Ned Tijdschr Klin Chem Labgeneesk 2007; 32: 79-82

Epidemiology of tumour markers: screening, diagnosis and prognosis

E.W. STEYERBERG

Epidemiology includes two main areas of research: explanatory research and prediction research. Explanatory research includes searching for causes of diseases (etiology), and the effect of treatments (therapeutic research). An example of an etiologic question is "Does smoking cause colorectal cancer?"; a therapeutic question is "Does high-dose chemotherapy increase survival in poor prognosis testicular cancer patients?".
Tumour markers are especially relevant for prediction research, which includes questions regarding diagnosis (including early diagnosis of disease, i.e. screening), and prognosis (the outcome of a disease process). After presenting various introductory examples, some epidemiological principles are discussed, followed by challenges in the scientific development of clinically useful new tumour markers.

Tumour markers in screening, diagnosis and prognosis

With screening (early diagnosis), we aim to detect a disease early in its course, before it would manifest itself clinically by signs and symptoms of the patient. If detected earlier, we expect that the disease can be treated better, such that long-term outcomes are improved. Screening can be done in asymptomatic subjects in the population, for example, the prostate-specific antigen (PSA) level in the blood can be used to detect prostate cancer. Subjects with high PSA are evaluated further with biopsy to diagnose presence or absence of cancer. Screening can also be done after treatment of cancer to detect recurrent disease during follow-up. For example, we can test PSA levels after radical prostatectomy to monitor patients. Or we use carcinoembryonic antigen (CEA) to detect relapse of colorectal cancer.
Tumour markers may also assist in diagnosing disease, for example in the case of unknown primary tumours. Also, marker decline is a sign of response to treatment. A complete remission of disease often requires normalization of markers if these were initially elevated, in combination with other criteria, such as CT scan assessments. Other examples of diagnosing disease are the prediction of relevant cancer. In prostate cancer, we can consider small, confined cancers without poorly differentiated characteristics as probably indolent. These cancers do not require radical prostatectomy. The likelihood of an indolent cancer can be estimated by a combination of PSA levels, prostate volume, and biopsy features (1). Similarly, patients with metastatic testicular cancer often have small, residual masses after treatment with chemotherapy. If these masses are necrotic, surgical resection is not necessary. The likelihood of a necrotic residual mass can be estimated by a combination of tumour marker levels (alpha-fetoprotein, AFP, and the beta subunit of human chorionic gonadotropin, beta-HCG), CT scan measurements (pre- and postchemotherapy size), and initial histology (presence of teratoma elements) (2, 3).
For prognosis, tumour markers are often important predictors, since they reflect the extent and aggressiveness of the cancer. A good prognosis can often be based on the combination of patient, tumour, and treatment characteristics. Solely using extent of disease as with TNM staging can often be improved by considering more predictive characteristics (4). For example, the survival of testicular cancer patients can be predicted based on AFP, HCG and LDH levels, combined with extent of disease characteristics (5). Prognostic classifications can be devised, which are helpful for informing patients, decision-making on treatment, and in medical research, e.g. for stratification in randomized clinical trials.

Epidemiological aspects of screening

Screening requires a number of conditions to be fulfilled. First, we require that the tumour marker can separate subjects without from those with the cancer. This is similar for diagnostic studies. For example, a high PSA is found among patients with prostate cancer, but also among some subjects without cancer. The latter are false-positive classifications by PSA testing. A low PSA is found among most subjects without prostate cancer, but also among some with prostate cancer (false-negatives). The characteristics of a screening test can be summarized in sensitivity and specificity (table 2).
A good test has a high sensitivity and high specificity. Especially a high specificity is important in screening settings, to prevent the finding of many false-positive results among the many subjects without cancer (cell d in table 2). The trade-off between sensitivity and specificity can well be visualized in a receiver operating characteristic (ROC) curve. We plot the true-positive rate (sensitivity) against the false-positive rate (1 – specificity), for consecutive cut-points of the tumour marker (figure 1). The area under the ROC curve is a measure for the overall diagnostic performance of the test.

Correspondence: dr Ewout W. Steyerberg, epidemiologist. Department of Public Health, Erasmus MC, room AE-236, Postbus 2040, 3000 CA Rotterdam
E-mail: e.steyerberg@erasmusmc.nl

Table 1. Diagnostic and prognostic epidemiologic research with examples of tumour markers

Area	Question	Example
Screening	Can we detect the disease early in its course, such that it can be treated with better prognosis?	Does PSA testing lead to early detection and better treatment of prostate cancer? Does CEA testing lead to early detection and better treatment of recurrent colorectal cancer, after curative treatment?
Diagnosis	Can we make a diagnosis, which guides treatment choice?	Does this patient have a complete remission of his/her disease, including normalization of markers? Is the cancer relevant to treat, or is the marker profile that favorable that the disease may be treated conservatively?
Prognosis	What is the likely outcome of the disease?	Given AFP and HCG levels and other characteristics, what is the 5-year survival of this testicular cancer patient?

In screening, we moreover need to have high specificity at a time before the cancer is clinically diagnosed. A careful selection of cases and controls is hence important. Often, diagnostic qualities are less at longer time before diagnosing the cancer. See Figure 1 for an illustration of PSA and prostate cancer (6).

A further requirement of a useful screening test is that the detected disease has a better prognosis than when detected later. Lead time bias and length bias are among the major problems in the evaluation of effectiveness of treatment in screen-detected cancers.

Lead time bias refers to the apparent increase in survival time as calculated from date of diagnosis, while no true increase is caused by the earlier detection of the cancer. By screening, we intend to diagnose a cancer earlier than it would be without screening. Without screening, the disease may be discovered later once symptoms appear. Even if in both cases a person will die at the same time, because we diagnosed the disease early with screening, the survival time since diagnosis is longer with screening. No additional life years have been gained, while we may have added anxiety as the patient lives with knowledge of the disease for longer. Length bias relates to the type of cancers that are detected by screening. For many cancers, fast growing tumours have worst prognosis. Screening is more likely to catch the slower growing cancers, which have a higher survival rate.

Before a screening program is implemented, it should hence be ensured that putting it in place would do more good than harm. A cross-sectional study showing reasonable sensitivity and specificity is not enough. The best studies for assessing whether a screening test will increase a population's health are rigorous randomized controlled trials. Indeed, for some cancers randomized trials have been performed or are underway (e.g. breast cancer, prostate, lung).

Table 2. Sensitivity and specificity of a tumour marker in screening or diagnosing cancer; sensitivity, or true positive rate, is defined as a/(a+c); specificity, or true negative rate, is defined as d/(b+d)

	Cancer	No cancer
Tumour marker positive	a	b
Tumour marker negative	c	d

Epidemiological studies of diagnosis

As for screening, we require that the tumour marker can separate subjects without from those with the cancer. However, setting a diagnosis is usually not possible with a single characteristic, and has a sequential nature, starting with simple tests. Moreover, some diagnoses are probabilistic in nature, i.e. that we can give a probability of a certain condition rather than 100% certainty. An example is the qualification of a screen-detected prostate cancer as probably indolent. PSA levels, prostate volume, and biopsy features were combined in a logistic regression model to estimate this probability (1). This model was presented as a nomogram, such that it would be relatively easy to apply by physicians. A nomogram is a graphical presentation of the model (figure 2); alternatives include tables of predicted probabilities according to predictive characteristics, and score charts.

The nomogram from figure 2 was validated recently in patients from the European screening trial on prostate cancer (ERSPC), and found systematically invalid (7). The probability of indolent cancer was around 50% in the ERSPC, while the average predicted probability was 20%. An updated version of the model was constructed, which may guide deci-

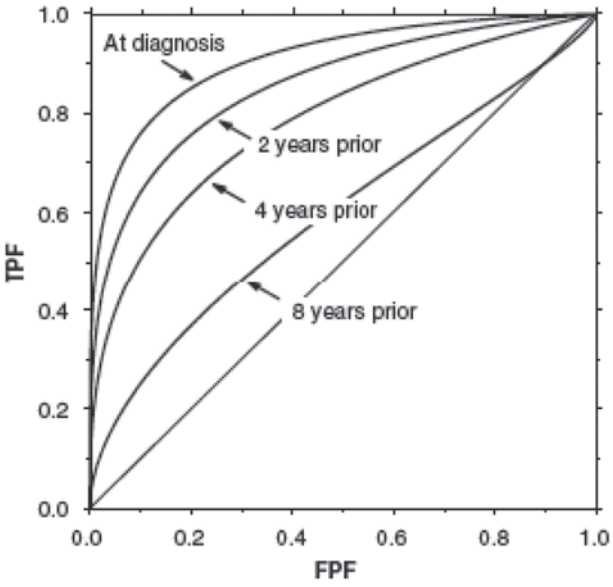


Figure 1. Time-dependent ROC curves for PSA in screening for prostate cancer. TPF: true-positive fraction; FPT: false-positive fraction. From (6).

sion making in screen-detected prostate cancer after further validation.

Epidemiological studies of prognosis

For prognostic studies, we usually consider a cohort of patients, and follow them in time for the occurrence of a relevant medical outcome. In cancer, common outcome measures are recurrence of disease, and (cancer-specific) survival. Tumour marker values can be related to these outcomes in statistical analyses. A simple cross-table can be constructed for high and low tumour marker values versus the outcome, when we know the outcome for all patients. For example, we may want to study 1 year survival and have at least 1 year follow-up for all patients. Survival analysis is required when the outcome is not observed for all patients, e.g. 5-year survival. Kaplan-Meier curves can then be constructed, which consider patients who did not die as censored observations. They are considered in the analysis until the end of their follow-up period (figure 3).

Combinations of characteristics can be considered in the Cox regression model. In such a regression model, we do not have to categorize tumour marker values, but can analyze them as continuous variables. For example, we can consider the log(PSA) in predicting survival of prostate cancer patients. Predictions from such models usually need validation in new patients before we can use them in clinical practice, especially when the number of events in the analysis was relatively small, and new markers were considered in the model rather than well-established markers.

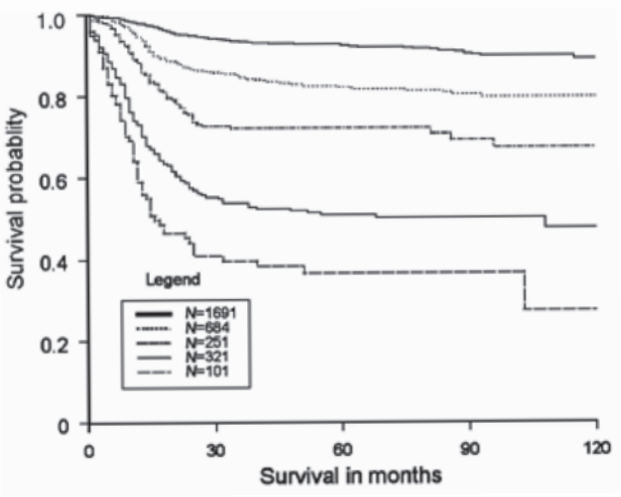
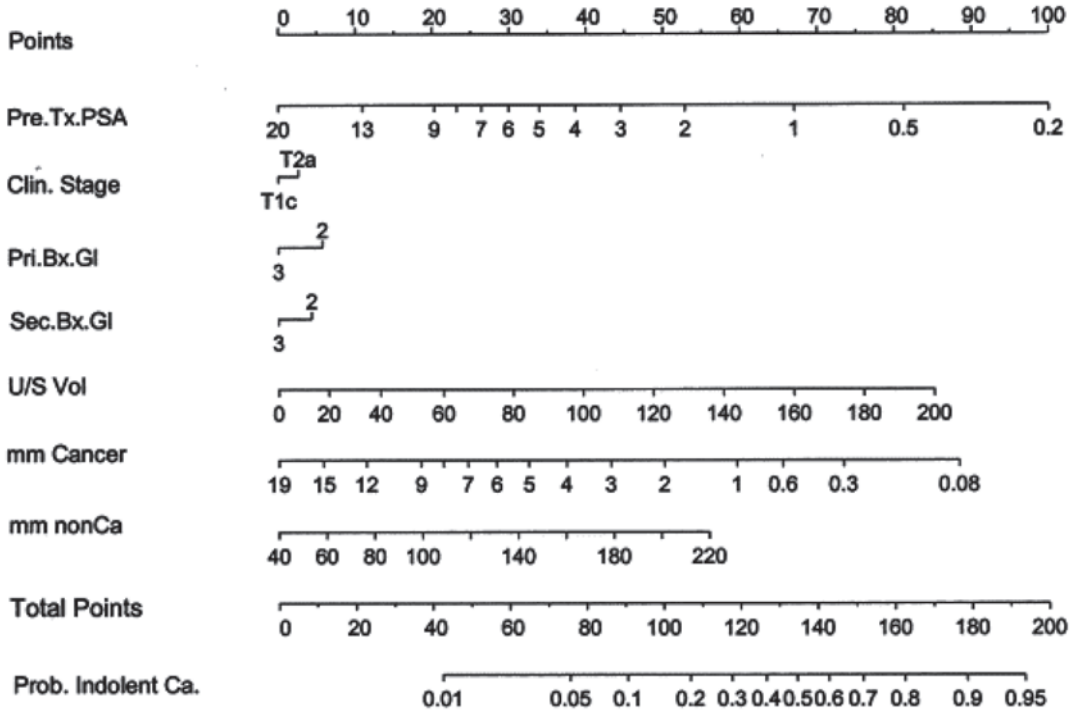


Figure 3. Survival curves for patients with nonseminomatous testicular cancer, as published by Van Dijk et al (8). Five groups were created based on the IGCC classification, which includes tumour markers AFP and HCG (5).



Instructions for Physician: Locate the patient's PSA on the **PreTx PSA** axis. Draw a line straight upwards to the **Points** axis to determine how many points towards having an indolent cancer the patient receives for his PSA. Repeat this process for the remaining axes, each time drawing straight upward to the **Points** axis. Sum the points achieved for each predictor and locate this sum on the **Total Points** axis. Draw a line straight down to find the patient's probability of having indolent cancer.

Instruction to Patient: "Mr. X, if we had 100 men exactly like you, we would expect <predicted probability from nomogram * 100 > to have indolent cancer."

Figure 2. Nomogram to predict the probability of indolent cancer, as published by Kattan et al (1). Pre.Tx., pre-treatment; Clin., clinical; Pri.Bx.Gl, primary biopsy Gleason score; Sec.Bx.Gl, secondary biopsy Gleason score; U/S, ultrasound; Prob., probability.

Developing useful new tumour markers

The development of clinically useful new tumour markers poses a number of challenges. Technological advancements allow for a increasing number of markers to be studied, including biochemical and genetic markers. Standardization of the tests and reproducibility of results are minimum requirements before thinking about wider applications of new markers in medicine. Case-control studies can subsequently be conducted using 'convenient samples', e.g. from stored tissue, but better with population based samples (6). Specific problems arise in the current evaluation of new markers (9). First, a great number of markers are potential candidates for use in screening, diagnosis, or prognosis. This leads to a multiple testing problem. If we evaluate 1000 markers, we expect that 50 of them reach statistical significance with the classical $p<0.05$ criterion, even if none of them is truly associated with cancer. Genome-wide searches include even more potential markers. Several approaches can be followed to address the multiple testing problem, including Bonferroni's correction, but such approaches limit the statistical power of a study, i.e. the probability of finding a significant result if a true association was present. Power in marker studies is typically already quite limited because of small sample size, e.g. less than 100 cases, and less than 100 controls. Initiatives have been proposed for analyses with larger sample sizes (10). Further, results of marker evaluations may only be reported if statistically significant (11). The effects are then overestimated, since non-significant results are by definition closer to 'no effect', and are not reported ('publication bias') (12).

A specific issue is how we deal with continuous values of a marker. It may be convenient to search for a dichotomization as positive versus negative. Searching for an optimal cut-off is fraught with statistical problems; the association will be exaggerated (13). In general, dichotomization is discouraged; continuous versions of a marker contain often much more information. In addition to the linear and logarithmic transformation, spline functions may well be used to study diagnostic and prognostic relationships (14).

Finally, new markers should be judged for their incremental value over classical markers and other diagnostic or prognostic characteristics. This can well be studied with regression models, where first a model is made with the traditional predictors, and next a model with these traditional predictors plus the new marker. If the latter model is promising, it should be validated in new patients to assess generalizability. In many clinical areas, the choice of more modern statistical methods, e.g. a classification tree or neural network instead of a regression model, did not improve the quality of predictions (15).

Summary points

- Tumour markers have an important role in screening, diagnosis, and prognosis of cancer
- Sensitivity and specificity are important characteristics of a screening or diagnostic test; they can jointly be considered in a ROC curve
- Lead time bias and length bias are among the major problems in the evaluation of screening programs; randomized controlled trials are the ultimate way to evaluate benefits and harms of screening
- Regression models may combine tumour markers with other characteristics to estimate a probability of a diagnosis (e.g. indolent cancer) or a prognostic outcome (e.g. survival). These models require validation before widespread use is possible
- The development of clinically useful new tumour markers is associated with several problems, including multiple testing, limited power because of small sample sizes, publication bias, and having to deal with continuous marker values. The incremental value should be assessed over traditional predictors, with rigorous validation of initially (too?) promising results.

References

1. Kattan MW, Eastham JA, Wheeler TM, et al. Counseling men with prostate cancer: a nomogram for predicting the presence of small, moderately differentiated, confined tumours. J Urol 2003;170:1792-7.
2. Steyerberg EW, Keizer HJ, Fossa SD, et al. Prediction of residual retroperitoneal mass histology after chemotherapy for metastatic nonseminomatous germ cell tumour: multivariate analysis of individual patient data from six study groups. J Clin Oncol 1995;13:1177-87.
3. Steyerberg EW, Keizer HJ, Habbema JD. Prediction models for the histology of residual masses after chemotherapy for metastatic testicular cancer. ReHiT Study Group. Int J Cancer 1999;83:856-9.
4. Kattan MW. Nomograms are superior to staging and risk grouping systems for identifying high-risk patients: preoperative application in prostate cancer. Curr Opin Urol 2003;13:111-6.
5. International Germ Cell Consensus Classification: a prognostic factor-based staging system for metastatic germ cell cancers. International Germ Cell Cancer Collaborative Group. J Clin Oncol 1997;15:594-603.
6. Pepe MS. Evaluating technologies for classification and prediction in medicine. Stat Med 2005;24:3687-96.
7. Steyerberg EW, Roobol MJ, Kattan MW, van der Kwast TH, de Koning HJ, Schröder FH. Prediction of indolent prostate cancer: validation and updating of a prognostic nomogram. J Urol 2007; Jan. 177(1):107-12.
8. van Dijk MR, Steyerberg EW, Stenning SP, Dusseldorp E, Habbema JD. Survival of patients with nonseminomatous germ cell cancer: a review of the IGCC classification by Cox regression and recursive partitioning. Br J Cancer 2004;90:1176-83.
9. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.
10. Ioannidis JP, Gwinn M, Little J, et al. A road map for efficient and reliable human genome epidemiology. Nat Genet 2006;38:3-5.
11. Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. J Natl Cancer Inst 2005;97:1043-55.
12. Ioannidis JP, Trikalinos TA. Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. J Clin Epidemiol 2005;58:543-9.
13. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumour marker prognostic studies (REMARK). J Natl Cancer Inst 2005;97:1180-4.
14. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer, 2001:xxii, 568.
15. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Stat Med 1998;17:2501-8.