PROBLEMS ENCOUNTERED IN USING A CLINICAL DATABASE FOR A LONGITUDINAL STUDY IN CYSTIC FIBROSIS

ANTHONY WOHLERS

INTRODUCTION

According to Date (1982), p.7; "A database is a collection of stored operational data used by the application systems of some particular enterprise." In this definition 'enterprise' refers to a self-contained organisation such as a manufacturing company, bank, hospital, etc. 'Operational data' refers to information that is needed for the routine operation of the enterprise and is stored in a readily retrievable form. 'Application systems' are software dedicated to a specific task like invoicing, stock control etc. For the purposes of this discussion a clinical database is a database established and maintained at a hospital.

Typically the information (operational data) recorded in such a database consists of (Safran, 1991);

- patient identification (i.e. sex, age, address, data and place of birth, etc.)
- diagnosis and procedure codes
- laboratory data

- medications
- results of diagnostic procedures
- billing/insurance information

The following example of a clinical database is given in Safran (1991). Since 1984, the Beth Israel Hospital, Boston, has stored all clinical data from more than 177,000 consecutive admissions in a readily retrievable form on computer. Data are collected at the point of transaction either directly on 1 of 800 terminals located throughout the hospital or from autoanalysers. During an average week in 1988, 137,526 entries and corrections, and 40,958 clinical inquiries were made.

Pryor and Lee (1991), p. 617, succinctly explain the incentive for analysing clinical databases:

> Carefully and appropriately analysed information from clinical databases permit accurate recall of a broader experience than that available to an individual clinician, and thus has the potential to substantially enhance the information used by a clinician in the decision making process.

Clinical decision making usually refers to situations where there is a therapeutic choice but other decision making processes such as those related to the allocation/scheduling of hospital staff and resources could also be examined. Analysis of such data can also be a form of quality control/assurance of health care. Finally, the work of the statistical analyst can be seen as a way of adding value to a potentially useful resource which when compared to the cost of that resource is relatively inexpensive.

Clinical databases are longitudinal in the sense that the operational data enters the database sequentially in time. However, of more immediate interest in this discussion are patients with chronic conditions that repeatedly visit the hospital. Within the database these patients generate longitudinal information that should reflect the course of their disease. If their records can be extracted from the database then we have a *cohort* pertaining to a particular disease which with appropriate modelling may yield valuable information about the treatment and progress of the disease.

THE PROBLEMS

From the description of databases in the previous section, it can be seen that such data are *not* usually gathered with a research question or hypothesis in mind. Thus, the researcher is analysing data for a purpose different from that for which it was gathered; this presents problems. The following list (after Pryor and Lee, (1991)) gives some idea of the variety of problems:

- Important questions of relevance to research issues are not always framed before data are collected. The researcher can exercise little control over the form, method and frequency with which the data are collected.
- From the research point of view, important events are often only partially observed. For instance, time of infection is seldom observed, rather it is the appearance of the first major symptoms that is recorded on the database. Another example; the time of alteration in the concentration of certain substances in the blood which may presage important changes in some chronic condition, could go unobserved until the time of the routine checkup.
- Therapies patients receive are not usually randomised, the effects of different therapies are thus *confounded.*
- There are a variety of biases (geographical, socio-economic, demographic etc.,) that may limit generalizability.
- There are often so many variables that data reduction strategies are needed.
- There may be repeated (possibly irregular) observations made over long periods of time.

- Databases can sometimes be very large and seldom are all the data relevant to a particular study. Thus their manipulation, extraction and checking can consume a major part of a study's resources.

- Supervision and maintenance of the database may change over time and there may be no provison for noting such changes *within* the database.

- Similarly, measurement procedures or record coding may change over time or vary from place to place without adequate notification *within* the database.

- There may be concerns about data quality or missing observations on a sizable proportion of the patients.

CASE STUDY: CYSTIC FIBROSIS (CF)

In order to illustrate the points raised above, an example of a longitudinal study at present in progress will be examined. However to better appreciate the issues involved, some brief comments about CF and the aims of the study are required.

BACKGROUND AND AIMS OF THE STUDY

CF is a genetic disease (inherited autosomal recessive trait) and is now the commonest cause of chronic suppurative lung disease in Caucasian children (Phelan, Landau and Olinsky, 1990, Chapter 10). The incidence in the Australian community is approximately one in 2500 live births (Allan, Robbie, Phelan and Danks, 1980). Thirty years ago few children with CF lived much beyond 5 to 10 years, but now most patients can expect to reach adult life, often with relatively little disability (Phelan, Landau and Olinsky, 1990, Chapter 10). Although in general terms, this increased longevity can be attributed to earlier diagnosis, improved therapy and better medications, the definite reasons for the better outcome are not known. Therefore, the ideal form or forms of management remain in doubt (Phelan et al, 1990). As the CF patient population ages, the need to remove this uncertainty increases. In the light of these comments, any statistical method that can supply useful prognoses to those charged with the care of CF sufferers is potentially important. This is especially true if such methods use existing data accumulated as part of the ongoing treatment of CF at specialist clinics. Therefore, such an investigation of statistical methods of prediction for CF is both timely and likely to be of practical importance.

The aim of this study is summarised as follows; To use a variety of statistical models and an existing CF patient database to predict individual outcomes (prognoses) in terms of regular measurements such as lung function, height and weight, and patient characteristics such as sex etc. Thus, the assumption is that the database contains information that can describe the course of the disease with sufficient clarity so that useful predictions for individuals can be made.

METHODOLOGY

In this study it is proposed to use the methodological framework of *Clinical Decision Analysis* (CDA) (Weinstein, Fineberg et al, 1980, Kassirer et al, 1987). CDA has been developed to explicate competing therapies (or no therapy) and their differing outcomes. Its attraction is that it provides an objective and communicable framework around which to organize one's thinking about the problem of therapeutic choice. In the case of long term illness where certain disease events may recur, it has been found convenient to reformulate CDA using *stochastic process* models. A stochastic process is any process whose evolution in time depends on chance (Syski, 1988); clearly the course of a chronic disease fits this description. Thus the CDA framework can be used to explicate competing courses of a disease and their differing outcomes. Furthermore, this framework allows for the incorporation of utilities, thus explicitly taking 'Quality of Life' into account (Gelber, 1989, Glasziou, Simes and Gelber, 1990).

How does a stochastic process describe a chronic disease? Usually, it approximates the course of the disease by a sequence of transitions between various 'states' of health. For example, in the 'disability' or 'three state illness model' (Andersen and Borgan, 1985, Andersen, Hansen and Keiding, 1991a, 1991b) a patient's experience of a disease consists of a sequence of transitions between 'not ill' and 'ill' states, possibly finishing with a transition to the 'dead' state. The states are described by a *marker* (Andersen, 1991) which is usually some function of one or more clinical measurements. For instance, being in the 'ill' state could mean that the value of some clinical measurement has dropped below a certain level. If information on this measurement is stored in the clinical database then the history, in terms of state transitions, can be recovered. Furthermore, if there are sufficient patient records stored in the database, then the parameters of the stochastic process (i.e. transition probabilities, time spent in various states (sojourn times) etc.) can be estimated from a subset of the data. A subset of the available data is used in the estimation stage so that the remaining data can be used in the model testing stage. In the testing stage model predictions (outcomes) are compared with actual outcomes as recorded on the database. If the predictions are usually in agreement, then the stochastic model can be used as a clinical decision support tool, if there is substantial disagreement the model needs to be reformulated and the estimation and testing stages need to be repeated. There are many types of stochastic process but three types that are particulary popular in chronic disease studies are *Markov*, *semi-Markov* and *Multivariate Counting* processes.

In simple *Markov* models (Beck, 1983, Kay, 1984, 1986), the probability of the present 'value' of the process (in the 'disability' model the values are; 'not ill', 'ill' or 'dead') depends only on the previous value of the process rather than on the whole sequence or history of previous values.

The simple Markov model is unrealistic in many contexts and various extensions have been proposed (Ochi, 1990). One of the most common is the *Semi-Markov* model (Lagakos, Sommer and Zelen, 1978, Dinse, 1986, Dunsmuir et al, 1989). This model extends the previous formulation by making the time in the present state influence the likelihood of transition (Andersen et al. 1991b). For instance, it has been observed that for some therapies, longer remission times seem to be related to fewer relapses.

Multivariate Counting Process Models (Aalen, 1978, Andersen and Gill, 1982, Gill, 1984, Andersen and Borgan, 1985) describe a stochastic process with two or more components that can be thought of as counting the occurences (as time proceeds) of two or more different types of event. For instance, in the disability model types of event would be transitions; 'not ill' to 'ill', 'ill' to 'not ill', 'not ill' to 'dead' and 'ill' to 'dead' (i.e. a four-component counting process). There are many attractions to the counting process formulation, most importantly it can be used to model a wide variety of multistate disease models *nonparametrically*; its theory and estimation methods can be developed without specifying a probabilistic structure. Other attractions include, flexibility in the way of specifying the number of subjects under quite arbitrary schemes of recruitment and censoring, and extension to regression type models such as Cox's Model (Cox, 1972, Andersen and Gill, 1982). Regression models allow clinical variables (covariates), other than those which are used to define the states of the process, to be used to adjust for heterogeneity amongst the patient group. Thus transition rates or survival probabilities can be made functions of sex, age, weight and any other attributes which vary amongst patients. The multivariate counting process can therefore be used to compute the Markov transition rates and probabilities (Aalen and Johansen, 1978, Andersen et al., 1991b), check the Markovian assumption and examine 'causality' or 'local dependence' between events (Aalen et al., 1980). With some restrictions multivariate counting processes can be used for semi-markov models (Voelkel and Crowley, 1984, Andersen and Rasmussen, 1986, Clayton, 1988). Thus, by viewing the course of CF as a realisation of a counting process, a rich source of theory and method is available to explore the data, test hypotheses and make predictions.

THE DATA

The data that will be analysed using the methods outlined above come from the Royal Children's Hospital Cystic Fibrosis database. This database contains records for patients under the care of the CF clinic, Royal Children's Hospital, Melbourne and an associated clinic for adults at the Alfred Hospital, Melbourne. These two centres care for about 90% of known living CF sufferers in the state of Victoria. There are slightly more than 600 patients on the database of whom about 30% have died. Many of these patients have been observed for over ten years. Data so far extracted from the database consist of the following variables

- numeric patient identification
- sex
- date of birth
- date of diagnosis
- date of death (if available)
- mode of presentation (at diagnosis)
- date of regular assessment (repeated)
- weight (repeated)
- height (repeated)
- FEV_1: forced expiry volume in 1 sec. (repeated)
- Vital Capacity: max volume that can be expired after complete inspiration (repeated)
- Pulmonary Status (Holzer, Olinsky and Phelan, 1981): a graded scale, 0 (good) to 4 (severe) made up of observations on cough persistence, amount of sputum, chest X-ray, FEV_1 (repeated)
- Sputum: a categorical description of an assay of sputum (repeated)
- Isolation of *Pseudomonas aeruginosa* in sputum, categorical (repeated)

In addition to these *explicit* variables, there are *implicit* variables such as survival time, time until diagnosis etc., which can be calculated from the date information in the database. Data on those variables with 'repeated' in brackets have been collected at yearly intervals since 1974. Since these variables exhibit large fluctuations that may be unrelated to the course of the disease, the best value in the six month period preceeding the annual assessment is taken (Hibbert, 1984). Before 1974, only non-repeated data on those that had not died before 1974 are on the database.

How will this information be used, in terms of the methods discussed in the previous section? The first task in applying the stochastic models is to identify different states of CF. For instance, pulmonary status is a composite measure of respiratory health which could be directly used as a marker to define 'states' of the disease process. However, this variable does not take into account weight changes, the type of micro-organisms found in the sputum assay or the presence and type of *Pseudomonas aeruginosa*[1] Thus more elaborate models may have distinct states defined by categories of weight and lung function loss or important bacteriological findings. Once the states or different groups of states have been identified (using expert clinical advice and graphical analysis), the parameters of the transitions can be estimated using the timing information implicit in the data. These parameters themselves can be made functions of variables not used in defining the disease states. However before these models can be tested and fitted, errors and inadequacies within the data need to be detected and resolved.

STRUCTURAL AND DATA QUALITY PROBLEMS

The CF database is not typical of clinical databases in that its coverage is almost an entire population, rather than a sample, and it is specific to one disease which is contracted at conception. Thus it is not unreasonable to hope that the results of the study will generalize to other CF sufferers. Although large enough for statistical modelling purposes, it is relatively small so that data may be checked against original hardcopy of the measurements which are still on file. Another advantage with its size is that the clinicians can recall a lot of anecdotal material about a large proportion of the patients which often explains unusal entries in the database. Such characteristics

[1] "Chronic *P. aeruginosa* lung infection caused by the mucoid strains is a predominant cause of death in patients with CF." (Phelan, Landau and Olinsky, 1990, p. 196)

certainly make it easier to analyse, but there are still major problems which it shares with general clinical databases.

The difficulties faced in implementing this analysis are two fold; structural and data quality problems. Structural problems arise because the data exhibit right censoring and interval censoring of timing information, and left truncation of repeated measurements. *Right censoring* arises because only a small proportion of patients have died during the history of the database. Therefore, for the majority of patients we only have a partial record of the course of CF which in terms of a horizontal time line is still continuing to the right. *Interval censoring* arises because apart from date of birth, death and diagnosis, measurements are taken at yearly intervals. This means that the exact time at which a state transition happens is unknown. *Left truncated* data belong to patients who were alive before 1974. For these patients repeated measurements start in 1974 although they had CF before this time (i.e. further to the left on a horizontal time line). What can one do about these structural problems? One of the attractive features of the multivariate counting process formulation is that it easily handles right censoring (Aalen, 1978). Interval censoring provides a bigger problem and the usual 'solution' is to make the approximation that changes happen exactly at the time of measurement (Andersen et al. 1991a, for a comparative study of this approach). Left truncation is more problematic since we have no idea of what the missing data may have been. The simplest solution is to leave out those patients whose histories show a high proportion of left truncation (e.g. an individual who was born in the early 1950s and died in the 1970s).

The other major problem area is data quality. The problem has two major facets; how to detect suspect data and then what to do about it. The researcher needs to have a 'feel' for the data but this develops slowly when the researcher is not involved in gathering the data. There are no short cuts but the following points may be useful to consider:

- The researcher needs to start with a clear and precise *written* description of the database; its history, its variables, their units, their method of measurement, their range of variation, any reasons for not taking measurements or failure to followup, etc. When this information is not available in a concise and clear form then it must be prepared through consultation with relevant staff and examination of past research papers (if any) that used the same data. The preparation of this material is a good example of research 'adding value' to the database.

- Variables are usually related; such relationships can be used to detect errors. For instance with regard to CF data, Vital Capacity (VC) must always be greater than Forced Expiry Volume (FEV_1); one cannot expire more than one's total lung capacity. Thus a plot of VC versus FEV_1 will show where this is not the case thus revealing mistakenly entered data.
- There may be bivariate and trivariate outliers that reveal errors. For instance, large weight loss and lung function gains may be separately acceptable, but nonsensical when simultaneous on one individual. Thus simple bivariate plots may illustrate suspect outliers.
- Graphical examination of individual patient data may reveal erroneous reversals or fluctuations that do not show up as outliers.
- The dynamics of the disease process may be more usefully visualised by graphically examining the first ('velocity') and second ('acceleration') differences of the repeated measurements. Such an examination may highlight sudden or peculiar changes.

Having found missing or unusual data, the next stage is to identify correct data from incorrect data and the circumstances under which data may be expected to be missing. This stage requires expert help since the researcher may not be able to distinguish clerical error from clinical peculiarity. In the case of CF data we have the luxury of being able to check the data against original hardcopy but with very large databases this may be impractical and sampling may be required. Where original data are not available, less direct methods need to be used. Various *imputation* schemes are available (Rubin and Schenker, 1991) but their implementation depends on some knowledge of measurement levels and their variation. Another approach could be interpolation by nonparametric smoothing which uses regression splines (Wegman and Wright, 1983, Wahba, 1989, Friedman, 1991). Unfortunately there is a shortage of software for simply implementing this approach, so one must adapt existing software to this purpose (Freund and Littell, 1986) or write programs. In addition, these methods tend to be computationally intensive which may render them impractical for large databases unless sampling is used. At present the CF study is still at the stage of developing diagnostic tools for outlier detection and checking database entries against raw data. It is hoped that the quality of the resulting data will preclude the use of less direct methods of error correction.

DISCUSSION

By way of conclusion some general comments from the Data Analysis Workshop (DAW) are in order:

- Considerable discussion was generated by a plot of a patient's height record which showed a fluctuation of six centimetres at age 22 years. The discussion showed that an apparently obvious transcription error could conceivably be a genuine clinical observation, thus emphasizing the care that needs to be taken in interpreting outliers.
- Several presentations foreshadowed or described the use of large databases for statistical analysis in Australia.
- The increasing local use of clinical and other databases illustrates a potentially wider use of counting process methods in longitudinal research.

These general comments on the local scene reflect the comments of Moses (1991) p. 633;

The emergence of the field of Health Service Research, and the increasing use of computer technology, ensure that accessible databases will be more and more widely used in research. Conclusions based on the results of such research will influence budgets, norms of practice, corporate behaviour, and presumably much more.

Such research is difficult to do, indeed may be full of traps, but is inevitably going to be done, more and more, never mind the risks. It is therefore urgent that we develop ways of doing such research as well as possible.

REFERENCES

AALEN, O.O. (1978) Nonparametric Inference for a Family of Counting Processes, *The Annals of Statistics*, **6** pp.701–726.

AALEN, O. O., BORGAN, O., KEIDING, N. and THORMANN, J. (1980) Interaction between Life History Events. Nonparametric Analysis for Prospective and Retrospective Data in the Presence of Censoring, *Scandinavian Journal of Statistics*, **7** pp.161–171.

AALEN, O. O. and JOHANSEN, S. (1978) An Empirical Transition Matrix for Non-homogeneous Markov Chains Based on Censored Observations, *Scandinavian Journal of Statistics*, **5** pp.141–150.

ALLAN, J., ROBBIE, M., PHELAN P.D. and DANKS, D.M. (1980) The incidence and presentation of cystic fibrosis in Victoria 1955-1978, *Aust. Paediatr. J*, **16**,pp.270–273.

ANDERSEN, P.K. and BORGAN, O. (1985) Counting Process Models for Life History Data: A Review (with discussion), *Scandinavian Journal of Statistics*, **12** pp.97–158.

ANDERSEN, P.K. and GILL, R. (1982) Cox's Regression Model for Counting Processes: A Large Sample Study, *Annals of Statistics*, **10** pp.1100–1120.

ANDERSEN, P.K., HANSEN, L.S. and KEIDING, N. (1991a) Assessing the Influence of Reversible Disease Indicators on Survival, *Statistics in Medicine*, **10** pp.1061–1067.

ANDERSEN, P.K., HANSEN, L.S. and KEIDING, N. (1991b) Non-and Semi-parametric Estimation of Transition Probabilities from Censored Observation of a Non-homogeneous Markov Process, *Scandinavian Journal of Statistics*, **18** pp.153–167.

ANDERSEN, P.K. and RASMUSSEN, N.K. (1986) Psychiatric Admissions and Choice of Abortion, *Statistics in Medicine*, **5** pp.243–253.

BECK, J.R. and PAUKER, S.G. (1983) The Markov Process in Medical Prognosis, *Medical Decision Making*, **3**, pp.419-458.

CLAYTON, D. (1988) The Analysis of Event History Data: A Review of Progress and Outstanding Problems, *Statistics In Medicine*, **7**, pp.819–841.

COX, D.R. (1972) Regression Models and Life-Tables (with discussion), *Journal of the Royal Statistical Society, Series B*, **34**, pp.187–220.

DATE, C.J. (1982) *An Introduction to Database Systems*, Addison-Wesley.

DINSE, G.E. and LARSON, M.G. (1986) A Note on semi-Markov Models for Partially Censored Data, *Biometrika*, **73**, pp.379–386.

DUNSMUIR, W., TWEEDIE, R, FLACK, L. and MENGERSEN, K. (1989) Modelling of the Transitions Between Employment States for Young Australians, *Australian Journal of Statistics*, **31A**, pp.165–196.

FREUND, R.J. and LITTELL, R.C. (1986) *SAS System for Regression*, SAS Institute Inc..

FREIDMAN, J.H. (1991) Multivariate Adaptive Regression Splines (with discussion), *The Annals of Statistics*, **19**,pp.1–141.

GELBER, R.D., GELMAN, R.S. and GOLDHIRSCH, A. (1989) A Quality-of-Life-Oriented Endpoint for Comparing Therapies, *Biometrics*, **45**,pp.781–795.

GILL, R.D. (1984) Understanding Cox's Regression Model: A Martingale Approach, *Journal of the American Statistical Association*, **79**,pp.441–447.

GLASZIOU, P.P., SIMES, R.J. and GELBER, R.D. (1990) Quality Adjusted Survival Analysis, *Statistics in Medicine*, **9**, pp.1259–1276.

HIBBERT, M.E. (1984) *A Statistical Analysis of Cystic Fibrosis*, Student Project: Royal Melbourne Institute of Technology, Australia.

HOLZER, F., OLINSKY, A. and PHELAN P.D. (1981) Variability of airway hyperreactivity in cystic fibrosis, *Arch Dis Childh*, **56**, pp.455–459.

KASSIER, J.P., MOSKOWITZ, A.J., LAU, J. and PAUKER, S.G. (1987) Decision Analysis: A Progress Report, *Annals of Internal Medicine*, **106**, pp.275–291.

KAY, R. (1984) Multistate Survival Analysis: An Application to Breast Cancer, *Methods of Information in Medicine*, **23**, pp.157–162.

KAY, R. (1986) A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies, *Biometrics*, **42**, pp.855–865.

LAGAKOS, S.W., SOMMER, C.J. and ZELEN, M. (1978) Semi-Markov Models for Partially Censored Data, *Biometrika*, **65**,pp.311–317.

MOSES, L.E. (1991) Innovative Methodologies for Research Using Databases, *Statistics in Medicine*, **10**, pp.629–633.

OCHI, M.K. (1990) *Applied Probability and Stochastic Processes in Engineering and Physical Sciences*, J. Wiley & Sons.

PHELAN, P.D., LANDAU, L.I. and OLINSKY, A. (1990) *Respiratory Illness in Children*, Blackwell Scientific Publications.

PRYOR, D.B. and LEE, K.L. (1991) Methods for the analysis and assessment of clinical databases: The clinician's perspectives, *Statistics in Medicine*, **10**, pp.617–628.

RUBIN, D.B. and SCHENKER, N. (1991) Multiple Imputation in Health-Care Databases: An Overview and Some Implications, *Statistics in Medicine*, **10**, pp.585–598.

SAFRAN, C. (1991) Using Routinely Collected Data for Clinical Research, *Statistics in Medicine,*, **10**, 559–564.

SYSKI, R. (1988) Stochastic Processes, In, *The Enclopeadia of Statistical Sciences*, **8**, pp.836–850.

VOELKEL, J.G. and CROWLEY, J. (1984) Nonparametric Inference for a Class of Semi-Markov Processes with Censored Observations, *The Annals of Statistics*, **12** pp.142–160.

WAHBA, G. (1989) Spline Functions, *The Encyclopedia of Statistical Sciences*, **Supplement**, pp.148–160.

WEGMAN, E.J. and WRIGHT, I.W. (1983) Splines in Statistics, *Journal of the American Statistical Association*, **78**, pp.351–365.

WEINSTEIN, M.C., FINEBERG, H.V., ELSTEIN, A.S., FRAZIER, H.S., NEUHAUSER, D., NEUTRA, R.R. and MCNEIL, B.J. (1980) *Clinical Decision Analysis*, W.B. Saunders Company.

Department of Paediatrics
The University of Melbourne
c/o The Royal Children's Hospital
Flemington Road
Parkville, Victoria, 3052
Australia