3, 4-19, ≥ 20) improved the model significantly (p=.15 and p=.34 respectively, Likelihood Ratio test).

Table 4 Multivariate analysis of prognostic factors for relapse, based on the univariately significant factors from Table 3. P-values are calculated for the Likelihood Ratio (LR) statistic with 1 or 2 degrees of freedom (df). Step 1, 2, and 3 refer to the inclusion of variables in the stepwise forward selection procedure.

Step	Variable (categorization)	LR statistic	df	p-value
1	Completeness of resection (complete, incomplete)	8.3	1	.004
2	Prechemotherapy HCG serum level (0-9999, ≥ 10.000 IU/l)	7.6	1	.006
3	Residual lung nodules without abdominal masses (yes/no)	5.4	1	.02
3	Size of lung metastases (≤ 1 cm, > 1 cm)	2.1	1	.15
3	Number of lung metastases (≤3, 4-19, ≥20)	2.2	2	.34

We used the first two factors from the stepwise selection procedure to define a simple prognostic classification. A good and a poor prognosis group were distinguished based on the prechemotherapy level of HCG and the completeness of resection (Figure 4). The good prognosis group was defined by prechemotherapy HCG values under 10,000 IU/l and a complete resection, and had an estimated 5y-RF% of 95% (95%-CI: 85%-98%). The majority of patients in this study (69/86=80%) had this very favorable prognosis. The poor prognosis group was formed by patients with HCG ≥ 10,000 IU/l and complete resection (5y-RF%: 65%), patients with HCG < 10,000 IU/l and incomplete resection (5y-RF%: 57%) and patients with both HCG ≥ 10,000 IU/l and incomplete resection (relapsed: 1/1). This poor prognosis group of 17 patients had a 5y-RF% of 58% (95%-CI: 31%-77%).

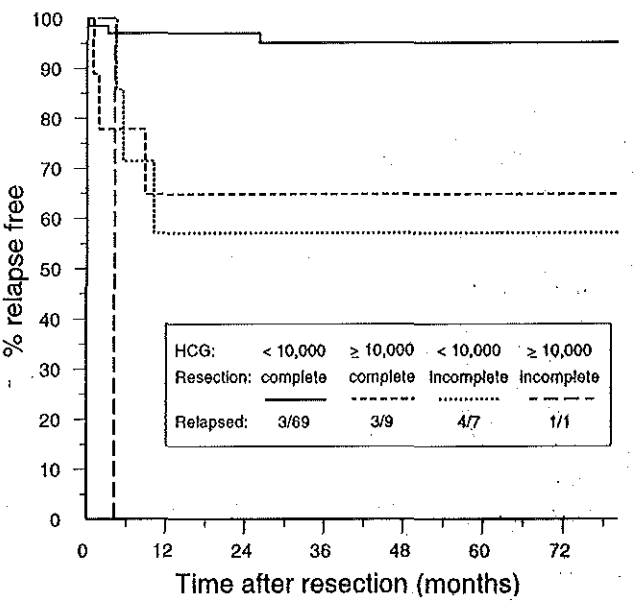

Figure 4 Relapse after resection, stratified for prechemotherapy HCG level and completeness of resection.

4.4 Discussion

This paper describes the prognosis of 86 patients with NSGCT of the testis, who underwent resection of residual masses after chemotherapy, while tumor markers were normal. Residual masses had a minimum size of 1 cm. These 86 patients make up 41% of the total group of 210 patients who received chemotherapy during the study period of approximately 11 years. Viable cancer cells were found in 16 patients (19%).

A recent review[7] showed that the percentage of resected specimens containing viable cancer cells is around 20%, and this percentage was also found in some other recent publications[13,14,15] and in our series. However, the fraction of resected patients varies widely between these studies, e.g. from around 20%[15,16] to over 85%[17,18]. The wide variation in the fraction of patients in whom it was deemed necessary to undergo resection may partly be explained by the heterogeneity of the patient groups, but also reflects the lack of agreement on the selection criteria for surgery after chemotherapy. For instance, there is disagreement in the definition of a postchemotherapy normal CT scan, varying from "absolutely normal"[7] to smaller than 2 cm[15,19]. Further, it has been advocated to perform a laparotomy in any patient with initial abdominal lymph nodes > 3cm, even if no pathologic mass could be detected on postchemotherapy CT scan[8]. Thus, the fraction of resected patients was as high as 51%[8] or, when resection was performed in practically all patients with "absolutely normal" CT scans: 86%[18]. Large European studies reported 31%[14] and 20%[15], reflecting the policy to resect CT-detectable residual masses only if these exceed an arbitrarily chosen size (≥ 1 cm, ≥ 2 cm). Further, subgroups of patients have been defined for whom the mortality and morbidity of resection may not be balanced by the small risk of leaving tumor unresected[6,7].

Our analysis of prognostic factors for relapse after resection showed (Table 3) that significant prechemotherapy factors were the size (> 3 cm) and number of lung metastases. Although the cut-off point for the prechemotherapy number of lung metastases at ≥ 20 is applied rather sharply in clinical practice, it is obvious that the change in prognosis has a more gradual course; we found that a more accurate categorization of the number of lung metastases is in three groups (≤3, 4-19, ≥20). Also, the initial serum value of the tumor marker HCG (≥ 10,000 IU/l) has major prognostic impact. These factors were also found in other studies to predict relapse[1,3,15,20,21], or to predict a complete clinical response after initial chemotherapy[20,21]. Postchemotherapy adverse prognostic factors were the size (> 1cm) of lung metastases, or the presence of any residual lung metastasis without detectable residual abdominal lymph metastases. Thus, the most important factors after chemotherapy and before resection, were the level of HCG and the extent of lung metastases.

The influence of the resection factors (completeness and histology) was studied in detail. Incompletely resected patients had a poor prognosis (5y-RF%: 50%), as was found in other studies[13,16,22]. The size of retroperitoneal metastases and the size of lung metastases were significantly correlated with the completeness of resection. However, the adverse prognosis of incompletely resected patients was not explained by these factors, nor the prechemotherapy HCG level, nor the center where the patient was

resected. Thus, the patient had a poorer prognosis if the surgeon was unable to perform a complete resection, independent of other potential prognostic factors. An explanation for this finding might be that intrinsic tumor characteristics, such as chemosensitivity[13] or grade of malignancy of distinct tumor cell populations, are different in patients who could not be resected completely. This explanation is supported by the observation that of the five relapses in incompletely resected patients only one was definitely in the resection area.

The histology at resection was not related to relapse in our patients, similar to one other report[22], but in contrast to the observations in several other studies[13,15,16,23]. The observation in our study may be explained by lack of power to detect an existing difference due to the relatively low number of relapses. A more interesting explanation is that the additional treatment after resection has been more effective than in other studies in controlling remaining microscopic disease, since a salvage chemotherapy regimen was used rather than two further cycles of the initial chemotherapy in ten of the 16 patients with viable cancer cells resected. The other six patients received two further cycles of the initial regimen (n=4), radiation therapy after eight courses of chemotherapy before resection (n=1), or no further treatment of a mesenchymal tumor (n=1). This observation supports the recommendation to change the chemotherapy regimen after resection[8].

The question rises whether incompletely resected patients might also benefit from a salvage regimen immediately after resection, even when no viable cancer cells are found in the resected material. The following observation suggests that benefit of additional chemotherapy might be obtained in these patients: six of the eight incompletely resected patients had no residual malignancy diagnosed (one necrosis, five mature teratoma). Five of these did not receive any additional chemotherapy after resection, and four relapsed. The other three patients received additional chemotherapy (one mature teratoma and two viable cancer cells resected). Of these, only one relapsed (123 months after resection).

According to our simple prognostic model (Figure 4), a poor prognosis is expected in patients with prechemotherapy HCG values over 10,000 IU/l or an incomplete resection. The poor prognosis of patients with a high prechemotherapy level of HCG is already being recognized by a number of treatment groups[15,20,21] and these patients are candidates to receive more intensive induction chemotherapy. Improvement of the prognosis of incompletely resected patients might be obtained by the administration of salvage chemotherapy after resection, although further research has to confirm this suggestion. The use of a salvage regimen after resection rather than two further cycles of the same chemotherapy is also subject to further investigation as well as more detailed recommendations for the selection of patients who would benefit from surgical resection.

We would like to thank Jo Hermans, PhD, Dept. of Medical Statistics, University of Leiden, for statistical support.

References

1. Peckham M. Testicular cancer. Rev Oncol: 1, 439-453, 1988
2. Einhorn LH. Treatment of testicular cancer: a new and improved model. J Clin Oncol 8, 1777-1781, 1990
3. Stoter G, Koopman A, Vendrik CPJ, Struyvenberg A, Sleijfer DTh, Willemse PHB, et al. Ten-year survival and late sequelae in testicular cancer patients treated with cisplatin, vinblastine, and bleomycin. J Clin Oncol 7, 1099-1104, 1989
4. Donohue JP & Rowland TG. The role of surgery in advanced testicular cancer. Cancer 54: 2716-2721, 1984
5. Levitt MD, Reynolds PM, Sheiner HJ, Byrne MJ. Non-seminomatous germ cell testicular tumours: residual masses after chemotherapy. Br J Surg 72: 19-22, 1985
6. Donohue JP, Rowland RG, Kopecky K, Steidle CP, Geier G., Ney KG, Einhorn L., Williams S, Loehrer P. Correlation of computerized 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. Fosså SD, Qvist H, Stenwig AE, Lien HH, Ous S, Giercksky KE. Is postchemotherapy retroperitoneal surgery necessary in patients with nonseminomatous testicular cancer and minimal residual tumor masses? J Clin Oncol 10: 569-573, 1992
8. Toner GC, Panicek DM, Heelan RT, Geller NL, Lin S-Y, Bajorin D et al. Adjunctive surgery after chemotherapy for nonseminomatous germ cell tumors: recommendations for patient selection. J Clin Oncol 8: 1963-1694, 1990
9. Pugh RCB. Testicular tumours, introduction. In *Pathology of the testis,* Pugh RCB. et al. (eds) pp. 139-162. Blackwell Scientific Publishers: Oxford, 1976
10. Kaplan EL & Meier P. Nonparametric estimates from incomplete observations. J Am Stat Assoc 53: 457-481, 1958
11. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 50: 163-170, 1966
12. Cox DR. Regression models and life tables (with discussion). J Roy Stat Soc, B 34: 187-220, 1972
13. Jansen RLH, Sylvester R, Sleijfer DTh, Ten Bokkel Huinink WW, Kaye SB, Jones WG et al. Long-term follow-up of non-seminomatous testicular cancer patients with mature teratoma or carcinoma at postchemotherapy surgery. Eur J Cancer 27: 695-698, 1991
14. Dearnaley DP, Horwich A, A'Hern R, Nicholls J, Jay G, Hendry WF, Peckham MJ. Combination chemotherapy with bleomycin, etoposide and cisplatin (BEP) for metastatic testicular teratoma: long-term follow-up. Eur J Cancer 27: 684-691, 1991
15. Mead GM, Stenning SP, Parkinson MC, Horwich A, Fosså SD, Wilkinson PM et al. The second medical research council study of prognostic factors in nonseminomatous germ cell tumors. J Clin Oncol 10: 85-94, 1992
16. Tait D, Peckham MJ, Hendry WF, Goldstraw P. Post-chemotherapy surgery in advanced non-seminomatous germ-cell testicular tumours: the significance of histology with particular reference to differentiated (mature) teratoma. Br J Cancer 50: 601-609, 1984
17. Mulders PFA, Oosterhof GON, Boetes C, De Mulder PHM, Theeuwes AGM, Debruyne FMJ. The importance of prognostic factors in the individual treatment of patients with disseminated germ cell tumours. Br J Urol 66: 425-429, 1990
18. Aass N, Klepp O, Cavallin-Ståhl E, Dahl O, Wicklund H, Unsgaard B, et al. Prognostic factors in unselected patients with nonseminomatous metastatic testicular cancer: a multicenter experience. J Clin Oncol 9: 818-826, 1991

19. Newlands ES & Reynolds KW. The role of surgery in metastatic testicular germ cell tumours (GCT). Br J Cancer 59: 837-839, 1989

20. Stoter G, Bosl GJ, Droz JP, Fosså SD, Freedman LS, Geller NL, et al. Prognostic factors in metastatic germ cell tumors. In *EORTC Genitourinary Group Monograph 7: Prostate Cancer and Testicular Cancer,* Newling DWW, Jones WG (eds) pp. 313-319. Wiley-Liss, Inc.: New York, 1990

21. Bajorin DF, Geller NL, Bosl GJ. Assessment of risk in metastatic testis carcinoma: impact on treatment. Urol Int 46: 298-303, 1991

22. Harding MJ, Brown IL, Macpherson SG, Turner MA, Kaye SB. Excision of residual masses after platinum based chemotherapy for non-seminomatous germ cell tumours. Eur J Cancer Clin Oncol 25: 1689-1694, 1989

23. Geller NL, Bosl GJ, Chan EYW. Prognostic factors for relapse after complete response in patients with metastatic germ cell tumors. Cancer 63: 440-445, 1989

5 Predictors of residual mass histology following chemotherapy for metastatic nonseminomatous testicular cancer: a quantitative overview of 996 resections

E.W. Steyerberg, H.J. Keizer, G. Stoter & J.D.F. Habbema.
Eur J Cancer 1994; 30A: 1231-1239

Abstract

Following chemotherapy for metastatic nonseminomatous testicular cancer, surgical resection may demonstrate that residual masses contain purely benign tissue (necrosis), or potentially malignant tissues (histologically viable cancer cells or mature teratoma). The morbidity, mortality and costs of resection demand that resection is based on empirical data rather than on subjective judgments. We reviewed 996 resections from 19 studies to quantify predictors of the histology at resection. Predictors were analyzed for each study and combined in a pooled Odds Ratio (OR). Predictors of necrosis were:

- a teratoma negative primary tumor (OR=5.1)
- normal tumor markers before chemotherapy (AFP: OR=2.8; HCG: OR=1.9; both AFP and HCG: OR=5.7)
- a smaller postchemotherapy abdominal mass (e.g. ≤ 20mm: OR=3.7)
- a large shrinkage (≥ 70%: OR=3.1)
- lung resections versus abdominal resections (OR=1.7)

Cancer was found in only 4% of residual retroperitoneal masses ≤ 20mm. Further research may combine the primary tumor histology, marker level and mass size to improve clinical guidelines, which define subgroups of patients for whom the benefits of resection do not outweigh the risks.

5.1 Introduction

Surgical resection is widely accepted as the treatment of choice in the presence of residual masses following chemotherapy for metastatic testicular nonseminomatous germ cell tumors (NSGCT)[1,2]. Resection provides the histological diagnosis of the residual mass, which may be purely benign with necrotic and/or fibrotic remnants only ('necrosis'), may contain mature teratoma elements ('mature teratoma'), or viable cancer cells/active malignancy ('cancer'). Resection of masses containing necrosis only is assumed to have no therapeutic benefit and is usually not followed by additional treatment. Resection of mature teratoma or cancer is considered to be beneficial as it prevents growth of (potentially) malignant cells[3]. Finally, the presence of viable cancer cells in the residual mass directs the decision to administer additional chemotherapy[4]. The prognosis after resection is generally favorable, with 5 year relapse free survival over 85% after resection of necrosis or mature teratoma[3,5,6,7,8], and between 50%[3,5,6,7] and 80%[8,9,10] after resection of cancer followed by additional chemotherapy. Another aspect of resection is that incompletely resected patients have a poor prognosis[5,8,9].

As the benefit of resection depends on the histology present in the residual mass, attention has been paid to factors associated with the histology at resection [3,11,12,13,14,15,16]. These analyses have focused on groups of patients with a high probability of necrosis, in whom resection might be omitted. In the present study, we analyzed both the probability of necrosis and the probability of cancer, as both are important in the decision to perform a surgical resection. For example, it is clear that a patient with probabilities of 90% necrosis, 1% mature teratoma and 9% cancer should more definitely undergo resection than a patient with probabilities of 90% necrosis, 9% mature teratoma and only 1% cancer, as leaving cancer unresected may be considered more serious than leaving mature teratoma unresected.

Recommendations for resection of abdominal residual masses vary to a considerable extent. For example, the size of the residual mass influences the decision to perform resection, but recommendations vary from laparotomy in any patient with initial abdominal lymph nodes > 3 cm, even if no pathologic mass could be detected on the postchemotherapy CT scan[3], to resection of residual abdominal masses only if they exceed 20mm[6]. Other factors which have been considered for patient selection include the presence of teratoma elements in the primary tumor[11,14,17], the reduction in size of the mass ('shrinkage')[11], and the prechemotherapy level of tumor markers like alpha-fetoprotein (AFP) and human gonadotrophin (HCG)[14].

The associations of these factors with the histology at resection have been observed in relatively small studies. In the present paper, we therefore have combined the data from several published studies to obtain larger numbers and hence more precise estimates of the predictors. Moreover, the published studies differ with respect to the selection of patients and the chemotherapy regimens used. The predictive value of factors may depend on these study characteristics (heterogeneity of effect). This potential heterogeneity is explicitly analyzed in this study. For example, we investigate whether the effects of predictors are different in lung and abdominal resections, or different in more recently published studies, where newer chemotherapy regimens were applied.

5.2 Patients and Methods

5.2.1 Predictors

Associations with the histology at resection ('effects') were quantified using techniques of meta-analysis[18]. The following factors (or predictors) were considered: primary tumor histology, tumor markers before chemotherapy, abdominal mass size, shrinkage during chemotherapy, and type of resection. The primary tumor histology was defined as teratoma-positive or teratoma-negative[3,11], indicating whether mature teratoma elements were present. The highest tumor marker level of AFP and HCG before chemotherapy was classified as elevated or normal, although normal values differed between studies (e.g. AFP < 5 ng/ml[5] or < 20 ng/ml[14,26]). The effects of primary histology and tumor marker level were analyzed in studies including laparotomies or thoracotomies only, or both. In those studies which included laparotomies only, associations with the histology at resection were quantified for pre- and postchemotherapy mass size as measured on CT scan, and for the reduction in size of the mass during chemotherapy ('shrinkage'). A large shrinkage has been defined as a reduction of more than 90% in area[3]. The corresponding reduction in one dimension is around 70% (exact: 68.4%), which was used if only measurements in transversal direction were available. The factor 'type of resection' indicated if abdominal or lung masses were resected. The histology of the resected material was classified according to the worst histologic element, either as cancer, mature teratoma, or necrosis. Thus, 'mature teratoma' refers to masses which contained mature teratoma and possibly also necrosis/fibrosis, but no viable cancer cells, and 'cancer' refers to masses which contained viable cancer cells and possibly mature teratoma and/or necrosis.

5.2.2 Study selection

Studies were selected using MEDLINE medical database and via references in articles. The studies had to contain frequency data on the association of a factor with the histology found at resection, either in tables or mentioned in the text. From some authors additional information on their patient series was obtained (Dr G. Pizzocaro, Milan, Italy, and Dr P.F.A. Mulders, Nijmegen, The Netherlands). Further, data were included from a series of 86 patients resected in three Dutch centers between July 1980 and June 1991. Details on the treatment of these 86 patients were described elsewhere[8]. The data used from each study are listed in the Appendix.

5.2.3 Statistical analysis

The probability of necrosis and the ratio of cancer and mature teratoma were related with factors known before resection (Figure 1). The Odds Ratio (OR) was used as the measure of association in 2 x 2 tables. The OR is the ratio of the odds of necrosis or cancer in one category divided by the odds in the other category. The OR may be interpreted as a relative risk, when the probabilities are small (e.g. < 10%). When the probabilities are larger, the OR is larger than the relative risk. An OR of more than one for a category of patients means that the odds (or risk) is increased compared to the other category. Contrary, an OR smaller than one indicates a lower risk. ORs were

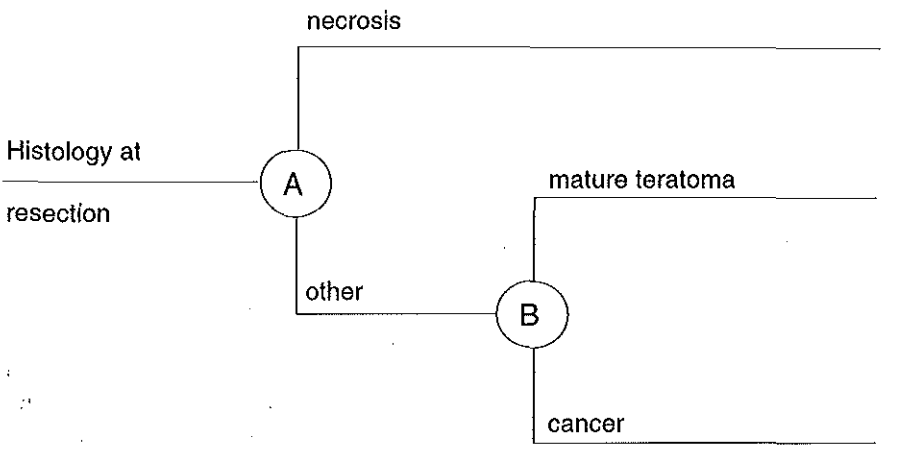

Figure 1 Schematic presentation of the statistical analysis of the histology at resection. A denotes the probability of necrosis at resection; B denotes the ratio of cancer and mature teratoma, or the relative probability of cancer.

calculated within each patient series (study OR) and subsequently pooled (pooled OR) using the Mantel-Haenszel method (StatXact version 2, CYTEL Software Corporation, Cambridge, MA, USA). The 95%-confidence intervals (95%-CI) of the pooled ORs were calculated with the exact variance estimate of Robins et al[19]. In studies with a zero cell frequency the variance and study OR were estimated by the procedure of Peto[20]. Factors have statistically significant effects ($p < 0.05$) if the 95%-CI of the pooled OR does not include one.

Since the studies differed with respect to a number of relevant characteristics, it was investigated whether the effect of the predictors depended on any of these characteristics (heterogeneity of effect[18]). The following study characteristics were considered (see Table 1): the selection of patients (type of resection, markers at resection, size of resected masses), the time of treatment, which is related to the type of chemotherapy regimen used, and study size. The characteristic 'study size' is used to detect publication bias, i.e. the phenomenon that statistically significant results have a higher chance of being published than insignificant results, leading to on average higher effect estimates in smaller studies. The heterogeneity of effect was tested for statistical significance by fitting a weighed linear regression equation of ln(study OR) on the study characteristics[21], where each study OR was weighed by the reciprocal of its variance. If significant heterogeneity existed ($p < 0.10$[18]), the pooled OR was calculated for each category of the study characteristic.

Table 1 Study characteristics considered for heterogeneity of effect of the associations with the histology at resection. The codes of each characteristic are listed with the studies in Figure 2, Table 2 and the Appendix.

Characteristic	Coding
Type of resection	1. abdominal resections only; 2. lung resections only; 3. both abdominal and lung resections or unclear
Markers at resection	0. markers normal before resection; 1. some patients with elevated markers; 2. unclear
Size of the resected masses	0. small masses only (CT normal / < 20mm); 1. larger masses only (CT abnormal / ≥ 20mm); 2. both small and larger masses
Time of treatment	Year of publication
Study size	Number of patients

5.3 Results

The analysis included 901 resections from 18 articles[3,5,9,11,12,13,14,15,16,17,22,23,24,25,26,27,28,29] published between 1983 and 1992 and 95 (75 abdominal, 20 lung) from a Dutch series[8]. Table 2 shows that the overall distribution of the histologies at 996 resections was necrosis in 480 (48%), mature teratoma in 361 (36%) and cancer in 155 (16%).

5.3.1 Probability of necrosis

The relation between the finding of necrosis only in the resected material and the primary tumor histology was described in many publications. Figure 2 depicts the OR for each study with the corresponding 95%-confidence intervals (95%-CI) and the pooled OR with its 95%-CI. The effect of teratoma elements in the primary tumor is consistent in all analyses; no heterogeneity of effect was found in relation to any of the study characteristics. The pooled OR was 5.1 (Figure 2, Table 3), which means that patients without teratoma elements in their primary tumor (teratoma-negative) have more often necrosis in their residual masses (see Appendix: 289/451=64%), compared to patients with teratoma elements in their primary tumor (126/438=29%).

Similarly, the associations of other factors with the finding of necrosis only at resection were summarized. As no heterogeneity of effect was found, one pooled OR is presented for each factor in Table 3. Patients with normal tumor markers AFP or HCG, or both AFP and HCG before chemotherapy had necrosis more often at resection. Smaller abdominal masses before chemotherapy had a higher probability of necrosis ($p > 0.10$). Smaller postchemotherapy masses (normal CT scan or ≤10mm, ≤15mm, ≤20mm, ≤50mm) contained necrosis more often than larger masses (abnormal CT scan or >10mm, >15mm, >20mm, >50mm, resp.). A large shrinkage indicated a higher probability of finding necrosis only. The type of resection was associated with the probability of necrosis: necrosis was found more often at lung resections ($p=0.04$).

Table 2 Distribution of the histology at resection for the studies analyzed. The first column ('Code') indicates the major study characteristics, as shown in Table 1. The studies are ordered according to these study characteristics and year of publication. 'Ref' denotes the reference number, 'N' the total number of resections in the patient series, 'Nec' necrosis, 'Ter' mature teratoma and 'Can' cancer.

Code	Author	Ref	Year	N	Nec		Ter		Can	
100	Bracken et al	27	1983	22	14	64%	3	14%	5	23%
100	Gelderman et al	22	1986	24	17	71%	7	29%	0	0%
100	Fosså et al	13	1989	37	24	65%	12	32%	1	3%
100	Fosså et al	14	1992	76	51	67%	22	29%	3	4%
101	Bracken et al	27	1983	22	8	36%	7	32%	7	32%
102	Stomper et al	15	1985	30	12	40%	10	33%	8	27%
102	Donohue et al	11	1987	80	35	44%	33	41%	12	15%
102	Sagalowsky et al	16	1990	15	6	40%	5	33%	4	27%
102	Mulders et al	26	1990	34	20	59%	11	32%	3	9%
102	Toner et al	3	1990	122	57	47%	48	39%	17	14%
102	Steyerberg et al	8	1993	75	31	41%	30	40%	14	19%
112	Carter et al	23	1987	26	7	27%	10	38%	9	35%
122	Suurmeijer et al	28	1984	33	20	61%	12	36%	1	3%
202	Bracken et al	27	1983	15	5	33%	4	27%	6	40%
202	Mulders et al	26	1990	8	5	63%	2	25%	1	13%
202	Toner et al	3	1990	39	25	64%	10	26%	4	10%
202	Steyerberg et al	8	1993	20	12	60%	6	30%	2	10%
302	Gelderman et al	17	1988	35	15	43%	16	46%	4	11%
302	Harding et al	9	1989	39	17	44%	14	36%	8	21%
312	Garnick et al	29	1983	29	9	31%	13	45%	7	24%
312	Talt et al	5	1984	73	25	34%	32	44%	16	22%
312	Pizzocaro et al	24	1985	34	14	41%	10	29%	10	29%
312	Fosså et al	12	1989	92	47	51%	34	37%	11	12%
322	Dexeus et al	25	1989	16	4	25%	10	63%	2	13%
Total				996	480	48%	361	36%	155	16%

5.3.2 Ratio of cancer and mature teratoma

The probabilities of cancer and mature teratoma have a ratio of 1:2 on average in the studies analyzed. It was investigated which factors were associated with this ratio (Table 3). Heterogeneity of effect was found for the primary histology and the prechemotherapy HCG level in relation to the type of resections included in the study (abdominal/abdominal and lung/lung resections only). For these factors, a pooled OR is presented per category of the study characteristic.

In abdominal resections, the ratio of cancer and mature teratoma did not differ clearly between patients with a teratoma-negative or a teratoma-positive primary tumor (see Appendix: 83:44 vs 108:47). In lung resections, a teratoma-negative primary tumor was strongly related to the finding of cancer at resection. In 37 teratoma-negative patients who underwent lung resection, no mature teratoma (0%) was found and cancer in 9 (24%), compared to 22 (49%) and 4 (9%) respectively in 45 teratoma-positive patients. Further, a normal prechemotherapy level of HCG was associated with a relatively higher risk of cancer in lung resections.

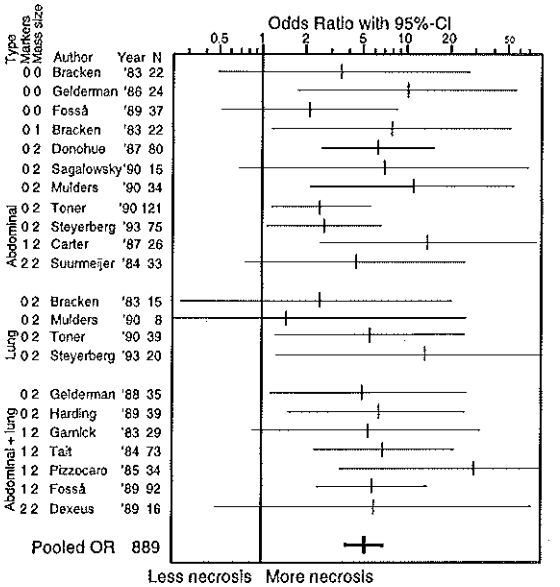

Figure 2 The relation of a teratoma-negative primary tumor and the probability of necrosis at resection. Studies are ordered according to the study characteristics as shown in Table 1.

No significant heterogeneity of effect was found for the other factors (normal AFP before chemotherapy, smaller pre-or postchemotherapy masses, a large shrinkage, lung resection compared to abdominal resection). Smaller pre- or postchemotherapy retroperitoneal masses were statistically significant associated with a lower relative probability of cancer.

5.4 Discussion

This study is the first quantitative overview of predictors for the histology at resection of residual masses in patients with NSGCT. Results from 19 published studies were summarized with statistical techniques. In 996 resections, necrosis was found in 48%, mature teratoma in 36% and cancer in 16%. In a recently published large study[30] the frequency of cancer was comparable (21%). The frequency of necrosis was however somewhat lower (22%) and the frequency of mature teratoma was somewhat higher (57%), probably reflecting the selection policy for resection (masses > 20mm[30]).

Predictors for necrosis did not significantly depend on study characteristics like the selection of patients or the time of treatment. Predictors for cancer, however, appeared to have different effects in lung and abdominal resections. Heterogeneity based on study size was not found, indicating that no major publication bias was present in the articles analyzed.

Table 3 Pooled Odds Ratios with the corresponding 95% confidence intervals for the factors described in the literature combined with own data. An OR > 1 indicates that the probability was higher in the first category of a factor, e.g. the probability of necrosis was higher if the primary histology was teratoma-negative (OR=5.1).

Factor	OR [95%-CI] Necrosis vs other	OR [95%-CI] Cancer vs Teratoma
Primary tumor histology		
Teratoma-negative vs positive	5.1 [3.8 - 6.9]	abdominal: 1.1 [.63 - 1.8] lung+abd.: 2.1 [1.1 - 4.2] lung: 25 [5.6 - 96]
Prechemotherapy markers		
AFP normal vs elevated	2.8 [1.7 - 4.6]	1.04 [0.45 - 2.4] abdominal: .67 [.19 - 2.4]
HCG normal vs elevated	1.9 [1.2 - 3.0]	lung+abd.: 4.3 [1.6 - 12] lung: 5.0 [.15 - 99]
Both AFP and BHCG normal vs one or both elevated	5.7 [2.5 - 13]	2.70 [.62 - 11]
Retroperitoneal mass size		
Prechemotherapy size*:		
≤ 20mm vs > 20mm	1.3 [0.54 - 3.1]	0.29 [.04 - 2.3]
≤ 50mm vs > 50mm	1.6 [0.81 - 3.3]	0.24 [.07 - .83]
Postchemotherapy size*:		
≤ 10mm vs > 10mm	3.6 [2.1 - 5.9]	1.04 [.41 - 2.7]
≤ 15mm vs > 15mm	8.4 [3.2 - 22]	1.45 [.29 - 7.1]
≤ 20mm vs > 20mm	3.7 [2.0 - 6.8]	0.20 [.05 - .77]
≤ 50mm vs > 50mm	4.3 [2.0 - 9.1]	0.53 [.21 - 1.3]
Shrinkage:		
≥ 70% vs < 70%	3.1 [2.0 - 4.8]	1.49 [.67 - 3.2]
Type of resection		
Lung vs abdominal	1.7 [1.0 - 2.7]	1.12 [.53 - 2.4]

* Note that the ORs calculated at the different cut-off points are not totally independent, as the same observations of some studies were used in multiple calculations.

The predictor of necrosis which appeared from this analysis as the most important is the absence of mature teratoma elements in the primary tumor. The fact that this primary histology is related to the finding of necrosis at resection was already evident from many individual studies, but the magnitude of the association could now be estimated more precisely (pooled Odds Ratio: 5.1, 95%-CI: [3.8-6.9]). Secondly, in this meta-analysis normal levels of tumor markers AFP and/or HCG before chemotherapy were clearly associated with a higher probability of necrosis at resection. Note that this association was statistically significant in two out of seven individual studies only ([5,24] and [14,29]). Thirdly, the analysis indicated that smaller residual retroperitoneal masses contained necrosis more often at resection. Associations of similar magnitude were found for several cut-off points (10mm, 15mm, 20mm or 50mm). Finally, other predictors for necrosis were a large shrinkage (≥ 70%) and the location of the mass: resection of pulmonary nodules versus retroperitoneal masses.

Predictors for cancer were not considered explicitly in any previous publication, as it has been argued that both patients with mature teratoma or cancer in a residual mass require resection[3]. However, the consequences of leaving viable cancer cells unresected are not equivalent to the consequences of leaving mature teratoma unresected. Resection of residual cancer is usually followed by two additional chemotherapy courses, which is effective in the majority[8,10] of these patients. A much lower efficacy may be expected in these patients if they relapse, due to development of drug resistance and a more extensive tumor bulk. Although the risks of leaving mature teratoma unresected include a more difficult resection after a rapid growth of the mass and malignant change[3], most patients with mature teratoma may presumably be resected successfully at a later date. Thus, a higher risk of leaving mature teratoma unresected might be accepted compared to the risk of cancer. The probability of cancer therefore requires explicit analysis.

Remarkably, the association of the primary tumor histology with the probability of cancer appeared different in lung and abdominal resections. At lung resection, the absence of maturé teratoma elements in the primary tumor was found related to a higher probability of necrosis, but also to a higher relative probability of cancer. This implies that the risk of mature teratoma is low in lung metastases, if the primary tumor is teratoma-negative. The same observation was made for a normal prechemotherapy level of HCG. The explanation for this association is unclear and might be explained by coincidence, because of the large number of tests for heterogeneity that were performed. Further empirical or histopathological research is therefore required that may confirm or explain for example that no mature teratoma elements are found in lung masses of patients with initially teratoma-negative tumors.

Smaller pre- or postchemotherapy abdominal masses were associated with a lower relative probability of cancer, especially residual masses ≤ 20mm. This is illustrated by the absolute frequencies of the histology at resection in our series[8]: necrosis 57%, mature teratoma 43%, and cancer in 0% of 23 retroperitoneal masses ≤ 20mm after chemotherapy. Combining these data with some recent publications presenting data on the histology found in small residual masses (≤ 15mm[3], < 20mm[14], ≤ 20mm[26]) leads to the following distribution for the total of 155 small masses: necrosis 71%, mature teratoma 25% and cancer 3.9% (6/155). The policy to resect residual abdominal masses only if they exceed 20mm[6,30] thus implies that masses will be left unresected which have a risk of circa 25% of containing mature teratoma, but a risk of only 4% of containing cancer.

This analysis identified no single factor that indicated subgroups with a probability of necrosis only as high as 80% or 90%, nor were factors identified which excluded the finding of residual cancer. Presently, the size of the residual mass is the foremost important factor to select patients for resection[3,6]. This analysis confirms that several other factors might be taken into account when considering resection, especially the absence of teratoma elements in the primary tumor. The association of the primary tumor histology with the finding of cancer at resection may however be markedly different in lung resections. This illustrates that selection guidelines for laparotomy may not apply to thoracotomy.

The combination of factors will possibly allow the definition of subsets of patients in whom resection might be omitted. Fossa[14] found necrosis only in 15 patients with residual masses smaller than 2 cm, with MTU (embryonal carcinoma) in their primary tumor and with both AFP and HCG normal before chemotherapy. In our study group of 75 patients, only 2 fulfilled these criteria and 1 had necrosis and 1 had mature teratoma at laparotomy. Donohue[11] found necrosis only in 15 patients who showed a shrinkage over 90% in volume and without teratoma elements in their primary tumor. Five of our patients met these criteria: 4 had necrosis and 1 had cancer at laparotomy. Thus, attempts to define groups of patients who will not have mature teratoma or cancer based on a few factors are not very reliable sofar and include a small fraction only of the total number of patients with CT scan detected residual masses.

It may be expected that refinement and extension of the number of factors can make more accurate predictions of the histology at resection than the above mentioned studies[11,14]. Therefore, we initiated a collaborative effort with other research groups to perform a multivariate analysis, using primary tumor histology, the levels of tumor markers before chemotherapy (AFP, HCG, LDH[3]), mass size after chemotherapy and shrinkage of the mass to estimate the probability of necrosis, mature teratoma and cancer at resection. These multivariate estimates may guide the decision to resect a residual mass. Optimal treatment is then determined on a more individual basis by weighing the benefits of resecting mature teratoma or cancer against the morbidity, mortality, financial costs and the patient's personal preferences.

The authors would like to thank Dr Liesbeth Bergman, Rotterdam, The Netherlands, for helpful comments; Dr Cees J. van Groeningen, Amsterdam, The Netherlands, for assistance in data collection; Dr G. Pizzocaro, Milan, Italy and Dr Peter EA. Mulders, Nijmegen, The Netherlands, for supplying additional data of their published patient series for the meta-analysis; and Dr Sophie D. Fossa, Oslo, Norway for checking her data.

Appendix

Distribution of the histology at resection according to the predictors analyzed. The first column ('Code') indicates the major study characteristics, as shown in Table 1. The studies are ordered according to these study characteristics and year of publication. 'N' denotes the number of resections, 'Nec' necrosis, 'Ter' mature teratoma and 'Can' cancer.

			Prechemotherapy highest AFP level: AFP normal				**AFP elevated**			
Code	Author	Year	N	Nec	Ter	Can	N	Nec	Ter	Can
100	Fossa et al	1992	35	77%	17%	6%	41	59%	39%	2%
102	Mulders et al	1990	13	77%	23%	0%	19	47%	37%	16%
102	Steyerberg et al	1993	16	63%	25%	13%	59	36%	44%	20%
202	Steyerberg et al	1993	7	71%	14%	14%	13	54%	38%	8%
312	Garnick et al	1983	16	31%	38%	31%	10	30%	40%	30%
312	Tait et al	1984	16	56%	31%	13%	58	28%	45%	28%
312	Pizzocaro et al	1985	10	70%	20%	10%	24	29%	33%	38%
Total			**113**	**65%**	**24%**	**12%**	**224**	**39%**	**41%**	**20%**

Prechemotherapy highest HCG level:			HCG normal				HCG elevated			
Code	Author	Year	N	Nec	Ter	Can	N	Nec	Ter	Can
100	Fosså et al	1992	35	83%	17%	0%	41	54%	39%	7%
102	Mulders et al	1990	11	55%	36%	9%	22	59%	32%	9%
102	Steyerberg et al	1993	23	57%	30%	13%	52	35%	44%	21%
202	Steyerberg et al	1993	6	67%	17%	17%	14	57%	36%	7%
312	Garnick et al	1983	11	55%	18%	27%	18	17%	56%	28%
312	Tait et al	1984	18	33%	22%	44%	55	35%	49%	16%
312	Pizzocaro et al	1985	9	33%	22%	44%	25	44%	32%	24%
Total			**113**	**59%**	**23%**	**18%**	**227**	**41%**	**42%**	**16%**

Prechemo highest AFP and HCG:			AFP & HCG normal				AFP or HCG elevated			
Code	Author	Year	N	Nec	Ter	Can	N	Nec	Ter	Can
100	Fosså et al	1992	24	92%	8%	0%	52	56%	38%	6%
102	Mulders et al	1990	6	67%	33%	0%	27	56%	33%	11%
102	Steyerberg et al	1993	8	75%	13%	13%	67	37%	43%	19%
202	Steyerberg et al	1993	2	100%	0%	0%	18	56%	33%	11%
312	Garnick et al	1983	7	57%	0%	43%	22	23%	59%	18%
312	Pizzocaro et al	1985	3	100%	0%	0%	31	35%	32%	32%
Total			**50**	**82%**	**10%**	**8%**	**217**	**44%**	**40%**	**16%**

Primary tumor histology:			Teratoma-negative				Teratoma-positive			
Code	Author	Year	N	Nec	Ter	Can	N	Nec	Ter	Can
100	Bracken et al	1983	17	71%	6%	24%	5	40%	40%	20%
100	Gelderman et al	1986	10	100%	0%	0%	14	50%	50%	0%
100	Fosså et al	1989	23	74%	22%	4%	14	50%	50%	0%
101	Bracken et al	1983	16	50%	25%	25%	6	0%	50%	50%
102	Donohue et al	1987	48	60%	27%	13%	32	19%	63%	19%
102	Sagalowsky et al	1990	6	67%	17%	17%	9	22%	44%	33%
102	Mulders et al	1990	18	83%	6%	11%	16	31%	63%	6%
102	Toner et al	1990	75	55%	33%	12%	46	33%	50%	17%
102	Steyerberg et al	1993	35	54%	40%	6%	40	30%	40%	30%
112	Carter et al	1987	13	54%	23%	23%	13	0%	54%	46%
122	Suurmeijer et al	1984	11	82%	18%	0%	22	50%	45%	5%
202	Bracken et al	1983	7	43%	0%	57%	8	25%	50%	25%
202	Mulders et al	1990	3	67%	0%	33%	5	60%	40%	0%
202	Toner et al	1990	18	83%	0%	17%	21	48%	48%	5%
202	Steyerberg et al	1993	9	89%	0%	11%	11	36%	55%	9%
302	Gelderman et al	1988	10	70%	10%	20%	25	32%	60%	8%
302	Harding et al	1989	18	67%	22%	11%	21	24%	48%	29%
312	Garnick et al	1983	15	47%	27%	27%	14	14%	64%	21%
312	Tait et al	1984	34	56%	24%	21%	39	15%	62%	23%
312	Pizzocaro et al	1985	19	68%	16%	16%	15	7%	47%	47%
312	Fosså et al	1989	39	74%	15%	10%	53	34%	53%	13%
322	Dexeus et al	1989	7	43%	43%	14%	9	11%	78%	11%
Total			**451**	**64%**	**22%**	**14%**	**438**	**29%**	**53%**	**18%**

Prechemotherapy maximum transversal abdominal lymph node size

Code	Author	Year	Category	N	Nec	Ter	Can
100	Fosså et al	1992	< 20 mm	13	54%	38%	8%
			≥ 20 mm	63	70%	27%	3%
102	Mulders et al	1990	20 - 49 mm	14	71%	29%	0%
			≥ 50 mm	20	50%	35%	15%
102	Steyerberg et al	1993	≤ 20 mm	10	40%	60%	0%
			21 - 49 mm	24	50%	42%	8%
			≥ 50 mm	15	37%	34%	29%
312	Pizzocaro et al	1985	≤ 20 mm	1	100%	0%	0%
			21 - 49 mm	6	33%	33%	33%
			≥ 50mm	27	41%	30%	30%

Postchemotherapy maximum transversal abdominal lymph node size

Code	Author	Year	Category	N	Nec	Ter	Can
100	Fosså et al	1992	0 - 10 mm	49	71%	22%	6%
			11 - 20 mm	27	59%	41%	0%
102	Bracken et al	1983	Clinical complete resp.	22	64%	14%	23%
			Residual mass	22*	36%	32%	32%
102	Stomper et al *	1985	11 - 20 mm	17	59%	35%	6%
			21 - 50 mm	10	50%	30%	20%
			> 50 mm	18	33%	39%	28%
102	Toner et al	1990	0 - 15 mm	39	79%	13%	8%
			> 15 mm	60	32%	48%	20%
102	Mulders et al	1990	0 - 10 mm	9	89%	11%	0%
			11 - 20 mm	8	88%	13%	0%
			21 - 50 mm	11	45%	36%	18%
			> 50 mm	6	0%	83%	17%
102	Steyerberg et al	1993	10 - 20 mm	23	57%	43%	0%
			21 - 50 mm	38	39%	37%	24%
			> 50 mm	14	21%	43%	36%
312	Garnick et al	1983	CT normal	7	43%	43%	14%
			CT abnormal	15	27%	33%	40%
312	Fosså et al	1989	CT normal	34	79%	18%	3%
			CT abnormal	58	34%	48%	17%
312	Pizzocaro et al	1985	0 - 10 mm	8	75%	13%	13%
			11 - 20 mm	4	50%	25%	25%
			21 - 50 mm	7	43%	29%	29%
			> 50 mm	15	20%	40%	40%

Type of resection:			**Lung resection**				**Abdominal resection**			
Code	Author	Year	N	Nec	Ter	Can	N	Nec	Ter	Can
302	Bracken et al	1983	15	33%	27%	40%	22*	36%	31%	32%
302	Mulders et al	1990	8	63%	25%	13%	34	59%	32%	9%
302	Toner et al	1990	39	64%	26%	10%	122	46%	40%	14%
302	Steyerberg et al	1993	20	60%	30%	10%	75	41%	40%	19%
312	Fosså et al	1989	9	56%	33%	11%	92	51%	47%	12%
Total			**91**	**57%**	**27%**	**15%**	**345**	**47%**	**38%**	**15%**

* N indicates in this study the number of residual masses

Shrinkage of the mass:			Shrinkage ≥ 70%				Shrinkage < 70%			
Code	Author	Year	N	Nec	Ter	Can	N	Nec	Ter	Can
100	Fosså et al	1992	34	71%	24%	6%	42	64%	33%	2%
102	Stomper et al *	1985	10	70%	10%	20%	21	57%	29%	14%
102	Donohue et al †	1987	24	71%	25%	4%	56	32%	48%	20%
102	Sagalowsky et al	1990	7	40%	40%	60%	10	40%	30%	30%
102	Mulders et al	1990	12	83%	17%	0%	22	45%	41%	14%
102	Toner et al	1990	25	72%	16%	12%	61	39%	44%	16%
102	Steyerberg et al	1993	11	73%	9%	18%	64	36%	45%	19%
312	Pizzocaro et al	1985	13	69%	15%	15%	21	24%	38%	38%
Total			**134**	**71%**	**19%**	**11%**	**297**	**41%**	**41%**	**17%**

* N indicates the number of residual masses;
† reduction over 90% in volume was used as a criterium for 'large reduction in size'

References

1. Einhorn LH. Treatment of testicular cancer: a new and improved model. J Clin Oncol 1990, 8, 1777-1781.
2. Peckham M. Testicular cancer. Rev Oncol 1988, 1, 439-453.
3. Toner GC, Panicek DM, Heelan RT, et al. Adjunctive surgery after chemotherapy for nonseminomatous germ cell tumors: recommendations for patient selection. J Clin Oncol 1990, 8, 1683-1694.
4. Donohue JP, Rowland RG. The role of surgery in advanced testicular cancer. Cancer 1984, 54, 2716-2721.
5. Tait D, Peckham MJ, Hendry WF, Goldstraw P. Post-chemotherapy surgery in advanced non-seminomatous germ-cell tumours: the significance of histology with particular reference to differentiated (mature) teratoma. Br J Cancer 1984, 50, 601-609.
6. Mead GM, Stenning SP, Parkinson MC, et al. The second medical research council study of prognostic factors in nonseminomatous germ cell tumors. J Clin Oncol 1992, 10, 85-94.
7. Jansen RLH, Sylvester R, Sleyfer DT, et al. Long-term follow-up of non-seminomatous testicular cancer patients with mature teratoma or carcinoma at postchemotherapy surgery. Eur J Cancer 1991, 27, 695-698.
8. Steyerberg EW, Keizer HJ, Zwartendijk J, et al. Prognosis after resection of residual masses following chemotherapy for metastatic nonseminomatous testicular cancer: a multivariate analysis. Br J Cancer 1993, 68, 195-200.
9. Harding MJ, Brown IL, Macpherson SG, et al. Excision of residual masses after platinum based chemotherapy for non-seminomatous germ cell tumours. Eur J Cancer Clin Oncol 1989, 25, 1689-1694.
10. Fox EP, Weathers TD, Williams SD, et al. Outcome analysis for patients with persistent nonteratomous germ cell tumor in postchemotherapy retroperitoneal lymph node dissections. J Clin Oncol 1993, 11, 1294-1299.
11. Donohue JP, Rowland RG, Kopecky K, et al. Correlation of computerized tomographic changes and histological findings in 80 patients having radical retroperitoneal lymph node dissection after chemotherapy for testis cancer. J Urol 1987, 137, 1176-1179.

12. Fosså SD, Aass N, Ous S, et al. Histology of tumor residuals following chemotherapy in patients with advanced nonseminomatous testicular cancer. J Urol 1989, 142, 1239-1242.
13. Fosså SD, Ous S, Lien HH, et al. Post-chemotherapy lymph node histology in radiologically normal patients with metastatic nonseminomatous testicular cancer. J Urol 1989, 141, 557-559.
14. Fosså SD, Qvist H, Stenwig AE, et al. Is postchemotherapy retroperitoneal surgery necessary in patients with nonseminomatous testicular cancer and minimal residual tumor masses? J Clin Oncol 1992, 10, 569-573.
15. Stomper PC, Jochelsen MS, Garnick MB, et al. Residual abdominal masses after chemotherapy for nonseminomatous testicular cancer: correlation of CT and histology. AJR 1985, 145, 743-746.
16. Sagalowsky AI, Ewalt DH, Molberg K, et al. Predictors of residual mass histology after chemotherapy for advanced testis cancer. Urol 1990, 35, 537-542.
17. Gelderman WAH, Schraffordt Koops H, Sleijfer DTh, et al. Results of adjuvant surgery in patients with stage III and IV nonseminomatous testicular tumors after cisplatin-vinblastine-bleomycin chemotherapy. J Surg Oncol 1988, 38, 227-232.
18. Dickersin K, Berlin JA. Meta-analysis: state-of-the-science. Epidemiol Rev 1992, 14, 154-176.
19. Robins JM, Greenland S, Breslow NE. A general estimator for the variance of the Mantel-Haenszel odds ratio. Am J Epidem 1986, 124, 719-723.
20. Yusuf S, Peto R, Lewis J, et al. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Prog Cardiovasc Dis 1985, 27, 335-371.
21. Brand R, Kragt H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. Stat Med 1992, 11, 2077-2082.
22. Gelderman WAH, Schraffordt Koops H, Sleijfer DTh, et al. Treatment of retroperitoneal residual tumor after PVB chemotherapy of nonseminomatous testicular tumors. Cancer 1986, 58, 1418-1421.
23. Carter GE, Lieskovsky G, Skinner DG, et al. Reassessment of the role of adjunctive surgical therapy in the treatment of advanced germ cell tumors. J Urol 1987, 138, 1397-1401.
24. Pizzocaro G, Salvioni R, Pasi M, et al. Early resection of residual tumor during cisplatin, vinblastine, bleomycin combination chemotherapy in stage III and bulky stage II nonseminomatous testicular cancer. Cancer 1985, 56, 249-255.
25. Dexeus FH, Shirkhoda A, Logothetis CJ, et al. Clinical and radiological correlation of retroperitoneal metastasis from nonseminomatous testicular cancer treated with chemotherapy. Eur J Cancer Clin Oncol 1989, 25, 35-43.
26. Mulders PFA, Oosterhof GON, Boetes C, et al. The importance of prognostic factors in the individual treatment of patients with disseminated germ cell tumours. Br J Urol 1990, 66, 425-429.
27. Bracken RB, Johnson DE, Frazier OH, et al. The role of surgery following chemotherapy in stage III germ cell neoplasms. J Urol 1983, 129, 39-43.
28. Suurmeijer AJH, Oosterhuis JW, Sleijfer DTh, et al. Non-seminomatous germ cell tumors of the testis: morphology of retroperitoneal lymph node metastases after chemotherapy. Eur J Cancer Clin Oncol 1984, 20, 727-734.
29. Garnick MB, Canellos GP, Richie JP.Treatment and surgical staging of testicular and primary extragonadal germ cell cancer. JAMA 1983, 250, 1733-1741.
30. Hendry WF, A'Hern RP, Hetherington JW, Peckham MJ, Dearnaley DP, Horwich A. Para-aortic lymphadenectomy after chemotherapy for metastatic non-seminomatous germ cell tumours: prognostic value and therapeutic benefit. Br J Urol 1993, 71, 208-213.

6 Prediction of residual retroperitoneal mass histology following chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups

E.W. Steyerberg, H.J. Keizer, S.D. Fossâ, D.T. Sleijfer, G.C. Toner, H. Schraffordt Koops, P.F.A. Mulders, J.E. Messemer, K. Ney, J.P. Donohue, D.F. Bajorin, G. Stoter, G.J. Bosl and J.D.F. Habbema.
J Clin Oncol 13:1177-1187, 1995

Abstract

Purpose: To develop a statistical model which predicts the histology (necrosis, mature teratoma or cancer) of residual retroperitoneal masses following cisplatin-based induction chemotherapy for metastatic nonseminomatous germ cell tumor.
Patients and Methods: We have collected an international data set comprising individual patient data from six study groups. Logistic regression analysis was used to estimate the probability of necrosis and the ratio of cancer and mature teratoma.
Results: Of 556 patients, 250 (45%) had necrosis at resection, 236 (42%) mature teratoma, and 70 (13%) cancer. Predictors of necrosis were the absence of teratoma elements in the primary tumor, prechemotherapy normal AFP, normal HCG, elevated LDH, a small pre- or postchemotherapy mass, and a large shrinkage of the mass during chemotherapy. Multivariate combination of predictors yielded reliable models (goodness-of-fit tests: $p > .20$), which discriminated necrosis well from other histologies (area under the receiver operating characteristic curve .84), but which discriminated cancer only reasonably from mature teratoma (area .66). Internal and external validation confirmed these findings.
Conclusions: The validated models estimate the histology at resection, especially necrosis, with high accuracy, based on well-known and readily available predictors. The predicted probabilities may help to choose between immediate resection of a residual mass or follow-up, taking into account the expected benefits and risks of resection, feasibility of frequent follow-up, the financial costs and the patient's individual preferences.

6.1 Introduction

Surgical resection is a generally accepted treatment for residual retroperitoneal masses following chemotherapy for metastatic testicular nonseminomatous germ cell tumor (NSGCT)[1,2]. Resection may reveal necrosis/fibrosis, mature teratoma or cancer. As these three histologies are not considered to have a similar necessity of resection, attempts have been made to predict the postchemotherapy histology[3,4,5]. For patients with a very high likelihood of necrosis the risk of leaving mature teratoma or cancer unresected might not be balanced by the disadvantages of resection (morbidity, mortality, financial costs).

The decision to perform surgical resection is especially difficult in patients with small residual masses. In some centers resection is performed if the residual mass exceeds an arbitrarily chosen size (e.g. 20 mm[6], 10 mm[7,8]), arguing that the probability of cancer or mature teratoma is very low in smaller masses. In case no pathologic mass can be detected on the postchemotherapy CT scan, some still advocate resection: if a teratomatous component was present in the primary tumor[9], if a prechemotherapy abdominal lymph node metastasis exceeded 30mm in size[3], or if resection is performed as a principle in all patients[10]. The fraction of resected patients varies in accordance with the selection criteria from around 25% (masses > 20 mm[6]) to 86%[10]. Also, the extent of surgery is debatable: some excise only visible abnormal masses[6], while others perform a more extensive retroperitoneal lymph node dissection[11].

Besides the size of the residual mass, other characteristics may be considered in the decision to resect a residual mass. Previously recognized[12] predictors of necrosis include the absence of teratoma elements in the primary tumor[3,4,5,8,9,13,14,15,16,17,18,19], prechemotherapy tumor marker levels (AFP[5], HCG[5], LDH[3]), prechemotherapy and postchemotherapy mass size[3,5,8,13,18] and shrinkage during chemotherapy[3,4,5,8]. The same predictors may help to distinguish cancer from teratoma. In this analysis our aim is to estimate the probabilities of necrosis, mature teratoma and cancer in residual retroperitoneal masses, based on these well-known and readily available predictors. To obtain a sufficiently high number of patients for statistical analyses, we have collected an international data set comprising data from six study groups.

6.2 Patients and Methods

6.2.1 Inclusion criteria

The international data set consisted of patients with metastatic nonseminomatous testicular cancer, including patients with histologically seminoma and elevated prechemotherapy tumormarkers, who underwent resection of retroperitoneal residual masses after induction chemotherapy with cisplatin-based chemotherapy. Excluded were patients with elevated tumor markers AFP or HCG at the time of surgery, patients with extragonadal tumors, patients with pure seminoma (normal prechemotherapy AFP and HCG) and patients resected after relapse of tumor following initial chemotherapy.

6.2.2 Study descriptions

Individual patient data were retrieved according to a data form, which included basic patient identification, year and type of treatment, histology, and information on predictors: presence of teratoma elements in the primary tumor, prechemotherapy tumor marker levels (AFP, HCG, LDH), and pre- and postchemotherapy mass size. Consistency of the data was checked with the participants (agreement with published figures, completion of missing values as far as possible). The studies are numbered 1 to 6 (Table 1) and briefly described in the following, while details of treatment and resection policy can be found in the original publications. Study #1[3] included 122 patients who all fulfilled the inclusion criteria. Study #2[5,13,14] included 149 patients of whom 22 were excluded (11 extragonadal tumors, 10 elevated postchemotherapy markers (AFP/HCG) and 1 both characteristics). Study #3[9,15,20,21] contained 137 patients. Study #4[8] included 49 patients, of whom 15 were excluded (11 pure seminomas, 4 elevated markers before surgery). Study #5[4] contained 80 cases with initial stage B3 disease (palpable prechemotherapy mass > 10 cm). Excluded were 17 patients who underwent salvage chemotherapy for recurrent disease, 5 with pure seminoma and 1 with elevated postchemotherapy AFP. Of the 51 remaining patients, 50 missed the prechemotherapy LDH value. Study #6[7] included 85 Dutch patients, all fulfilling the inclusion criteria. All European centers participated in consecutive EORTC or MRC trials. In total 556 patients were included in the analysis.

Table 1 Characteristics of the six participating study groups.

#	Principal investigator	Reference	Study group *	N
1	Toner	3	MSKCC, New York	122
2	Fossá	5,13,14	Norwegian Radium Hospital, Oslo	127
3	Sleijfer	9,15,20,21	UH, Groningen	137
4	Mulders	8	UH, Nijmegen	34
5	Donohue	4	Indiana university, Indianapolis	51
6	Steyerberg	7	UH, Leiden; RCI, Rotterdam; FU, Amsterdam	85

* MSKCC: Memorial Sloan-Kettering Cancer Center; UH: university hospital; RCI: Rotterdam Cancer Institute; FU: Free University

6.2.3 Definitions of predictors and histology

The primary tumor histology was defined as teratoma-positive or teratoma-negative[3,4], indicating whether teratomatous elements were present. The tumor marker levels of AFP, HCG and LDH before chemotherapy were classified as elevated or normal using the normal values of each center. Higher cut-off points and transformations (square root, log) were evaluated[22] for the absolute values of AFP and HCG, and for standardized values of LDH (LDH value divided by normal value per study). Prechemotherapy and

postchemotherapy mass size were measured in transversal direction on CT scan. Shrinkage was calculated as the percentage reduction in size: 100•(Presize-Postsize) /Presize. Shrinkage was 100% if the postchemotherapy CT scan did not show any residual mass. Various cut-off values and transformations (square root, log) were assessed[22].

The histology of the resected material was classified according to the worst histologic element, either as cancer, mature teratoma, or necrosis. Thus, 'mature teratoma' refers to masses which contained mature teratoma and possibly also necrosis/fibrosis, but no viable cancer cells, and 'cancer' refers to masses which contained viable cancer cells and possibly mature teratoma and/or necrosis.

6.2.4 Missing values

Of the 556 included patients, 429 had complete values for all predictors. Instead of discarding the 127 patients with any missing value, missing values were filled in (or *imputed*) in 115 patients missing one single value, discarding only 12 with two or more missing values. Imputation was based on the correlation between each variable with missing values and the other predictors[23]. The correlation was estimated from the 429 complete cases. The method is described in detail in the Appendix. The results of this analysis were compared with the results obtained when complete cases only were considered.

6.2.5 Statistical analysis

The histology at resection (necrosis, mature teratoma, cancer) was predicted using two statistical models. The first model estimated the probability of necrosis by comparing patients with necrosis at resection with patients showing other histologies (teratoma or cancer). The second model aimed to distinguish between cancer and teratoma in the patients who did not have necrosis at resection. This second analysis estimated the ratio of cancer and teratoma or the *relative* probability of cancer. The use of these two models agrees with the clinical notion that the probability of necrosis is of predominant importance for the decision to resect a residual mass[3] and that the ratio of cancer and teratoma is a second consideration.

The probability of necrosis and the relative probability of cancer were related to factors known before resection (predictors). The Odds Ratio (OR) was used as the measure of association. Relations between predictors and outcomes were first estimated univariately within each study. If a test for homogeneity indicated no major heterogeneity ($p>.10$) of the relations, the data were pooled using the Mantel-Haenszel method (EGRET statistical package[24]). Predictors have statistically significant effects ($p<0.05$) if the 95%-CI of the OR does not include the value one. Multivariate logistic regression analysis was applied to estimate the probability of necrosis and the relative probability of cancer based on the combination of predictors. As the aim of this analysis is prediction, all variables contributing information should preferably be included in the models[25]. The multivariate analyses therefore included all predictors with p-values below 0.50 in the univariate analysis. Of the three variables related to mass size (prechemo-

therapy size, postchemotherapy size and shrinkage), the postchemotherapy size and shrinkage were candidates for use in the multivariate models. It was checked if the coefficients of the predictors differed between the studies by adding interaction terms of the predictors and study number. Also, it was checked if the relations were constant in time by adding interaction terms of the predictors and year of treatment.

6.2.6 Evaluation of model performance

Predictive accuracy of the multivariate models can be distinguished in reliability (or calibration) and discrimination. Reliability refers to the amount of agreement between predicted and observed outcomes. If, for instance, patients with certain characteristics are predicted to have a 70% chance of necrosis at resection, then 70% of such patients should actually have necrosis at resection. A graphical impression of reliability was obtained by plotting observed frequencies of the outcome (necrosis/cancer) against predicted probabilities. Reliability was tested by the Hosmer-Lemeshow goodness-of-fit test[26] (BMDP module LR[27]), which evaluates the correspondence between a model's predicted probabilities and the observed frequencies over groups spanning the entire range of probabilities.

Discrimination was assessed using receiver operating characteristic (ROC) analysis. The ROC curve is a plot of the true positive rate (sensitivity) versus the false positive rate (1 -specificity) evaluated at consecutive cut-off points of the predicted probability. The area under the ROC curve forms a suitable single number to summarize the discriminative ability of a predictive model[28,29]. The area represents, for all possible pairs of patients, the proportion in which the patients with that outcome (necrosis/cancer) had a higher probability than the patients without the outcome. A useless predictive model, such as a coin flip, would yield an area of 0.5. When the area is 1.0, the model discriminates perfectly. For our prediction problem, a value over 0.6 may be interpreted as reasonably, over 0.7 as satisfactory and over 0.8 as good with respect to discriminative ability.

Validity of model performance was distinguished in internal and external validity. Internal validity indicates if the results of the analysis hold for the data under study. Internal validity was assessed with bootstrapping techniques[30]. Random bootstrap samples were drawn with replacement from the full sample consisting of all patients (200 replications). Models were estimated on these bootstrap samples and evaluated on the full sample. In this way, the discriminative ability of the models in future but similar patients is estimated. Moreover, bootstrap estimates were used to derive the final predictive models by correcting the logistic regression coefficients for overoptimism[31].

External validity refers to the validity of the results of this analysis when applied to patients in other centers. To assess external validity, each study was left out of the full sample once. The models were fit on the remaining studies. Discriminative power was tested on the study not included in the fitting procedure (test sample).

Table 2 Patient characteristics per study group.

	#1 N=122	#2 N=127	#3 N=137	#4 N=34	#5 N=51	#6 N=85	TOTAL N=556
Primary tumor histology							
Teratoma-positive	46 (38%)	76 (60%)	84 (61%)	16 (47%)	26 (51%)	45 (53%)	293/555 (53%)
Prechemotherapy markers							
AFP elevated	72 (59%)	73 (58%)	97 (71%)	19 (59%)	33 (72%)	65 (77%)	359/548 (66%)
median (ng/ml)	134	44	48	28	-	121	69
HCG elevated	67 (55%)	71 (56%)	91 (66%)	22 (67%)	34 (74%)	58 (68%)	343/550 (62%)
median (IU/l)	2	19	5	15	-	42	11
LDH elevated	93 (76%)	86 (68%)	93 (68%)	17 (52%)	1 (-)	32 (71%)	322/465 (69%)
median (U/l)	357	496	285	542	-	203	399
Prechemotherapy mass size							
0 - 20 mm	9 (8%)	21 (17%)	18 (13%)	1 (3%)	- (0%)	12 (13%)	61/542 (11%)
21 - 50 mm	61 (54%)	61 (48%)	68 (50%)	13 (38%)	- (0%)	34 (40%)	237/542 (44%)
51 - 100 mm	33 (29%)	34 (27%)	40 (29%)	16 (47%)	10 (21%)	33 (39%)	166/542 (31%)
> 100 mm	9 (8%)	11 (9%)	11 (8%)	4 (12%)	37 (79%)	6 (7%)	78/542 (14%)
Postchemotherapy mass size							
0 - 10 mm	30 (30%)	54 (43%)	55 (40%)	9 (27%)	- (0%)	14 (17%)	162/532 (31%)
11 - 20 mm	19 (19%)	31 (24%)	25 (18%)	8 (24%)	- (0%)	30 (36%)	113/532 (21%)
21 - 50 mm	32 (32%)	26 (21%)	41 (30%)	11 (32%)	- (0%)	23 (27%)	133/532 (25%)
51 - 100 mm	17 (17%)	12 (9%)	14 (10%)	4 (12%)	8 (16%)	15 (18%)	70/532 (13%)
> 100 mm	1 (1%)	4 (3%)	2 (2%)	2 (6%)	43 (84%)	2 (2%)	54/532 (10%)
Shrinkage							
>= 70%	30 (33%)	49 (39%)	50 (37%)	12 (35%)	- (0%)	14 (17%)	155/521 (30%)
50 - 69%	21 (23%)	35 (28%)	45 (33%)	9 (27%)	- (0%)	28 (33%)	138/521 (27%)
30 - 49%	17 (19%)	15 (12%)	8 (6%)	4 (12%)	3 (6%)	15 (18%)	62/521 (12%)
0 - 29%	13 (14%)	23 (18%)	19 (14%)	5 (15%)	35 (74%)	25 (30%)	120/521 (23%)
< 0% (increase)	11 (12%)	5 (4%)	15 (11%)	4 (12%)	9 (19%)	2 (2%)	46/521 (9%)
Year of treatment							
1975 - 1980	16 (13%)	7 (6%)	22 (16%)	3 (9%)	11 (22%)	3 (4%)	62/556 (11%)
1981 - 1985	83 (68%)	67 (53%)	48 (35%)	21 (62%)	34 (67%)	32 (38%)	285/556 (51%)
1986 - 1993	23 (19%)	53 (42%)	67 (49%)	10 (29%)	6 (12%)	50 (59%)	209/556 (38%)
Histology at resection							
Necrosis	57 (47%)	66 (52%)	61 (45%)	20 (59%)	10 (20%)	36 (42%)	250/556 (45%)
Mature teratoma	48 (39%)	51 (40%)	70 (51%)	11 (32%)	23 (45%)	33 (39%)	236/556 (42%)
Cancer	17 (14%)	10 (8%)	6 (4%)	3 (9%)	18 (35%)	16 (19%)	70/556 (13%)

Table 3 Relations of predictors with the histology at resection.

		Necrosis N=250 (45%)	Teratoma N=236 (42%)	Cancer N=70 (13%)	OR Necrosis vs Other	p-value	OR Cancer vs Teratoma	p-value
Primary tumor histology								
Teratoma-negative		155 (60%)	78 (30%)	29 (11%)	3.35 [2.3-5.0]	p < .001	-#	-#
Teratoma-positive		94 (32%)	158 (54%)	41 (14%)				
Prechemotherapy markers								
AFP	normal	116 (61%)	56 (30%)	17 (9%)	2.74 [1.9-4.1]	p < .001	1.05 [.52-2.1]	p = .87
	elevated	130 (36%)	176 (49%)	53 (15%)				
HCG	normal	119 (58%)	67 (32%)	21 (10%)	2.17 [1.5-3.2]	p < .001	1.17 [.61-2.3]	p = .60
	elevated	128 (37%)	166 (48%)	49 (14%)				
LDH	elevated	165 (51%)	120 (37%)	37 (12%)	1.69 [1.2-2.7]	p = .011	2.62 [1.1-6.4]	p = .020
	normal	56 (39%)	78 (55%)	9 (6%)				
Prechemotherapy mass size						Trend:		Trend:
0 - 20 mm		35 (57%)	23 (38%)	3 (5%)	1.0 (rc)*	p = .008	1.0 (rc)*	p = .16
21 - 50 mm		120 (51%)	98 (41%)	19 (8%)	.76 [.41-1.4]		1.32 [.31-6.5]	
51 - 100 mm		66 (40%)	74 (45%)	26 (16%)	.51 [.29-1.1]		2.21 [?-?]$	
> 100 mm		24 (31%)	36 (46%)	18 (23%)	.34 [?-?]$		1.87 [?-?]$	
Postchemotherapy mass size						Trend:		Trend:
0 - 10 mm		117 (72%)	38 (24%)	7 (4%)	1.0 (rc)*	p < .001	1.0 (rc)*	p = .46
11 - 20 mm		62 (55%)	43 (38%)	8 (7%)	.45 [.26-.80]		1.00 [.25-4.0]	
21 - 50 mm		42 (32%)	69 (52%)	22 (17%)	.17 [.09-.29]		1.27 [.42-4.0]	
51 - 100 mm		11 (16%)	47 (67%)	12 (17%)	.05 [.03-.17]		0.99 [.38-5.1]	
> 100 mm		10 (19%)	25 (46%)	19 (35%)	.08 [?-?]$		2.96 [?-?]$	
Shrinkage						Trend:		Trend:
>= 70%		114 (74%)	33 (21%)	8 (5%)	1.0 (rc)*	p < .001	1.0 (rc)*	p = .19
50 - 69%		72 (52%)	51 (37%)	15 (11%)	.42 [.25-.72]		1.03 [.32-3.4]	
30 - 49%		28 (45%)	26 (42%)	8 (13%)	.24 [.14-.59]		.69 [.15-3.0]	
0 - 29%		26 (22%)	69 (58%)	25 (21%)	.09 [?-?]$		.66 [?-?]$	
< 0% (increase)		- (0%)	38 (83%)	8 (17%)	.01 [?-?]$		.35 [?-?]$	

The Odds Ratios were significantly heterogenous between the 6 studies (p=.016)
* rc: reference category
$ The 95%-confidence intervals could not be calculated because of empty cells in some studies

6.3 Results

6.3.1 Patient characteristics

Table 2 shows the distribution of patient characteristics in each of the 6 study groups. Overall, half of the patients had a teratoma-positive primary tumor histology (53%). Tumor markers AFP, HCG and LDH were elevated before chemotherapy in about two thirds of all patients (66%, 62% and 69%). Half of the patients had a prechemotherapy mass size ≤ 50 mm (55%), a postchemotherapy mass size ≤ 20 mm (52%), or a shrinkage in mass size during chemotherapy ≥ 50% (56%). A minority of the patients was treated before 1981 (11%). The histology at resection was 45% necrosis, 42% mature teratoma and 13% cancer. The relative probability of cancer was 70/(70+236)=23% on average. Study #5 contained patients with larger masses compared to the other studies, with less shrinkage during chemotherapy and with necrosis in 20% only and cancer in 35%.

6.3.2 Univariate analysis

Table 3 shows the results of the univariate analyses. Odds Ratios (ORs) for necrosis were reasonably homogenous between studies: all tests for homogeneity had p-values >0.15. All predictors had significant relations with the finding of necrosis at resection. Patients with a teratoma-negative primary tumor histology had necrosis at resection in 60% of the cases, compared to 32% in patients with teratoma-positive primary tumor histology. Prechemotherapy tumormarkers were related to the finding of necrosis at resection: normal AFP, normal HCG, or *elevated* LDH. Smaller pre- or postchemotherapy masses contained necrosis more often, as well as masses that reduced largely in size during chemotherapy (large shrinkage). An increase in mass size during chemotherapy precluded the finding of necrosis (0 out of 46 patients).

Cancer could be distinguished from teratoma by a higher prechemotherapy LDH level (Table 3, p=.020). The OR of the primary tumor histology was significantly heterogenous (p=.016): The OR was larger than one in studies #1 to #4 and smaller than one in study #5 and #6. Therefore the primary tumor histology cannot be used to distinguish between cancer and teratoma. Prechemotherapy AFP and HCG were excluded as predictors in the multivariate models (p>.50). Note that cancer was found in 4% (7/162) only of the patients with residual masses ≤ 10 mm and in 5% (8/155) of the patients with a shrinkage ≥ 70%.

6.3.3 Multivariate analysis

The multivariate model for necrosis included 544 patients, where 115/3264 (3.5%) of the values were imputed (Table 4). All predictors for necrosis were significant ($p<.003$) as well as the multivariate model as a whole ($p<.0001$). Dichotomous characteristics (present/not present) which predict necrosis were a teratoma-negative primary tumor histology, normal AFP and normal HCG. Three other predictors were used as continuous variables. The natural logarithm (ln) appeared to be the optimal transformation of standardized LDH (LDH_{st}). The square root (sqrt) was taken of the residual mass size. Shrinkage of the mass during chemotherapy was used as a continuous, untransformed variable and the OR was calculated per 10% decrease. Thus, a decrease of 10% indicates a circa 1.2 times higher probability of necrosis compared to no decrease, simultaneously adjusting for the other predictors.

The multivariate model for the relative probability of cancer included 299 patients (77/897=8.6% missing values imputed). Prechemotherapy LDH, postchemotherapy mass size and shrinkage were used as predictors, resulting in a significant model ($p=.003$). The relations of the predictors did not vary significantly with study nor year of treatment in both multivariate models ($p>.10$).

Table 4 Results of the logistic regression analysis.

	Necrosis vs Other N=544	Cancer vs Teratoma N=299
Primary tumour histology		
Teratoma-negative vs positive	2.46 [1.6 - 3.7]	-
Prechemotherapy markers		
AFP normal vs elevated	2.49 [1.6 - 3.9]	-
HCG normal vs elevated	2.22 [1.4 - 3.5]	-
LDH: $\ln(LDH_{st})$*	2.76 [1.8 - 4.2]	1.58 [.93 - 2.7]
Postchemotherapy mass size		
Sqrt(transversal diameter)*	.744 [.63 - .87]	1.17 [.99 - 1.4]
Shrinkage		
Per 10% decrease*	1.17 [1.1 - 1.3]	1.06 [.95 - 1.2]

* Continuous predictors

6.3.4 Model evaluation

Reliability of the multivariate models is shown in Figure 1. Overall, the correspondence between observed and expected probabilities is good. In patients with a predicted probability of necrosis over 80%, the observed probability was 66/77=86%. The goodness-of-fit tests indicated no lack of fit of the models (necrosis $p=.59$, cancer $p=.34$). Discriminative ability of the multivariate models is shown in Figure 2. Clearly, discrimination of necrosis from other histology is much better feasible than discrimination of cancer from teratoma.

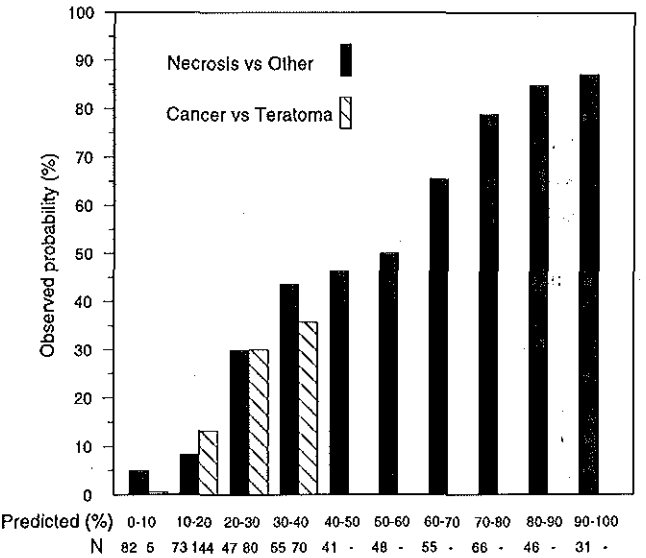

Figure 1 Reliability of the models predicting necrosis and distinguishing cancer from teratoma. Number of patients in each group of the predicted probability is listed after N.

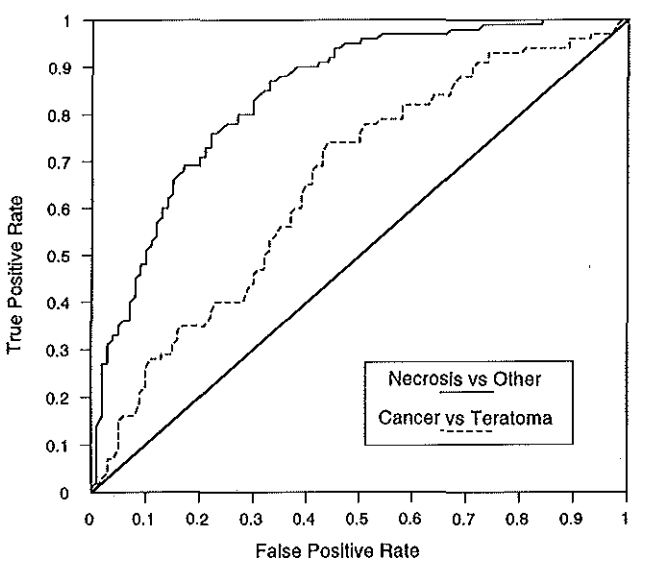

Figure 2 ROC curves of the models predicting necrosis and distinguishing cancer from teratoma, indicating discriminative ability. The areas under the curves are 0.839 and 0.661.

Validity of the models is shown in Table 5. Assessment of internal validity indicates that overall discriminative ability of the model for necrosis is expected to be good in similar patients (area=.83). The models for cancer will have less discriminative ability (area=.66). The external validation procedure mimics the situation that one of the studies would not have been included in the analysis. Table 5 shows that the necrosis model discriminates well in all studies, with some higher discrimination in study #4 and some lower discrimination in study #6. The second model discriminates reasonably in study #3 to #6.

Table 5 Internal and external validation of the models predicting necrosis and distinguishing cancer from teratoma (areas under the ROC curve).

	Necrosis vs Other	Cancer vs Teratoma
Internal validation		
Bootstrapping	.83 (N=544)	.65 (N=299)
External validation		
Study #1	.83 (N=121)	.58 (N=65)
Study #2	.85 (N=127)	.55 (N=61)
Study #3	.84 (N=137)	.60 (N=76)
Study #4	.91 (N=33)	.53 (N=14)
Study #5	.85 (N=42)	.67 (N=34)
Study #6	.75 (N=84)	.64 (N=49)

When the multivariate analysis is restricted to the patients with complete values only, the necrosis model remains very similar with respect to both the ORs and model performance. The model distinguishing between cancer and teratoma has a smaller OR for LDH in the complete case analysis (OR 1.50 vs OR 1.58, Table 4) and its discriminative ability is inferior to the model with some imputed missing values (area≈.60 vs area≈.65).

6.3.5 Practical application

Finally, the two multivariate models are presented in a prognostic score chart (Table 6). This score chart is intended to facilitate the estimation of the probabilities of necrosis, mature teratoma and cancer at resection in clinical practice, using the final models. Scores for each predictor were derived from the logistic regression coefficients, which were reduced by a correction for overoptimism (multiplied by 0.955 in the necrosis model and by 0.870 in the cancer model), and subsequently multiplied by 10 and rounded to whole numbers. Ten points on the score chart correspond to an OR of $e^1 = 2.72$; an OR of two (doubling of the odds) is obtained by a score of 7 points. Values for continuous predictors are given with such intervals that the scores show small steps, but scores for intermediate values may well be estimated by linear interpolation.

Table 6 Prognostic score chart for the probability of necrosis and the relative probability of cancer at resection of residual masses in NSGCT patients with normal tumormarkers AFP and HCG before resection.

Predictor	Value	Necrosis	Cancer
Primary tumour histology			
Presence of teratoma elements	If negative	+9	-
Prechemotherapy markers			
AFP	If normal	+9	-
HCG	If normal	+8	-
LDH_{st} (= LDH / normal value)	0.6	-5	-2
	0.8	-2	-1
	1.0	0	0
	1.2	+2	+1
	1.5	+4	+2
	2.0	+7	+3
	3.0	+11	+4
	4.5	+15	+6
	7.0	+19	+8
Postchemotherapy mass size			
Transversal diameter	2 mm*	-4	+2
	5 mm	-6	+3
	10 mm	-9	+4
	15 mm	-11	+5
	20 mm	-13	+6
	30 mm	-16	+7
	50 mm	-20	+10
	70 mm	-24	+11
	100 mm	-28	+14
Shrinkage			
100•(Presize-Postsize)/Presize	-50%	-7	-3
	0%	0	0
	50%	+7	+3
	75%	+11	+4
	100%	+15	+5
Constant		-10	-24
Sumscore: add relevant scores**		.....	.....

* If no mass is detectable on the postchemotherapy CT scan, a size of 2 mm is assumed.

** The exact formulas to calculate the sumscores are:

Sumscore(Necrosis): -9.78 + 8.58•'teratoma-negative'+ 8.70•'AFPnormal'+ 7.61•'HCGnormal' + 9.69•ln(LDH_{st}) - 2.83•Sqrt(postsize) + 0.147•shrinkage;

Sumscore(Cancer): -24.18 + 3.95•ln(LDH_{st}) + 1.36•Sqrt(postsize) + 0.053•shrinkage,

where the variables 'teratoma-negative', 'AFPnormal' and 'HCGnormal' are 1 if true, 0 if false, ln(LDH_{st}) is the natural logarithm of LDH/normal value, postsize is expressed in mm and shrinkage is expressed as %.

The corresponding probabilities are calculated with the formulas:

Probability(Necrosis): $1 / [1+e^{-(\text{Sumscore(Necrosis)}/10)}]$

Probability(Cancer): [1 - Probability(Necrosis)] • $[1/(1+e^{-(\text{Sumscore(Cancer)}/10)})]$

Probability(Teratoma): 1 - [Probability(Necrosis)+Probability(Cancer)]

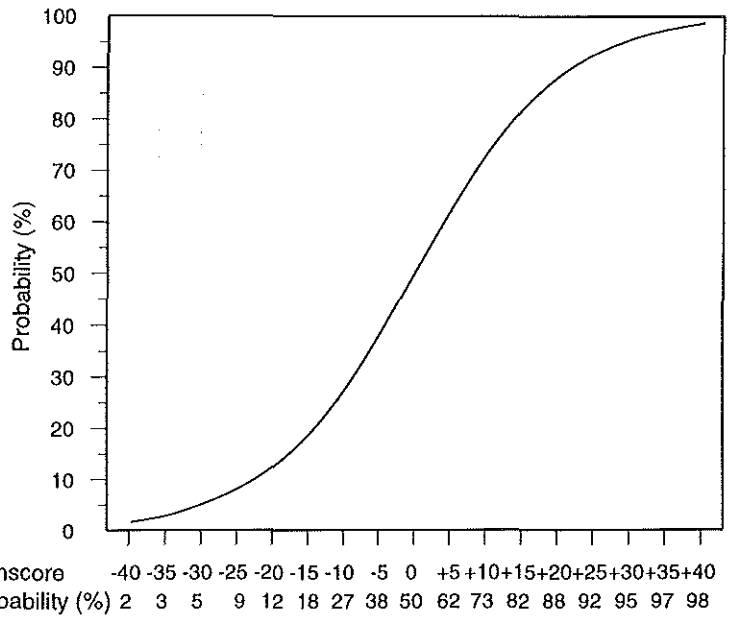

Figure 3 Predicted probabilities corresponding to the sumscores as calculated with the prognostic score chart (Table 6; see text). For example, a sumscore of +15 corresponds to a probability of 82%.

For individual patients, the scores corresponding to the values of the predictors can be filled in on the score chart. An individual sumscore consists of the sum of all scores and a constant which represents the score if all values were zero. Figure 3 shows the probabilities corresponding to this sumscore.

The use of the score chart is illustrated with two patients (Table 7). Patient 1 had a teratoma-negative primary tumor, normal AFP and normal HCG before chemotherapy, LDH three times the normal level, a residual mass of 10 mm, which measured 50 mm before chemotherapy (shrinkage (50–10)/50 = 80%). Patient 2 had a teratoma-positive primary tumor, elevated AFP, but other characteristics as patient 1. The sumscores for patient 1 are +30 and -12 (necrosis and cancer respectively), and for patient 2 +12 and -12. The corresponding probabilities can be read from Figure 3. Patient 1 has a probability of necrosis of 95%, leaving around 5% for the sum of the probabilities of teratoma and cancer. His relative probability of cancer is 27% (Figure 3). This means that the absolute probability of cancer is around 27%•5% ≈ 1.35%, and complementary, the probability of teratoma is around 3.65%. Alternatively, the exact probabilities can be calculated using the formulas below Table 6: necrosis: 94.6%, teratoma 4.1%, cancer 1.3%. The probabilities for patient 2 are 75.5%, 18.5% and 6.0% respectively.

Table 7 Illustration of the application of the prognostic score chart in two hypothetical patients.

	pt 1	pt 2	nec_1	nec_2	can_1	can_2
Primary tumour histology						
Presence of teratoma elements	neg.	pos.	+9	0	-	-
Prechemotherapy markers						
AFP	normal	elev.	+9	0	-	-
HCG	normal	normal	+8	+8	-	-
LDH_{st} (=LDH/normal value)	3.0	3.0	+11	+11	+4	+4
Postchemotherapy mass size						
Transversal diameter	10 mm	10 mm	-9	-9	+4	+4
Shrinkage						
100•(Presize-Postsize)/Presize	80%	80%	+12	+12	+4	+4
Constant			-10	-10	-24	-24
Sumscore: add relevant scores			+30	+12	-12	-12
Probabilities						
Using Figure 3			95%	≈ 75%	≈ 25%	≈ 25%
Using formulas						
Necrosis	94.6%	75.5%				
Teratoma	4.1%	18.5%				
Cancer	1.3%	6.0%				

6.4 Discussion

We developed two models to predict the histology of residual retroperitoneal masses in patients who were treated with cisplatin-based chemotherapy for metastatic NSGCT and who obtained normal values of AFP and HCG before resection. The first model aimed to predict the finding of necrosis only, while the second model was developed to separate cancer from mature teratoma in patients without necrosis. Individual patient data from six study groups [3,4,5,7,9,13,14,15,20,21] were available, providing the largest data set of this type of patients thus far.

The model predicting necrosis consisted of six predictors, which were highly significant when analyzed alone (univariately) or combined (multivariately). Predictors for necrosis were a teratoma-negative primary tumor, normal prechemotherapy AFP or HCG, elevated prechemotherapy LDH, a relatively small residual mass, and a large shrinkage of the mass during chemotherapy. The model distinguishing cancer from teratoma consisted of three predictors, which together constituted a statistically significant multivariate model. Cancer was found relatively more often if prechemotherapy LDH was elevated, the residual mass was larger, and if a large shrinkage of the mass occurred during chemotherapy.

It thus appears that higher values of prechemotherapy LDH are related with a higher probability of necrosis but also with a higher relative probability of cancer. This implies that the probability of mature teratoma decreases with higher prechemotherapy LDH values. The absolute probability of cancer is lowered by higher LDH values in most

patients, as the multivariate Odds Ratio of LDH for cancer is smaller than for necrosis (Table 4). In agreement with previous statements[3], a prechemotherapy higher LDH thus is a fortuitous prognostic sign in patients with normal tumormarkers AFP and HCG after chemotherapy. On the other hand, higher LDH has been found to indicate a lower probability of complete response[32,33] and a worse survival[10] in patients with metastatic disease, analyzing from the start of primary chemotherapy. It may therefore be postulated that patients with residual masses and responding to chemotherapy (as indicated by normalization of AFP/HCG) form a favorable subgroup of patients with prechemotherapy high LDH.

The two models were reliable, which means that the observed probabilities agreed with predicted probabilities. Internal and external validation procedures showed that necrosis could well be discriminated from other histologies, but that cancer could only reasonably be distinguished from teratoma. It should be realized that the probability of cancer strongly depends on the probability of necrosis, as estimated with the first model. For example, if the estimated probability of necrosis is very high, e.g. 95%, the probability of cancer can never be higher than 5%. If we use the average relative probability of cancer (23%) instead of the second model, this results in a risk of cancer of only 23%•5%=1.2%. Thus, even when no predictors for the ratio of cancer and mature teratoma are used, the risk of cancer will be very low once the probability of necrosis is high.

The observation that mature teratoma can difficultly be distinguished from cancer may partly be explained by the fact that the histological distinction between cancer and mature teratoma is less clearly made than the distinction between purely benign tissue (necrosis/fibrosis) and other tissue. For example, one study[19] described that the histological classification changed in a substantial proportion of the patients, when reviewed with the Indiana criteria[34] for the diagnosis of 'cancer': in 5 patients (12%), the diagnosis changed from cancer to atypia and in 3 (7%) from other diagnosis to cancer[19].

The final models were presented in a prognostic score chart, which use was illustrated in two patients, both with residual masses of 10 mm. Taking into account all prognostic factors, the probability of necrosis was very high in the first patient (95%), and somewhat lower (76%) in the second patient. The probabilities for cancer and teratoma were 1.3% and 4.1%, and 6% and 18% respectively. Should both patients undergo laparotomy? Or more in general, can we define thresholds for the decision to resect a residual mass? Should resection not be performed if the probability of necrosis exceeds e.g. 80%, or is the threshold as high as 90%? And what should we accept as the risk of cancer?

These thresholds need to be determined by the balance between the expected benefits and risks of resection. The risks of resection include surgery caused long-term morbidity, especially ejaculation problems[6,20,35,36], which depend on the size and location of the residual mass and the extent of surgery[37]. Short-term morbidity consists of hospital stay itself, and complications like hemorrhage, renal failure and lymphocele[36]. The mortality of laparotomy is low, presumably below 1% in most experienced centers, but risk estimates may sometimes be higher for individual patients.

If mature teratoma or cancer is present in the residual mass, the patient is expected to benefit from resection. The prognosis then is generally favorable, with 5 year relapse free survival over 85% after resection of mature teratoma[6,7,8,16], and 50%[6,8,16,38,39,40] to 80%[7,19] after resection of cancer. Resection of viable cancer cells is usually followed by two additional cycles of chemotherapy. It has been suggested that this postresection chemotherapy should preferably be a different regimen than before resection[3,7]. Yet, an expectant policy after complete resection of cancer has also been followed (Horwich A., personal communication 1992).

If resection is not performed, masses containing mature teratoma may start to grow during a follow-up of months, or even years ('growing teratoma syndrome')[41]. Resection may then be more complicated than it would have been shortly after the end of chemotherapy. Although residual mature teratomas have a less abnormal karyotype than the primary tumor[42], a risk of malignant transformation has been reported[43,44,45,46]. Leaving masses with residual cancer unresected is considered to harbour a serious risk. It is however uncertain how many of the patients with histologically viable cancer cells will eventually relapse. It may be hypothesized that the risk of relapse also depends on the extensiveness of cancer cells in the residual mass, with a low probability of relapse if only small foci of malignancy remain. No data are currently available about this relation. When relapse of malignancy occurs after not resecting mature teratoma or cancer, salvage chemotherapy will be given. However, these regimens have rather limited efficacy (around 25%[1,2,47,48]) and late relapses frequently have a high degree of chemotherapy resistance.

Given these uncertainties in the benefits of resection, it is difficult to indicate thresholds for the probability of necrosis and cancer. These thresholds may also depend on country or center specific circumstances, like the feasibility of frequent follow-up visits with high-quality CT scanning of the abdomen[5]. If frequent follow-up is impossible, any residual mass should be resected. Finally, in a health care environment with limited resources, financial costs of surgery and subsequent hospital stay may argue against resection in patients with a high probability of necrosis or a low probability of cancer.

In conclusion, the histology at resection, especially necrosis, can be predicted with high accuracy: the models predict reliably and discriminate rather well. The models cannot exclude the presence of cancer, but the probability is very low in some patients. The predicted probabilities are easily calculated with the prognostic score chart and may help to choose the optimal treatment, taking into account the potential benefits, morbidity and mortality of resection, feasibility of frequent follow-up, the financial costs and the patient's individual preferences.

We would like to thank Nancy L. Geller, PhD, Biostatistics Research Branch, National Heart, Lung, and Blood Institute, Bethesda, for statistical discussions and René Eijkemans, Center for Clinical Decision Sciences, Erasmus University, Rotterdam for mathematical support.

Appendix

The missing values are filled in (or *imputed*) assuming random missingness. Regression models for the variables with missing values were estimated on the complete cases. Multiple linear regression models were estimated for the continuous predictors mass size (on the pre- or postchemotherapy CT scan) and prechemotherapy marker values (AFP, HCG, LDH_{st}, where LDH_{st} means the LDH value divided by the normal level). All continuous predictors had skewed distributions, which became more normally distributed by log-transformation. Independent variables were selected in a stepwise manner, with $p<.05$ for entry of variables and $p<.10$ for removal of variables. A logistic regression model was used to estimate the presence of teratoma elements in the primary tumor.

The correlation matrix between the predictors is shown below (Table A1). It appears that strong correlations exist between several predictors, and thus imputation based on the values of the other predictors is attractive.

Table A1 Correlations between predictors. All continuous predictors are log-transformed.

	PRESIZE	POSTSIZE	AFP	HCG	LDH_{st}	TERATOMA
PRESIZE	1.00					
POSTSIZE	.57**	1.00				
AFP	.31**	.32**	1.00			
HCG	.22**	.20**	.24**	1.0		
LDH_{st}	.51**	.23**	.20**	.24**	1.00	
TERATOMA	.10	.22**	.04	.13*	-.07	1.00

2-tailed Significance: * - .01 ** - .001

Table A2 shows the independent variables used to impute the missing values, as estimated from the complete cases. Missing values were imputed in 115 cases. For the patient with missing primary histology, the predicted probability was 0.20, and hence the value '0' was imputed (no teratoma). The cases with imputed values were assigned a weight less than the cases with complete values for all predictors (*downweighted*). This weight is calculated as $1 - \rho^2_{1y*s}$, where ρ^2_{1y*s} is the partial correlation of the predictor with missing values and Y given the other predictors. The partial correlation was approximated by the ratio of the Wald statistic and -2 times the log likelihood of a base model that contains only the intercept in the logistic regression models predicting necrosis and cancer respectively.

Table A2 Regression models for imputation of missing values, estimated from the complete cases.

Dependent	Independent	r^2	N	wNec	wCan
PRESIZE	AFP, LDH_{st}, POSTSIZE	.50	7	.981	.997
POSTSIZE	AFP, LDH_{st}, PRESIZE	.37	23	.985	.997
AFP	HCG, PRESIZE, POSTSIZE	.15	1	.979	-
HCG	AFP, LDH_{st}, TERATOMA	.13	0	.984	-
LDH_{st}	HCG, PRESIZE, POSTSIZE, TERATOMA	.27	83	.971	.992
TERATOMA	HCG, LDH_{st}, POSTSIZE -	1	.974	-	

Dependent: (log-transformed) dependent variables; Independent: Independent variables in the regression equations used to impute missing values; r^2: adjusted multiple correlation coefficient, indicating the variance explained by the model; N: Number of cases where values were imputed; wNec: weight in logistic regression model predicting necrosis; wCan: weight in logistic regression model distinguishing cancer from teratoma. If shrinkage could not be calculated because of missing presize, the weight of presize was used.

References

1. Einhorn LH. Treatment of testicular cancer: a new and improved model. J Clin Oncol 8: 1777-1781, 1990
2. Peckham M. Testicular cancer. Rev Oncol 1: 439-453, 1988
3. Toner GC, Panicek DM, Heelan RT, et al. Adjunctive surgery after chemotherapy for nonseminomatous germ cell tumors: recommendations for patient selection. J Clin Oncol 8: 1683-1694, 1990
4. Donohue JP, Rowland RG, Kopecky K, et al. Correlation of computerized 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
5. Fosså SD, Qvist H, Stenwig AE, et al. Is postchemotherapy retroperitoneal surgery necessary in patients with nonseminomatous testicular cancer and minimal residual tumor masses? J Clin Oncol 10: 569-573, 1992
6. Hendry WF, A'Hern RP, Hetherington JW, et al. Para-aortic lymphadenectomy after chemotherapy for metastatic non-seminomatous germ cell tumours: prognostic value and therapeutic benefit. Br J Urol 71: 208-213, 1993
7. Steyerberg EW, Keizer HJ, Zwartendijk J, et al. Prognosis after resection of residual masses following chemotherapy for metastatic nonseminomatous testicular cancer: a multivariate analysis. Br J Cancer 68: 195-200, 1993
8. Mulders PFA, Oosterhof GON, Boetes C, et al. The importance of prognostic factors in the individual treatment of patients with disseminated germ cell tumours. Br J Urol 66: 425-429, 1990
9. Gelderman WAH, Schraffordt Koops H, Sleijfer DTh, et al. Results of adjuvant surgery in patients with stage III and IV nonseminomatous testicular tumors after cisplatin-vinblastine-bleomycin chemotherapy. J Surg Oncol 38: 227-232, 1988
10. Aass N, Klepp O, Cavillin-Ståhl E, et al. Prognostic factors in unselected patients with nonseminomatous metastatic testicular cancer: a multicenter experience. J Clin Oncol 9: 818-826, 1991.
11. Donohue JP, Schraffordt Koops H, Hendry WF, DeBruyne FMJ. Surgery in advanced disease for testicular cancer in Newling DWW, Jones WG (eds): EORTC Genitourinary Group Monograph 7: Prostate Cancer and Testicular Cancer. Wiley-Liss, Inc., New York, 1990

12. Steyerberg EW, Keizer HJ, Stoter G, Habbema JDF. Predictors of residual mass histology following chemotherapy for metastatic nonseminomatous testicular cancer: a quantitative overview of 996 resections. Eur J Cancer, 30A: 1231-1239, 1994
13. Fosså SD, Aass N, Ous S, et al. Histology of tumor residuals following chemotherapy in patients with advanced nonseminomatous testicular cancer. J Urol 142: 1239-1242, 1989
14. Fosså SD, Ous S, Lien HH, et al. Post-chemotherapy lymph node histology in radiologically normal patients with metastatic nonseminomatous testicular cancer. J Urol 141: 557-559, 1989
15. Gelderman WAH, Schraffordt Koops H, Sleijfer DTh, et al. Treatment of retroperitoneal residual tumor after PVB chemotherapy of nonseminomatous testicular tumors. Cancer 58: 1418-1421, 1986
16. Tait D, Peckham MJ, Hendry WF, Goldstraw P. Post-chemotherapy surgery in advanced non-seminomatous germ-cell tumours: the significance of histology with particular reference to differentiated (mature) teratoma. Br J Cancer 50: 601-609, 1984
17. Pizzocaro G, Salvioni R, Pasi M, et al. Early resection of residual tumor during cisplatin, vinblastine, bleomycin combination chemotherapy in stage III and bulky stage II nonseminomatous testicular cancer. Cancer 56: 249-255, 1985
18. Bracken RB, Johnson DE, Frazier OH, et al. The role of surgery following chemotherapy in stage III germ cell neoplasms. J Urol 129: 39-43, 1983
19. Harding MJ, Brown IL, Macpherson SG, et al. Excision of residual masses after platinum based chemotherapy for non-seminomatous germ cell tumours. Eur J Cancer Clin Oncol 25: 1689-1694, 1989
20. Nijman JM, Schraffordt Koops H, Kremer J, et al. Gonadal function after surgery and chemotherapy in men with stage II and III nonseminomatous testicular tumors. J Clin Oncol 5: 651-656, 1987
21. De Graaf WE, Oosterhuis JW, Van der Linden S, Homan van der Heide JN, Schraffordt Koops H, Sleijfer DTh. Residual mature teratoma after chemotherapy for nonseminomatous germ cell tumors of the testis occurs significantly less often in lung than in retroperitoneal lymph node metastases. J Urogen Pathol 1:75-81, 1991
22. Harrell FE, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst 80: 1198-1202, 1988
23. Little RJA. Regression with missing X's: a review. J Am Stat Assoc 87: 1227-1237, 1992
24. EGRET statistical package, Statistics and Epidemiology Research Corporation, Seattle, Washington, USA, 1990
25. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 5: 421-433, 1986
26. Hosmer DW, Lemeshow S. Applied logistic regression. New York, NY: John Wiley & Sons Inc, 1989
27. BMDP statistical software, Inc. Los Angeles, CA, USA, 1990
28. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 247: 2543-2546, 1982
29. SAS Institute Inc., SAS® Technical Report P-200, SAS/STAT® Software: CALIS and LOGISTIC Procedures, Release 6.04. Cary, NC: SAS Institute Inc., USA, 1990. pp. 194-195
30. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 78: 316-331, 1983
31. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med, 9: 1303-1325, 1990
32. Bosl GJ, Geller NL, Cirrincione C, et al. Multivariate analysis of prognostic variables in patients with metastatic testicular cancer. Cancer Res 43: 3403-3407, 1983

33. Bajorin D, Katz A, Chan E, et al. Comparison of criteria for assigning germ cell tumor patients to "good risk" and "poor risk" studies. J Clin Oncol 6: 786-792, 1988
34. Davey DD, Ulbright TM, Loehrer PJ et al. The significance of atypia within teratomatous metastases after chemotherapy for malignant germ cell tumours. Cancer 59: 533-539, 1987
35. Fosså SD, Kreuser ED, Roth GJ, et al. Long-term side effects after treatment of testicular cancer in Newling DWW, Jones WG (eds): EORTC Genitourinary Group Monograph 7: Prostate Cancer and Testicular Cancer. Wiley-Liss, Inc., New York, 1990
36. Horwich A. Testicular cancer: investigation and management. Horwich A (ed), Chapman & Hall, London, 1991. Abdominal surgery postchemotherapy: 278-280
37. Donohue JP, Foster RS, Rowland RG, et al. Nerve-sparing retroperitoneal lymphadenectomy with preservation of ejaculation. J Urol 144: 287-292, 1990
38. Mead GM, Stenning SP, Parkinson MC, et al. The second medical research council study of prognostic factors in nonseminomatous germ cell tumors. J Clin Oncol 10: 85-94, 1992
39. Geller NL, Bosl GJ, Chan EYW. Prognostic factors for relapse after complete response in patients with metastatic germ cell tumors. Cancer 63:440-445, 1989
40. Fox EP, Weathers TD, Williams SD, et al. Outcome analysis for patients with persistent nonteratomous germ cell tumor in postchemotherapy retroperitoneal lymph node dissections. J Clin Oncol 11: 1294-1299, 1993
41. Logothetis CJ, Samuels ML, Trindade A, et al. The growing teratoma syndrome. Cancer 50: 1629-1635, 1982
42. Castedo SMMJ, de Jong B, Oosterhuis JW, et al. Chromosomal changes in mature residual teratomas following polychemotherapy. Cancer Res 49: 672-676, 1989
43. Ulbright TM, Loehrer PJ, Roth LM, et al. The development of non-germ malignancies within germ cell tumors: a clinicopathologic study of 11 cases. Cancer 54: 1824-1833, 1984
44. Ahlgren AD, Simrell CR, Triche TJ, et al. Sarcoma arising in a residual testicular teratoma after cytoreductive chemotherapy. Cancer 54: 2015-2018, 1984
45. Ahmed T, Bosl GJ, Hajdu SI. Teratoma with malignant transformation in germ cell tumors in men. Cancer 56: 860-863, 1985
46. Molenaar WM, Oosterhuis JW, Meiring A, et al. Histology and DNA contents of a secondary malignancy arising in a mature residual lesion six years after chemotherapy for a disseminated nonseminomatous testicular tumor. Cancer 58: 264-268, 1986
47. Dearnaley DP, Horwich A, A'Hern R, et al. Combination chemotherapy with bleomycin, etoposide and cisplatinum (BEP) for metastatic testicular teratoma: long-term follow-up. Eur J Cancer 27: 684-691, 1991
48. Dimipoulos MA, Amato RJ, Logothetis CJ. Predictive factors for effective salvage therapy of nonseminomatous germ cell tumors of testis. Urol 38: 351-354, 1991

7 Resection of residual retroperitoneal masses in testicular cancer: evaluation and improvement of selection criteria

E.W. Steyerberg, H.J. Keizer, S.D. Fossâ, D.T. Sleijfer, D.F. Bajorin, J.P. Donohue and J.D.F. Habbema for the ReHiT study group.
Submitted for publication

Abstract

Background: Most patients with metastatic non-seminomatous testicular cancer can currently be cured with cisplatin based chemotherapy. After a successful response to chemotherapy (as apparent from normal tumor markers AFP and HCG), retroperitoneal lymph nodes may harbour residual tumor or totally benign tissue (necrosis/fibrosis). Surgical resection is an effective way to remove these residual masses, but the selection criteria vary widely between centers. These criteria were evaluated in this study.
Methods: Individual patient data were available from 544 patients, who had retroperitoneal lymph node dissection of residual masses. Six currently applied resection policies were identified from the literature and were evaluated in this data set. Two alternative policies were developed, based on logistic regression analysis with well-known predictors of the histology at resection. Evaluation of the policies included the true positive rate (resection in case of tumor), and the false positive rate (resection in case of necrosis).
Results: Most current policies use the size of the residual mass (≥ 10mm/≥ 20mm) as the predominant selection criterion. This results in high true positive rates (most >90%), but false positive rates between 37% and 87%. The alternative strategy included five well-known predictors of necrosis in addition to residual mass size (primary tumor histology, prechemotherapy levels of the three tumor markers AFP, HCG and LDH, and mass shrinkage during chemotherapy). This strategy resulted in improved true and false positive rates, even when categories of the predictors were simplified for practical application. Moreover, the alternative policies allow for the choice of a center-specific combination of the true and false positive rate, based on the relative weight of missing tumor versus unnecessary resection of necrosis.
Conclusion: A simple statistical model, based on a limited number of patient characteristics, provides better guidelines for patient selection than those currently used in clinical practice. This implies that guidelines for resection of retroperitoneal residual masses may need to be reconsidered. Further validation of the statistical models is required to confirm the findings.

7.1 Introduction

Testicular cancer is the most common malignancy among men in the age between 20 and 35. Fortunately, even metastatic disease can currently be cured in the majority (60-80%) of patients with nonseminomatous germ cell tumor since the introduction of cisplatin based chemotherapy[1,2]. After chemotherapy, surgical resection is a generally accepted treatment to remove residual retroperitoneal lymph node masses, since these masses still harbour residual tumor in about half of the patients. Alternatively, patients may be treated conservatively, which includes follow-up with regular blood tests and CT scans of the abdomen. A uniform approach to the selection of patients for resection is lacking[3,4,5,6,7,8,9] and percentages of surgically treated patients vary between 20%[7,10] and 86%[11]. Therefore, several large cancer centers cooperated to evaluate the selection for resection.

Resection of residual masses provides the histological diagnosis, which may be purely benign with necrotic and/or fibrotic remnants only (necrosis), or residual tumor (mature teratoma or viable cancer). In case of viable cancer, two additional courses of chemotherapy are usually recommended[12,13]. This additional therapy probably reduces the risk of relapse, in addition to the resection itself. Resection of mature teratoma prevents growth of the residual mass[14]. In contrast, resection of benign masses has no therapeutic benefit. An ideal resection policy would therefore result in surgical removal of all masses with residual tumor (mature teratoma or cancer) and in a conservative treatment of all masses with necrosis.

Currently used selection policies were evaluated in an international data set from 6 study groups. A statistical model was developed from this same data set, using several well-known predictors of the histology of residual masses[3,4,5,8,10,15,16,17,18]. Easy-to-use alternative selection criteria were based on this analysis and compared with the current policies.

7.2 Patients and Methods

7.2.1 Patients

An international data set was collected, consisting of patients with metastatic nonseminomatous testicular cancer, including patients with pure seminoma and elevated levels of prechemotherapy tumor markers, who underwent resection of retroperitoneal residual masses after induction chemotherapy with cisplatin based chemotherapy[19]. Excluded were patients with elevated tumor markers AFP or HCG at the time of surgery, patients with extragonadal tumors, patients with pure seminoma and patients resected after relapse of tumor following initial chemotherapy. This selection resulted in a reasonable homogeneous patient group with respect to the decision to resect retroperitoneal masses.

Individual patient data included basic patient identification, histology at resection, and the following predictors: presence of teratoma elements in the primary tumor, prechemotherapy tumor marker levels (AFP, HCG, LDH), and prechemotherapy and

postchemotherapy mass size. Patients were included from Memorial Sloan-Kettering Cancer Center (MSKCC)[3] (N=121), Norwegian Radium Hospital (NRH)[5,20,21] (N=127), Indiana University Hospital (IUH)[15] (N=42), University Hospital Groningen (UHG)[4,22,23,24] (N=137), and four other Dutch centers (University Hospitals of Nijmegen[8], Leiden, Amsterdam, Rotterdam[9]: N=117). Most European patients were treated according to trial protocols of the EORTC and MRC. In all centers, patients with residual abnormalities on radiologic studies were recommended to undergo resection. Adherence to this recommendation was not evaluated in this study. In addition, patients with initial bulky retroperitoneal disease (diameter ≥30 mm) were candidates for resection at MSKCC[3], as well as UHG-patients with teratoma elements in their primary tumor from 1988 onwards[4]. At NRH, resection was performed routinely in all patients with retroperitoneal lymph node enlargement at diagnosis[11]. The 51 patients included in this analysis from Indiana (IUH) all had a palpable prechemotherapy mass larger than 10 cm[15]. This series thus represents a small part only of the experience at IUH with resection of residual masses. 544 patients were available for analysis: 245 (45%) with necrosis and 299 (55%) with residual tumor. 68 of the latter 299 patients had cancer (23%) and 231 had mature teratoma (77%). Patients were treated between 1975 and 1993, with a minority (11%) treated before 1981, and most between 1981 and 1985 (51%).

7.2.2 Methods

Currently used resection policies were evaluated in the international data set. The probabilities of each residual histology (necrosis, mature teratoma, viable cancer) were calculated in masses that would be selected for resection and in masses that would be treated conservatively according to each policy. Implicitly, the currently used policies uses cut-off values for the probability of necrosis below which resection is performed. These cut-offs were approximated with logistic regression analysis models including the predictors in the policy considered.

The policies were further evaluated as diagnostic tests, using the histology at resection as the gold standard diagnosis[25]. The true positive rate (or sensitivity) of a policy referred to the fraction of resected patients among those with residual tumor. The false positive rate (or 1 minus specificity) referred to the fraction of patients who would undergo resection among the patients with necrosis. A perfect resection policy would have a true positive rate of 100% and a false positive rate of 0%. Areas under the receiver operating characteristic (ROC) curve were estimated to facilitate comparison of the diagnostic quality of the policies, assuming a logistic distribution of the data[26]. An area of 0.5 would arise if patients with and without residual tumor were equally likely to undergo resection. An area of 1.0 corresponds to a perfect policy.

Alternative resection criteria were developed with logistic regression analysis (SPSS/PC+ v5.01 software; SPSS Inc, Chicago, IL, USA, and SAS v6.04 software; SAS Institute Inc, Cary, NC, USA). The probability of necrosis was estimated for combinations of characteristics known before resection (predictors). A previous analysis of the data set showed that important predictors of necrosis were: the absence of

teratoma elements in the primary tumor, prechemotherapy normal AFP, normal HCG, high LDH, a small postchemotherapy mass size and a large shrinkage during chemotherapy[19]. The latter three predictors were modeled as continuous variables, including transformations of postchemotherapy size (square root) and prechemotherapy LDH (logarithmic). This model showed good results with extensive validation procedures, including bootstrapping[27] and leave-one-study-out evaluations[19]. To facilitate application in clinical practice, we simplified the analysis by categorizing the prechemotherapy LDH value (elevated versus normal), postchemotherapy size (0-9 mm, 10-19 mm, 20-29 mm, 30-49 mm, ≥50 mm) and shrinkage (<0%, 0-69.9%, ≥70%). Both for the original and the simplified model, we calculated areas under the ROC curve[28]. True and false positive rates were calculated with increasing cut-off values for the probability of necrosis.

7.2.3 *Comparison of policies*

The diagnostic quality of the policies could be compared with the area under the ROC curve, with larger areas indicating better policies. A limitation of the area under the ROC curve is however that it does not consider the frequency of the outcome (necrosis / tumor at resection) nor the relative importance of misclassifications[29]. We therefore calculated a weighted classification error. The relative importance (or weight) of missing residual tumor was set as 1, 2, 4, 8 and 16 times that of unnecessary resection. The weighted classification error was expressed as the number of unnecessary resections of necrosis and was calculated as:

(# unnecessary resections) + (weight • # missed resections of tumor).

McNemar's test for paired observations was used for statistical comparisons between the policies[30]. Since the test assumes equal weights for false positive and false negative misclassifications, fair comparisons could only be made if one policy dominated, i.e. had both a higher true and a lower false positive rate.

7.2.4 *Verification bias*

In this analysis, data are only available from patients where the residual histology was verified by resection. These patients were selected from the total population of patients with normal tumor markers after chemotherapy according to the center-specific selection policies. Verification bias is therefore likely, which would lead to overestimated true and false positive rates, but to largely unbiased predicted probabilities of necrosis[31,32]. Correction for verification bias in the international data set was difficult, since six different centers participated. Fortunately, in one center resection was performed routinely (NRH, N=127)[11], such that virtual absence of verification bias might be assumed here. The policies were therefore also evaluated separately in these 127 patients.

7.3 Results

7.3.1 Current policies

Table 1 shows the currently used resection policies that were evaluated. The histological distribution is shown in masses that would be resected or treated conservatively according to each policy. The first policy (resection of all masses ≥ 10 mm) has widely been applied in European centers[8,9,33]. Masses ≥10 mm are generally detected on CT scans, and this practice thus corresponds to resection if residual masses are detected on CT scans. It can be read from Table 1 that the probability of necrosis was 38% in masses ≥ 10mm, in contrast to 72% in masses < 10mm. The second policy (resection of masses ≥ 20mm) has especially been used in British centers[6,7,10]. It would leave masses unresected with a low risk of viable cancer (4%), but a considerable risk of mature teratoma (30%). Policy 3 to 5 use one or more patient characteristics in addition to residual mass size. If resection is performed in all patients with a teratoma positive primary tumor (policy 3[4]), the risk of leaving tumor unresected reduces to 23% (15%+8%) compared to 28% with policy 1. Policy 4[3] leads to a similar risk of missing residual tumor compared to policy 1 (30% vs 28%). Policy 5[5] consists of resection in all patients, except a small subgroup with residual masses < 20mm and three favorable characteristics (primary tumor teratoma negative and prechemotherapy AFP and HCG normal). This stringent practice does not guarantee that no tumor is missed, but the risk is low (6%+6%=12%). Policy 6[15] consists of conservative treatment of patients with a shrinkage over 70% and a teratoma negative primary tumor. Residual tumor was found in 24% (17%+7%) of these patients.

Table 1 Resection policies and the histology of residual masses. All patients had normal tumor markers AFP and HCG after chemotherapy for metastatic non-seminomatous testicular cancer.

Policy	Resection if	R/ C*	Total N=544 100%	Necrosis N=245 45%	Teratoma N=231 43%	Cancer N=68 13%
1.	Residual masses ≥ 10 mm	R	437	38%	47%	14%
		C	107	72%	22%	6%
2.	Residual masses ≥ 20 mm	R	313	29%	52%	19%
		C	231	67%	30%	4%
3.	Residual masses ≥ 10 mm or primary tumor teratoma positive	R	482	41%	46%	13%
		C	62	77%	15%	8%
4.	Residual masses ≥ 10 mm or prechemotherapy mass > 30 mm	R	480	42%	45%	14%
		C	64	70%	25%	5%
5.	Residual masses ≥ 20 mm or primary tumor teratoma positive or prechemo-therapy AFP/HCG elevated	R	508	42%	45%	13%
		C	36	89%	6%	6%
6.	Shrinkage in size < 70% or primary tumor teratoma positive	R	456	39%	47%	14%
		C	88	76%	17%	7%

* R: patients fulfilling resection criteria; C: patients fulfilling conservative treatment criteria

7.3.2 Alternative resection policies

Alternative resection policies were based on statistical analysis of the international data set. The results of an analysis with continuous predictors are presented in Table 2[19]. The probability of necrosis corresponds to the sum score and can readily be calculated for individual patients. Exact formulas to calculate the probability of necrosis, mature teratoma and cancer are presented in the Appendix.

Table 2 Prognostic score chart to estimate the probability of necrosis in residual retroperitoneal masses.

Predictor	Value							Score
Primary tumor histology								
Teratoma-negative	+9							.....
Prechemotherapy markers								
Normal AFP	+9							.....
Normal HCG	+8							.....
LDH/normal value**	.6	.8	1.0	1.5	2.0	3.0	4.5	
Score	-5	-2	0	+4	+7	+11	+15	.....
Postchemotherapy mass size								
Transversal diameter (mm)**	2*	5	10	20	30	50	100	
Score	-4	-6	-9	-13	-16	-20	-28	.....
Shrinkage								
100•(presize-postsize)/presize**	-50	0	50	75	100			
Score	-7	0	+7	+11	+15			.....
Estimated individual probability of necrosis					**Sum score (add)**			.....
Sum score	10	15	20	25	30	35	40	
Probability (%)	51	63	74	82	88	93	95	

* If no mass is detectable on the postchemotherapy CT scan, a size of 2 mm is assumed.
** Continuous variables; scores for intermediate values can be estimated with linear interpolation

A simplified model used categories instead of the continuous predictors in the original model. It was anticipated that the performance of this model would only be slightly worse than the original model, while the application in clinical practice would be facilitated. The categorized predictors as shown in Table 3 were analyzed simultaneously with residual mass size. All five predictors had similar Odds Ratios (Table 3: range 2.2-2.8). Therefore, a 'simple score' was constructed by counting the number of favorable characteristics.

Next, we used the two models (Table 2 and 3) to derive alternative resection strategies. These alternative strategies use a cut-off value for the probability of necrosis. If the predicted probability of necrosis is lower than the cut-off value, resection is performed; if not, conservative treatment will follow. The choice of the cut-off values was based on the estimated cut-off values for the currently used policies. With 60% and 90% as extremes of the probability of necrosis, two areas with a clear treatment advice

evolve. If the probability of necrosis is less then 60%, resection should follow; if the probability exceeds 90%, conservative treatment is advised. In between is a grey area, where the decision to resect a residual mass depends on the cut-off value applied (60%, 70%, 80% or 90%). Table 4 shows the resection strategies for the simplified model from Table 3. It can for instance be read that the probability of necrosis is less than 60% in patients with a residual mass ≥50 mm, in patients with a mass that increased during chemotherapy, in patients with a low score (0 or 1 point), in patients with a mass of 20–29 mm and a score of 2 points, and in patients with a mass of 30–49 mm and a score of 3 points.

Table 3 Categorized predictors of necrosis in addition to residual mass size. Odds Ratios and 95% confidence intervals were calculated with logistic regression analysis (N=544).

Characteristic	OR	95%-CI	Score
Primary tumour teratoma-negative	2.7	[1.8-4.2]	0/1
Prechemotherapy AFP normal	2.4	[1.5-3.9]	0/1
Prechemotherapy HCG normal	2.2	[1.4-3.4]	0/1
Prechemotherapy LDH *elevated*	2.8	[1.6-4.7]	0/1
Shrinkage in mass size ≥ 70%	2.2	[1.3-3.9]	0/1
		Simple score	0-5

Table 4 Resection policies with cut-off values for the probability of necrosis of 60%, 70%, 80% and 90%.*

Mass size	Simple score					
	0	1	2	3	4	5
0-9 mm			Nec>60%	Nec>70%	Nec>80%	Conservative treatment (Nec>90%)
10-19 mm						
20-29 mm				Nec>60%		
30-49 mm	Resection (Nec ≤ 60%)				Nec>70%	Nec>80%
≥ 50 mm or increased mass						

* The probability of necrosis is estimated by the simple score (Table 3) and residual mass size. Resection is advised if the probability is lower than 60% and conservative treatment if the probability exceeds 90%. If e.g. a cut-off value of 70% is applied, a predicted probability lower than 70% implies resection and a probability over 70% implies conservative treatment.

** Nec: Predicted probability of necrosis

7.3.3 Evaluation of policies

Table 5 shows the results of the evaluation of the currently used policies, the alternative policies, and the extreme policy of resection in all patients. Figure 1 displays the results graphically. The true positive (TP) rate of the currently used policies (except policy 2) exceeds 90%. This means that over 90% of the patients with residual tumor would be resected with these policies and that less than 10% of the masses with tumor would be missed. The false positive (FP) rate varies between 37% and 87%, which means that a large proportion of the patients with necrosis would undergo resection unnecessarily. Policy 2 is remarkable, as both the TP and FP rate are relatively low (74% and 37%). For the alternative policies (7 and 8), it is clear that an increase of the cut-off values for the probability of necrosis, leads to a larger fraction of resected patients and to higher TP and FP rates. Thus, the higher the required probability of necrosis for conservative treatment, the lower the risk of missing tumor, but the higher the risk of unnecessary resection. The diagnostic performance of the policies was further compared by the areas under the ROC curve. The performance of policy 1, 2, 3, 4 and 6 was more or less similar (area .72, .74, .75, .69 and .75). Policy 5 had a better diagnostic ability, similar to the alternative resection policies (7 and 8).

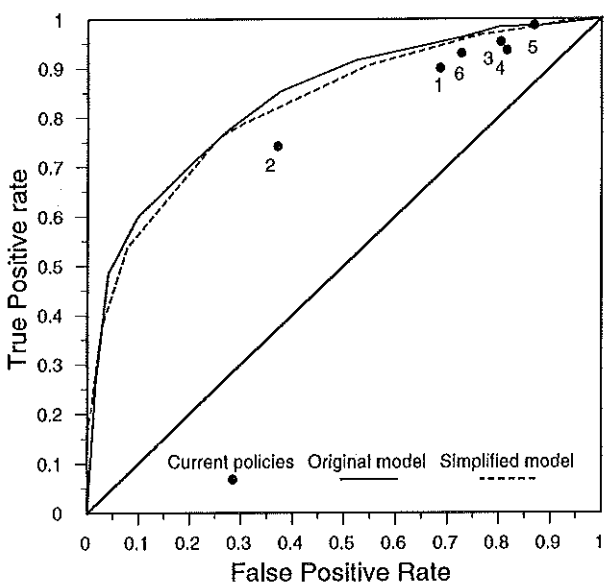

Figure 1 ROC curves of the original and simplified model to distinguish non-benign from benign tissue. The dots (•) indicate the true and false positive rates of the current policies.

7.3.4 Dominance and cut-off values

For the alternative policies, cut-off values for the probability of necrosis could be found where these policies dominated over the current policies except policy 5. For example, a cut-off value of 70% with policy 7 resulted in a higher TP rate and a lower FP rate than policy 1 ($p<.001$). Similar comparisons were made between the alternative policies and the current policies 2, 3, 4 and 6, which were statistically significant ($p<.05$). A cut-off value of 90% leads to a similar performance as policy 5.

The misclassification error shown in Table 5 indicates that the optimal cut-off value for the probability of necrosis in policy 7 and 8 increases with the relative weight of missing tumor. For example if 2, 4 or 8 unnecessary resections are judged to be worth 1 case of tumor, optimal cut-off values are 70%, 80% and 90% respectively. If the ratio is increased to 16:1 or higher, resection in all patients (policy 9) is the optimal strategy, since this strategy then has the lowest misclassification error among the policies.

Evaluation of the policies in the 127 largely unselected patients confirms that verification bias is present in the true and false positive rates (Table 6). As expected, the true and false positive rate are lower than when evaluated on the total data set for most policies. The areas under the ROC curve are however comparable to the initial estimates. Also, the alternative policies 7 and 8 still dominate over the other policies (higher FP and lower FP), except policy 5. Therefore, verification bias does not influence our main findings substantially.

7.4 Discussion

In this study we evaluated several selection policies for surgery in patients who were successfully treated for metastatic testicular cancer, as apparent from normal tumor markers after chemotherapy. In 45% of these patients resection was unnecessary, since only totally benign tissue was present. We found that currently used policies would lead to resection in between 37% and 87% of these patients. This variation is explained by the patient characteristics considered for selection and the varying degree of certainty that tumor is not missed. Alternative strategies were developed that combine more characteristics than most current policies and hence have a better inherent diagnostic ability (area under the ROC curve). Moreover, the degree of certainty that tumor is not missed can be decided on by weighing the relative importance of missing tumor against unnecessary resection.

Currently used resection policies are mainly based on a single characteristic, i.e. the size of the residual mass. The policy to resect CT scan detected masses of 10 mm or larger is probably the most frequently used nowadays. Some strategies include additional characteristics for the selection of patients. Indeed, our previous analyses[18,19] indicate that other equipotent predictors include the absence of teratoma elements in the primary tumor, prechemotherapy tumor marker levels (AFP, HCG, LDH), and shrinkage. Therefore, alternative criteria for resection can be developed, so that small residual masses (< 10mm) are resected if an unfavorable combination of other characteristics is

Table 5 Evaluation of resection policies. For each policy, the table shows the cut-off value for the probability of necrosis, true-positive (TP) rate, the false-postive FP) rate, the area under the ROC curve AUC), the percentage of patients undergoing resection, and the number of weighted misclassifications resection of tumor: resection of necrosis).

							Classification error				
Policy	Selection criteria	Cut-off*	TP	FP	AUC	%Resected	1:1	2:1	4:1	8:1	16:1
1.	Residual masses ≥10 mm	62%	90%	69%	.72	80%	198	228	288	408	648
2.	Residual masses ≥20 mm	49%	74%	37%	.74	58%	168	245	399	707	1323
3.	Residual masses ≥10 mm or primary ter+	73%	95%	80%	.75	89%	211	225	253	309	421
4.	Residual masses ≥10 mm or presize >30 mm	59%	94%	82%	.69	88%	219	238	276	352	504
5.	Residual masses ≥20 mm or primary ter+ or prechemotherapy AFP or HCG elevated	85%	98.7%	87%	.84	93%	217	221	229	245	277
6.	Shrinkage <70% or primary ter+	72%	93%	73%	.75	84%	199	220	262	346	514
7.	Table 2: original model	60%	85%	38%	.84	64%	136	180	268	444	796
		70%	92%	52%		74%	153	178	228	328	528
		80%	96%	73%		86%	190	201	223	267	355
		90%	98.7%	89%		94%	222	226	234	250	282
8.	Table 4: simplified model	60%	79%	30%	.82	57%	138	202	330	586	1098
		70%	91%	55%		74%	162	190	246	358	582
		80%	97%	76%		88%	197	207	227	267	347
		90%	99.7%	94%		97%	231	232	234	238	246
9.	All patients	100%	100%	100%	.5	100%	245	245	245	245	245

* Cut-off estimated with logistic regression models containing the predictors of each policy

Table 6 Evaluation of resection policies in 127 largely unselected patients from the Norwegian Radium Hospital, Oslo, Norway.*

Policy	Selection criteria	TP	FP	AUC	%Resected	Classification error 1:1	2:1	4:1	8:1	16:1
1.	Residual masses ≥10 mm	80%	53%	.70	66%	47	59	83	131	227
2.	Residual masses ≥20 mm	57%	17%	.78	36%	37	63	115	219	427
3.	Residual masses ≥10 mm or primary ter+	93%	70%	.77	81%	50	54	62	78	110
4.	Residual masses ≥10 mm or presize >30 mm	89%	65%	.72	76%	50	57	71	99	155
5.	Residual masses ≥20 mm or primary ter+ or prechemotherapy AFP or HCG elevated	100%	76%	1.0	87%	50	50	50	50	50
6.	Shrinkage <70% or primary ter+	89%	70%	.69	79%	53	60	74	102	158
7.	Table 2: probability of necrosis; sum score			.86						
	≤60%; ≤ 13	84%	31%		59%	34	44	64	104	184
	≤70%; ≤ 18	93%	49%		70%	36	40	48	64	96
	≤80%; ≤ 23	98%	70%		84%	47	48	50	54	62
	≤90%; ≤ 32	100%	91%		95%	60	60	60	60	60
8.	Table 4: probability of necrosis			.82						
	≤60%	69%	20%		43%	32	51	89	165	317
	≤70%	84%	50%		66%	43	53	73	113	193
	≤80%	95%	73%		84%	51	54	60	72	96
	≤90%	100%	88%		94%	58	58	58	58	58
9.	All patients	100%	100%	.5	100%	66	66	66	66	66

* For each policy, the table shows the true-positive (TP) rate, the false-postive (FP) rate, the area under the ROC curve (AUC), the percentage of patients undergoing resection, and the classification error for varying weights (non-resection of tumor: resection of necrosis) of misclassification.

present and, on the other hand, larger masses (e.g. 10-19 mm or 20-29 mm) are treated conservatively if other predictors are favorable.

The diagnostic performance of the policies was evaluated with false positive (FP) and true positive (TP) rates, the area under the ROC curve (AUC) and the classification error. The FP rate should be low (no resection of necrosis) and the TP should be as high as possible (resection of residual tumor), resulting in a large AUC. The AUC indicates the diagnostic ability of a selection strategy, but does not consider the prevalence of the residual histologies (necrosis or tumor) nor the relative importance of beneficial and unnecessary resections[29]. These two considerations are taken into account in the weighted classification error.

Most of the currently used policies lead to resection in the majority of patients with residual tumor (TP rates > 90%). Resection of masses ≥ 20 mm (policy 2) however resulted in a relatively low TP rate (74%), which meant that 26% of the masses with residual tumor were left unresected. If the latter error is more important than unnecessary resection, policy 2 has to be rejected, although the diagnostic ability as measured by the ROC curve was reasonable (area .74). Further, a slightly less favorable performance was observed with the policy to resect small residual masses if the initial mass was relatively large (>30 mm)[3]. This is explained by the finding that a large shrinkage is a predictor of necrosis (multivariate p-value: 0.003), rather than a predictor of tumor. The most stringent currently applied selection policy (# 5)[5], resulted in a combination of the FP and TP rate similar to the use of a high cut-off for the probability of necrosis in the alternative policies (>90%). The similar diagnostic ability is explained by the fact that the three predictors used in this policy in addition to mass size (primary tumor teratoma negative, prechemotherapy AFP and HCG normal), were also used in the alternative strategies. At lower cut-off values, these alternative strategies had better TP and FP rates than the other currently used policies. For example, the policy to resect masses ≥10 mm is dominated by using Table 2 or Table 4 with a cut-off value of 70% for the predicted probability of necrosis ($p<.001$).

Although the alternative selection strategies have better diagnostic properties than most currently used policies, a dilemma remains on the optimal cut-off value for the probability of necrosis, which is determined by the relative importance of missing tumor and unnecessary resection. The disadvantages of unnecessary resection include short-term and long-term morbidity (especially retrograde or anejaculation[6,23]), mortality, and financial costs. Residual mature teratoma or viable cancer may grow, and increase the risk of relapse[3,14]. These risks cannot be quantified easily to derive a cut-off value for the probability of necrosis in the alternative selection policies. On the other hand, a plausible estimate may be that missing residual tumor is at least 4 times as important as an unnecessary resection. This leads to an optimal cut-off value of at least 80% for the probability of necrosis with the alternative criteria. If frequent follow-up is difficult[5], the risk of missing tumor may be worth 8 or even 16 unnecessary resections, which leads to more aggressive selection with a cut-off value of 90% or resection in all patients as the preferred strategy.

Another consideration is the probability of cancer, which can be estimated with the formulas in the Appendix[19]. If the probability of cancer exceeds 5%, resection may be indicated, although this implies a value judgement for resection of cancer relative to teratoma and necrosis.

Two limitations of this study have to be considered. First, only operated patients were included and these patients were selected with different criteria in the six participating centers. Evaluation on a subsample without selection showed that this verification bias had resulted in overestimated true and false positive rates. The areas under the ROC curve were however largely unaffected, resulting in the same ordering of the diagnostic performance of the policies. Second, the alternative resection policies have not yet been validated on a new, independent data set. Although several less rigorous validation procedures showed only minor overoptimism of model performance[19], further conformation is required.

We conclude that a policy that takes into account all currently known predictors may result in improved selection of patients for resection. This means that the balance between the number of beneficial and unnecessary resections will favorably be influenced by the clinical application of such a policy.

The ReHiT ('Re-analysis of histology in testicular cancer') study group consists of E.W. Steyerberg, MSc, and Prof. J.D.F Habbema, PhD, Erasmus University, Rotterdam; H.J. Keizer, MD, University Hospital, Leiden; D.T. Sleijfer, MD, and Prof. H. Schraffordt Koops, MD, University Hospital, Groningen; P.F.A. Mulders, MD, University Hospital, Nijmegen; Prof. G. Stoter, MD, Rotterdam Cancer Institute, The Netherlands; Prof. S.D. Fossâ, MD, The Norwegian Radium Hospital, Oslo, Norway; G.C. Toner, MD, Peter MacCallum Cancer Institute, Melbourne, Australia; D.F. Bajorin, MD, Prof. G.J. Bosl, MD, Memorial Sloan Kettering Cancer Center, New York; J.E. Messemer, BSc, K. Ney, MD, and Prof. J.P. Donohue, MD, Indiana University School of Medicine, Indianapolis, USA

Appendix

The formulas to calculate the probability of each histology are shown below. These formulas are implemented in a simple spreadsheet program available from the authors (E-mail: steyerberg@ckb.fgg.eur.nl).

Sumscore(Necrosis): -9.78 + 8.58•'teratoma-negative' + 8.70•'AFPnormal' + 7.61•'HCGnormal' + 9.69•ln(LDH_{st}) - 2.83•Sqrt(postsize) + 0.147•shrinkage

Sumscore(Cancer): -24.18 + 3.95•ln(LDH_{st}) + 1.36•Sqrt(postsize) + 0.053•shrinkage

The variables 'teratoma-negative', 'AFPnormal' and 'HCGnormal' are 1 if true, 0 if false, ln(LDH_{st}) is the natural logarithm of LDH/normal value, postsize is expressed in mm and shrinkage is expressed as %.

The corresponding probabilities are calculated with the formulas:

Probability(Necrosis): $1/[1+e^{-(\text{Sumscore(Necrosis)}/10)}]$

Probability(Cancer): [1 - Probability(Necrosis)] • $[1/(1+e^{-(\text{Sumscore(Cancer)}/10)})]$

Probability(Teratoma): 1 - [Probability(Necrosis)+Probability(Cancer)]

References

1. Einhorn LH. Treatment of testicular cancer: a new and improved model. J Clin Oncol 8: 1777-1781, 1990
2. Peckham M. Testicular cancer. Rev Oncol 1: 439-453, 1988
3. Toner GC, Panicek DM, Heelan RT, et al. Adjunctive surgery after chemotherapy for nonseminomatous germ cell tumors: recommendations for patient selection. J Clin Oncol 8: 1683-1694, 1990
4. Gelderman WAH, Schraffordt Koops H, Sleijfer DTh, et al. Results of adjuvant surgery in patients with stage III and IV nonseminomatous testicular tumors after cisplatin-vinblastine-bleomycin chemotherapy. J Surg Oncol 38: 227-232, 1988
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. Hendry WF, A'Hern RP, Hetherington JW, et al. Para-aortic lymphadenectomy after chemotherapy for metastatic non-seminomatous germ cell tumours: prognostic value and therapeutic benefit. Br J Urol 71: 208-213, 1993
7. Mead GM, Stenning SP, Parkinson MC, et al. The second medical research council study of prognostic factors in nonseminomatous germ cell tumors. J Clin Oncol 10: 85-94, 1992
8. Mulders PFA, Oosterhof GON, Boetes C, et al. The importance of prognostic factors in the individual treatment of patients with disseminated germ cell tumours. Br J Urol 66: 425-429, 1990
9. Steyerberg EW, Keizer HJ, Zwartendijk J, et al. Prognosis after resection of residual masses following chemotherapy for metastatic nonseminomatous testicular cancer: a multivariate analysis. Br J Cancer 68: 195-200, 1993
10. Tait D, Peckham MJ, Hendry WF, Goldstraw P. Post-chemotherapy surgery in advanced non-seminomatous germ-cell tumours: the significance of histology with particular reference to differentiated (mature) teratoma. Br J Cancer 50: 601-609, 1984
11. Aass N, Klepp O, Cavillin-Ståhl E, et al. Prognostic factors in unselected patients with nonseminomatous metastatic testicular cancer: a multicenter experience. J Clin Oncol 9: 818-826, 1991
12. Einhorn LH, Williams SD, Mandelbaum I, Donohue JP. Surgical resection in disseminated testicular cancer following chemotherapeutic cytoreduction. Cancer 48: 904-908, 1981
13. Fox EP, Weathers TD, Williams SD et al. Outcome analysis for patients with persistent nonteratomatous germ cell tumor in postchemotherapy retroperitoneal lymph node dissections. J Clin Oncol 11: 1294-1299, 1993
14. Logothetis CJ, Samuels ML, Trindade A, et al. The growing teratoma syndrome. Cancer 50: 1629-1635, 1982
15. Donohue JP, Rowland RG, Kopecky K, et al. Correlation of computerized 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
16. Harding MJ, Brown IL, Macpherson SG, et al. Excision of residual masses after platinum based chemotherapy for non-seminomatous germ cell tumours. Eur J Cancer Clin Oncol 25: 1689-1694, 1989
17. Gerl A, Clemm C, Schmeller N, et al. Outcome analysis after post-chemotherapy surgery in patients with non-seminomatous germ cell tumours. Ann Oncol 6: 483-488, 1995

18. Steyerberg EW, Keizer HJ, Stoter G, Habbema JDF. Predictors of residual mass histology following chemotherapy for metastatic nonseminomatous testicular cancer: a quantitative overview of 996 resections. Eur J Cancer, 30A, 1231-1239, 1994
19. 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 6 study groups. J Clin Oncol 13: 1177-1187, 1995
20. Fosså SD, Aass N, Ous S, et al. Histology of tumour residuals following chemotherapy in patients with advanced nonseminomatous testicular cancer. J Urol 142: 1239-1242, 1989
21. Fosså SD, Ous S, Lien HH, et al. Post-chemotherapy lymph node histology in radiologically normal patients with metastatic nonseminomatous testicular cancer. J Urol 141: 557-559, 1989
22. Geldeman WAH, Schraffordt Koops H, Sleijfer DTh, et al. Treatment of retroperitoneal residual tumour after PVB chemotherapy of nonseminomatous testicular tumors. Cancer 58: 1418-1421, 1986
23. Nijman JM, Schraffordt Koops H, Kremer J, et al. Gonadal function after surgery and chemotherapy in men with stage II and III nonseminomatous testicular tumors. J Clin Oncol 5: 651-656, 1987
24. De Graaf WE, Oosterhuis JW, Van der Linden S, Homan van der Heide JN, Schraffordt Koops H, Sleijfer DTh. Residual mature teratoma after chemotherapy for nonseminomatous germ cell tumors of the testis occurs significantly less often in lung than in retroperitoneal lymph node metastases. J Urogen Pathol 1:75-81, 1991
25. Sox HCJr, Blatt MA, Higgins MC, Marton KI. Medical decision making. Butterworths, Boston, 1988
26. Van der Schouw YT, Straatman H, Verbeek ALM. ROC curves and the areas under them for dichotomized tests: emperical findings for logistically and normally distributed diagnostic test results. Med Decis Making 14: 374-381, 1994
27. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 78: 316-331, 1983
28. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 247: 2543-2546, 1982
29. Hilden J. The area under the ROC curve and its competitors. Med Decis Making 11: 95-101, 1991
30. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrica 12: 153-157, 1947
31. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 299: 926-930, 1978
32. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 39: 207-215, 1983
33. Jansen RLH, Sylvester R, Sleijfer DT, et al. Long-term follow-up of non-seminomatous testicular cancer patients with mature teratoma or carcinoma at postchemotherapy surgery. Eur J Cancer 27: 695-698, 1991

8 Residual pulmonary masses following chemotherapy for metastatic nonseminomatous germ cell tumor: prediction of histology

E.W. Steyerberg, H.J. Keizer, J.E. Messemer, G.C. Toner, H. Schraffordt Koops, S.D. Fossâ, A. Gerl, D.T. Sleijfer, R.S. Foster, J.P. Donohue, D.F. Bajorin, G.J. Bosl and J.D.F. Habbema
Submitted for publication

Abstract

Purpose: The goal of this analysis was to predict the histology (necrosis, mature teratoma or cancer) of residual pulmonary masses and to support decision making on the need and sequence of thoracotomy and retroperitoneal lymph node resection.
Patients and Methods: Individual patient data were available of 215 patients who had a thoracotomy following cisplatin-based induction chemotherapy for metastatic testicular nonseminomatous germ cell tumor. Logistic regression analysis was used to estimate the probability of necrosis, mature teratoma and cancer in relation to predictors known before resection.
Results: Of the 215 patients, 116 (54%) had necrosis at thoracotomy, 70 (33%) mature teratoma, and 29 (13%) cancer. Necrosis was found at thoracotomy in 89% of the patients with necrosis at retroperitoneal lymph node dissection (RPLND). Other predictors included the primary tumor histology, prechemotherapy tumor marker levels, change in mass size during chemotherapy, and the presence of a single, unilateral mass. Multivariate combination of predictors yielded reliable models (goodness-of-fit tests: $p>.20$), which discriminated necrosis well from other histologies, especially if RPLND histology was available (area under the receiver operating characteristic (ROC) curve .86).
Conclusions: This analysis indicates subgroups of patients with a high probability of necrosis and a low risk of cancer, where close follow-up of the residual pulmonary mass might be considered. In most patients, a RPLND should be performed before a thoracotomy is considered, since the probability of necrosis is generally higher at thoracotomy than at RPLND and the histology at RPLND is a strong predictor of the histology at thoracotomy.

8.1 Introduction

Surgical resection of residual masses is a generally accepted treatment after chemotherapy for metastatic testicular nonseminomatous germ cell tumor (NSGCT)[1,2]. A thoracotomy to resect residual pulmonary nodules is performed in about 10% of the patients[3,4,5,6,7,8,9,10], while retroperitoneal lymph node dissection is the more common procedure (frequency between 25%[11] and 86%[7]) and resection of other sites (liver, neck, bone, brain) is less frequent. The histology of the resection specimen may reveal necrosis/fibrosis, mature teratoma or cancer. Resection of necrosis (totally benign tissue) is considered to have no therapeutic benefit, in contrast to resection of mature teratoma or cancer. Therefore, attempts have been made to identify patients pre-operatively with necrosis only at thoracotomy, who might be spared the burden of an unnecessary resection[3,12,13].

The selection of patients for pulmonary resection is generally based on the presence of persisting lung nodules after chemotherapy, while serum tumor markers have normalized. However, subgroups might be defined where the risk of leaving mature teratoma or cancer unresected is so small that it does not outweigh the disadvantages of resection (morbidity, mortality, financial costs). These subgroups might be defined by previously recognized predictors of necrosis[14], such as the absence of teratoma elements in the primary tumor[3,4,5,6,8,12,13,15,16,17], and prechemotherapy tumor marker levels (AFP[16,17], HCG[16,17], LDH[3]). Also, the histology as found at RPLND has proved to be a strong predictor of the histological content of pulmonary masses[12,13,18,19].

In this analysis our aim is to estimate the probabilities of necrosis, mature teratoma and cancer in residual pulmonary masses, based on these predictors. We will consider both the situation that a RPLND was performed before thoracotomy and the situation that the RPLND histology is not available at the time when the decision to perform a thoracotomy is made. To obtain a sufficiently high number of patients for statistical analyses, we have collected an international data set comprising data from several study groups.

8.2 Patients and methods

8.2.1 Patients

The international data set consisted of patients with metastatic nonseminomatous testicular cancer, including patients with histologically pure seminoma and elevated prechemotherapy serum tumor markers. All patients underwent resection of residual lung masses after induction chemotherapy with cisplatin-based chemotherapy. Exclusion criteria were: elevated levels of tumor markers AFP or HCG at the time of thoracotomy; extragonadal primaries; pure seminoma; resection after relapse of tumor following initial chemotherapy.

Seven study groups contributed individual data on 215 patients. The numbers of patients in each study group were as follows: group 1 (Memorial Sloan Kettering Cancer Center), N=39[3], group 2 (Norwegian Radium Hospital), N=22[12], group 3 (University Hospital Groningen), N=23[20], group 4 (University Hospital Nijmegen), N=7 [5], group 5

(Indiana University Medical School), N=71[18], group 6 (University Hospital Munich), N=26[13], group 7, N=27 (University Hospitals Leiden, Rotterdam, Amsterdam)[10]. The numbers of patients in studies #2, #3, #6 and #7 were extended since the original publications. Study #5 contained all patients with both RPLND and thoracotomy at Indiana between 1977 and 1994.

The individual patient data consisted of basic patient identification, year of treatment, information on potential predictors and the histological outcome of pulmonary resection. The predictors included the primary tumor histology (teratoma-positive or teratoma-negative[3,15]), prechemotherapy tumor marker levels of AFP, HCG and LDH (elevated or normal according to normal values of each center), prechemotherapy and postchemotherapy maximum transversal size as measured on CT scan, location and number of metastases, and the histology of retroperitoneal lymph node dissection (RPLND), if performed shortly before or after the pulmonary resection. The pulmonary histology was classified according to the worst histology, either as cancer, mature teratoma, or necrosis. If multiple thoracotomies were performed, the first one only was considered for analysis.

8.2.2 Missing values

Missing values were present in a large number of the patients (190/215). Group 5 missed prechemotherapy LDH values in the majority of cases (68/71), and had registered the location (unilateral/bilateral) and number (single/multiple) of lung metastases instead of size. Group 1 to 4 had registered the prechemotherapy and postchemotherapy size (in mm) of lung masses, which were used to calculate the shrinkage of the mass, but not the location nor number of lung metastases. Group 6 and 7 had registered both size, location and number. Primary histology was missing in 4 patients and prechemotherapy AFP and HCG in 9. The RPLND histology was available for all patients from group 5 and 6. RPLND was not performed in 56 of the 118 patients in the other studies.

Missing values were imputed with regression techniques, using the correlation of the variable with missing values and other predictors[21,22]. Linear regression was used for continuous predictors, e.g. tumor marker levels, and logistic regression for dichotomous predictors, e.g. presence of teratomatous elements in the primary tumor (SPSS/PC+ software; SPSS Inc, Chicago, IL). It appeared that the prechemotherapy LDH level was significantly correlated with prechemotherapy HCG levels and the presence of teratoma elements in the primary tumor. The presence of a single, unilateral metastasis was negatively correlated with prechemotherapy HCG levels. No predictors correlated with an increase in size during chemotherapy. Therefore, the mean prevalence of an increase in size was imputed if no size measurements were available. Teratomatous elements were found more often in the primary tumor if AFP was elevated. No attempt was undertaken to impute the RPLND histology values in the 56 patients where no RPLND was performed.

Excluded from multivariate analysis were those patients with all three prechemotherapy tumor markers missing (both AFP, HCG and LDH missing, N=9).

In the remaining 206 patients, LDH levels were imputed in 83, 'increase' in 69, 'location and number of metastases' in 97 and primary tumor histology in 4. 110 cases had imputation of one value, 70 of 2 values and 1 of 3 values.

8.2.3 Statistical analysis

The histology at resection (necrosis, mature teratoma, cancer) was predicted using two statistical models. The first model estimated the probability of necrosis by comparing patients with necrosis at resection with patients showing other histologies (teratoma or cancer). The second model aimed to distinguish between cancer and teratoma in the patients who did not have necrosis at resection. This second analysis estimated the ratio of cancer and teratoma or the *relative* probability of cancer. The combination of the estimates of the two models results in predicted probabilities for necrosis, mature teratoma and cancer at thoracotomy.

The probability of necrosis and the relative probability of cancer were related to patient characteristics known before resection (predictors). The Odds Ratio (OR) was used as the measure of association. Univariate relations between predictors and outcomes were estimated within each study and pooled using the Mantel-Haenszel method, provided that no major heterogeneity between studies was found (test for homogeneity $p>0.10$). Predictors have statistically significant effects ($p<0.05$) if the 95% confidence interval (CI) of the OR does not include the value 1. Logistic regression analysis was applied to estimate the prognostic effect of combinations of predictors, with the study number as the grouping variable in conditional logistic regression analyses (EGRET statistical package; Statistics and Epidemiology Research Corporation, Seattle, WA). Selection of variables was based on statistical grounds (p-value and magnitude of OR in univariate and multivariate analyses), the number of missing values, and agreement with a previous analysis on retroperitoneal mass histology[21]. A conservative attitude was taken towards the inclusion of interaction terms in the models. We considered the overall p-value of all two-way interaction terms of the predictors in a model (criterium $p<0.10$) and of specific two way interaction terms (criterium $p<0.05$).

Since the number of missing values was substantial, we made a comparison with an alternative analysis using no imputed values at all. In this analysis, we calculated univariate and multivariate logistic regression coefficients in subgroups of patients with complete values for a set of predictors. The changes in coefficients were subsequently used to adapt the original univariate coefficients[23]. The analysis with imputed values yielded more conservative estimates (ORs closer to 1) and was therefore preferred over this alternative method. Also, the correlations between most predictors were relatively small (Pearson correlation coefficients <0.20). The multivariate ORs were therefore expected to be close to the univariate ORs, which was indeed found in the analysis with imputed values.

8.2.4 Evaluation of model performance

Predictive accuracy of the multivariate models can be distinguished in reliability (or calibration) and discrimination. Reliability refers to the amount of agreement between predicted and observed outcomes. If, for instance, patients with certain characteristics are predicted to have a 70% chance of necrosis at resection, then 70% of such patients should actually have necrosis at resection. Reliability was tested by the Hosmer-Lemeshow goodness-of-fit test[24] (BMDP module LR; BMDP statistical software, Inc. Los Angeles, CA).

Discrimination was assessed using the area under the receiver operating characteristic (ROC) curve, which forms a suitable single number to summarize the discriminative ability of a predictive model[25,26]. A useless predictive model, such as a coin flip, would yield an area of 0.5. When the area is 1.0, the model discriminates perfectly. We interpret a value over 0.6 as reasonable, over 0.7 as satisfactory and over 0.8 as good. An indication of the discriminative ability of the models in future patients was obtained with bootstrapping techniques[27]. Random bootstrap samples were drawn with replacement from the full sample consisting of all patients (200 replications). Models were estimated on these bootstrap samples and evaluated on the full sample. With the same procedure, a correction for overoptimism[28] was estimated for the logistic regression coefficients in the full sample. Correction of the coefficients for RPLND histology was excluded because of the strong predictive value of RPLND histology, which was already known before this analysis[12,13,18,19].

8.3 Results

A detailed overview of the distribution of patient characteristics in each of the 7 participating study groups is shown in the Appendix. The differences between the studies were relatively small. Table 1 shows predictors considered in the analysis and the relation of the predictors with the histology as found at pulmonary resection. Overall, nearly half of the patients had a teratoma-negative primary tumor (95/211=45%). Prechemotherapy AFP, HCG and LDH tumor markers were elevated in over two thirds of the patients (69%, 74% and 72%). Around 40% had lung masses ≤ 20mm on pre-chemotherapy CT scan, or ≤ 10 mm on postchemotherapy CT scan. A major reduction in size (≥ 70%) occurred in 20% of the patients, while in 11%, the mass increased during chemotherapy. One third of the patients had a single, unilateral residual mass after chemotherapy (30%). A RPLND was performed in 159 patients, showing necrosis in 34%, mature teratoma in 52% and cancer in 14% of the patients. At pulmonary resection, these histologies were found in 54%, 33% and 13% respectively. Necrosis thus was more frequent and mature teratoma less frequent at thoracotomy than at RPLND (54% vs 34% and 33% vs 52%, $p<.001$). The relative probability of cancer was 29/(29+70)=29% on average at thoracotomy.

Table 2 shows the results of the univariate analysis of the relations of the predictors with the histology of pulmonary nodules. Odds Ratios (ORs) were calculated with aggregated categories of a predictor if the ORs between categories were similar, or if the

Table 1 Overview of the predictors considered and the relations with the histology at pulmonary resection.

	Necrosis N=116 (54%)	Teratoma N=70 (33%)	Cancer N=29 (14%)	Total N=215 (100%)
Primary tumour histology				
Teratoma-negative	63 (66%)	16 (17%)	16 (17%)	95/211 (45%)
Teratoma-positive	50 (43%)	53 (46%)	13 (11%)	116/211 (55%)
Pre-AFP normal (ng/ml)	42 (67%)	13 (21%)	8 (13%)	63/206 (31%)
norm-100	13 (43%)	15 (50%)	2 (7%)	30/206 (15%)
100-1000	36 (55%)	22 (34%)	7 (11%)	65/206 (32%)
>1000	21 (44%)	17 (35%)	10 (21%)	48/206 (23%)
Pre-HCG normal (IU/l)	24 (44%)	20 (37%)	10 (19%)	54/206 (26%)
norm-100	21 (44%)	23 (48%)	4 (8%)	48/206 (23%)
100-1000	16 (50%)	12 (38%)	4 (13%)	32/206 (16%)
1000-10000	20 (59%)	8 (24%)	6 (18%)	34/206 (17%)
>10000	31 (82%)	4 (11%)	3 (8%)	38/206 (18%)
Pre-LDH normal (U/l)	14 (40%)	18 (51%)	3 (9%)	35/126 (28%)
1 - 2 x normal	16 (52%)	10 (32%)	5 (16%)	31/126 (25%)
2 - 4 x normal	17 (63%)	5 (19%)	5 (19%)	27/126 (21%)
> 4 x normal	22 (67%)	7 (21%)	4 (12%)	33/126 (26%)
Prechemotherapy size				
0 - 20 mm	22 (41%)	23 (43%)	9 (17%)	54/138 (39%)
21 - 50 mm	42 (71%)	12 (20%)	5 (9%)	59/138 (43%)
> 50 mm	14 (56%)	8 (32%)	3 (12%)	25/138 (18%)
Postchemotherapy size				
0 - 10 mm	33 (57%)	16 (28%)	9 (16%)	58/140 (41%)
11 - 20 mm	27 (61%)	13 (30%)	4 (9%)	44/140 (31%)
> 20 mm	19 (50%)	15 (39%)	4 (11%)	38/140 (27%)
Shrinkage				
≥ 70%	15 (56%)	7 (26%)	5 (19%)	27/138 (20%)
0 - 69%	60 (63%)	27 (28%)	9 (9%)	96/138 (70%)
< 0% (increase)	3 (20%)	9 (60%)	3 (20%)	15/138 (11%)
Location and number of mets				
Unilateral, single	21 (60%)	13 (37%)	1 (3%)	35/117 (30%)
Unilateral, multiple	7 (29%)	12 (50%)	5 (21%)	24/117 (21%)
Bilateral	34 (59%)	14 (24%)	10 (17%)	58/117 (50%)
Histology at RPLND				
necrosis	48 (89%)	4 (7%)	2 (4%)	54/159 (34%)
mature teratoma	31 (38%)	42 (51%)	9 (11%)	82/159 (52%)
cancer	7 (30%)	5 (22%)	11 (48%)	23/159 (14%)

number of patients in a category was very small. All ORs were reasonably homogenous between studies ($p>0.10$). If the primary tumor was teratoma-negative, thoracotomy showed mature teratoma infrequently (Table 1: 17%). Hence, both necrosis and cancer were found more often compared with patients with teratoma-positive primary tumors (OR 2.58 and OR 3.84, respectively). In case of a normal prechemotherapy level of AFP, both necrosis and cancer were found slightly more often than with elevated levels of AFP. Higher prechemotherapy values of HCG or LDH showed a clear relation with

Table 2 Univariate analysis of the relations of predictors with the histology at pulmonary resection.

	OR Necrosis vs Other [95% CI][$]	p-value	OR Cancer vs Teratoma [95% CI][$]	p-value
Primary tumour histology				
teratoma-negative	2.58 [1.4-5.0]	p = .001	3.84 [2.4-29]	p=.004
teratoma-positive	1.0			
Pre-AFP normal	1.90 [.98-3.8]	p = .045	1.85 [1.0-13]	p = .26
elevated	1.0		1.0	
Pre-HCG norm-1000 IU/l	1.0	p < .001	1.0	p = .10
1000-10000	1.89 [.78-4.5]		}	
>10000	5.42 [2.0-16]		} 2.22 [.72-11]	
Pre-LDH normal	1.0	p = .018	1.0	p = .18
1 - 2 x normal	1.61 [.87-12]		2.70 [.66-352]	
2 - 4 x normal	2.23 [.60-8.9]		}	
> 4 x normal	3.67 [? - ?]		} 2.91 [? - ?]	
Prechemotherapy size		p = .08		p = .80
0 - 20 mm	1.0		1.0	
21 - 50 mm	3.39 [1.5-9.0]		1.70 [.23-13]	
> 50 mm	1.49 [.46-5.2]		1.03 [.09-10]	
Postchemotherapy size		p = .41		p = .46
0 - 10 mm	1.0		1.0	
11 - 20 mm	1.10 [.43-2.8]		.39 [.04-2.2]	
> 20 mm	.65 [.24-1.7]		.49 [.06-2.7]	
Shrinkage ≥0	1.0	p = .003	1.0	p = .98
< 0% (increase)	0.13 [.02-.62]		.97 [.15-6.0]	
Location and number of mets				
Unilateral, single	1.51 [.62 -3.7]	p = .32	0.12 [.01 - 1.1]	p = .036
Uni/bilateral or multiple	1.0		1.0	
Histology at RPLND		p < .001		p = .001
necrosis	1.0		1.0	
mature teratoma	.07 [.03-.24]		1.0	
cancer	.03 [.01-.28]		15.6 [3.3 - 133]	

Abbreviations: OR, Odds Ratio; 95% CI, 95% confidence interval

[$] Some studies had empty cells, which made the computation of the confidence intervals impossible for some of the predictors (95% CI indicated with [?-?]) and made the upper limits of the confidence intervals irrealistically high for some other predictors.

a more frequent finding of necrosis at resection ($p<0.001$ and $p<0.018$ respectively), but also with a relatively more frequent finding of cancer ($p=0.10$ and $p=0.18$ respectively). Neither prechemotherapy size nor postchemotherapy size showed a clear association with the histology, but an increase in size indicated a low likelihood of necrosis at resection (OR=0.13). A unilateral, single metastasis had a slightly higher probability of necrosis (60% vs 50%) and a much lower risk of cancer (3% vs 18%), compared to patients with multiple or bilateral metastases. The histology as found at RPLND proved to be the strongest predictor of the histological content of a residual pulmonary mass.

Of 54 patients with necrosis at RPLND, 48 (89%) also had necrosis at thoracotomy. In case of mature teratoma or cancer at RPLND, only 38% or 30% had necrosis at thoracotomy respectively. In 101 patients (64%), the histology at RPLND and at thoracotomy were identical.

The results of the multivariate logistic regression analyses are presented in Table 3. Models were constructed with and without available RPLND histology. The model predicting necrosis without knowledge of the RPLND histology included 5 characteristics: primary tumor histology, prechemotherapy AFP, prechemotherapy HCG, increase in size, and number and location of residual metastases. The prechemotherapy LDH level was not included in the model as LDH level was of minor prognostic importance once the HCG level was taken into account ($p>0.50$). Inclusion of the RPLND histology substantially improved the model ($p<0.001$). Statistically significant interaction was found between the RPLND histology and the predictors 'prechemotherapy HCG level' and 'number and location of residual metastases' (both $p<0.001$). If teratoma or cancer was found at RPLND, the probability of necrosis clearly increased with higher HCG levels or in the presence of a single, unilateral metastasis (Table 3: OR=5.5 and OR=8.1 respectively). If necrosis was found at RPLND, these variables had no incremental effect on the (already high) probability of necrosis. Once the RPLND histology was known, the predictive value of a teratoma-negative primary tumor decreased, which is explained by its correlation with the finding of necrosis at RPLND[21]. Prechemotherapy AFP level was left out of the model because of its neglectable predictive influence once the RPLND histology was known.

Table 3 Multivariate analysis of the relations of predictors with the histology at pulmonary resection.

		Necrosis vs Other		Cancer vs Teratoma	
		Without RPLND N=206	With RPLND N=150	Without RPLND N=94	With RPLND N=68
*Predictors**					
RPLND histology	teratoma	-	0.02 [.01 - .08]	-	-
	cancer	-	0.01 [.00 - .06]	-	9.6 [1.9 - 49]
Primary tumour ter.-negative vs positive		3.6 [1.8-7.1]	2.2 [.81 - 5.9]	4.4 [1.5 - 13]	1.7 [.41 - 6.8]
Prechemotherapy AFP norm. vs elev.		1.7 [.85-3.5]	-	1.4 [.43 - 4.4]	-
Prechemotherapy HCG high vs low [$]		3.0 [1.8-4.9]	5.5 [2.5 - 12]**	2.8 [.78 - 10]	4.5 [.93 - 22]
Increase in size vs decrease/same size		.12 [.03-.54]	.24 [.03 - 1.8]	-	-
Single metastasis vs multiple/bilateral		3.1 [1.1-8.5]	8.1 [2.2 - 30]**	.13 [.01 - 1.3]	.23 [.02 - 2.4]
Model performance					
Goodness-of-fit [#]		p-value .39	p-value .98	p-value .52	p-value .93
Discriminative ability [¶]		area .77	area .86	area .73	area .75
Correction factor for overoptimism		.90	.86	.77	.79

* The multivariate Odds Ratios are shown with the corresponding 95% confidence intervals for the models predicting necrosis and distinguishing cancer from teratoma, both with and without RPLND histology.

[$] For necrosis model coded as 0 (normal to 1000 IU/l), 1 (1000 to 10000 IU/l), 2 (>10,000 IU/l); for cancer model coded as 0 (normal to 1000 IU/l), 1 (>1000 IU/l); ** Odds Ratio in patients with teratoma/cancer at RPLND; [#] Hosmer-Lemeshow test; [¶] Area under the ROC curve

The relative probability of cancer was related to the primary tumor histology, prechemotherapy AFP, prechemotherapy HCG, and number and location of residual metastases. Knowledge of the RPLND histology improved the model significantly ($p=0.007$).

Goodness-of-fit was adequate for all models (Table 3: $p>0.20$). Discriminative ability was expressed as the area under the ROC curve. Discriminative ability is expected to be good for the model predicting necrosis including the RPLND histology (area 0.86), while the model without this predictor is expected to perform satisfactory (area 0.77). The models for cancer vs teratoma are also expected to perform satisfactory, with a modest improvement in prediction by knowledge of the RPLND histology (area increases from 0.73 to 0.75).

Table 4 presents the estimated probabilities of necrosis, mature teratoma and cancer for all combinations of predictors, provided that the lung metastases did not increase in size during chemotherapy. In case of an increase, the probability of necrosis is generally low (<60%). The probabilities were calculated with the multivariate ORs as shown in Table 3, which were multiplied by correction factors to compensate for overoptimism of the models. The presented probabilities have a range of uncertainty, which necessitates a cautious interpretation. Table 5 shows the 95%-confidence intervals for the predicted probabilities of necrosis in the various subgroups. Some subgroups contain a small number of patients (e.g. only 10 patients had HCG >1000 IU/L and a single, unilateral residual nodule), which is reflected in the relatively wide confidence intervals. Despite these limitations, Table 4 provides a simple way to estimate the histological content of residual lung masses. For example, it can be read that a patient with a decrease in size of the mass, a teratoma-negative primary tumor, prechemotherapy AFP elevated, HCG <1000 IU/L, and a single metastasis after chemotherapy, has predicted probabilities of necrosis, mature teratoma and cancer of 76%, 20% and 4% respectively. These probabilities change to 93%, 6% and 1% if necrosis was found at RPLND. From Table 5 it may be read that the initial estimate of a 76% chance of necrosis is very likely between 57 and 87%, while the 95% confidence interval for the estimate after finding necrosis at RPLND (93%) ranges from 82 to 97%.

Table 4 Predicted probabilities of necrosis, mature teratoma and cancer at resection of residual lung metastases, which did not increase during chemotherapy.

Primary tumor histology		Teratoma-positive						Teratoma-negative					
Prechemotherapy HCG		<1000		1000-10000		>10000 IU/l		<1000		1000-10000		>10000 IU/l	
Single, unilateral residual nodule?		Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Prechemotherapy AFP normal	Nec	61%	36%	81%	60%	92%	80%	83%	64%	93%	83%	97%	93%
	Ter	36%	46%	16%	21%	7%	10%	13%	16%	4%	5%	2%	2%
	Can	3%	18%	3%	19%	1%	9%	3%	20%	3%	13%	1%	5%
Prechemotherapy AFP elevated	Nec	49%	26%	72%	48%	87%	71%	76%	53%	89%	75%	96%	89%
	Ter	48%	56%	24%	31%	11%	17%	20%	24%	7%	8%	3%	4%
	Can	3%	18%	4%	21%	2%	12%	4%	23%	3%	17%	1%	8%
RPLND: necrosis	Nec	87%	87%	87%	87%	87%	87%	93%	93%	93%	93%	93%	93%
	Ter	12%	11%	11%	8%	11%	8%	6%	5%	5%	3%	5%	3%
	Can	1%	2%	2%	5%	2%	5%	1%	2%	2%	4%	2%	4%
RPLND: mature teratoma	Nec	47%	13%	79%	39%	94%	73%	64%	22%	88%	55%	97%	84%
	Ter	50%	72%	17%	36%	5%	16%	33%	59%	9%	22%	2%	8%
	Can	3%	15%	4%	25%	1%	11%	3%	19%	3%	23%	1%	8%
RPLND: cancer	Nec	31%	7%	66%	24%	89%	58%	47%	13%	79%	39%	94%	73%
	Ter	49%	41%	15%	15%	5%	8%	33%	30%	7%	8%	2%	4%
	Can	20%	52%	19%	61%	6%	34%	20%	57%	14%	53%	4%	23%

Table 5 Predicted probabilities of necrosis with 95% confidence intervals.

Primary tumor histology		Teratoma-positive						Teratoma-negative					
Prechemotherapy HCG		<1000		1000-10000		>10000 IU/l		<1000		1000-10000		>10000 IU/l	
Single, unilateral residual nodule?		Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
	Nec	61%	36%	81%	60%	92%	80%	83%	64%	93%	83%	97%	93%
Prechemotherapy AFP normal	[95%-CI]	41-78	21-57	61-92	43-76	83-98	62-90	72-94	42-83	82-98	72-91	92-100	84-97
	Nec	49%	26%	72%	48%	87%	71%	76%	53%	89%	75%	96%	89%
Prechemotherapy AFP elevated	[95%-CI]	31-68	16-39	52-85	34-61	67-96	49-86	57-87	36-67	68-95	60-85	87-99	76-95
	Nec	87%	87%	87%	87%	87%	87%	93%	93%	93%	93%	93%	93%
RPLND: necrosis	[95%-CI]	69-95	69-95	69-95	69-95	69-95	69-95	82-97	82-97	82-97	82-97	82-97	82-97
	Nec	47%	13%	79%	39%	94%	73%	64%	22%	88%	55%	97%	84%
RPLND: mature teratoma	[95%-CI]	26-69	5-25	56-90	24-57	85-98	48-89	36-84	9-45	59-99	42-68	86-100	58-95
	Nec	31%	7%	66%	24%	89%	58%	47%	13%	79%	39%	94%	73%
RPLND: cancer	[95%-CI]	14-60	2-26	34-89	8-48	57-99	43-72	24-71	4-27	54-91	22-60	83-99	46-90

8.4 Discussion

In this study, we analyzed over 200 testicular cancer patients undergoing thoracotomy to remove residual lung nodules after cisplatin-based chemotherapy. The histology at resection was estimated using several well-known predictors in statistical models. The large number of patients enabled an accurate definition of subgroups with a high likelihood of necrosis at resection and/or a low risk of cancer.

The predictors included the absence of teratoma elements in the primary tumor, prechemotherapy tumor marker levels, increase of the mass during chemotherapy, the presence of a single, unilateral residual mass, and the histology at RPLND (if available). The relations of the predictors with the histology were in general similar to our previous analysis on retroperitoneal mass histology[21]. For example, predictors for necrosis were a teratoma-negative primary tumor, normal prechemotherapy AFP, elevated LDH. Also, an increase in size during chemotherapy indicated a low likelihood of necrosis (3/15=20%). This predictor was not as strong as in case of retroperitoneal residual disease, where an increase excluded the finding of necrosis (0/42). Remarkably, the size of the nodules (prechemotherapy or postchemotherapy) was not related to the histology. This is in contrast to retroperitoneal masses, where postchemotherapy size is a strong predictor[21], but in agreement with a previous analysis[3]. Probably the small size of most nodules (96% of residual masses < 50mm) also accounts for the absence of a clear association. Prechemotherapy high values of HCG were a strong predictor of necrosis ($p<0.001$), but the effect was opposite to that in case of RPLND[21]. This discrepancy may partly be explained by the higher frequency of high HCG levels in the thoracotomy patients (18% vs 7% with values >10.000 IU/L). Upon re-examination of the RPLND data we found a slightly higher probability of necrosis in patients with HCG >10.000 IU/L compared to patients with HCG between 1000 and 10.000 IU/L. The magnitude of the effect in this group was however so small (OR 1.3), that it was not considered previously in the classification of HCG values (normal versus elevated)[21]. A fortuitous sign was the presence of a single, unilateral residual lung nodule, which was both associated with a higher probability of necrosis as well as with a lower probability of cancer. The strongest predictor of the pulmonary histology was the histology as found at RPLND. If RPLND showed necrosis only, the probability of necrosis at thoracotomy was 89%. If RPLND showed mature teratoma or cancer, this probability was much lower (38% and 30% respectively).

The results of the statistical analyses were tabulated in such a way that the probability of necrosis, mature teratoma, and cancer could be determined easily (Table 4). Subgroups with a high likelihood of necrosis can directly be distinguished in this table. If the RPLND histology was necrosis and the primary tumor was teratoma-negative, the predicted probability of necrosis at thoracotomy is as high as 93%. In case of a teratoma-positive tumor, the probability is slightly lower (87%). The actually observed frequencies largely agree with these predictions: 94% (30/32) and 82% (14/17). Previously, careful analysis of a small number of patients already led to the hypothesis that those patients with necrosis at RPLND and a teratoma-negative primary tumor

formed a favorable subgroup[12]. The present analysis confirms this hypothesis on a much larger number of patients, providing strong argument for debate on the need of resection in this subgroup. Other subgroups with a high likelihood of necrosis could be defined even among those patients with mature teratoma or cancer at RPLND. These subgroups are infrequent, making the predictions uncertain. The infrequency also limits the relevance for clinical practice in general. On the other hand, if a thoracotomy is considered in a patient from such a rare subgroup, the predictions may play a role in the decision making process, especially if surgical risk is increased or the procedure is technically difficult.

In the situation where RPLND is not performed before thoracotomy, the order and necessity of these two procedures merits attention. For thoracotomy, a few favorable subgroups can be indicated without availability of the RPLND histology by combinations of predictors such as a teratoma-negative primary tumor, elevated HCG, a single metastasis, or normal AFP (Table 4). For RPLND, the predicted probability of necrosis can be estimated with the prognostic score chart as presented before [21]. In most patients, the likelihood of necrosis at RPLND is lower than at thoracotomy. This logically leads to the general rule that a RPLND should be performed before a thoracotomy is considered. If the RPLND shows necrosis only, the need of thoracotomy can be reconsidered, while teratoma or cancer at RPLND argue for a thoracotomy in most instances. If the order of procedures is reversed, RPLND cannot be omitted safely in most patients, even if necrosis is found at thoracotomy[13]. Thus, the advantage of performing a RPLND first is that the patients with necrosis at RPLND might be spared a thoracotomy. Note that this advantage is missed if a thoracotomy is combined with a RPLND in one session.

Three exceptions on this general guideline can be thought of. Firstly, the risk of missing teratoma or cancer at thoracotomy may be judged too high, even if RPLND would show necrosis. In this case, the order of the procedures can freely be decided on by the treating physicians, and a combined procedure for RPLND and thoracotomy might be considered. Secondly, some patients may have a higher predicted probability of necrosis at RPLND than at thoracotomy. As an example, we may think of patients with a teratoma-negative primary tumor, normal postchemotherapy CT abdomen, but multiple residual lung metastases on CT thorax. In such patients, thoracotomy might be preferred as the first procedure. Finally, both RPLND and thoracotomy may be considered unwarranted. This may for example be the case in a patient with a teratoma-negative primary tumor, a normal postchemotherapy CT abdomen, and a single residual lung mass on CT thorax. With respect to the order of RPLND and thoracotomy it can thus be concluded that the procedure with the lowest likelihood of necrosis should be performed first, unless both or none of the procedures are considered necessary.

Thus far, subgroups of patients were described according to the probability of necrosis. However, the relative risks of cancer and mature teratoma should also be considered. For example, if RPLND showed necrosis and the primary tumor was teratoma-negative, Table 4 shows that the probability of necrosis is 93%, while the risk of cancer may vary between 1 and 4%, depending on the values of other predictors.

Obviously, a thoracotomy may more readily be omitted in the patient with the lower risk of cancer.

The exact thresholds for resection (minimum probability of necrosis, maximum risk of cancer) are difficult to determine, but should weigh the expected benefits, risks and financial costs of resection compared to an observation strategy with frequent follow-up (physical examination, tumor marker measurements, CT scanning). The risks of thoracotomy include short-term morbidity (hospital stay and complications like pneumothorax, pneumonia) and mortality (around 1%[29,30]). The benefits of thoracotomy relate to the patients with mature teratoma or cancer. If resection is performed, the prognosis generally is favorable with 5 year relapse free survival over 85% after resection of mature teratoma [5,8,10,11], and 50%[5,8,11,31,32,33] to 70/80%[10,17,34] after resection of cancer. If resection is not performed, masses containing mature teratoma may start to grow during a follow-up of months, or even years ('growing teratoma syndrome')[35]. Also, a risk of malignant transformation has been reported[36,37,38,39]. Leaving masses with residual cancer unresected is considered to increase the risk of relapse substantially. Malignant relapses can be treated with salvage chemotherapy regimens, which have rather limited efficacy (around 25%[1,2,9,40,41]). Moreover, salvage treatment may be even less effective in patients presenting with pulmonary metastases, since a 5-year survival of only 6% has been reported for this group of patients[41]. Other considerations include technical aspects of surgery, the patient's personal preferences and country- or center-specific circumstances like the feasibility of frequent follow-up visits with high-quality CT scanning[16].

In conclusion, this analysis may assist in decision making on the necessity and order of pulmonary residual mass resection. The necessity of thoracotomy is doubtful in a number of subgroups where the probability of necrosis is high and the risk of cancer is low. Patients in these subgroups might benefit more from close follow-up than resection. With respect to the order of sequential resections, RPLND should generally be performed before a thoracotomy is considered, because a purely benign histology at RPLND is highly predictive for necrosis in residual lung nodules. Decision making on residual mass resection however remains complex, as it should take into account the potential benefits, technical feasibility, morbidity, and mortality of resection, feasibility of close follow-up, financial costs, and the patient's individual preferences.

Appendix

Table A Patient characteristics per study group.

		group 1 N=39	group 2 N=22	group 3 N=23	group 4 N=7	group 5 N=71	group 6 N=26	group 7 N=27	TOTAL N=215
Primary tumour histology									
Teratoma-positive		21 (55%)	11 (50%)	7 (30%)	2 (29%)	35 (49%)	10 (43%)	13 (48%)	95/211 (45%)
Prechemotherapy markers									
AFP	elevated	25 (64%)	16 (73%)	20 (87%)	5 (83%)	43 (68%)	19 (73%)	15 (56%)	143/206 (69%)
	median (ng/ml)	78	200	130	434	-	80	159	120 ng/ml
HCG	elevated	24 (62%)	15 (68%)	18 (78%)	4 (67%)	50 (79%)	24 (92%)	17 (63%)	152/206 (74%)
	median (IU/l)	6	44	11	41	-	44000	120	48 IU/l
LDH	elevated	27 (69%)	16 (73%)	12 (52%)	1 (17%)	3 (100%)	16 (94%)	16 (100%)	91/126 (68%)
	median (U/l)	431	906	243	231	-	631	252	431 U/l
Prechemotherapy size									
0 - 20 mm		14 (36%)	8 (36%)	12 (52%)	3 (43%)	-	6 (23%)	11 (48%)	54/138 (39%)
21 - 50 mm		13 (35%)	6 (27%)	9 (39%)	4 (57%)	-	15 (58%)	12 (52%)	59/138 (43%)
> 50 mm		10 (27%)	8 (36%)	2 (9%)	-	-	5 (19%)	-	25/138 (18%)
Postchemotherapy size									
0 - 10 mm		14 (36%)	9 (41%)	13 (57%)	4 (57%)	-	11 (42%)	7 (30%)	58/140 (41%)
11 - 20 mm		12 (31%)	8 (36%)	4 (17%)	2 (29%)	-	8 (31%)	10 (44%)	44/140 (31%)
> 20 mm		13 (34%)	5 (23%)	6 (26%)	1 (14%)	-	7 (27%)	6 (26%)	38/140 (27%)
Shrinkage									
>= 70%		7 (19%)	5 (23%)	5 (22%)	2 (29%)	-	6 (23%)	2 (9%)	27/138 (20%)
0 - 69.9%		26 (70%)	16 (73%)	14 (61%)	3 (43%)	-	19 (73%)	18 (78%)	96/138 (70%)
< 0% (increase)		4 (11%)	1 (4%)	4 (17%)	2 (29%)	-	1 (4%)	3 (13%)	15/138 (11%)
Location and number of mets									
Unilateral, single		-	-	-	-	21 (30%)	6 (23%)	8 (40%)	35/117 (30%)
Unilateral, multiple		-	-	-	-	11 (16%)	8 (31%)	5 (25%)	24/117 (21%)
Bilateral		-	-	-	-	39 (55%)	12 (46%)	7 (35%)	58/117 (50%)
Histology at RPLND									
Necrosis		13 (57%)	5 (50%)	5 (33%)	1 (33%)	17 (24%)	8 (31%)	5 (46%)	54/159 (34%)
Mature teratoma		8 (35%)	5 (50%)	10 (67%)	1 (33%)	36 (51%)	16 (62%)	6 (55%)	82/159 (52%)
Cancer		2 (9%)	0 (0%)	0 (0%)	1 (33%)	18 (25%)	2 (8%)	0 (0%)	23/159 (14%)
Histology at thoracotomy									
Necrosis		25 (64%)	12 (55%)	7 (30%)	5 (71%)	34 (48%)	16 (62%)	17 (63%)	116/215 (54%)
Mature teratoma		11 (28%)	4 (18%)	13 (57%)	2 (29%)	25 (35%)	8 (31%)	7 (26%)	70/215 (33%)
Cancer		3 (8%)	6 (27%)	3 (13%)	0 (0%)	12 (17%)	2 (8%)	3 (11%)	29/215 (13%)
Year of treatment									
1977 - 1980		9 (23%)	3 (14%)	3 (13%)	0 (0%)	9 (13%)	1 (4%)	1 (4%)	26/215 (12%)
1981 - 1985		24 (62%)	14 (64%)	6 (26%)	6 (86%)	34 (48%)	11 (42%)	7 (26%)	102/215 (47%)
1986 - 1994		6 (15%)	5 (23%)	14 (61%)	1 (14%)	28 (39%)	14 (54%)	19 (70%)	87/215 (41%)

References

1. Einhorn LH. Treatment of testicular cancer: a new and improved model. J Clin Oncol 8: 1777-1781, 1990
2. Peckham M. Testicular cancer. Rev Oncol 1: 439-453, 1988
3. Toner GC, Panicek DM, Heelan RT, et al. Adjunctive surgery after chemotherapy for nonseminomatous germ cell tumors: recommendations for patient selection. J Clin Oncol 8: 1683-1694, 1990
4. Fosså SD, Aass N, Ous S, et al. Histology of tumor residuals following chemotherapy in patients with advanced nonseminomatous testicular cancer. J Urol 142: 1239-1242, 1989
5. Mulders PFA, Oosterhof GON, Boetes C, et al. The importance of prognostic factors in the individual treatment of patients with disseminated germ cell tumours. Br J Urol 66: 425-429, 1990
6. Gelderman WAH, Schraffordt Koops H, Sleijfer DTh, et al. Results of adjuvant surgery in patients with stage III and IV nonseminomatous testicular tumors after cisplatin-vinblastine-bleomycin chemotherapy. J Surg Oncol 38: 227-232, 1988
7. Aass N, Klepp O, Cavillin-Ståhl E, et al. Prognostic factors in unselected patients with nonseminomatous metastatic testicular cancer: a multicenter experience. J Clin Oncol 9: 818-826, 1991
8. Tait D, Peckham MJ, Hendry WF, Goldstraw P. Post-chemotherapy surgery in advanced non-seminomatous germ-cell tumours: the significance of histology with particular reference to differentiated (mature) teratoma. Br J Cancer 50: 601-609, 1984
9. Dearnaley DP, Horwich A, A'Hern R, et al. Combination chemotherapy with bleomycin, etoposide and cisplatinum (BEP) for metastatic testicular teratoma: long-term follow-up. Eur J Cancer 27: 684-691, 1991
10. Steyerberg EW, Keizer HJ, Zwartendijk J, et al. Prognosis after resection of residual masses following chemotherapy for metastatic nonseminomatous testicular cancer: a multivariate analysis. Br J Cancer 68: 195-200, 1993
11. Hendry WF, A'Hern RP, Hetherington JW, et al. Para-aortic lymphadenectomy after chemotherapy for metastatic non-seminomatous germ cell tumours: prognostic value and therapeutic benefit. Br J Urol 71: 208-213, 1993
12. Qvist HL, Fosså SD, Ous S, Høie J, Stenwig AE, Gierckscky KE. Post-chemotherapy tumor residuals in patients with advanced nonseminomatous testicular cancer. Is it necessary to resect all residual masses? J Urol 145: 300-303, 1991
13. Gerl A, Clemm C, Schmeller N, et al. Sequential resection of residual abdominal and thoracic masses after chemotherapy for metastatic non-seminomatous germ cell tumours. Br J Cancer 70: 960-965, 1994
14. Steyerberg EW, Keizer HJ, Stoter G, Habbema JDF. Predictors of residual mass histology following chemotherapy for metastatic nonseminomatous testicular cancer: a quantitative overview of 996 resections. Eur J Cancer 30A: 1231-1239, 1994
15. Donohue JP, Rowland RG, Kopecky K, et al. Correlation of computerized 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
16. Fosså SD, Qvist H, Stenwig AE, et al. Is postchemotherapy retroperitoneal surgery necessary in patients with nonseminomatous testicular cancer and minimal residual tumor masses? J Clin Oncol 10: 569-573, 1992
17. Gerl A, Clemm C, Schmeller N, et al. Outcome analysis after post-chemotherapy surgery in patients with non-seminomatous germ cell tumours. Ann Oncol 6: 483-488, 1995

18. Mandelbaum I, Yaw PB, Einhorn LH, Williams SD, Rowland RG, Donohue JP. The importance of one-stage median sternotomy and retroperitoneal node dissection in disseminated testicular cancer. Ann Thor Surg 36: 524-528, 1983
19. Tiffany P, Morse MJ, Bosl G, et al. Sequential excision of residual thoracic and retroperitoneal masses after chemotherapy for stage III germ cell tumors. Cancer 57: 978-983, 1986
20. De Graaf WE, Oosterhuis JW, Van der Linden S, Homan van der Heide JN, Schraffordt Koops H, Sleijfer DTh. Residual mature teratoma after chemotherapy for nonseminomatous germ cell tumors of the testis occurs significantly less often in lung than in retroperitoneal lymph node metastases. J Urogen Pathol 1: 75-81, 1991
21. Steyerberg EW, Keizer HJ, Fosså SD, et al. Prediction of residual retroperitoneal mass histology following chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from 6 study groups. J Clin Oncol 13: 1177-1187, 1995
22. Little RJA. Regression with missing X's: a review. J Am Stat Assoc 87: 1227-1237, 1992
23. Steyerberg EW, Kievit J, De Mol Van Otterloo JCA, Van Bockel JH, Eijkemans MJC, Habbema JDF. Perioperative mortality of elective abdominal aortic aneurysm surgery: a clinical prediction rule based on literature and individual patient data. Arch Int Med, 155: 1998-2004, 1995
24. Hosmer DW, Lemeshow S. Applied logistic regression. New York, NY: John Wiley & Sons Inc, 1989, pp 140-145
25. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 247: 2543-2546, 1982
26. SAS Institute Inc., SAS® Technical Report P-200, SAS/STAT® Software: CALIS and LOGISTIC Procedures, Release 6.04. Cary, NC: SAS Institute Inc., USA, 1990. pp. 194-195
27. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 78: 316-331, 1983
28. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 9: 1303-1325, 1990
29. van de Wal HJ, Verhagen A, Lecluyse A, von Lier HJ, Jongerius CM, Lacquet LK. Surgery of pulmonary metastases. Thorac Cardiovasc Surg 34 Spec No 2: 153-156, 1986
30. Shimizu J, Oda M, Hayashi Y, et al. Results of surgical treatment of pulmonary metastases. J Surg Oncol 58: 57-62, 1995
31. Mead GM, Stenning SP, Parkinson MC, et al. The second medical research council study of prognostic factors in nonseminomatous germ cell tumors. J Clin Oncol 10: 85-94, 1992
32. Geller NL, Bosl GJ, Chan EYW. Prognostic factors for relapse after complete response in patients with metastatic germ cell tumors. Cancer 63: 440-445, 1989
33. Fox EP, Weathers TD, Williams SD, et al. Outcome analysis for patients with persistent nonteratomous germ cell tumor in postchemotherapy retroperitoneal lymph node dissections. J Clin Oncol 11: 1294-1299, 1993
34. Harding MJ, Brown IL, Macpherson SG, et al. Excision of residual masses after platinum based chemotherapy for non-seminomatous germ cell tumours. Eur J Cancer Clin Oncol 25: 1689-1694, 1989
35. Logothetis CJ, Samuels ML, Trindade A, et al. The growing teratoma syndrome. Cancer 50: 1629-1635, 1982
36. Ulbright TM, Loehrer PJ, Roth LM, et al. The development of non-germ malignancies within germ cell tumors: a clinicopathologic study of 11 cases. Cancer 54: 1824-1833, 1984
37. Ahlgren AD, Simrell CR, Triche TJ, et al. Sarcoma arising in a residual testicular teratoma after cytoreductive chemotherapy. Cancer 54: 2015-2018, 1984

38. Ahmed T, Bosl GJ, Hajdu SI. Teratoma with malignant transformation in germ cell tumors in men. Cancer 56: 860-863, 1985

39. Molenaar WM, Oosterhuis JW, Meiring A, et al. Histology and DNA contents of a secondary malignancy arising in a mature residual lesion six years after chemotherapy for a disseminated nonseminomatous testicular tumor. Cancer 58: 264-268, 1986

40. Dimipoulos MA, Amato RJ, Logothetis CJ. Predictive factors for effective salvage therapy of nonseminomatous germ cell tumors of testis. Urol 38: 351-354, 1991

41. Gerl A, Clemm C, Schmeller N, Hartenstein R, Lamerz R, Wilmanns W. Prognosis after salvage treatment for unselected male patients with germ cell tumours. Br J Cancer 72: 1026-1032, 1995

9 Perioperative mortality of elective abdominal aortic aneurysm surgery: a clinical prediction rule based on literature and individual patient data

E.W. Steyerberg, J. Kievit, J.C.A. de Mol van Otterloo, J.H. van Bockel, M.J.C. Eijkemans, J.D.F. Habbema.
Arch Int Med 1995; 155: 1998-2004

Abstract

Background: Abdominal aortic aneurysm surgery is a major vascular procedure with a considerable risk of (mainly cardiac) mortality.
Objective: To estimate elective perioperative mortality, we aimed to develop a clinical prediction rule based on several well established risk factors: age, gender, a history of myocardial infarction (MI), congestive heart failure (CHF), ischemia on the electrocardiogram (ECG), pulmonary impairment and renal impairment.
Methods: Two sources of data were used: 1) Individual patient data from 246 patients operated at the University Hospital Leiden and 2) Studies published in the literature between 1980 and 1994. The Leiden data were analyzed with univariate and multivariate logistic regression. Literature data were pooled with meta-analysis techniques. The clinical prediction rule was based on the pooled Odds Ratios from the literature, which were adapted by the regression results of the Leiden data.
Results: The strongest adverse risk factors in the literature were CHF and cardiac ischemia on ECG, followed by renal impairment, history of MI, pulmonary impairment and female gender. The literature data further showed that a 10-year increase in age more than doubled surgical risk. In the Leiden data, most multivariate effects were smaller than the univariate effects, which is explained by the positive correlation between the risk factors. In the clinical prediction rule, cardiac, renal and pulmonary comorbidity are the most important risk factors, while age per se has a moderate effect on mortality. For practical application, a center specific average surgical risk can be taken into account.
Conclusions: A readily applicable clinical prediction rule can be based on the combination of literature data and individual patient data. The risk estimates may be useful for clinical decision making in individual patients.

Abbreviations: CHF, congestive heart failure; ECG, electrocardiogram; MI, myocardial infarction, OR, Odds Ratio; LR, likelihood ratio; CI, confidence interval

9.1 Introduction

Elective surgery is the preferred therapy for abdominal aortic aneurysms, as until now reconstructive vascular surgery is the only effective prevention of rupture. Generally, the cumulative mortality risk of rupture is much higher than the risk of elective surgery[1,2]. A conservative policy may however be considered in case of small aneurysms, where the risk of rupture is low[3,4,5], in older patients who suffer from cardiac, pulmonary or renal comorbidity, and in patients who have a short life-expectancy because of a malignancy[6,7]. In these categories of patients the mortality risk of elective surgery may exceed the cumulative mortality risk of aneurysm rupture. Therefore, reliable estimates of surgical mortality are crucial for clinical decision making.

In the literature numerous studies can be found that relate surgical mortality to a single patient characteristic[1,8,9,10,11,12,13,14,15]. For example, by comparing the surgical mortalities in patients with and without congestive heart failure (CHF), the univariate effect of CHF on mortality can be determined. Such literature data may be pooled using meta-analysis techniques, resulting in more precise estimates of the univariate effect of risk factors. However, as risk factors tend to occur in association, the combination of univariate risk factor effects generally leads to a falsely high mortality estimate.

Surgical risk estimates based on combinations of characteristics are rare[16]. One well-known multifactorial risk estimator is the Goldman index[17,18]. However, this index was not developed specifically for aneurysm surgery and predicts cardiac mortality rather than surgical mortality as a whole. To relate the mortality of aortic aneurysm surgery to combinations of characteristics, one possibility is to analyze individual patient data with multivariate statistical techniques. Traditionally, such an analysis ignores data from the literature, because univariate data from the literature cannot be combined directly into a multivariate analysis. This approach has the disadvantage that small numbers lead to wide confidence intervals of risk factors and unreliable risk estimates. Moreover, variable selection may be difficult[19,20], e.g. for infrequent patient characteristics.

In this article, we introduce a method to combine results from literature data and individual patient data. The risk estimates can thus be based on a larger number of patients than available from one single institution. We present the results of the analysis as a clinical prediction rule that estimates the mortality of elective aneurysm surgery in the individual patient.

9.2 Material and Methods

9.2.1 Leiden data

We collected individual patient data on 246 consecutive patients, who underwent primary elective surgery for abdominal aortic aneurysm at the University Hospital Leiden between 1977 and 1988[21]. Preoperative examination consistently comprised history, physical examination and standard resting ECG. Standard registration forms were used, and all data collection was supervised by one of the authors (J.C.A.M.O.).

All patients were examined prior to operation by the same physician (JDM Feuth, MD), who also reviewed all ECGs (blinded to surgical outcome).

Surgical or perioperative mortality was defined as in-hospital mortality, independent of duration of hospital stay, or death within 30 days after surgery, when the patient was discharged earlier. Of the 246 patients who underwent elective surgery, 18 died (7.3%). Nine (50%) of these patients died within 14 days, six (33%) between 14 and 30 days, and 3 (17%) died in hospital 34, 39 and 95 days after surgery. Most patients (11/18=61%) died of cardiac causes.

9.2.2 Literature data

Published studies were selected using MEDLINE medical database, from 1980 up to July 1994, and via cross-references between articles. The selection was limited to English language studies, which had to contain frequency data on the association of a potential risk factor and surgical mortality, either in tables or mentioned in the text (see Appendix 1). Further on, analysis of the literature data refers to the analysis of these published studies combined with the Leiden data.

9.2.3 Definitions of risk factors

Risk factors considered were age, gender, and cardiac, renal or pulmonary comorbidity. Cardiac comorbidity consisted of three factors. 'History of MI' was defined as a documented history of a myocardial infarction, regardless of findings on the present preoperative ECG. Congestive heart failure ('CHF') was defined as cardiogenic pulmonary edema and/or jugular vein distension, or presence of a gallop rhythm regardless of treatment. 'ECG: Ischemia' was present if ST-depression was over 2 mm on the standard resting ECG. Renal function used the cut-off value of 160 μmol/l or 1.8 mg/dl for the preoperative creatinine level. Pulmonary comorbidity was present if patients suffered from COPD, emphysema or dyspnoea, or had undergone previous pulmonary surgery. Definitions of renal and pulmonary impairment varied to some extent in the literature (see Appendix 1).

9.2.4 Definitions of effect measures

The Odds Ratio (OR) and the Likelihood Ratio (LR) were used as effect measures of the risk factors for surgical mortality. The OR indicates to what extent the risk in patients with a risk factor is higher than in those without. The OR is calculated as the ratio of the mortality odds in categories of patients with and without a risk factor being present. Likelihood Ratios (LRs) indicate to what extent the average or prior surgical risk has to be corrected to a higher or a lower probability in the presence or absence of a risk factor. Likelihood ratios can be calculated for the presence (LR+) and absence (LR-) of dichotomous risk factors. The LR+ is the probability of the presence of a risk factor in the patients who died divided by the probability of the presence of that risk factor in the patients who survived: LR+ = p(+|died) / p(+|survived). The LR- is defined analogously, with absence of the risk factor instead of presence. The relation between the LR+, LR- and the OR is straightforward: OR=LR+/LR-.

9.2.5 *Statistical analysis*

In the Leiden data both univariate and multivariate ORs were determined using logistic regression analysis[22]. All risk factors that were significant in the meta-analysis were included in the multivariate model, as emphasis is on the combined effect of the risk factors and not on statistical significance of individual factors[19]. A graphical impression of the goodness-of-fit of the multivariate model was obtained by plotting the observed versus expected cumulative number of deaths[23]. Goodness-of-fit was tested by the Hosmer-Lemeshow test[22] (BMDP module LR[24]), which evaluates the correspondence between a model's predicted probabilities and the observed frequencies over groups spanning the entire range of probabilities.

Literature data were summarized using techniques of meta-analysis. For the dichotomous (+/-) characteristics gender, history of MI, CHF, ECG: Ischemia, pulmonary impairment and renal impairment, ORs were calculated within each study (study OR) and pooled subsequently (pooled OR) using exact methods[25,26]. Factors have statistically significant effects ($p < 0.05$) if the 95%-CI of the pooled OR does not include 1. An exact test for homogeneity was used to test if one pooled OR across studies might be assumed[27]. Pooled Likelihood Ratios were calculated, using the method described by Simel[28].

The univariate effect of age as a continuous variable could not be estimated directly from the literature, as age is typically reported in categories (e.g. <70 vs >70, or <60, 60-80, >80 years). For quantitative analysis, numeric values have to be assigned to the categories, e.g. the mean of the category[29]. Mean age was estimated using study-specific descriptions of the age distribution (mean and standard deviation, or mean only, using standard deviations of the Leiden data). Logistic regression analysis was used to estimate the effect of age on mortality.

9.2.6 *Clinical prediction rule*

A clinical prediction rule was developed for the estimation of elective surgical risk in individual patients. The development of the prediction rule involved four steps, as shown in Table 1. In step 1a, the univariate and multivariate logistic regression coefficients ($\ln(OR_{Uni})$ and $\ln(OR_{Mult})$) are calculated from the individual patient data, in this case the 'Leiden data'. Step 1b involved the calculation of an adaptation factor for each risk factor, i.e. the difference between the univariate and multivariate logistic regression coefficients[29]. In the second step, pooled ORs and pooled LRs were calculated from the literature data (containing both published studies and individual patient data). In the third step, the results of the univariate literature data analysis and multivariate individual patient data analysis were combined. The natural logarithms of the pooled ORs ($\ln(OR_{Pooled})$) were adapted with the adaptation factor for each risk factor. Finally, adapted ln(LR)s or 'adapted weights' were calculated by using the ratio of the pooled ln(LR+) and the pooled ln(LR-). These adapted weights allow the estimation of the risk of a base-line case (a patient without comorbidity) in the clinical prediction rule. 95% confidence intervals (95%-CI) were calculated for the adapted ORs. The calculation is described in detail in Appendix 2.

Table 1 Derivation of the clinical prediction rule.*

Step	Data	Technique	Calculation
1a+1b	Individual patient	Logistic regression	Adaptation factor = $\ln(OR_{Mult}) - \ln(OR_{Uni})$
2	Literature	Meta-analysis	OR_{Pooled}, $LR+_{Pooled}$, $LR-_{Pooled}$
3	Lit+Individual pt	Adaptation of OR	$\ln(OR_{Adapted}) = \ln(OR_{Pooled}) + \text{Adaptation factor}$
4	Literature	Ratio of LR+ and LR-	$\ln(LR+_{Adapted}) = \ln(OR_{Adapted}) / [1-(\ln(LR-_{Pooled})/\ln(LR+_{Pooled}))]$ $\ln(LR-_{Adapted}) = \ln(OR_{Adapted}) / [(\ln(LR+_{Pooled})/\ln(LR-_{Pooled}))-1]$

* Literature and individual patient data are combined in step 3, under the assumption that the change in logistic regression coefficient in the individual patient data may be applied to the literature data.

The development of the clinical prediction rule involved the following assumptions. First, it is assumed that the logistic regression model with multiplicative effects adequately models the surgical mortality as a function of the risk factors. Reliability or goodness-of-fit of the multivariate model on the own data was therefore assessed[22,23]. The meta-analysis assumed fixed effects of the risk factors across studies. This assumption was assessed by tests of homogeneity. In the third step, the meta-analysis results were adapted, based on the difference in uni- and multivariate results of the individual patient data. This assumed that the correlation between factors in the individual patient data was similar to the correlation between factors in the studies. We therefore compared the correlation between the risk factors in the individual patient data (Table 3) to the correlation (as far as reported) in the literature. We also assessed whether the correlation between univariate and multivariate regression coefficients was so strong that the adaptation method was expected to result in better estimates of the multivariate regression coefficients (Appendix 2). Finally, adapted weights were calculated, assuming that the ratio of the weights in the univariate meta-analysis are equal to the ratio of the weights in the multivariate analysis. This final assumption also underlies other methods that estimate multivariate weights for the presence and absence of risk factors[19].

9.2.7 Prognostic score chart

We present the resulting clinical prediction rule as a prognostic score chart with rounded values of $10 \cdot \ln(OR_{Adapted})$ as scores. Before an individual mortality risk can be quantified, an average hospital mortality has to be estimated[18]. We assume that surgical mortality is around 5% nowadays for a patient population with an average prevalence of risk factors. If a higher or lower average surgical mortality is observed at a particular institution, this may be explained by a different prevalence of risk factors, which is related to patient selection. If this explanation is insufficient, a different center specific average surgical risk has to be considered in the clinical prediction rule, which is determined by factors not considered here (e.g. definition of surgical mortality, hospital volume, experience of surgeon[30,31]).

9.3 Results

9.3.1 Leiden data

The results of the analysis of our individual patient data are shown in Table 2. Of 246 elective patients, 238 had no missing value on any of the risk factors. Uni- and multivariate analyses were based on these 238 patients. If the multivariate ORs are smaller than the univariate ORs, the adaptation factors are negative. This is the case for all risk factors except gender. A large, negative, adaptation factor was found for each of the factors related to cardiac comorbidity (history of MI, CHF and ECG: ischemia), which is explained by the positive correlation between these risk factors (Table 3). Also, the OR of age was reduced considerably, because age was significantly associated with the presence of cardiac comorbidity (Table 3). The model as a whole was significant (model χ^2 25.3, df 7, $p < 0.001$). Goodness-of-fit was adequate (Figure 1, Hosmer-Lemeshow test: p=.79), indicating that the multivariate logistic regression model reliably fitted the Leiden data.

Table 2 Results of the univariate and multivariate analysis of the Leiden data.*

Prognostic factor	OR_{Uni} [95%-CI]	OR_{Mult} [95%-CI]	Adaptation factor [95%-CI]
Age (per decade)	2.67 [1.3-5.7]	1.79 [.83-3.9]	-0.40 [-.80 - -.01]
Female gender	1.32 [.28-6.2]	1.34 [.25-7.3]	+0.02 [-.90 - +.94]
History of MI	4.48 [1.7-12]	2.07 [.69-6.4]	-0.77 [-1.4 - -.13]
CHF	5.94 [2.0-17]	2.83 [.89-9.0]	-0.74 [-1.3 - -.17]
ECG: Ischemia	5.57 [1.9-16]	2.73 [.80-9.1]	-0.71 [-1.5 - +.04]
Impaired renal function	3.47 [.88-14]	3.07 [.68-14]	-0.12 [-.98 - +.74]
Impaired pulmonary function	2.32 [.82-6.6]	1.83 [.58-5.8]	-0.23 [-.84 - +.38]

* Odds Ratios with 95%-confidence intervals were calculated by logistic regression analysis. The adaptation factor was calculated as the difference in multivariate and univariate regression coefficient: $\ln(OR_{Mult}) - \ln(OR_{Uni})$. The 95%-confidence intervals for the adaptation factor were based on 500 bootstrap samples.

Table 3 Correlation between risk factors in the Leiden data (N=238).

	Age	Female gender	MI	CHF	ECG: Isch.	Renal	Pulm. function
Age	1.00						
Female gender	.05	1.00					
History of MI	.17**	-.04	1.00				
CHF	.20**	-.00	.26**	1.00			
ECG: Ischemia	.15*	-.01	.45**	.32**	1.00		
Renal function	.01	-.02	.05	.14*	.03	1.00	
Pulm. function	.10	.00	.08	.13*	.03	.01	1.00

2-tailed Significance: * - .01 ** - .001

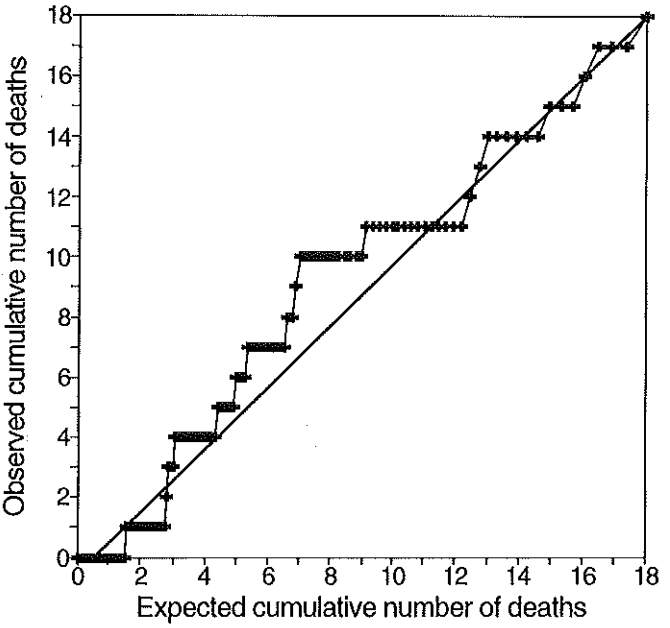

Figure 1 Reliability curve of the multivariate logistic regression analysis of the Leiden data. 18 patients died after surgery. The observed curve closely follows the ideal straight line.

9.3.2 Literature data

The studies forming the literature data are shown in Table 4. Fifteen published studies and one thesis were identified, which presented data on the relationship between at least one of the risk factors and surgical mortality[1,8,9,10,11,12,13,14,15,30,31,32,33,34,35]. Details are shown in Appendix 1. From the largest study included (N=8185[31]), only the effects of age and gender were analyzed. The definitions of other variables (e.g. renal failure) in this study were considered incompatible with the definitions used in the meta-analysis. This incompatibility was confirmed by tests for homogeneity. Overall, mortality was 6.8% (1171/16067) in the studies included in the meta-analysis.

The effect of age was estimated from fourteen studies (Table 4), in which patients had a mean age of 69.2 years. A ten-year increase in age more than doubled surgical risk (OR 2.20 per decade), with a narrow confidence interval ([1.9-2.5]) because of the large number of patients analyzed (N=13336). The effects of the other characteristics are shown in Table 5. A smaller number of studies was available (Appendix 1)[1,10,11,12,30,31,33,34,35], in which the average prevalence of risk factors was between 15% and 25%. The strongest adverse risk factors in the literature were CHF and ischemic changes on the preoperative ECG, followed by an impaired renal function,

history of MI, impaired pulmonary function and female gender. We found no indications for heterogeneity of effects (all tests for homogeneity: p>.10).

Table 4 Studies included in the meta-analysis of preoperative characteristics and elective surgical mortality.

First author	Reference	Study period	N	Mortality	OR_{age}*
Lundell	8	1971-80	35	5 (14%)	1.2
Noppeney	9	1984-86	100	2 (2.0%)	1.3
Morishita	10	1968-90	110	3 (2.7%)	0.74
Chang	32	1973-77	120	11 (9.2%)	2.5
Fielding	1	1960-79	222	18 (8.1%)	1.8
De Mol Van Otterloo	21	1977-88	246	18 (7.3%)	2.7
AbuRahma	33	1983-87	332	7 (2.1%)	-
Diehl	11	1974-78	350	18 (7.3%)	4.4
Bosman	34	1958-78	360	30 (8.3%)	1.5
McCabe	12	1972-77	364	9 (2.5%)	5.5
D'Angelo	13	1966-90	590	27 (4.6%)	¶0.87
Johnston	35	1986-86	666	32 (4.8%)	2.2
Paty	14	1978-91	699	16 (2.3%)	1.1
Akkersdijk	15	1990-90	1289	88 (6.8%)	2.8
Hannan	30	1985-87	3570	273 (7.7%)	2.2#
Katz	31	1980-90	8185	614 (7.5%)	2.3
Total			16067	1171 (6.8%)	2.2

* Odds Ratio of age (per decade); ¶ Analysis with stratification in three reported time-periods;
# Reported logistic regression result; not used in meta-analysis of the effect of age

Table 5 Meta-analysis of literature data, with pooled Odds Ratios and pooled Likelihood Ratios.

Prognostic factor	OR_{Pooled} [95%-CI]	$LR+_{Pooled}$	$LR-_{Pooled}$	Average*
Age (per decade)	2.20 [1.9 - 2.5]	-	-	6.92±.63
Female gender	1.44 [1.2 - 1.7]	1.34	0.96	19%
History of MI	2.80 [1.6 - 4.7]	2.12	0.76	22%
CHF	4.89 [2.4 - 9.8]	2.74	0.70	15%
ECG: Ischemia	4.58 [2.4 - 8.8]	2.55	0.62	20%
Impaired renal function	3.75 [2.3 - 6.1]	2.35	0.85	17%
Impaired pulmonary function	2.43 [1.6 - 3.8]	1.76	0.79	25%

* Averages of the risk factors in the studied population (mean ± standard deviation or prevalence).

9.3.3 Clinical prediction rule

Table 6 shows the adapted ORs and LRs for the clinical prediction rule. Remarkably, the adapted OR of a history of MI appeared to be close to one ($OR_{Adapted}$=1.31), indicating that the additional risk due to a history of a myocardial infarction is small when the other risk factors are taken into account. The small adapted OR is explained

by the fact that a large reduction in the OR was observed in the Leiden data (Table 2: OR_{Uni}=4.48, OR_{Mult}=2.07), and the fact that the pooled OR in the meta-analysis was not very large (Table 5: OR_{Pooled}=2.80). For the same reason, the adapted multivariate effect of age is relatively small ($OR_{Adapted}$=1.47 per decade).

Table 6 Results of the derivation of the clinical prediction rule.*

Prognostic factors	$OR_{Adapted}$ [95%-CI]	$LR+_{Adapted}$	$LR-_{Adapted}$
Age (per decade)	1.47 [1.0 - 2.1]	-	-
Female gender	1.47 [.53 - 4.0]	1.40	0.95
History of MI	1.31 [.63 - 2.7]	1.22	0.93
Congestive heart failure (CHF)	2.33 [1.0 - 5.2]	1.86	0.80
ECG: Ischemia	2.22 [.95 - 5.2]	1.69	0.76
Impaired renal function	3.32 [1.4 - 7.7]	2.73	0.82
Impaired pulmonary function	1.93 [.91 - 4.1]	1.59	0.82

* The adapted Odds Ratios are calculated from the results of the meta-analysis (Table 5) combined with the adaptation factor, as calculated from the Leiden data (Table 2). Adapted Likelihood Ratios are calculated using the ratio of the (natural logarithm of the) pooled LRs (Table 5).

9.3.4 Prognostic score chart

The clinical prediction rule is presented as a score chart (Table 7), based on the adapted ORs. First, a center specific average surgical mortality has to be estimated. A score of 0 points corresponds to an average risk of 5%. Next, we defined a base-line case as a 70-year-old male patient without comorbidity. The score of this patient is set at 0, which corresponds to a risk of 1.9%. Older patients attain a higher score (1 point per 2.5 year), as well as female patients and patients with cardiac, pulmonary or renal comorbidity. Patients with cardiac comorbidity may have any combination of MI, CHF and ECG: Ischemia. The scores add up to a sumscore and the corresponding individual surgical risk can be read from the final part of Table 7.

The use of the score chart is illustrated with a 80-year-old male patient, who has congestive heart failure and renal impairment, but no other comorbidity. The score of this patient is +4 for age, +8 for CHF, and +12 for renal impairment, which adds to 24 points. If this patient is operated in a center with an average surgical mortality of 5%, the sumscore is 24 points and corresponds to a risk near 19%. If this patient is operated in a center that attains an average surgical mortality of 5% in patients 72 years of age (±2.5 years older than the mean of 69.2 years found in this analysis), the center specific surgical risk is 1 point lower. This 80-year-old patient then has a score of +23 points, or a risk of ±16%.

Table 7 Score chart for mortality of elective abdominal aortic aneurysm surgery.*

1. Center specific average surgical mortality							
%	3%	4%	5%	6%	8%	12%	
Score	-5	-2	0	+2	+5	+10	.....
2. Individual prognostic factors							
Age	60	70	80				
Score	-4	0	+4				.....
Gender	female						
Score	+4						.....
Cardiac comorbidity	MI	CHF	ECG: ischemia				
Score	+3	+8	+8				.....
Renal comorbidity	impairment						
Score	+12						.....
Pulm. comorbidity	impairment						
Score	+7						.....
3. Estimated individual surgical mortality						**Sumscore (add)****	.....

Sumscore	-5	0	5	10	15	20	25	30	35	40
Probability	1%	2%	3%	5%	8%	12%	19%	28%	39%	51%

* After estimation of the average surgical mortality (**1.**), individual patient characteristics are taken into account (**2.**). The resulting sumscore is translated into a probability (**3.**).

** The formula to calculate the mortality risk is: 1/[1+exp(-((sumscore/10)-3.95))]

9.4 Discussion

In this paper we present a clinical prediction rule to estimate perioperative mortality of elective abdominal aortic aneurysm surgery. Risk factors were identified both from individual patient data collected in one institution (University Hospital Leiden) and from the literature. Risk factors comprised demographic data (age, gender) and comorbidity (cardiac, pulmonary or renal). We used a new statistical methodology to quantify the combined effect of these risk factors on surgical mortality.

A number of problems arose in the analysis of the literature data. First, the relation of risk factors to surgical mortality was described in a number of papers, but definitions were not consistent (see Appendix 1). For example, pulmonary function was defined by different criteria, and an impaired renal function was defined at different cut-off values of the creatinine level. Despite these differences in definition, one single effect could be assumed for each risk factor across the studies (tests for homogeneity). The analysis of the continuous variable age was based on a large number of patients (N=13336), but the analysis was hampered by the fact that mortalities were described in relatively large age-

intervals, e.g. younger or older than 70 years. For logistic regression analysis, we estimated the average ages in these age-intervals using study-specific descriptions as far as available. The effect of age would have been estimated more accurately if smaller age-intervals had been reported and more study characteristics had been described in the publications.

Ideally, risk estimates for different centers are obtained by multivariate analysis of the original individual patient data from published series. In the absence of these data, only the published figures can be used for a univariate meta-analysis. To adapt for the associations between risk factors, we used an adaptation factor, which was obtained from the Leiden data. This method has several advantages over using the Leiden data only. First, the literature provides a much larger number of patients than available from one single institution, making the estimates more precise and variable selection more straightforward. Moreover, the estimates represent the experience of several centers and hence are more generally applicable.

Disadvantages of the method are that publication bias may be present when assessing the literature, i.e. the phenomenon that the relation of a risk factor with mortality has only been reported when found significant. Publication bias leads to higher effect estimates in smaller studies. Examination of Table 4 or Appendix 1 shows no clear relation between study size and effect size, suggesting that no strong publication bias is present in this meta-analysis. Next, the correlation of risk factors in the Leiden data may be different from the correlation in the literature. For example, in one series[31] female patients were 3 years older age than male patients, while the age difference was only 2 months in our patient series (correlation +.05). Such differences in correlation between risk factors lead to inaccurate adaptations of the meta-analysis by the Leiden data.

The resulting clinical prediction rule quantifies the prognostic impact of age, gender, and cardiac, renal, and pulmonary comorbidity. The most important risk factors are renal function, CHF and ischemic changes on the ECG. Age has a limited effect on mortality, when corrected for cardiac, renal and pulmonary comorbidity (OR 1.47 per decade).

Average risk of elective surgery was found to be 6.8% (1171/16067) in the studies included in the meta-analysis. This figure is dominated by three large studies[15,30,31], which are all population-based. On the one hand, a clear reduction of elective surgical mortality has been shown during the past decade[13,30,31,36], leading to a lower average mortality estimate nowadays, e.g. 5%. On the other hand, considerable variation is found in elective mortality between studies and centers. For example, an enquiry in district hospitals in the U.K. revealed a mortality rate up to 16%[37]. Variation in reported surgical mortality may be explained by several factors[38]. Firstly, patient selection influences reported figures, as some high risk patients will be operated on in some centers but not in other centers. Patient selection may however not be sufficient to explain differences in average mortality between centers. Another factor is that most centers define surgical mortality as mortality within 30 days, while others also included

in-hospital deaths after 30 days[1], which is preferred in our view. Finally, hospital volume and experience of the surgeon have been found as important risk factors[30,31].

The large variation in average surgical risk has to be taken into account when making predictions for individual patients[18]. We assumed a center specific risk of 5% for a population of patients with the characteristics as found in the meta-analysis. Higher or lower average risks, if not explained by patient selection, can easily be accounted for in the prognostic score chart.

Our score chart enables the estimation of surgical risk for individual patients. This risk estimate has to be viewed with some caution. First, the limited number of patients, especially in the Leiden data, causes uncertainty in the estimates. Prospective validation of the prediction rule is therefore necessary. Further, statistical modeling can never completely substitute for definitive clinical judgment, which may involve specific patient characteristics that were not considered in the model[18]. However, the score chart is helpful in identifying high risk patients, who may be candidates for further diagnostic work up, like coronary angiography, and subsequent therapy like CABG[39]. The risk estimates can also be applied in decision making in situations where the risk of elective surgery may exceed that of follow-up, e.g. in small aneurysms. According to one model[3], early surgery is preferred in male patients with aneurysms 40–49 mm in size until the age of 93, assuming an age-independent surgical mortality of 4.6%. According to our score chart, surgical risk may vary between 4.4% and 75% for a 93-year-old patient, assuming a center specific mortality of 5% and depending on the presence of comorbidity. Obviously, this individualized surgical risk estimate may direct optimal treatment of a particular 93-year-old patient (surgery or follow-up). Note that age-thresholds for early surgery depend on the presence of comorbidity, because of a higher perioperative mortality, but also because of a lower life-expectancy[2].

In conclusion, our clinical prediction rule estimates surgical mortality based on the combination of literature data and data collected in one institution using a multivariate statistical model. This method may also be useful in other clinical fields where prediction is at issue and patient series are published in the literature which are comparable to the own series. The resulting score chart is expected to be easy to use in daily clinical practice. Firstly, because the risk factors are all standard diagnostic examinations or readily available patient characteristics. Secondly, because the individual patient scores can simply be added in a score chart and the corresponding risk estimate can be read from a table, taking into account center specific surgical risk.

We would like to thank J.D.M. Feuth, Department of Surgery, University Hospital Leiden, The Netherlands, for assistance with data collection and Houke M. Klomp, Department of Surgery, University Hospital Rotterdam, for helpful comments

Appendix 1

Surgical mortality in relation to the preoperative characteristics gender, renal function, pulmonary function, history of MI, CHF and ECG: Ischemia. Published studies and Leiden data (De Mol Van Otterloo) are shown, ordered according to study size.

Gender:	**Women**		**Men**		
First author	Dead	N	Dead	N	OR
De Mol Van Otterloo	9.5%	21	7.1%	225	1.4
AbuRahma	5.3%	76	1.2%	256	4.5
Bosman	2.4%	41	9.1%	320	.25
McCabe	3.7%	54	2.3%	310	1.7
Hannan	8.5%	778	7.4%	2792	1.2
Katz	10.6%	1469	6.8%	6716	1.6

Renal function:	**Impaired**		**Unimpaired**		
First author	Dead	N	Dead	N	OR
Morishita #	2.4%	42	2.9%	68	0.8
De Mol Van Otterloo **	20%	15	6.7%	223	3.5
Bosman **	25%	8	6.6%	319	4.7
McCabe *	6.7%	45	1.8%	319	3.7
Diehl ##	19%	31	4.6%	521	5.0
Johnston ***	9.9%	223	2.9%	350	3.7

Creat > 2.0 mg/ml and clearance < 50 ml/min, or a PSP (15 min) < 25%; ## Creat >2.0mg/dl; * Creat > 1.8 mg/dl or BUN > 40 mg/dl; ** Creat ≥ 1.8 mg/dl (≥ 160 μmol/l); *** Creat > 1.25 mg/dl

Pulmonary function:	**Impaired**		**Unimpaired**		
First author	Dead	N	Dead	N	OR
Morishita #	8.3%	36	0%	74	∞
De Mol Van Otterloo **	13%	47	6.0%	199	2.3
Diehl *	6.5%	77	2.7%	222	2.5
Bosman **	11%	71	7.0%	287	1.7
Johnston ***	8.2%	184	3.2%	475	2.7

FEV1 < 70% or VC < 80%; * FEV1 < 60%; ** COPD or emphysema or previous pulmonary surgery (lobectomy/pneumectomy) / dyspnoea; *** COPD or abnormal PO_2, PCO_2 or FEV1

History of MI:	**MI**		**No MI**		
First author	Dead	N	Dead	N	OR
Fielding	15%	26	7.1%	196	2.4
De Mol Van Otterloo	17%	58	4.3%	188	4.7
Johnston	8.1%	160	3.8%	506	2.3

Congestive Heart Failure: First author	**CHF** Dead	N	**No CHF** Dead	N	OR
De Mol Van Otterloo	16%	81	3.0%	165	6.1
Johnston	15%	54	3.9%	612	4.3

Electrocardiogram: First author	**Ischemia** Dead	N	**No Ischemia** Dead	N	OR
De Mol Van Otterloo	15%	85	3.1%	156	5.6
Johnston	13%	92	3.5%	574	4.2

Appendix 2

The adaption method calculates the adapted multivariate regression coefficients as

$$\beta_{MULT|LIT} = \beta_{UNI|LIT} + (\beta_{MULT|IND} - \beta_{UNI|IND}) \quad (1).$$

The variance of the adapted multivariate regression coefficients can be formulated as

$$\begin{aligned} var(\beta_{MULT|LIT}) = {} & var(\beta_{UNI|LIT}) + var(\beta_{MULT|IND}) + var(\beta_{UNI|IND}) \\ & +2 \cdot covariance(\beta_{UNI|LIT}, \beta_{MULT|IND}) - 2 \cdot covariance(\beta_{UNI|LIT}, \beta_{UNI|IND}) \\ & -2 \cdot covariance(\beta_{MULT|IND}, \beta_{UNI|IND}) \end{aligned} \quad (2).$$

Since the individual patient data set is a random sample from the literature, it may be assumed that a positive correlation exists between $\beta_{UNI|IND}$ and $\beta_{UNI|LIT}$. Moreover, it may be assumed that

$$covariance(\beta_{MULT|IND}, \beta_{UNI|LIT}) < covariance(\beta_{UNI|IND}, \beta_{UNI|LIT}) \quad (3).$$

Leaving out both covariances leads to a slight overestimation of the variance of the $var(\beta_{MULT|LIT})$, since the larger term has a minus sign in (2).

The $covariance(\beta_{MULT|IND}, \beta_{UNI|LIT})$ can be written as
$\rho(\beta_{MULT|IND}, \beta_{UNI|IND}) \cdot (SE(\beta_{UNI|IND}) \cdot SE(\beta_{MULT|IND}))$.

This leads to

$$\begin{aligned} var(\beta_{MULT|LIT}) = {} & var(\beta_{UNI|LIT}) + var(\beta_{MULT|IND}) + var(\beta_{UNI|IND}) \\ & -2 \cdot \rho(\beta_{MULT|IND}, \beta_{UNI|IND}) \cdot (SE(\beta_{UNI|IND}) \cdot SE(\beta_{MULT|IND})) \end{aligned} \quad (4).$$

Table A1 shows the variances of the coefficients involved, which are all expressed as standard errors (SE). The $\rho(\beta_{MI}, \beta_{UI})$ was estimated from 500 bootstrap samples of the Leiden data set with individual patient data and was large for all prognostic factors. The $SE(\beta_{MI|LIT})$ was calculated with formula (4). The $SE(\beta_{MI|LIT})$ is smaller than $SE(\beta_{MI|IND})$ for all prognostic factors.

Table A1 Standard errors and correlation of prognostic factors.

Prognostic factor	$SE(\beta_{U|LIT})$	$SE(\beta_{M|IND})$	$SE(\beta_{U|IND})$	$\rho(\beta_{M|I},\beta_{U|I})$	$SE(\beta_{M|LIT})$
Age (per decade)	.06	.39	.38	.91	.17
Female gender	.08	.86	.79	.81	.51
History of MI	.27	.57	.50	.88	.38
CHF	.33	.59	.55	.91	.41
ECG: Ischemia	.31	.62	.55	.87	.43
Impaired renal function	.25	.77	.70	.85	.43
Impaired pulmonary function	.23	.59	.53	.90	.38

References

1. Fielding JWL, Black J, Ashton F, Slaney G, Campbell DJ. Diagnosis and management of 528 abdominal aortic aneurysms. Br Med J, 1981, 283: 355-359
2. Crawford ES. Ruptured abdominal aortic aneurysm: an editorial. J Vasc Surg, 1991, 13: 348-350
3. Katz DA, Littenberg B, Cronenwett JL. Management of small abdominal aneurysms. Early surgery vs watchful waiting. JAMA, 1992, 268: 2678-2686
4. Michaels JA. The management of small abdominal aortic aneurysms: a computer simulation using Monte Carlo methods. Eur J Vasc Surg, 1992, 6: 551-557
5. Geroulakos G, Nicolaides A. Infrarenal abdominal aortic aneurysms less than five centimeters in diameter: the surgeon's dilemma. Eur J Vasc Surg, 1992, 6: 616-622
6. Morris DM, Colquitt J. Concomitant abdominal aortic aneurysm and malignant disease: a difficult management problem. J Surg Oncol, 1988, 39: 122-125
7. Lierz MF, Davis BE, Noble MJ, Wattenhofer SP, Thomas JH. Management of abdominal aortic aneurysm and invasive transitional cell carcinoma of bladder. J Urol, 1993, 149: 476-479
8. Lundell L, Norbäck B. Abdominal aortic aneurysm - results of treatment in nonspecialized units. Acta Chir Scand, 1983, 149: 695-702
9. Noppeney T, Raithel D. Age as a high-risk factor in the treatment of abdominal aortic aneurysm. Vascular Surg, 1990, 24: 271-276
10. Morishita Y, Toyohira H, Yuda T, Yamashita M, Shimokawa S, Saigenji H, Hashiguchi M, Kawashima S, Moriyama Y, Taira A. Surgical treatment of abdominal aortic aneurysm in the high-risk patient. Jap J Surg 21: 595-599, 1991
11. Diehl JT, Cali RF, Hertzer NR, Beven EG. Complications of abdominal aortic reconstruction. An analysis of perioperative risk factors in 557 patients. Ann Surg, 1983, 197: 49-56
12. McCabe CJ, Coleman WS, Brewster DC. The advantage of early operation for abdominal aortic aneurysm. Arch Surg, 1981, 116: 1025-1029
13. D'Angelo F, Vaghi M, Zorzoli C, Gatti S, Tacconi A. Is age an important risk factor for the outcome of elective abdominal aneurysm surgery? J Cardiovasc Surg, 34: 153-155, 1993
14. Paty PSK, Lloyd WE, Chang BB et al. Aortic replacement for abdominal aortic aneurysm in elderly patients. Am J Surg, 1993, 166: 191-193
15. Akkersdijk GJM, Van der Graaf Y, Van Bockel JH, De Vries AC, Eikelboom BC. Mortality rates associated with operative treatment of infrarenal abdominal aortic aneurysm in the Netherlands. Br J Surg, 1994, 81: 706-709

16. Johnston KW. Multicenter prospective study of nonruptured abdominal aortic aneurysm. Part II. Variables predicting morbidity and mortality. J Vasc Surg, 1989, 9: 437-447
17. Goldman L, Caldera DL, Nussbaum SR, et al. Multifactorial index of cardiac risk in noncardiac surgical procedures. N Engl J Med, 1977, 297: 845-850
18. Detsky AS, Abrams HB, Forbath N, et al. Cardiac assessment for patients undergoing noncardiac surgery: a multifactorial clinical risk index. Arch Intern Med, 1986, 146: 2131-2134
19. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med, 1986, 5: 421-433
20. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosatie RA. Regression modelling strategies for improved prognostic prediction. Stat Med, 1984, 3: 143-152
21. De Mol Van Otterloo JCA. Risk analysis of aortic reconstruction. Thesis Leiden University 1995
22. Hosmer DW, Lemeshow S. Applied logistic regression. John Wiley & Sons, NY, NY, USA, 1989
23. Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis: Trustworthiness of the exact values of the diagnostic probabilities. Meth Inf Med, 1978,17: 227-237
24. BMDP statistical software, Inc. Los Angeles, CA, USA, 1990
25. Cox DR. The analysis of binary data. Methuen, London, 1970
26. StatXact Statistical Software for Exact Nonparametric Inference. CYTEL Software Corporation, Cambridge, MA, USA, 1991
27. Zelen M. The analysis of several 2 x 2 contingency tables. Biometrika, 1971, 58: 129-137.
28. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J Clin Epidemiol, 44, 1991: 763-770
29. Greenland S. Quantitative methods in the review of epidemiologic literature. Epidemiologic Rev, 1987, 9: 1-30
30. Hannan EL, Kilburn H, O'Donnell, et al. A longitudinal analysis of the relationship between in-hospital mortality in New York state and the volume of abdominal aortic aneurysm surgeries performed. Health Serv Res, 1992, 27: 517-542.
31. Katz DJ, Stanley JC, Zelenock GB. Operative mortality rates for intact and ruptured abdominal aortic aneurysms in Michigan: an eleven-year statewide experience. J Vasc Surg, 1994, 19: 804-817
32. Chang FC, Smith JL, Rahbar A, Farha GJ. Abdominal aneurysms. A comparative analysis of surgical treatment of symptomatic and asymptomatic patients. Am J Surg, 1978, 136: 705-708
33. AbuRahma AF, Robinson PA, Boland JP, et al. Elective resection of 332 abdominal aortic aneurysms in a southern West Virginia community during a recent five-year period. Surgery, 1991, 109: 244-251
34. Bosman CHR. Report on 20 years of experience with atherosclerotic aneurysms of the abdominal aorta. Thesis Leiden University 1983
35. Johnston KW. Multicenter prospective study of nonruptured abdominal aortic aneurysm. I. Population and operative management. J Vasc Surg, 1988, 7: 69-81
36. Crawford ES, Saleh SA, Babb JW, Glaeser DH, Vaccaro PS, Silvers A. Infrarenal abdominal aortic aneurysms. Factors influencing survival after operations performed over a 25 year period. Ann Surg, 1981, 193: 699-709
37. Guy AJ, Lambert D, Jones NAG, Chamberlain J. After the confidential enquiry into perioperative deaths - aortic aneurysm surgery in the Northern Region. Br J Surg, 1990, 77: A344-A345 (abstract)
38. Campbell WB. Mortality statistics for elective aortic aneurysms. Eur J Vasc Surg, 1991, 5: 111-113
39. Hertzer NR, Beven EG, Young JR, et al. Coronary artery disease in peripheral vascular patients. A classification of 1000 coronary angiograms and results of surgical management. Ann Surg, 1984, 199: 223-233

10 Age thresholds for prophylactic replacement of Björk-Shiley convexo-concave heart valves: a clinical and economic evaluation

J.H.P. van der Meulen, E.W. Steyerberg, Y. van der Graaf, L.A. van Herwerden, C.J. Verbaan, J.J.A.M.T. Defauw, J.D.F. Habbema.
Circulation 1993; 88: 156-164

Abstract

Background: Björk-Shiley convexo-concave heart valves have an increased risk of mechanical failure. One might consider prophylactic re-replacement as a preventive measure to avert the disastrous consequences of these failures. We investigated the effect that prophylactic re-replacement has on survival of individual patients and on the medical costs.

Methods: Quantitative estimates for the surgical risks of prophylactic replacement of Björk-Shiley valves, long-term survival and the risk of outlet strut fracture were as much as possible derived from a detailed analysis of a follow-up study conducted in The Netherlands, including 2303 patients with a mean follow-up of 6.6 years. On the basis of these estimates, we calculated the life-expectancy with and without prophylactic replacement. For the various valve types, age thresholds were determined, below which re-replacement prolongs the (discounted quality-adjusted) life-expectancy. We also calculated the cost per year of life gained as a function of age.

Results: The age thresholds below which prophylactic re-replacement increases life-expectancy (expressed in simple future years of life) for male patients without comorbidity, if the surgical mortality after re-replacement is equivalent to that of primary replacement, are 27, 48, 51 and 65 years, for small and large 60° and for small and large 70° mitral valves, respectively. For aortic valves these age thresholds lie somewhat higher: 39, 52, 56 and 76 years, respectively. Repeat analyses indicated that for females all aga thresholds lie about one or two years higher. These age thresholds decrease considerably if the surgical mortality after re-replacement is considered to be higher after prophylactic re-replacement than after primary replacement, or if comorbidity is present. The costs per discounted and quality-adjusted year of life gained depend on type and position of the Björk-Shiley convexo-concave heart valve and rise steeply as the patient's age approaches the threshold for re-replacement.

Conclusions: The results of the Dutch follow-up study allow guidance for prophylactic replacement of the Björk-Shiley convexo-concave valve on an individual basis. Re-replacement compares favorably with expectant management in some patient subgroups with both 60° and 70° valves. Age thresholds may serve as a first step to identify patients in whom re-replacement might be beneficial.

10.1 Introduction

The Björk-Shiley convexo-concave heart valve was withdrawn from the market in 1986 after repeated reports of mechanical failure. This type of heart valve had been developed in the early 1970s by Shiley Inc. as an improvement on the spherical disc valve. About 86,000 patients had received the Björk-Shiley convexo-concave (BScc) valve worldwide: about 82,000 with an opening angle of 60° and about 4,000 with an opening angle of 70°. By November, 1991, 466 outlet strut fractures had been reported to Shiley, which has to be considered as an underestimate of the true incidence[1].

Recently, the results of a retrospective cohort study were reported that provided detailed information on all patients in The Netherlands with a BScc valves[2]. It describes the experience of 2588 BScc valves implantated in 2303 patients between 1979 and 1985, followed-up for a mean of 6.6 years. Information on vital status was obtained from municipality registers and information about the cause and mode of death was obtained from the patient's general practitioner or retrieved from clinical records. The yearly risk of strut fracture appeared to be constant over time. It was demonstrated that the risk was greater for larger valves (≥29 mm), for valves with an opening angle of 70°, for valves implanted in the mitral position and for valves of younger patients.

Prophylactic replacement of BScc valves is generally not recommended[3,4,5]. It has only been suggested for patients with early production 70° BScc valves with a diameter ≥ 29mm (group I valves), which are known to be especially vulnerable[5]. The findings in the Dutch follow-up study however show that a high risk of strut fracture is not limited to these early production series of the 70° valves.

In this study, we evaluate the effects of prophylactic re-replacement using prognostic information obtained from the Dutch follow-up study. For each valve type, the age of the patients is determined, below which re-replacement is beneficial. Further evaluation included the cost-effectiveness of re-replacement as a function of the patient's age.

10.2 Methods

10.2.1 Structure

The structure of the problem is represented by the decision tree in Figure 1. The model contains four health states ('alive with a BScc valve', 'alive without a BScc valve', 'alive with severe morbidity without a BScc valve' and 'death'). A Markov process was applied to calculate the patient's life-expectancy[6]. In a Markov process the patient's prognosis is represented as a sequence of particular states of health and the possible transitions

among them during fixed time intervals (Markov cycle)[7]. The duration of the Markov cycle in the present model is one year. The crux of this approach is that we estimate after each subsequent year the probability that a patient is in one of the defined health states; in other words, we construct hypothetical survival curves. These survival estimates allow us now to calculate the expected lifetime a patient will spend in each of the health states. The calculations were performed with the Decision Maker computer software (New England Medical Center, 1988).

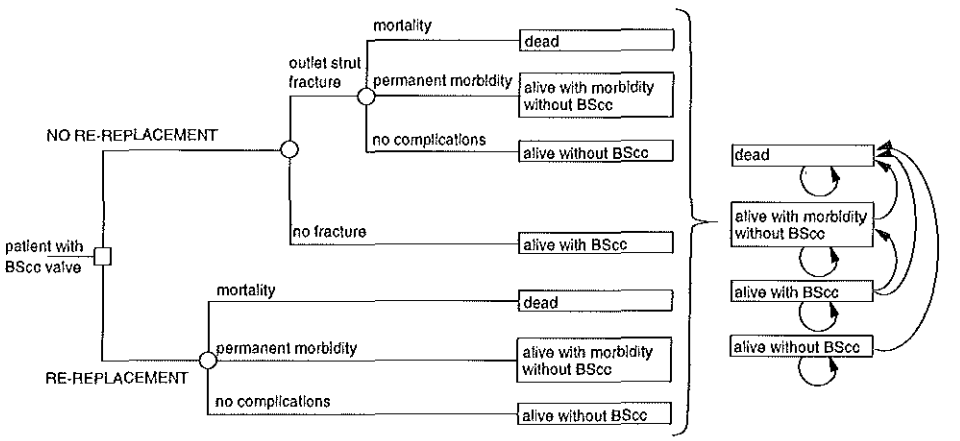

Figure 1 Decision tree for prophylactic replacement of a BScc valve. A representation of the four possible health states and the transitions among them is given on the right.

10.2.2 Probability estimates

The probability estimates required for this analysis were 1) the surgical mortality and morbidity after prophylactic re-replacement, 2) the age-specific annual risk of death, 3) the annual risk of strut fracture and 4) the mortality and morbidity after strut fracture. If possible, these probabilities were derived from the Dutch follow-up study, including 2303 patients with a mean duration of follow-up of 6.6 years[2]. We re-analyzed the data from this study using logistic regression and Poisson regression to derive prognostic models for the surgical mortality, age-specific risk of death and the risk of fracture. Variable selection was performed with a forward stepwise procedure based on the significance level of the partial likelihood ratio test (limit for significance to enter 0.10). In the first place, we will estimate the effects of prophylactic re-replacement for patients without comorbidity. Later we will also explore to some extent the effect of comorbidity, such as a poor ventricular function.

10.2.3 Surgical mortality and morbidity

Data about the risks associated with prophylactic replacement of artificial valves is scarce[8,9]. It has been emphasized that, when estimating the risks of prophylactic re-replacement, one must not only take into account the increased hazard for death during the early post-operative period (for example the first 30 days after surgery), but also the increased hazard during the entire first postoperative year[3]. This opinion however is not supported by the observation that mortality after re-replacement, for other reasons than an increased risk of mechanical failure, seems to decline rapidly to a constant level already after two weeks[10]. Therefore, we defined surgical mortality of re-replacement as that occurring during the first 30 days after surgery. Furthermore, we assumed it to be equivalent with the 30-day mortality after primary valve replacement.

Surgical mortality after re-replacement is then estimated with a logistic model derived from the Dutch follow-up study. Surgical mortality for a 40-year-old patient without any risk factor is 1.5% (odds 0.015). Age (OR 1.022, 95%CI 1.000 - 1.044, for each additional year), a BScc in the mitral position (OR 2.6, 95%CI 1.6 - 4.1), concomitant bypass surgery during valve replacement (OR 1.5, 95%CI 1.0 - 2.5), acute endocarditis (OR 2.2, 95%CI 1.2 - 4.4), a poor ventricular function (OR 2.9, 95%CI 1.5 -5.7) and valve replacement as emergency treatment (OR 6.3, 95%CI 2.6 - 15.5) are incremental risk factors. In the Dutch cohort study the left ventricular function was classified from the right oblique view of the left ventricular angiogram as good, reduced or poor[2].

For example, the surgical mortality of a 50-year-old male patient with a large mitral BScc valve without comorbidity can be estimated as 4.6%. (odds $0.015 \times 1.022^{10} \times 2.6$).

Information on the risk of permanent morbidity after valve surgery is derived from the results of a follow-up study on neurological complications of coronary bypass surgery by Shaw and coworkers[11]. In this study four of the 304 patients who survive surgery were considered to have severe permanent neurological disability. Thus the risk of permanent severe morbidity is estimated to be 1.3%.

10.2.4 Age-specific annual risk of death

In general, the life-expectancy of patients with mechanical valves is lower than that of the general population[10,12]. To obtain age-specific mortality rates we carried out Poisson regression for death after primary valve replacement. Patients with strut fractures were considered as censored observations. Patients who did not survive the first year were left out, because of the reported higher mortality during the first post-operative year (compared with mortality during later years) after *primary* valve replacement (see also above)[3]. The annual mortality rate for a female patient younger than 40 without any risk factor is 0.0061 (annual risk of death is 1-exp(-0.0061)=0.6%). Age (hazard ratios for patients between 40 and 49 1.26, 95%CI 0.66-2.44, between 50 and 59 2.16, 95%CI 1.23 - 3.79, between 60 and 69 3.64, 95%CI 2.08 - 6.37 and between 70 and 79 7.32, 95%CI 3.94 - 13.60), concomitant bypass surgery (HR 1.45, 95%CI 1.14 - 1.84), a BScc valve in the mitral position (HR 1.62, 95%CI 1.26 - 2.08) and male gender (HR 1.29, 95%CI 1.00 - 1.67) are incremental risk factors. From this Poisson model we

approximated the age-specific annual risk of death for patients after valve replacement. This approximation is based on the assumption that mortality after the operative period depends on the attained age and the condition of the patient rather than on the time elapsed since valve replacement. The age-specific hazard rates were assumed to be constant for patients younger than 35, while those for patients older than 80 were estimated on the basis of exponential extrapolation. The age-specific annual risks of death are based on the condition of the patient at the time of primary valve replacement and the present age. For example, the annual risk of death for a 20-year-old male patient with a large mitral BScc valve without comorbidity is 1-exp(-0.0061x1.62x1.29)=1.3%; for a 65-year-old female patient with a small aortic BScc valve this risk is 1-exp(-0.0061x3.64)=2.2%.

Table 1 Poisson regression model of strut fracture of Björk-Shiley convexo-concave heart valves.*

Risk factor	Rate Ratio
Age at valve implantation (years)	
40-50	0.42 (0.017-1.05)
> 50	0.30 (0.015-0.59)
Position of BScc valve	
mitral vs aortic	3.25 (1.3 - 8.3)
Valve size	
≥ 29mm vs < 29mm	3.75 (1.6 - 8.7)
Opening angle	
70° vs 60°	5.82 (3.1 - 11)

* Values in parentheses are 95% confidence intervals

10.2.5 Annual risk of outlet strut fracture

The results of the Dutch follow-up study indicated that the annual risk of strut fracture is constant over time and depends on valve characteristics and age at implantation. We performed Poisson regression to estimate these effects on the annual risk of valve fracture (see Table 1). The baseline risk of strut fracture was 0.09%/year (95%-CI: 0.03-0.22%/year) for patients younger than 40 years with a 60° aortic valve < 29mm. Table 1 can be used to calculate the risks for patients of other ages with other types of valves. For example, the annual risk of strut fracture for a 20-year-old male patient with a large 60° mitral valve, who was 12 at implantation, is 1 - exp(-0.0009•3.75•3.25)=1.1% (95%-CI: 0.6-2.1%); for a 65-year-old female patient with a small aortic 60° valve, who was 57 at implantation, this risk is 1-exp(-0.0009•0.030)=0.03% (95% CI: 0.01-0.06%).

Table 2 Observed and expected 8-year probabilities of outlet strut fracture for the various Björk-Shiley convexo-concave valve types.*

Valve type		8-year % Observed		Expected
mitral valves				
small 60°	(n=305)	0.0	(0.0-1.5)	0.9
large 60°	(n=677)	4.2	(2.7-6.5)	3.3
small 70°	(n=55)	7.2	(3.4-21)	5.8
large 70°	(n=93)	17.3	(9.1-32)	19.8
aortic valves				
small 60°	(n=1241)	0.2	(0.0-0.8)	0.3
large 60°	(n=86)	0.0	(0.0-5.1)	1.5
small 70°	(n=115)	3.7	(1.2-11)	1.7
large 70°	(n=16)	8.3	(1.2-46)	8.1

* Values in parentheses are 95% confidence intervals

We present the observed 8-year risk of strut fracture in Table 2 together with the predicted 8-year risk for the various valve types. The predicted probabilities were calculated with the split-quarter method. This implies that the cohort was randomly split into four groups of equal size and that a model was estimated on three of the four groups (training set). The fourth group (test set) was then used to predict the annual fracture risk. This was repeated four times so that all four groups served as a test set once. The results of this cross-validation procedure indicate that the performance of this prognostic model is adequate.

10.2.6 Mortality and morbidity after fracture

A patient sustaining an outlet strut fracture of a mechanical valve may die immediately or after an attempted emergency valve replacement. The mortality after aortic strut fracture is high: in the Dutch cohort study, 6 out of 7 reported patients died (86%). Mortality after mitral strut fracture was lower: of the 35 reported patients 18 died (51%). These mortality rates were adopted in the present analysis.

It is assumed that 50% of the survivors of an outlet strut fracture will have severe permanent morbidity. This estimate is based on an evaluation of the functional status of the Dutch patients who survived the outlet strut fracture.

10.2.7 Outcomes

We calculated life expectancy with and without replacement of the BScc valve. To account for the fact that most patients are risk averse (in other words, they attach more value to nearby life years than to life years in the distant future) we investigated the effects of discounting future life years at 5% per year[13] (this implies that the value of each additional year decreases with 5%) and also the effects of adjusting for the quality of life by weighing the time spent with severe permanent morbidity from valve surgery or outlet strut fracture with a quality adjustment factor of 0.5 (each year for a patient with severe morbidity is worth half a year in full health).

We represent the direct medical costs for re-replacement and expectant management in 1990 Dutch guilders (*f*)[14,15,16] (1 US Dollar was approximately 2 Dutch guilders). The costs of prophylactic re-replacement are estimated to be *f*20,000.- (angiography, surgery, 3 days intensive care and 10 days low care). The costs of an outlet strut fracture amount to *f*15,000.- for a patient who dies after admission to hospital (surgery and 5 days intensive care). Furthermore, it is assumed that 50% of the patients who die after an outlet strut fracture die outside the hospital. The costs for patients who survive after an outlet strut fracture are *f*45,000.- (surgery for valve replacement, 5 days intensive care and 20 days low care followed by surgery for removal of the fractured strut, 1 day intensive care and 12 days low care). Future costs were discounted to present value at a 5% per year discount rate.

10.2.8 Simplifications

Our analysis was subject to the following simplifying assumptions: 1) the surgical mortality of prophylactic re-replacement is equivalent to the 30-day mortality after primary valve replacement; 2) survival of patients with an artificial valve is determined by their attained age and clinical condition and not by the time elapsed since primary valve replacement; 3) replacement of the BScc valves obviates the risk of strut fracture without affecting long-term mortality and morbidity; 4) the annual risk of strut fracture is constant over time; 5) mechanical heart valves have an infinite life span, except for the risk of strut fracture in BScc valves.

We calculated the (discounted and quality-adjusted) life-expectancy of patients with one BScc valve; the effects of prophylactic valve re-replacement in patients with more than one artificial valve have not been dealt with.

10.3 Results

The effects of prophylactic re-replacement on the life-expectancy of patients with BScc valves are first presented for a fictitious 40-year-old male patient with a 29 mm 60° mitral BScc valve and no comorbidity. According to the prognostic models presented earlier we estimated for this patient the annual risk of strut fracture to be 1.1%, surgical mortality 3.7% and the "basal life-expectancy", that is the life-expectancy, if the strut fracture risk is assumed to be zero, 25.0 years. The life-expectancy of this patient increases from 23.1 to 24.1 years if prophylactic re-replacement is performed. So, prophylactic re-replacement adds 1.0 years (or 4.3%) to the life-expectancy. Expressed in terms of loss, prophylactic re-replacement gives rise to a 53% reduction of the loss in life-expectancy that is attributable to outlet strut fracture (1.0 from 1.9 years).

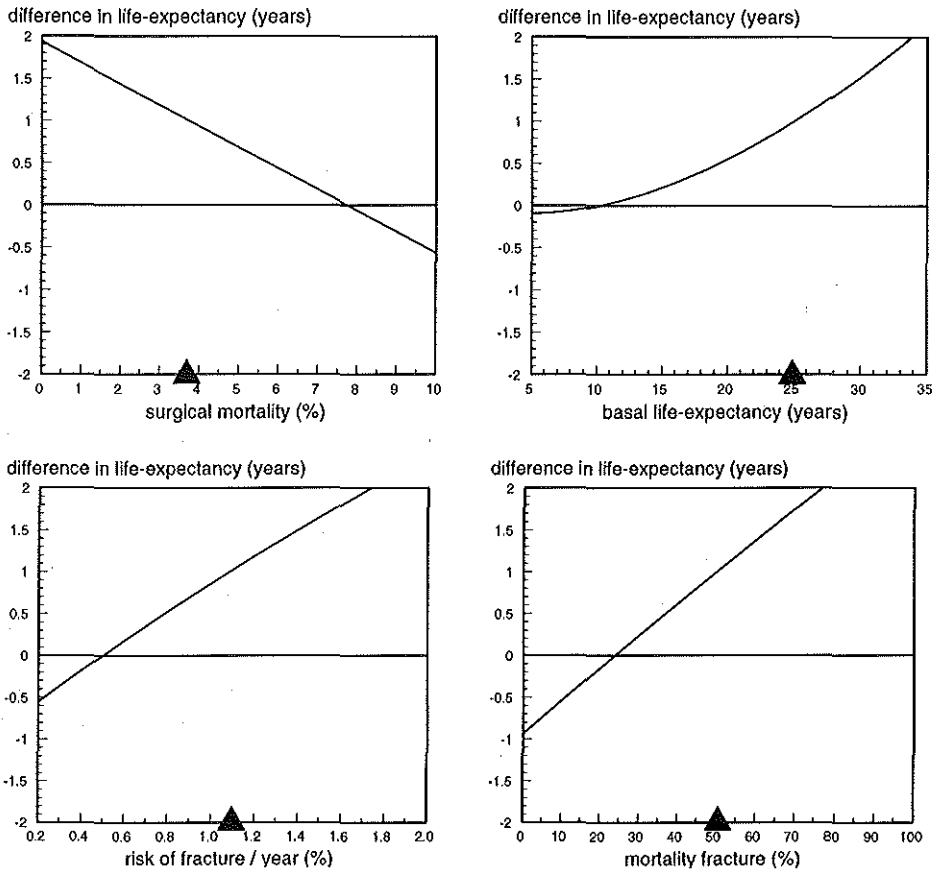

Figure 2 Sensitivity analysis of the difference in life-expectancy with and without prophylactic re-replacement for a 40-year-old male patient with a 29mm 60° mitral BScc valve.

To investigate how sensitive these results are to variations in the quantitative estimates, we varied each estimate one by one over wide ranges. In Figure 2 we show the effects of these variations on the difference in life-expectancy. A positive difference indicates that prophylactic re-replacement results in an extension of the life-expectancy. This Figure demonstrates that all estimates have a substantial independent effect on this difference. Re-replacement gives the higher life-expectancy in the 40-year-old patient, taken as an example, if the surgical mortality after valve re-replacement is 7.7% or less, if the basal life-expectancy is 10.3 years or more, if the annual risk of strut fracture is 0.50% or more, or if the mortality after strut fracture is 23.9% or more.

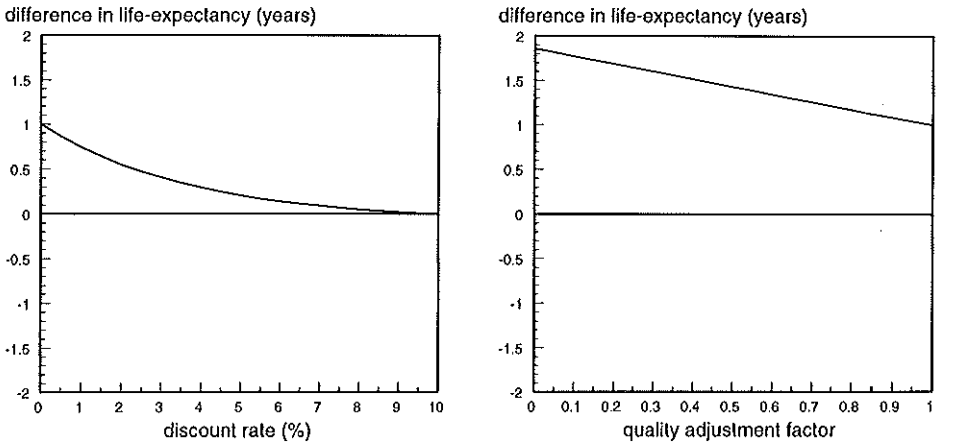

Figure 3 Sensitivity analysis for variation in the yearly discount rate of future life-years and in the quality adjustment factor for permanent morbidity caused by prophylactic re-replacement or strut fracture. Same patient as Figure 2.

In Figure 3 we present the effects of variations in the preferences for the length and quality of life. The difference in discounted life-expectancy decreases to zero if the discount rate of future years of life increases from 0% to 10%. If the quality adjustment factor decreases from one (morbidity is equivalent with full health) to zero (morbidity is equivalent with death) the advantageous effect of re-replacement increases.

In Figure 4 the age thresholds for prophylactic replacement of BScc valves are shown for patients without comorbidity. For each valve type we calculated age thresholds, firstly using simple future years of life, secondly using discounted years of life and thirdly using discounted and quality-adjusted years of life. To account for the statistical uncertainty we indicated confidence intervals for these age thresholds using the upper and lower limits of the 95% CI interval of the estimated strut fracture risk. It can be read from the upper panel in Figure 4 that the life-expectancy (in simple future years of life) for male patients with a small 60° mitral valve is higher with than without prophylactic re-replacement if they are younger than 27 years. For male patients with a 60° large mitral BScc valve this age threshold is 48 years. The age thresholds are considerably higher for valves with an opening angle of 70°: 51 and 65 years for small and large mitral valves, respectively. The age-thresholds for aortic valves are somewhat higher than for mitral valves. They are 39 and 52 years for small and large 60° aortic valves, and 56 and 76 years for small and large 70° BScc aortic valves, respectively. For female patients all age thresholds lie one or two years higher (lower panel in Figure 4).

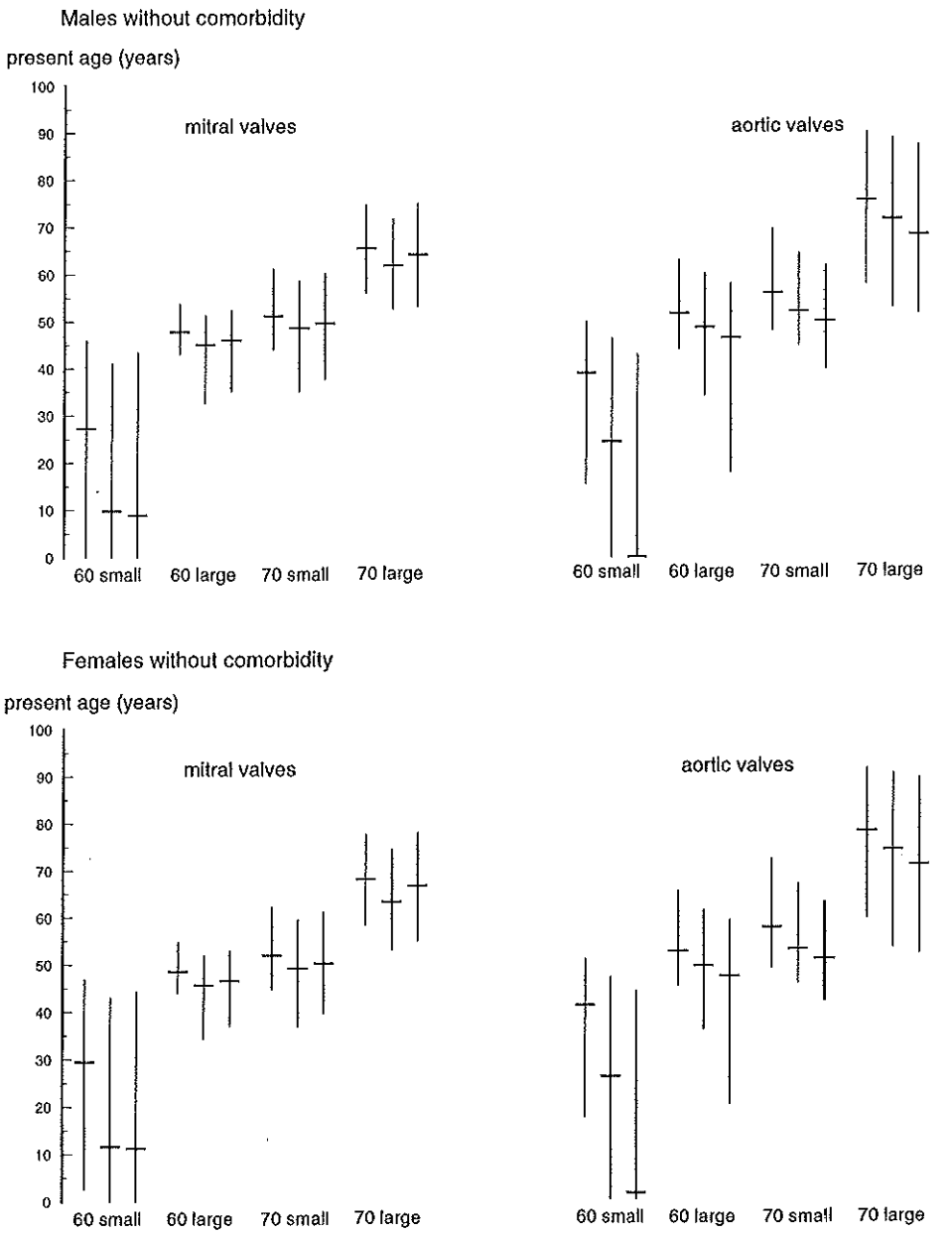

Figure 4 Age thresholds for prophylactic re-replacement in male and female patients, according to simple years of life, discounted years of life, and discounted and quality-adjusted years of life (from left to right).

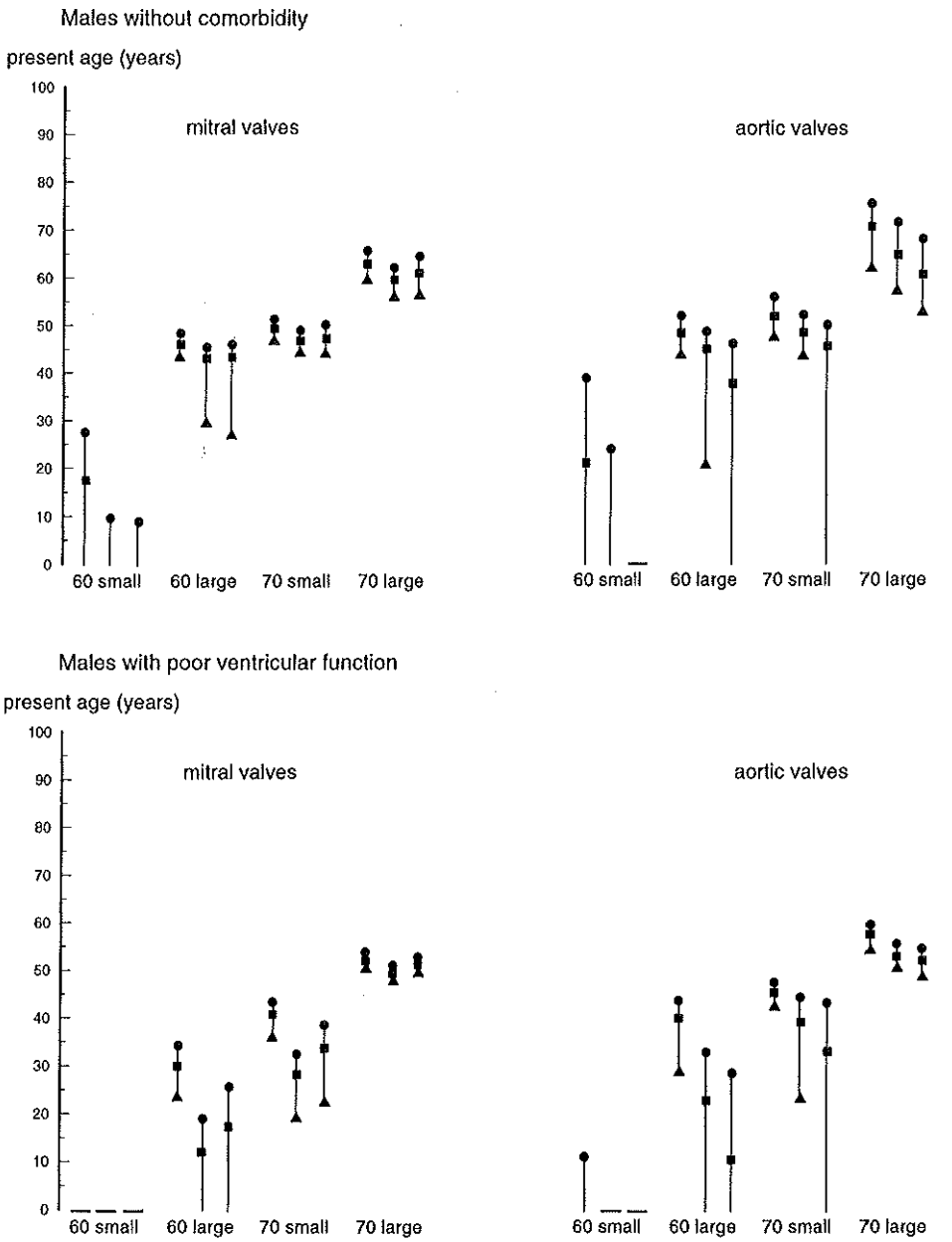

Figure 5 Age thresholds for prophylactic re-replacement in male patients, without comorbidity, and with a poor left ventricular function, when 0% (●), 1% (■), or 3% (▲) is added to the surgical mortality.

The effect of discounting future years of life and adjusting for the quality of life on the age thresholds is relatively small for most valve types. Only for small 60° valves either in the mitral or aortic position, the age thresholds decrease considerably when discounted future years of life are used.

The age thresholds presented so far are based on the assumption that the surgical mortality for prophylactic re-replacement is the same as that following primary valve replacement. In the upper panel of Figure 5 we present the age thresholds for male patients without comorbidity, if 1% or 3% is added to the surgical mortality as well as to the risk of permanent morbidity after prophylactic re-replacement. If 3% is added to these surgical risks, prophylactic re-replacement always gives a lower life-expectancy for male patients with small 60° mitral valves. With this increase of 3%, the age thresholds (based on simple future years of life) decrease with approximately 5 years for the other valve types in the mitral position. For small 60° aortic valves, the life-expectancy with prophylactic replacement is always lower than with an expectant management. For the other valve types in the aortic position, the effect of an increase in the surgical risks is somewhat larger than for mitral valves: the age thresholds decrease with about 8 years for large 60° and small 70° valves and with 16 years for large 70° valves.

The lower panel of Figure 5 demonstrates the effect of a poor left ventricular function on the age thresholds for prophylactic re-replacement (a poor ventricular function is an increasing risk factor for surgical mortality; Odds Ratio 2.9). The age thresholds (based on simple future years of life) decrease with at least 13 years for the mitral valves and with at least 8 years for the aortic valves.

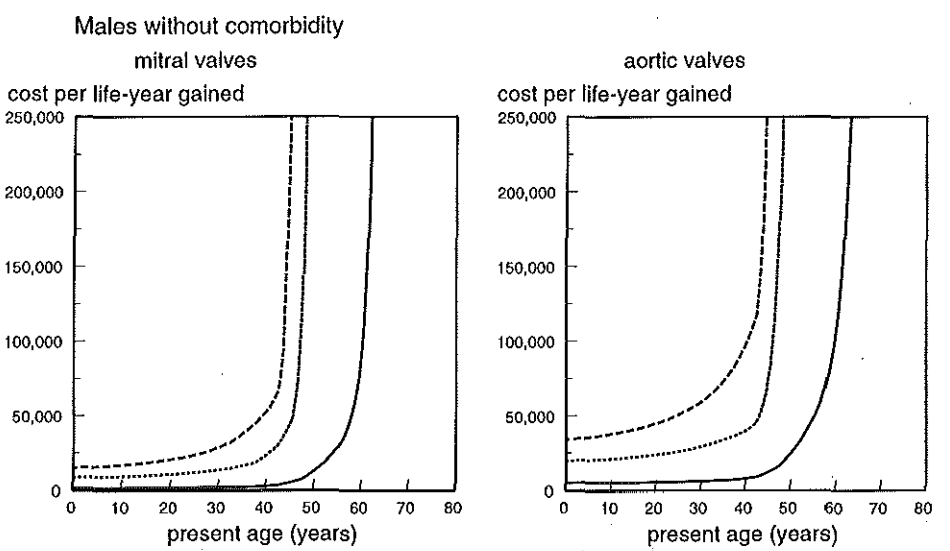

Figure 6 Costs per discounted and quality-adjusted life-year gained, according to age for male patients with large 60° (dashed line), small 70° (dotted line) and large 70° (continuous line) BScc valves.

Figure 6 shows the marginal cost-effectiveness of prophylactic re-replacement for male patients without comorbidity as a function of age for the various valve types. The costs per discounted and quality-adjusted life-year gained depend upon valve type and position. Replacement of mitral BScc valves produce lower cost-effectiveness ratios than replacement of aortic valves, indicating that re-replacement is more cost-effective. The costs per discounted and quality-adjusted life-year gained rise steeply as the patient's age approaches the threshold for re-replacement. Repeat analyses for females gave similar results.

10.4 Discussion

Until recently, prophylactic replacement of BScc valves was recommended only for patients with large (≥ 29 mm) 70° BScc valves of an early production series (group I 70° BScc valves)[5]. Our study indicates that also patients with other BScc valve types may benefit from prophylactic re-replacement. Prophylactic replacement of BScc valves may increase the discounted and quality-adjusted life-expectancy in patients without comorbidity with large 60° mitral and aortic BScc valves, if they are younger than about 45 and in patients with small 70° mitral and aortic valves, if they are younger than about 50. The age thresholds for prophylactic replacement are high for aortic BScc valves, considering the relatively low strut fracture risk is taken into account. The explanation for this result is the high mortality following aortic strut fracture on the one hand and the relatively low surgical risks and high life-expectancy in patients with aortic valves on the other. The slightly higher age thresholds that we established for female patients can be explained by the higher basal life-expectancy of female patients. The cost-effectiveness of prophylactic re-replacement depends strongly on age, valve type and valve position. This is to be expected, because these factors determine the extent of the survival advantage of valve replacement.

If the surgical risks after prophylactic re-replacement are thought to be higher than after primary replacement however, the age thresholds, below which prophylactic re-replacement prolongs the discounted and quality-adjusted survival, decrease considerably for the large 60° and small 70° aortic valves and to a lesser extent for the large 60° mitral BScc valve. Conceivably, the age thresholds for prophylactic replacement are also significantly lower, if the presence of a comorbid factor, such as a poor left ventricular function, is present. For patient with a large 70° BScc valve however with its high risk of strut fracture the effect of comorbidity on the age thresholds is small.

In a recent study Birkmeyer and coworkers reported the operative risk thresholds below which replacement of a BScc valves increases life-expectancy expressed in simple future years of life[17]. The estimates of the fracture rate in this study were derived from Dutch and Swedish follow-up studies and from an international multi-institutional follow-up study of patients with 70° BScc valve[2,3,18]. The strut fracture risks they used for the 70° mitral valve were on average about 40% lower than we used in our study and their fracture risks for 60° aortic valves were more than twice as high as ours. Their recommendations for re-replacement agree with ours to a large extent. The operative

risk thresholds they present in their study however are, even for the 70° mitral valves, rather high. For example, according to their study, replacement of large 70° mitral BScc valves seems advantageous in patients up to 70 years of age (the operative risk thresholds are 5.3% and 6.7% for male and female patients, respectively). For patients with small 70° mitral BScc valves re-replacement seems beneficial in patient up to 50 years of age (operative risk thresholds are 5.0% and 5.9% for male and female patients, respectively). Their results indicate that for large 70° aortic BScc valves replacement is advantageous in patients of even older age (operative risk thresholds are 4.1% and 5.6% for 80-year-old male and female patients, respectively). Furthermore, this study confirms our conclusion that re-replacement of the large 60° mitral valves may be advantageous in patients without comorbidity up to 45 years old. The high operative risk thresholds for older patients in the study of Birkmeyer and coworkers can partly be explained by their assumption that the strut fracture rates are age-independent, whereas the Dutch follow-up study indicated a strong decrease of the strut fracture rates with age at implantation.

In our study we derived the quantitative estimates from a detailed multivariate analysis of the Dutch follow-up study. This allowed us to develop prognostic models for valve fracture as well as for surgical mortality and life-expectancy. Another reason for not using the results of studies from different countries on the strut fracture rate are indications that the fracture rate depends on batch-related manufacturing deficiencies and that different countries have received different production series[12,19]. This implies in our view that recommendations on prophylactic re-replacement have to take into account the batch-specific strut fracture risk estimates.

When interpreting the results of the present study one must also take into account that a number of strut fractures have remained undetected[2,3], thereby leading to an underestimate of the risk of strut fracture as well as of the mortality of strut fracture. Another element of uncertainty in this respect is the risk of strut fracture in the distant future. Our estimations of the effect of re-replacement are based upon a constant fracture risk over time, which is supported by the observations during the follow-up period and also by the metallurgical investigations, which indicated fatigue at the welding sites of the outlet strut as a possible cause of fracture[21,22]. If however a distinct rise or fall of the number of strut fractures is observed in the future, the indications for replacement of the BScc valves have to be adjusted accordingly. It is shown in our study that variations in the annual risk of strut fracture have a considerable effect on the indication for prophylactic re-replacement (see Figure 2).

When considering prophylactic replacement of BScc heart valves, the patient and his or her doctor have to balance the consequences of cardiac surgery against the possibility that a strut fracture may occur at some moment in the future with its associated high mortality. In general, most patients tend to be risk averse and consider current benefits preferable to future benefits[13]. In other words, they consider life during the next few months more important than during later years. We estimated therefore also the age thresholds discounting future years of life. Next to this attitude towards the surgical risk, a patient may wish the BScc valve to be replaced, because of the fear and anxiety evoked by the possibility that the artificial valve may fail mechanically. This risk attitude is not

explicitly modelled in our study. In this respect, it is important to note however that mechanical failure is only one of the dangers that threaten patients with artificial heart valves[23]. We demonstrated for instance that the life-expectancy of a 40-year-old patient with a BScc mitral valve *without* any risk of valve fracture and, who does not have concomitant morbidity, is considerably lower than the life-expectancy according to the vital statistics of the general Dutch population (24.8 compared to 34.9 years). Finally, one has to account for the postoperative morbidity, which may interfere with a patient's normal activities of daily life during at least a few months.

The results of the Dutch follow-up study allow recommendations for prophylactic replacement of the BScc valves. The age thresholds presented in this study may serve as rough guidelines for patient selection but can never substitute definitive decision making that should be based on an individual evaluation of the strut fracture rate (which may vary between countries), the risks of surgery, the life-expectancy and the patients's attitude towards the alternative options.

This study was supported in part by a grant from the Netherlands' Health Research Promotion Programme (SGO). The authors wish to thank M.A. Koopmanschap, Dept of Public Health, Erasmus University Rotterdam, for his contribution to the economic evaluation, and Professor J. Benbassat, Faculty of Health Sciences, Ben Gurion University, Beer Sheva, Israel, for his helpful comments.

References

1. Pfizer / Shiley Heart Valve Research Centre. Dear Doctor letter. March, 1992.
2. Van der Graaf Y, De Waard F, Van Herwerden LA, Defauw J. Risk of strut fracture of Björk-Shiley valves. Lancet 1992; 339: 257-261.
3. Lindblom D, Björk VO, Semb BKH. Mechanical failure of the Björk-Shiley valve. J Thorac Cardiovasc Surg 1986; 92: 894-907.
4. Hiratzka LF, Kouchoukos NT, Grunkemeyer GL, Miller DC, Scully HE, Wechsler AS. Outlet strut fracture of the Björk-Shiley 60° convexo-concave valve; current information and recommendations for patient care. J Am Coll Cardiol 1988; 11: 1130-1137.
5. Lindblom D, Rodriguez L, Björk VO. Mechanical failure of the Björk-Shiley valve. J Thorac Cardiovasc Surg 1989; 97: 95-97.
6. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical Decision Making. Boston MA: Butterworths, 1988.
7. Beck JR, Pauker SG. The Markov process in medical prognosis. Med Decis Making 1983; 419-458.
8. Wideman FE, Blackstone EH, Krklin JW, Karp RB, Koucoukos NT, Pacifico AD. The hospital mortality of re-replacement of the aortic valve. Incremental risk factors. J Thor Cardiovasc Surg 1982; 82: 692-698.
9. Husebye DG, Pluth JR, Peihler JM, et el. Reoperation on prosthetic heart valves. J Thorac Cardiovasc Surg 193; 86: 543-552.
10. Blackstone EH, Kirklin JW. Death and other time-related events after valve replacement. Circulation 1985; 72: 753-767.

11. Shaw PJ, Bates D, Cartlidge NEF, Heaviside D, French JM, Juliand DG, Shaw DA. Neurological complications of coronary artery bypass graft surgery: six month follow-up study. British Medical Journal 1986; 293: 165-167.
12. Kirklin JW, Barratt-Boyes BG. Cardiac Surgery. New York NY: John Wiley & Sons, 1986.
13. McNeil BJ, Weichselbaum R, Pauker SG. Fallacy of the five-year survival in lung cancer. N Eng J Med 1978; 229: 1397-1401.
14. Vrieze OJ, Boas GM, Janssen JHA. An econometric model for a scenario analysis of coronary heart disease (in Dutch). Maastricht: Rijksuniversiteit Limburg, 1992.
15. Financial statistics 1990 (in Dutch). Utrecht: Nationaal Ziekenhuisinstituut, 1991.
16. Van Hout BA. Heart transplantation; costs, effects and prognosis (thesis in Dutch). Rotterdam: Erasmus Universiteit, 1990.
17. Birkmeyer JD, Marrin CAS, O'Connor GT. Should patients with Björk-Shiley valves undergo prophylactic replacement? Lancet 1992; 340: 520-523.
18. Ericsson A, Lindblom D, Semb G, Huysmans HA, Thulin LI, Scully HE, Bennett JG, Ostermeyer J, Grunkemeyer GL. Strut fracture with Björk-Shiley 70° convexo-concave valve; an international multi-institutional follow-up study. Cardiothoracic Surgery 1992; 6: 339-346.
19. Pfizer / Shiley Heart Valve Research Centre. Dear Doctor letter. September, 1992.
20. Woodyard C, 1991. Firm told to warn 350 with heart valves. Los Angeles Times. April, 28, 1991.
21. Sacks SH, Harrison M Bischler PJ, Martin JW, Watkins, Gunning A. Metallurgical analysis of failed Björk-Shiley cardiac valves prostheses. Thorax 1986; 41: 142-147.
22. Van Swieten HA, De Mol BA, Defauw J, Overkamp PJ, Vermeulen FEE. Metallurgical analysis of the Björk-Shiley convexo-concave valve prosthesis to assess the cause of late outlet structure. In: Bodinat E, ed. Surgery for heart valve disease. London UK, ICR Publishers, 1990: 616-627.
23. Taylor K. Acute failure of artificial heart valves. The risk is small. Br Med J 1988; 297: 996-997.

11 Prophylactic replacement of Björk-Shiley convexo-concave heart valves: an easy-to-use tool for decision support

E.W. Steyerberg, J.H.P. van der Meulen, L.A. van Herwerden, J.D.F. Habbema.
Br Heart J, accepted for publication

Abstract

Background: Björk Shiley convexo-concave heart valves have a considerable risk of strut fracture, which is associated with a high lethality. If one considers prophylactic replacement of the valve, this fracture risk has to be weighed against the risks of reoperation. Estimates of strut fracture risk and reoperation mortality have recently been subject to revisions.
Objective: This study aimed to develop an easy-to-use tool for decision making on prophylactic replacement of Björk Shiley convexo-concave heart valves.
Methods: A decision analysis model was used to calculate the loss of life-expectancy caused by strut fracture and by elective prophylactic replacement. Quantitative estimates for the model were obtained from a large follow-up study in The Netherlands and recently published studies.
Results: A simple graph was constructed that presents the loss of life-expectancy (LE) caused by strut fracture for combinations of basal LE (LE without strut fracture) and lethal fracture risk (strut fracture risk multiplied by lethality of fracture). This loss of LE can directly be compared with the loss caused by surgical mortality. The calculations take individual patient characteristics into account, such as age, gender, cardiac comorbidity, position of the valve, and can easily be made by hand or with a simple computer application.
Conclusions: This decision support tool enables the direct estimation of the gain or loss of life-expectancy by replacement of a Björk-Shiley convexo-concave heart valve. The tool may be useful for evaluation of individual patients as well as groups of patients, and allows for easy incorporation of revisions of fracture risk estimates.

Abbreviations: LE: life-expectancy; Bscc valve: Björk-Shiley convexo-concave valve

11.1 Introduction

Björk Shiley convexo-concave (BScc) heart valves have a considerable risk of mechanical failure[1,2,3,4]. Prophylactic replacement of a BScc valve may be considered to avert this risk. We aimed to develop a simple tool that quantifies the gain or loss of life-expectancy by replacement of a BScc valve. This tool should support decision making in individual patients. Moreover, it should assist the treating clinician in screening groups of patients on the consequences of recent and forthcoming revisions of fracture risks.

BScc heart valves were withdrawn from the market in 1986 after reports of mechanical failure (outlet strut fracture). Fracture risk estimates were reported in several studies[1,2,3,4]. The largest of these is a follow-up study of 2588 BScc valves implanted in 2303 patients[3]. This study calculated the risk of strut fracture on the basis of valve characteristics (site of implantation, size of the valve, opening angle) and the patient's age at implantation. Recent revisions[5,6] of the fracture risk estimates include production characteristics such as weld date and 'remilling' status, thus distinguishing a large number of subgroups of BScc valves. Also, the welder of the valve is considered a risk factor for strut fracture[7], and the most recent estimates for 60° valves incorporate this characteristic[6].

Two decision analyses[8,9] have quantitatively compared the risk of strut fracture, which accumulates over time if the valve is not replaced, with the elective surgical risk of prophylactic replacement. These decision analyses presented surgical risk thresholds[8] or age-thresholds[9] below which elective replacement increases the life-expectancy. The published presentation of these analyses does not allow the calculation of the number of years expected to be gained or lost by replacement for individual patients with specific risk profiles. Moreover, both analyses used risk estimates that differ from recent estimates, both for fracture risk and surgical risk of reoperation[10]. We aimed to overcome these drawbacks with a flexible and easily applicable decision support tool. This tool quantifies the benefit of elective replacement compared with no replacement, taking into account individual fracture risk estimates, the patient's age, gender, position of the valve (aortic/mitral) and cardiac comorbidity.

11.2 Methods

11.2.1 Loss of life-expectancy

The loss of life-expectancy caused by replacement and the loss caused by strut fracture were calculated with a previously published Markov model[9,11]. It is assumed that the replacing valve has the same hemodynamic characteristics as the BScc valve, but carries no risk of strut fracture. Further, it is assumed that the fracture risk is constant over time, in agreement with published figures thus far[2,3,4].

The loss of life-expectancy was calculated relative to the basal life-expectancy, which is the life-expectancy of a patient with a mechanical heart valve that is similar to a BScc valve, but has no risk of strut fracture. The loss of life-expectancy caused by replacement

is by definition equal to the elective surgical mortality. The loss of life-expectancy caused by strut fracture depends on the combination of the yearly fracture risk and the lethality of fracture. Moreover, the fracture caused loss of life-expectancy decreases with decreasing basal life-expectancy, as the yearly lethal fracture risk is relatively less important if the basal life-expectancy is low. We calculated the fracture caused loss of life-expectancy as a function of basal life-expectancy and the lethal fracture risk (yearly fracture risk multiplied by lethality of fracture). This approach is based on the simplification that the loss is independent of the specific combination of fracture risk and lethality or specific determinants of basal life-expectancy. For example, a lethal fracture risk of 1%/year is present in a valve with a fracture risk of 2% and a lethality of 50%, or in a valve with a fracture risk of 1% and a lethality of 100%. And a life-expectancy of 25 years may be estimated for a 40-year-old male patient with a mitral valve and no comorbidity, but also for a 48-year-old male patient with an aortic valve and no comorbidity. To evaluate the effect of this simplification, we varied the combinations of fracture risk, lethality, and basal life-expectancy and we found that the differences in life-expectancy were always very small (<0.1 year).

11.2.2 Quantification

The presented estimates of basal life-expectancy and lethality of fracture were obtained from a large follow-up study in the Netherlands[3]. Basal life-expectancy was dependent on age, gender, position of the valve and concomitant bypass surgery. The lethality of strut fracture was found to be 86% (6 died out of 7, 95% confidence interval 47-99%) and 51% (18 died out of 35, 95% confidence interval 34-69%) in case of aortic and mitral valves, respectively.

Recently published revised estimates of fracture risk[5,6] include production characteristics such as weld date (5 periods for 60° aortic and 6 periods for 60° mitral valves), welder (group A, B or C for 60° valves), and 'remilling' status (for 70° valves). Other valve characteristics are site of implantation (aortic/mitral), size, and opening angle (60°/70°). Table 1 shows these recent risk estimates for the 49 aortic valve subgroups and 56 mitral valve subgroups.

The estimates of elective surgical mortality in case of replacement were based on a recent analysis of 2246 prosthetic valve reoperations in 1984 patients[10]. Twelve risk factors were distinguished in a logistic regression function to estimate surgical mortality of reoperation. Only age, weight, NYHA class and number of previous operations are relevant for most BScc patients, assuming that the other risk factors are absent[10].

11.3 Results

Figure 1 forms the central element of the decision support tool. It shows the loss of life-expectancy caused by fracture in relation to basal life-expectancy and lethal fracture risk. This loss can directly be compared with the surgical mortality of elective replacement. We illustrate the practical use of the tool with a fictitious 40-year-old male patient without comorbidity, who is considered for elective replacement of a mitral BScc valve, opening angle 60°, size 29mm[9].

Table 1 Fracture risk estimates (%/year) for BScc valves according to position, opening angle and size. Estimates for 60° valves were further subdivided by weld date and welder (group A/B/C, if available)[6]. Estimates for 70° valves were subdivided by remilling status[5].

	Aortic BScc valves						
	Opening angle						
	60°					70°	
	Valve weld date					Remilling	
Aortic valve size	<1/80	1/80-12/80	1/81-6/82	7/82-3/84	>3/84	remilled	non-remilled
≤21 mm	.01	.01	.01	.01	.01	.19	.19
23 mm	.09	.09	.09	.09	.01	.67	.29
25 mm	.01	.01	.01	.01	.01	.19	.19
27 mm	.09	.09	.09	.09	.01	.67	.29
29 mm	.03	.11	.17/.21/.57	.03	.01	1.33	.72
31 mm	.03	.11	.17/.21/.57	.03	.01	1.33	.72
33 mm	.03	.11	.17/.21/.57	.03	.01	1.33	.72

	Mitral BScc valves							
	60°						70°	
	Valve weld date						Remilling	
Mitral valve size	<1/80	1/80-12/80	1/81-6/81	7/81-6/82	7/82-3/84	>3/84	remilled	non-remilled
≤21 mm	.01	.01	.01	.01	.01	.01	.19	.19
23 mm	.09	.09	.09	.09	.09	.01	.67	.29
25 mm	.01	.01	.01	.01	.01	.01	.19	.19
27 mm	.09	.09	.09	.09	.09	.01	.67	.29
29 mm	.13	.11	.19/.35/1.05	.19/.35*/1.05	.13	.01	1.33	.72
31 mm	.18/.71/.88	.11	.43/.87/1.39	.43/.87/1.39	.18/.71/.88	.01	2.25	1.36
33 mm	.46/1.36/2.82	.11	1.28/1.36/2.82	1.28/1.36/2.82	.46/1.36/2.82	.01	2.25	1.36

* Hypothetical patient (see text)

Basal life-expectancy can be read from Table 2. The fictitious 40-year-old male mitral valve patient has a life-expectancy of 25.0 years. Next, we calculate the loss of life-expectancy caused by replacement, i.e. surgical mortality. Surgical mortality is presented in Table 2 according to combinations of age, number of previous operations, NYHA class and weight (60 kg may be considered as an average weight for females, 80 kg for male patients). For the fictitious 40-year-old male patient we assume a weight of 80 kg, NYHA class I and one previous open heart operation and estimate surgical mortality as 0.9%.

The loss of life-expectancy caused by fracture is calculated in four steps. First, the yearly fracture risk has to be estimated, e.g. from Table 1[5,6]. The mitral valve of our fictitious patient with an opening angle of 60°, size 29mm, if produced in December 1981 by a welder from group B, has an estimated fracture risk of 0.35%/year (Table 1). Second, the lethality of fracture has to be estimated. Using the average lethality from the Dutch follow-up study yields 51% for mitral valves. Third, the lethal fracture risk is cal

Table 2 Estimation of the basal life-expectancy[9] (life-expectancy without fracture risk) and surgical mortality[10] of BScc valve replacement.

Patient characteristics	Basal life-expectancy (years)						
	Age						
Position, gender	30	40	50	60	70	80	90
Aortic valve							
Male	38.2	31.0	23.9	17.6	12.1	8.0	4.8
Female	41.9	34.3	26.9	20.1	14.1	9.4	5.6
Mitral valve							
Male	31.3	**25.0***	18.8	13.4	8.8	5.7	3.4
Female	34.9	28.2	21.5	15.6	10.5	6.9	4.2

Patient characteristics **			Surgical mortality (%)						
			Age						
Reop #, NYHA, weight			30	40	50	60	70	80	90
First reoperation									
NYHA class	1, weight	60 kg	0.9	1.1	1.4	1.8	2.5	3.6	5.5
	1	80 kg	0.8	**0.9***	1.2	1.5	2.2	3.1	4.8
	2	60 kg	1.6	1.9	2.4	3.2	4.4	6.3	9.4
	2	80 kg	1.4	1.6	2.1	2.7	3.8	5.4	8.2
	3	60 kg	2.8	3.3	4.2	5.5	7.5	10.6	15.5
	3	80 kg	2.4	2.9	3.6	4.7	6.5	9.3	13.6
Second reoperation									
NYHA class	1, weight	60 kg	1.3	1.5	1.9	2.5	3.4	5.0	7.5
	1	80 kg	1.1	1.3	1.6	2.1	3.0	4.3	6.5
	2	60 kg	2.2	2.6	3.3	4.3	6.0	8.5	12.6
	2	80 kg	1.9	2.3	2.8	3.7	5.2	7.4	11.0
	3	60 kg	3.9	4.6	5.7	7.5	10.1	14.2	20.4
	3	80 kg	3.3	4.0	5.0	6.5	8.8	12.4	18.0

* Hypothetical patient (see text)

** It is assumed that other patient characteristics are: a stable hemodynamic state, single valve replacement, no other comorbidity (chronic renal failure, tricuspid incompetence, endocarditis), no concomitant procedures (repair of an ascending aortic aneurysm, resection of a left ventricular aneurysm, coronary artery bypass grafting)[10].

culated by multiplying the fracture risk and the lethality of fracture: 0.35•51%=0.18% /year. Finally, we use Figure 1 to determine the fracture caused loss of life-expectancy. Our hypothetical patient with a basal life-expectancy of 25 years and a lethal fracture risk of .18%/year, has an estimated loss of around 2.7%.

Comparison of the fracture caused loss of life-expectancy (2.7%) with the surgical mortality (0.9%) reveals that replacement of this particular BScc valve yields a lower loss of life-expectancy in this patient. The magnitude of this difference in years is calculated by multiplying the relative losses and the basal life-expectancy. The expected number of years lost due to surgery is 0.9%•25=0.23 year, while the loss due to strut fracture would be 2.7%•25=0.68 year. The advantage of surgery thus is relatively small: 0.68-0.23=0.45 year.

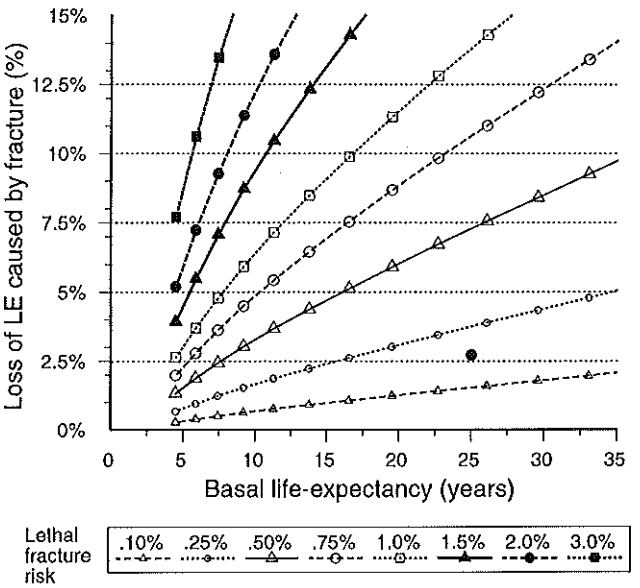

Figure 1 Loss of life-expectancy caused by fracture in relation to basal life-expectancy and lethal fracture risk. The lethal fracture risk is calculated by multiplying the fracture risk (e.g. from Table 1) and the lethality of fracture (.51 for mitral valves and .86 for aortic valves). The loss of life-expectancy caused by fracture may directly be compared with surgical mortality (e.g. from Table 2). Hypothetical patient (see text, •)

Figure 1 can be used to evaluate one single patient at a time. To evaluate groups of patients, we present formulas in the Appendix which can easily be implemented in a computer application. For example, we used a spreadsheet program to assess the consequences of revisions of fracture risk for the BScc patients in our centre. We evaluated the patients with actual age, gender and sex-specific weight, but optimistically assuming no comorbidity. Subsequently, we performed a more detailed and individualized examination of those patients with a calculated benefit of replacement.

11.4 Discussion

In this paper we present an easy-to-use tool to calculate the life-expectancy of replacement or observation for patients with a BScc valve. This tool presents the output of a previously developed model[9] in such a way that newly available fracture risk estimates and surgical risk estimates can easily be included in decision making in individual patients as well as in groups of patients.

Selection of candidates for replacement can be based on the estimated fracture risks, with a closer examination of patients with valves of relatively high risk[5]. The impact of a certain fracture risk depends however on individual patient characteristics, especially

age, as a higher age both increases surgical risk and diminishes life-expectancy. For example, Blackstone at al.[13] considered two hypothetical female patients of 38 and 67 years old, without comorbidity, with a fracture risk of 2% in a BScc mitral valve and a lethality of fracture of 50%. Their analysis indicated that the advantage of surgery in the 67 old patient was minimal, and that the advantage in the 38 year old patient was somewhat larger. Our decision tool confirms these findings qualitatively, but indicates that the magnitude of the advantage of replacement in the 38 year old female patient (assuming NYHA class 1, first reoperation, 60 kg) is as large as 4.3 years, which strongly supports replacement. We propose that the selection of patients for replacement should use this expected benefit as the starting point. This benefit can easily be calculated with our decision tool, either by hand (Figure 1) or computerized (formulas in Appendix).

Considerable uncertainty may be present about the true values of estimates required for decision making in BScc valves. The advantage of our tool is that the impact of variations in estimates can directly be explored. If, for example, the estimated fracture risk is varied to the extremes of a wide plausible range, the loss of life-expectancy corresponding to these extremes is directly available.

Uncertainty in the estimates of fracture risks is firstly caused by underreporting, which leads to a systematic bias both in the reported fracture risks and the lethality of fracture. Underreporting may especially be a problem in aortic BScc valves, as patients suffering from strut fracture usually die within 2 hours[3], without distinct symptoms of mechanical failure. Secondly, the estimates of strut fracture risks are uncertain because of the limited number of fractures available for multivariate statistical analysis. Estimates of surgical mortality were taken from a large recently published series[10], but may vary because of centre-specific circumstances or the presence of risk factors not considered in the model. Next, basal life-expectancy may be estimated lower than the figures in Table 2[3], e.g. because of the presence of risk factors not considered[12]. Further, the lethality of fracture may vary because of the patient's age, clinical condition, and feasibility and time to reach medical facilities for urgent surgery. The effect of these uncertainties can directly be quantified by our tool.

This analysis did not consider decision making in patients undergoing bypass surgery or in patients with aortic as well as mitral BScc valves, because no reliable data on basal life-expectancy and surgical mortality were available for these types of patients. The approach is however identical to the approach followed in patients who are considered for elective replacement of one BScc valve, without concomitant bypass surgery. Again the surgical risk has to be weighed against the cumulative fracture risk. In patients undergoing bypass surgery, surgical mortality refers to the additional risk of valve surgery compared to bypass surgery alone and basal life-expectancy refers to the life-expectancy after successful bypass surgery. In patients with two BScc valves, surgical mortality of replacement of the aortic, the mitral, or both valves has to be weighed against the cumulative fracture risk of the aortic, the mitral or both valves respectively.

Decision making in BScc valves is complex. The problem has been labelled 'tough'[13], since the risks of strut fracture are relatively low, while a major heart reoperation is required. In the future, the choice between prophylactic surgery and observation may

be extended with the option to screen patients for defects in mitral BScc valves[14]. Selection of patients for such radiographic screening might be helped by our tool to assess the impact on life-expectancy of a certain fracture risk. Besides life-expectancy, other aspects may be considered in the decision making process. For example, neurological deficits may remain after reoperation (on average 1.1% in a recent analysis, correlating with estimated surgical risk[10]). Such permanent morbidity may however remain as well after strut fracture. Also, time-preference may play a role. Most patients are risk averse and attach more value to nearby years than years in the distant future. This implies that replacement, which causes a short term risk, would be less attractive. Finally, the patient's personal preferences will influence in the decision making process. In this context, we expect that our tool can serve as a first step in the decision making process by supplying information about the expected benefit or harm of prophylactic replacement.

Appendix

For computerized calculation of the relative loss of life-expectancy with surgery or observation we derived the following formulas. The formulas for basal life-expectancy and for the fracture caused loss of life-expectancy use regression analysis techniques to create a 'meta-model' of original Markov model.

- Basal life-expectancy was calculated with the Markov model for male and female patients with mitral or aortic valves, not undergoing bypass surgery. We varied the patients age between 25 and 90 years, with steps of 5 year (Decision Maker software, version 7.0, New England Medical Centre, 1988). Linear regression analysis was subsequently used to estimate the basal life-expectancy for aortic and mitral valve patients separately and as a function of age and gender (SPSS/PC+ statistical package, version 5.0.1). The regression models explained a very high proportion of the variance (r^2>.99), which indicated that the formulas closely describe the original estimates of the basal life-expectancy as calculated with the Markov model.
- Surgical mortality was estimated with the logistic regression formula from the paper by Piehler et al[10], where also data are presented to estimate the confidence limits around the estimated surgical mortality.
- The fracture caused loss of life-expectancy (see Figure 1) was calculated for patients without comorbidity, with age between 25 and 90 years (steps of 5 year), average mortality of male and female patients, and average lethality of fracture (75%). Linear regression analysis was used to estimate the loss due to strut fracture as a function of the combination of the lethal fracture risk and basal life-expectancy. The regression model explained over 99% of the variance, indicating that the formula closely describes the curves in Figure 1.

Using these formulas, the life-expectancies with surgery and with observation can be calculated. A spreadsheet program is available from the authors (E-mail: steyerberg @ckb.fgg.eur.nl), which includes the formulas presented below.

- Basal life-expectancy aortic valve patients =
  66.66 - 1.069 • Age + 10.522 • Age2yrs + 5.31 • Female - .04903 • Female • Age
  Basal life-expectancy mitral valve patients =
  57.65 - 1.002 • Age + 11.038 • Age2yrs + 5.10 • Female - .04880 • Female • Age
- Surgical mortality risk = $e^{(PI)} / [1+e^{(PI)}]$, where PI =
  -6.444 + .5744 • NYHA + .5647 • Hdstate + .6427 • Age2yrs + .5270 • InvWgtKg + .8859 • Renal + .5088 • Double + .8647 • TVIncomp + 1.1512 • PVE + .3311 • OpenNumb + 1.4985 • AAA + 1.9928 • LVA + .6005 • CABG [10]
- % fracture caused loss of LE =
  .5134 • LEbasal • LethFrac + .1591 • LEbasal • Sqrt(LethFrac) - $9.59 \cdot 10^{-3}$ • $LEbasal^2$ • LethFract + $5.69 \cdot 10^{-3}$ • $LEbasal^2$ • Sqrt(LethFract)
- LEsurgery = LEbasal • (1 - Surgical mortality risk)
  LEobservation = LEbasal • (100 - % fracture caused loss of LE)/100
  Advantage of surgery = LEsurgery - LEobservation

where Age is expressed in years; Age2yrs is [patient's age (years)/50]^2; Female is 1 if female, 0 if male; NYHA is NYHA class in numerical terms (1 through 5); Hdstate is hemodynamic state (0=stable, 1=unstable, 4=cardiogenic shock); InvWgtKg is 70/[patient's weight (kg)]; Renal is 1 in case of chronic renal failure or creatinine > 2.5 mg/dl, 0 if not; Double is 1 in case of multiple valve disease, 0 if not; TVIncomp is 1 in case of present or previous tricuspid valve incompetence, requiring intervention, 0 if not; PVE is 1 in case of active prosthetic valve endocarditis, 0 if not; OpenNumb is number of previous open heart operations (1 for first reoperation); AAA is 1 in case of repair of an ascending aortic aneurysm, 0 if not; LVA is 1 in case of resection of a left ventricular aneurysm, 0 if not; CABG is 1 in case of coronary artery bypass grafting [10]; Mitral is 1 if mitral valve, 0 if aortic; poorLV is 1 if left ventricular function (as classified from the right oblique view of the ventricular angiogram) is poor, 0 if good or reduced; LEbasal is expressed in years; LethFract is the lethality of fracture (.51 for mitral valves and .86 for aortic valves) multiplied by the annual strut fracture risk, expressed as a percentage.

References

1. Lindblom D, Björk VO, Semb BKH. Mechanical failure of the Björk-Shiley valve. J Thorac Cardiovasc Surg 1986;92:894-907.
2. Lindblom D, Rodriguez L, Björk VO. Mechanical failure of the Björk-Shiley valve. J Thorac Cardiovasc Surg 1989;97:95-97.
3. Van der Graaf Y, De Waard F, Van Herwerden LA, Defauw JJAMT. Risk of strut fracture of Björk-Shiley valves. Lancet 1992;339:257-261.
4. Ericsson A, Lindblom D, Semb G, Huysmans HA, Thulin LI, Scully HE, Bennett JG, Ostermeyer J, Grunkemeier GL. Strut fracture with Björk-Shiley 70° convexo-concave valve. Eur J Cardiothorac Surg 1992;6:339-346.
5. Pfizer/Shiley Heart Valve Research Centre. Dear Doctor letter. August, 1994.
6. Pfizer/Shiley Heart Valve Research Centre. Dear Doctor letter. January, 1995.
7. De Mol BA, Kallewaard M, McLellan RB, Van Herwerden LA, Defauw JJAMT, Van der Graaf Y. Single-leg strut fracture in explanted Björk-Shiley valves. Lancet 1994;343:9-12.
8. Birkmeyer JD, Marrin CAS, O'Connor GT. Should patients with Björk-Shiley valves undergo prophylactic replacement? Lancet 1992;340:520-523.
9. Van der Meulen JFP, Steyerberg EW, Van der Graaf Y, Van Herwerden LA, Verbaan CJ, Defauw JJAMT, Habbema JDF. Age thresholds for prophylatic replacement of Björk-Shiley convexo-concave heart valves: A clinical and economic evaluation. Circulation 1993;88:156-164.
10. Piehler JM, Blackstone EH, Bailey KR, Sullivan ME, Pluth JR, Weiss NS, et al. Reoperation on prosthetic heart valves: patient-specific estimates of in-hospital events. J Thorac Cardiovasc Surg 1995;109:30-48.
11. Beck JR, Pauker SG. The Markov process in medical prognosis. Med Decis Making 1983;3: 419-458.
12. McGiffin DC, O'Brien MF, Galbraith AJ, McLachlan GJ, Stafford EG, Gardner MAH, et al. An analysis of risk factors for death and mode-specific death after aortic valve replacement with allograft, xenograft, and mechanical valves. J Thorac Cardiovasc Surg 1993;106:895-911.
13. Blackstone EH, Kirklin JW. Recommendations for prophylactic removal of heart valve prostheses. J Heart Valve Dis 1992;1:3-14.
14. O'Neill WW, Chandler JG, Gordon RE, Bakalyar DM, Abolfathi AH, Castellani MD, et al. Radiographic detection of strut separations in Björk-Shiley convexo-concave mitral valves. N Engl J Med 1995;333:414-419.

General discussion

12 Prognostic modeling for clinical decision making: discussion

This thesis describes theoretical aspects of prognostic modeling for clinical decision making as well as several practical applications. In this chapter the theoretical aspects are recapitulated. Subsequently, the practical applications will be reviewed in the light of the theoretically desired properties. Other issues addressed are the evaluation of some of the prognostic models as clinical prediction rules, topics for further research, and the impact of prognostic models on clinical decision making.

12.1 Theoretical aspects of prognostic modeling

Prognostic modeling is a complex issue, of which some theoretical aspects were addressed. Especially problematic is the selection of variables to be used as predictors in a regression equation. It was shown in chapter 2 that the use of standard stepwise selection methods carries a high risk of leading to unreliable models in small data sets with many potentially predictive variables: the regression coefficients are imprecise and, more important, they are biased to larger values. Alternative selection strategies were discussed, including the use of a higher p-value in stepwise selection (e.g. 50%), selection of variables on the basis of the literature (preferably in a formal review or 'meta-analysis'), and the clustering of related variables in groups. It may be concluded that variable selection should be based as limited as possible on the regression results obtained in a relatively small data set.

A second point of interest is the estimation of the regression coefficients in a fixed set of predictors. The statistical fitting procedures lead to a slight overestimation of the regression coefficients. This overoptimism can be corrected with a shrinkage factor. In chapter 3, a new method is described to obtain more precise estimates by combining literature data and individual patient data.

It was noted that the selection bias and the estimation bias in the regression coefficients are especially relevant in small data sets. In large data sets these biases are much less a problem.

Model performance is usually distinguished in reliability and discriminative ability. Reliability or 'goodness-of-fit' is determined by the correctness of the model specification (fulfillment of assumptions) and by the values of the regression coefficients. Discriminative ability was shown to be less sensitive to the values of the regression coefficients. Discrimination was relatively poor if a limited number of variables was selected by standard stepwise selection methods. When discrimination is the main objective of the model, the selection of predictors thus merits primarily attention and the exact values of the regression coefficients are less important.

Further, validation was discussed, which relates to an unbiased assessment of modeling steps and model performance. Several approaches were reviewed, including

the split-sample approach, cross-validation and the bootstrap re-sampling method. It was noted that these methods generally address only internal validity (performance within the same patient population), in contrast to external validity (performance in slightly different patients, e.g. other centers, other interpretations of predictive characteristics).

Finally, the presentation of prognostic models deserves attention, especially if the prognostic models are intended as tools for busy clinicians. Suggested presentations include the construction of a table or score chart to facilitate practical application.

12.2 Applications and theory

In this section, the prognostic modeling strategies as applied in the practical applications will be reviewed in the light of the theoretical desired properties. Some discrepancies between theory and practice are described. These are partly explained by the specific objectives of the prognostic model. Further, the development of a prognostic model is a team effort, with input from both methodologists and clinicians. The merits of a prognostic model should therefore not be judged by methodological standards alone.

12.2.1 Prognosis after resection in testicular cancer

The first application of a prognostic model concerned the prediction of relapse of malignant disease after resection of a residual mass in testicular cancer patients. Standard forward stepwise variable selection was used to identify the most important predictors for relapse. A prognostic classification was subsequently proposed, using only two of the three selected predictors (HCG level and completeness of resection, but not a variable related to the presence of residual lung metastases). It may be anticipated that the selection strategy has led to overestimated regression coefficients, the identification of a random variable as an important predictor, and to the exclusion of some important predictors. Previous research had however often shown that the first two of the selected variables were prognostically important. The clinical implication of the analysis is that additional chemotherapy is indicated in those patients with incomplete resection. This predictor was highly significant (p=.004, *see section 4.3.4*). Moreover, later publications have confirmed the importance of incomplete resection as a predictor of relapse[1,2].

12.2.2 Prediction of residual mass histology in testicular cancer

The second research topic in testicular cancer concerned the prediction of residual mass histology. The histology can be distinguished in three groups: purely benign (necrosis/fibrosis), mature teratoma (potentially malignant tissue) and cancer (viable cancer cells). This is essentially a polychotomous outcome. The histology was analyzed with two outcome definitions: one distinguishing necrosis from other tissue, and a second distinguishing cancer from teratoma. Alternatively, the histology might have been analyzed as mature teratoma versus necrosis and cancer versus necrosis, for example with polytomous logistic regression.

Several advantages can be formulated for the chosen type of analysis. The distinction of necrosis versus other tissue is clinically the most relevant, since both mature teratoma

and cancer should preferably be resected while necrosis should not. The regression results for the distinction of necrosis from other histology are directly interpretable. Next, the presentation of a polytomous model for the histological probabilities would have been more complex than the presentation of two models. Further, the variables used to distinguish between the histologies need not be identical in both models. Three predictors with a trivial association with the ratio of cancer versus mature teratoma ($p>.50$) were excluded from the second, while they were included in the first model.

A disadvantage of the second model is that it is not readily interpretable, since the probability of cancer is calculated conditional on the probability of necrosis. This disadvantage would have been overcome by a direct calculation of the absolute probability of cancer. This could be achieved by defining the outcome as cancer or other histology. If regression models were constructed for cancer versus other histology and necrosis versus other histology, we found, unfortunately, that some combinations of predictors led to a sum of probabilities larger than 1. Therefore, the chosen classification may be judged an optimal combination of mathematical correctness and clinical interpretability.

The steps taken in the modeling process for the retroperitoneal residual mass histology come close to the theoretical ideal. First a meta-analysis was performed on the published literature to explore the value of potential predictors of the histology at resection. Since several strong predictors could be identified, cooperation with other centers was sought to obtain a large data set with individual patient data, where reliable multivariate prognostic models could be developed. The regression coefficients were only based on the latter data base, because inclusion of patients in this analysis could better be controlled than in the meta-analysis. The meta-analysis for example included some patients with elevated tumour markers at resection, or patients with pure seminoma. The multivariate analysis used a p-value of 50% for inclusion of variables. For the model predicting necrosis at retroperitoneal resection, the chosen p-value would not have affected the selection of variables, as all variables had p-values less than 1% (*section 6.3.3*). For the model distinguishing cancer from teratoma, the standard p-value would have led to inclusion of one predictor only (LDH, $p=.02$). In the prediction of the histology at pulmonary resection, several variables would have been excluded at the 5% level. Two strong interaction terms were found and included in the prognostic model for necrosis at pulmonary resection. Bootstrapping techniques were used to correct the regression coefficients for overoptimism, and to estimate model performance in future patients. External validity of the model for the histology at retroperitoneal resection was assessed by leaving each participating center out once. For the thoracotomy model, external validity could not be assessed since most centers contained a too limited number of patients. The final prognostic models were presented as a prognostic score chart (retroperitoneal resection) or table (pulmonary resection) to facilitate practical application.

The prognostic models for the residual mass histology are intended to improve the selection of patients for surgery. It was shown in chapter 7 that selection for retroperitoneal resection would considerably improve if the model were applied

compared to several current selection policies. If the prognostic model was simplified by categorizing the predictors, the performance was still substantially better than most current policies. These findings were theoretically expected, since discriminative ability of a prognostic model depends largely on the predictors included and less on the specific regression coefficients of the predictors (*section 2.4.4*).

12.2.3 Mortality of elective aortic aneurysm surgery

The third problem concerned the prediction of surgical mortality in elective abdominal aortic aneurysm surgery. The method applied was described in chapter 3. Literature data and individual patient data were combined, which resulted in substantially smaller confidence intervals around the regression coefficients than when only individual patient data were used. A major challenge was the estimation of the constant or intercept in the prognostic model, which involved calculations with likelihood ratios. For application in clinical practice, it may as a standard be assumed that the center-specific surgical mortality is 5%. In a highly specialized center, the risk may be lower, e.g. 3%, while a non-specialized setting may be associated with a higher risk, e.g. 8%. A proper assessment of the center-specific risk is difficult. In the presence of empirical data, this risk may be calculated while correcting for the prevalence of prognostic factors in the patient population in the center.

The performance of the model has not yet been evaluated and merits attention. It is presumed that the presented model, which incorporates literature data, will perform better than the multivariate regression model based solely on the individual patient data.

12.2.4 Replacement of mechanical heart valves

The fourth problem concerned decision making in patients with a Björk-Shiley convexo-concave heart valve, a mechanical valve with increased risk of failure. Three prognostic models were developed: one for the risk of failure, a second for the survival of patients with a mechanical valve, and a third for surgical mortality. The focus of the first model was to estimate the risk of failure as accurately as possible. The stepwise selection strategy followed led to inclusion of all plausible predictors. The prognostic value of one of these variables (age at valve implantation) has not been reconfirmed in other analyses. The focus of the other two models was to derive estimates for patients with a good risk profile, which means for patients without major comorbidity. The estimates were used in a decision analysis model to calculated age-thresholds, below which prophylactic surgery might be contemplated, and above which surgery would lead to a lower life-expectancy than an expectant management. If more risk factors had been used in the estimation of survival and surgical mortality, the thresholds would have been somewhat lower.

In recent years, many more models have been developed for the risk of mechanical failure, incorporating more characteristics of the valve than previously available, such as production date and welder information[3,4]. The modeling strategies applied are however dubious. The categorization of variables seems severely based on the observed failure rates, as well as the inclusion of numerous interaction terms. The risk of reoperation was

recently analyzed in a large multicenter study, which may provide better estimates for the decision analysis than previously used[5].

Based on the essence of the decision analysis, we developed a graphical tool that allows for an easy incorporation of the new prognostic estimates (chapter 11). We also used regression analysis to derive a meta-model of the decision analysis. This model has the special characteristic that the variables influencing the model outcome are known a priori. For example, it could analytically be derived that the model should have no intercept. Variable selection aspects hence only related to non-linearity and additivity (inclusion of interaction terms).

12.3 Evaluation of clinical prediction rules

Several prognostic models were presented as clinical prediction rules (chapter 6: prediction of retroperitoneal mass histology; chapter 8, prediction of lung mass histology; chapter 9, prediction of mortality from elective aneurysm surgery). Methodological standards have been described before that can be used to evaluate these prediction rules[6]. The results are shown schematically in Table 1.

Table 1 Evaluation according to quality criteria for clinical prediction rules*.

Methodological standard	Retroperitoneal mass histology	Lung mass histology	Aneurysm operative mortality
Definition of outcome	+	+	+
Definition of predictors	+	+	+
Patient characteristics	+	+	+
Study site described	+	+	+
Accuracy, e.g. misclassification rate	+/-	+/-	-
Effects on patient care prospectively measured	+/-	-	-
Mathematical technique described	+	+	+

* +: standard was fulfilled; -: standard was not yet evaluated; +/-: preliminary evaluated

The first methodological standards relate to a clear definition of the outcome and the predictors. All rules fulfilled these criteria. Further, the applicability of the prediction rule may depend on the patient population used to derive the rule. This may be read from patient characteristics like age and sex. For the testicular cancer prediction studies, all patients were (obviously) male. Age was not described in these papers, since age has never been found as an important predictor of residual mass histology. Age and other characteristics like the specific type of chemotherapy were described in the original publications from the participating centers. It was further proposed that the accuracy of the prediction rule should be determined[6]. In the prediction rules for testicular cancer, the area under the ROC curve was used as the measure for discriminative ability. For the model predicting aneurysm mortality, discriminative ability could not be reported

for the final model. Next, it was stated that model performance should preferably be determined in an independent data set[6]. This has not yet been performed for any of the prediction rules. Neither have the rules been evaluated prospectively for their effects on patient care. A retrospective comparison was however performed (chapter 7) between the retroperitoneal mass histology model and current selection guidelines, which showed that the prediction rule would improve patient care considerably. Finally, the mathematical technique should be described. All prediction rules fulfilled this criterium.

12.4 Further research

Several methodological aspects of prognostic modeling require further study. Further evaluation is required of the adaptation method to combine literature data and individual patient data. Evaluation might be performed in a large multi-institutional data set with individual patient data available. Criteria for performance might include the estimates of the coefficients, and the discriminative ability. The technique was described here for logistic regression analysis, but application might probably be easily extended to survival analysis models like Cox or Poisson regression.

Further research is also required on variable selection strategies. A comparison between strategies might include bootstrapping as a method for variable selection, while using stepwise selection. The frequency of obtaining a different selection of variables than with the original selection strategy is anticipated to be rather low, except when the choice of the last selected variables is rather arbitrary. Other variable selection strategies will probably appear preferable, such as variable clustering or selection on the basis of a literature review[7]. The preference for a strategy should be investigated in relation to the sample size.

Further research on the prognostic models presented in the applications should focus on external validation. External validation includes evaluation of the models on patients not included in the modeling phase, preferably also from other centers than those participating in the development phase. We may expect two problems with the transfer of the prognostic models[8]. First, the prevalence of the outcome may be different in another setting, even after correction for the prognostic factors. This means that other variables, which are not considered among the prognostic factors, are relevant. In the prediction of surgery from elective aneurysm surgery, such variables may include the degree of specialization of the hospital and the individual surgeon. The second problem may relate to the definition and scaling of the predictors. For example, normal values for laboratory test results should match before the prognostic model is applied.

Patient groups for external validation may retrospectively be identified from existing databases, or from prospective research. In the case of testicular cancer, a validation study for the retroperitoneal histology has been set up. This study will include recently operated patients, who were not included in the previous analyses. Initial results are very favorable. Prospective validation will take place with patients included in a recently started randomised trial (protocol MRC-EORTC 30941). Over 700 patients will be included in this trial, of which around 200 are expected to be operated on.

For the prediction of mortality from aortic aneurysm surgery, retrospective evaluation might be performed in patients operated on in the Rotterdam Dijkzigt hospital during recent years. Also, a randomised trial has started which may allow for prospective validation.

12.5 Impact on clinical decision making

It can be concluded from the applications that prognostic models may provide information to clinicians on treatment decisions in a wide variety of situations. This information is meant to support the clinician in the decision making process. It may not be expected that decision making is based solely on the prognostic models. First, the models may not include all relevant prognostic characteristics. For example, a very infrequent condition may influence decision making strongly, while such a characteristic may not be included in the prognostic model. Further, reliability of the model may be insufficient, as discussed before. Nevertheless, the support of the treating clinician by a prognostic model is expected to result in better decision making than is possible without this support.

The prognostic models may also be useful in the communication between a clinician and his/her patient, and the attitudes and behavior of the patient. For example, knowledge of a good or poor prognosis after resection may influence short-term decisions of a testicular cancer patient like plans for holidays, or long-term personal decisions like marriage and procreation. A high probability of totally benign tissue after chemotherapy may be sufficient for a testicular cancer patient to refrain from surgery, while a high risk of malignancy may convince a reluctant patient that resection is necessary. Other examples can be imagined in aortic aneurysms, where an 80-year-old patient with a small aneurysm and comorbidity might initially like to be operated on to get rid of the risk of rupture of the aneurysm. After being told that the surgery would have a considerable risk, he would probably reconsider his preference. In contrast, a healthy 70-year-old patient might be reassured by knowing that surgery in his case is expected to bear a risk of around 2%. Finally, patients with a Björk-Shiley convexo-concave valve may feel reassured once told that the cumulative risk of mechanical failure is very low and does not warrant prophylactic surgery.

We might wonder whether the developed prognostic models have indeed supported decision making by clinicians. In general, it is known that conclusive clinical trials have a more direct impact than decision analyses or epidemiological studies. The decision analysis on elective replacement of Björk-Shiley convexo-concave valves certainly did have an impact on decision making at the Thoraxcenter Rotterdam. Patients were called back for further physical examination and discussion on the basis of the model results. The model was used again with inclusion of the updated surgical risk estimates from the physical examination. Also, more detailed evaluations, especially sensitivity analyses, were performed. Of course, the model results often agreed with the clinical impression regarding the benefit of replacement. Definitive decision making was left to the treating physicians (cardiologist and surgeon) and the patient.

The direct impact of the testicular cancer and aneurysm models is not known, but is probably smaller. In the case of testicular cancer, the suggestion of administering additional chemotherapy to patients with an incomplete resection will probably be followed in the future, especially if clinical studies keep confirming the poor prognosis after incomplete resection. The models for retroperitoneal and lung histology should lead to a revision of the resection policies in many centers. Demonstration of validity in new patients, and in other centers may promote more widespread application of these prognostic models in clinical practice. We are also working on a decision analysis concerning the question which small residual masses (≤ 20mm) should be resected and which might safely be observed.

The prediction model for elective surgical mortality of aneurysm surgery was intended to contribute to decision making on surgery in patients with small aneurysms, in patients with a high surgical risk, and in patients with a limited life-expectancy. For this purpose, the risk estimates have been incorporated in a decision analysis model developed at the Leiden University Hospital[9]. This model will be used in the near future to advise on the advantage of elective surgery in individual patients, and may thus directly affect decision making. Like in the case of replacement of a risky heart valve, final decision making is left to the treating physician and the patient.

In conclusion, the developed models have probably contributed to prognostic knowledge in the diseases involved, which is useful for both clinicians and patients. The prognostic models may have supported decision making in some cases directly. Further proof of the validity may enhance more widespread use of prognostic models, which will improve clinical care.

References

1. Hendry WF, A'Hern RP, Hetherington JW, Peckham MJ, Dearnaley DP, Horwich A. Para-aortic lymphadenectomy after chemotherapy for metastatic non-seminomatous germ cell tumours: prognostic value and therapeutic benefit. Br J Urol 1993; 71: 208-213
2. Gerl A, Clemm C, Schmeller N, et al. Outcome analysis after post-chemotherapy surgery in patients with non-seminomatous germ cell tumours. Ann Oncol 1995; 6: 483-488
3. Pfizer/Shiley Heart Valve Research Centre. Dear Doctor letter. August, 1994
4. Pfizer/Shiley Heart Valve Research Centre. Dear Doctor letter. January, 1995
5. Piehler JM, Blackstone EH, Bailey KR, et al. Reoperation on prosthetic heart valves: patient-specific estimates of in-hospital events. J Thorac Cardiovasc Surg 1995; 109: 30-48
6. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules: applications and methodological standards. N Engl J Med 1985; 313: 793-9
7. Harrell FE, Lee K, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984; 3: 143-152
8. Segaar RW, Wilson JHP, Habbema JDF, Malchow-Møller A, Hilden J, Maas PJ van der. Transferring a diagnostic decision aid for jaundice. Neth J Med 1988; 33: 5-15
9. Kievit J, Kitslaar PJEHM.Decision making in asymptomatic abdominal aneurysm: who should be operated upon? Med Decis Making 1991; 11: 332 (abstract)

Appendices

Summary

The determination of the prognosis for a patient is central to clinical decision making, since all diagnostic or therapeutic interventions eventually aim to influence the patient's disease process in a favorable way. Prognosis may refer to several medical outcomes, like mortality, complications, or complete recovery from disease.

Estimates or predictions of prognosis can be based on statistical models (*Chapter 1*). These models often use the technique of regression analysis to predict the outcome of the disease process on the basis of a number of patient characteristics. Such a regression analysis can be performed on a data set where, for each patient, a number of potential predictive characteristics (predictors) is registered in combination with the outcome of the disease. A regression coefficient which quantifies the prognostic value can be estimated for each predictor. Several theoretical aspects of this prognostic modeling process are discussed in this thesis, as well as a number of clinical applications.

I Theory

Theoretical aspects of prognostic modeling are discussed in *Chapter 2* of this thesis, starting with the basic assumptions underlying regression analysis. These assumptions include the mathematical form of the relationship of the predictors with the outcome, linearity of continuous predictors, and additivity of the predictors in the prognostic model. Attention is given to 'overoptimism', which is here defined as the phenomenon that a prognostic model tends to perform better on the data set which was used to derive the model than on a new data set. Causes of overoptimism may be found in three modeling phases: selection, estimation and evaluation.

The first phase concerns the selection of predictors for the prognostic model. It is shown that standard stepwise selection methods have a high risk of leading to overoptimism. The main problem is that the regression coefficients of the selected predictors are biased to more extreme values, especially in relatively small data sets with a large number of potential predictors. Alternative selection strategies are discussed; these try to limit the dependence of the selection of predictors on the regression results in the data set under study. For example, the predictors may be selected on the basis of a systematic literature review ('meta-analysis'). The second modeling phase may also contribute to overoptimism. Even when a fixed set of predictors is used in a regression analysis, the coefficients are biased to more extreme values. This bias is however neglectable in large data sets with a limited number of predictors. The third phase, the evaluation of model performance, may show overoptimistic results with regard to the discriminative ability and calibration of the model predictions.

Several validation procedures, which aim to limit the biases in these three modeling phases (selection, estimation, evaluation), are discussed. Re-sampling techniques like the bootstrap procedure may be helpful, especially to limit the overoptimism in phases two and three, and are therefore advocated for routine use in prognostic modeling. Finally,

alternative presentations of a prognostic model are suggested, such as the construction of a table or score chart to facilitate the clinical application of the prognostic model.

In *Chapter 3*, a new method for prognostic modeling is presented. This entails logistic regression analysis, in the situation that a data set with individual patient data is available together with one or more comparable patient series in the literature. Multivariate regression coefficients can be estimated in the individual patient data, in contrast to the literature data, where only univariate coefficients can be estimated. Following this method, predictors are selected with a formal meta-analysis of the literature data combined with the own data. The univariate regression coefficients for single predictors can be estimated more reliably with this meta-analysis than from the data set under study alone, since a larger number of patients is analyzed. The univariate regression coefficients from the literature are usually ignored for further modeling in the data set under study. It was however shown that it is sensible to combine the regression results from the literature with those from the data set under study, since this combination generally leads to more precise estimates than obtained without the literature data. Statistical requirements for the data set under study were formulated. These requirements can be checked empirically with re-sampling methods like the bootstrap procedure.

II Applications

The clinical applications in this thesis relate to three clinical decision problems: patients with testicular cancer, patients with an aortic aneurysm, and patients with an artificial valve which has an increased risk of mechanical failure.

Residual masses in testicular cancer

The first application concerns patients with metastatic (non-seminomatous) testicular cancer. After hemi-orchidectomy of the testicle with the primary tumor, these patients are usually treated with several courses of chemotherapy. After chemotherapy, remnants of the metastases may still be present. These so-called residual masses can effectively be removed by a surgical resection. In *Chapter 4*, we analyze the long-term prognosis after such a resection, which appears very good: 87% of the patients was free of relapse after 5 years. A simple prognostic classification was proposed, using two predictors (level of the tumormarker HCG and completeness of resection). Finally, it was suggested that patients with an incomplete resection benefit from additional chemotherapy.

Chapters 5 to 8 address the prediction of the histology of residual masses. Three histologies can be distinguished: totally benign tissue (necrosis/fibrosis), for which resection implies no therapeutic benefit; potentially malignant tissue (mature teratoma), which should be removed since it may grow and become harder to resect; and residual malignancy (viable cancer cells), which should be removed since a) growth and further metastasizing are prevented, and b) further therapy can be given which treats remaining malignancy at other sites. *Chapter 5* contains a meta-analysis of predictors for the residual mass histology. Several strong predictors are identified, especially for distinguishing necrosis from other histology. These predictors are subsequently used in a prognostic

model for the probability of each of the three histologies in residual masses in the abdomen (*Chapter 6*) and in the lung (*Chapter 8*). International cooperation was sought to obtain a sufficient number of patients for statistical analyses. The models performed well with regard to their ability to differentiate between patients with totally benign tissue and patients with mature teratoma or viable cancer. *Chapter 7* compares the statistical model for residual abdominal masses with several current policies. It was shown that the statistical model would lead to a better selection for resection than most current policies. If the model had been used, resection would have been avoided more often in patients with totally benign tissue, while at the same time more patients with malignancy would have undergone resection.

Elective aortic aneurysm surgery

The second clinical application concerns the prediction of surgical mortality in elective abdominal aortic aneurysm surgery (*Chapter 9*). When deciding on surgery, this risk has to be weighed against the cumulative risk of rupture of the aneurysm and its associated very high lethality. A prognostic model is developed to estimate this risk using the method presented in *Chapter 3*. Publications in the literature comprised a large number of patients (N=17000), in contrast to 238 patients in the data set under study with complete individual patient data. These two sources of data were combined to derive a presumably more reliable predictive model than a model based on the 238 patients only.

Replacement of mechanical heart valves

The third decision problem concerns patients with a mechanical heart valve with an increased risk of failure (Björk-Shiley convexo-concave valve). Acute failure may occur by fracture of the strut supporting the disk in this valve. When considering replacement, the elective surgical risk has to be weighed against the cumulative risk of failure of the valve with its associated high lethality. Three prognostic models were developed for this decision problem: one for the risk of failure, a second for the survival of patients with a mechanical valve, and a third for elective surgical mortality. The prognostic estimates were used in combination in a decision analysis model to calculate age-thresholds, below which prophylactic surgery might be contemplated, and above which surgery would lead to a lower life-expectancy than conservative treatment (*Chapter 10*).

The prognostic estimates used in a decision analysis will be subject to discussion, especially when empirical evidence arises from new studies. Therefore, the essential weighing in the decision analysis was implemented in a graphical tool that allows for an easy incorporation of alternative prognostic estimates (*Chapter 11*). Further, a 'meta-model' was derived by regression analysis on the results of the original decision analysis. This meta-model can easily be implemented in a spreadsheet programme for an efficient evaluation of a larger number of patients.

Conclusion

This thesis aims to contribute to knowledge on prognostic modeling by addressing a number of theoretical aspects and by illustrating the practical usefulness in a number of clinical problems. *Chapter 12* contains a general discussion of the theoretical results and their application in the three clinical decision problems. It is noted that several issues require further study. Theoretical issues include the proposed method for combining literature data and individual patient data, and alternative strategies for selection of predictors in a prognostic model. The developed prognostic models should be validated on new patients, also from centers not involved in the model development phase. Further proof of validity may contribute to a more widespread use of prognostic models to support clinical decision making, resulting in improved clinical care.

Samenvatting

Het bepalen van de prognose van een patiënt bekleedt een centrale positie in de klinische besliskunde, aangezien het doel van alle diagnostische of therapeutische handelingen uiteindelijk is om de prognose van een ziekteproces in gunstige zin te beïnvloeden. De prognose kan betrekking hebben op geheel verschillende medische uitkomsten, zoals korte- of lange-termijnsterfte, het optreden van complicaties, of compleet herstel.

De prognose kan worden voorspeld met behulp van statistische modellen (*hoofdstuk 1*). Deze modellen maken vaak gebruik van regressieanalyse om de uitkomst van een ziekteproces te voorspellen op basis van een aantal patiëntkenmerken. Met deze analyse worden in een dataset regressiecoëfficiënten bepaald die voor elk kenmerk de prognostische waarde weergeven. Zo ontstaat een regressievergelijking, waarmee voor nieuwe patiënten een kwantitatieve prognose kan worden gegeven. In *hoofdstuk 2 en 3* van dit proefschrift worden theoretische aspecten van dit modelleringsproces besproken. Daarna volgen toepassingen van prognostische modellen bij drie klinische beslissingsproblemen.

I Theorie

Enige theoretische aspecten van prognostische modellering worden besproken in *hoofdstuk 2*, beginnend met de aannames die ten grondslag liggen aan regressieanalyse. Deze aannames bevatten de mathematische vorm van de relatie tussen de voorspellers en de uitkomst, lineariteit van continue voorspellers, en optelbaarheid van verschillende voorspellers. Verder wordt veel aandacht gegeven aan 'overoptimisme', dat hier gedefinieerd wordt als het verschijnsel dat een prognostisch model de neiging heeft beter te voldoen in de dataset die werd gebruikt om het model te construeren dan in een nieuwe dataset. Overoptimisme kan optreden in drie fasen van modellering: selectie, schatting, en evaluatie.

De eerste fase bij modellering is het selecteren van variabelen om te gebruiken als voorspellers in het prognostische model. Het blijkt dat een veelgebruikte selectiemethode, nl. stapsgewijze selectie, substantieel bijdraagt aan overoptimisme van een prognostisch model. De regressiecoëfficiënten van de geselecteerde voorspellers worden systematisch overschat, met name in relatief kleine datasets met een groot aantal potentiële voorspellers. Daarom worden alternatieve selectieprocedures besproken, die als karakteristiek hebben te proberen de selectie van voorspellers zo min mogelijk te baseren op de resultaten in de bestudeerde dataset. De voorspellers kunnen bijvoorbeeld worden geselecteerd op basis van een systematische literatuurstudie ('meta-analyse'). Overoptimisme kan ook ontstaan in fase twee: schatting van de regressiecoëfficiënten. Het blijkt dat overschatting van de coëfficiënten zelfs optreedt wanneer selectie van voorspellers geheel onafhankelijk van de eigen dataset plaatsvindt. Deze overschatting is groter naarmate de dataset kleiner is, en het aantal voorspellers groter. Ten derde kan

bij evaluatie van het model overoptimisme optreden wanneer dezelfde dataset wordt gebruikt als waarmee het model werd geschat.

Een aantal validatieprocedures wordt besproken, dat kan worden gebruikt om het optreden van overoptimisme te voorkomen of te corrigeren. Herbemonsteringstechnieken zoals de 'bootstrap'-methode kunnen hierbij nuttig zijn, met name om overoptimisme in fase twee en drie te beperken. Deze techniek verdient daarom aanbeveling om als standaardmethode bij prognostische modellering toe te passen.

Tenslotte wordt een aantal suggesties gedaan voor de presentatie van prognostische modellen, zoals het construeren van een scoretabel (toegepast in *hoofdstuk 6 en 9*) of een tabel met voorspelde kansen (toegepast in *hoofdstuk 8*). Deze presentatie kan de klinische toepassing van een prognostisch model aanzienlijk vergemakkelijken.

In *hoofdstuk 3* wordt een nieuwe methode gepresenteerd om de gegevens uit de literatuur te combineren met eigen gegevens in een logistisch regressiemodel. Hierbij wordt uitgegaan van de situatie dat een eigen dataset ter beschikking staat met de gegevens van individuele patiënten, terwijl in de literatuur ook patiëntengroepen beschreven zijn die vergelijkbaar zijn met de groep eigen patiënten. In de eigen dataset kan een analyse worden uitgevoerd met steeds één voorspeller (univariate analyse), maar ook met meerdere voorspellers tegelijk, waarbij rekening wordt gehouden met de onderlinge afhankelijkheden tussen voorspellers (multivariate analyse). De literatuurgegevens kunnen meestal alleen gebruikt worden voor een univariate analyse. Met de hier voorgestelde methode worden voorspellers geselecteerd op basis van een formele meta-analyse, uitgevoerd op de combinatie van literatuurgegevens en eigen gegevens. Deze meta-analyse schat de univariate regressiecoëfficiënten voor elke voorspeller betrouwbaarder dan mogelijk is in de eigen dataset, aangezien grotere patiëntenaantallen worden gebruikt. Multivariate coëfficiënten worden vervolgens geschat via een eenvoudige transformatie van deze univariate coëfficiënten, nl. door aanpassing met het verschil tussen univariate naar multivariate coëfficiënt zoals waargenomen in de eigen dataset. De methode is alleen valide indien de literatuurgegevens vergelijkbaar zijn met de eigen gegevens. Bovendien moet aan bepaalde statistische voorwaarden worden voldaan, hetgeen te controleren is met herbemonsteringstechnieken, zoals de 'bootstrap'-methode. In dat geval leidt de voorgestelde methode tot nauwkeuriger schattingen van de multivariate regressiecoëfficiënten.

II Toepassingen

De klinische toepassingen in dit proefschrift hebben betrekking op drie beslissingsproblemen: het opereren van restmassa's bij de behandeling van patiënten met testiscarcinoom, het electief opereren van patiënten met een verwijding (aneurysma) van de grote lichaamsslagader (aorta), en ten derde het vervangen van een risicodragende mechanische hartklep.

Restmassa's bij testiscarcinoom
De eerste toepassing betreft patiënten met een gemetastaseerd (uitgezaaid) testiscarcinoom. Bij deze patiënten wordt eerst de zaadbal met de primaire tumor verwijderd. Vervolgens wordt de patiënt behandeld met chemotherapiekuren, waarna nog restmassa's aanwezig kunnen zijn van de metastasen. Deze restmassa's kunnen effectief worden verwijderd met een chirurgische ingreep. In *hoofdstuk 4* wordt de langetermijnprognose geanalyseerd na operatie van een restmassa. Het bleek dat de prognose zeer goed was: 87% van de patiënten was na 5 jaar nog ziektevrij. De totale groep patiënten kon worden onderscheiden in een groep met een hoog en één met een laag risico, op basis van twee kenmerken: hoogte van een tumormarker en volledigheid van de operatie. Verder werden aanwijzingen gevonden dat patiënten met een onvolledige resectie baat hebben bij additionele chemotherapie na de operatie.

Hoofdstuk 5, 6, 7 en 8 hebben betrekking op het voorspellen van de histologie in de restmassa. Drie celtypes kunnen worden onderscheiden: geheel goedaardig weefsel (necrose en/of fibrose), in potentie maligne weefsel (matuur teratoom), en tumorcellen (meestal van hetzelfde type als de primaire tumor). Voor patiënten met alleen goedaardig weefsel is een operatie onnodig: de patiënt wordt nodeloos blootgesteld aan de risico's van de operatie (ziekenhuissterfte, complicaties) en aan een extra ziekenhuisopname van ongeveer één week, met een daaropvolgend herstelproces van enkele weken tot maanden. Voor patiënten met (in potentie) maligne weefsel is opereren wel nodig, aangezien de restmassa zou kunnen gaan groeien of aanleiding zou kunnen geven tot nieuwe metastasering. Aan patiënten met tumorcellen in de restmassa wordt bovendien additionele chemotherapie gegeven, hetgeen de prognose positief beïnvloedt.

Hoofdstuk 5 geeft een systematisch overzicht van de literatuur met betrekking tot voorspellende kenmerken voor de histologie van restmassa's. Er werd een aantal sterke voorspellers gevonden, met name om geheel goedaardig weefsel te onderscheiden van (in potentie) kwaadaardig weefsel. Deze voorspellende kenmerken werden vervolgens gebruikt in een aantal prognostische modellen, waarmee de kans op elk van de drie celtypes kan worden geschat. *Hoofdstuk 6* beschrijft modellen voor restmassa's in de buik en *hoofdstuk 8* voor restmassa's in de longen. Er werd (inter)nationaal samengewerkt met een aantal ziekenhuizen om voldoende aantallen patiënten te verkrijgen voor de complexe statistische analyses. Met de modellen kon vrij goed onderscheid worden gemaakt tussen patiënten met goedaardig en kwaadaardig weefsel. In *hoofdstuk 7* worden de statistische modellen voor restmassa's in de buik vergeleken met een aantal in de huidige praktijk gehanteerde selectiestrategieën. Het bleek dat de statistische modellen leidden tot een betere selectie van patiënten die voor operatie in aanmerking komen. Dit betekent dat, indien gebruikt gemaakt zou zijn van het ontwikkelde model, er enerzijds minder patiënten met goedaardig weefsel ten onrechte zouden zijn geopereerd en er anderzijds meer patiënten met kwaadaardig weefsel zouden zijn geopereerd.

Electieve aortale aneurysmaoperatie
De tweede klinische toepassing betreft het risico op operatiesterfte bij een electieve (geplande) operatie van een aneurysma van de abdominale aorta (*hoofdstuk 9*). Bij de beslissing om te opereren of af te wachten moet dit risico worden afgewogen tegen het

cumulatieve risico van barsten van het aneurysma met een zeer hoge acute sterfte. Voor de operatiesterfte werd een prognostisch model ontwikkeld met de in *hoofdstuk 3* beschreven methode. Er was een zeer groot aantal patiënten beschreven in de literatuur (N=17000), in tegenstelling tot een relatief klein aantal patiënten (N=238) in de eigen dataset met complete gegevens op individueel niveau. Deze twee informatiebronnen werden gecombineerd om een zo betrouwbaar mogelijk model te verkrijgen voor het schatten van de mortaliteit voor individuele patiënten.

Vervangen van mechanische hartkleppen
Het derde beslissingsprobleem betreft patiënten met een mechanische hartklep met een risico op breuk (Björk-Shiley convexo-concave hartkleppen). Wanneer vervangen van deze klep overwogen wordt, moeten de risico's van electief vervangen (m.n. de operatiemortaliteit) worden afgewogen tegen het cumulatieve risico op breuk met de bijbehorende hoge acute sterfte. Voor dit probleem werd een beslissingsanalyse uitgevoerd, waarvoor drie prognostische modellen werden ontwikkeld: één voor het risico op breuk, een tweede voor de overleving van patiënten met een mechanische hartklep, en een derde voor de electieve operatiesterfte. De resultaten van de beslissingsanalyse werden gepresenteerd als leeftijdsdrempels (*hoofdstuk 10*). Bij een patiënt met een leeftijd onder deze drempel leidt vervanging van de hartklep tot een hogere levensverwachting dan een afwachtend beleid.

De prognostische schattingen zoals gebruikt in een beslissingsanalyse kunnen ter discussie staan, met name indien nieuwe gegevens beschikbaar komen. Deze nieuwe informatie moet bij voorkeur kunnen worden gebruikt in de beslissingsanalyse. Daarom werd in *hoofdstuk 11* de afweging van dit beslissingsprobleem weergegeven op een manier die het mogelijk maakt actuele, zo individueel mogelijke, prognostische schattingen direct te incorporeren. Een grafische weergave werd ontwikkeld voor toepassing op één patiënt per keer. Bovendien werd een zgn. 'meta-model' ontwikkeld door een regressieanalyse uit te voeren op de resultaten van het oorspronkelijke model. Dit meta-model kan worden geautomatiseerd (bijvoorbeeld in een 'spreadsheet'), zodat evaluatie van grotere aantallen patiënten op een efficiënte manier mogelijk is.

Conclusie

Dit proefschrift poogt bij te dragen aan kennis over prognostische modellering door een aantal theoretische aspecten te bespreken en door de praktische bruikbaarheid van prognostische modellen te illustreren aan de hand van concrete klinische beslissingsproblemen. De theoretische en praktische resultaten worden besproken in *Hoofdstuk 12*. Verder theoretisch onderzoek is gewenst naar de voorgestelde methode om literatuurgegevens te combineren met individuele patiëntgegevens, en naar alternatieve methoden voor het selecteren van voorspellers in een prognostisch model. Verder praktisch onderzoek is nodig naar de validiteit van de ontwikkelde prognostische modellen. Empirische ondersteuning van de validiteit zal zeker bijdragen aan een breder gebruik van prognostische modellen om klinische besluitvorming te ondersteunen, hetgeen zal leiden tot betere klinische zorg.

Co-authors

Name	Affiliation
D.F. Bajorin, MD, PhD	Division of Solid Tumor Oncology, Memorial Sloan Kettering Cancer Center, New York, USA
J.H. van Bockel, MD, PhD	Department of Vascular Surgery, University Hospital Leiden
G.J. Bosl, MD, PhD	Division of Solid Tumor Oncology, Memorial Sloan Kettering Cancer Center, New York, USA
J.J. Defauw, MD, PhD	Department of Thoracic Surgery, Antonius Hospital Nieuwegein
J.P. Donohue, MD, PhD	Department of Urology, Indiana University Medical Center, Indiana, USA
M.J.C. Eijkemans, MSc	Department of Public Health, Erasmus University Rotterdam
S.D. Fosså, MD, PhD	Norwegian Radium Hospital, Oslo, Norway
R.S. Foster, MD	Indiana University Medical Center, Indiana, USA
A. Gerl, MD, PhD	Klinikum Grosshadern III, University Hospital Munich, Germany
Y. van der Graaf, MD, PhD	Department of Epidemiology, University of Utrecht
C.J. van Groeningen, MD	Department of Internal Oncology, Free University Amsterdam
J.D.F. Habbema, PhD	Department of Public Health, Erasmus University Rotterdam
L.A. van Herwerden, MD, PhD	Department of Cardio-thoracic Surgery, University Hospital Dijkzigt Rotterdam
H.J. Keizer, MD, PhD	Department of Clinical Oncology, University Hospital Leiden
J. Kievit, MD, PhD	Department of General Surgery and Medical Decision Making Unit, University Hospital Leiden
J.E. Messemer, BSc	Department of Urology, Indiana University Medical Center, Indiana, USA
J.H.P. van der Meulen, MD, PhD	Department of Epidemiology and Biostatistics, University of Amsterdam

J.C.A. de Mol van Otterloo, MD, PhD	Department of Vascular Surgery, University Hospital Leiden
P.F.A. Mulders, MD, PhD	Department of Urology, University Hospital St Radboud Nijmegen
K.G. Ney, MD	Department of Urology, Indiana University Medical Center, Indiana, USA
G.L. van Rijk, MD	Department of Thoracic Surgery, University Hospital Leiden
H. Schraffordt Koops, MD, PhD	Department of Surgical Oncology, University Hospital Groningen
D.Th. Sleijfer, MD, PhD	Department of Internal Medicine, University Hospital Groningen
G. Stoter, MD, PhD	Department of Internal Oncology, University Hospital Rotterdam and Rotterdam Cancer Institute
G.C. Toner, MD, PhD	Peter MacCallum Cancer Institute, Melbourne, Australia
C.J. Verbaan, MD	Thoraxcenter, University Hospital Dijkzigt Rotterdam
J. Zwartendijk, MD	Department of Urology, University Hospital Leiden

List of publications

by March 31, 1996

First authorships

STEYERBERG EW, Keizer HJ, Zwartendijk J, Van Rijk GL, Van Groeningen CJ, Habbema JDF, Stoter G. Prognosis after resection of residual masses following chemotherapy for metastatic nonseminomatous testicular cancer: a multivariate analysis. Br J Cancer 1993; 68: 195-200

STEYERBERG EW, Keizer HJ, Fosså, Mulders PFA, Stoter G, Messemer JE, Ney K, Donohue JP, Toner GC, Bajorin D, Bosl GJ, Habbema JDF. Predictors of residual mass histology following chemotherapy for metastatic nonseminomatous GCT: univariate and multivariate meta-analysis. In: Jones WG, Harnden P, Appleyard I (eds.), Germ cell tumours III. Advances in the biosciences vol 91. Oxford, Elsevier Science, 1994: pp 239-240

STEYERBERG EW, Keizer HJ, Stoter G, Habbema JDF. Predictors of residual mass histology following chemotherapy for metastatic non-seminomatous testicular cancer: a quantitative overview of 996 resections. Eur J Cancer 1994; 30A: 1231-1239

STEYERBERG EW, Kievit J, Mol van Otterloo JCA de, Bockel JH van, Eijkemans MJC, Habbema JDF. Perioperative mortality of elective abdominal aortic aneurysm surgery: a clinical prediction rule based on literature and individual patient data. Arch Intern Med 1995; 155: 1998-2004

STEYERBERG EW, Eijkemans MJC, Habbema JDF. Strut separations in Björk-Shiley mitral valves [letter]. N Engl J Med 1995; 333: 1714-1715

STEYERBERG EW, Keizer HJ, Fosså SD, Sleijfer DT, Toner GC, Schraffordt Koops H, Mulders PFA, Messemer JE, Ney K, Donohue JP, Bajorin D, Stoter G, Bosl GJ, Habbema JDF. Prediction of residual retroperitoneal mass histology after chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups. J Clin Oncol 1995; 13: 1177-1187
Also published as:
STEYERBERG EW, Keizer HJ, Fosså SD, et al. Predizione dei rilievi istologici della massa retroperitoneale residua dopo chemioterapia per tumore metastatico a cellule germinali non seminomatoso: analisi multivariata dei dati dei singoli pazienti appartenanti a 6 gruppi di studio. Urology Digest 1996; 1: 7-9

STEYERBERG EW, Meulen JHP van der, Herwerden LA van, Habbema JDF. Prophylactic replacement of Björk-Shiley convexo-concave valves: an easy-to-use tool for decision support. Br Heart J, accepted for publication

STEYERBERG EW, Keizer HJ, Messemer JE, Toner GC, Schraffordt Koops H, Fosså SD, Gerl A, Sleijfer DT, Foster RS, Donohue JP, Bajorin DF, Bosl GJ, Habbema JDF. Residual pulmonary masses following chemotherapy for metastatic nonseminomatous germ cell tumor: prediction of histology. Submitted.

STEYERBERG EW, Keizer HJ, Fossa SD, Sleijfer DT, Bajorin DF, Donohue JP, Habbema JDF (ReHiT study group). Resection of residual masses in testicular cancer: evaluation and improvement of selection criteria. Submitted.

Co-authorships

Bonjer HJ, Lange JF, Kazemier G, Herder WW de, STEYERBERG EW, Bruining HA. Adrenal surgery: keyhole or backdoor access? a comparison of 3 techniques. Submitted

Buskens E, STEYERBERG EW, Hess J, Wladimiroff JW, Grobbee DE. Routine screening for congenital heart disease; what can be expected? A decision analytic approach. Submitted

Esch A van, STEYERBERG EW, Berger MY, Offringa M, Derksen-Lubsen G, Habbema JDF. Family history and recurrence of febrile seizures. Arch Dis Child 1994; 70: 395-399

Esch A van, Steensel-Moll HA, STEYERBERG EW, Offringa M, Habbema JDF, Derksen-Lubsen G. Antipyretic efficacy of ibuprofen and acetaminophen in children with febrile seizures. Arch Pediatr Adolesc Med 1995; 149: 632-637

Esch A van, Ramlal IR, Steensel-Moll HA van, STEYERBERG EW, Derksen-Lubsen G. Outcome after febrile status epilepticus. Dev Med Child Neurol 1996; 38: 19-24

Esch A van, STEYERBERG EW, Steensel-Moll HA van, Offringa M, Hoes AW, Habbema JDF, Derksen-Lubsen G. Efficacy of antipyretics in the prevention of febrile seizure recurrences. Submitted

Hokken RB, STEYERBERG EW, Verbaan N, Herwerden LA van, Domburg R van, Bos E. 25 years of aortic valve replacement with mechanical valves: risk factors for early and late mortality. Submitted

Keizer HJ, STEYERBERG EW. Resection of small masses of residual NSGCT and subsequent therapy: dilemmas in clinical decision making. In: Jones WG, Harnden P, Appleyard I (eds.), Germ cell tumours III. Advances in the biosciences vol 91. Oxford, Elsevier Science, 1994: p 240

Klomp HM, Spincemaille GHJJ, STEYERBERG EW, Berger MY, Habbema JDF, Urk H van (ESES study group). ESES-trial: evaluation of epidural spinal cord electric stimulation (ESES) in critical limb ischemia - a randomized controlled clinical trial. In: Horsch-S, Claes-L (eds.), Spinal cord stimulation: an innovative method in the treatment. Darmstadt, Steinkopff 1994: 173-182

Klomp HM, Spincemaille GHJJ, STEYERBERG EW, Berger MY, Habbema JDF, Urk H van (ESES study group). Design issues of a randomised controlled clinical trial on spinal cord stimulation in critical limb ischaemia. Eur J Vasc Endovasc Surg 1995; 10: 478-485

Krijnen P, Kaandorp CJE, STEYERBERG EW, Schaardenburg DJ van, Bernelot Moens HJ, Habbema JDF. Antibiotic prophylaxis for prevention of bacterial arthritis in joint disease patients: when and to whom? A decision analysis. Submitted

Kroon HM, STEYERBERG EW, Schultze-Kool LJ, Hilkens CMU, Seeley GW. Considerations in compiling a database of clinical test images. Invest Radiol 1992; 27: 255-263

Meulen JHP van der, STEYERBERG EW, Graaf van der Y, Herwerden LA van, Verbaan CJ, Defauw JJAMT, Habbema JDF. Age thresholds for prophylactic replacement of Björk-Shiley convexo-concave heart valves: a clinical and economic evaluation. Circulation 1993; 88: 156-164

Mol Van Otterloo JCA de, Van Bockel JH, STEYERBERG EW, Feuth JDM, Weeda HWH, Brand R. The potential of simple clinical information and electrocardiogram to predict mortality of primary elective abdominal aortic reconstruction. Eur J Vasc Endovasc Surg 1995; 10: 470-477

Mol van Otterloo JCA De, Bockel JH van, STEYERBERG EW, Brand R, Feuth JDM, Wall EE van der, Weeda HWH, Blokland JAK, Pauwels EKJ. Prospective risk analysis by exercise radionuclide angiography before elective abdominal aortic reconstruction. Submitted

Schouw YT van der, Graaf Y van der, STEYERBERG EW, Eijkemans MJC, Banga JD. Age at menopause as a risk factor for cardiovascular mortality. Lancet 1996; 347: 714-18

Severijnen AJ, STEYERBERG EW, Huisman J. Chlamydia trachomatis-infectie: complicaties, kosten en effecten van screening. SOA-Bulletin 1992; 13: 19-21

Severijnen AJ, STEYERBERG EW. Screening op chlamydia trachomatis: de baten zijn de kosten niet. Infectieziekten Bulletin 1993; 4: 47-49

Veelen LR van, STEYERBERG EW, Cleton FJ, Keizer HJ. Het testiscarcinoom: een witte raaf onder de kwaadaardige tumoren. Submitted

Willems TP, Herwerden LA van, STEYERBERG EW, Taams MA, Keyburg VE, Hokken RB, Roelandt JRTC, Bos E. Subcoronary implantation or aortic root replacement for human tissue valves: sufficient data to prefer either technique? Ann Thorac Surg 1995; 60: S83-S86.

Willems TP, Bogers AJJC, Cromme-Dijkhuis AH, STEYERBERG EW, Herwerden LA van, Hokken RB, Hess J, Bos E. Allograft reconstruction of the right ventricular outflow tract. Eur J Cardio-Thorac Surg, accepted for publication

Curriculum vitae

De schrijver van dit proefschrift werd geboren op 26 juli 1967. Hij bezocht het Maerlant Lyceum te 's-Gravenhage, waar hij in 1985 zijn gymnasium β diploma behaalde. Aansluitend studeerde hij aan de Rijksuniversiteit Leiden; in 1986 behaalde hij zijn propedeuse Geneeskunde. Omdat de prognose van het werken als arts hem minder aansprak dan het werken als wetenschapper stapte hij over naar de studie Gezondheidswetenschappen, later Biomedische Wetenschappen geheten. Tijdens de eindfase van deze studie was hij student-assistent bij de afdeling Medische Statistiek, waar ook zijn afstudeerstage plaatsvond. In februari 1991 behaalde hij cum laude het doctoraal examen.

Hierna verkreeg hij een aanstelling aan de Erasmus Universiteit Rotterdam bij het Centrum voor Klinische Besliskunde, tegenwoordig deel van het instituut Maatschappelijke Gezondheidszorg. Hier is hij ook nu nog werkzaam. Hij onderzocht diverse besliskundige problemen, waarover werd gerapporteerd op verschillende internationale congressen en in publicaties in medisch-wetenschappelijke tijdschriften. Sinds 1994 is hij vooral betrokken bij consultatiewerkzaamheden ten behoeve van het Academisch Ziekenhuis Rotterdam. Dit biedt hem de mogelijkheid om een bijdrage te leveren aan een breed scala van klinische onderzoeksvragen.

In zijn vrije tijd speelt hij viool in diverse gezelschappen en speelt hij hockey.

Dankwoord

Bij de totstandkoming van dit proefschrift zijn velen betrokken geweest. Dank gaat uit naar allen die op één of andere manier een bijdrage hebben geleverd. Een aantal personen wil ik hier met name noemen.

In de eerste plaats wil ik mijn promotor Dik Habbema bedanken voor de vrijheid die hij mij heeft gegund voor het uitwerken van mijns insziens interessante onderzoeksvragen. Het opbouwende commentaar ter afronding van de verschillende onderzoeken heeft in belangrijke mate bijgedragen aan het hier gepresenteerde werk. Mijn co-promotor Jan Keizer (Klinische Oncologie, AZL) zal het aantal voorlopige versies van manuscripten over testiscarcinoom wel niet meer kunnen tellen. Jan, je was steeds weer bereid om je vrij te maken van het drukke klinische werk, en je bleef altijd optimistisch over de kansen voor acceptatie van onze stukken: dank!

Verder denk ik natuurlijk aan mijn collega's bij het Centrum voor Klinische Besliskunde van het instituut voor Maatschappelijke Gezondheidszorg. In mijn beginjaren daar was met name Jan van der Meulen mij tot steun bij het 'brainstormen' over vele onderwerpen, en ook na zijn vertrek naar het AMC kon ik steeds rekenen op deze steun. Jan, dank voor het altijd weer kritisch meedenken en het formuleren van alternatieve gezichtspunten, ook in de eindfase van dit proefschrift! Veel dank ben ik verschuldigd aan mijn paranimf René Eijkemans. René, jouw geduld werd vaak op de proef gesteld bij het aanhoren van mijn wilde ideeën, maar jouw mathematische onderbouwing van een aantal analysemethoden was onmisbaar voor vele delen van dit boekje. Ook bedank ik mijn overige directe collega's Pieta Krijnen, Kees van Bezooijen, Mona Richter en Tineke Kurtz voor hun hulp en steun in het algemeen, en Paul Krabbe voor de lay-out adviezen.

De samenwerking met in de kliniek werkzame onderzoekers en artsen heb ik altijd zeer gewaardeerd. Deze contacten waren vaak inspirerend voor het uitwerken van mijn eigen onderzoeksvragen. Hierbij denk ik met name aan Tineke Willems, Raymond Hokken en Lex van Herwerden (Thoraxcentrum, AZR), Houke Klomp (ESES-trial, Heelkunde, AZR) en Arjen van Esch, Margriet van Stuijvenberg en Henriëtte van Steensel-Moll (Kindergeneeskunde, Sophia Kinderziekenhuis).

Ook buiten Rotterdam zijn velen mij behulpzaam geweest. In Leiden zijn mijn afstudeercontacten met de afdeling Heelkunde op een zeer prettige en stimulerende manier voortgezet met Alexander de Mol van Otterloo, Hayo van Bockel en Job Kievit. Ook de afdeling Medische Statistiek, in de personen van Ronald Brand en Hans van Houwelingen, was zeer behulpzaam bij het verfijnen van het theoretisch kader van dit proefschrift.

A prominent place in this thesis is taken by the results of the cooperative study on the re-analysis of residual mass histology in testicular cancer ('ReHit' study). I would therefore like to thank the various participants for their contributions.

Tenslotte bedank ik mijn familie en vrienden, die indirect hebben bijgedragen aan dit proefschrift of de juiste omstandigheden voor wetenschappelijk werk hebben gecreëerd. In de eerste plaats denk ik hierbij aan mijn vader. Lieve Wim, dank voor de vanzelfsprekendheid waarmee je mij hebt laten studeren en je niet aflatende steun en betrokkenheid. Mijn broers Maarten en Rutger en zus Iris, mijn paranimf, hebben het gelukkig nooit vreemd gevonden dat ik onderzoekertje aan het spelen was in de ivoren universiteitstoren, en hebben mij op die manier gesteund. Dit laatste geldt ook voor mijn vrouw Aleida, die van nabij het totstandkomen van dit proefschrift heeft meegemaakt. Lieve Aleida, door jouw eigen achtergrond wist jij mij vaak te inspireren en een andere kijk te geven op dit werk: dankjewel!