SURVIVAL ANALYSIS IN JOINT REPLACEMENT SURGERY

ANDREW J. CARR, RICHARD W. MORRIS, DAVID W. MURRAY, PAUL B. PYNSENT

From the Nuffield Orthopaedic Centre, Oxford, Guy's and St Thomas' Hospitals, London, and the Royal Orthopaedic Hospital, Birmingham, England

After the widespread introduction of joint replacement surgery the early reports of outcome described success in terms of relief of pain, restoration of movement and the ability to return to normal activities. These were subsequently refined to include some assessment of the longevity of the implant and many reports now describe excellent success rates over long periods of time. The pioneering age of joint replacement surgery appears to be over.

Alterations to the design of the original prostheses have become increasingly numerous and the clinician is faced with the difficult task of choosing the best implant for his patients. Innovation is more common than evaluation and new designs can be introduced with no published results available or, at best, short follow-ups on small numbers of cases. Surgeons using these new prostheses risk encountering abnormally high rates of failure. Well-designed and well-conducted clinical trials are the best method of evaluating new prostheses and should be mandatory before their widespread use is allowed.

The most common method employs a survival or life table, first put into effective use by the astronomer Edmond Halley (1656 - 1742). The statistical methods used in survival analysis have been developed considerably in the last four decades, principally for cancer research, but only recently have they been adapted for use in orthopaedic surgery. This form of analysis defines a failure point or terminal event and provides an assessment, not only of how many patients failed, but also of how long after the operation failure occurred. It does not measure function or pain unless they are included in the definition of failure.

CALCULATING RATES OF SUCCESS OR FAILURE

The survival table. A survival table allows us to estimate the probability that an implant will survive for a given length of time. If a joint replacement trial were to be continued until all the prostheses had failed, the success rate for each time period could be calculated by simply subtracting the number of failures from the number at risk at each time period.

Failure of a prosthesis, however, unlike death, is not an inevitable event. Since some operations have still not failed at the time of analysis the observations are incomplete and are termed censored observations.

The data from a hypothetical trial of 1000 joint replacements performed over a period of 15 years have been used to calculate a survival table (Table I) and a survival curve (Fig. 1). The methods used are based on those described by Dobbs (1980) and Armitage and Berry (1987).

The cases are first sorted by follow-up time and grouped into intervals, in this example, of one year. For each interval of follow-up the number at the start, the number of failures and the number of withdrawals are calculated. The last includes those patients who have died, those who have reached the end of their follow-up without failing and those who have been lost to follow-up for more than a year. The number of cases at risk during any interval is calculated as the number at the start less half the number of withdrawals. The latter is used as a reasonable compromise; clearly not all the withdrawals were exposed to risk for the whole time interval. Each case is recorded in the table at the time of the most recent review.

The proportions which failed and which succeeded are then determined. The cumulative success rate at the

A. J. Carr, ChM, FRCS, Senior Registrar
D. W. Murray, MD, FRCS, Senior Registrar
Nuffield Orthopaedic Centre, Windmill Road, Headington, Oxford OX3 7LD, UK.
R. W. Morris, MSc, Lecturer in Medical Statistics
United Medical and Dental Schools, Guy's Hospital, St Thomas Street, London SE1 9RT, UK.
P. B. Pynsent, PhD, Director of Research
Royal Orthopaedic Hospital, Bristol Road South, Northfield, Birmingham B31 2AP, UK.
Correspondence should be sent to Mr A. J. Carr.

©1993 British Editorial Society of Bone and Joint Surgery
0301-620X/93/2545 $2.00

178

THE JOURNAL OF BONE AND JOINT SURGERY
end of a particular year is the product of the proportion succeeding for all previous years multiplied by the proportion succeeding at the end of that year. The calculated cumulative proportion may be expressed as a percentage by multiplying it by 100.

A check list for conducting (or criticising) a joint replacement trial is provided in the appendix.

The follow-up interval. With the life-table method described above the cumulative estimate of success will change only at the end of each chosen interval. With the product-limit method, described by Kaplan and Meier in 1958, the cumulative estimate of success is calculated on a daily basis and changes with each failure. The former method is more easily portrayed in tabular form and is suitable for graphical representation (Fig. 1).

**Definition of failure.** In joint replacement surgery the usual end-point is the decision to remove the prosthesis and revise or convert it to some other form of treatment. This end-point has been criticised because the criteria used to decide the need for removal will vary between patients and surgeons. Another criticism is that the threshold for revision surgery may be different for one form of implant compared with another, for example unicompartmental and total knee replacement. There

![Cumulative survival estimate (per cent) vs Years since operation](image)

**Fig. 1**

Survival curve derived from the hypothetical data used in Table I (including 95% confidence intervals). The number of patients at risk in the first year of follow-up and in succeeding years was 984, 915, 803.5, 711.5, 609, 489, 345, 169, 86.5, 63, 34, 19, 7, 1.5, 0.

<table>
<thead>
<tr>
<th>Year (post-operative)</th>
<th>Number at start</th>
<th>Withdrew at last review</th>
<th>Number at risk</th>
<th>Proportion failing (%)</th>
<th>Proportion succeeding (%)</th>
<th>Cumulative estimate of survival (%)</th>
<th>Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 1</td>
<td>1000</td>
<td>30</td>
<td>984</td>
<td>0.3</td>
<td>99.7</td>
<td>99.7</td>
<td>0.2</td>
</tr>
<tr>
<td>&gt; 1 to 2</td>
<td>965</td>
<td>85</td>
<td>915</td>
<td>2.2</td>
<td>97.8</td>
<td>97.6</td>
<td>0.5</td>
</tr>
<tr>
<td>&gt; 2 to 3</td>
<td>845</td>
<td>70</td>
<td>803.5</td>
<td>1.5</td>
<td>98.5</td>
<td>96.1</td>
<td>0.7</td>
</tr>
<tr>
<td>&gt; 3 to 4</td>
<td>750</td>
<td>65</td>
<td>711.5</td>
<td>1.4</td>
<td>98.6</td>
<td>94.7</td>
<td>0.9</td>
</tr>
<tr>
<td>&gt; 4 to 5</td>
<td>663</td>
<td>90</td>
<td>609</td>
<td>2.5</td>
<td>97.5</td>
<td>92.4</td>
<td>1.1</td>
</tr>
<tr>
<td>&gt; 5 to 6</td>
<td>540</td>
<td>82</td>
<td>489</td>
<td>2.4</td>
<td>97.6</td>
<td>90.2</td>
<td>1.4</td>
</tr>
<tr>
<td>&gt; 6 to 7</td>
<td>426</td>
<td>105</td>
<td>345</td>
<td>2.9</td>
<td>97.1</td>
<td>87.5</td>
<td>2</td>
</tr>
<tr>
<td>&gt; 7 to 8</td>
<td>239</td>
<td>120</td>
<td>169</td>
<td>2.9</td>
<td>97.1</td>
<td>85</td>
<td>3.4</td>
</tr>
<tr>
<td>&gt; 8 to 9</td>
<td>94</td>
<td>10</td>
<td>86.5</td>
<td>2.3</td>
<td>97.7</td>
<td>83</td>
<td>3.9</td>
</tr>
<tr>
<td>&gt; 9 to 10</td>
<td>77</td>
<td>10</td>
<td>63</td>
<td>0</td>
<td>100</td>
<td>83</td>
<td>4.9</td>
</tr>
<tr>
<td>&gt; 10 to 11</td>
<td>49</td>
<td>20</td>
<td>34</td>
<td>0</td>
<td>100</td>
<td>83</td>
<td>6.4</td>
</tr>
<tr>
<td>&gt; 11 to 12</td>
<td>29</td>
<td>15</td>
<td>19</td>
<td>0</td>
<td>100</td>
<td>83</td>
<td>11</td>
</tr>
<tr>
<td>&gt; 12 to 13</td>
<td>9</td>
<td>1</td>
<td>7</td>
<td>42.8</td>
<td>57.2</td>
<td>47.5</td>
<td>24</td>
</tr>
<tr>
<td>&gt; 13 to 14</td>
<td>2</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
<td>100</td>
<td>47.5</td>
<td>34</td>
</tr>
<tr>
<td>&gt; 14 to 15</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>47.5</td>
<td>–</td>
</tr>
</tbody>
</table>
does seem to be a good case for identifying patients for whom revision surgery is indicated but prevented because they are otherwise unfit. Attempts have also been made to extend the definition of failure by including an assessment of pain or radiological signs of failure (Tew and Waugh 1982), but these alternatives may reduce the objectivity of the end-point. It may be useful to define two or more end-points and to carry out separate analyses for each.

Secondary outcome measures. Since there are often several mechanisms of failure (infection, component loosening, etc) it may be interesting to analyse these as secondary outcome measures. It is also reasonable to ask the question whether the probability of failure varies with time. The answer can be represented graphically and is termed the hazard rate (Dobbs 1980). If the hazard rate is not constant then predictions of outcome beyond the time limits of the trial are unreliable. The probability of failure is likely to change during the course of the trial due to the increasing experience of the operator as well as to modifications in the prosthetic design.

Loss to follow-up. The trial begins at the time of the first operation and ends when it is decided to analyse the results. Ideally, the period at risk for every patient would run from the date of operation to the end date of the trial — the elapsed time. If, however, the patient is assessed some time before the end date the period at risk only runs to the date of the last follow-up; this is the follow-up time. The follow-up time should of course be as close as possible to the elapsed time because there is no satisfactory way of allowing for any shortfall.

Patients who are lost to follow-up are withdrawn from the study at the time of their last assessment, and are entered in a separate column in the survival table. The assumption has to be made that they would have behaved in the same way as those patients not lost to follow-up (Armitage and Berry 1987; Dorey and Amstutz 1989), but this may not be correct. In addition, the proportion of failures that occurs in patients lost to follow-up is unlikely to be the same in different hospitals or countries. One way of dealing with them is to assume that they all fail. This ‘worst case’ method, although it is unlikely to represent the true failure rate, would certainly encourage completeness of follow-up. Carr et al (1992) have suggested that the elapsed time and the follow-up time should each be portrayed graphically. The area between the two curves, representing the period in which the outcome of the implant is unknown, offers the reader a measure of the completeness of the follow-up.

Small numbers and standard errors. Conclusions based on cumulative survival curves must be treated with caution, particularly from data on the right-hand side of the curve where the number of patients at risk may be small. Such graphs should never be interpreted as evidence that after a certain period of time the chance of failure of a joint replacement is effectively nil. Often a dramatic (and meaningless) drop occurs at the end of the curve (Fig. 1) as a result of one failure among a very small number of cases at risk. If the cumulative estimate of success towards the end of a trial is 0.8 and one case fails when there are only four at risk then the cumulative success rate will decrease to 0.6 (0.75 × 0.8). It is important to include information about the number of patients at risk in the life table and at the base of the graph.

Kaplan and Meier (1958) used the Greenwood equation to calculate standard errors. It is based on the variance of results in each period of the trial in which there was a failure. This can produce a spuriously low standard error on the right-hand side of the curve where the number of failures and the number at risk may be low. More recently, Knutsen, Lindstrand and Lindgren (1986) and Lettin, Ware and Morris (1991) have reported survival rates with confidence limits using the method suggested by Peto et al (1977). Their equation uses approximate binomial methods and takes account of numbers at risk. This method and that described by Rothman (1978) are conservative and deal appropriately with the increasing uncertainty to be expected on the right-hand side of survival curves. With the low failure rates usually encountered in joint replacement surgery, however, there is no completely satisfactory way of obtaining confidence limits.

How much data? In general, as much data as possible should be collected at the outset which can be retained and used in the later analysis, but it is advisable to collect only limited data at follow-up assessment. The more concise the proforma the better it will be completed and essential pieces of information will not be omitted. All complications should be described, any radiographic projections used should be defined and the causes of failure recorded. Information about continuous variables should be recorded exactly but it is useful to group subjective assessments and assign them numbers which can be stored on a computer. The advice of a statistician should be sought when setting up the trial if mistakes are to be avoided.

The construction of a life table, the calculation of standard errors and the simple comparison of survival between two groups may be accomplished without a computer if the number of patients is fairly small. A very helpful menu-driven program, however, usable on a PC, is Confidence Interval Analysis (Gardner, Gardner and Winter 1989). The use of Cox’s method will probably require collaboration with a statistician.

COMPARING RESULTS

The most important use of clinical trials is to compare different methods of treatment. Survival curves allow such comparisons to be made even although different numbers of patients may be at risk at different times.

The log-rank test. The simplest statistical method that can be applied is the log-rank test, suggested by Mantel (1966) and described with fully worked examples by Peto
et al (1977). Its underlying principle is simple: if, at some
given time, two-thirds of the patients still at risk were
treated by operation A and one-third by operation B then
we should expect, if operations A and B are equally
effective, that two-thirds of the failures in the next period
would be in group A and one-third in group B.

The number of patients at risk of failure at any time
is calculated as for the survival table and the expected
number of failures in any time period can then be
calculated for each operation group. The total expected
failures is the sum of the expected failure values calculated
for each of the various time points. The log-rank test
compares the observed failures in the two groups with
the expected failures for the two groups. As with the life
tables, the expected numbers of failures may be computed
for the end of each time period, or re-calculated every
time a failure occurs. The former method was used by
Curtis, Bland and Ring (1992) in comparing the survival
of several designs of knee replacement. In general,
however, it is recommended that the exact length of time
that each replacement survives should be used in all
formal statistical tests, thereby maximising the power of
the analysis.

Stratification. The log-rank method can be used to
investigate the effect of variables such as age, sex, weight
or disease type on survival. Statisticians refer to these as
explanatory variables and to the process of forming groups
as stratification. Ideally, the groups should be created
from data collected at the outset of the trial. These
analytical methods are designed to discover prognostic
determinants so that they can be taken into consideration
when two joint replacements are compared. It is
important to know if an apparent difference between
the survival of two prostheses is in fact due to the presence
of more ‘good-prognosis’ cases in one group than in the
other.

An advantage of dividing patients into groups is
that the performance of one prosthesis may only be
noticeably better than another in one group of patients.
The process of stratification, however, diminishes the
numbers and spuriously significant results can be pro-
duced. If differences are found then it is useful to perform
a number of different comparisons, first treating losses to
follow-up as all failing and then as if they were all
successful. If significant differences persist then the
conclusion is more robust.

Cox’s proportional hazards model. Another method
for taking into account the effect of prognostic variables is
the proportional hazards model of Cox (1972). It has
been described with worked examples by Christensen
(1987) and Kalbflleisch and Prentice (1980). Like the log-
rank test it makes the assumption that the ratio of the
hazard for operations in group A to that in group B
remains constant. The value of the model is that it can
be used to assess the influence of several variables at
once, including that of the use of either joint replacement
A or B. It is therefore possible to assess a difference in
outcome between two joint replacements while taking
account of the effects of other influential variables.

These methods of analysis are complex and require
the co-operation of a statistician. There is a considerable
risk of finding spurious associations, particularly if many
variables are examined and if the failure rate is low.

Randomised and multicentre trials. Orthopaedic surgeons
must be aware of the serious errors which may arise when
comparison is made with the published results of other
investigations (Sacks, Chalmers and Smith 1982). When
analysing the survival of joint replacements it is prefer-
able to use randomised rather than historical controls.

Stirrat et al (1992) gave a number of reasons why
randomised trials have seldom been used in evaluating
surgical procedures. Neither blinding of the subjects nor
the use of a placebo is possible, and when an established
technique is compared with a new method, surgical
experience tends to favour the accepted method. There
may also be a number of less valid reasons why surgeons
do not perform randomised clinical trials.

The manufacturers of prostheses are perhaps in the
best position to control the use of the implants that they
sell and to finance proper trials, but if surgeons are
prepared to use new implants without proper evaluation
of their efficiency then there is little incentive for
manufacturers to be involved in such activities.

In joint replacement surgery the reported rates of
failure are low, the detection of significant differences
between prostheses requires the study of large numbers
of patients, and it may take ten or more years before the
outcome can be properly assessed. Such numbers may
not be attainable by one surgeon or even one centre and
the case has been made for incorporating data from
different trials by using the technique of meta-analysis
(Chalmers 1990).

Conclusions. Any new surgical procedure or implant
should be properly compared with currently accepted
standards, before it is accepted into general use. This
should not be regarded as an optional requirement;
failure to make such comparisons is unethical. There
remain many unresolved questions in joint replacement
surgery; well-constructed clinical trials offer the best
hope of answering them. Such trials are onerous and time
and money are needed to collect, record and analyse the
data. The relatively few clinicians who undertake such
trials must have an understanding of the statistical
methods used and the problems that may be encountered.
It is equally important that the much larger number of
clinicians who read the reports of such trials should be
able to criticise them in an informed way so that the
practice of joint replacement rests on a sound scientific
basis.

The authors are grateful for the help of Dr R. Petro, Dr J. J. O’Connor
and Mr C. J. K. Bulstrode.

No benefits in any form have been received or will be received
from a commercial party related directly or indirectly to the subject of
this article.
APPENDIX

A joint replacement trial checklist

Planning the trial

1) Entry criteria and a trial protocol should be defined.
2) Although most joint replacement trials involve case studies with historical controls, randomised controls are preferable.
3) Statistical help should be obtained when planning the number of patients to be included in the trial.
4) The definition(s) of failure must be made.
5) Approval of the ethics committee should be obtained, and informed consent from the patient, if a new prosthesis is used.
6) Provision must be made for adequate follow-up assessment.
7) In a multicentre trial a co-ordinator should be appointed.

Performing the trial

1) Adequate information should be collected at the outset of the trial and at subsequent follow-up including data regarding possible explanatory variables. A standard assessment form is preferable.
2) If radiographic assessment is made then the projections used must be defined.
3) Every effort must be made to trace patients lost to follow-up. The use of a central register and national health number may help.
4) The protocol should not be changed during the trial without proper justification and clear documentation.

5) Complications should be recorded.

Analysing the trial

1) The dates of analyses should be planned at the outset. Repeated interim analyses run the risk of generating spurious differences (type I errors) and special statistical methods are required to allow for this.
2) At the end of the trial the results should be presented in the form of a survival analysis. We suggest that the trial should last a minimum of five years.
3) Confidence limits should be calculated and numbers at risk displayed.
4) Patients lost to follow-up should be represented in the life table.
5) Analysis using several definitions of failure may be useful.
6) The survival curves of the trial prosthesis (and the control) should be compared using established statistical methods.

Interpreting results of the trial

1) The completeness of follow-up should be taken into consideration.
2) If the number at risk is small then the cumulative estimate of success should be interpreted with caution.
3) Trials demonstrating no difference from established results may have inadequate periods of follow-up.
4) Comparisons with historical controls should be interpreted with caution.

REFERENCES


