

**ERA-NET SCHEME**  
**COORDINATION ACTION**

**ALLIANCE-O**



**European Group for Coordination of National Research Programmes on  
Organ Donation and Transplantation**

Project/Contract Number: 0011853

**Work Package 5:  
Coordination of evaluation of methodologies of transplantation performance**

**Deliverable 5.2: Proposal for standardised methods to monitor the performance of  
different transplantation teams and standardised risk factors**

**Work Package Leader:** UK Transplant, United Kingdom

| <b>Participant name</b>                                | <b>Abbreviation</b> | <b>Country</b> |
|--|---------------------|----------------|
| Agence de la Biomédecine                               | ABM                 | France         |
| Deutsche Stiftung Organtransplantation                 | DSO                 | Germany        |
| Hungarotransplant                                      | Hu-T                | Hungary        |
| Centro Nazionale Trapianti                             | CNT                 | Italy          |
| Organização Portuguesa de Transplantação               | OPT                 | Portugal       |
| Centro Nacional de Trasplantes y Medicina Regenerativa | CENATMER            | Spain          |
| UK Transplant  | UKT                 | United Kingdom |

## TABLE OF CONTENTS

|       |  |    |
|-------|--|----|
| 1     | INTRODUCTION .....                                   | 5  |
| 1.1   | Background .....                                     | 5  |
| 1.2   | Purpose of document .....                            | 6  |
| 1.3   | Data auditing and data quality .....                 | 6  |
| 1.4   | Outline of content .....                             | 8  |
| 2     | STRATEGY FOR DATA ANALYSIS .....                     | 9  |
| 2.1   | General strategy for data analysis .....             | 9  |
| 2.2   | Data and types of outcome .....                      | 10 |
| 2.3   | Choice of cohort .....                               | 10 |
| 2.4   | Choice of method .....                               | 11 |
| 2.5   | Data checking and validation .....                   | 11 |
| 2.6   | Discussion .....                                     | 13 |
| 3     | SUMMARISING AND MODELLING TRANSPLANT OUTCOMES .....  | 15 |
| 3.1   | Summarising transplant outcomes .....                | 15 |
| 3.1.1 | <i>Binary outcomes</i> .....                         | 15 |
| 3.1.2 | <i>Time-to-event outcomes</i> .....                  | 15 |
| 3.1.3 | <i>Limitations of descriptive methods</i> .....      | 16 |
| 3.2   | Modelling transplant outcomes .....                  | 16 |
| 3.2.1 | <i>General modelling strategy</i> .....              | 17 |
| 3.2.2 | <i>Modelling binary outcomes</i> .....               | 18 |
| 3.2.3 | <i>Modelling time-to-event outcomes</i> .....        | 18 |
| 3.3   | Discussion .....                                     | 19 |
|       | References .....                                     | 19 |
| 4     | RISK FACTORS .....                                   | 20 |
| 4.1   | Potential risk factors .....                         | 20 |
| 5     | COMPARING OUTCOME RATES ACROSS CENTRES .....         | 22 |
| 5.1   | Variation in outcome measures .....                  | 22 |
| 5.2   | Outcome measures .....                               | 23 |
| 5.2.1 | <i>Unadjusted outcome measures</i> .....             | 23 |
| 5.2.2 | <i>Risk-adjusted outcome measures</i> .....          | 23 |
| 5.2.3 | <i>Calculation of expected outcome rates</i> .....   | 24 |
| 5.3   | Presentation of results .....                        | 24 |
| 5.3.1 | <i>League tables</i> .....                           | 24 |
| 5.3.2 | <i>Funnel plots</i> .....                            | 25 |
| 5.4   | Discussion .....                                     | 26 |
|       | References .....                                     | 27 |
| 6     | MONITORING TRANSPLANT OUTCOMES WITHIN CENTRES .....  | 29 |
| 6.1   | Rationale for sequential monitoring .....            | 29 |
| 6.2   | Methods for sequential monitoring .....              | 29 |
| 6.2.1 | <i>Observed – Expected CUSUM chart</i> .....         | 30 |
| 6.2.2 | <i>SPRT and re-setting SPRT (RSPRT) charts</i> ..... | 31 |
| 6.2.3 | <i>Tabular CUSUM chart</i> .....                     | 32 |
| 6.2.4 | <i>Other methods</i> .....                           | 33 |
| 6.2.5 | <i>Choice of method</i> .....                        | 33 |
| 6.3   | Planning and setting up a monitoring procedure ..... | 33 |
| 6.4   | Discussion .....                                     | 34 |
|       | References .....                                     | 35 |

|       |  |    |
|-------|--|----|
| 7     | MISSING DATA TECHNIQUES .....                                | 37 |
| 7.1   | Why worry about missing data? .....                          | 37 |
| 7.2   | When to exclude variables .....                              | 37 |
| 7.3   | The missingness mechanism .....                              | 38 |
| 7.3.1 | <i>Missing completely at random</i> .....                    | 38 |
| 7.3.2 | <i>Missing at random</i> .....                               | 38 |
| 7.3.3 | <i>Missing not at random (MNAR)</i> .....                    | 38 |
| 7.4   | How to deal with missing data: the most common methods ..... | 38 |
| 7.4.1 | <i>Complete case analysis or listwise deletion</i> .....     | 38 |
| 7.4.2 | <i>Missing category</i> .....                                | 39 |
| 7.4.3 | <i>Mean / Median Imputation</i> .....                        | 39 |
| 7.4.4 | <i>Regression imputation</i> .....                           | 39 |
| 7.5   | More sophisticated methods .....                             | 39 |
| 7.5.1 | <i>Hot Deck imputation</i> .....                             | 39 |
| 7.5.2 | <i>EM Algorithm</i> .....                                    | 40 |
| 7.5.3 | <i>Multiple Imputation</i> .....                             | 40 |
| 7.6   | Statistical software .....                                   | 40 |
| 7.7   | Discussion .....   | 41 |
|       | References .....   | 41 |
| 8     | ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS .....            | 42 |
| 8.1   | Origins of the methods .....                                 | 42 |
| 8.2   | Theory .....   | 43 |
| 8.3   | AI in medicine .....   | 43 |
| 8.4   | AI in survival analysis .....                                | 45 |
| 8.5   | A review of some of some specific papers .....               | 46 |
| 8.6   | An example of the use of AI methods .....                    | 51 |
| 8.7   | Conclusions .....  | 54 |
| 9     | COMPARISON OF STATISTICAL SOFTWARE .....                     | 55 |
| 9.1   | Available software .....                                     | 55 |
| 9.2   | Software capability .....                                    | 56 |
| 9.3   | Choice of software .....                                     | 58 |
| 10    | CONCLUDING REMARKS .....                                     | 60 |

# 1 INTRODUCTION

A state of the art review of methodology used by European organisations in Alliance-O for the monitoring and analysis of transplant outcomes led to the publication of Deliverable 5.1 of the project. The main aim of this review was to summarise the techniques that are currently being used by the partner organisations in the analysis of outcomes following transplantation and in the comparison and monitoring of centre performance. The review included details on all risk factors that are taken account of in these analyses. We also took the opportunity to outline the data collection process in each country, the funding arrangements for data collection and statistics and audit, the computer software used in data analysis, and how statistical results are made available to the public.

This document, which forms Deliverable 5.2, summarises statistical methods that are helpful in summarising, analysing and monitoring data on transplant outcomes.

## 1.1 Background

In the field of organ transplantation, quantitative methods are widely used in processing the large amounts of data that are collected on patients who are registered for a transplant, transplant recipients, and organ donors. Some of the aims in analysing such data are listed below.

- Summarise graft and patient survival rates following transplantation.
- Investigate how graft and patient survival depend on factors associated with the donor, the recipient and the transplant procedure.
- Monitor and evaluate transplant outcomes in particular centres in order to identify if there has been an underlying change in graft or patient survival.
- Ensure that organ allocation procedures are operating in the desired manner.
- Investigate the extent to which there is geographical equity of access to both registration and transplantation.
- Identify how treatment changes following transplantation, such as changes to immunosuppression, impact on graft or patient survival.
- Audit the incidence of damaged in organs retrieved for transplantation, and complications arising from transplantation.
- Summarise the mortality of patients on the waiting lists for a transplant.
- Quantify the benefits of transplantation in terms of survival time and quality of life.

- Provide criteria that may be used to ensure consistency of approach between centres in registration for transplantation.

These analyses serve many different purposes, but in particular they allow national organisations responsible for the conduct of organ transplantation to ensure that organ allocation is equitable, and that optimal use is made of organs that become available. The results of analyses such as these also lead directly to the development of clinical practice for the greater benefit of patients. Proper provision of a statistical analysis and audit service in a national transplant organisation is therefore essential for the development of organ transplantation

## **1.2 Purpose of document**

In the past 50 years, many statistical methods have been used to address a wide variety of issues and problems stemming from organ donation and transplantation. However, in many situations it is not straightforward to select the most appropriate technique. We also have to bear in mind that most European national organisations that oversee transplantation have few, if any, statisticians (see Deliverable 5.1) and yet they are nonetheless required to give quantitative information on many aspects of transplantation in their respective countries.

Recognising this, the main aim of this document is to provide a summary of methods that have proved to be helpful in the analysis of transplant data. This summary is not designed to be comprehensive, and we certainly have not reviewed all statistical methods that have the potential for use in this area. Moreover, we have not given much detail on the techniques themselves, but have made reference to published textbooks and papers where full details can be found. What we have done is to identify methods that are of value, that are widely referred to in the transplant literature, and that are readily implemented in computer software that is commonly available. Our hope is that this review will enable countries with a less well-developed transplant infrastructure to adopt suitable statistical procedures without the need for research in this area. In other words, the methods are presented in the spirit of sharing best practice.

## **1.3 Data auditing and data quality**

The shortage of organs for transplantation and the continual increase in the size of the waiting lists means that best use must be made of those organs that do become available. Consequently, it is vital to audit transplant activity and outcomes following transplantation so that any unavoidable causes of transplant failure are quickly identified and remedied. Data auditing is therefore very important in the conduct and development of organ transplantation.

In general, the procedure involves the collection of data on various aspects of the transplant process to ensure compliance with specified standards. For example, data may be obtained on the characteristics of donors and recipients of a particular organ in order to confirm that allocation rules have been correctly implemented. Clearly, if there are instances of non-compliance, these need to be thoroughly investigated so that those who oversee the allocation process can determine the reasons for the deviation from clinical practice and take any necessary action. By the same token, an audit may identify unanticipated effects of an allocation scheme. For example, in a

recent audit of the UK heart allocation scheme, it was found that the practice of allocating hearts from blood group O donors to recipients with a compatible blood group was disadvantaging blood group O patients. The audit led to a review of the allocation scheme and consequential modifications.

An audit may be set up to study whether the results being obtained by a particular transplant centre are in line with past performance or national rates. Commissioners of transplantation, and also potential patients, will need to be assured that all funded centres are producing similar results, after suitable account has been taken of the fact that some centres may have a greater proportion of more difficult cases. A transparent monitoring process will help to ensure prompt investigation of the cause of any change in surgical performance, or patient management following transplantation, which may in turn lead to procedural change.

The success of transplantation as a treatment for a number of life threatening conditions means that an increasing number of patients are being deemed suitable for transplantation, and the criteria for assessing suitability are constantly evolving. Moreover, the shortfall in the number of donors to meet the demand for transplantation means that it is essential to ensure that best use is made of all organs that become available. The setting of standards for best practice is fundamental to the development of transplantation, and so professional bodies, national and international, produce guidelines and standards. Audit procedures enable an assessment to be made of the extent to which such guidelines are followed. In addition, an established audit procedure can help to ensure compliance with these standards.

Audit procedures require data that have a high degree of accuracy with missing data kept to a minimum. It is therefore important that organisations that maintain transplant registries strive for both accuracy and completeness in the data collected. Data may be received by a transplant registry in a number of ways, including paper forms, and through web based file transfer system. Data returns need to be constantly monitored to ensure that the return rate is maximised. Remote data collection using a data-collecting agency may be necessary. The development of strong relationships between those responsible for data collection on the transplant units, and the national organisation, is essential to ensure that there is the necessary level of cooperation.

Within the transplant organisation itself, procedures need to be in place to ensure that no errors are introduced in data processing. This generally means that there should be a dedicated data entry team, with proper support and training. Double data entry will be needed for paper records, and a system for contacting individual centres about missing data values needs to be in place. A process for tracking the data entry and validation process is needed to ensure that timescales for data processing are met.

In addition, validation rules must be in place to test the accuracy of data. These rules will be specific to each data item, and will typically be used to examine whether the recorded value is in the valid range for that variable. For example, the height range of an adult may be specified as 1 metre to 2 metres; values found to be outside this range may be referred back to the transplant centre for correction or verification. There may also be checks across a number of variables to ensure that there is not an impossible combination of values. For example, a record of the incidence of ovarian cancer in following up a transplant recipient would only be expected in a female patient.

It is only after each data form has passed all the validation checks that data should be committed to the transplant database. Such procedures ensure that data of the highest quality are available for analysis.

#### **1.4 Outline of content**

Following this introductory chapter, Chapter 2 outlines the general strategy in data analysis. Included in this chapter is material on the different types of data that are encountered in transplantation, and the outcome variables that are of most interest. There is also a discussion of the importance of data checking and validation. In Chapter 3, methods for summarising and modelling transplant outcomes are described, with particular emphasis on techniques for analysing survival data and categorical data.

The modelling approach to the analysis of transplant data enables account to be taken of those factors that affect the outcome. Often, information on a large number of such factors is available, and so in Chapter 4, we summarise risk factors that ought to be considered when modelling outcome data.

It is important to audit centre outcomes in order to ensure that these are in line with expectation, and that any variation in the results from different centres that can be explained by the fact that different centres may see patients with different characteristics, that is case mix. Consequently, Chapter 5 describes methods that are helpful in analysing the extent of differences in outcomes between transplant centres. In Chapter 6, techniques developed for monitoring the outcomes of centres over time are summarised and illustrated.

Missing values are frequently encountered in many areas of application, and particularly in the analysis of transplant data. Accordingly, methods for handling such values are described in Chapter 7.

Recently, so-called neural network methods have found application in the formulation of survival rates that a transplanted patient may expect, as a function of the characteristics of possible donors, the potential recipient. These methods are at the early stages of development and so at this stage it not practicable to provide definitive guide to their use in connection with transplantation. Instead, in Chapter 8, an overview of the area is given in the hope that this will lead to further development and application of the techniques.

Chapter 9 reviews computer software that may be used to implement the methods described in this document. This is followed by some concluding remarks in Chapter 10.

#### *A note on authorship:*

The main part of this document is contained in Chapters 2-6, that were written by Nokuthaba Sibanda, formerly of UK Transplant and now with the Clinical Effectiveness Unit of the Royal College of Surgeons of England. Ylana Chalem of Agence de la Biomédecine, France, wrote Chapter 7 on missing values. Chapter 8 on artificial intelligence methods and Chapter 9 on computer software was prepared by Dino Mattucci of Centro Nazionale Trapianti, Italy. Dave Collett of UK Transplant wrote Chapters 1 and 10.

## 2 STRATEGY FOR DATA ANALYSIS

Data on organ transplants and their outcomes are often collected and submitted to transplant registries where they are stored in large databases. The primary aim in collecting the data is to facilitate the efficient allocation of organs to patients listed for transplant and to allow an evaluation of the outcomes of this process [1]. Analyses of these data are then carried out at a later stage with the aim of, say, providing descriptive summaries of transplant outcomes, determining the existence of association between variables, evaluating the effect of particular factors on outcome or comparing outcomes across different transplant teams. The results of these analyses have the potential to influence decision-making and practice patterns at various stages of the transplant process, such as organ allocation and post-transplant care [2,3,4].

In this section we consider a strategy that might be used to undertake an analysis of transplant outcome data. It is assumed that the data being analysed are observational, where the data are collected without intentionally interfering with any part of the transplant process as would be done in a clinical trial. A general strategy for data analysis is set out first, and then some of the aspects of the analysis are considered in more detail.

### 2.1 General strategy for data analysis

The following are the steps that one might follow in undertaking an analysis of transplant outcome data [5]. It is assumed that the data have already been collected and are stored in a database or spreadsheet.

**Step 1:** Formulate an aim or objective for the analysis. This requires a clear definition of the question to be answered through the analysis, including the outcome of interest and the characteristics of the transplants to include in the analysis. This is an important step in the analysis as it determines the approach taken.

**Step 2:** Design the analysis. The design must at least consider the following:

- a) Identification of the ‘cohort’ (group of patients or transplants) to be used in the analysis. In addition, the variables measuring the outcome and those necessary for use in the selection of transplants for inclusion must also be identified.
- b) The methods that will be used in the analysis.
- c) The output that should be generated in order to answer the question of interest.
- d) How the results will be presented and reported.

**Step 3:** Obtain, check and validate the data.

**Step 4:** Analyse the data and use the results to draw conclusions and inferences.

**Step 5:** Report the results.

The steps presented above are interrelated and it is not always possible to follow them exactly as has been shown. For example, an analysis to answer a specific question can only be done if the necessary data are available. Therefore, the aims and objectives that can be set are limited to those that can be answered by the data that are available. Similarly, a check of the data in Step 3 may

reveal problems with some of the variables that mean they cannot be used, and it may become necessary to change the aim of the analysis. The steps have been presented in a simplified manner as shown to isolate the different aspects that should be considered in undertaking statistical analysis of data.

## 2.2 Data and types of outcome

Data on transplant outcomes collected by transplant agencies is largely time-to-event data, where patients are followed up from a pre-defined start point, such as registration on the transplant list or the transplant itself, to an event of interest. The term ‘time-to-event’ refers to the length of time between the starting point and the event of interest. Patients may be followed up for different lengths of time and for some, the event of interest does not occur during the time they are followed-up. Observations on patients for whom the event does not occur during follow-up are said to be ‘censored’. The methods used for analysis must address these features of the data.

When the starting point is registration of a patient on the transplant list, the event of interest might be a transplant, removal from the list or death before transplant. When the starting point is the time of transplant, the event of interest might be failure of the grafted organ or death of the patient. We will focus on outcomes where the starting point is a transplant, although the methods used are applicable to outcomes of patients on the transplant list. There are three main types of post-transplant outcomes that are commonly analysed. These are (i) *patient survival*, the time to death from transplantation, (ii) *graft survival*, the time to failure of the graft, and (iii) *transplant survival*, the time to the earlier of graft failure or patient death.

## 2.3 Choice of cohort

Data stored by a transplant registry may contain a very large number of records for transplants undertaken over a long period of time. In addition, the data have a longitudinal structure, whereby each transplant has a record of information collected at each successive follow-up time. When carrying out an analysis, a cohort of patients must be selected with care from the database. In selecting a cohort, consideration must be given to several issues.

***Aim of analysis*** Before undertaking an analysis, the aim must be clearly defined as this determines what transplants will be included in the cohort. For example, the age groups, date of transplant donor and transplant type should preferably be defined, so that the cohort reflects the aims of the analysis. If the transplant characteristics are clearly defined then cohort selection will be relatively straightforward. In some cases judgement is required by the analyst about the most appropriate transplants to select. For example, an analysis to estimate the current survival rate may imply that only the most recent transplants should be included, but this must be balanced with having a cohort that is large enough to allow reliable estimation.

***Size of cohort*** The size of the cohort must be large enough to ensure reliable statistical inferences can be drawn from the analysis [1]. In determining the appropriate size, consideration is given to the number of observations and events. Note that reliability of the results is related to the number of events rather than the number of observations. If the event rate is small, a larger number of observations may be needed or a more common event may be considered [6]. For example, transplant survival may be analysed in place of patient survival for live donor kidney

transplants, where survival rates are high. The number of events may also be increased by using a longer follow-up period for each transplant so that longer-term outcomes are analysed, or by using a cohort that spans a longer period of time. Longer follow-up times and time periods may mean that less recent transplants are included and the results may therefore be less relevant to current experiences due to changes in medical practice. A trade-off between improved reliability of the results and relevance of the analysis is required.

***Follow-up period*** If analysis at a particular post-transplant period is specified e.g. 5-year survival, the cohort must contain some patients who were transplanted at least five years before the end of the time period being considered. For example, if 5-year survival is being analysed for transplants up to 2005, then the cohort should at least contain some transplants in 2000. However, not all transplants need to have had at least five years of potential follow-up; some may be censored less than five years after transplant. Methods for time-to-event data analysis are designed to deal with such censored observations.

## 2.4 Choice of method

The method used for analysis is determined by the aim of the analysis and the type of data to be analysed. Table 2.1 [7] provides a summary of how one might use the aim ('Goal' in Table 2.1) of analysis and the type of data, to choose a method of analysis. The term 'Gaussian population' in Table 2.1 refers to a variable whose statistical distribution is said to be 'Normal'. A histogram of a variable that follows a Normal distribution shows a symmetric pattern. There are statistical tests that can be used to determine whether or not a variable follows a normal distribution. A time-to-event variable, such as survival time does not follow a Normal distribution, and would therefore be described using a median and interquartile range rather than a mean and a standard deviation (SD). Details of the methods shown in Table 2.1 can be found in a number of statistical texts such as Fisher and van Belle (1993) [5], Sheskin (2000) [8], Collett (2003) [9,10] and Harrell (2001) [11].

## 2.5 Data checking and validation

The aim in checking and validating data is to identify and, if possible, rectify errors in the data. This is an important step of the analysis process since if the data are not correct, the results will not reflect the true nature of the data being analysed. There are different types of checks that can be done, some of which may not involve the person undertaking the analysis. However, it is important that the analyst be aware of the checks that are carried out elsewhere, so that they know what further checks they need to do.

***Check for typing errors*** Errors in typing can be checked at the data entry stage by using double data entry, for example, where the data are re-typed and the two versions compared.

***Validity checks*** A check that only valid values or codes have been recorded for each of the variables. For example, a negative age is not valid. Such errors become apparent when looking at data summaries and frequency tables, or plots such as scatter plots, dot plots and box and whisker plots, which show the entire range of values of a variable. Any clear outliers should be verified as they may represent genuine errors.

|  | <b>Type of Data</b>   |  |  |   |
|--|---|--|--|---|
| <b>Goal</b>  | <b>Measurement (from Gaussian Population)</b>               | <b>Rank, Score, or Measurement (from Non- Gaussian Population)</b> | <b>Binomial (Two Possible Outcomes)</b>      | <b>Survival Time</b>                        |
| <b>Describe one group</b>  | Mean, SD  | Median, interquartile range  | Proportion                                   | Kaplan Meier survival curve                 |
| <b>Compare one group to a hypothetical value</b>                 | One-sample <i>t</i> test                                    | Wilcoxon test  | Chi-square or Binomial test                  |   |
| <b>Compare two unpaired groups</b>                               | Unpaired <i>t</i> test                                      | Mann-Whitney test  | Fisher's test (chi-square for large samples) | Log-rank test or Mantel-Haenszel            |
| <b>Compare two paired groups</b>                                 | Paired <i>t</i> test  | Wilcoxon test  | McNemar's test                               | Conditional proportional hazards regression |
| <b>Compare three or more unmatched groups</b>                    | One-way ANOVA   | Kruskal-Wallis test  | Chi-square test                              | Cox proportional hazard regression          |
| <b>Compare three or more matched groups</b>                      | Repeated-measures ANOVA                                     | Friedman test  | Cochrane Q                                   | Conditional proportional hazards regression |
| <b>Quantify association between two variables</b>                | Pearson correlation   | Spearman correlation   | Contingency coefficients                     |   |
| <b>Predict value from another measured variable</b>              | Simple linear regression or Nonlinear regression            | Nonparametric regression   | Simple logistic regression                   | Cox proportional hazard regression          |
| <b>Predict value from several measured or binomial variables</b> | Multiple linear regression or Multiple nonlinear regression |  | Multiple logistic regression                 | Cox proportional hazard regression          |

<sup>1</sup> Excerpt from Motulsky (1995) [7]. Table available from: <http://www.graphpad.com/www/Book/Choose.htm>. Accessed 1 June 2006.

**Consistency checks** A check should be made that records are consistent across different variables. For example, a transplant date that is earlier than the date of birth of a patient would be an inconsistent record. Such checks can be incorporated into the data collection process by collecting crucial data in two different ways e.g. request both date of birth and age [5].

**Missing data checks** A check should be made for any data that may be missing. This should identify the variables for which there are missing observations and what proportion these missing data are. Thought should be given to why the data are missing and whether they really are missing. For example, data may not be available for a particular variable for all patients transplanted during a particular time period because no information was requested on the variable at that time. Consideration must then be given to whether the variable should be excluded from analysis. This may not be appropriate as the variable may have an important influence on the outcome. Excluding the patients with the missing data may introduce bias in the results. In some cases the data may be missing randomly across patients, thus allowing some patients to be excluded without biasing the results.

Transplantation data is usually collected retrospectively, for example from patients' hospital notes. In such cases, for information that is coded 'yes' or 'no', such as the presence of a particular symptom, for example diabetes, it is sometimes considered appropriate to replace missing values by 'no' on the grounds that the information would have been recorded if the symptom had been present. This assumption should not be made lightly or as a matter of course, as it may not necessarily be true [6].

Dealing with missing data is a complex issue to which careful thought needs to be given. More detailed consideration is given to ways of dealing with missing data in Chapter 7.

## 2.6 Discussion

Transplantation data are usually collected in an observational manner where no attempt is made to interfere with the transplant process or control any of the variables. Therefore any analysis done cannot show 'causality', but can only identify associations between variables. It is therefore necessary to incorporate subject matter knowledge provided by clinicians in determining whether any associations found are genuine or whether they may be due to the presence of unmeasured confounding variables.

## References

1. Schaubel DE, Dykstra DM, Murray S et al. Analytical approaches for transplant research, 2004. *American Journal of Transplantation* 2005; 5(Part 2): 950 – 957. Available from <http://www.ustransplant.org/publications.aspx> (Accessed 1 June 2006).
2. Rosati A, Salvadori M. Donor characteristics can influence overall transplant activities: The Italian experience. *Journal of Nephrology* 2003; 16:342 – 349.
3. Wolfe RA, Ashby VB, Milford EL. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *New England Journal of Medicine* 1999; 341: 1725 – 1730.

4. Evrard V, Otte JB, Sokal E et al. Impact of surgical and immunological parameters in paediatric liver transplantation: a multivariate analysis in 500 consecutive recipients of primary grafts. *Annals of Surgery* 2004; 239: 272 – 280.
5. Fisher LD, van Belle G. *Biostatistics: A methodology for the health sciences*. Wiley, New York, 1993.
6. Altman DG. *Practical Statistics for Medical Research*. Chapman & Hall / CRC, London, 1991.
7. Motulsky H. *Intuitive Biostatistics*. Oxford University Press, USA, 1995.
8. Sheskin DJ. *Handbook of parametric and non-parametric statistical procedures*. Chapman & Hall / CRC, New York, 2000.
9. Collett D. *Modelling binary data*. Chapman & Hall / CRC, London, 2003.
10. Collett D. *Modelling survival data in medical research*. Chapman & Hall / CRC, London, 2003.
11. Harrell FE. *Regression modelling strategies: With applications to Linear Models, Logistic Regression and Survival Analysis*. Springer, New York, 2001.

### 3 SUMMARISING AND MODELLING TRANSPLANT OUTCOMES

Transplant outcome data are largely of the ‘time-to-event’ type. However, for convenience, time-to-event outcome variables are sometimes converted to binary variables (referred to as binomial in Table 2.1) that allow only two possible outcomes. In this section we consider the methods that may be used to analyse these two data types and the issues related to such analyses.

#### 3.1 Summarising transplant outcomes

In this section, methods for summarising outcome following transplantation are reviewed.

##### 3.1.1 Binary outcomes

Time-to-event outcome variables are sometimes converted to binary type data, where unity (1) is recorded if, say, the event of interest, e.g. graft failure, occurs within a given time period, and zero (0) if the event does not occur during the specified time period. Records for patients whose follow-up terminates before the end of the time period, and whose times are therefore censored, are excluded, as it is not known whether the event has occurred or not, and they cannot therefore be classed as either a ‘1’ or a ‘0’. However, if any censoring occurs after the time point of interest, it will be known that the person was alive at that time.

Proportions, which measure the event rate, are used to summarise such data and statistical tests can be carried out to compare the event rate of one group to a particular value or to the event rate of one or more other groups. See Table 2.1 for the appropriate statistical test in each case. In this context, a group consists of patients with a particular characteristic in common. For example, one might wish to compare the mortality rate for patients from one centre to another, or the graft failure rate for heartbeating donor transplants to non-heartbeating donor transplants.

##### 3.1.2 Time-to-event outcomes

Time-to-event outcome data are largely summarised in one of two ways. The average length of time to the event of interest is estimated using the median, quoted with an interquartile range. The median time is the time that 50% of the patients are expected to reach before they experience the event. See Collett (2003) [1] for methods of estimating the median. The most common way of summarising time-to-event data is the estimation of ‘survival’ probabilities (see for example 1,2,3).

A survival probability is an estimate of the proportion of patients who would reach a given time point without experiencing the event. For example, one might wish to estimate the probability of patients surviving 5 years after a heart transplant. The methods used for estimating survival probabilities make allowances for patients who are censored and make use of their information up to the time they are censored. The Kaplan-Meier method is the most commonly used method for estimating survival rates in transplantation data and is the most suitable method when individual event times are reported. In cases where individual event times are unknown and the data are grouped into a series of time intervals, with the number of events and censored observations in each interval given, the actuarial or life-table method may be used [1,3].

The survival experience of two or more groups can be compared using various tests, the most common of which is the log-rank test. It may be necessary to compare survival across two or more groups after taking account of other confounding variables. For example, one may be interested in comparing the survival experience of patients who receive kidneys from heartbeating

and non-heartbeating donors. In this case, a comparison that accounts for donor age would be more reliable. When comparisons take account of other variables, they are said to be stratified by the additional variable accounted for. Stratified comparisons, e.g. a stratified log-rank test, give more precise summaries of the effect of the main variable of interest.

There are several incorrect approaches to the analysis of time-to-event data that are sometimes taken. Altman (1999) [3, pg 385 – 387] discusses and warns against these incorrect approaches.

### 3.1.3 Limitations of descriptive methods

The methods described above may be used to summarise transplant outcomes for a single group of patients, with the possibility of extension to a comparison of two or more groups. However, there may be additional variables that should be accounted for to ensure that the comparison reflects the true effect of the main variable of interest and is not confounded by the effect these other additional variables. When there are numerous additional variables to account for, one might be tempted to carry out several stratified log-rank test. There are two problems with such an approach. When numerous comparisons are made, some of the comparisons may falsely show a significant difference by chance. In addition, some of the variables that are used for stratification may be correlated and this would not be taken into account. A better procedure is to adopt a modelling approach that allows several variables to be accounted for simultaneously.

## 3.2 Modelling transplant outcomes

Statistical regression models are used to express the relationship between an outcome variable and several explanatory variables in a mathematical form as follows:

$$E[Y] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p,$$

where  $E[Y]$  represents the expected outcome or a transformation of the expected outcome,  $X_1, X_2, \dots, X_p$  represent  $p$  explanatory variables, and  $\alpha, \beta_1, \beta_2, \dots, \beta_p$  are fixed constant values called coefficients. This allows a patient's expected outcome to be calculated based on what their values of the explanatory variables are. The explanatory variables included in a model are often selected from a larger set of candidate variables using statistical methods. This process is referred to as model development. Explanatory variables are also referred to as predictors or risk factors. These terms will be used interchangeably.

Regression models have several uses, including:

**Effect estimation** One may wish to evaluate the effect of a particular factor on the outcome, while simultaneously accounting for the effect other variables that influence the outcome.

**Predicting outcomes** A regression model may be used to predict outcomes for individual patients. For example, a model may be used to predict post-transplant survival for patients on the transplant list, when allocating organs. Regression models may also be used in risk-adjustment, where the model is used to predict outcomes of patients for individual transplant teams in order to obtain an expected outcome rate. This expected outcome rate is then used to obtain risk-adjusted outcome measures to allow a fair comparison of outcomes across the teams.

The type of model used and the way in which it is developed depends on the type of outcome variable and the purpose of the model. However, for all models there is a general strategy to follow when modelling outcome variables.

### *3.2.1 General modelling strategy*

The process of obtaining a model for an outcome variable can be divided into two main stages of model development and model validation. The following steps may be followed when developing a model.

***Statement of clinical aim*** The aim of the modelling will influence the choice of candidate risk factors. For example, if the aim is to compare transplant teams, then there may be no need for a complex model as a model that accurately predicts overall numbers of events for different transplant teams will be adequate.

***Identify potential risk factors*** This can be done using clinical knowledge or published literature. The choice of potential risk factors depends on the purpose of the model. For example, if the purpose of the model is to predict post-transplant outcome, then variables that are measured after the transplant cannot be used.

***Select a suitable cohort*** The number of events in the cohort must be large enough to allow reliable selection of a model and to avoid overfitting. It is advised that there should be at least 10 events for each *candidate* risk factor variable [4]. The number of candidate risk factors must include the number of categories for categorical variables and any nonlinear terms and interactions.

***Treatment of missing values*** Plans must be made about how missing values in the risk factors will be dealt with.

***Select final set of predictors*** The final set of predictors should be selected in a systematic manner that includes a check that the model selected fits the data well. See, for example, [4] for a discussion of methods for predictor variable selection.

***Fit final model*** Once the final set of predictors has been selected, the risk factor coefficients and any other model quantities required can be calculated.

When a model is developed from data for one set of patients, it may make poor predictions for a different set of patients. Model validation is the process of establishing that a model “works satisfactorily for patients other than those from whose data it was developed” [5] and is an essential part of the modelling process. Validation looks at how well the model predicts outcomes and discriminates between patients at high and low risk of having an event. There are several ways in which model validation can be done. These will be mentioned briefly here. See for example [4,5] for a detailed discussion of model validation methods.

**Data-splitting** A data set is split at random into a training subset and a test subset. The model is developed using the training set and validated on the test set, and the two subsets are combined for fitting the final model. The main disadvantage in using this method is that the data set for model development is greatly reduced in size. This may be a problem, particularly if the event of interest has a low rate of occurrence. In addition, the model that is validated is not the final model, but that developed from the test set.

**External validation** Here the model is validated using a completely independent set of data. For example, a model may be developed using data from one centre and validated on data for patients from another centre.

**Re-sampling** There are other techniques, such as the bootstrap and cross-validation, in which the model development is done on the complete data set and then the modelling process is repeated on multiple re-samples from the original data set to mimic the process that produced the final model [4]. For the bootstrap, re-sampling is done by sampling with replacement to give a data set of the same size as the original, so that some observations appear more than once. In cross-validation, the re-sampling is done by leaving out one or more observations at a time, and then predicting the outcome for the omitted patients using a model developed from the remaining observations.

### *3.2.2 Modelling binary outcomes*

When the outcome variable is binary in nature, the type of model used is a logistic regression model. This expresses the risk (or probability) of an event occurring in terms of the factors that are included in the model. If a factor has a positive coefficient in the model, the factor is assumed to be associated with increased risk of the event occurring. A negative coefficient is associated with a reduced risk. The effects of individual factors are often expressed as odds ratios. Details of how logistic models are developed and the specific considerations that should be made are contained in various statistical texts. See for example [3,4,6].

When evaluating centre differences in outcome rates, a variable that identifies the centre that each patient belongs to can be included either as a fixed effect variable or a random effect variable. A random effect variable is preferred as it accounts for correlation that exists between patients from the same centre. However, when the number of centres is small, the estimate of the level of variation between centres may be unreliable and a fixed effect variable may be more appropriate. When the number of centres is large, a random effect variable is preferred, as a fixed effect model would result in a large increase in the number of coefficients that require estimation.

### *3.2.3 Modelling time-to-event outcomes*

There are several types of models that can be used to analyse time-to-event outcomes. These models are referred to as survival models. The most commonly used model is the Cox proportional hazards model. Survival models express the hazard of the event occurring in terms of the factors included in the model. The hazard of an event is the probability that the event occurs at a given time point, on condition that it has not occurred before that time point. If a factor has a positive coefficient in the model, the factor is assumed to be associated with increased hazard for the event. A negative coefficient is associated with a reduced hazard. The effects of individual factors are often expressed as hazard ratios. Details of other types of survival models

and how they are developed and the specific considerations that should be made are contained in various statistical texts. See for example [1,3,4].

Evaluation of centre differences is more complex for survival models than for logistic models, particularly when the number of centres is large. To account for differences among centres without evaluating them, a model that is stratified by centre can be used. To explicitly evaluate centre differences in outcome rates, a variable that identifies the centre that each patient belongs to can be included either as a fixed effect variable or a random effect variable. The inclusion of a random effect variable in a model gives a more complex model called a frailty model. A fixed effect model can also be used.

### 3.3 Discussion

In this section we presented a summary of methods that can be used to summarise and model transplant outcomes. Transplant outcome data are largely of the ‘time-to-event’ type but are sometimes converted to binary data. When the data are converted in this manner, the observations that are censored before the time point of interest are excluded. Such a conversion is not suitable for long-term outcomes as there is likely to be a larger number of censored observations. In addition, information about the length of time to the event is lost and there is therefore no differentiation between events that occur shortly after the transplant and those that occur a long time after the transplant. Time-to-event analysis is therefore preferred.

### References

1. Collett D. *Modelling survival data in medical research*. Chapman & Hall / CRC, London, 2003.
2. Fisher LD, van Belle G. *Biostatistics: A methodology for the health sciences*. Wiley, New York, 1993.
3. Altman DG. *Practical Statistics for Medical Research*. Chapman & Hall / CRC, London, 1991.
4. Harrell FE. *Regression modelling strategies: With applications to Linear Models, Logistic Regression and Survival Analysis*. Springer, New York, 2001.
5. Altman DG, Royston P. What do we mean by validating a prognostic model. *Statistics in Medicine* 2000; 19: 453 – 473.
6. Collett D. *Modelling binary data*. Chapman & Hall / CRC, London, 2003.

## 4 RISK FACTORS

Statistical regression models are used to express the relationship between an outcome variable and several explanatory variables or risk factors in a mathematical form. This allows the effect of each risk factor on the outcome to be evaluated, with the effects of other factors that also influence the outcome being accounted for. These models have several uses, including risk adjustment when comparing outcomes across different transplant teams, or to assess the effect of a particular factor on outcome.

A statistical model is however, an imperfect tool. The factors included in the model are limited to those on which data have been collected. Furthermore, not all factors for which data are available should be included in a statistical model. For example, when developing a model for risk adjustment for use in comparing outcomes across centres, factors that are under centres' control should not be adjusted for, as this will mask differences that arise from different centre practices.

The aim in this section is to provide a minimum set of risk factors that should be considered when developing statistical models from transplant outcome data. The list of factors given should not be regarded as prescriptive: much depends on the context of the analysis.

### 4.1 Potential risk factors

In Deliverable 5.1 of Work package 5 of the Alliance-O project, each country was asked to identify the factors used in their statistical models. The factors that were common to all countries are presented in Table 1 as a minimum set of factors that should be included in a model for the organ specified. Any other factors used are included in the column for 'Other factors'. Deliverable 5.1 also included some discussion on risk factors and differences between countries. Consequently, the table of risk factors is presented without comment in this chapter.

| <b>Organ</b>  | <b>Minimum set</b>  | <b>Other factors</b>  |
|---------------|---|---|
| <b>Kidney</b> | <p><i>Recipient:</i> Age, Gender, Number of previous transplants, Primary diagnosis</p> <p><i>Donor:</i> Age, Gender</p> <p><i>Other:</i> Cold ischaemia time, HLA mismatch</p> | <p><i>Recipient:</i> Blood group, Ethnicity, Peak PRA, Waiting time, Time on dialysis, Diabetes</p> <p><i>Donor:</i> Blood group, Type of donor, Cause of death</p> <p><i>Other:</i> Donor-recipient gender match, Year of graft, Type of transplant, Organ origin (local/imported), Days of delayed graft function, case mix</p>   |
| <b>Liver</b>  | <p><i>Recipient:</i> Age, Primary diagnosis</p> <p><i>Donor:</i> Age</p> <p><i>Other:</i></p>   | <p><i>Recipient:</i> Gender, Blood group, Number of previous grafts, Waiting time, Urgency status, Gravity of illness, C Virus marker, Weight</p> <p><i>Donor:</i> Gender, Cause of death, Quality of donor, Type of donor</p> <p><i>Other:</i> Blood group compatibility, Year of graft, Type of graft, Ischaemia time, Type of transplant, Centre activity, Type of bypass, Organ origin (local/imported)</p>   |
| <b>Heart</b>  | <p><i>Recipient:</i> Age, Gender, Primary diagnosis</p> <p><i>Donor:</i> Age, Gender</p> <p><i>Other:</i></p>   | <p><i>Recipient:</i> Blood group, Urgency status, Previous heart surgery, Transpulmonary gradient (mmHg) risk, Waiting time, Vascular resistance, Size (BMI, BSA, height or weight), Creatinine clearance, Diabetes, Ventilation, Peripheral vascular disease</p> <p><i>Donor:</i> Blood group, Cause of death, Quality of donor, Size (BMI, BSA, height or weight), Diabetes, Drug abuse</p> <p><i>Other:</i> Donor-recipient gender match, Donor-recipient blood group match, Year of transplant, Ischaemia time, Type of transplant, Donor-recipient size match, Donor-recipient CMV match</p> |
| <b>Lung</b>   | <p><i>Recipient:</i> Age, Gender, Primary diagnosis</p> <p><i>Donor:</i> Age, Gender</p> <p><i>Other:</i></p>   | <p><i>Recipient:</i> Blood group, Number of previous grafts, Size (BMI, BSA, height or weight), Waiting time, Vital capacity, Forced expiratory volume, PaO<sub>2</sub>, PaPm pre-transplant</p> <p><i>Donor:</i> Blood group, Cause of death, Quality of donor, Size (BMI, BSA, height or weight)</p> <p><i>Other:</i> Donor-recipient gender match, Donor-recipient blood group match, Donor-recipient size match, Donor-recipient CMV match, Year of transplant, Ischaemia time, Type of transplant</p>  |

## 5 COMPARING OUTCOME RATES ACROSS CENTRES

Centre-specific data on post-transplant outcomes are now routinely made publicly available in some countries in response to increasing calls for more open reporting of the quality of patient care. For example in the UK [1] and USA [2], post-transplant survival rates for individual centres are routinely published on the Internet. However, the publication of centre-specific data inevitably leads to comparisons being made and is open to misinterpretation by the media, patients and clinicians themselves if not carefully done. Indeed it is feared that publication of such data may unwittingly lead to ‘gaming’, whereby clinicians avoid high-risk patients in an attempt to achieve favourable outcome measures [3]. For this reason it is crucial that centre comparison data are presented in “an easily interpretable fashion, so as to encourage rigorous, but fair evaluation of clinical performance, whilst highlighting pitfalls in interpretation”[4].

In this section we present methods that can be used for presenting comparisons of transplant centres with respect to quantitative outcome measures. The methods we present use grouped data and are more suitable for a retrospective, rather than prospective, comparison. They are suitable for analysing short or long-term outcomes.

### 5.1 Variation in outcome measures

When a quantitative outcome measure, for example overall graft survival, is listed for a group of centres, the values for individual centres will almost always differ. The causes of the differences seen may be classed as follows [5]:

- a) Random fluctuations – these are differences that occur in the outcome measure due to natural chance variation, whereby the outcome measures differ even if all centres have acceptable outcomes.
- b) Case-mix – centres may serve communities that differ with respect to the distribution of risk factors. For example, one centre may have proportionately more diabetic kidney transplant patients than another.
- c) Unmeasured factors – these are (known or unknown) factors that are not easily measurable but influence outcome, such as centre policies in determining eligibility for transplant. These may include other differences in case-mix for which data are not available.
- d) Unknown causes of differences that are within the centre’s control and are associated with only a few centres.
- e) Nationwide changes in practice that lead to changes in outcome rates in all or some centres. For example, changes in the organ allocation scheme or introduction of improved immunosuppressive medication may lead to a general improvement in outcomes in all centres.

In comparing centres, the aim is usually to detect differences that result from any unmeasured factors (c) and unknown causes (d), while adjusting for random fluctuations (a), differences in case-mix (b) and changes in practice (e).

Identification of centres that seem to differ from others, the so-called ‘divergent’ centres, is only the first step. It is recommended that if a centre is identified as being divergent, investigations are carried out together with the centre to determine the cause for the difference. In some cases, an acceptable explanation is found and in others, the centre may already be aware of an existing issue and may have taken steps to address it.

## 5.2 Outcome measures

Outcome measures take different forms depending on the type of data being analysed. They can be an average of a continuous variable (e.g. serum creatinine), a rate for binary or time-to-event outcomes (e.g. mortality), or a count (e.g. number of rejection episodes). The outcome measures used by Alliance-O partners tend to be for time-to-event data, which are at times converted to binary data [see Deliverable 5.1]. We therefore focus on the outcome measures related to these two data types.

### 5.2.1 Unadjusted outcome measures

Depending on the type of outcome considered, centres may be compared with respect to a rate of the form  $y_j / n_j$ , derived from binary data, where  $y_j$  denotes the number of patients who have the outcome of interest out of a total of  $n_j$  patients, and  $j$  denotes the  $j^{\text{th}}$  centre. Centres may also be compared with respect to a percent ‘survival’ derived from time-to-event data. To account for random fluctuations, confidence intervals of these outcome measures are often used. However, such a comparison does not adjust for differences in case-mix, which in most cases will explain the variation seen from one centre to the next. Comparison of unadjusted measures is to be avoided where possible as it might lead to centres being penalised for treating high-risk patients. In some cases, data may not be available to allow adjustment for case-mix to be done. In such instances, it is important to highlight the lack of adjustment and the implications of any differences that may be seen.

### 5.2.2 Risk-adjusted outcome measures

To account for differences in case-mix, risk-adjusted outcomes measures can be used [see for example 1,2,6]. Two types of risk-adjusted outcome measures are often used. The first, called the Standardised Mortality Ratio (SMR) when comparing mortality rates, is a ratio of the observed to the expected outcome rate [2,7]. The SMR gives a measure of how the outcome rate observed at a centre differs from that expected. Confidence intervals about the ratio are normally used to determine whether any difference seen between the two is statistically significant. If the confidence interval includes unity (observed and expected outcome rates are equal) then there are no statistically significant differences between the two rates. A second type of measure is a conversion of the SMR-type measure into a risk-adjusted rate (Risk-adjusted Mortality Rate) [6,8]. This is achieved by multiplying the ratio of observed to expected outcome by a crude national rate and gives an estimate, based on the risk-adjustment scheme, of what the outcome rate at a centre would have been had they had the same case-mix as that seen nationally.

### *5.2.3 Calculation of expected outcome rates*

The expected outcome rate for a centre can be calculated as an average of individual predicted risks for its patients. Individual patient risks can be calculated using an existing risk score (e.g. EuroSCORE for cardiac surgery) [4] or using a regression model developed from national data. In risk scoring, patients achieving a score within a given range have an associated risk of experiencing the adverse outcome. Such risk scores, if available, need to be regularly updated to reflect changes in the relative impact of the risk factors on outcome. In addition, assessments are needed to ensure that the risk score, which may have been developed in a different country, is applicable to the patients whose data are being analysed. When a statistical model is used, individual predicted risks are calculated using the mathematical representation of the model.

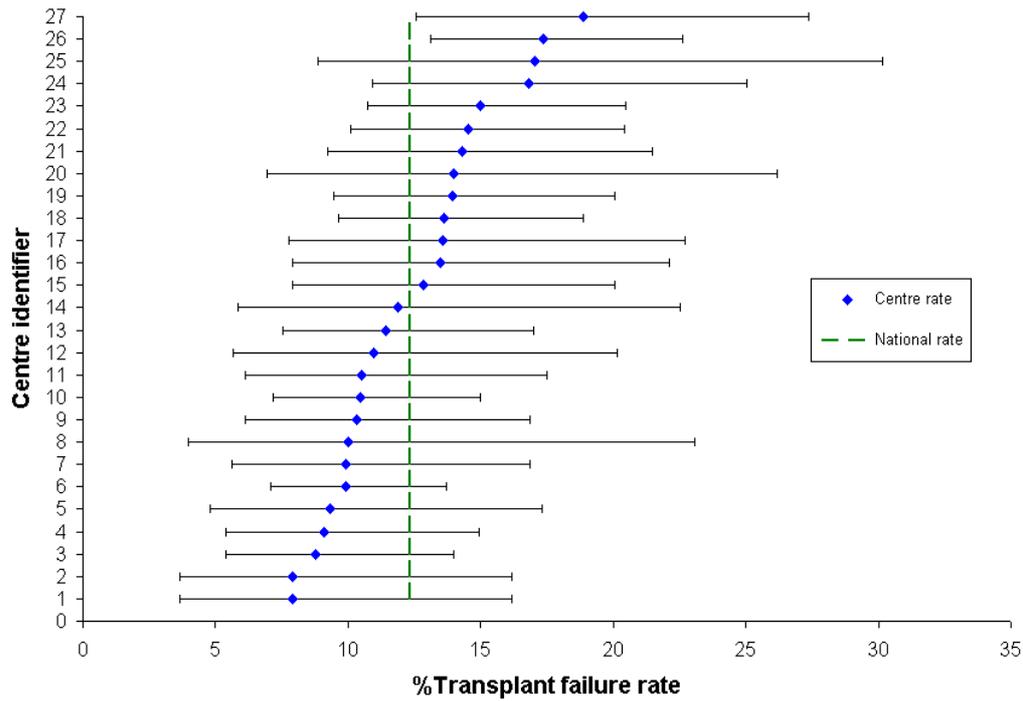
## **5.3 Presentation of results**

Centre-specific outcome measures can be presented in table or graphical format. If the number of centres involved is large, then a table may not be the best approach. In this case, a graphical format is preferred.

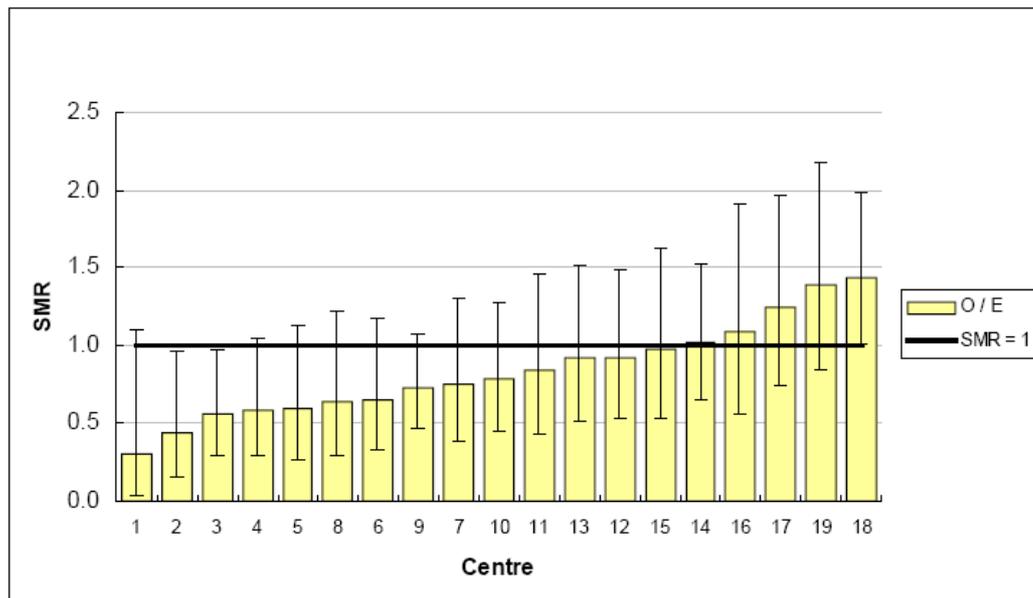
### *5.3.1 League tables*

League tables have until recently, been the most common approach for comparing outcome measures across hospitals or surgeons. A league table lists the outcome measure of interest, such as a mortality rate, in order from the highest ranked to the lowest ranked. The outcome measure may be unadjusted or risk-adjusted and confidence intervals may be used to account for random variation. Examples of graphical league tables are shown in Figures 5.1 and 5.2.

League tables have been criticised for their use of ranking [9]. If, at the end of a 5-year period, two hospitals have the same rate of graft failure, they would be given the same rank. However, if the graft failure rates were calculated on a yearly basis then on any given year, one centre would, due to random variation, be ranked above the other by chance. The true rank of a centre is difficult to estimate and the issue of constructing confidence intervals of these ranks has not been addressed well [10]. Furthermore, interpretation of the results may not be clear. For example, in Figure 7.1 the confidence intervals for centres 26 and 27 do not include the national average. Should these centres be considered to have considerably high mortality rates even though their confidence intervals overlap with those of other centres?



**Figure 5.1** ‘Caterpillar plot’ comparing transplant failure rates across centres in order



**Figure**

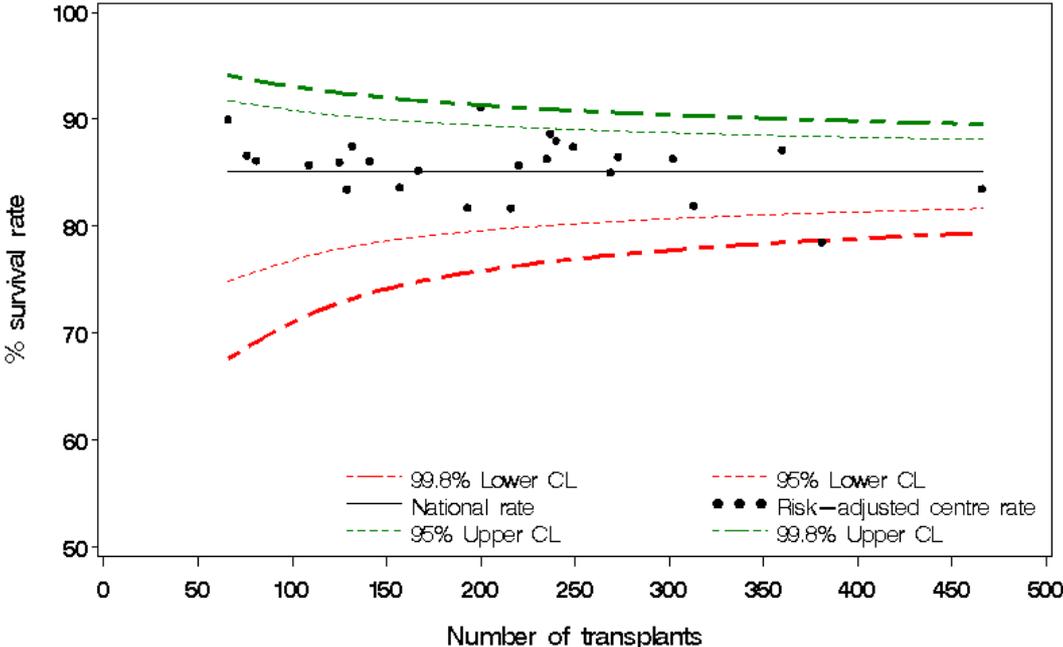
**Figure 5.2** A plot showing ranked standardised mortality ratios. Source: Keogh and Kinsman (2002) [4]

An alternative tool to league tables, the funnel plot, has been proposed [11,12].

### 5.3.2 Funnel plots

A funnel plot is a graphical display that can be used to compare clinical outcome measures across centres or individual surgeons [11-14]. The centre outcome rates are plotted against numbers of

patients and the national average rate and its confidence limits are superimposed. A centre that falls outside the limits is taken to have an outcome rate that is significantly different from the national average. 1 in 20 centres will fall outside the 95% confidence limits by chance and for the 99.8% limits, the corresponding figure is 1 in 1000. The limits are based on numbers of patients at each centre, so that the width of the confidence interval decreases as the number of patients increases, thus forming a funnel. By presenting outcome rates in this manner, normal variation is displayed and those centres that diverge significantly from the national average are easily identified. An example is shown in Figure 5.3.



**Figure 5.3** Funnel plot of risk-adjusted survival rates

There are several advantages in using the funnel plot. The spurious ranking seen in league tables is avoided. By focusing on whether a centre lies within the confidence limits or not, attention is drawn to identifying divergent centres, rather than on which the best and worst centres are. This gives a clear indication when investigation of a centre may be required and helps to direct audit exercises. Secondly, the larger variation normally seen in outcome rates of smaller centres is taken into account by having wider confidence intervals as the transplant volume decreases.

**5.4 Discussion**

It is important that centres whose post-transplant outcome rates are not inline with what is expected for them are identified and investigations carried out to determine the cause. Outcome measures are often compared across centres to identify such centres. However, any such comparison exercise should incorporate adjustment for differences in case-mix and the limitations of the comparison should be made clear. In particular, it may not be possible to adjust for all pertinent risk factors and therefore any differences seen may reflect variation in unmeasured risk factors as well as quality of care. For that reason, any attempt to compare centres can only be a

first step in trying to identify centres whose outcome rates are truly not as good as they should be. It is only through investigations involving the centre concerned that 'poor performance' can be confirmed.

Risk-adjustment is an imperfect tool. Only those factors on which data are available can be adjusted for. Furthermore, there may be other factors that influence outcome that are not easily measurable and therefore cannot easily be adjusted for. Despite these limitations, it is preferable to use an imperfect risk-adjustment model than none at all.

It is recommended that any methods that are based on the ranking of centres are avoided and that funnel plots are a suitable alternative to ranking. Funnel plots help focus the comparison on identifying centres whose outcome rates are not as good as expected in relation to the experience of patients seen throughout the country, rather than in classifying centres as the 'best' or 'worst'. To ensure transparency of the comparison exercise, it is recommended that any exercise to compare centres and any risk-adjustment models used, be developed in consultation with the centres or individuals being compared.

## References

1. UK Transplant centre-specific report website:  
[http://www.uktransplant.org.uk/ukt/statistics/centre-specific\\_reports/centre-specific\\_reports.jsp](http://www.uktransplant.org.uk/ukt/statistics/centre-specific_reports/centre-specific_reports.jsp). Accessed 8 May 2006.
2. US Scientific Registry of Transplant Recipients website:  
[http://www.ustransplant.org/annual\\_reports/current/chapter\\_viii\\_AR\\_cd.htm](http://www.ustransplant.org/annual_reports/current/chapter_viii_AR_cd.htm). Accessed 8 May 2006.
3. Goddard M, Davies HTO, Dawson D, Mannion R, McInnes F. Clinical performance measurement: part 2 – avoiding the pitfalls. *Journal of the Royal Society of Medicine*, 2002; 95: 549 – 551.
4. Keogh BE, Kinsman R. The Society of Cardiothoracic Surgeons of Great Britain and Ireland National Adult Cardiac Surgical Database Report, 2000 – 2001, Dendrite Clinical Systems Ltd, Henley-on-Thames (2002) p 200 -223.
5. Marshall C, Best N, Bottle A, Aylin P. Statistical issues in the prospective monitoring of health outcomes across multiple units. *Journal of the Royal Statistical Society A* 2004; 167: 541 – 559.
6. New York State Department of Health. *Adult Cardiac Surgery in New York State, 2001 – 2003*. October 2005. Available at:  
[http://www.health.state.ny.us/nysdoh/heart/heart\\_disease.htm](http://www.health.state.ny.us/nysdoh/heart/heart_disease.htm). Accessed 8 May 2006.
7. Armitage B, Berry G, Matthews JNS. *Statistical Methods in Medical Research* 4<sup>th</sup> ed. Blackwell Science, Oxford, 2002. pp 660 – 667.
8. Goldstein H, Spiegelhalter DJ. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society A* 1996; 159: 385 – 443.
9. Adab A, Rouse AM, Mohammed MA, Marshall T. Performance league tables: the NHS deserves better. *British Medical Journal* 2002; 324: 95 – 98.
10. Woodall WH. The use of control charts in health-care and public health surveillance. *Journal of Quality Technology*, 2006; 38: 88 – 102.

11. Stark J, Gallivan S, Lovegrove J, Hamilton JRL, Monro JL, Pollock JCS, Watterson KG. Mortality rates after surgery for congenital heart defects in children and surgeons' performance. *Lancet* 2000; 355: 1004 – 1007.
12. Spiegelhalter DJ. Letter to the editor: Funnel plots for institutional comparisons. *Quality and Safety in Health Care* 2002; 11: 390 – 391.
13. Tekkis PP, McCulloch P, Steger AC, Benjamin IS, Poloniecki JD. Mortality control charts for comparing performance of surgical units: validation study using hospital mortality data. *British Medical Journal* 2003; 326: 786 – 788.
14. Spiegelhalter D. Funnel plots for comparing institutional performance. *Statistics in Medicine* 2004; 24: 1185 - 1202.

## 6 MONITORING TRANSPLANT OUTCOMES WITHIN CENTRES

In this section we focus on methods that can be used to detect as quickly as possible when increases in rates of adverse outcomes occur. The focus is on monitoring outcomes within centres rather than a comparison across centres as presented in the previous section. The methods we consider are sequential in nature and if applied prospectively, can be used for ‘real-time’ surveillance.

### 6.1 Rationale for sequential monitoring

Under clinical governance requirements, National Health Service (NHS) organisations in the UK are accountable for ‘continuously improving the quality of patient care and developing the capacity of the NHS in England to maintain high standards (including dealing with poor professional performance)’ [1]. There are similar requirements to maintain high standards of quality of patient care in other countries e.g. Spain, and one of the ways in which this is achieved is through clinical audit [2]. In spite of this requirement, the Enquiry into paediatric cardiac surgery at the Bristol Royal Infirmary [3] found that one of the issues was a lack of a ‘systematic mechanism for monitoring the clinical performance of healthcare professionals or of hospitals’. It was recommended that effective systems be put in place within hospitals to ensure that clinical performance is monitored and that ‘there must also be a system of independent external surveillance to review patterns of performance over time and to identify good and failing performance’.

Sequential monitoring tools provide a solution for continuous on-going monitoring of health outcomes. These tools use data sequentially, adding the information from each new patient as it becomes available. Therefore decisions about changes in outcome rates may be made after each successive patient and, if the tools are implemented in a prospective manner, changes in outcome rates can be detected in a timely manner. An additional advantage of these tools is that they can be adapted to adjust for individual patient pre-operative risk, therefore allowing the monitoring to be done with adjustment for high-risk cases. Therefore those units that undertake experimental procedures with a high risk of failure or those units that generally undertake high-risk transplants can be assured that the monitoring procedures are able to take this in to account.

### 6.2 Methods for sequential monitoring

Statistical Process Control (SPC) methods, which have their origin in quality control of industrial manufacturing processes, have recently been adapted for use in monitoring of health outcomes such as mortality following vascular surgery [4]. In transplantation, these methods have been used in the UK to monitor 30 and 90-day mortality [5] and in the USA to monitor 1-year mortality and graft failure rates [6,7]. In France, monitoring procedures based on these methods are currently being developed. In Spain, similar methods for monitoring health outcomes such as graft and patient survival are being used for liver transplants. These methods have been developed by scientific societies in collaboration with ONT and will be extended to kidney, lung and pancreas transplant outcomes.

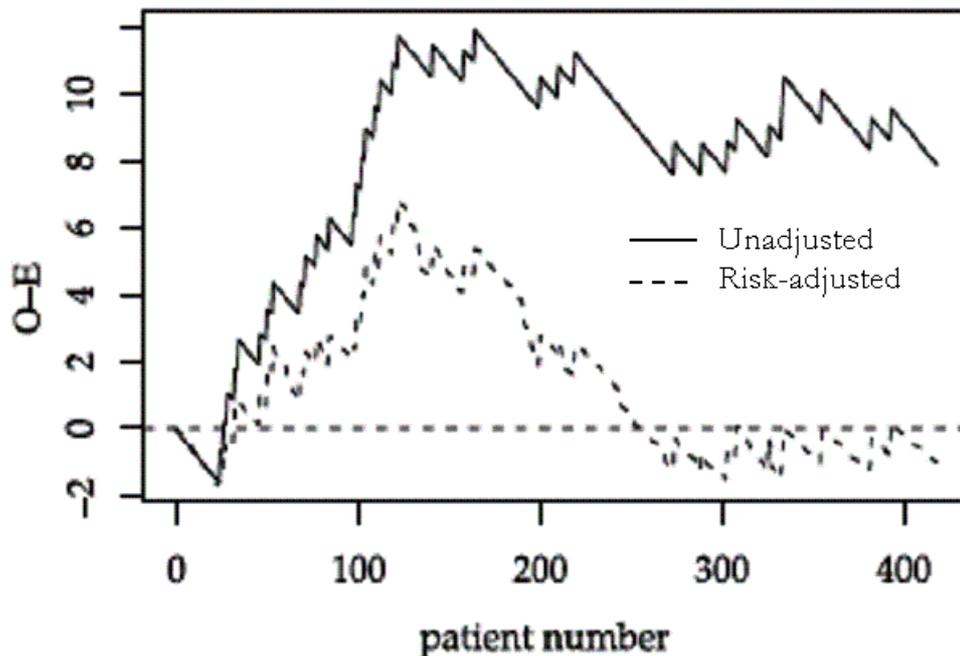
There are a considerable number of SPC tools that can be used and some of those most widely used for monitoring health outcomes are described in the next section.

### 6.2.1 Observed – Expected CUSUM chart

The Observed – Expected ( $O - E$ ) Cumulative Sum (CUSUM) chart plots the cumulative difference between the observed and expected number of adverse events and offers an intuitively appealing way of presenting trends in outcome rates over time.

Consider, for example, that 90-day graft failure rates are being monitored at a single centre. Each graft that fails within 90 days is assigned a value of 1 and a graft that functions for more than 90 days is assigned a value of 0, and this denotes their observed outcome,  $O$ . For each transplant, there is an associated pre-operative risk that the graft will fail within 90-days, and this denotes the expected outcome,  $E$ . When the analysis is risk-adjusted, the value of  $E$  varies from patient to patient. When the analysis is unadjusted, the value of  $E$  is the same for all patients.

To construct the  $O - E$  chart, a cumulative sum of the quantity  $O - E$  is obtained for each successive transplant and plotted against the patient number. The date of transplant may be used in place of the patient number. An example illustrating unadjusted and risk-adjusted plots is shown in Figure 6.1.



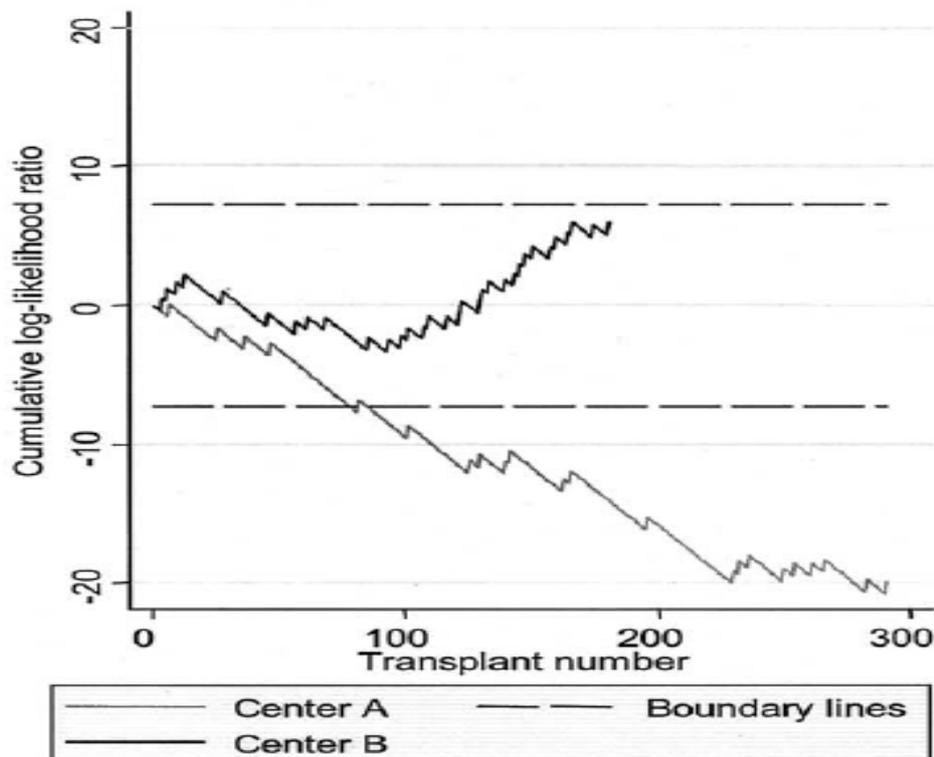
**Figure 6.1** Example of an  $O - E$  chart showing unadjusted and risk-adjusted plots. Source: Grigg and Farewell (2004) [10]

If the observed number of adverse outcomes is similar to that predicted from the expected outcomes, the plot would fluctuate around 0. If more failures occur than are expected, the curve will have an upward trend. Conversely, if fewer failures occur than predicted, the curve will have a downward trend. The  $O - E$  chart does not provide a statistical test that can be used to

determine whether the number of adverse outcomes observed is significantly greater or lower than that expected. The Sequential Probability Ratio Test (SPRT) and tabular CUSUM chart provide such a test.

### 6.2.2 SPRT and re-setting SPRT (RSPRT) charts

The sequential probability ratio test (SPRT) chart is based on a statistical test of a null hypothesis that the outcome rate is at an acceptable level against an alternative hypothesis that the outcome rate is unacceptably high or low. When a risk-adjusted chart is used, the hypotheses are based on an odds ratio rather than the outcome rate itself. A quantity, called the log-likelihood ratio, which assesses which of the two hypotheses the data support the most is calculated, added cumulatively for each new patient and plotted against the patient number. If the data are in favour of the alternative hypothesis, suggesting that the outcome rate may no longer be at an acceptable level, there will be an upward trend in the chart. If the data are in favour of the null hypothesis, suggesting that the outcome rate is at an acceptable level, there will be a downward trend. Thresholds are used to indicate when sufficient evidence has been accrued to support one hypothesis in favour of the other. The thresholds are based on probabilities of Type I (false positive) and Type II (false negative) errors. Once one of the thresholds has been crossed, the test terminates. Spiegelhalter et al (2003) [8] gave a readable account of this method and an example of an SPRT plot is shown in Figure 6.2.



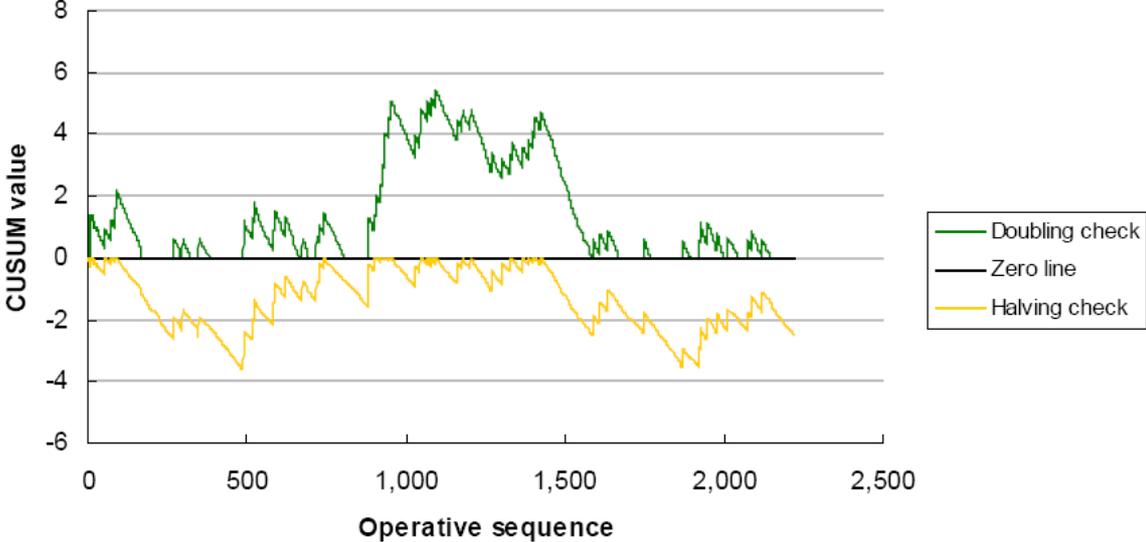
**Figure 6.2** Example of an SPRT chart. Source Rogers et al. 2004 [5]

The SPRT chart presents an appealing approach, particularly to clinicians and statisticians familiar with hypothesis testing, because of the use of error probabilities in setting the thresholds. It has

been noted, however, that the SPRT chart has the disadvantage of allowing a centre that has acceptable outcomes to build up ‘credit’ during a good run. This is because the natural progression for such a centre is downwards towards the lower boundary. This could cause a delay in detecting important changes in outcome rates. Furthermore, the SPRT chart is not suitable for continuous monitoring as it is more of a decision-making tool designed to test hypotheses only once [9].

6.2.3 *Tabular CUSUM chart*

The tabular CUSUM chart is, like the SPRT and RSPRT charts, based on a log-likelihood ratio test [13, 15 (theoretical background)] and also has a risk-adjusted analogue. In monitoring health outcomes, a “one-sided” chart is typically used to detect increases in rates of adverse outcomes [for example 2,6]. It is also possible to use two-sided charts that detect both increases and reductions in the rate being monitored [for example 13]. Figure 6.3 shows an example of a risk-adjusted tabular CUSUM in which the top part is used to detect increases in the odds ratio and the lower part detects reductions in the odds ratio.



**Figure 6.3** Example of a two-sided risk-adjusted tabular CUSUM. The top part tests for a doubling of the odds ratio and the bottom part tests for a halving of the odds ratio. Source: Keogh and Kinsman (2002) [14].

Only one threshold is specified for each chart. When this threshold is crossed, the chart is said to ‘signal’ or ‘trigger’ that the rate being monitored may have changed. The cumulative sum of the log-likelihood ratio score is reset to zero whenever it goes below or above zero, depending on whether the chart is designed to detect increases or decreases. This property of the tabular CUSUM makes it quicker at detecting changes in outcome rates than the SPRT or RSPRT [9]. The value of the threshold is chosen in such a way that the number of false signals is minimised without significantly increasing the chance of not detecting a centre with unacceptable performance.

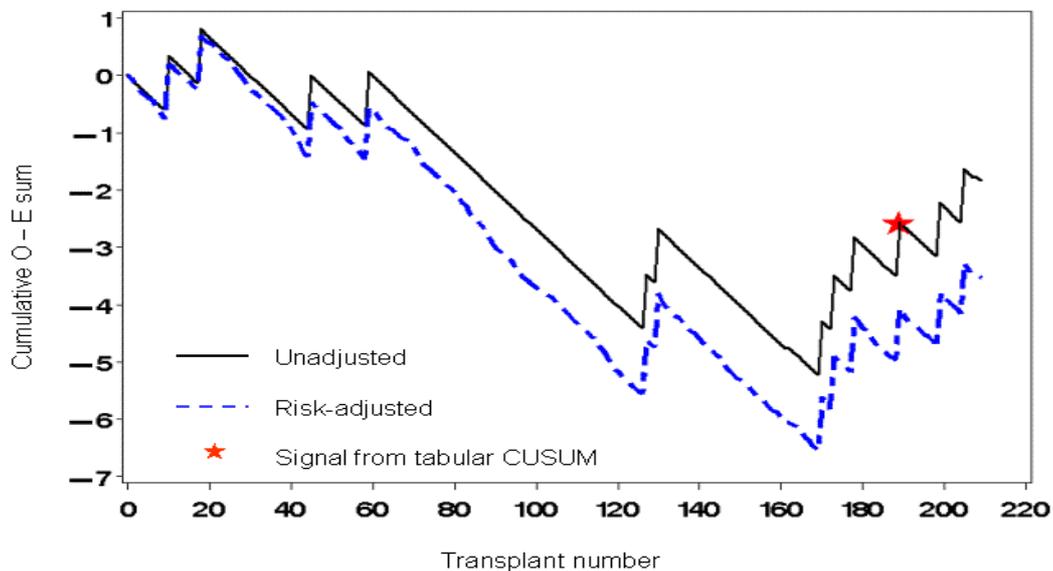
The performance of the tabular CUSUM chart is assessed using the average run length (ARL), the average number of observations seen before a signal occurs. The ARL is typically evaluated using simulations. Some software has been provided to evaluate the ARL for different types of outcomes [15].

### 6.2.4 Other methods

Other methods that can be used for sequential monitoring of health outcomes have been developed. These include the Sets method [16, 17] and the Scan statistic [18], which is a modification of the Sets method. The Sets method is used for monitoring rates of rare adverse outcomes and may not be suitable for monitoring transplant outcomes. The Scan statistic has recently been proposed as an additional tool for monitoring transplant outcomes [19].

### 6.2.5 Choice of method

It is recommended that more than one method be used in order to aid interpretation. The tabular CUSUM plots the same quantity as the SPRT and RSPRT, but has the advantage of not allowing a build up of excess ‘credit’ and is recommended. The  $O - E$  chart is more easily interpretable than any of the log-likelihood based charts. The  $O - E$  chart can be used to present the results, with the tabular CUSUM running in the background to provide a formal trigger mechanism. An example of such a combination is shown in Figure 6.4. Risk-adjusted analogues of the charts are recommended for use where possible. This may be in addition to or in place of the unadjusted charts.



**Figure 6.4**  $O - E$  chart with position of signal (\*) from unadjusted tabular CUSUM superimposed

## 6.3 Planning and setting up a monitoring procedure

Before introducing a procedure for monitoring outcomes, consideration should be given the issues listed below. It is recommended that the centres being monitored are consulted on these issues before the procedure is set up and that a protocol based on the issues is written at that stage [20]. The protocol may include:

- a) The objectives of the procedure.
- b) Outcomes and patients to be monitored – clinical input in this is important as the outcomes that are most appropriate to monitor differ depending on the type of organ transplant.
- c) Construction of charts – careful consideration needs to be given to the choice of parameters used for the charts. In particular, when multiple centres are monitored simultaneously, the parameters should be chosen so that there is consistency across centres and adjustments are made for multiple testing [11].
- d) Data sourcing – data may already be routinely available from existing data collection systems or there may be a need to collect new or additional data.
- e) Risk-adjustment – including details of how this has been developed and what factors will be adjusted for. While adjustment for case-mix is crucial in accounting for factors that predispose patients to adverse outcomes, it may mask unsuitable centre practice if overdone, or encourage clinicians to avoid high-risk patients if done inadequately. The limitations of the risk-adjustment should be made clear and any risk-adjustment tools used should be updated regularly, as practices change and outcomes tend to improve over time.
- f) Practical implementation – how the procedure is to be implemented, including the frequency with which results are produced and settings are updated, the procedure to follow when signals occur.
- g) Limitations of the procedure – including the potential for clinicians avoiding high-risk patients. Monitoring procedures such as these should be seen as ‘screening’ tools that form a first-pass analysis that can then be used to direct further investigation and to target clinical audit [11].
- h) The procedure to be adopted if a signal is indicated in the monitoring procedure, indicating that a centre is not performing as well as expected. This includes a detailed description of who should be informed at what stage, and any action that would be expected.

#### **6.4 Discussion**

The tools for monitoring outcomes described here can be used at individual centres for internal monitoring or by a central agency to monitor several centres simultaneously. When used by a central agency it is likely that the monitoring will be done retrospectively as data may only be collected at set intervals. The advantage of the later approach is that there will be consistency across centres with regards to the procedure that is followed.

It is important to adjust for factors that influence outcome. However the risk-adjustment employed will be limited by the data available. Care must also be taken not to adjust for factors that are under centres’ control, otherwise the risk-adjustment may mask unsuitable practice.

The impact of ‘false positive’ signals should be assessed. Too many false alarms may lead to doubt about the credibility of the procedure. In addition, resources would be dedicated to unnecessary investigation of ‘poor performance’ when outcome rate are actually as good as they can be. Care must be taken that in an attempt to reduce the number of false positives, the risk of missing genuine changes is not significantly increased.

The limitations of any monitoring procedure applied should be clearly stated and its reliability explained in an easily interpretable manner to avoid misinterpretation of the results.

### Recommended reading

- Hawkins DM, Olwell DH. *Cumulative Sum Charts and Charting for Quality Improvement*. Springer, New York, 1998.
- Schold JD, Howard RJ. Prediction models assessing transplant centre performance: Can a little knowledge be a dangerous thing? *American Journal of Transplantation* 2006; 6: 245 – 246.
- Spiegelhalter, D. Monitoring clinical performance: A commentary. *Journal of Thoracic and Cardiovascular Surgery* 2004; 128: 820–825.

### References

1. Scally G, Donaldson LJ. Looking forward: Clinical governance and the drive for quality improvement in the new NHS in England. *British Medical Journal* 1998; 317: 61 – 65.
2. Copeland G. A practical handbook for clinical audit: Guidance published by the NHS Clinical Governance Support Team, 2005. Available from: <http://www.cgsupport.nhs.uk/Resources/default.asp> (Accessed 17 May 2006)
3. Department of Health. *Learning from Bristol: the Report of the Public Inquiry into Children’s Heart Surgery at the Bristol Royal Infirmary 1984-1995*. Command paper CM 5207. London: The Stationery Office, 2001.
4. Beiles CB, Morton AP. Cumulative sum control charts for assessing performance in arterial surgery. *ANZ Journal of Surgery* 2004; 74: 146 – 151.
5. Rogers CA, Ganesh JS, Banner NR, Bonser RS. Cumulative risk adjusted monitoring of 30-day mortality after cardiothoracic transplantation: UK experience. *European Journal of Cardiothoracic Surgery* 2005; 27:1022 – 1029.
6. Axelrod DA, Guidinger MK, Metzger RA, Wiesner RH, Webb RL, Merion RM. Transplant centre quality assessment using a continuously updatable, risk-adjusted technique (CUSUM). *American Journal of Transplantation* 2006; 6: 313 – 323.
7. Schold JD, Howard RJ. Prediction models assessing transplant centre performance: Can a little knowledge be a dangerous thing? *American Journal of Transplantation* 2006; 6: 245 – 246.
8. Spiegelhalter D, Grigg O, Kinsman R, Treasure T. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. *International Journal for Quality in Healthcare* 2003; 15: 7 – 13.
9. Grigg OA, Farewell VT, Spiegelhalter DJ. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Statistical Methods in Medical Research* 2003; 12: 147 – 170.
10. Grigg O, Farewell V. An overview of risk-adjusted charts. *Journal of the Royal Statistical Society A* 2004; 167: 523–539.
11. Marshall C, Best, N, Bottle A, Aylin P. Statistical issues in the prospective monitoring of health outcomes across multiple units. *Journal of the Royal Statistical Society A* 2004; 167: 541–559.

12. Woodall WH. The use of control charts in health-care and public health surveillance. *Journal of Quality Technology* 2006; 38: 88 – 102.
13. Steiner SH, Cook R J, Farewell VT, Treasure T. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 2000; 1: 441–452.
14. Keogh BE, Kinsman R. The Society of Cardiothoracic Surgeons of Great Britain and Ireland National Adult Cardiac Surgical Database Report, 2000 – 2001, Dendrite Clinical Systems Ltd, Henley-on-Thames (2002) p 200 -223.
15. Hawkins DM, Olwell DH. *Cumulative Sum Charts and Charting for Quality Improvement*. Springer, New York, 1998.
16. Chen R. A surveillance system for congenital malformations. *Journal of the American Statistical Association* 1978; 73: 323–327.
17. Grigg O, Farewell V. A risk-adjusted sets method for monitoring adverse medical outcomes. *Statistics in Medicine* 2004; 23: 1593–1602.
18. Ismail NA, Pettitt AN, Webster RA. ‘On-line’ monitoring and retrospective analysis of hospital outcomes based on a scan statistic. *Statistics in Medicine* 2003; 22: 2861 – 2876.
19. Rogers CA, Alfred T, Banner NR, Bonser RS, van der Meulen J. UK Cardiothoracic Transplant Audit Report 2005.
20. Bird SM, Cox D, Farewell VT, Goldstein H, Holt T, Smith PC. Performance indicators: Good, bad and ugly. *Journal of the Royal Statistical Society A* 2005; 168: 1 – 27. (Report of the Royal Statistical Working Party on Performance Monitoring in Public Services. Available from: <http://www.rss.org.uk/PDF/PerformanceMonitoring.pdf>).

## 7 MISSING DATA TECHNIQUES

Missing data are a common issue and part of almost all research, especially when analyzing registry data (large retrospective data sets). Missing or incomplete data have to be treated with care in any model otherwise they can seriously affect the validity of your results. There are a number of alternative ways of dealing with missing data, and this chapter is an attempt to outline the most commonly used approaches.

### 7.1 Why worry about missing data?

Missing data are simply observations that we intended to be made but did not. If you ignore missing data or assume that excluding missing data is sufficient, you risk reaching invalid results by affecting parameter estimates and statistical tests of significance. An additional complication is that the more data is missing, the more likely it is that you will need to address the problem of incomplete cases. Those situations are those where imputing or filling in values for the missing data is questionable due to the reduced proportion of valid data available.

Statisticians have to deal both with Unit non-response (complete non-response for a particular observation) and item non-response (partial data are available for subjects). We will focus on item non-response, the most frequent type of missing data in registries.

Before making the decision on what method to use, preliminary data analysis should be undertaken to determine :

- a. the percentage of missing data on each variable of interest
- b. distribution of the missing data : is there much data missing from a few subjects or a little data missing from each or several subjects?
- c. the number of variables that are concerned
- d. the missing data pattern (e.g. a particular time period, centre or type of recipient)

This analysis would be expected to provide information about the missing data mechanism.

### 7.2 When to exclude variables

Variables of interest should be described using simple statistical methods (Ie, Table 2.1, Chapter2) in order to determine the shape of the distribution as well as the percentage of missing data.

There is no simple decision rule for whether to leave data as they are, to drop cases with missing values, or to impute values to replace missing values. In fact, there is no consensus concerning the proportion of missing data from which the variable has to be excluded from the analysis. The final choice of excluding a variable presenting a substantial proportion of missing data also depends on the importance of its effect on the outcome (clinical relevance, statistical significance proven in previous studies).

### 7.3 The missingness mechanism

It is important to determine the missing data mechanism, or in other words if missingness is related to study variables or not, in order to decide what method can be used to deal with the missing data problem.

There are 3 categories of mechanisms which we will briefly describe below.

#### 7.3.1 *Missing completely at random*

Exists when missing values are randomly distributed across all observations. Cases with complete data are indistinguishable from cases with incomplete data. The probability that the observation is missing does not depend on the missing data, or on the observed data, or any other data.

Missing data are considered as ignorably missing. Complete case analysis can be undertaken as estimations would not be biased. It is the least problematic type of missing data, but also the least realistic situation.

#### 7.3.2 *Missing at random*

Exists when missing values are not randomly distributed across all observations but are randomly distributed within one or more sub-samples. In fact, cases with incomplete data differ from cases with complete data, but the pattern of data missingness is traceable or predictable from other variables in the database rather than being due to the specific variable on which the data are missing. In this case, missing data can be considered MCAR within strata. The probability that the observation is missing does not depend on the missing data or unobserved data, but might depend on values of the observed data.

In practice it is usually difficult to meet the MCAR assumption. MAR is an assumption that is more often, but not always tenable. The more relevant and related predictors one can include in statistical models, the more likely it is that the MAR assumption will be met.

#### 7.3.3 *Missing not at random (MNAR)*

The pattern of data missingness is non-random and it is not predictable from other variables in the database. The probability that the observation is missing depends on the missing data or unobserved data.

It is necessary to model both the missingness mechanism and the observed values. It is the most problematic type of missing data and it is very difficult to identify the appropriate model for the missingness mechanism. Missing data are considered as non-ignorable.

### 7.4 How to deal with missing data: the most common methods

Once you have determined the missingness mechanism, it is important to keep in mind when choosing a missing data handling approach, to maintain the shape of the original distribution of responses.

#### 7.4.1 *Complete case analysis or listwise deletion*

Consists in deleting the entire record presenting a missing value on a covariate. This is the method by default in most statistical software. It often results in a substantial decrease in the sample size

(loss of statistical power) but it does have important advantages. In fact, if the assumption that data are MCAR or MAR is met and if the percentage of missing data is small (and concentrated on a small number of subjects), the resulting dataset should be a representative subsample of the full dataset and lead to unbiased parameter estimates.

The following methods all use the complete data set as they replace the missing values by another value or set of values. Imputation needs to be considered carefully since a real value is inserted to replace a missing value. It is therefore important to choose the most appropriate value as using an inappropriate measure can increase error and distort findings.

#### *7.4.2 Missing category*

When a variable is categorical, it is of common use to create a missing category including all missing values. This technique might pool very different individuals. The impact of this technique depends on the missingness mechanism and produces biased estimates.

#### *7.4.3 Mean / Median Imputation*

Consists in replacing the missing value by group's mean or median value. It has been one of the most common methods of imputation but is no longer. Not only does it assume that data are MCAR but it also tends to reduce the variance and covariance of the variable. Mean substitution in the case of one variable can lead to bias in estimates of the effects of other or all variables in the regression analysis, because bias in one correlation can affect the beta weights of all variables. Somewhat better is substitution of the group mean for a categorical variable known to correlate highly with the variable which has missing values (conditional mean).

#### *7.4.4 Regression imputation*

Consists in developing a regression equation on complete case data for each variable with missing data, based on all the other variables as predictors and then substitute the predicted mean for each unit with a missing value. The regression method has the problem that all cases with the same values on the independent variables will be imputed with the same value on the missing variable, causing a portion of the same problems as mean imputation. Variance understatement can be reduced by adding a random residual to an expected conditional mean before the value is imputed.

It is important to indicate that these procedures only produce unbiased estimates of parameters under MCAR condition. Moreover, the standard errors associated with imputed values are always underestimated (all missing data are replace by the same value) and the associated test statistics overestimated. These statistical problems become more severe as the percentage of missing data increases.

### **7.5 More sophisticated methods**

#### *7.5.1 Hot Deck imputation*

Hot-deck imputation fills in missing values on incomplete records using values from complete records for "similar" people in the same dataset. The difficulty of this method is to define similarities. Variance understatement can be reduced by choosing a random respondent from among the similar respondents to donate values for the imputation

### 7.5.2 EM Algorithm

EM is a class of iterative algorithms for Maximum Likelihood Estimation (MLE). On each iteration of the algorithm, there are two steps: the expectation (E-) step and the maximization (M-) step. In the **E-step**, the expected values of missing data are calculated from the observed data so as to maximize the total log-likelihood. In the **M-step**, the expected values of the missing data computed in the E-step are used to re-estimate the parameters and to update the total log-likelihood. The steps are iterated until the difference between current and subsequent estimates is small.

MLE makes fewer demands of the data in terms of statistical assumptions and takes advantage of relationships in the data. It generates reliable parameter estimates under MAR condition but standard errors are underestimated. It can also be computationally intensive if data sets are large.

### 7.5.3 Multiple Imputation

Missing values are replaced by a set of plausible values that represent the uncertainty about the right value to impute. The different steps are described below :

- 1- Missing data are filled in  $m$  times to generate  $m$  complete data sets.
- 2- The  $m$  complete data sets are analyzed by using standard procedures.
- 3- The results from the  $m$  complete data sets are combined for the inference.

The results of the analyses in step 2 are pooled in step 3 to provide point and variance estimates for the effects of interest. The variance is not biased any longer if MCAR, MAR and has the advantage of simplicity over MLE, making it particularly suitable for large data sets.

Computation process is complicated and intensive although specialized programs are available.

## 7.6 Statistical software

Most procedures in statistical software exclude observations with any missing variable values from the analysis (Complete Case Analysis). Single imputation technics can all easily programmed under any software. Concerning multiple imputation, many solutions have been developed. A non exhaustive list is given below:

- **SAS Proc MI and PROC MIANALYZE.**

- **MICE** by Stef van Buuren and Karin Oudshoorn contains *S-PLUS* software for flexible generation of multivariate imputations.

- **WinMICE** by Gert Jacobusse is a stand-alone program under Windows that implements imputation on the linear mixed model. Needs Windows.

- **IVEWARE** by Raghunathan, Solenberger and John Van Hoewyk is a *SAS-based* application for creating multiple imputations.

- **Missing Data Library in S-Plus 6.** It features Gaussian, Loglinear and Conditional Gaussian. Performing multiple complete data analysis after multiple imputation, and consolidating results, is simplified by using the library.

- **SOLAS** for Missing Data Analysis is a commercial Windows program by Statistical Solutions Limited.

## 7.7 Discussion

The best situation is not have any missing data, but it is in general far from being realistic. There is no “magical recipe “ and there is therefore no unique solution to missing data problems. In each study, missing data have to be explored before choosing what method to use, or to exclude the concerned variables from the analysis.

Part only of all possible techniques to treat missing data have been presented in this chapter. In fact, theory and practice for treating missing values is far beyond the “ad hoc” stage, and development continues.

## References

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage Publications.

Little, R. J. A. and D. B. Rubin (1987). *Statistical analysis with missing data*. John Wiley & Sons, New York.

Pickles, Andrew (2005). Missing data, problems and solutions. Pp. 689-694 in Kimberly Kempf-Leonard, ed., *Encyclopedia of social measurement*. Amsterdam: Elsevier.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, New York.

Schafer, J.L. (1997) *Analysis of incomplete multivariate data*. Chapman & Hall, London. Book No. 72, Chapman & Hall series Monographs on Statistics and Applied Probability.

Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*. 8: 3-15.

Schafer, J.L. and M. K. Olsen (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*. 33: 545-571.

Truxillo, C. (2005). Maximum Likelihood Parameter Estimation with Incomplete Data, SAS Users Group International Conference, Philadelphia PA, April 10-13.

Software Documentation : SAS®, S-PLUS® and SPSS®.

<http://www.lshtm.ac.uk/msu/missingdata/index.html>

<http://www.multiple-imputation.com/>

## 8 ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS

The aim of this chapter is to show how in the framework of a general progress of scientific research it is possible to consider the techniques commonly referred to as Artificial Intelligence methods. This approach provides a possible alternative to the methods described in Chapters 2-6. We have devoted one chapter to these techniques because they have different and significant features compared to the techniques used for statistical analysis already described. Also, since they are relatively new methodologies it is considered also useful to give a brief historical introduction to them with reference to the "Philosophy of Science". The main part of this chapter will be a review of the most important articles about the use of this method in survival analysis and comparing outcomes between centres.

### 8.1 Origins of the methods

Artificial Intelligence (AI), also some times called Synthetic Intelligence, is defined as intelligence exhibited by an artificial (non-natural, man-made) entity and forms a vital branch of computer science, dealing with intelligent behaviour in machines. When was for the first time the term Artificial Intelligence used? John McCarthy was responsible for the coining of the term "Artificial Intelligence" in his 1955 proposal for 1956 Dartmouth Conference. He has been a prominent computer scientist who received the Turing Award in 1971 for his major contributions to the field of mathematical logic.

The first ideas related to this concept can be found early in the 17th century when René Descartes envisioned the bodies of animals as complex but reducible machines, thus formulating the mechanistic theory, also known as the "clockwork paradigm". Wilhelm Schickard created the first mechanical digital calculating machine in 1623, followed by machines of Blaise Pascal (1643) and Gottfried Wilhelm von Leibniz (1671), who also invented the binary system. Bertrand Russell and Alfred North Whitehead published *Principia Mathematica* in 1910-1913, which revolutionized formal logic. In 1941 Konrad Zuse built the first working program-controlled computers. Warren McCulloch and Walter Pitts published "A Logical Calculus of the Ideas Immanent in Nervous Activity" in 1943, laying the foundations for neural networks. Norbert Wiener's *Cybernetics or Control and Communication in the Animal and the Machine* (MIT Press, 1948) popularises the term "cybernetics".

The 1950s were a period of active efforts in AI. In 1950, Alan Turing introduced the "Turing test" as a way of implementing a test of intelligent behaviour. The first working AI programs were written in 1951 to run on the Ferranti Mark I machine of the University of Manchester: a draughts-playing program written by Christopher Strachey and a chess-playing program written by Dietrich Prinz. At the same time, John von Neumann, who had been hired by the RAND Corporation, developed the game theory, which would prove invaluable in the progress of AI research.

During the 1960s and 1970s, Joel Moses demonstrated the power of symbolic reasoning for integration problems in the Maccsma program, the first successful knowledge-based program in mathematics. Leonard Uhr and Charles Vossler published "A Pattern Recognition Program That Generates, Evaluates, and Adjusts Its Own Operators" in 1963, which described one of the first

machine learning programs that could adaptively acquire and modify features. Ted Shortliffe demonstrated the power of rule-based systems for knowledge representation and inference in medical diagnosis and therapy in what is sometimes called the first expert system.

A key development took place in the 1980s, when neural networks became widely used due to the back propagation algorithm, first described by Paul Werbos in 1974. Today, the AI debate in the philosophy of science ('can a man-made artefact be conscious?') is still a hot topic; among the others most notably Roger Penrose in his book *The Emperor's New Mind* and John Searle with his "Chinese room" thought experiment argue that true consciousness cannot be achieved by formal logic systems. Douglas Hofstadter in *Gödel, Escher, Bach* and Daniel Dennett in *Consciousness Explained* argue in favour of functionalism. In many AI supporters' opinion, artificial consciousness is considered as the holy grail of artificial intelligence.

## 8.2 Theory

AI divides roughly into two schools of thought: *Conventional AI* (also called symbolic AI or logical AI) and Computational Intelligence *CI*.

Conventional AI mostly involves methods now classified as machine learning, characterized by formalism and statistical analysis. Methods include:

*Expert systems*: apply reasoning capabilities to reach a conclusion. An expert system can process large amounts of known information and provide conclusions based on them.

*Case based reasoning*.

*Bayesian networks*.

*Behaviour based AI*: a modular method of building AI systems by hand.

Computational Intelligence involves iterative development or learning (e.g. parameter tuning as in connectionist systems). Learning is based on empirical data and is associated with non-symbolic AI, and methods mainly include:

*Neural networks*: systems with very strong pattern recognition capabilities.

*Fuzzy systems*: techniques for reasoning under uncertainty have been widely used in modern industrial and consumer product control systems.

*Evolutionary computation*: applies biologically inspired concepts such as populations, mutation and survival of the fittest to generate increasingly better solutions to the problem. These methods most notably divide into evolutionary algorithms (e.g. genetic algorithms) and swarm intelligence (e.g. ant algorithms).

With hybrid intelligent systems attempts are made to combine these two groups.

## 8.3 AI in medicine

Performing a term-related search in the well known search engine for scientific publications PUBMED (<http://www.ncbi.nlm.nih.gov/entrez>), we find (1st June, 2006) for the terms *Artificial + Intelligenc + Medicine* 1860 citations (but in this case we have to consider that it includes robotics

medical techniques) and in the more theme-specific terms *Artificial + Intelligence + Survival + Analysis*, 103 citations.

It also seems that very early on, scientists and doctors alike were captivated by the potential such a technology might have in medicine (e.g. Ledley and Lusted, 1959). With intelligent computers able to store and process vast stores of knowledge, the hope was that they would become perfect ‘doctors in a box’, assisting or surpassing clinicians with tasks like diagnosis.

AI in medicine at that time was a largely US-based research community. Work originated out of a number of campuses, including MIT-Tufts, Pittsburgh, Stanford and Rutgers (e.g. Szolovits, 1982; Clancey and Shortliffe, 1984; Miller, 1988). The field attracted many of the best computer scientists and by any measure their output in the first decade of the field remains a remarkable achievement. In 1984, Clancey and Shortliffe defined Artificial Intelligence in Medicine:

*Medical artificial intelligence is primarily concerned with the construction of Artificial Intelligence programs that perform diagnosis and make therapy recommendations. Unlike medical applications based on other programming methods, such as purely statistical and probabilistic methods, medical Artificial Intelligence programs are based on symbolic models of disease entities and their relationship to patient factors and clinical manifestations.*

The study of artificial intelligence in medicine (AIM) is nearly 30 years old. The recent work in AIM has addressed problems common to the fields of medicine and artificial intelligence. With the emphasis in medicine shifting to more evidence-based practice, the field AIM is positioned to help provide solutions for the evolving field of medicine.

According to Coiera (2003), other important applications are as follows:

*Alerts and reminders.* In real-time situations, an expert system attached to a patient monitoring device like an ECG or pulse oximeter can warn of changes in a patient’s condition. In less acute circumstances, it might scan laboratory test results, drug or test order, or the EMR and then send reminders or warnings, either via immediate on-screen feedback or through a messaging system like e-mail. Reminder systems are used to notify clinicians of important tasks that need to be done before an event occurs. For example, an outpatient clinic reminder system may generate a list of immunizations that each patient on the daily schedule requires.

*Diagnostic assistance.* When a patient’s case is complex, rare or the person making the diagnosis is simply inexperienced, an expert system can help in the formulation of likely diagnoses based on patient data presented to it, and the systems understanding of illness, stored in its knowledge base. Diagnostic assistance is often needed with complex data, such as the ECG, where most clinicians can make straightforward diagnoses, but may miss rare presentations of common illnesses like myocardial infarction, or may struggle with formulating diagnoses, which typically require specialised expertise.

*Therapy critiquing and planning.* Critiquing systems can look for inconsistencies, errors and omissions in an existing treatment plan, but do not assist in the generation of the plan. Critiquing systems can be applied to physician order entry. For example, on entering an order for a blood transfusion a clinician may receive a message stating that the patient’s haemoglobin level is above the transfusion threshold, and the clinician must justify the order by stating an indication, such as active bleeding

(Randolph et al., 1999). Planning systems on the other hand have more knowledge about the structure of treatment protocols and can be used to formulate a treatment based upon a data on patient's specific condition from the EMR and accepted treatment guidelines.

*Prescribing decision support systems (PDSS).* One of the commonest clinical tasks is the prescription of medications, and PDSS can assist by checking for drug-drug interactions, dosage errors, and if connected to an EMR, for other prescribing contraindications such as allergy. PDSS are usually well received because they support a pre-existing routine task, and as well as improving the quality of the clinical decision, usually offer other benefits like automated script generation and sometimes electronic transmission of the script to a pharmacy.

*Information retrieval.* Finding evidence in support of clinical cases is still difficult on the Web, and intelligent information retrieval systems can assist in formulating appropriately specific and accurate clinical questions, they can act as information filters, by reducing the number of documents found in response to a query to a Web search engine, and they can assist in identifying the most appropriate sources of evidence appropriate to a clinical question. More complex software 'agents' can be sent to search for and retrieve information to answer clinical questions, for example on the Internet. The agent may contain knowledge about its user's preferences and needs, and may also have some clinical knowledge to assist it in assessing the importance and utility of what it finds.

*Image recognition and interpretation.* Many clinical images can now be automatically interpreted, from plane X-rays through to more complex images like angiograms, CT and MRI scans. This is of value in mass-screenings, for example, when the system can flag potentially abnormal images for detailed human attention.

#### **8.4 AI in survival analysis**

Before a review of the most important scientific publications in this field it is important to point out which are the most important features of the application of AI methods in the study of risk factors of transplanted patients and more generally of survival analysis:

*Conditions in which the AI approach is possible*

Difficult Analytical approach.

Complex or Chaotic System.

*Qualities of the methods*

Identify arbitrary nonlinear relations between the independent and dependent variables.

All possible interactions between the dependent variables. Standard statistical approaches (e.g., logistic or Cox regression) require additional modeling to allow this flexibility.

Do not require explicit distributional assumptions, such as normality.

## 8.5 A review of some of some specific papers

In this section, we review some of the papers that have used AI methods in survival analysis. So as to provide a detailed summary of the work, abstracts of the papers have been reproduced.

One of the earliest articles in this area was that of Faraggi and Simon (1995).

1. *Statistics in Medicine*. 1995 Jan 15;14(1):73-82. **A neural network model for survival data.** Faraggi D, Simon R. Biometric Research Branch, National Cancer Institute, Rockville, MD 20852.

*Abstract: Neural networks have received considerable attention recently, mostly by non-statisticians. They are considered by many to be very promising tools for classification and prediction. In this paper we present an approach to modelling censored survival data using the input-output relationship associated with a simple feed-forward neural network as the basis for a non-linear proportional hazards model. This approach can be extended to other models used with censored survival data. The proportional hazards neural network parameters are estimated using the method of maximum likelihood. These maximum likelihood based models can be compared, using readily available techniques such as the likelihood ratio test and the Akaike criterion. The neural network models are illustrated using data on the survival of men with prostatic carcinoma. A method of interpreting the neural network predictions based on the factorial contrasts is presented*

In the article of Ripley et al. (2004), the authors discuss the possibility of using AI methods to allow non-linear predictors to be fitted implicitly and the effect of the covariates to vary over time.

2. *Statistics in Medicine*. 2004 Mar 15;23(5):825-42. **Non-linear survival analysis using neural networks.** Ripley RM, Harris AL, Tarassenko L. Department of Statistics, University of Oxford, Oxford OX1 3TG, UK.

*Abstract: We describe models for survival analysis which are based on a multi-layer perceptron, a type of neural network. These relax the assumptions of the traditional regression models, while including them as particular cases. They allow non-linear predictors to be fitted implicitly and the effect of the covariates to vary over time. The flexibility is included in the model only when it is beneficial, as judged by cross-validation. Such models can be used to guide a search for extra regressors, by comparing their predictive accuracy with that of linear models. Most also allow the estimation of the hazard function, of which a great variety can be modelled. In this paper we describe seven different neural network survival models and illustrate their use by comparing their performance in predicting the time to relapse for breast cancer patients.*

In the paper of Fuller et al. (2005), it has been found that using Artificial Neural Networks (ANN) it is possible to predict in a more precise way the probability of survival of trauma victims.

3. *The West Virginia Medical Journal*. 2005 May-Jun;101(3):120-5. **A comparison of neural networks for computing predicted probability of survival for trauma victims.** Fuller JJ,

Emmett M, Kessel JW, Price PD, Forsythe JH. Marshall University, Joan C. Edwards School of Medicine, Huntington, USA

*Abstract: TRISS is a statistical method for predicting the probability of survival of trauma victims. Analysis of data from the Trauma Registry at Charleston Area Medical Center showed that only 48% of the trauma fatalities in the 5-year period 1992-1996 were correctly predicted by TRISS. Trauma practitioners from other Trauma Centers report similar problems with TRISS. Researchers have suggested improvements that range from simply changing the input variables and/or regression coefficients in TRISS to using an entirely different model. In this study we describe a method of calculating survival probabilities using Artificial Neural Networks (ANN). This method was chosen because of the similarity of the ANN output function to the function that produces the TRISS probability of survival. Additional variables were added based on the results of other research efforts as well as analysis of the CAMC Trauma Registry. A comparison was made between the abilities of TRISS to predict fatalities and to approximate probability of survival. The ANN outperformed TRISS in predicting fatalities in a training set (68.1% correct vs. 47.9% correct) and in a testing set (61.3% correct vs. 51.3% correct). More importantly, the ANN produced better estimates of predicted deaths. Using a data set that included 119 deaths, the ANN model predicted 125 deaths for a 5% relative error. The predicted number using TRISS was 86 for a relative error of 27.7%. Since effective quality improvement for trauma care depends on accurately identifying cases that fall outside the expected results, a more accurate predictive tool allows a more focused review of those significant cases, thus conserving resources without compromising quality. Neural Networks appear to be a predictive tool that can provide probability of survival estimates that are more accurate than TRISS.*

In the article of Dekker et al. (2002), it is shown that the fuzzy method yields the highest predictive accuracy for both nodal involvement and survival analyses.

4. *Anticancer Research*. 2002 Jan-Feb;22(1A):433-8. **Assessment of nodal involvement and survival analysis in breast cancer patients using image cytometric data: statistical, neural network and fuzzy approaches.** Seker H, Odetayo MO, Petrovic D, Naguib RN, Bartoli C, Alasio L, Lakshmi MS, Sherbet GV. BIOCORE, School of Mathematical and Information Sciences, Coventry University, UK.

*Abstract: Accurate and reliable decision making in breast cancer prognosis can help in the planning of suitable surgery and therapy and, generally, optimise patient management through the different stages of the disease. In recent years, several prognostic factors have been used as indicators of disease progression in breast cancer. In this paper we investigate a fuzzy method, namely fuzzy k-nearest neighbour technique for breast cancer prognosis, and for determining the significance of prognostic markers and subsets of the markers, which include histology type, tumour grade, DNA ploidy, S-phase fraction, G0G1/G2M ratio, and minimum (start) and maximum (end) nuclear pleomorphism indices. We also compare the method with (a) logistic regression as a statistical method, and (b) multilayer feed forward backpropagation neural networks as an artificial neural network tool, the latter two techniques having been widely used for cancer prognosis. Nodal involvement and survival analyses in breast cancer are carried out for 100 women who were clinically diagnosed with breast disease in the form of carcinoma and benign conditions, and seven prognostic markers collected for each patient. For nodal involvement analysis, node positive and negative patients are predicted whereas survival analysis is carried out for two categories: whether a patient is alive or dead within 5 years of diagnosis. The results obtained show that the fuzzy method yields the highest predictive accuracy of 88% for both nodal involvement and survival analyses*

*obtained from the subsets of [tumour grade, S-phase fraction, minimum (start) nuclear pleomorphism index] and [tumour histology type, DNA ploidy, S-phase fraction, G0G1/G2M ratio], respectively. We believe that this technique has produced more reliable prognostic factor models than those obtained using either the statistical or artificial neural networks-based methods.*

The work of Snow et al. (2001), indicates the adoption of AI methods by an important national organization, the American Commission on Cancer, and the fact that AI methods yield a greater sensitivity to mortality compared to logistic regression.

5. *Cancer*. 2001 Apr 15;91(8 Suppl):1673-8. **Neural network and regression predictions of 5-year survival after colon carcinoma treatment.** Snow PB, Kerr DJ, Brandt JM, Rodvold DM. Xaim, Inc., Colorado Springs, Colorado, USA. psnow@xaim.com

*Abstract: The Commission on Cancer data from the National Cancer Data Base (NCDB) for patients with colon carcinoma was used to develop several artificial neural network and regression-based models. These models were designed to predict the likelihood of 5-year survival after primary treatment for colon carcinoma.*

*METHODS: Two modeling methods were used in the study. Artificial neural networks were used to select the more important variables from the NCDB database and model 5-year survival. A standard parametric logistic regression also was used to model survival and the two methods compared on a prospective set of patients not used in model development. RESULTS: The neural network yielded a receiver operating characteristic (ROC) area of 87.6%. At a sensitivity to mortality of 95% the specificity was 41%. The logistic regression yielded a ROC area of 82% and at a sensitivity to mortality of 95% gave a specificity of 27%. CONCLUSIONS: The neural network found a strong pattern in the database predictive of 5-year survival status. The logistic regression produced somewhat less accurate, but good, results*

In the article of Suka et al. (2004), it has been found that the highest predictive accuracy in estimating prognosis was found in an AI technique.

6. *Statistics in Medicine*. 2004; 11(Pt 1):741-5. **Comparison of proportional hazard model and neural network models in a real data set of intensive care unit patients.** Suka M, Oeda S, Ichimura T, Yoshida K, Takezawa J. Department of Preventive Medicine, St. Marianna University School of Medicine, 2-16-1 Sugao, Miyamae-ku, Kawasaki, Kanagawa 216-8511, Japan

*Abstract: There has been increased interest in using neural network model (NNM) for prognosis tasks. However, the performance of NNM has seldom been compared with that of traditional statistical models such as proportional hazard model (PHM) in real data sets. We conducted a comparative study of PHM and two types of NNM, that is, aggregate single point model (ASPM) and multiple point model (MPM), using a real data set of intensive care unit patients. The three models were developed using the 70% training subset and their predictive accuracy were assessed using the 30% testing subset according to classification accuracy, area under receiver operating curve, and concordance index. Overall, the highest predictive accuracy was found in MPM,*

*followed by PHM and ASPM. MPM is likely to have the potential ability to provide more accurate estimation of prognosis than PHM and ASPM*

In the article of Bakker et al. (2004), it is shown that traditional Cox survival analysis can be improved within a neural Bayesian framework to give more reliable predictions, in particular for relatively small data sets.

7. *Statistics in Medicine*. 2004 Oct 15;23(19):2989-3012. **Improving Cox survival analysis with a neural-Bayesian approach.** Bakker B, Heskes T, Neijt J, Kappen B. Theoretical Foundation SNN Laboratory, University of Nijmegen, Geert Grooteplein 21, 6525 EZ Nijmegen.

*Abstract: In this article we show that traditional Cox survival analysis can be improved upon when supplemented with sensible priors and analysed within a neural Bayesian framework. We demonstrate that the Bayesian method gives more reliable predictions, in particular for relatively small data sets. The obtained posterior (the probability distribution of network parameters given the data) which in itself is intractable, can be made accessible by several approximations. We review approximations by Hybrid Markov Chain Monte Carlo sampling, a variational method and the Laplace approximation. We argue that although each Bayesian approach circumvents the shortcomings of the original Cox analysis, and therefore yields better predictive results, in practice the use of variational methods or Laplace is preferable. Since Cox survival analysis is infamous for its poor results with (too) many inputs, we use the Bayesian posterior to estimate p-values on the inputs and to formulate an algorithm for backward elimination. We show that after removal of irrelevant inputs Bayesian methods still achieve significantly better results than classical Cox.*

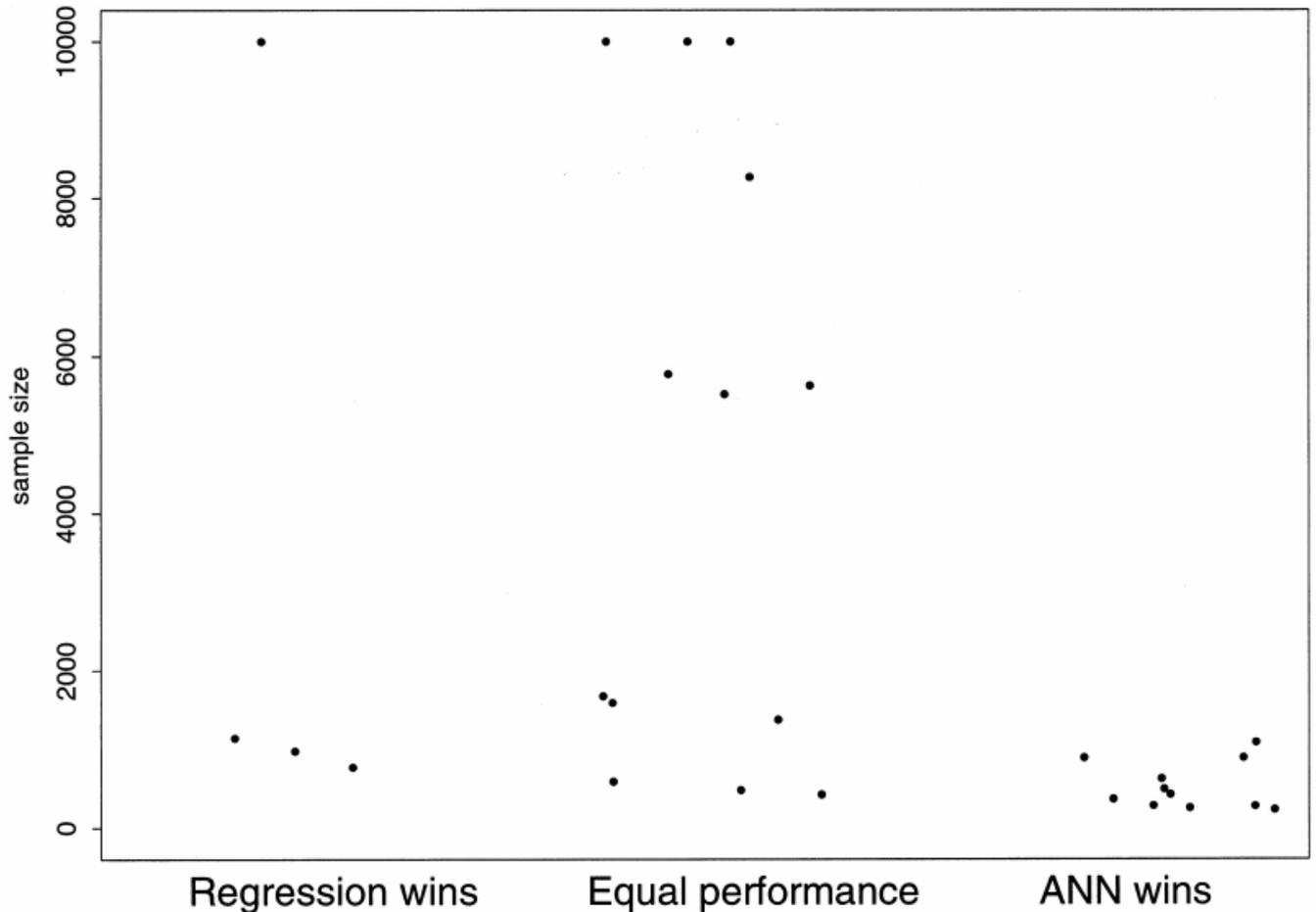
The interesting work of Sargent is meta-analysis like paper comparing AI methods with standard statistical methods, such as logistic or Cox regression.

8. *Cancer*, 91; 1636-164. **Comparison of artificial neural networks with other statistical approaches Results from medical data set.** Daniel J. Sargent, Ph.D. Cancer Center Statistics, Mayo Clinic, Rochester, Minnesota

*Abstract: In recent years, considerable attention has been given to the development of sophisticated techniques for exploring data sets. One such class of techniques is artificial neural networks (ANNs). Artificial neural networks have many attractive theoretic properties, specifically, the ability to detect non predefined relations such as nonlinear effects and/or interactions. These theoretic advantages come at the cost of reduced interpretability of the model output. Many authors have analyzed the same data set, based on these factors, with both standard statistical methods (such as logistic or Cox regression) and ANN. METHODS: The goal of this work is to review the literature comparing the performance of ANN with standard statistical techniques when applied to medium to large data sets (sample size > 200 patients). A thorough literature search was performed, with specific criteria for a published comparison to be included in this review. RESULTS: In the 28 studies included in this review, ANN outperformed regression in 10 cases (36%), was outperformed by regression in 4 cases (14%), and the 2 methods had similar performance in the remaining 14 cases (50%). However, in the 8 largest studies (sample size > 5000), regression and ANN tied in 7 cases, with regression winning in the remaining*

case. In addition, there is some suggestion of publication bias. CONCLUSIONS: Neither method achieves the desired performance. Both methods should continue to be used and explored in a complementary manner. However, based on the available data, ANN should not replace standard statistical approaches as the method of choice for the classification of medical data.

Figure 8.1 Comparison of AI methods with traditional methods



In this article by Haydon et al. (2005), an AI approach is used to match donor livers and recipients so it is possible to inform transplant clinicians about the optimum use of donor livers and thereby effectively make the best use of a scarce resource.

9. *Transplantation*. 2005 Jan 27; 79(2): 213-8. **Self-organizing maps can determine outcome and match recipients and donors at orthotopic liver transplantation.** Haydon GH, Hiltunen Y, Lucey MR, Collett D, Gunson B, Murphy N, Nightingale PG, Neuberger J. Liver Unit, Third Floor, Nuffield House, The Queen Elizabeth Hospital, Birmingham, UK

*Abstract: There is a relative lack of donor organs for liver transplantation. Ideally, to maximize the utility of those livers that are offered, donor and recipient characteristics should be matched to ensure the best possible posttransplant survival of the recipient. METHODS: With prospectively collected data on 827 patients receiving a primary liver graft for chronic liver disease, we used a self-organizing map (SOM) (one form of a neural network) to predict outcome after transplantation using both donor and recipient factors. The SOM was then validated using a data set of 2622 patients undergoing transplantation in the United Kingdom at other centers. RESULTS: SOM analysis using 72 inputs and two survival intervals (3 and 12 months) yielded three neurons with either higher or lower probabilities of survival. The model was validated using the independent data set. With 20 patients on the waiting list and 10 sequential donor livers, it was possible to demonstrate that the model could be used to identify which potential recipients were likely to benefit most from each liver offered. CONCLUSIONS: With this approach to matching donor livers and recipients, it is possible to inform transplant clinicians about the optimum use of donor livers and thereby effectively make the best use of a scarce resource.*

In the book of Edmunds and Cohen (2003), there is an interesting part regarding risk-adjustment methods.

**10. Cardiac Surgery in the Adult** by Louis Henry Edmunds, Lawrence H. Cohn, 2003. Published by McGraw-Hill Education

### ***New Risk Stratification Methods: Neural Networks and Other Computer-Intensive Methods***

*The goal of risk adjustment is to account for the contribution of patient-related risk factors to the outcome of interest. This allows patient outcomes to be used as an indicator of the care rendered by physicians or administered by hospitals. This chapter has outlined some of the risk-adjustment methods commonly used for this purpose, including use of multivariate analyses to predict patient outcomes based on patient risk factors. Inevitably, refinements and use of newer techniques will be brought to bear on the problem of risk adjustment. One of the most promising newer risk-adjustment methods is the use of neural networks (also termed artificial intelligence) to develop prediction models based on patient risk factors.*

*A weakness of multivariate regression techniques is that some variables have too low an incidence to be used in multivariate regression models but still contribute significantly to outcome. This weakness is overcome by use of neural networks and cluster analysis and are far less affected by low frequency of a particular variable.*

### **8.6 An example of the use of AI methods**

This example on the use of neural networks in a clinical setting is taken from a study carried out at the National Transplant Centre of Italy, and is concerned with estimating the waiting time for a patient with defined clinical variables. A Neural Network model can be used to model complex relationships between inputs and outputs or to find patterns in data, and is implemented in order to evaluate the mean waiting time of patients on the transplant list.

The specific aim of this study is to obtain a useful tool to estimate the patient waiting time on the transplant list, when we have a series of clinical parameters for the patient. A neural network

model with 10 neurons (nodes) was adopted with 4 neurons in the input layer, 5 neurons in the hidden layer, and one neuron in the output layer.

The four input neurons contain the following variables (integer variable , after the necessary information coding) : Blood Group, Patient Age, PRA, Transplant Centre. The output neuron gives the Waiting Time on the list.

We used a sample of about 4000 selected events, divided in a training set and a validation set each composed of 2000 transplant. Each event is made of the four input variables and the one output variable. The one output variable is the “expected” waiting time.

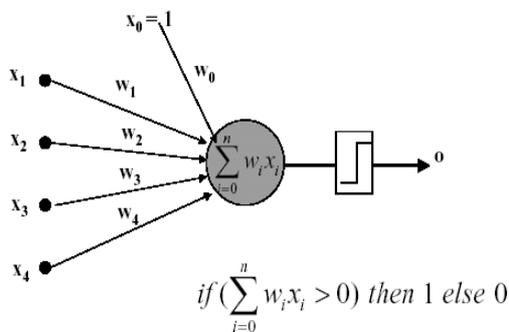
*Training phase:*

For each event in the training set, we submit the four inputs values to the neural network and then we obtain, from the network, the one output “waiting time”. This last value will be compared with the “expected” waiting time from the training set. The difference between these values, using the back-propagation algorithm, is used to adjust the network's weights so to minimize the error in the “waiting time list” prediction on the training set. Every time the “training set” sample is submitted to the neural network we have an epoch. The training stops when a given number of epochs elapses, or when the error reaches an acceptable level, or when the error stops improving.

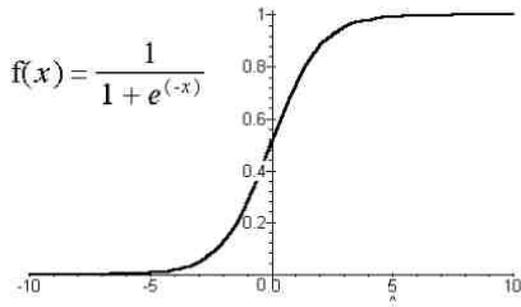
*Validation phase:*

After the validation phase, we submit the “validation set” sample to the neural network to evaluate its prediction power by estimating the error between the neural network output and the “expected” value.

Neural Networks are made of fundamental units , said “neurons” or nodes, that are connected among them in two ways, so there are two networks types: "Perceptron" or feed-forward and "Hopfield ". In the "Hopfield" model, every neuron is connected to all the others nodes, while in the "Perceptron", the neurons are gathered in layers and the information flow from a layer to the next one in unidirectional way. In both cases the basic unit , the neuron, works (figure 8.4) in very simple way: it receives a series of inputs ( $x_i$ ) which are summed with their own weights  $w_i$  . That weighted sum is used to obtain an output value according to a determined function, said “activation function”(each unit generates output or "fires" if the activation is above a threshold value). The activation functions are mainly the “step function” and the “sigmoid” or logistic function (figure 8.2).



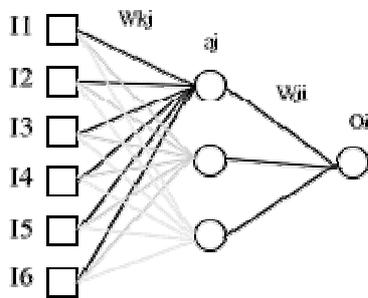
**Figure 8.2** Basic unit of a neural network



**Figure 8.3** Example of activation function

In particular, a generic feed-forward network (Multi Layer Perceptron) is constituted by an input layer (Ii), one or more hidden layers (aj) and an output layer (Oj), as shown in Figure 8.3. Every unit receives a linear combination of the preceding layer outputs neurons. The connections among the neurons are modulated by weights (“synaptic weights”  $w_{ij}$ ) that store "knowledge" of the network.

Networks learn by changing the weights of the connections. In Figure 8.3,  $w_{kj}$  shows the weights between the k-th input and the j-th neuron of the hidden layer, while  $w_{ji}$  shows the weights between the j-th neuron of the hidden layer and the i-th neuron of the output layer.



**Figure 8.6** Multi Layer Perceptron

The weights optimisation is obtained by different algorithms, but the most used is said "Back-Propagation". This algorithm allows the network to learn in four phases:

1. Network initialisation by choosing random synaptic weights;
2. Training by using a sample composed by couples (Input, expected Output);
3. Comparing the output of the network with the expected one;
4. Changing synaptic weights in relation to differences between the network output and expected output

## **8.7 Conclusions**

As can be seen from this chapter, the use of so called Artificial Intelligence techniques cannot be considered as a substitute for more classical methodologies. Instead, they provide an alternative approach that complements standard methods. In-depth analysis and comparison with standard methodologies provides a means of evaluating results. A further consideration is how these methodologies could in the future cope with the rising complexity of the interaction between clinical parameters and associated problems in interpretation.

## 9 COMPARISON OF STATISTICAL SOFTWARE

The focus of this chapter is on software for survival analysis, since this is the most commonly needed technique in the analysis of transplant data.

### 9.1 Available software

For each package mentioned below, we include contact information (US information is supplied where available; many have numerous contact sites outside the US also) as well as a table showing which of the above modelling capabilities. The list is not complete; emphasis was placed on packages that include modelling procedures, especially Cox's proportional hazards model. The packages are in alphabetical order.

1. BMDP, SPSS Inc., 444 N. Michigan Ave., Chicago, IL 60611, USA; (312) 329-4000.
2. Egret, Cytel Software Corp., 675 Massachusetts Avenue, Cambridge, MA 02139, USA; (617) 661-2011.
3. Epicure, HiroSoft International Corp., 1463 E. Republican Ave., Suite 103, Seattle, WA 98112, USA; (206) 328-5301.
4. Epilog Plus, Epicenter Software, PO Box 90073, Pasadena, CA 91109, USA; (818) 304-9487.
5. Limdep, Econometric Software, Inc., 15 Gloria Place, Plainview, NY 11803, USA; (516) 938-5254.
6. NCSS, 329 North 1000 East, Kaysville, UT 84037, USA; (801) 546-0445.
7. SAS, SAS Institute, Inc., Box 8000, Cary, NC 27511, USA; (919) 677-7000.
8. Spida, The Statistical Laboratory, Macquarie University, NSW 2109 Australia; 02-850-8792.
9. S-PLUS, MathSoft, Inc., Data Analysis Products Division, 1700 Westlake Avenue North, Suite 500, Seattle, WA 98109, USA; (800) 569-0123.
10. SPSS, SPSS Inc., 444 N. Michigan Ave., Chicago, IL 60611, USA; (312) 329-4000.
11. Stata, Stata Corp., 702 University Drive East,

College Station, TX 77849, USA; (800) 782-8272.

12. Statistica, StatSoft, 2325 East 13th Street, Tulsa, OK 74104, USA; (918) 583-4149.

13. Survival, Salford Systems, 8880 Rio San Diego Dr., Suite 1045, San Diego, CA 92108, USA; (619) 543-8880.

14. True Epistat, Epistat Services, 2011 Cap Rock Circle, Richardson, TX 75080-3417, USA; (214) 680-1376.

## 9.2 Software capability

Table 1 shows some of the capabilities of each of the packages, and uses the abbreviations described below.

1. Under “Nonparametric”: “K-M” is the Kaplan Meier product limit estimator; “N-A” is the Nelson Aalen estimator.

2. Under “Semiparametric”: “CPH” is Cox’s proportional hazards model; “t-d” means that one can use time-dependent variables in CPH; “t-v” means that one can use time-varying (or, step function) variables in CPH; “H-H” means the Han Hausman estimator.

3. Under “Types of Residuals”; note that the martingale and Schoenfeld residuals appear to be the most important types as they are particularly useful for checking major regression assumptions; the score residuals are useful in checking for influential observations, as well as for computing sandwich covariance matrix estimates: “C-S” is Cox Snell; “Std” means standardized residuals; “M” is martingale residuals (these are useful for assessing the adequacy of an hypothesized transformation of a predictor variable); “D” is deviance residuals; “Score” is score residuals; “Sch” is Schoenfeld residuals, which are useful for testing the PH assumption.

4. Under “Doc[umentation] for checking assumptions”, the possible entries are: “Yes”, meaning that the testing of both the proportional hazards (PH) assumption and the loglinear assumption are discussed in the manual; “PH”, meaning that only the testing of the PH assumption is discussed; “No”, meaning there is no discussion of testing assumptions in the manual (the user may still be able to test assumptions, but the manual provides no help or guidance).

5. Under “Left-trunc[ation]-delayed entry”: “DE” means that the package can handle observations whose entry into the risk set is delayed (software packages that have this capability have it for the Cox model only, not for parametric models); “LT” means that the package can handle left-truncated data; “neither” means that the package can’t deal with either delayed entry or left-truncation.

6. While there are no abbreviations under “Type of package”, it is worthwhile to note that packages that are extendable, and have active user communities making extensions, may well have additional techniques; two of the included packages have such user communities:

(i) S-PLUS: not only was the original routine written by a user (Terry Therneau), but the author of this article has written a coordinated set of functions for in-depth survival modelling which includes:

- (a) relaxing and testing linearity assumptions using cubic splines
- (b) automatic pooled Wald tests (see Likelihood)
- (c) bootstrap model validation
- (d) presentation graphics for describing models

(ii) Stata: there were many additions to the prior version, including Bailey-Makeham models, which will also work with the current version; the first author has added cubic splines to Stata's toolbox.

**Table 1:** a description of the main important software packages:

| Package      | Nonparametric                | Semiparametric   | Parametric                  | Type of Residuals     |
|--------------|------------------------------|------------------|-----------------------------|-----------------------|
| BMDP         | K-M                          | CPH,t-d,t-v      | Several                     | C-S,Std               |
| Egret        | K-M                          | CPH,t-d,t-v      | Several                     | None                  |
| Epicure      | K-M, N-A                     | CPH,t-d,t-v      | Several                     | M,D,Score,Sch         |
| Epilog Plus  | K-M                          | CPH,t-d,t-v      | None                        | None                  |
| Limdep       | K-M                          | CPH,t-d,t-v, H-H | Several                     | None                  |
| NCSS         | K-M                          | CPH              | Some (but not covariates)   | None                  |
| SAS          | K-M                          | CPH,t-d,t-v      | Several                     | M,D,Score,Sch         |
| Spida        | K-M                          | CPH,t-d,t-v      | None                        | None                  |
| S-PLUS       | K-M, N-A                     | CPH,t-v          | Several                     | M,D,Score,Sch         |
| SPSS         | K-M                          | CPH,t-d,t-v      | None                        | C-S,M,Sch             |
| Stata        | K-M                          | CPH,t-d,t-v      | Exp. Weib.                  | M,Score               |
| Statistica   | K-M                          | CPH,t-d          | Several                     | None                  |
| Survival     | K-M                          | CPH,t-v          | Several                     | None                  |
| True Epistat | K-M                          | CPH,t-v          | None                        | None                  |
| R            | K-M, N-A                     | CPH,t-d,t-v      | Several                     | M,D,Score,Sch         |
|              |                              |                  |                             |                       |
| Package      | Doc. for Checking assumption | Counting process | Left-trunc./ delayed entry  | Type of Package       |
| BMDP         | Yes                          | No               | DE                          | Complete,general      |
| Egret        | PH                           | No               | DE                          | Special purpose       |
| Epicure      | PH                           | Yes              | DE                          | Special purpose       |
| Epilog Plus  | PH                           | No               | Neither                     | Complete, biostat     |
| Limdep       | PH                           | No               | LT (parametric models only) | Complete, econometric |
| NCSS         | No                           | No               | Neither                     | Complete,general      |
| SAS          | PH                           | No               | DE                          | Complete,general      |
| Spida        | PH                           | Yes              | DE                          | Complete,general      |
| S-PLUS       | Yes                          | No               | DE                          | Complete,general      |
| SPSS         | PH                           | No               | neither                     | Complete,general      |
| Stata        | PH                           | Yes              | DE                          | Complete,general      |
| Statistica   | PH                           | No               | Neither                     | Complete,general      |
| Survival     | PH                           | No               | Neither                     | Special purpose       |
| True Epistat | No                           | No               | Neither                     | Complete, biostat     |
| R            | PH                           | Yes              | DE                          | Complete,general      |

### 9.3 Choice of software

In choosing software for statistical analysis and in particular for survival analysis, the following features need to be considered:

*Cost* – Obviously the cost is one of the factor to be taken into consideration, in particular if the software licence foresee an expiration date or some free updating, some kind of support and some paper or electronic documentation; many excellent software use the Open Source modality which is free and as we will see also has some interesting characteristics.

*Expandability* – that means the possibility to integrate or modify the integrated statistical methodologies as needed.

In particular the software related to Open Source licence are those with the major possibility to fit to the different needs.

*Possibility to foster the computational possibilities* – In this case in particular we refer to the possibility of elaboration that foresee an elevated time machine to carry out the application, e.g. in a cluster environment.

*Calculation speed* – the difference with the above mentioned characteristic is the efficiency of the calculation in the same operative system conditions and hardware.

*Updating* – that means that the application is updated, fostered in all its characteristics at regular interval; these characteristic could affect the costs due to the fact that the acquired licence could or could not foresee those updating.

*Support* – that means the availability of paper of electronic manual supplied by the software house, and, more in general, to all the information that could be found on the Internet like Forum full of indications and useful indications.

*Versatility with operating system and/ or hardware* –the possibility to install or use the same software regardless of the operative systems or hardware used in the organization could be an important element of homogenisation in order to allow the exchange and sharing of the activities carried out.

*The interaction with other data bases* – for each data processing the simplification and the creation of the data sample could be extremely important, for this reason it could a important feature that the software allows and is provided of the tools for data enquiry.

*Possible web oriented development* – In particular lately it could be useful that an application of data processing is provided with characteristics that allows data publishing on the net and, this is a more sophisticated feature, the possibility to carry out the processing through a web interface.

*Integration with OLAP system* – In a more sophisticated environment the IT System data are commonly extracted and processed through query with what is called OLAP (On Line Analytical Processing) applicative. In these cases it is important that the system of data processing could be integrated either with this applicative.

*Graphic features* – this is a sometimes disregarded feature but the graphic characteristics of an applicative for statistical elaboration are important; e.g. if it is possible and if it is easy to customise a graphic and the possible exportation formats.

*Integration with other software* – it is a complex issue but for example the possibility to integrate the possible statistical methodologies with other development environments, including C++, Perl, Python and others.

## 10 CONCLUDING REMARKS

In this review of statistical methods used in the analysis of data on transplant outcomes, we have described a number of issues that need to be considered. However, the main part of this review outlines the statistical methods that are in common use in the analysis of particular types of outcome, particularly survival rates and survival times.

Although it is standard practice to report survival rates following transplantation, and possibly for sub-groups of the population defined by recipient age, primary disease etc, it is important to investigate factors that affect survival of the graft or patient using the methods described in Chapter 3. The identification of factors that affect survival offers the potential to change clinical practice to improve the quantity or quality of life of a patient. As a specific example, it was found in the UK that the period of cold ischaemic time following retrieval of a kidney needs to be kept below 22 hours; after this time, the period of cold ischaemia has a significant detrimental effect on graft outcome. This result was communicated to individual transplant units, so that necessary resources, particularly theatre availability, could be secured in a timely manner. Similarly, analyses that provide information about the factors affecting the chance of a transplant has an impact on organ allocation schemes. If for example, it is found that certain sectors of the population do not have equity of access, steps may be taken to address this.

The material on monitoring centre performance is of particular importance. Indeed, the methods based on CSUSUM charts described in Chapter 4 are now well established in certain areas of medicine. However, these techniques are not widely in transplantation at present, and yet they are fundamental to determining whether organs allocated to particular centres are being transplanted safely. This is of course an area that is particularly sensitive. Centres may be concerned at the very idea of a monitoring process being in place. However, as long as the introduction of such schemes is handled delicately, and clinicians are involved at all stages of the process, it comes to be viewed as non-threatening. In order to alleviate the concerns of centres who undertake more difficult cases, and who therefore expect inferior results, and risk adjustment for case mix is essential.

Of necessity, this review is highly selective, and concentrates on the most commonly encountered techniques in the analysis of transplant data. Moreover, most of the methods can be implemented using software that is readily available. It is therefore hoped that this review will serve as a guide to transplant units who are introducing methods for monitoring and analysis.