Mathematisches Forschungsinstitut Oberwolfach

Report No. 06/2010

DOI: 10.4171/OWR/2010/06

Statistical Issues in Prediction: what can be learned for individualized predictive medicine?

Organised by
Leonhard Held (Zürich)
Robin Henderson (Newcastle upon Tyne)
Ulrich Mansmann (München)

January 24th – January 30th, 2010

Abstract. Error is unavoidable in prediction. And it is quite common, often sizable, and usually consequential. In a clinical context, especially when dealing with a terminal illness, error in prediction of residual life means that patients and families are misinformed about their illness, that they may take foolish actions as a result, and that they may be given inappropriate or needlessly painful treatments or denied appropriate ones. In meteorology, error in prediction of storm paths or extreme events can have devastating consequences. In finance and economics, major policy decisions are taken on the basis of predictions and forecasts. Rational approaches to reduce and assess error in prediction are presented. Ideas are introduced how to relate these statistical strategies with clinical and medical concepts in particular and how to integrate ideas from apparently different areas.

Mathematics Subject Classification (2000): 62N01, 62P10, 92C60.

Introduction by the Organisers

There is a recent resurgence of interest and activity in probability forecasting, which encompasses a wide range of sciences [1]. As far as medicine is concerned, this has been motivated in part by the more routine availability of individual–level genetic information and consequent potential for improved prognosis, diagnosis, and individualized therapy [2]. Further motivation has arisen from dramatic increases in power of the computationally intensive statistical methods needed to determine predictive probability distributions.

The goal of a good probabilistic prediction is to maximize the sharpness of the predictive distributions subject to calibration [3]. Calibration refers to the statistical consistency between the probabilistic forecasts and the observations, and is a joint property of the predictive distributions and the events that materialize. Sharpness refers to the concentration of the predictive distributions, and is a property of the forecasts only: The sharper the distributional forecast, the less the uncertainty, and the sharper, the better, subject to calibration. A number of alternative prediction accuracy measures have been suggested, for example skill scores [4] and the notion of predictiveness [5].

The workshop studied recent developments of tools to quantify the quality of a prediction strategy. These tools have additional impact on guidance in a wealth of applied statistical problems for count data [6], multivariate continuous data [7] and survival data [8]. They range from the evaluation of probabilistic forecasts to model criticism, model comparison and model choice.

Predictive distributions arise naturally in Bayesian modelling approaches. Therefore, the workshop looked at their state–of–the–art and explored interaction between Bayesian ideas and alternative approaches.

The intersection of genomics and medicine has the potential to yield a new set of molecular tools that can be used to individualize and optimize therapy as well as prognosis [9]. Specific prediction problems in individualized medicine are related to individual prognosis. Sharpness of individual prognosis is hampered by the intrinsic large uncertainty of point processes which are so far the methodological backbone of the predictive models. Biomarkers and their relevance for diagnosis, prognosis and clinical patient management pose new challenges on the development of statistical methods for joint models [10, 11, 12]. High–dimensional data from genetic screens are a further aspect to be integrated into the theoretical basis of prediction models for individualized medicine [13, 14].

The workshop also offered contributions in prediction strategies applied in the atmospheric sciences, economics, and finance. Their relevance for the solution of the problems to be handled in individualized predictive medicine was discussed.

The general aspects discussed during the workshop were producing and assessing probabilistic forecasts. Probabilistic forecasts arise natural in a Bayesian framework, taking automatically parameter uncertainty into account. A number of alternative likelihood–based approaches also exist [15]. Technically, probabilistic forecasts are often based on a random sample from the predictive distribution, due to the non–accessibility of the predictive distribution in a closed form. For example, in meteorology so–called ensemble forecasts are often used. These technical problems increase for multivariate forecasts where a closed form of the predictive distribution is rarely available. The selection of useful criteria for the assessment of the quality of probabilistic forecasts is of paramount importance. Proper scoring rules, for example the logarithmic or the Brier score, are accepted tools that address both calibration and sharpness of a prediction. The mathematical theory of proper scoring rules is related to the theory of convex functions, to information measures, and entropy functions [16, 3].

Further complications arise in the selection of the validation set in order to quantify the quality of a predictive model. Cross–validation is at the one end of the spectrum, while truly external validation sets are at the other extreme. As shown by Stone [17], the cross–validated logarithmic score is asymptotically equivalent to Akaike's information criterion (AIC), commonly used for model selection [18]. On the other hand, the Bayesian information criterion (BIC) can be viewed as an approximation to the log marginal likelihood, the sum of the one-step-ahead logarithmic scores [19, 20].

The specific aspects for individualized predictive medicine considered *individual prognosis for event times*, *joint modelling of biomarkers and event time data*, *high–dimensional data for predictive models*, as well as *global assessment of strategies in predictive medicine, clinical studies, and meta–analysis.*

Individual prognosis for event times: Prediction of time–to–event is particularly challenging yet fundamentally important, especially in medicine when there is a need to predict residual lifetime following diagnosis of a potentially terminal disease [8]. Complications include censoring of available data and heteroscedasticity. A variety of measures of predictive accuracy have been suggested e.g. [21, 22, 23] but none has been uniformly accepted. Furthermore, the availability of high dimensional genetic information has brought two unsolved challenges: how best to measure the additional value of genetic information over perhaps simpler to measure characteristics and how to deal with the well-known $p \gg n$ problem for predictive purposes as opposed to estimation and model selection. Dealing with high–dimensional covariate data for non–linear models is beginning to attract attention but numerous problems remain [24]. The usefulness of genetic information for individual level prediction was a core feature of the workshop.

Joint modelling of biomarkers and event time data: Biomarkers — measures of biological health — can be used as measures of disease progression and as prior surrogates for long term events. Examples are CD4 cell counts or other blood markers for HIV/AIDS e.g. [25] or reduction in telomere length as a measure of aging [26]. Methods of the joint analysis of the evolution of longitudinal biomarkers and time–to–event data are being developed [27, 28] but what is not yet clear is how best to exploit biomarker trajectories — not just current values — for predictive purposes. A comprehensive Bayesian approach for individual prediction based on longitudinal biomarker measurements is a promising theoretical approach [10].

High–dimensional data for predictive models: There is a series of studies which demonstrate the clinical value of prognostic models based on data measured by high–throughput biotechnologies. These models are in general derived by applying black–box model free algorithms which do not incorporate subject matter knowledge on the disease of interest [13]. The common feature of these algorithms is a mechanism of regularization which helps to handle the high–dimensionality of the measurements used to derive the prognosis [14]. Research on theoretical grounds of these regularization techniques is of high interest for mathematical statistics with eminent implication for practical applications. Besides using data to predict there is also usually an aims of obtaining information about the underlying data generation mechanism. The first successes in establishing clinically valuable gene signatures are not matched by an elucidation of the disease processes which produce the data. Theoretical guidance and appropriate statistical tools are needed to build the bridge between a gene signature and the functional aspects of the disease under consideration [29].

Global assessment of strategies in predictive medicine, clinical studies, and meta–analysis: The availability of gene signatures which predict specific risks or the response on therapeutic substances is the building stone of individualized medicine. They need a thorough assessment of their clinical relevance. The mathematical and statistical foundations of new design ideas for clinical trials have to be developed and implemented in biometrical practice [2]. Furthermore, there is a lack of mathematical and statistical tools for combining knowledge over a large literature on the relationship between specific biomarkers and response on specific substances. First examples of these meta–analyses are being published and started the discussion on methodological development [30].

References

[1] T. Gneiting, *Editorial: Probabilistic forecasting*, Journal of the Royal Statistical Society: Series A (Statistics in Society) **171** (2008), 319–321.

[2] R. Simon, *Roadmap for developing and validating therapeutically relevant genomic classifiers*, Journal of Clinical Oncology **23** (2005), 7332–7341.

[3] T. Gneiting, A.E. Raftery, *Strictly proper scoring rules, prediction, and estimation*, Journal of the American Statistical Association **102** (2007), 359–378.

[4] W. Briggs, D. Ruppert, *Assessing the skill of yes/no predictions*, Biometrics **61** (2005), 799–807.

[5] Y. Huang, M.S. Pepe, Z. Feng, *Evaluating the predictiveness of a continuous marker*, Biometrics **63** (2007), 1181–1188.

[6] C. Czado, T. Gneiting, L. Held, *Predictive model assessment for count data*, Biometrics **65** (2009), 1254–1261.

[7] T. Gneiting, L.I. Stanberry, E.P. Grimit, L. Held, N.A. Johnson, *Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds (with discussion)*, Test **17** (2008), 211–264.

[8] R. Henderson, N. Keiding, *Individual survival time prediction using statistical models*, Journal of Medical Ethics **31** (2005), 703–706.

[9] A. Dupuy, R. Simon, *Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting*, Journal of the National Cancer Institute **99** (2007), 147–157.

[10] J.M. Taylor, D.P. Ankerst, R.R. Andridge, *Validation of biomarker–based risk prediction models*, Clinical Cancer Research **14** (2008), 5977–5983.

[11] C. Proust-Lima et J.M.G. Taylor, *Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach.* Biostatistics, **10** (2009), 535–549.

[12] Y. Zheng, T. Cai, M.S. Pepe, W.C. Levy, *Time–dependent predictive values of prognostic markers with failure time outcome*, Journal of the American Statistical Association **103** (2008), 362–368.

[13] M. Schumacher, H. Binder, T. Gerds, *Assessment of survival prediction models based on microarray data*, Bioinformatics **23** (2007), 1768–1774.

[14] A. Benner, M. Zucknick, T. Hielscher, C. Ittrich, U. Mansmann, *High-dimensional Cox models: the choice of penalty as part of the model building process*, Biometrical Journal **52** (2010), 50–69.

[15] G.A. Young, R.L. Smith, *Essentials of Statistical Inference*, (2005), Cambridge University Press, Cambridge.

[16] L. J. Savage, *Elicitation of personal probabilities and expectations*, Journal of the American Statistical Association **66** (1971), 783–801.

[17] M. Stone, *An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion*, Journal of the Royal Statistical Society, Ser. B **39** (1977), 44–47.

[18] G. Claeskens and N. J. Hjort, *Model Selection and Model Averaging*, (2008), Cambridge University Press, Cambridge.

[19] A. P. Dawid, *Statistical theory: The prequential approach*, Journal of the Royal Statistical Society, Ser. A **147** (1984), 278–292.

[20] R. E. Kass, A. E. Raftery, *Bayes factors*, Journal of the American Statistical Association **90** (1995), 773–795.

[21] P.J. Heagerty, Y. Zheng, *Survival model predictive accuracy and ROC curves*, Biometrics **61** (2005), 92–105.

[22] S. Rosthoj, N. Keiding, *Explained variation and predictive accuracy in general parametric statistical models: the role of model misspecification*, Lifetime Data Analysis **10** (2004), 461–472.

[23] P. Royston, W. Sauerbrei, *A new measure of prognostic separation in survival data*, Statistics in Medicine **23** (2004), 723–748.

[24] S. Nygaard, O. Borgan, O.C. Lingjaerd, H.L. Storvold, *Partial least squares Cox regression for genome–wide data*, Lifetime Data Analysis **14** (2008), 179–195.

[25] X. Song, C.Y. Wang, *Semiparametric approaches for joint modeling of longitudinal and survival data with time–varying coefficients*, Biometrics **64** (2008), 557–568.

[26] T. DeMeyer, E.R. Rietzshel, M.L. De Buyzere, E. Van Criekinge, S. Bekaert, *Studying telomeres in a longitudinal population based study*, Frontiers in Bioscience **13** (2008), 2960–2970.

[27] R.M. Elashoff, G. Li N. Li, *A joint model for the longitudinal and survival data in the presence of competing failure types*, Biometrics **64** (2008), 762–771.

[28] P. Diggle, D. Farewell, R. Henderson, *Analysis of longitudinal data with dropout: objectives, assumptions and a proposal (with discussion)*, Applied Statistics **56** (2007), 499–550.

[29] M. Hummel, K.H. Metzeler, C. Buske, S.K. Bohlander, U. Mansmann, *Association between a prognostic gene signature and functional gene sets*, Bioinformatics and Biology Insights **2** (2008), 329–341.

[30] C.Y. Li, M. Mao, L. Wei, *Genes and (common) pathways underlying drug addiction*, PLoS Computational Biology **4** (2008), e2.