HANS C. VAN HOUWELINGEN

Predicting Patient Survival using Genomic Data

The role of medical statistics

Clinical (medical) research is centered around clinical trials that compare different treatments. The role of medical statistics is i) to design such studies (how many patients etc.), ii) to compare the treatments (is there a significant difference?), iii) to use the data to predict outcome of treatment, and iv) to help to select the best treatment.

In cancer, the main primary outcome is SURVIVAL, if possible adjusted for QUALITY OF LIFE. Survival data are special because *it takes time to observe time*. Hence, the data are always incomplete (censored) and, moreover, the data are dynamic over time.

Special techniques have been developed to analyze survival data. The most well-known are the Kaplan-Meier survival curves and the Cox regression model for survival data. One prime example is given by the following graph (Figure 1) based on data from the Dutch Gastric Cancer Trial (Putter et al., 2005). There, two different surgical strategies, denoted by D1 and D2 are compared. At first sight there is little difference. Risk avoiders might opt for D1, the less aggressive treatment, because its short term prospects are better, while risk seekers might opt for D2, the more aggressive treatment, which might give better long term survival.

Professor Hans C. van Houwelingen
Medical Statistics and Bioinformatics,
Leiden University Medical Center,
The Netherlands
E-mail: J.C.van_Houwelingen@lumc.nl
CAS Fellow 2005/2006

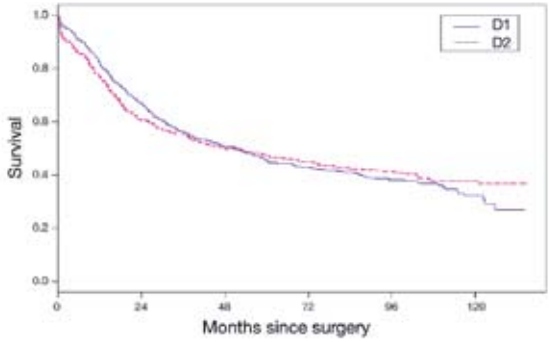

Figure 1. Kaplan-Meier plots of the survival curves for D1-and D2-dissection. The survival curves cross after 53 months.

Building prognostic models

Treatment is usually not the most relevant predictor of outcome. More important are clinical factors such as i) stage of the disease, ii) residual tumor after surgery, iii) subtype of tumor (histology), and iv) condition of the patient (at start of treatment or dynamic during follow-up).

The general structure of a prognostic study is

- find out what information is relevant for a prognosis
- define a score (Prognostic Index) that summarizes the information
- use that score to classify patients in prognostic categories
- communicate the relationship between score and prognosis by means of tables or graphs (or interactive software)

One example of such a prognostic model can be found in one of my first papers after I moved into medical statistics in 1986 (Van Houwelingen et al., 1989). It deals with the prognosis of patients with advanced ovarian cancer. On the basis of the clinical information, six prognostic groups can be distinguished. The survival curves as shown (Figure 2) are obtained from the Cox regression model. They are based on the percentiles of the prognostic index derived from the Cox model (0–10 per cent, 10–25 per cent, 25–50 per cent, 50–75 per cent, 75–90 per cent, 90–100 per cent). Thus, for example, the upper curve of the plot is the survival curve of the patients having the most favorable prognosis. The curve implies that in this group approximately 60 per cent of the patients will survive for six years.

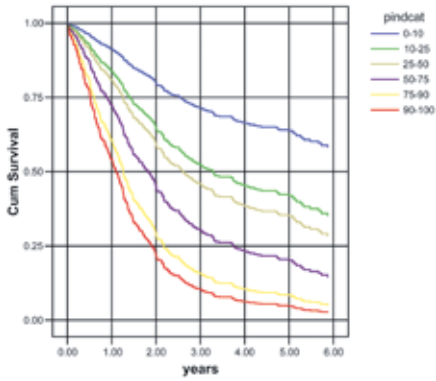

Figure 2. Survival function for patterns 1–6.

Although such studies are popular and useful, the 'predictability' of patient survival should not be overestimated. The differences seen in survival curves are of very limited relevance for the patient. There is a lot of variation around 'mean survival' within groups with similar prognoses. Instead of showing survival curves, it would be better to give survival margins (prediction intervals). Figure 3 shows the survival of the Dutch ovarian cancer patients as the 25 per cent –75 per cent prediction interval in years per category. For example, in the fourth group (50–75) 25 per cent of the patients will die within one year and 75 per cent will die within 3.5 years, so the prediction interval runs from 1 to 3.5 years.

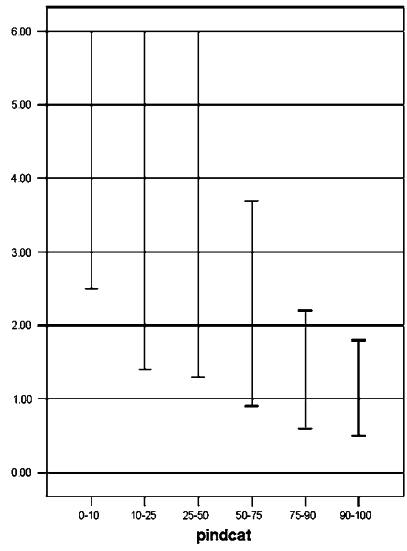

Figure 3.

Using genetic and genomic information

Can genetics and genomics help improve prognoses and, hopefully, offer better treatment for patients? Genetics (inherited DNA) is relevant for the susceptibility. (However, cancer is less genetic than most people think). Genomics (RNA , gene expression obtained through micro-arrays that

measure the amount of RNA in tissue or serum, proteins measured by mass spectrometry) might be helpful for prediction. The success story in using gene expression data for prediction is the breast cancer data set of the Netherlands Cancer Institute in Amsterdam (van de Vijver, 2002).

I reanalyzed those data with colleagues in Amsterdam (Van Houwelingen et al., 2006), using statistical methodology one of my PhD students and I developed about a decade ago. The problem with using such data for prediction is that there are many more predictors (gene expression for about 5000 genes) than patients (295 for the breast cancer data).

Conventional methodology (step-wise regression) that works well for small number of predictors will fail. We can perfectly 'predict' what we have, but fail to predict new observations. In order to get answers that make sense, we have to 'tame' the statistical algorithm by penalization to prevent over-fitting and check continuously what we would get in new data by crossvalidation.

As a result, we obtained four prognostic groups (see the flanking graph, figure 4) that show some differentiation with respect to survival.

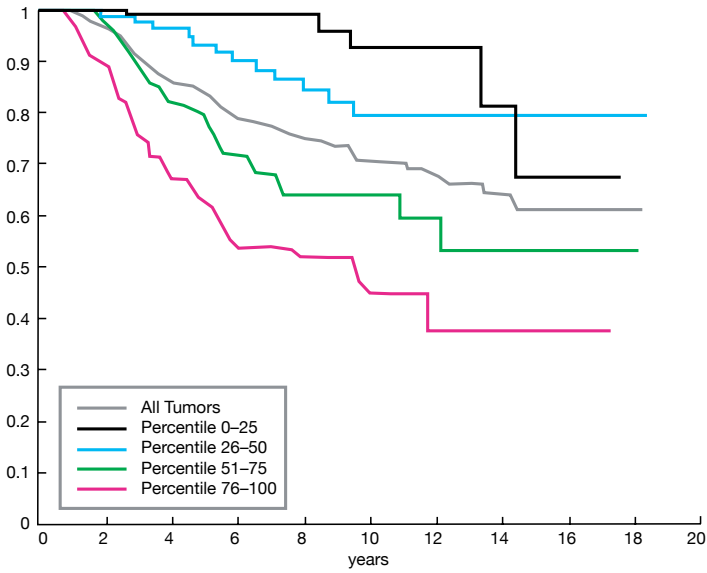

Figure 4. Kaplan-Meier curves for four equal sized subgrups.

However, the genomic information appears to be correlated with the clinical information (tumor stage etc.) and gene expression does not add very much to the prognostic model based on clinical information. Currently, I am working on methods to improve the prognostic model by using biological information that helps to group the genes in pathways. It looks as though that might help considerably.

References

Putter, H; Sasako, M; Hartgrink, HH; van de Velde, CJH; van Houwelingen, JC. "Long-term survival with non-proportional hazards: results from the Dutch Gastric cancer Trial" *Statistics in Medicine* 24 (18), 2005: 2807–2821.

van de Vijver, MJ; He, YD; van 't Veer, LJ; Dai, H; Hart, AAM; Voskuil, DW; Schreiber, GJ; Peterse, JL; Roberts, C; Marton, MJ; Parrish, M; Atsma, D; Witteveen, A; Glas, A;

Delahaye, L; van der Velde, T; Bartelink, H; Rodenhuis, S; Rutgers, ET; Friend, SH; Bernards, R. "A gene-expression signature as a predictor of survival in breast cancer." *New England Journal of Medicine* (25) 2002: 1999–2009.

van Houwelingen, JC; Huinink, WWT; van der Burg, MEL; van Oosterom, AT; Neijt, JP. "Predictability of the survival of patients with advanced ovarian-cancer" *Journal of Clinical Oncology* 7 (6), 1989: 769–773.

van Houwelingen HC, Bruinsma T, Hart AA, Van't Veer LJ, Wessels LF. "Cross-validated Cox regression on microarray gene expression data" *Statistics in Medicine*, 25 (13), 2006, 3201–3216.