

**Challenges in design, analysis and reporting of prognostic and predictive marker research – from single studies to an EBM based assessment**

Satellite workshop to the Cochrane Colloquium

Organisers:

Willi Sauerbrei (Institut für Medizinische Biometrie und Medizinische Informatik, Freiburg),

Doug Altman (Centre for Statistics in Medicine, Oxford)

Freiburg, 8 October 2008

Hörsaal Virologie, Institut für Medizinische Mikrobiologie und Hygiene,

Hermann-Herder-Str. 11, 79104, Freiburg

## Program

<b>9.00</b>	<b><i>Welcome</i></b>	
<b>9.05 – 10.35</b>	<b><i>Introduction</i></b> Doug Altman (Oxford)	From single studies to an EBM based assessment – some central issues
	<b><i>Analysis of individual studies</i></b> Karel Moons (Utrecht)	Design, analysis and impact of single prognostic studies
	Patrick Royston (London)	Multivariable modelling of continuous markers, with extension to interactions with treatment
<b>11.00 – 12.30</b>	<b><i>Quality of the literature</i></b> Jill Hayden (Toronto)	Evaluation of the quality of prognosis studies in systematic reviews
	Panayiotis Kyzas (Ionnina)	Selective reporting, quality of reporting and statistical significance chasing in prognostic marker studies
	<b><i>Validation</i></b> Marc Buyse (Brussels)	On the need for external validation in biomarker research
<b>12.30 – 13.30</b>	<b><i>Lunch break with sandwiches</i></b>	
<b>13.30 – 15.00</b>	<b><i>Evaluating and comparing prediction models</i></b> Ewout Steyerberg (Rotterdam)	Assessment of performance and decision curve analysis
	<b><i>High-dimensional data</i></b> Ulrich Mansmann (Munich)	High-dimensional prognosis
	Elia Biganzoli (Milan)	High-dimensional data
<b>15.30 – 17.00</b>	<b><i>Meta-Analysis</i></b> Simon Thompson (Cambridge)	Lessons from meta-analyses of observational studies in epidemiology
	Richard Riley (Liverpool)	The challenges of a systematic review and meta-analysis of prognosis studies
	<b><i>Summing up</i></b> Martin Schumacher (Freiburg)	Summary of main challenges and future directions

# From single studies to an EBM based assessment – some central issues

**Doug Altman,**

*Centre for Statistics in Medicine, Oxford*

Prognostic markers can help to identify patients at different degrees of risk for specific outcomes, facilitate treatment choice, and aid patient counselling. Numerous studies are carried out to investigate factors that could help to explain the large variation in patient prognosis. The principles of good study design and analysis are less well appreciated for prognostic factor studies than therapeutic studies,<sup>1,2</sup> so that the literature is full of conflicting results from studies of varying (and often poor) quality and inadequate size. Few recently proposed tumour markers have been shown to be clinically useful.

Among the difficulties are retrospective sampling, variety of ways of grouping continuous variables (some of which lead to bias), use of different assays, varying and often inadequately described patient cohorts, incomplete data, and adjustment for different other variables (often using data-dependent selection methods). In addition, poor reporting<sup>3</sup> and publication bias<sup>4</sup> are a major concern for such studies.

It is usually difficult to ascertain the benefit of a marker from single studies and a clear view is only likely to emerge from looking across multiple studies.<sup>5</sup> Current systematic reviews and meta-analyses often fail to provide clear answers, and rather only draw attention to the paucity of good-quality evidence.

Large protocol-driven, prospective studies are the ideal, with clear, unbiased reporting of the methods used and the results obtained.<sup>5</sup> Unfortunately, there are few such prognostic studies. Also, prospectively planned pooled analyses of high-quality studies, along with general availability of individual patient data and adherence to reporting guidelines, would help alleviate many of the problems.<sup>5</sup>

There is also concern about studies to develop prognostic models, reflecting many of the same issues of design and analysis, and also the paucity of good validation studies.<sup>6</sup> Studies of high dimensional data bring even more difficulties.<sup>7</sup>

Improvement in the coming years requires more careful attention to the design and analysis and reporting of prognostic studies.<sup>8,9</sup>

1. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer* 1994;69:979-85.
2. Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Res Treat* 1998;52:289-303.
3. Riley RD, Abrams KR, Sutton AJ, et al. Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *Br J Cancer* 2003;88:1191-8.
4. Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. *J Natl Cancer Inst* Jul 20 2005;97:1043-55.

5. Altman DG, Riley RD. Primer: an evidence-based approach to prognostic markers. *Nat Clin Pract Oncol* 2005;2:466-72.
6. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. In: Lyman GH, Burstein HJ (eds). *Breast cancer. Translational therapeutic strategies*. New York: Informa Healthcare, 2007: 11-25.
7. Koscielny S. Critical review of microarray-based prognostic tests and trials in breast cancer. *Curr Opin Obstet Gynecol* 2008;20:47-50.
8. Hayden JA, Côté P, Steenstra IA, Bombardier C, et al. Identifying phases of investigation helps planning, appraising, and applying the results of explanatory prognosis studies. *J Clin Epidemiol* 2008;61:552-60.
9. Rifai N, Altman DG, Bossuyt PM. Reporting bias in diagnostic and prognostic studies: time for action. *Clin Chem* 2008;54:1101-1103.

# Design and analysis of single prognostic studies

**Karel G.M. Moons, PhD.**

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht,  
The Netherlands. k.g.m.moons@umcutrecht.nl

In the medical literature, etiologic, therapeutic and laboratory (e.g. genetic and molecular) traditionally gained much more appreciation than prognostic studies, even though Hippocrates already recognised that setting a prognosis forms the basis of medical care. Setting a prognosis is estimating the probability or risk of developing a particular outcome in the (near) future. Practicing physicians estimate this probability in their patients to tailor subsequent management. Since the introduction of evidence based medicine there seems a paradigm shift from eminence and experience based medicine with largely implicit estimation of a patient's prognosis, to explicit estimations using properly developed and validated prognostic tools. The latter include notably multivariable prognostic models (or prediction rules or risk scores), presented in varying formats.

Studies to develop multivariable prognostic models are often incorrectly designed or analysed yielding invalid results with limited relevance to clinical practice This includes improper sampling of patients (e.g. a case control approach) that does not match the goals of prognosis, improper methods for developing a model (e.g. selection of too many predictors given the available data), problems of overfitting a model, and the simple deletion of records with missing values. Properly dealing with missing values – e.g. using multiple imputation - is obviously an issue for any type of medical research but I will focus on its merits for prediction research. I will briefly overview the desired designs and analysis to develop a prognostic model. Finally, I touch upon the phases to be conducted after model development, including model validation, updating and implementation.

1. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;**59**(10):1087-91.
2. Moons KG, Donders AR, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol* 2004;**57**(12):1262-70.
3. Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006;**59**(10):1092-101.
4. Moons KG, Harrell FE, Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? *J Clin Epidemiol* 2002;**55**(10):1054-5.
5. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* 1996;**49**:907-16.
6. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol* 2003;**56**(5):441-7.

7. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361-87.
8. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* 2000;**19**(8):1059-79.
9. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;**61**(1):76-86.

# **Multivariable modelling of continuous markers, with extension to interactions with treatment**

**Patrick Royston**

MRC Clinical Trials Unit, London ([pr@ctu.mrc.ac.uk](mailto:pr@ctu.mrc.ac.uk))

The usual approach to modelling continuous predictors in regression settings assumes linear covariate effects, but the linearity assumption may be violated. Alternatively, cut-points are often used, inappropriately implying that the dose-risk relationship is a step function. Use of cut-points has major disadvantages and often the fit can be much improved by explicit estimation of the functional form.

Approaches based on splines (e.g. in generalized additive models; Hastie & Tibshirani 1990) or fractional polynomials (FP) (Royston & Altman 1994) have also been used. Splines are considered to be more flexible because they model functional relationships locally (nevertheless often controlled by a global smoothing parameter) whereas FPs model relationships globally.

The multivariable fractional polynomial (MFP) procedure (Sauerbrei & Royston 1999; Royston & Sauerbrei 2008) combines backward elimination of weakly influential variables with a systematic search for well-fitting FP functions of continuous covariates. By illustrating the analysis of medical data in the framework of a regression model, several issues of multivariable model building with continuous data will be discussed. The emphasis will be on MFP.

An important issue in clinical trials and in clinical research generally is the possibility of interaction between a randomized treatment and a continuous covariate. We also present and discuss a primarily graphical approach to detecting and modelling a treatment-covariate interaction known as STEPP (Bonetti & Gelber 2000, 2004), and an extension of MFP known as MFPI (Royston & Sauerbrei 2004), which detects and models such interactions while retaining the continuous scale of the covariate and also models non-linearity if present. A randomized trial in advanced renal cancer is used to exemplify the methods.

Bonetti M, Gelber RD (2000): A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. *Statistics in Medicine*, 19: 2595-2609.

Bonetti M, Gelber RD (2004): Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*, 5:465-481.

Hastie, TJ, Tibshirani, RJ. *Generalized Additive Models*. Chapman & Hall, London, 1990.

Royston P, Altman DG (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion) *Applied Statistics* 43: 429-467.

Royston P, Sauerbrei W (2004). A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine* 23: 2509-2525.

Royston P, Sauerbrei W (2008) *Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Chichester, John Wiley & Sons.

Sauerbrei W, Royston P (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society, Series A* 162: 71-94.



# Exploration of Methods of Prognosis Studies and Reviews

**Hayden JA**

*Dept. of Community Health & Epidemiology  
Dalhousie University  
Halifax, Nova Scotia CANADA  
[JHayden@Dal.ca](mailto:JHayden@Dal.ca)*

Prognosis is the probable course and outcome of a health condition over time. Researchers gather information regarding prognosis from observational and experimental designs. Clinicians commonly use prognostic information to educate their patients, identify target groups for treatment, or to target specific factors to modify through intervention<sup>1</sup>. However, the published literature is not always easy to interpret and apply. Systematic reviews of prognosis have been increasingly published<sup>2</sup> and often highlight inconsistent and sometimes contradicting conclusions from prognosis studies of similar conditions. In this presentation I will discuss three recent projects that explore the methods of prognosis studies and reviews, and make recommendations that we hope will advance this literature.

First, I will present the results of a 'review of systematic reviews' on low back pain prognosis<sup>3</sup>. We observed important differences in the methods of 17 prognosis systematic reviews. A quarter of reviews did not clearly report their search strategies. The number of potential citations identified ranged from 15 to 4458, and the number of included prognosis studies ranged from 3 to 32 (of 162 distinct citations included across reviews). 70% of reviews assessed quality of included studies, but assessed only a median of 4 of 6 potential biases. All reviews reported associations based on statistical significance only; they used various strategies for syntheses. We found discrepancies in results across reviews: differences in some selection criteria influenced studies included, and various approaches to data interpretation influenced review conclusions about evidence for specific prognostic factors.

Second, I will discuss a proposed framework for prognosis research<sup>4</sup> that we adapted from earlier work<sup>5</sup>. We identify two main approaches to study prognosis, which influence the strength of evidence and interpretation of results: 1. explanatory analyses, and 2. outcome prediction. Explanatory studies focus on the causal association between prognostic factors and an outcome, while outcome prediction studies focus on variables taken together in order to identify the combination of factors that is most strongly associated with outcome and can be used to stratify patients on an outcome, often to triage them into treatment programs<sup>6, 7</sup>. We propose three phases of investigation that define the level of prognostic evidence for explanatory prognosis studies (Figure).

Third, I will discuss quality assessment of prognosis studies<sup>2</sup>. We found that quality assessment in systematic reviews of prognosis studies was often incomplete and that there is wide variation in current practice. We propose four distinct elements necessary to adequately assess the quality of prognosis studies: 1. operationalization of items to address potential opportunities for bias (including biases related to study participation, study attrition, measurement of prognostic factors, measurement of and controlling for

confounding variables, measurement of outcomes, and analysis approaches), 2. assessment of biases including judgments about risk, 3. synthesizing the evidence, and 4. reporting results.

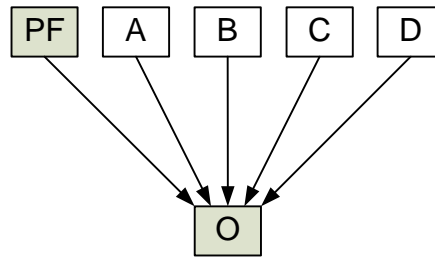
We will discuss potential impacts of prognosis study methods for each project reported, and make recommendations for future prognosis research.

### **References:**

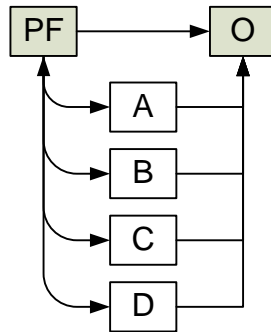
- (1) Croft PR, Dunn KM, Raspe H. Course and prognosis of back pain in primary care: the epidemiological perspective. *Pain* 2006; 122(1-2):1-3.
- (2) Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. *Annals of Internal Medicine* 2006; 144(6):427-437.
- (3) Hayden JA, Chou R, Hogg-Johnson S, Bombardier C. Systematic reviews of low back pain prognosis had variable methods and results - Guidance for future reviews. Accepted, *Journal of Clinical Epidemiology*, 2008.
- (4) Hayden JA, Côté P, Steenstra IA, Bombardier C, for the QUIPS-LBP Working Group. Identifying phases of investigation helps planning, appraising and applying the results of explanatory prognosis studies. *Journal of Clinical Epidemiology*. 2008;61(6):552-60.
- (5) Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Research & Treatment* 1998; 52(1-3):289-303.
- (6) Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of Internal Medicine* 2006; 144(3):201-209.
- (7) Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997; 277(6):488-494.

**Figure:** Graphical representations of the phases of explanatory prognosis studies. The prognostic factor of interest is indicated by 'PF'; 'O' represents the outcome of interest; 'A' to 'D' represent potential confounders. In Phase 3 studies 'PF\*' may represent a prognostic factor construct. The graphic illustrates one example. From Hayden et al. (2008)<sup>4</sup>.

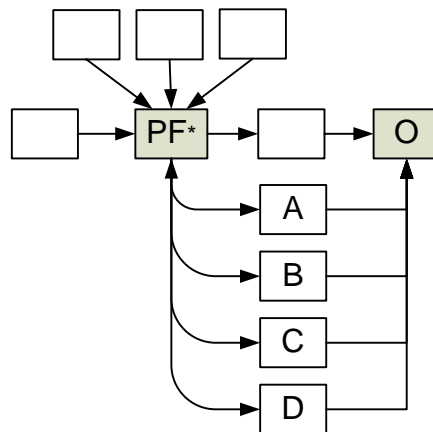
**A. Phase 1: Identifying associations**



**B. Phase 2: Testing independent associations**



**C. Phase 3: Understanding prognostic pathways**



# Selective reporting, quality of reporting and statistical significance chasing in prognostic marker studies

Panayiotis Kyzas<sup>1,2</sup>

<sup>1</sup>: Research Fellow, Department of Hygiene and Epidemiology, University of Ioannina, Medical School, Greece. <sup>2</sup>: Specialist trainee, Oral and Maxillofacial Surgery, North Manchester General Hospital, Manchester, UK.

## Background

Several methodological problems have been implicated for prognostic marker studies. We aimed to investigate three domains that may introduce biases in this literature. The first domain refers to non-reported and selectively reported information, the second deals with issues of reported study quality, and the third examines the extent of the pursuit for statistically significant results.

## Methods – Results

We probed selective reporting bias in a meta-analysis of a prognostic factor for head and neck squamous cell cancer mortality that has drawn wide attention — the status of the tumor suppressor protein TP53. We compared results of meta-analyses that included published data plus unpublished data retrieved from investigators; published data; and only published data indexed with “survival” or “mortality” in MEDLINE/ EMBASE, with or without standardized definitions. For studies that included published and indexed data, we obtained a highly statistically significant association between TP53 status and mortality. When we used the definitions preferred by each publication, the association was stronger ([RR] = 1.38, 95% [CI] = 1.13 to 1.67; P = .001) than when we standardized definitions (RR = 1.27, 95% CI = 1.06 to 1.53; P = .011). The addition of studies that included published but not indexed data reduced the observed association (RR = 1.23, 95% CI = 1.03 to 1.47; P = .02). Finally, when we obtained data from investigators and analyzed it with all other data, statistical significance was lost (RR = 1.16, 95% CI = 0.99 to 1.35; P = .06).

To address reported study quality, we evaluated eight quality measures pertaining to study design and assay methods of the REMARK consensus (i.e., blinding, prospective versus retrospective design, power calculations, outcomes’ definitions, time of enrollment, reporting of variables, assay description, and assay reference) in cancer prognostic marker studies included in meta-analyses<sup>2</sup>. We estimated the ratios of relative risks, which compared the overall prognostic effects (summary relative risks) between poor-quality and good-quality studies for each quality item. Only three (0.9%) of the studies presented power calculations, 129 (39.0%) studies stated that analyses were blinded, and 73 (21.5%) stated that they were prospective. Time of enrollment was defined in 232 (70.0%), 234 (70.7%) gave lists of candidate variables, and 254 (76.7%) defined outcomes. The assay used was described in 317 (95.8%), but only 177 (53.5%)

provided the assay reference. Summary ratios of relative risks of poor- versus good-quality studies for the seven quality measures ranged from 0.95 to but 1.26, but none was statistically significantly.

Finally, we aimed to understand the extent of the pursuit for statistically significant results in the prognostic literature of cancer<sup>3</sup>. We evaluated 340 articles included in prognostic marker meta-analyses (Database 1) and 1575 articles on cancer prognostic markers published in 2005 (Database 2). For each article, we examined whether the abstract reported any statistically significant prognostic effect for any marker and any outcome ('positive' articles). 'Negative' articles were further examined for statements made by the investigators to overcome the absence of prognostic statistical significance. 'Positive' prognostic articles comprised 90.6% and 95.8% in Databases 1 and 2, respectively. Most of the 'negative' prognostic articles claimed significance for other analyses, expanded on non-significant trends or offered apologies that were occasionally remote from the original study aims. Only five articles in Database 1 (1.5%) and 21 in Database 2 (1.3%) were fully 'negative' for all presented results in the abstract and without efforts to expand on non-significant trends or to defend the importance of the marker with other arguments.

### **Conclusion**

Selective reporting may spuriously inflate the importance of postulated prognostic factors for various malignancies. Among cancer prognostic marker studies, reported quality of design and assay information often appears suboptimal. Furthermore, almost all articles on cancer prognostic marker studies highlight some statistically significant results. All these biases indicate that this literature may be largely unreliable and that under these circumstances, statistical significance loses its discriminating ability for the importance of prognostic markers.

### **References**

1. Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. *J Natl Cancer Inst.* 2005;97(14):1043-55.
2. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Quality of reporting of cancer prognostic marker studies: association with reported prognostic effect. *J Natl Cancer Inst.* 2007;99(3):236-43.
3. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer.* 2007;43(17):2559-79.

# External validation in biomarker research: examples from gene profiling in early breast cancer

Marc Buyse, ScD

IDDI, Louvain-la-Neuve, and Hasselt University, Diepenbeek, Belgium

Despite 20 years of advances in molecular biology, few biomarkers have so far been shown to be clinically useful in predicting patient outcomes (*prognostic biomarkers*), let alone therapeutic response (*predictive biomarkers*) [1]. The present talk will illustrate some of the difficulties involved in identifying and validating clinically useful gene signatures [2,3]. Over the last few years, several signatures have been shown to have independent prognostic impact for patients with early breast cancer [4-7]. The prospective validation of these signatures on independent datasets [8,9] has shed light on several outstanding questions: Is the predictive accuracy of the signatures acceptable [10]? Is there a better signature, given that they involve different sets of genes [7]? Do they predict early and late events equally well [8,9]? Do they add to the clinico-pathological factors that are routinely used for cancer prognosis [11]? Trials are currently on-going to confirm the usefulness of these signatures in clinical practice.

## References

1. Hammond ME, Taube SE. Issues and barriers to development of clinically useful tumor markers: a development pathway proposal. *Semin Oncol* 2002;29:213-21
2. [Simon R](#), [Radmacher MD](#), [Dobbin K](#), [McShane LM](#). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 2003;95:14-8
3. [Lusa L](#), [McShane LM](#), [Radmacher MD](#), et al. Appropriateness of some resampling-based inference procedures for assessing performance of prognostic classifiers derived from microarray data. *Stat Med* 2007;26:1102-13
4. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;347:1999-2009
5. Wang Y, Klijn J, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;365:671-79
6. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;351:2817-26
7. Sotiriou C, Wirapati P, Loi S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006;98: 262-272
8. Buyse M, Loi S, van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006;98:1183-92

9. Desmedt C, Piette F, Loi S, et al. Strong time-dependency of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multi-centre independent validation series. *Clin Cancer Research* 2007;13:3207-3214
10. Pepe MS, Janes H, Longton G, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol.* 2004;159:882-90
11. Dunkler D, Michiels S, Schemper M. Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? *Eur J Cancer* 2007;43:745-51

# Assessment of performance and decision curve analysis

**Ewout W. Steyerberg, PhD<sup>1</sup>; Andrew J. Vickers<sup>2</sup>**

1 Dept of Public Health, Erasmus MC, Rotterdam, the Netherlands. 2 Dept of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, USA

From a research perspective, diagnosis and prognosis constitute a similar challenge: the clinician has some information and wants to know how this relates to the true patient state, whether this can be known currently (diagnosis) or only at some point in the future (prognosis). This information can take various forms, including a diagnostic test, a marker value, or a statistical model including several predictor variables. In all cases, the information can be expressed as probabilistic predictions for the outcome of interest. Predictions are hence absolute risks, which go beyond assessments of relative risks, such as regression coefficients, odds ratios or hazard ratios.

There are various ways to assess the performance of a marker, diagnostic test, or statistical model.<sup>1</sup> The traditional statistical approach to performance is to quantify how close our predictions are to actual outcome, using measures such as explained variation ( $R^2$ ) and the Brier score. Performance can further be quantified in terms of calibration (do close to  $x$  of 100 patients with a risk prediction of  $x\%$  have the outcome?) and discrimination (do patients who have the outcome have higher risk predictions than those who do not?), using “goodness-of-fit” statistics and measures such as sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC, or c statistic).

The problem with these measures is that they are uninformative as to clinical usefulness: how accurate is accurate enough to justify the clinical use of a test, marker or model? Evaluation in terms of clinical consequences is the remit of “decision analysis”. A problem with such an evaluation is that it requires additional information, for example, on the benefits, harms and costs of treatment, or on patient preferences for different health states. Decision curve analysis is however a recent decision analytic method that can be implemented without the need for substantive additional data.<sup>2,3</sup> The method starts from the observation that clinical decisions require a cutoff for the predicted probability (the ‘probability threshold’, or ‘classification cutoff’): patients with predictions above the cutoff are classified as positive; those below the cutoff as negative. Decision theory states that the odds at this threshold give the relative cost of a false positive compared to a false negative. For example, a patient who would accept treatment only if he had at least a 10% probability of disease, believes that not being treated when he needs it (false negative) is 9 times worse than being treated unnecessarily (false positive).

We can use this relationship to define the net benefit:

$$\text{Net benefit} = \frac{\text{True Positives}}{n} - \frac{\text{False positives}}{n} \left( \frac{p_t}{1-p_t} \right)$$

where  $n$  is the total number of patients in the study and  $p_t$  is the threshold probability. We can then compare the net benefits of various strategies – treat everyone, treat non-one,



treat according to a statistical model – and select the model with the highest net benefit. We create a decision curve by varying the threshold probability (e.g. 0 – 100%). Decision curves hence combine the mathematical simplicity of accuracy measures, such as sensitivity and specificity, with the clinical applicability of decision analytic approaches. We will illustrate the use of various performance measures using case studies of patients with cancer.

## References

1. Kattan MW (2003) Judging new markers by their ability to improve predictive accuracy. *J Natl Cancer Inst* 95: 634-635.  
<http://jnci.oxfordjournals.org/cgi/content/full/95/9/634>
2. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 26: 565-574.  
<http://mdm.sagepub.com/cgi/reprint/26/6/565>
3. Steyerberg EW, Vickers AJ (2008) Decision curve analysis: a discussion. *Med Decis Making* 28: 146-149. <http://mdm.sagepub.com/cgi/reprint/28/1/146>

# High-dimensional prognosis

Ulrich Mansmann

IBE, LMU, University of Munich, Germany

Starting point of my talk is the paper of Leo Breiman *Statistical Modeling: The Two Cultures* (2001, *Statistical Science*, Vol. 16, No. 3, 199–231). This paper describes the paradigm under which a high-dimensional prognosis will be successful.

- I will discuss the way how practical computational diagnosis and prognosis in molecular medicine complies with Breiman's ideas and were problems are met.
- While classical prognosis can rely on two basic methodological tools (*logistic regression* and the *proportional hazards model*), it is not obvious which model class should be chosen when working in a high-dimensional setting.
- The talk discusses problems of transportability of prognostic signatures in time, between institutions, and technologies (Justice et al.; *Ann Intern Med.* 1999; 130:515-524).
- Because of the complexity of the task, many people make up their own superstition on how to proceed and on what to look when developing a gene signature for prognosis or diagnosis. I will present a few examples for superstition in the *OMICS*, especially rituals which are related to a functional interpretation of a developed signature.
- I will discuss the problem of reproducible statistical analysis with specific focus on the development of high-dimensional classifier.
- Part of our own work on a prognostic gene signature for AML patients will be presented as example.

# **High dimensional data for supporting medical decision: are we meeting the challenge?**

**Elia Biganzoli, PhD**

Dept. of Medical Statistics and Biometry "Giulio A. Maccacaro",  
Istituto Nazionale Tumori - Università degli Studi di Milano  
Via Venezian 1, 20133 Milano (Italy)

In biomedical research, expectations concerning tailoring of therapies on a biological basis have been dramatically increased following the introduction of high throughput omic techniques that can simultaneously evaluate the mutation/expression of large numbers of genes. However, clinical decision-making still largely relies on classical information like pathological staging, grading and a limited number of clinical features, without clear indications on how to integrate the results of emerging techniques bioanalytical techniques.

Despite the strong expectations that biological markers could help in tailoring systemic treatments, the proper application of their information remains to be defined. A possible reason could be related to the large number of contrasting results. Unfortunately the advent of omic studies has not yet solved this issue. A concerning aspect of these studies, is their tendency in proposing new criteria for tumour sub-typing and prognostic classification “from the scratch”, without resorting to previous knowledge about the disease biology. This is potentially dangerous since their findings are actually based on a limited number of subjects with huge number of possibly inaccurate and/or imprecise measures. Moreover, few efforts have been done for the development of standardised criteria for the evaluation of the performances of diagnostic/prognostic classification criteria. Consequently, there seems to be an increasing gap between the resources employed for basic and translational research on biomarkers and actual patient benefits and overall social gain.

Until now, translational research focused on single biological markers which could putatively discriminate patients' prognosis or treatment response. Now, the limited power of single genes seems to be generally acknowledged, whilst highlighting the need for a classification based on optimised, quantitative analysis of many genes. However, it could be argued that different conventional biomarkers could still be useful to determine individual outcome and treatment response. The question could be whether traditional markers, if reliably measured and jointly analysed with suitable statistical methodologies would better support clinical decision, waiting for improving the reliability and reducing the costs of new molecular techniques. This would provide a benchmark to assess the information gain coming from future techniques.

The need for integrating exploratory studies addressing relevant biological issues possibly related to disease dynamics (*knowledge phase*) with subsequent prospective clinical studies (*decision phase*) must be carefully considered to exploit biological knowledge in a clinical context. It is unlikely that the physician would apply a decision

criterion without clearly understanding its biological and clinical bases, but this is the underlying risk of developing blind “black-box” classifications based on multiple markers, by means of sophisticated statistical techniques.

The talk will consider issues in the design and analysis of studies on biomarkers either classical or omic in a comparative perspective, showing how complex high-dimensional problems and the need of accounting for emergences even when limited features are considered.

Examples related to identification of distinct cancer profiles from routine and omic markers and disease dynamics will be presented. In breast cancer, gene expression profiling studies “rediscovered” a separation between tumour subtypes with steroid receptor absent and those with low or high levels of receptors, more distinct tumour subtypes may be identified within different levels of ER and PgR.

Exploring the complexity in high dimensional data coming from omics is a careful task when that underlying the role of the few acknowledged ones is still missed. The design and analysis shortcuts applied in most cases could overcome the benefits coming from the putative information of high dimensional data sets.

A rapid increase in the number of studies on markers identified by means of high throughput techniques at considerable expense is likely. It would therefore be relevant to promote the application of suitable study designs and statistical methods for the reliable assessment of data collected on biomarkers, either genomic or traditional, and a faster translation of basic research to medical decision-making.

## References

1. Biganzoli E, Boracchi P. *Old and new markers for breast cancer prognosis: the need for integrated research on quantitative issues*. Eur J Cancer. 2004; 40: 1803-1806.
2. Biganzoli E., Lama N., Ambrogi F., Antolini L., Boracchi P. *Prediction of cancer outcome with microarrays*. Lancet 2005; 365: 1683
3. Biganzoli E, Boracchi P, Coradini D, Grazia Daidone M, Marubini E. *Prognosis in node-negative primary breast cancer: a neural network analysis of risk profiles using routinely assessed factors*. Ann Oncol 2003; 14: 1484–1493.
4. Lama N., Boracchi P., Biganzoli E. *Exploration of distributional models for a novel intensity-dependent normalization procedure in censored gene expression data*, Collection of Biostatistics Research Archive The Berkeley Electronic Press. (October 2006) COBRA Preprint Series. Article 14. <http://biostats.bepress.com/cobra/ps/art14>. Accepted on “Computational Statistics and Data Analysis”
5. Taktak A.; Antolini L.; Aung H.; Boracchi P.; Campbell I.; Damato B.; Ifeakor E.; Lama N.; Lisboa P.; Setzkorn C.; Stalbovska V. And Biganzoli E. *Double-Blind Evaluation and Benchmarking of prognostic Survival Models in a Multi-Center Study*, Comput Biol Med. 37, 1108-1120
6. Ambrogi F., Biganzoli E. Querzoli P. Ferretti S. Boracchi P. Alberti S., Marubini E., Nenci I. *Molecular subtyping of breast cancer from traditional tumor marker*

*profiles using parallel clustering methods.* Clinical Cancer Research, 12: 781-790, 2006

7. Lusa L, Mcshane LM, Reid JF, De Cecco L, Ambrogi F, Biganzoli E, Gariboldi M, Pierotti MA. *Challenges in projecting clustering results across gene expression-profiling datasets.* J Natl Cancer Inst. 2007 Nov 21;99(22):1715-23.

# Lessons from meta-analyses of observational studies in epidemiology

**Simon G. Thompson**

MRC Biostatistics Unit, Cambridge, UK

Collations of individual participant data from epidemiological studies are increasingly common. Such extensive data can yield precise estimates of risk associations, help to resolve controversy, and enable exploration of heterogeneity. However, a number of statistical issues arise. These will be discussed with reference to the Fibrinogen Studies Collaboration, an individual participant data meta-analysis of the relationship between fibrinogen and coronary heart disease risk, based on 150,000 people in 31 prospective studies.

Heterogeneity across studies – in the distribution of fibrinogen, in the risk relationship, and in confounder effects – is a principal consideration. We favoured a two-stage random effects meta-analysis method, based on a sex-stratified proportional hazards model. Such analyses can be used to investigate the shape of the exposure-disease relationship, assess the effect of adjusting for confounders, and investigate interactions. When some of the desired confounders are not recorded in some studies, fully adjusted estimates encompassing information from all studies can still be derived. For interactions, it is important to separately identify within-study and between-study information. In particular, the assumption of proportional hazards is equivalent to the absence of a time interaction within studies. Measurement error needs to be addressed if the underlying aetiological relationship is to be estimated, but not if the focus is on risk prediction.

## *References*

- Fibrinogen Studies Collaboration. Plasma fibrinogen and the risk of major cardiovascular diseases and nonvascular mortality: an individual participant meta-analysis. *JAMA* 2005; **294**: 1799-1809.
- Pooling Project of Prospective Studies of Diet and Cancer (Smith-Warner SA *et al*). Methods for pooling results of epidemiologic studies. *Am J Epidemiol* 2006; **163**: 1053-64.

# **The challenges of a systematic review and meta-analysis of prognosis studies**

**Richard D. Riley\*, Willi Sauerbrei, Douglas G. Altman**

\* Centre for Medical Statistics and Health Evaluation  
University of Liverpool, England

An evidence-based approach to prognostic markers is needed<sup>1</sup>, as it is often difficult to ascertain the benefit of a marker from a single published study. In oncology, for example, hundreds of prognostic marker studies are published each year, often with small sample sizes and inconsistent findings. In principle at least, a clearer view is more likely to emerge by looking across multiple studies and by synthesising their results. For this purpose, a systematic review and meta-analysis are potentially important<sup>2</sup>, as they allow a transparent framework for identifying, evaluating and summarising an evidence-base. In this talk, I will discuss the challenges of performing a coherent and reliable evidence synthesis of prognostic marker studies. Using examples of published systematic reviews, I will show that they generally fail to provide clear evidence-based answers, and usually only highlight the paucity of good quality primary studies. Meta-analysis is also shown to be severely limited by heterogeneity and poor reporting across studies<sup>3</sup>, alongside the very real threats of publication bias<sup>4</sup> and selective reporting<sup>5</sup>. The talk will conclude on a more positive note by describing how we, the prognostic research community, have made progress toward addressing some of the inherent problems. Reporting guidelines<sup>6</sup>, availability of individual patient data, and a newly registered Cochrane Prognosis Methods Group<sup>7</sup> are just some of the encouraging steps being taken. Researchers are encouraged to continue such progress and work together, across multiple disciplines, to help realise the evidence-based use of markers in practice.

## References

1. Sauerbrei W, Holländer N, Riley RD, et al: Evidence-based assessment and application of prognostic markers: the long way from single studies to meta-analysis. *Communications in Statistics* 35:1333-1342, 2006
2. Altman DG: Systematic reviews of evaluations of prognostic variables. *BMJ* 323:224-8, 2001
3. Riley RD, Abrams KR, Sutton AJ, et al: Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *Br J Cancer* 88:1191-8, 2003
4. Kyzas PA, Denaxa-Kyza D, Ioannidis JP: Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer* 43:2559-79, 2007
5. Kyzas PA, Loizou KT, Ioannidis JP: Selective reporting biases in cancer prognostic factor studies. *J Natl Cancer Inst* 97:1043-55, 2005
6. McShane LM, Altman DG, Sauerbrei W, et al: REporting recommendations for tumor MARKer prognostic studies (REMARK). *J Natl Cancer Inst* 97:1180-4, 2005

7. Riley RD, Ridley G, Williams K, et al: Prognosis research: toward evidence-based results and a Cochrane methods group. *J Clin Epidemiol* 60:863-5; author reply 865-6, 2007



# Summary of main challenges and future directions

**Martin Schumacher**

*Institute of Medical Biometry and Medical Informatics,  
University Medical Center, Freiburg*

Starting with the situation about fifteen years ago where explicit criteria for confirmatory prognostic factor studies have been formulated<sup>1</sup> some methodological achievements will be highlighted. A prominent example is the problem of an “optimal” cutpoint<sup>2,3</sup> that is theoretically fully understood, published widespread in an accessible manner but still remains an issue in prognostic and predictive marker research. Thus although state-of-the-art methodology including reporting guidelines<sup>4</sup> is available the situation is still unsatisfactory, the medical literature is full of conflicting results and only few markers have proven to be clinically useful. In addition, it is often not fully appreciated that the predictive performance of prognostic models, when assessed in an unbiased way is at most moderate and predictions for individual patients are mostly poor<sup>5</sup>.

With the advent of high-dimensional genomic information in the last years we have seen a dramatic change of the old paradigm that the number of predictors investigated should be much smaller than the number of patients or events, respectively, in a study. High dimensionality and not only the insufficient or inadequate use of flexible statistical methods is now a major source of instability, overfitting and overoptimism of prognostic models. This leads to the requirements of stringent tools for their assessment<sup>6</sup> as well as of efficient algorithms for fitting sparse models that simultaneously consider high-dimensional genomic and “traditional” clinical information<sup>7</sup>.

An evidence-based assessment of prognostic models by meta-analytic techniques of multiple studies is further hampered by severe publication and reporting bias alongside with poor quality regarding conduct, analysis and reporting of single studies. Thus at present, it is often not possible to get a clear view on the prognostic and/or predictive relevance even of a single marker<sup>8</sup>. Based on the presentations given at the workshop and current status of methodological development I will try to summarize main challenges and future directions that an evidence-based approach to prognostic and predictive markers would require.

## References

1. Simon R, Altman DG. Statistical aspects of prognostic factor studies. *Br J Cancer* 1994; 69: 979-985.
2. Altman DG, Lausen B, Sauerbrei W, Schumacher M: Danger of using "optimal" cutpoints in the evaluation of prognostic factors. *J Nat Cancer Inst* 1994; 86: 829-835.
3. Schumacher M, Holländer N, Schwarzer G, Sauerbrei W: Prognostic Factor Studies. In: John Crowley and Donna Pauler Ankerst (Hrsg): *Handbook of Statistics in Clinical Oncology*. , 2. Auflage. Boca Raton, FL: Chapman & Hall /CRC, 2006; 289-333.

4. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, for the Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics: Reporting recommendations for tumor marker prognostic studies (REMARK). *J Nat Cancer Inst*, 2005; 97: 1180-1184.
5. Gerds TA, Cai T, Schumacher M: The performance of risk prediction models. *Biometrical J*, 2008; 50: 457-479.
6. Schumacher M, Binder H, Gerds T: Assessment of survival prediction models based on microarray data. *Bioinformatics*, 2007; 23: 1768-1774
7. Binder H, Schumacher M: Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *Bmc Bioinformatics*, 2008; 9: 10-19
8. Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HCW: Prognosis research: toward evidence-based results and a Cochrane methods group. *J Clin Epidemiol*, 2007; 60: 863-865.