8/97

Oberwolfach Conference on Medical Statistics, 23.02.-1.03.1997

Mathematical Models for Diagnosis and Prognosis

Organizers: Mitchell H. Gail, Rockville; Helmut Schäfer, Marburg

The topic of the conference is of great medical relevance, and at the same time is a field with important and interesting theoretical developments during the last years. It was the aim of the conference to present these recent developments, to discuss possible applications of new statistical methods, and to compare them to standard procedures.

The participation of outstanding experts in the field on an international level showed the attractivity of this concept, and also once again showed the attractivity of the Oberwolfach institute.

The conference was structured in ten sessions, each session covering a special part of the topic (see the schedule on the next two pages). The conference was closed by an open discussion, chaired by Leo Breiman, and discussing the questions: What did we learn? What needs to be done?

The conference has enabled a fruitful scientific exchange between American and European biostatisticians and has given impulses for further cooperations. An outstanding journal in the field (Statistics in Medicine) has accepted to publish the contributions, after peer-reviewing, in a special issue, with the organizers of the conference as guest editors.

OBERWOLFACH CONFERENCE ON MEDICAL STATISTICS, 23.02. - 1.03.1997
„MATHEMATICAL MODELS FOR DIAGNOSIS AND PROGNOSIS"

Organizers: Mitchell H. Gail, Rockville, Helmut Schäfer, Marburg

MONDAY

Morning session 9.00-12.30
Practical and theoretical motivation. *Chair: Helmut Schäfer*

- Prognosis - what does the clinician associate with this notion? *Jürgen Windeler (Heidelberg)*
- Linking prognostic models to therapeutic strategies - implications for analysis and design. *Rolf Holle (Oberschleißheim)*
- Decision-analytic critique of some popular evaluation methods for diagnostic and prognostic tests. *Jörgen Hilden (Hellerup)*
- Regulatory issues for diagnostic products. *Susan Ellenberg (Rockville)*

Afternoon session 16.00-18.00
Designs for assessing absolute and relative risk. *Chair: Colin Begg*

- Population-based case-control designs. *Jacques Benichou (Rouen)*
- Strategies for cohort sampling and estimation of relative and absolute risk for sample cohort data. *Oernulf Borgan (Oslo)*
- Sample size considerations for the evaluation of prognostic factors. *Claudia Schmoor (Freiburg)*

TUESDAY

Morning session 9.00-12.30
Model building and stability. *Chair: Klaus Krickeberg*

- What do we mean by validating a model ? *Douglas Altman (Cambridge)*
- Instability and variable selection in prediction. *Leo Breiman (Berkeley)*
- Stability of recursive partitioning methods. *F. Dannegger (München)*
- Building stable prognostic models in breast cancer. *Patrick Royston (London)*
- EM estimation of diagnosis. *Carla Rossi (Rom)*

Afternoon session 16.00-18.00
Criteria for evaluating diagnostic/prognostic systems. *Chair: Frank Harrell.*

- Bayesian model comparison: Aposteriori distribution on the hierarchy of log-linear models with incomplete data. *Ulrich Mansmann (Berlin)*
- Error rate estimation under variable selection in linear discriminant analysis. *Jochem König (Homburg/Saar)*
- Explained variation in survival analysis. *Michael Schemper (Wien)*

Evening session 19.30-21.30
Complex modelling. *Chair: Richard Olshen.*

- A learning diagnostic system for anemia based on Boltzmann machines. *Wim Wiegerinck (Nijmegen)*
- A model for sequential classification. *Guenter Tusch (Hannover)*
- Multiple correspondence analysis in the face of higher-order interactions. *Johannes Fassbinder (Essen)*

WEDNESDAY

Morning session 9.00-12.30
Epidemiologic methods. *Chair: Sam Wieand.*

- A new strategy for evaluating the impact of epidemiologic risk factors for cancer. *Colin Begg (New York)*
- The genotype proband design for estimating genotype-specific risk of disease. *Mitchell Gail (Rockville)*
- Incorporating of family history in prognostic models. *Hans van Houwelingen (Leiden)*
- The evaluation of screening tests and the problem of verification bias. *Anja Gebler, A. (Bochum)*
- Analysis of risk factors in pharmaco-epidemiologic case-control studies. *Hans-Helge Müller (Marburg)*

THURSDAY

Morning session 9.00-12.30
Dichotomous outcomes and paired outcomes. *Chair: Mitchell Gail.*

- ROC analysis for the evaluation of diagnostic tests. *Gregory Campbell (Rockville)*
- Simultaneous confidence bands for ROC curves. *Katrin Jensen (Marburg)*
- Nonparametric procedures for evaluating the performance of repeated markers used to predict a dichotomous endpoint. *Sam Wieand (Pittsburgh)*
- Statistical methods for the evaluation of diagnostic measurements concerning paired organs. *Peter Martus (Erlangen)*
- On the misuse of artificial neural nets in oncology. *Martin Schumacher (Freiburg)*

Afternoon session 16.00-18.00
Survival outcomes. *Chair: Martin Schumacher*

- Problems in the analysis of prognostic factors with reference to the Cox model. *Maria Grazia Valsecchi (Milano)*
- Cox-model and CART for the development of classification schemes in survival data. *Willi Sauerbrei, (Freiburg)*
- Marginal versus conditional modelling in bivariate survival. *Robin Henderson (Lancaster)*

Evening session 19.30-21.30
Ordinal outcomes. Missing values. *Chair: Susan Ellenberg*

- Development of a clinical prediction model for an ordinal diagnostic outcome. *Frank Harrell (Charlottesville)*
- Allocatability and distinguishability of ordinally scaled outcomes with special reference to grading systems in medicine. *Uwe Feldmann (Homburg/Saar)*
- Stabilized multivariate tests and the treatment of missing values. *Siegfried Kropf, S. (Magdeburg)*

FRIDAY

Morning session 9.00-12.30
Time-varying prognostic states. *Chair: Maria Grazia Valsecchi*

- A multi-state model for bleeding episodes and mortality in liver cirrhosis. *Per Kragh Andersen (Kobenhavn)*
- Longitudinal data concerning renal function: summarizing and predicting. *Richard A. Olshen (Stanford)*
- Time-varying effects of prognostic factors. *Kurt Ulm (München)*
- Serial marker measurements for prognosis. *Bruce W Turnbull (Ithaca)*
- Optimal REMODELling of prognostic tree models applied to longitudinal nutrition, growth and urinary data. *Berthold Lausen (Dortmund)*

Afternoon session 14.00-16.00
Open discussion: What did we learn ? What needs to be done ? Other topics.
Chair: *Leo Breiman.*

What do we mean by validating a model?

Douglas G. Altman
ICRF Medical Statistics Group, Centre for Statistics in Medicine,
Institute for Health Sciences, Oxford, UK

There are several reasons why we might create a prognostic model. Two common reasons are to determine which variables affect patient prognosis or to predict outcome for individual patients. The latter seems a more sensible objective.

Having obtained a model, it is commonly advised that the model should be validated, but there does not seem to be a clear view of what exactly this means. In prognostic studies a reasonable definition is that the model works well (i.e. predicts well) for further patients - those whose data were not used in to obtain the model. This definition begs the question of what we mean by satisfactory.

Three levels of validation may be considered - internal (using the original data set), temporal (using further data from the same source) and external (using data from elsewhere). Internal validation includes methods such as bootstrapping and cross-validation. External validation provides the hardest test of a model and is strongly recommended. Several examples of published temporal and external validation studies are considered, including some where the model did not translate successfully.

In both temporal and external evaluations the question arises of how to compare the observed data with the predictions of the model. It may seem appealing to quantify the prediction errors for individual patients, but this is not simple for survival models. If the data are grouped, then it is not clear how to make the groups. Testing is possible but of questionable relevance. Global measures of fit (such as R^2) do not address individual prediction error.

While the methods of validation remain uncertain, it seems wise to recommend that the modelling process aims to develop a model with the best chance of fitting successfully for further patients.

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

A multi-state model for bleeding episodes and mortality in liver cirrhosis

Per Kragh Andersen
Copenhagen

In a clinical trial in patients with liver cirrhosis and esophageal varices, 286 patients were randomized in a 2 by 2 factorial design to either prophylactic treatment with propranolol, to prophylactic sclerotherapy, to both or to neither. The purpose was to reduce mortality and incidence of bleeding episodes. Thus, data consist of times of bleeding and death and various approaches to the modelling of these data will be discussed, including a competing risks model with two end-points "bleeding" and "death", a standard survival analysis neglecting information on bleeding episodes, and a multi-state model with three possible transitions: "no bleeding" to "bleeding", "no bleeding" to "death", and "bleeding" to "death". In particular, it will be studied how survival probabilities for patients with given characteristics may be estimated in the various models and the uncertainty of the estimators will be compared.

A new strategy for evaluating the impact of epidemiologic risk factors for cancer

Colin H. Begg, Memorial Sloan-Kettering Cancer Center, New York, USA

The population coefficient of variation in risk of cancer is proposed as a useful measure for evaluating the individual and collective impact on cancer incidence of known and suspected risk factors. It is demonstrated that the contribution to this measure of a single risk factor can be estimated nonparametrically using an empirical Bayes methods. The statistical properties of this estimator are examined using simulations. The categorization of a continuous risk factor can attenuate the apparent coefficient of variation substantially. The joint contribution of several risk factors will usually require statistical modelling to obtain an appropriate ranking of the likelihood ratios for the sample. The total coefficient of variation in risk is estimable, in theory, from the standardized incidence ratio of second primary cancers, providing a benchmark of total potentially explainable variation in risk. This quantity is a measure of the potential preventability of the disease in a population.

One- and two-stage estimation of exposure-specific incidence rates and absolute risk of breast cancer from a case-control study within a cohort

Jacques Benichou
University of Rouen Medical School, France

The Breast Cancer Detection and Demonstration Project (BCDDP) included a large cohort of women followed for incidence of breast cancer and for whom an initial case-control sample was drawn and standard risk factors obtained. An approach to obtain point and variance estimates of exposure-specific incidence rates as well as absolute risk from data like the BCDDP is reviewed. Point estimates were obtained by combining relative risk estimates from the case-control data and composite incidence rate estimates from the cohort variability, namely the variance of relative risk estimates and of baseline incidence rate estimates, as well as the covariance between the two, the latter term being based on implicit delta-method arguments.

In order to study the effect of mammographic features on breast cancer risk, a nested subsample of cases and controls in the BCDDP was drawn. Therefore, these completed data can be viewed as two-stage case-control data within a cohort, or as cohort data with two nested levels of missingness, since basic characteristics like age were measured on all members of the cohort, standard risk factors were elicited only in the initial case-control sample, and mammographic features were assessed only in the nested subsample of cases and controls. A pseudo-likelihood approach to estimating exposure-specific incidence rates and absolute risk based on the three types of missingness, namely, that for basic variables and standard risk factors, some levels (i) were omitted by design in the nested subsample of case and controls or (ii) were empty because of the sparsity of the data in that subsample. Estimates of standard errors are obtained from a parametric bootstrap. This approach appears to be relatively efficient when applied to the BCDDP data and contrasted with the original approach based on the cohort and initial case-control data only. It is flexible for modelling breast cancer rates and taking the special missingness features of these data into account.

References

Gail MH, Brinton LA, Byar DP et *al.* Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 81:1879-1886, 1989.

Benichou J and Gail MH. A delta-method for implicitly defined random variables. *Am Stat* 43:41-44, 1989.

Benichou J. A computer program for estimating individualized probabilities of breast cancer. *Comp Biomed Res* 26:373-382, 1993.

Benichou J and Gail MH. Methods of inference for estimates of absolute risk derived from population-based case-control studies. Biometrics 51:182-194, 1995.

Benichou J. A complete analysis of variability for estimates of absolute risk from a population-based case-control study on breast cancer. *Biometrical J* 37:3-24, 1995.

Benichou J, Byrne C and Gail MH. An approach to estimating exposure-specific rates of breast cancer from a two-stage case-control study within a cohort. *Stat Med* 16:133-151, 1997.

Strategies for cohort sampling and estimation of relative and absolute risk for sampled cohort data

Ørnulf Borgan, Institute of Mathematics, University of Oslo, Oslo, Norway

Estimation in proportional hazards models, like Cox's regression model, is based on a partial likelihood which compares the covariate values of a failing individual to those of all individuals at risk at the time of the failure. In large epidemiological cohort studies of a rare disease, these methods require the collection of information on exposure variables and other covariates of interest for all individuals in the cohort even though only a small fraction of these actually get diseased. This may be very expensive, or even logistically impossible. Cohort sampling methods, in which each failing individual (case) is compared to a small sample of controls from those at risk at the case's failure time, may give a substantial reduction in the resources (time and money) that need to be allocated to a study.

In the talk I will describe a general framework for such cohort sampling methods, incorporating classical nested case-control sampling (with simple random sampling of the controls) and counter-matched sampling (with stratified random sampling of the controls) as special cases. Further, I will review methods for estimation of the regression parameters and the integrated baseline hazard and indicate how one may estimate absolute risk for given time-dependent covariate histories. The methods will be illustrated by data on lung cancer deaths in a cohort of uranium miners from the Colorado Plateau.

Instability and variable selection in prediction

Leo Breiman
University of California, Berkeley, USA

Given data of the form T = {(y(n),x(n)), n=1, ...,N} where y(n) is a numerical dependent variable and x(n) a vector of predictor variables. We want to use T to from a predictor f(x,T) of future y-values given an x vector of predictor variables. Some methods for constructing predictors are inherently unstable - that is, small changes in T can lead to large changes in f(x,T). Examples are variable selection in linear regression, tree-structured methods like CART, and neural nets. Unstable methods have some undesirable characteristics which we point out. Perhaps the most important is that instability results in a loss of predictive accuracy.

There are two ways to get around instability. One is to use less unstable methods that accomplish the same purpose. This is illustrated by the garrote method in linear regression which is a more stable method of variable selection. Another way is this: perturb T repeatedly to get multiple versions T1, ..., Tk of T, grow predictors f(x,Tj), j=1, ...,k on the perturbed T and define a stabilized predictor as the average, at each x, of the f(x,Tj). When the perturbed T are formed as boostrap samples from T, then the procedure is called bagging. We show that bagging produces dramatic improvements in predictive accuracy and explain this on the basis of a bias-variance decomposition of the prediction error.

ROC Analysis for the Evaluation of Diagnostic Tests

Gregory Campbell
U.S. Food and Drug Administration, Rockville, USA

The methodology of Receiver Operating Characteristic (ROC) analyses is quite useful in the evaluation for diagnostic tests, including cancer genetic markers, clinical laboratory tests, imaging technologies and even artificial neural networks. For two populations of discriminatory interest, called normal and diseased groups, and a univariate continuous diagnostic test, the ROC plot is a graph of observed sensitivity versus (1-specificity) over the range of the diagnostic test. Fixed-width simultaneous confidence bands for the ROC plot using the Kolmogorov statistic are presented. A bootstrap sampling procedure is used to generate simultaneous bands for the ROC plot ignoring the values of the diagnostic test. It can also be applied guaranteeing a fixed distance for each level of the diagnostic test. The ROC plot is nonparametric, depending only on the combined ranks of the data; its area is the Mann-Whitney version of the Wilcoxon statistic and its maximal vertical deviation from the diagonal is the two-sample Kolmogorov-Smirnov statistic. A decision theoretic minimum risk is obtained as the intersection from above the ROC plot by a line segment with slope cFP(1-p)/(cFN)p, where p is the prevalence and cFP/cFN is the ratio of the costs of false positives to false negatives. Bootstrap confidence intervals for the threshold, sensitivity, and specificity are easily obtained. Discretization of continuous data can lead to nonadmissible rules if the chosen points are not on the convex hull of the ROC plot. A bootstrap procedure to compare the ROC plots of two diagnostic tests on the same patients is presented that preserves the Spearman correlation. For ratings data, the observed sensitivities and specificities and their associated predicted values on the fitted curve should be plotted. An approach for non-dichotomous states of nature is discussed; for example, left ventricular hypertrophy (LVH) is a continuity not a dichotomy. For d in the interval [0,1] representing degreee of membership in LVH, ROC plots are presented to compare two diagnostic tests on Framingham Heart data using the bootstrap. This suggests the use of the bootstrap to study the ROC behavior is small samples when the sampling is fixed for the total but not the number of normals and diseased persons.

Stability of recursive partitioning methods

Felix Dannegger
Inst. f. Med. Statistik u. Epidemiologie, TU München, München, Germany

Analyzing stability of recursive partitioning methods is a current area of statistical research. Identifying trees as a generally high variance procedure, Breiman (1994) proposes to use bootstrap aggregation or other perturb and combine procedures to increase stability and thus predictive performance. This talk discusses stability diagnostics and alternative stabilizing measures.

Using non-binary splits, it is proposed to avoid artificial dichotomization of effects, when the data don't warrant it. To limit increased computational burdens, methods for preselecting potential cut-points are discussed.

Additionally, using node level bootstrap samples, diagnostic tools such as confidence intervals are developed to assess stability of cut-points, and covariate importance rankings. Finally, using this information, a simple node level split averaging procedure is proposed to obtain a more stable tree, which can lead to a marked improvement in predictive performance when compared to a regular, single tree based predictor. While the improvement is not as drastic as for bagging, one does in contrast retain a single, easy to interpret tree structure.

Regulatory considerations for diagnostic products

Susan S. Ellenberg
Center for Biologics Evaluation and Research, FDA, Rockville, USA

A great variety of diagnostic products are regulated by the U.S. Food and Drug Adminstration, such as imaging agents, nuclear medicine scans, test kits for infectious diseases, mammography machines, genetic tests and many others. An important issue in evaluating such products is the extent to which their clinical utility must be demonstrated. The most definitive approach to assessing clinical utility is a clinial trial with subjects randomized to the use (or not) of the diagnostic text in managing their disease, with all other aspects of management standardized and clinical outcome as the primary endpoint. In many cases, clinical utility is self-evident without such a trial; for example, when one demonstrates accurate and reliable diagnosis of a treatable condition. Accuracy and reliability must, of course, be demonstrated for all diagnostic products prior to marketing approval; the standards for „accuracy and reliability" will vary from case to case, depending on such factors as the consequences of each type of diagnostic error, the existence (or practicability) of a gold standard, etc. Another important issue is whether the product should be assessed as an independent diagnostic tool, with readers who are blinded to subjects' clinical history, the results of other diagnostic tests, and who are not permitted to confer with other readers. The alternative would be to assess the product in a „real world" setting in which access to the above would be permitted. A variety of problems are routinely seen in applications for marketing approval of diagnostic products, including lack of any prospective analytical plan, inadequate and/or oversimplified analysis plans, use of an inappropriate patient population and inadequate consideration of the problem created by „indeterminate" outcomes. Safety concerns must also be considered. Most diagnostic products do not pose immediate physical safety issues, but a few do raise such concerns; for example, allergic reactions to contrast agents and radiation exposure from scans. These products also pose some specialized safety issues, such as potentially emotional response to test results (genetic testing, HIV tests, drug use tests) and consequences of inaccurate results, leading to inadequate or delayed treatment, or to unnecessary treatment.

Multiple Correspondence Analysis in the face of higher-order interactions

Johannes Faßbinder,
Institut für Medizinische Informatik, Biometrie und Epidemiologie,
Medizinische Einrichtungen der Universität Essen, Germany

Correspondence analysis was devised as a tool to depict similarities and differences between levels of a categorical variable by their frequency profile with respect to other variable categories. Another approach lies in interpreting the linear combinations gained this way as scores for ordinal variables which can – slightly modified – be used as scores in an ordinal-by-ordinal analysis. Additionally correspondence analysis is used to investigate relations between categories of different variables. The profiles are interpreted as vectors in Euclidean space and analysed by means of multidimensional scaling. The feasibility of displaying several variables in one diagram is vividly discussed, yet widely used as an explorative means. The extreme approach is to use correspondence analysis for the Burt matrix which can be viewed as the simultaneous analysis of all one-way interactions. Similarly to the analysis of continuous multivariate data by means of multidimensional scaling, higher interactions remain unconsidered. In practice, however, it often happens that the quality of interaction between two variables changes when the levels of a third variable is held fixed.

By feeding correspondence analysis with Simpson's Paradox as an extreme example of this effect, it will be demonstrated how interactions of higher order are masked when the analysis is restricted to the Burt matrix. As Correspondence Analysis gets only the marginal tables as input information, it cannot detect the Simpson's Paradox. However, necessary conditions for this effect can be verified very easily, inducing the necessity of investigation into higher-order interaction by the simultaneous analysis of the stratified table.

References

KR Gabriel, Biplot Display of Multivariate Categorical Data, with Comments on Multiple Correspondence Analysis, in WJ Krzanowski (ed.) *Recent Advances in Descriptive Multivariate Analysis*, Clarendon Press, Oxford (1995)

MJ Greenacre, *Theory and Applications of Correspondence Analysis*, Academic Press, New York (1984)

MJ Greenacre, Correspondence Analysis of multivariate categorical data by weighted least-squares, *Biometrika* (1988) **75**, 457-467

E Künzel, Ueber Simpsons Paradoxon, *Stochastik in der Schule* **1**, Dortmund (1991)

Distinguishability and allocatability of ordinally scaled outcomes with special reference to grading systems in medicine.

Uwe Feldmann
Homburg/Saar, Germany

A generalization of J.A. Anderson's (J.R.Statist.Soc.1984) concept of stereotype logistic regression is used to assess distinguishability and allocatability of ordered outcomes. The concept of generalized linear models is applied for model extension. Threshold parameters are introduced which allow for the assessment of allocatability of ordinal outcomes and to conduct a Bayes allocation into these outcome categories. Gain functions with specified utilities may also be applied in this context. Scale parameters allow for the assessment of distinguishabilty of ordered outcomes. The proposed canonical discriminant model remains identifiable for partially ordered and even for non ordered outcome categories and leads to nonlinear discriminant functions. Applications to grading systems in medicine are investigated, worldwide used under routine conditions for medical diagnostics and prognostics.

A comparison of strategies to estimate the penetrance of a rare dominant gene

Mitchell Gail
National Cancer Institute, Rockville, USA

With the advent of methods to detect specific mutations in DNA samples from study subjects, it becomes feasible to estimate the probability that a subject with a mutated gene will develop disease (the penetrance) by several study designs. Cohort designs, case-control designs (if the probability of disease in the population is known) and genotyped proband designs are feasible approaches. Genotyped proband designs involve obtaining a representative sample of volunteers (probands) who are to be genotyped, and observing the disease status (phenotype) and, possibly, the genotype of first degree relatives. It is shown for rare mutations of a dominant gene that required sample sizes can be larger for case-control than for cohort designs. Genotyped proband designs with non-diseased probands can require even larger sample sizes, but if one employs diseased probands, the numbers of required genotypes can be substantially reduced, especially if relatives are also genotyped. Genotyped proband designs are, however, subject to ascertainment biases that do not afflict cohort and case-control designs.

The evaluation of screening tests and the problem of verification bias

Anja Gebler
Ruhr-University Bochum, Germany

In order to estimate the sensitivity of a diagnostic test all subjects of the study population have to referred to the „gold standard" assessment to register all persons with a true-positive and a false-negative test result. If the diagnostic test is already used in clinical routine, it is often impossible, however, that all subjects with a negative test undergo the „gold standard" assessment.

Especially, this problem exist in the evaluation of screening procedures. In this situation the diagnostic test is used in a population with a small prevalence. The „gold standard" assessment is mo[redacted] an invasive diagnostic method and a great number of healthy people must be assessed for obtaining a small number of persons with false-negative test results. Due to this dilemma, the investigator wants to estimate the rate of false-negative test results in a subsample of the persons with negative test results. Is a subsample made by consideration of symptoms or other characteristics of the disease, the estimations are biased. This bias is called „verification bias" or „work-up bias".

BEGG and GREENES proposed a solution of adjustment, assuming the conditional independence of verification and disease probabilities. CHOI proposed an estimator for the sensitivity which is based on unbiased estimations from a random subsample. Both proposals make no restriction on the number or proportion of the selected population. But this is a problem in situations where the disease is rare and the diagnostic test has an unknown high sensitivity, like in the screening situation.

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Development of a clinical prediction model for an ordinal diagnostic outcome

Frank E. Harrell Jr.
Division of Biostatistics and Epidemiology, Dept. of Health Evaluation Sciences, School of Medicine, University of Virginia, Charlottesville, Virginia, USA

This paper describes the methodologies used to develop a prediction model to assist health workers in developing countries in facing one of the most difficult health problems in all parts of the world: the presentation of an acutely ill young infant. Statistical approaches for developing the clinical prediction model faced at least two major difficulties. First, the number of predictor variables, especially clinical signs and symptoms, is very large, necessitating the use of data reduction techniques that are blinded to the outcome. Second, there is no uniquely accepted continuous outcome measure or final binary diagnostic criterion. For example, the diagnosis of neonatal sepsis is ill-defined. Clinical decision makers must identify infants likely to have positive cultures as well as to grade the severity of illness.

In the WHO/ARI Young Infant Multicentre Study we have found an ordinal outcome scale made up of a mixture of laboratory and diagnostic markers to have several clinical advantages as well as to increase the power of tests for risk factors. Such a mixed ordinal scale does present statistical challenges because it may violate constant slope assumptions of ordinal regression models. In this paper we develop and validate an ordinal predictive model after choosing a data reduction technique. We show how ordinality of the outcome is checked against each predictor. We describe new but simple techniques for graphically examining residuals from ordinal logistic models to detect problems with variable transformations as well as to detect non-proportional odds and other lack of fit. We examine an alternative type of ordinal logistic model, the continuation ratio model, to determine if it provides a better fit. We find that it does not but that this model is easily modified to allow the regression coefficients to vary with cutoffs of the response variable. Complex terms in this extended model are penalized to allow only as much complexity as the data will support. We approximate the extended continuation ratio model with a model with fewer terms to allow us to draw a nomogram for obtaining various predictions. The model is validated for calibration and discrimination using the bootstrap. We apply much of the modeling strategy described in Harrell, Lee and Mark (Stat in Med 15:361-387;1996) for survival analysis, adapting it to ordinal logistic regression and further emphasizing penalized maximum likelihood estimation and data reduction.

Key Words: clinical prediction, diagnostic outcomes, variable clustering, data reduction, scaling, multivariable regression models, ordinal response, proportional odds model, continuation ratio model, penalized maximum likelihood estimation, shrinkage, predictive accuracy, validation, model approximation, nomogram, diagnosis, screening, infectious disease, imputation, differential penalization, bootstrap, regression spline

Conditional versus Marginal Modelling in Multivariate Survival

Robin Henderson
Lancaster University, UK

There are two popular approaches to the analysis of multivariate survival data. One is to assume a frailty model applies, under which a shared random effect acts multiplicatively on the individual hazards, but responses are otherwise conditionally independent. The other, more recent, approach is to adopt a marginal modelling strategy, estimating regression coefficients under an independence working assumption using models for the marginal failure time distributions without specifying the dependence structure, but using robust variance estimators.

The marginal approach is usually simpler, requires no parametric assumptions about the form of frailty, and allows inference about the average effect of a covariate across a population, but reveals nothing about the dependence structure. A frailty approach on the other hand allows both the effect of covariates at the individual level and the association between related responses to be studied, but may be harder to implement and is also restrictive in the type of dependence that can be modelled.

The two approaches are compared and the consequences of adopting a marginal approach when association arises through frailty are investigated. This leads on to discussion of the circumstances in which a marginal approach may be appropriate.

Decision-analytic Critique of Some Popular Evaluation Methods for Diagnostic and Prognostic Tests

Jørgen Hilden
University of Copenhagen, Norway

Maximization of posterior medical utility must underlie clinical choice, and departures from this principle must be conscious and well-argued. This tenet applies to therapy and hence to diagnosis, prognosis and patient counselling.

Diagnostic uncertainty and the information afforded by tests must therefore be conceptualized in terms of pre-post-differences in expected utility, be it in a mathematically stylized form.

Clinical practice calls for individual, and medical policy-making for aggregate utilities. In either case, marginal utility/cost ratios may be appropriate. I shall mention four implications for current evaluation practice:

(i) It is often argued that, just as sensitivity and specificity are „prevalence-free", any combined measure of diagnostic power ought to be „prevalence-free" measures for comparison of clinical tests. However, such measures have a quite restrictive functional form and never satisfy the common-sense requirement that mathematical utility be bounded because clinical utility is (never are infinite losses at stake in clinical practice).

(ii) Under the pertinent assumptions, the ROC curve does indeed hold the requisite information: diagnostic power is reflected by a functional which takes the form of a curve integral involving the local slope (exclusively). To summarize the curve in terms of the area thereunder is not a rational procedure, not even after „concavification". My original counterexample (MDM, 1991) will be supplemented by a different, more startling, one.

(iii) Kappa, and other correlation-like measures of diagnostic (oberserver dis) agreement, are inappropriate because they do not reflect how observer variation hurts. Utility-based measures must emphasize the trade-off between perfection and cost by studying the marginal utility/cost ratios of obtaining a second opinion or referring uncertain cases to a staff conference.

(iv) A final implication is that, for those who undertake methods development in the area of diagnostic evaluation, it is vital to understand the theory of strictly proper scoring rules - a concept still unknown to many statisticians. Briefly, such scoring rules encourage honesty. They serve as a safeguard against health care people cheating themselves, or being cheated, into buying an inferior diagnostic service or product.

Linking prognostic models to therapeutic strategies - Implications for analysis and design

Rolf Holle

GSF - National Research Centre for Environment and Health
Neuherberg, Germany

Prediction models in medicine may be useful for various reasons: they supply information for individual medical decisions, they may reveal causal factors which influence the course of disease and they may help to define differential treatment strategies. However, the clinical usefulness of prognostic models is rarely demonstrated in a strict way. We discuss the implications for analysis of prognostic models and design of efficacy trials when the aim is to establish prognosis-based therapies.

In order to define a differential treatment strategy prognostic information from a study of one of the following types must be available: a randomized trial of two treatment alternatives (case 1), an observational study with the same treatment for all patients (case 2), or an observational study with differential treatment according to an established prognostic model (case 3). In the first case a statistical analysis of interaction effects is required, whereas in the second case the subgroup of patients with very good or very bad outcome has to be defined whom a new, perhaps more intensive treatment should be offered. In the third situation the evaluation of prognostic factors will become difficult, because the predictive value of established factors will be underestimated if treatment decisions are influenced by them.

When a treatment strategy based on prognostic information has been defined, it has to be compared to standard treatment in a randomized clinical trial. There are various design options at hand, the most straightforward being the randomization of prognosis-based treatment versus standard treatment in all patients. However, it has the disadvantage that a considerable number of patients in both groups are treated identically which leads to a loss in statistical power. The second alternative therefore first applies the prognostic model to all patients and then restricts the randomized treatment comparison to those patients for which model based therapy differs from standard therapy. However, under a pragmatic point of view the first design has the advantage, that the acceptance and practicability of a model based treatment choice is tested as well.

The problems are illustrated by a study on patients with non-metastatic breast cancer. Nodal status is widely accepted as the most important prognostic factor in this disease and the decision to give adjuvant systemic therapy is mainly based on this factor. However, since lymph node dissection is associated with surgery-related morbidity it is important to look for other prognostic factors which are at least as predictive as nodal status but easier to assess.

Simultaneous Confidence Bands for ROC Curves

Katrin Jensen, Hans-Christoph Nürk, Helmut Schäfer
Institute for Medical Biometry, Philipps-University Marburg, Germany

Specificity (the probability of a correct decision for diseased individuals) and sensitivity (the probability of a correct decision for non-diseased individuals) of a quantitative diagnostic test depend on the selected cut-off point. The receiver operating characteristic (ROC) curve is generated by plotting sensitivity versus specificity as the cut-off point runs through the whole range of possible test values. The ROC curve is a graphical approach to visualize the performance of the quantitative diagnostic test.

In practice, the ROC curve of a quantitative diagnostic test has to be estimated from clinical data. Searching an optimal cut-off point it is necessary to take the statistical variability of this empirical ROC curve into consideration. Schäfer (1994) discussed *pointwise* confidence bands. If the task is to find an optimal cut-off point along the entire ROC curve, *simultaneous* confidence bands should be used.

In this contribution a nonparametric asymptotic procedure to construct *regional* simultaneous confidence bands for fixed regions of interest is proposed and discussed. This new approach is based on the convergence in distribution of empirical processes. Using Monte Carlo simulations the actual confidence levels under the model of normal and lognormal data are investigated for small sample sizes and various regions of specificity.

The actual confidence levels of the *global* simultaneous confidence bands are compared to those of Campbell (1994). Campbell presented three methods to construct global simultaneous confidence bands: The first method of Campbell (1994) combines two separate confidence intervals for specificity and sensitivity. The second method uses confidence bands which are obtained by the Kolmogorov-Smirnov two-sample statistics and are constructed for the special case 'sensitivity + specificity = 1'. Campbell's third method is based on bootstrap resampling.

Campbell, G. (1994). *General Methodology I, Advances in Statistical Methodology for the Evaluation of Diagnostic and Laboratory Tests.* Statistics in Medicine, Vol. 13, 499-508.

Schäfer, H. (1994). *Efficient Confidence Bounds for ROC Curves.* Statistics in Medicine, Vol. 13, 1551-1561.

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Error Rate Estimation after Variable Selection in Linear Discriminant Analysis

Jochem König
Homburg/Saar, Germany

It is a well known fact in discriminant analysis that error rates are optimistically biased when estimated from the training sample in a naive manner. Corrections are well established: Lachenbruch's hold-one-out method, McLachlan's asymptotic formulae under normal assumptions. When data driven variable selection is performed, the variable subset , the allocation rule and the error rate estimator are subject to random fluctuation. Hence an optimistic bias is induced even on corrected error rate estimators.

As a remedy to the problem double crossvalidation (Stone), bootstrap methods and special smoothed estimators have been proposed.

In this talk the situation of two normally distributed groups with known trivial covariance structure is investigated more thoroughly. Here the selection procedure, the allocation rule and the estimated error rate (EER) depend on the vector D of mean differences only. The following results are presented

1. The norm of the post selection difference vector (usefull for selection-ignoring parametric EERs) induces a so called selection norm on the full sample space.
2. The selection procedure Sp(q:p), that selects q variable from p on the basis of the absolute mean differences, induces a selected normal distribution on the space of the selected variables. It is calculated for simple parameter sets.
3. A central and non-central χ^2 - distribution are derived as the distribution of the norm of selected normal variates.
4. The distribution of the actual error rate (true error rate of the estimated allocation rule) is derived from a non-central t-distribution and an estimator is presented that is median-unbiased when no variable selection is performed.
5. The distribution of the unbiased estimator under variable selection is calculated for the simple parameter sets $\Delta = 0, \Delta = (\delta,\ldots,\delta), \Delta = (\delta,0,\ldots,0)$.
6. Sample sizes necessary to bound estimation error are given for the unfavourable case $\Delta = 0$.

It is concluded, that bias can be strong, and that a validation sample has to be taken for well founded error rate estimation.

The problem of estimating the error rate on the training sample still remains and the object of further investigation is to derive correct error rate estimators. Some contributions are presented:

1. An adequate estimator has to use more information than the post selection norm and the number of variables.
2. A conditionally median unbiased estimator (and confidence interval) for the selection procedure Sp(1:2) is presented.
3. Another approach is presented for the selection of variables based on a fixed cutpoint.

It is an open question,

- whether confidence statements can be related to the selected variable subset and
- whether exact confidence intervals can be constructed.

References

1. König, J (1991). Fehlerratenschätzung in der Diskriminanzanalyse bei Variablenselektion. Dissertation, Univ. Heidelberg.
2. König,J. (1988). Error rate estimation in discriminant analysis in the presence of variable selection. In: P.L. Reicherts, D.A.B. Lindberg (eds.). Expert systems and decision support in medicine. Springer, Berlin.

Use of Stabilized Multivariate Procedures in Tests and Prediction

Siegfried Kropf and Jürgen Läuter
Institute of Biometrics and Medical Informatics
Otto von Guericke University Magdeburg, Germany

Small samples of high-dimensional data are typical for many analyses in medicine. This is just the situation, where traditional multivariate procedures run into trouble. Tests perform poor with respect to power, error rates in discriminant analyses are high, the prediction in regression analyses is bad, and so on.

The inclusion of model-building steps such as selection of variables or other methods of reducing the dimension gives biased results in the subsequent analyses. Special proposals for multivariate tests with so-called multiple endpoints (O'Brien, 1984, Wei and Lachin, 1984, among others) are based on heuristical or asymptotical arguments. Their null-distribution is known only approximately and the power for small samples is low for some of the proposals.

The class of tests, proposed by Läuter (1996) and Läuter, Glimm and Kropf (1996) provides both: The type I error of the tests is kept exactly, and they enable for a variety of stabilizing steps. The effect of the stabilization with respect to the power depends on a good characterization of the practical problem, which is a real challenge to biometricians.

In the practical execution, the tests run in two steps. In the first one, scores are derived from the high-dimensional data. Then these scores are analyzed in the second step with standard procedures. Special rules for the scores ensure, that the final analyses are exact despite of the preprocessing. The theoretical background is given by the theory of spherical distributions (Fang and Zhang, 1990).

The use of these scores is not restricted to tests. The scores are understood as latent factors, which determine the investigated biological effects. Thus, they can be applied also in prediction and classification, when a large number of explanatory variables is suspected to reflect only a small number of essential factors as, e.g., in neurobiological problems.

The treatment of missing values can be included into these considerations (Kropf, Läuter and Glimm, 1997). The high-dimensional situation has to be taken in account for this problem, too.

References

Fang, K.-T.; Zhang, Y.-T. (1990). *General Multivariate Analysis*. Science Press Beijing and Springer-Verlag Berlin Heidelberg.

Kropf, S.; Läuter, J.; Glimm, E. (1997). Stabilized Multivariate Tests - the Inclusion of Missing Values. *Biometrical Journal*, in press.

Läuter, J. (1992). *Stabile Multivariate Verfahren. Diskriminanzanalyse, Regressionsanalyse, Faktoranalyse*. Akademie Verlag, Berlin.

Läuter, J. (1996). Exact t and F Tests for Analyzing Studies with Multiple Endpoints. *Biometrics* **52**, 964-970.

Läuter, J.; Glimm, E.; Kropf, S. (1996). New Multivariate Tests for Data with an Inherent Structure. *Biometrical Journal* **38**, 5-22.

O'Brien, P.C. (1984). Procedures for Comparing Samples with Multiple Endpoints. *Biometrics* **40**, 1079-1087.

Wei, L.J.; Lachin, J.M. (1984). Two-Sample Asymptotically Distribution-Free Tests for Incomplete Multivariate Observations. *JASA* **79**,Theory and Methods Section, 653-661.

Optimal REMODELling of prognostic tree models applied to longitudinal nutrition, growth and urine data

Berthold Lausen
Dortmund, Germany

The robust modelling of longitudinal prognostic factors for longitudinal response variables is the main topic of my paper (cf. Lausen 1997d). The partitioning and recursive method of P-value adjusted regression trees (Lausen, Sauerbrei & Schumacher 1994) fits homogeneous subgroups, interaction terms and factors which were measured on different scales. A restrictive assessment of the confidence or of the stability of the prognostic tree model is provided by the combination of the P-value adjusted split criterion maximally selected rank statistic (Lausen & Schumacher 1992, 1996; Altman et al. 1994) and the conservative Bonferroni inequality regarding the effect of variable selection (Lausen, Kersting & Schöch 1997).

The additive combination of generalized linear models and tree models is a promising idea (REMODEL-method, Lausen 1997; Loh 1991). Often it makes sense to define the variable selection problem of the linear term for a finite set of model spezifications S_{lin}. For example this set S_{lin} consists of all model terms considered in an analysis using fractional polynomials (cf. Royston & Altman 1994) or S_{lin} could be defined as the models of an forward/backward variable selection procedure.

A straightforward definition of optimal REMODELling (Lausen 1997c) is given by defining a goodness of prognosis criterion (GOPC), for example sum of squared differences of estimated and observed response values. One possibility for validation of tree models is given by a random splitting of the data in a learning sample L and a test sample T (cf. Breiman et al. 1984). For each member of S_{lin} a REMODEL-prognostic tree is computed for given values P_{stop} und n_{min} on the data of the learning sample L. Afterwards the GOPC value is computed by the data of the test sample T. The overall minimum GOPC value defines the Optimal REMODELling of the prognostic tree model. Resampling, e.g. cross-validation, of the minimum GOPC value may be used for bias correction of the estimated GOPC of the final prognostic tree model. Compared to other new developments in the literature regarding tree regression my suggestion may be viewed as an algorithmic specification of a state of the art idea to fit, apply and interprete prognostic models (cf. also Breiman 1996; Schumacher, Holländer & Sauerbrei 1996). This argument is in contrast to suggestions (e.g. Chipman, George & McCulloch 1996; Sutton 1991) which provide a theory without answering applied problems like the quest of prognostic models in medicine.

The application of prognostic tree models is supported by various graphical display techniques of the tree model (Dirschedl 1992; Lausen, Kersting & Schöch 1997). The approach is applied to calcium excretion curves of healthy German children and adolescents (Kersting et al. 1997; Manz, Lausen & Kehrt 1997; Lausen 1997ab). The computed prognostic tree models show results which allow new insights in the calcium balance of children and adolescents. Furthermore the central statements of recent papers on the effect of sodium on calcium excretion have to be newly discussed (Matkovic et al. 1995; Antonius & McGregor 1996). These preliminary results of the first interim analysis of the Dortmund longitudinal study (DORLOS) show, that the inter-individual variation, the relative growth velocity and the dietary protein intake are the dominating prognostic factors of the urinary calcium excretion of healthy children and adolescents. Less important prognostic dietary factors are magnesium, sodium, calcium and phosphorus intake. Moreover, individual long term calcium intake is the dominating dietary factor during the period of most rapid growth.

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

References

1. **Altman, D., Lausen, B., Sauerbrei, W., and Schumacher, M. (1994).** Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors, *J Nat Canc Inst 86*, 829-835.
2. **Antonius, T.F.T., and MacGregor, G.A. (1996).** Salt - more adverse effects, *The Lancet 348*, 250-251.
3. **Breiman, L. (1996).** *Bagging predictors*, 20th Annual meeting Gesellschaft für Klassifikation e.V., March 1996, Freiburg.
4. **Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984).** *Classification and regression trees.* Wadsworth, Monterey.
5. **Chipman, H.A., George, E.I., and McCulloch, R.E. (1996).** *Bayesian CART.* Paper at the ASA meeting 1996.
6. **Dirschedl, P. (1992).** Tutoriell: Klassifikationsbäume - Grundlagen und Neuerungen, in: Fleischer, W., Nagel, M., und Ostermann, R. (Hrsg.). *Interaktive Datenanalyse mit ISP.* Westarp-Verlag, Essen, 15-30.
7. **Kersting, M., Sichert-Hellert, W., Lausen, B., Alexy, U., Manz, F., und Schöch, G. (1997),** Energy and macronutrient intakes of 1 to 18 year old German children and adolescents, submitted.
8. **Lausen, B. (1997a).** Generalized regression trees applied to longitudinal nutritional survey data, in: Klar, R., and Opitz, O. (eds.). *Classification and Knowledge Organisation.* Springer-Verlag, Heidelberg, 467-474.
9. **Lausen, B., Kersting, M., Manz, F., und Schöch, G. (1997b),** Inter-individual variation and growth velocity are the dominating prognostic factors of calcium excretion curves in healthy German children and adolescents - An early report on results of the first interim analysis of the Dortmund Longitudinal Study (DORLOS) (1985-1996), manuscript.
10. **Lausen, B. (1997c).** REMODELling of prognostic tree models applied to longitudinal nutrition, growth and urine data, manuscript.
11. **Lausen, B. (1997d).** Robuste prognostische Baummodelle und deren Anwendung in der Ernährungsmedizin, manuscript submitted to the gmds-97, 9/1997 University of Ulm.
12. **Lausen, B., Kersting, M., and Schöch, G. (1997).** The regression tree method and it's application in nutritional epidemiology, *Informatik, Biometrie und Epidemiologie in Medizin und Biologie 28*, 1, 1-13.
13. **Lausen, B., Sauerbrei, W., and Schumacher, M. (1994).** Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales, in: Dirschedl, P., and Ostermann, R. (eds.). *Computational Statistics.* Physica-Verlag, Heidelberg, 483-496.
14. **Lausen, B., and Schumacher, M. (1992).** Maximally selected rank statistics, *Biometrics 48*, 73-85.
15. **Lausen, B., and Schumacher, M. (1996).** Evaluating the effect of optimized cutoff values in the assessment of prognostic factors, *Computational Statistics and Data Analysis 21*, 307-326.
16. **Loh, W.-Y. (1991).** Survival modeling through recursive stratification, *Computational Statistics and Data Analysis 12, 295-313.*
17. **Manz, F., Lausen, B., and Kehrt, R. (1997).** Urinary calcium excretion in healthy children and adolescents, submitted.
18. **Matkovic, V., Ilich, J.Z., Andon, M.B., Hsieh, L.C., Tzagournis, M.A., Lagger, B.J., and Goel, P.K. (1995).** Urinary calcium, sodium, and bone mass of young females, *Am J Clin Nutr 62*, 417-425.
19. **Royston, P., and Altman , D.G. (1994).** Regression using fractional polynomials of continuous covariates: Parsimonious parametric modeling (with discussion), *Applied Statistics 43*, 3, 429-467.

20. **Schumacher, M., Holländer, N., & Sauerbrei, W. (1996).** Resampling and cross-validation techniques: A tool to reduce bias caused by model building?, submitted.
21. **Sutton, C. (1991).** Improving classification trees with simulated annealing, in: Keramidas, E. (ed.), *Proceedings of the 23rd Symposium on the Interface*, Interface Foundation of North America.

Baysian model comparison: Aposteriori distribution on the hierachy of log-linear models with incomplete data

Ulrich Mansmann, Institut für medizinische Statistik, Freie Universität Berlin, Universitätsklinikum Benjamin Franklin Berlin, Germany

We study the assessment of the quality of different diagnostic test devices with respect to the detection of well defined disease features. In the most simple situation different diagnostic tests are performed simultaneously on a patient and the results are recorded as binary data (positive/negative). The assessment of the quality is given in terms of *sensitivity* and *specificity*. To compute estimates of these measures of quality, information about the true disease state, *gold standard*, is needed. Generally the availability of the gold standard is restricted to certain patients with respect to the outcomes of the diagnostic tests applied. E.g. for a patient with consistently negative test results the gold standard information will not be available.

R. Sundberg [1] presents a Maximum-Likelihood theory which can be used to calculate estimates of *sensitivity* and *specificity* for the situation described above. The estimates depend on the model structure on which the calculation is based on. Model selection has to be performed in order to find unbiased estiamtes of the true quality of the diagnostic test devices considered. We could show that model selection based on the Bayesian Information Criterion performes better as model selection based on ML-Test or AIC.

If in addition covariates describing features of the patient are recorded and the influence of the covariates on the test quality is studied, model selection has to be performed in a very large hierachy of models. We developed a Gibbs sampler in order to select the appropriate log-linear model in such a large model hierachy. At the model sites in the hierachy independent Gibbs sampler are running to estimate the relevant model coefficients. A random walk will be performed on the model hierachy graph. The jumps of the walker are influenced by the state of the site's Gibbs sampler. This produces a random walk on a random enviroment whose asymptotic measure equals the *aposteriori distribution* on the hierachy of log-linear models with incomplete data. This approach differs from the proposal given by Green [2].

[1] Sundberg R., *Maximum Likelihood Theory for incomplete data from an exponential family*, Scand. J. Statist, 1, 49-58,1974
[2] Green PJ., *Reversible jump MCMC computation and Bayesian model determination*, Biometrika, 82, 711-32, 1995

Statistical methods for the evaluation of diagnostic measurements concerning paired organs

Peter Martus
University of Erlangen-Nürnberg, Germany

The evaluation of diagnostic procedures for paired organs, e.g. in ophtalmology, requires specific methodological approaches. (1) Marginal models for the probability of disease, knowing the diagnostic variables, are constructed. The dependency of measurements on the same patient is treated to maximum likelihood methods or generalized estimating equations. (2) If there is specific interest in the interaction of both eyes, especially in intraindividual differences, the polychotomous logistic regression model, with realizations ++, +-, -+, -- may be used. By examining several contrasts of the underlying parameters, this model may be simplified to the Rosner model or even to the independence model using a stepwise procedure.

In the talk pros and cons of the several approaches are discussed, referring to different clinical situations. In a screening situation primarily the most affected eye is of interest. In the clinical situation acute, typically monocular disease has to be distincted from chronical diseases like most of the glaucomas with irreversible damages and binocular manifestation. Depending on the sampling scheme the stochastic nature of the diagnostic measurements has to be taken into account. The proposed methods may also be applied to prognostic studies.

The results are illustrated using data of the Erlangen Glaucoma study. Important diagnostic measurements for this disease are the visual field testing and the assessment of optical nerve damage. In both of these procedures spatial data are obtained. This second level of dependency has to be treated by using methods of data reduction.

Analysis of Risk Factors in Pharmaco-Epidemiologic Case Control Studies

Hans-Helge Müller and Helmut Schäfer
Institute of Medical Biometry, University of Marburg, Germany

When an association between drug exposure and an adverse event has been established in pharmaco-epidemiologic case-control studies, the additional relevance of risk factors is often studied. A recent example is given by the international case-control studies on oral contraceptives and venous thrombosis. Risk factors such as body-mass-index, smoking, age, and previous use of other oral contraceptives were investigated in these studies. In consequence, regulatory authorities have formulated restricted indications for oral contraceptives of the third generation in special subgroups defined by some of these factors. The relevance of genetic factors, like the factor V Leiden mutation, is discussed controversially [1] [2].

In the present contribution, the question is examined whether the results from the published case control studies actually support these regulatory decisions and what these data can contribute to the discussion about genetic factors. Conventional statistical concepts like interaction terms in the risk model modelling the interaction between the potential risk factor and intake of the drug have been mentioned (but not applied) to analyse these factors in case control studies [2]. It is shown that these types of analysis do not meet the problem. A decision-theoretic approach is proposed. A utility parameter is defined which quantifies the preventive effect (risk reduction) that can be achieved by selective prescription of the drug depending on the presumed risk factor. The proposed parameter can be estimated in the case control design. Methods of statistical inference (asymptotic tests and confidence intervals) for this parameter are developped. The approach is applied to the example of oral contraceptives and venous thrombosis.

References:

[1] Bloemenkamp et al., Enhancement by factor V Leiden mutation of risk of deep vein thrombosis associated with oral contraceptives containing a third-generation progestagen. The Lancet 346, 1995, p. 1593.
[2] Letters to the editor and discussion on [1]. The Lancet 347, 1996, p. 396.

Longitudinal data concerning renal function: summarizing and predicting

Richard A. Olshen
Stanford University, Stanford, California, USA

The talk was a report on a collaboration with Jonathan Buckheit, Kristina Blouch, and Bryan Myers of Stanford. We studied glomerular filtration longitudinally for 36 - 120 months in patients undergoing treatment for diffuse, proliferative lupus nephritis. Attention focused upon flomerular filtration rate (GFR) and several other related quantities. Patients were divided into progressors (9) and non-progressors (9) according to the presence or absence of an irrevocable decline in GFR over time. Early identification of who will and will not progress has important implications for medical management of the patient. We adapted techniques pioneered by B. Silverman and colleagues on the one hand, and by J.O. Ramsey and C.J. Dalzell on the other to our irregularly measured data. Thus, we found a new approach to regularizing regression splines for each variable and subject, imputed data to desired grids, and proceeded to find principal curves for studying variables marginally and canonical curves for looking at joint behavior, too. We were able to discriminate between progressors and non-progressors by cross-validated linear discriminant analysis applied to the logarithms of fractional albumin and IgG clearance curves from the first 36 months of followup; the coefficients make biological sense. Covariation among pairs of GFR and two other curves also inform biological understanding. Simulations confirm some aspects of the methodology.

See http://www-isl.stanford.edu/~gray/olshenWWW/lupus.html

EM estimation of diagnosis

Carla Rossi
Università degli Studi di Roma, Italy

The diagnosis/prognosis model has already been introduced in previous papers, each of them dealing with a particular aspect, such as Bayesian expert systems, latent diagnostic variables, classificiation of survival data via regression trees and more. The present contribution start from the general mathematical diagnostic/prognostic model and is particularly focussed on the specific aspect of the estimation of the risk function within diagnostic classes, with particular reference to the latent variable (the unobservable diagnosis) regarded as a missing value. An EM procedure is derived and proposed to solve the estimation problems. An application to AIDS survival data (detection of long survivors) is presented to show the model in action.

Building stable prognostic models in breast cancer

Patrick Royston
Royal Postgraduate Medical School, University of London

It may be argued that to be useful to clinicians, prognostic indices must be derived from accurate models developed using appropriate datasets. We show that fractional polynomials, which extend ordinary polynomials by including non-positive and fractional powers, may form the basis of such models. We show how to fit fractional polynomials in several covariates and propose ways of ensuring the resulting models are parsimonious and consistent with basic medical knowledge. The methods are applied in a dataset of patients with node-positive breast cancer. We adopt bootstrap resampling to assess model stability, and compare our new approach with conventional modelling methods which apply stepwise variable selection to categorised covariates. We conclude that fractional polynomial methodology can be successful in generating simple and appropriate prognostic models.

Use of the Cox-model and CART for the development of classification schemes in survival data

Willi Sauerbrei
Institute of Medical Biometry and Informatics, University of Freiburg, Freiburg, Germany

Classification schemes are important for the diagnosis and treatment of patients. In breast cancer, several schemes are published, but with the exception of the Nottingham Prognostic Index none of them seems to be sufficiently validated in the literature. We used a study in patients with node negative breast cancer of the German Breast Cancer Study Group (GBSG) to validate some of the literature proposals and to develop new ones (Sauerbrei et al 1997). In contrast to the NPI, two other classificaton schemes for node negative patients did not discriminate the different groups as proposed in the original papers.

Concerning survival data, two common methods for the development of classification schemes are based either on Cox regression models or on classification and regression trees (CART). Although both approaches are conceptionally very different, both share the common problem of overoptimism concerning the predicitve ability, if the 'final' models are based on too many factors. Despite of many criticisms stepwise selection methods are the most popular approaches to select the final regression model. The predefined nominal selection level can be used as the parameter to control model complexity. Furthermore additional constraints based on the specific aim of a study may influence the complexity of regression models and trees. Using crossvalidation, we will discuss the importance of aspects of model and tree complexity and stability for the development and validation of classification schemes.

LAUSEN B, SAUERBREI W, SCHUMACHER M (1994): Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In: Dirschedl R, Ostermann R (ed.), Computational Statistics. Physica-Verlag. Heidelberg, 483-496.

SAUERBREI W, HÜBNER K, SCHMOOR C, SCHUMACHER M (1997): Validation of existing and development of new prognostic classification schemes in node negative breast cancer, Breast Cancer Research and Treatment, 42, 149-163.

VERWEIJ PJM, VAN HOUWELINGEN HC (1993): Cross-validation in survival analysis, Statistics in Medicine, 9, 487-503.

SAUERBREI W (1997): On the development and validation of classification schemes in survival data. In: Klar R, Opitz O (ed.), Classification, data analysis and knowledge organisation, Springer, to appear.

Explained variation in survival analysis

Michael Schemper
University of Vienna, Austria

Current knowledge of the course of chronic diseases in individual patients can be quantified by measures of explained variation. Several of such measures have been suggested for the Cox proportional hazards regression model. These measures can be categorized into three classes which correspond to three different definitions of multiple R^2 of the general linear model. The performance of these measures was compared in an empirical study and they were classified by their adherence to a set of criteria which should be met by a measure of explained variation for survival data.

It is suggested that currently there is no uniformly superior measure, particularly as the concepts of either uncensored or censored populations may lead to different choices. In particular the approaches by Kent & O'Quigley (1988), Korn & Simon (1990), Schemper (1990), Schemper (1992) and Schemper & Kaider (1996) are reviewd and the underlying concepts clarified.

Sample size considerations for the evaluation of prognostic factors in survival analysis

Claudia Schmoor, Freiburg, Germany

If the role of a new prognostic factor shall be investigated a careful planning of an appropriate study is required. This includes an assessment of the power of the study in terms of sample sizes. An adequate analysis of the independent prognostic effect of a new factor has to be adjusted for the existing standard factors. With survival time as endpoint this will usually be done with the Cox proportional hazards model. Sample size and power formulae in survival analysis have been developed by Schoenfeld for randomized treatment comparisons. In the analysis of prognostic factors the covariates included are expected to be correlated with the factor of primary interest. In this situation, the existing sample size and power formulae may not be applied. In this talk, first Schoenfeld's formula is extended to the situation that a correlated factor is included in the analysis. The validity of the resulting approximate asymptotic formula is investigated for its asymptotic behaviour by numerical intgration and for its finite behaviour by simulation. Second, an approximate formula for sample size and power is provided to detect an interactive effect between the interesting and a second correlated factor. This extents the formula for independent effects. Finally, the approach is illustrated by an example on the prognostic impact of DNA ploidy in advanced ovarian cancer.

On the misuses of artificial neural networks in oncology

Martin Schumacher
Inst. f. Med. Biometrie u. Med. Informatik, Albert-Ludwigs-Universität Freiburg, Germany

The applications of artificial neural networks (ANN's) for prognostic and diagnostic classification in clinical medicine has become very popular. Some indications might be derived from a recent "mini-series" in the Lancet with three more or less enthusiastic review articles and an additional commentary expressing at least some sceptism. In this paper, the essentials of feed-forward neural networks and their statistical counterparts (e. g. logistic regression models) are reviewed. The problems associated with the application of ANN's to survival data are outlined: These consist mainly in the fact that the resulting predicted survival probabilities are not necessarily monotone functions of time and that censored observations are not properly incorporated. A very serious problem is the estimation of misclassification rates. Here, often naive estimates as the apparent error rate are calculated and reported. The inherent overoptimism is only seldomly corrected by means of cross-validation or resampling techniques or by using an independent study as test set.

Finally, the results of a search in the medical literature from 1991 to 1993 are reported and the most frequently occuring mistakes are summarized. It is concluded that the application of ANN's to problems of diagnostic and prognostic classification in oncology deserve more thoughtful planning and analysis as it has been observed so far.

Models for longitudinal biomarkers of disease onset

Bruce W. Turnbull
Steve Gulyas & Elizabeth Slate Statistics Center, Cornell University, Ithaca, USA

We consider the analysis of serial biomarkers to screen and monitor individuals in a given population for onset of a specific disease of interest. The readings are subject to error. Two models are proposed. The first is a fully Bayesian hierarchical structure for a mixed effects segmented regression model. Posterior estimates of the change-point (onset time) distribution are obtained by Gibbs sampling. A second approach involves a hidden change-point model in which the onset time distribution is estimated by maximum likelihood using the EM algorithm. Both methods lead to a dynamic index that represents a strength of evidence that onset has occurred by the current time in an individual subject. The methods are applied to some large data sets concerning prostate specific antigen (PSA) as a serial marker for prostate cancer. The indices are compared to some standard criteria though the use of ROC curves adapted for longitudinal data.

A Model for Sequential Classification

Günter Tusch
Klinik für Abdominal- und Transplantationschirurgie,
Medizinische Hochschule Hannover, Germany

Given a sequence of K normally distributed (prognostic) discriminant scores under the homoscedastic linear model. A sequential classification procedure for two classes is proposed controlling both (conditional) error rates. The procedure is intended for clinical use. It is based on methods from classification, partial classification, and group sequential tests. There is an important difference to group sequential testing: on each step of the procedure not an additional group of patients is tested, but for the same patient an additional (independent) set of variables is included into the procedure. The admissibility of the values for the given error rates depends on the available information of the given variables. An optimisation problem is formulated for the given framework to achieve the earliest possible classification. A heuristic solution based on the error probability spending function approach of Lan and DeMets and the power function approach of Wang and Tsiatis is proposed. The procedure is extended to the case when the independence assumption is not valid. The aspect of stability of the classification is also considered. The method is demonstrated on clinical data sets.

Time-varying effects of prognostic factors

K. Ulm
Institute for Medical Statistics and Epidemiology,
Technical University Munich, Germany

The statistical analysis of prognostic factors is mainly performed by the Cox-model. One of the key assumption for the Cox-model is the proportionality of the hazards. This means the effect of a certain factor on the hazard rate has to be constant over time. It is more natural to assume that a factor is loosing its influence when time is increasing. Also other situations may occur, an increase of the effect after some time etc.. Ignoring a possible change over time can result in overlooking some factors which may be of relevance.

In order to identify such factors possible changes over time have to be considered. The usual regression coefficient has to be extended to be a function of time (b(t)). This function can be fully parameterized or estimated by some smoothing techniques. The problems in the later approach are related to perform a test on the hypothesis of b(t) = b and how to extend a model allowing a change over time at least in some of the factors. There are several proposals in the literature how to test for proportionality. But all the test proposed are optimal for certain situations. Applying spline-functions leads to the problems of selecting the appropriate test statistics. I want to show some examples from breast and ovarian cancer trials and to point to some of the possible solutions. The results may have clinical impacts on the schedule for visits at the hospitals or the treatment decisions and can help to get more insight in the biology of the tumour.

Problems in the analysis of prognostic factors with reference to the Cox model

Maria Grazia Valsecchi
Istituto di Statistica Medica e Biometria, Università di Milano, Milano, Italy

This work considers model selection and robust estimation in the context of clinical studies on survival which aim at the evaluation of treatment effect and of prognostic factors.

The use of several methods for checking the proportional hazards (PH) assumption is examined. Prognostic factors which have a non PH effect on outcome are often found in evaluating long-term survival. A graph based on the local estimation of the hazards ratio is introduced: it is obtained by fitting the basic Cox model in overlapping time windows, suitably defined. It is shown to be a useful alternative to PH checking based on time dependent covariates. This approach is compared with the one based on the smoothing of the sum of the regression coefficients, estimated under PH, and the scaled Shoenfeld residuals from the corresponding Cox model. Another graphical approach that takes advantage of the martingale residual process in time is used to represent the lack of fit, with a metric of the type "observed-expected". Several methods that extend the basic Cox regression analysis are examined. Robust estimation is carried out to downweight the contribution to estimation of influential observations, typically confined to a few long term survivors. These approaches are illustrated with a data set arising from a clinical trial on patients with advanced ovarian cancer, followed for between 7 and 12 years from randomization.

Incorproration of family history in prognostic models

Hans van Houwelingen
Dept. of Medical Statistics, Leiden University, The Netherlands

"Family history" stands for the information from a patient's pedigree. Such pedigrees can vary in size and structure. It will be discussed how such data can be used in prognostic models. Suggestions for a proper fixed dimensional score statistic can be obtained from first order approximations in a polygenic genetic model.

Reference:

J.J. Houwing-Duistermaat, H.C. van Houwelingen et. al., Testing familial aggregation, Biometrics, 51, 1292-1301, 1995

Nonparametric Procedures for Evaluating the Performance of Repeated Markers Used to Predict a Dichotomous Endpoint

Sam Wieand
Department of Biostatistics, University of Pittsburgh, USA

When evaluating repeated markers, we have found traditional statistical approaches such as the use of time dependent covariates in a proportional hazards model, while useful, fail to provide inference regarding utility measures such as specificity, sensitivity, and predictive values.

We provide nonparametric methods for estimating these parameters in the repeated marker case and derive the theory required to make inference statements. We also define an asymptotically normal statistic for comparing the sensitivities of two markers at a fixed specificity. The theory allows correlations introduced by the fact that markers may be obtained from the same patient at multiple visits and that both markers may be obtained from the same patient. We present an example from a breast cancer trial for which the utility measures give a very different perspective than results of proportional hazards model analyses.

Earlier work by Wieand, Gail, James and James (Biometrika, 1989) regarding inference for single markers and by Beam and Wieand (Biometrics, 1991) regarding methods for comparing a discrete marker to a continuous marker are special cases of this theory.

A learning diagnostic system for anemia based on Boltzmann machines

Wim Wiegerinck
Foundation for Neural Networks, Nijmegen , The Netherlands

This presentation addresses the problem of lab test selection in medical diagnosis: it has been noticed hat inexperienced physicians tends to request more lab tests than necessary, while experienced physicians use a more efficient strategy. The efficiency of the diagnostic process might be improved by a system that advices both on diagnosis and in subsequent lab test.

It is proposed to base such a system on a probability model (e.g. a Boltzmann machine) of diseases and lab test results. The system will be trained on patient data from hospital records, combined with expert knowledge. The feasibility and the usefulness of such a system is under investigation for diagnosis of anemia.

Prognosis - what does the clinician associate with this notion?

Jürgen Windeler
Institute for Medical Biometry, University of Heidelberg, Germany

Prognostication - The making of prognosis - is aimed at informing doctors and patients. According to the available literature there is a large amount of work on prognostic information and this indicates a great need: Doctors seem to think that prognostic models will automatically help them with their job which is to advise patients. Unfortunately, they are misled. Firstly they are not aware of the fact that there is no such thing like "the" prognosis of a patient. Prognoses depend on a bundle of possible future actions and their consequences.

Secondly the doctors job is acting: The aim of prognosis is to improve decisions and the treatment of patients. Apart from all statistical problems with estimation in and validation of models, however, for the vast majority of prognostic models published it is by no means evident that knowledge of prognosis leads to better treatment choices. Sometimes there are actually no treatment alternatives available.

It is argued that prognostic models have to incorporate therapeutic information and/or have to be validated in therapeutic decisions to be of help to doctors and to meet the associations doctors have with the term "prognosis".

Tagungsteilnehmer

Dr. Douglas Altman
ICRF Medical Statistics Group
Centre for Statistics in Medicine
Institute of Health Sciences
PO Box 777, Headington

GB-Oxford OX3 7LF

Prof.Dr. Leo Breiman
Department of Statistics
University of California
367 Evans Hall

Berkeley , CA 94720-3860
USA

Prof.Dr. Per Kragh Andersen
Department of Biostatistics
University of Copenhagen
Blegdamsvej 3

DK-2200 Kobenhavn N

Dr. Gregory Campbell
Food and Drug Administrati[redacted] (FDA)
HFZ-542
1350 Piccard Drive

Rockville , MD 20850
USA

Dr. Colin B. Begg
Memorial Sloan-Kettering
Cancer Center
1275 York Avenue, Box 44

New York , NY 10021
USA

Felix Dannegger
Klinikum rechts der Isar
Inst. for Medical Statistics and
Epidemiology
Ismaninger Str. 22

81675 München

Dr. Jacques Benichou
2771 Route de Neuchatel

F-76230 Isneauville

Dr. Susan S. Ellenberg
U.S.Dept. of HHS
Public Health Service
Food and Drug Administration
1401 Rockville Pike, HFM-210

Rockville , MD 20852-1448[redacted]
USA

Dr. Ornulf Borgan
Dept. of Mathematics
Section for Statistics
P.O.Box 1053 Blindern

N-0316 Oslo

Johannes Fassbinder
Inst. für Med. Informatik
Biometrie u. Epidemiologie
Hufelandstr. 55

45122 Essen

Prof.Dr. Uwe Feldmann
Institut für medizinische Biometrie
Epidemiologie und Informatik
Universitätskliniken des Saarlandes
Haus 61

66424 Homburg/Saar

Dr. Jörgen Hilden
Department of Biostatistics
University of Copenhagen
3, Blegdamsvej

DK-2200 Copenhagen N

Dr. Mitchell H. Gail
National Cancer Institute
Executive Blvd. 6130 EPN 403

Rockville , MD 20892
USA

Dr. Rolf Holle
Medis - Institut
Gesellschaft für Strahlen- und
Umweltforschung mbH
Ingolstädter Landstraße 1

85764 Oberschleißheim

Anja Gebler
Abt. für Sozialmedizin und
Epidemiologie
Ruhr-Universität Bochum
Overbergstr. 17

44780 Bochum

Prof.Dr. Hans van Houwelingen
Department of Medical Statistics
University of Leiden
P.O.Box 9604

NL-2300 RC Leiden

Prof.Dr. Frank E. Harrell jr.
Div. of Biostatistic & Epidemiology
Dept. of Health Evaluation Sciences
School of Medicine
Box 600

Charlottesville , VA 22908
USA

Katrin Jensen
Institut für Medizinische Biometrie
der Philipps-Universität Marburg
Bunsenstr. 3

35037 Marburg

Dr. Robin Henderson
Dept. of Mathematics & Statistics
University of Lancaster
Fylde College
Bailrigg

GB-Lancaster , LA1 4YF

Dr. Jochem König
Inst. für Med. Biometrie, Epidemio-
logie und Informatik
Universitätskliniken des Saarlandes
Haus 61

66424 Homburg/Saar

gefördert durch
DFG Deutsche Forschungsgemeinschaft

Prof.Dr.Dr.h.c. Klaus Krickeberg
3, rue de l'estrapade

F-75005 Paris

Prof.Dr. Jochen Mau
Institut für Statistik in der Medizin
Universität Düsseldorf
Postfach 101007

40001 Düsseldorf

Dr. Siegfried Kropf
Institut für Biometrie
u. Medizinische Informatik
Otto-von-Guericke-Universität
Leipziger Str. 44

39120 Magdeburg

Dr. Hans-Helge Müller
Institut für Medizinische Biometrie
der Philipps-Universität Ma[redacted]g
Bunsenstr. 3

35037 Marburg

Dr. Berthold Lausen
Forschungsinstitut für
Kinderernährung
Heinstück 11

44225 Dortmund

Prof.Dr. Richard A. Olshen
Department of Statistics
Stanford University
Sequoia Hall

Stanford , CA 94305-4065
USA

Dr. Ulrich Mansmann
Institut f. Med. Statistik und
Informationsverarbeitung
Hindenburgdamm 30

12200 Berlin

Prof.Dr. Carla Rossi
Dipartimento di Matematica
Universita degli Studi di Roma
Tor Vergata
Via della Ricerca Scientifica

I-00133 Roma

Dr. Peter Martus
Inst. für Medizinische Statistik
und Dokumentation
Universität Erlangen-Nürnberg
Waldstr. 6

91054 Erlangen

Dr. Patrick Royston
Dept. of Med.Stat. and Evaluation
Royal Postgraduate Medical School
DuCane Road

GB-London W12 ONN

Wilhelm Sauerbrei
Institut für Medizinische Biometrie
und Medizinische Informatik
Klinikum der Universität
Stefan-Meier-Str. 26

79104 Freiburg

Prof.Dr. Helmut Schäfer
Institut für Medizinische Biometrie
der Philipps-Universität Marburg
Bunsenstr. 3

35037 Marburg

Prof.Dr. Michael Schemper
Abt. für Klinische Biometrie
Inst. für Med. Computerwissensch.
der Universität Wien
Spitalgasse 23

A-1090 Wien

Claudia Schmoor
Institut für Medizinische Biometrie
und Medizinische Informatik
Klinikum der Universität
Stefan-Meier-Str. 26

79104 Freiburg

Prof.Dr. Martin Schumacher
Institut für Medizinische Biometrie
und Medizinische Informatik
Klinikum der Universität
Stefan-Meier-Str. 26

79104 Freiburg

Prof.Dr. Winfried Stute
Mathematisches Institut
Universität Gießen
Arndtstr. 2

35392 Gießen

Prof.Dr. Bruce W. Turnbull
Dept. of Operation Research and
Industrial Engineering
Cornell University
227 E&Tc

Ithaca , NY 14853-3801
USA

Dr. Günter Tusch
Klinik für Abdominal- und
Transplantationschirurgie
Medizinische Hochschule Hannover

30623 Hannover

Dr. Kurt Ulm
Institut für Medizinische Statistik
und Epidemiologie
Technische Universität
Ismaninger Str. 22

81675 München

Dr. Maria Grazia Valsecchi
Istituto di Biometria e Statistica
Medica
Universita degli Studi di Milano
Via G. Venezian 1

I-20133 Milano

gefördert durch
DFG Deutsche Forschungsgemeinschaft

Prof.Dr. Sam Wieand
Biostatistical Center
NSABP
Treatment Trials
230 McKee Place, Suite 600

Pittsburgh , PA 15213
USA

Dr. Wim A.J.J. Wiegerinck
Dept. of Med. Physics & Biophysics
University of Nijmegen
Geert Grooteplein 21

NL-6525 EZ Nijmegen

Dr. Jürgen Windeler
Abt. Medizinische Biometrie
Universität Heidelberg
Im Neuenheimer Feld 305

69120 Heidelberg