Predicting survival after treatment for fracture of the proximal femur and the effect of delays to surgery


Published in:
Journal of Clinical Epidemiology

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
Predicting survival after treatment for fracture of the proximal femur and the effect of delays to surgery

J. Elliot, T. Beringer, F. Kee, D. Marsh, C. Willis, M. Stevenson

Department of Trauma and Orthopaedics, Royal Victoria Hospital, Grosvenor Road, Belfast BT 12 6BJ, Ireland
The Elliot Dynes Rehabilitation Unit, Royal Victoria Hospital, Grosvenor Road, Belfast BT 12 6BJ, Ireland
Department of Epidemiology and Public Health, Queen\'s University of Belfast, Musgrave Park Hospital, Stockmans Lane, Belfast BT 9 7JB, Ireland
Eastern Health, Social Services Board, Champion House, Linenhall St, Belfast BT 2 8BS, Ireland

Accepted 2 October 2002

Abstract

Background and Objectives: The aim of this study was to evaluate the prognosis after treatment for femoral neck fracture, to assess the impact of delay to surgery, and to devise a clinical prediction rule and score.

Methods: A prospective observational study was conducted in which 1780 patients treated surgically in two teaching hospitals between 1 November 1997 and 31 October 1999 were followed over 12 months. Logistic regression was used to distinguish the effects of predictor variables on survival. Using a probit transformation of the predicted posterior probabilities of death, a prognostic score was devised with scores constrained so that a nominal score of approximately 90 represented a 50:50 chance of survival over 12 months.

Results: Mortality was 30.1% in men and 19.5% in women. Increasing age, male gender, longer pre-operative delay, a higher American Society of Anesthesiology score, a lower Mental Test score, and a lower activities of daily living (Barthel) score were associated with increased risks of death. Of those waiting between 1 and 5 days for surgery, approximately 8 medium-risk and 17 high-risk patients (with prognostic scores of 90 and 120, respectively) would have to have their delay reduced to 24 hours to yield one additional survivor.

Conclusion: The application of prediction rules must be guided by ethical, social, and scientific concerns. © 2003 Elsevier Inc. All rights reserved.

Keywords: Femoral neck fracture; Prognosis; Operative delay

1. Introduction

Although we hope that risk factors such as poor diet and inactivity may be improved by lifestyle changes, current demographic trends guarantee that the health service will have to deal with growing numbers of patients with hip fractures for the foreseeable future. For example, over the last decade, caseloads have risen by between 5% and 10% per year, and it is estimated that by 2011 nearly 70,000 patients will require admission for this condition in England alone [1].

A concerted effort has been made to enhance services for these patients, with guidelines being produced by different bodies to improve their organization, delivery, and efficiency [2–4]. Although the objective of these services is to restore functional independence as much as possible, the consequence of the fracture can be life threatening for a substantial minority of patients. Although evidence from randomized trials is scarce, if patients whose mortality risk was highest could be identified earlier, it might be possible to marshal resources and target preventive interventions to reduce the risk. Knowing the magnitude of the risk attributable to the main prognostic factors is essential for the design of appropriate intervention trials. The use of risk scores, however, can also help us understand trends in outcomes over time and between treatment units. Consequently, the longer-term outcomes of hip fracture have been studied in a number of prospective studies with a view to identifying the key prognostic variables. Some of these studies have been modest in size, and few have attempted to validate the predictive accuracy of the resulting prognostic score.

In 1997, the fracture surgeons and the ortho-geriatric service in the Royal Victoria Hospital, Belfast, set up a large prospective Audit of Fractures of the Proximal Femur, with the express intention of determining the early and late consequences of hip fracture in their catchment area and of deriving an improved prognostic scoring system. Our objectives were twofold: (1) to determine what were the most important
prognostic variables in our population for outcome 12 months after hip fracture repair and (2) to devise a clinical prediction rule. We report here on the results for patients ascertained during the first 2 years of the audit’s operation. Given the local health Minister’s Priorities for Action (to reduce the maximum delay to treatment to no more than 48 hours), our null hypothesis was that delay was not associated with outcome at 12 months.

2. Methods

The fracture units based at the Belfast City and the Royal Victoria hospital co-operated in the establishment of the Audit before their merger and amalgamation on the Royal Victoria site in November 1999. Together they serve a population of approximately 800,000. A range of socio-demographic and medical information is routinely collected on admission and at 4, 6, and 12 months after injury. The main variables used in this study were age, sex, marital status, a proxy measure of material affluence based on the Townsend deprivation score of the electoral ward [5], an abbreviated Mini-Mental State score on admission [6], pre-injury Barthel (activities of daily living) score [7], and the American Society of Anesthesiology (ASA) Physical Status grading [8] (categories 1 [least impaired physical status] to 4 [most impaired physical status]). In addition, the delay from injury to operation was known. The Townsend Deprivation score is derived for each census ward and has four components: (1) the proportion of the labor force who were unemployed, (2) the proportion of households with no access to a car, (3) the proportion of households not owner occupied, and (4) the proportion of employed men and women in social classes IV and V.

This article is based only on patients who came to surgery and whose date of injury was known. Of the 1876 patients on the database for the years 1998 and 1999, 78 had no surgery, and 18 had injury dates that were clearly estimates (their admission dates were approximately 3 weeks or more from the inferred time of injury). Vital status 1 year after injury was ascertained by regular mailing and telephoning of the patients and their general practitioners. When patients did not reply (by mail or phone), their named caretaker was contacted. When no contact could be established through this route, the practice register was checked. The Central Services Agency maintains the practice registers and, by using the Registrar General’s data on death registrations, estimates that 99% of deceased patients’ records are removed within 100 days of the death (Dr. David Marshall, CSA, personal communication, 2003).

Univariate contingency tables using chi-squared statistics were calculated to initially assess the effect of possible prognostic factors on vital status (live/dead) 12 months after surgery. Multiple logistic regression was then used to distinguish independent effects on outcome. The odds ratios represent the multiplicative effect of a one-unit change in the independent variable (operation delay is on a square root scale). A backward procedure was used for variable selection with an initial F test value of $P > .05$ for exclusion. We also sought bi-variate interaction terms for main effects.

The performance of a prognostic model can be assessed in a number of ways. One commonly used method is to derive the c-statistic, which is identical to the area under the receiver operator characteristic curve. Essentially, all possible pairs of individuals where one is dead after 12 months and one is alive after 12 months are considered. Then, the number of such pairs where the posterior probability for the dead person is higher than the posterior probability for the person alive is counted. The latter as a ratio of the former is defined as the c-statistic [9].

We have assessed the performance of our models by randomly splitting the data set into two, deriving independent logistic regression models for each separately, and comparing the predictive performance of each (deriving both a c-statistic and a kappa value) using first their own group’s and then the alternate group’s data set. Finally, we compared the outcome of these models with a single model using the whole data set.

The logistic regression model provides posterior probabilities of death within 12 months as final output. Although these probabilities are useful to epidemiologists and statisticians, they are not user friendly to clinicians. We have further explored the utility of our model by deriving a simpler prognostic score, based on the magnitude of the coefficients in the finally selected logistic model. To check the validity of the scoring system, the scores were regressed against the probit transformation of the posterior probabilities. The probit transformation also normalizes the posterior probabilities (Figs. 1, 2), which is advantageous when there is some skew in the data. For instance, a model that is asked to discriminate between a large number of cases in the middle of the risk distribution has a much harder task than one asked to discriminate between a distribution of cases in which posterior probabilities close to 0 or 1 occur frequently.

Because probits are standard normal variates ($z$ values), a value of 0 equates to a cumulative probability of 0.5.

3. Results

A total of 415 men and 1365 women sustained hip fracture that was surgically treated in the Belfast City and Royal Victoria hospitals between 1 November 1997 and 31 October 1999. Twelve-month follow-up on all patients showed 30.1% mortality in men and 19.5% in women. The characteristics of these patients and their effects on 12-month mortality are shown in Table 1. The most obvious effects are the higher mortality with increasing age, male gender, ASA status, and lower mini-mental state and Barthel scores. Longer operation delays were also associated with poorer outcome, but there was a better outcome among single and divorced subjects.
In the modeling process, we found a linear relationship with the square root of delay. Although the effect is illustrated in Table 1 using five categories (ranging from <1 day to >10 days), our most parsimonious model was based on a single parameter \text{SQRT-delay} as a continuous variable. Although post-coding and Townsend scoring was unsuccessful in 7% of cases, there was no apparent effect of deprivation on outcome.
Is your patient at risk?

Following fractured neck of femur some patients have much reduced life span. The following ready reckoner is designed as a guide to indicate those patients who would most benefit from prioritisation for surgery.

Is the patient male? If so award 15 points

For ASA stage 2 award 20 points; at stage 3 award 25 points and at stage 4 award 35 points

For every completed two years of life award a point.

For every point lost on the Mental Test Score and Barthel scales, award a point.

Is the patient single or divorced? If so deduct 8 points.

How did your patient score? If the total is below 50, the patient will almost certainly survive and a small delay in time to surgery will be of small consequence.

If the total is between 50 and 120 your patient is in the group that would most benefit from priority – particularly so if the score falls between 70 and 100.

If the score is over 120 the odds are now stacked against survival and the effect of speedy surgery, while obviously still the goal, is nevertheless diminishing.

Fig. 3. Illustrative derivation of “at-risk” score.

Tables 2, 3, and 4 give the results of logistic regression for the data set overall and when divided randomly into two groups (group 1 and group 2). Because there were no deaths in ASA category 1 in group 2, the comparison category is ASA group 4. Largely, the same variables were selected by each data set. In group 1, the effect of marital status was not significant, and the confidence intervals for the Barthel score effect were 0.90 to 1.00. In group 2, the main (and interaction) effect of delay was not significant.

Table 5 summarizes the results of our data-splitting validation procedure. All of the c-statistics were in the range 0.73 to 0.80. Group 2 had a slightly wider range of explanatory variables than group 1, and predictions of the status of its members tended to be a little better. There was little deterioration in predicting individual status from an independently derived set of data when compared with predicting status directly from the data used to derive the model.

The distribution of posterior probabilities (Fig. 1) is considerably skewed to the right. The typical person admitted for neck of femur fracture had a small probability of not surviving 12 months, and hence the mode of the distribution was approximately 0.1. A minority of patients had a large likelihood of nonsurvival, with posterior probabilities in excess of 0.8. The distribution can be normalized by applying a probit transformation as shown in Fig. 2. Because probits are standard normal variates (z values), a value of 0 equates to a cumulative probability of 0.5, $-1$ equates to a cumulative probability of 0.16, and $-2$ equates to a cumulative probability of approximately 0.025. A value of $+1$ equates to a probability of 0.84.

Because it is rarely desirable for a clinician to feed data into a logistic regression “black box” to produce a posterior probability, we have devised a moderately simple scoring system that has a strong relationship with the probit scale described above. We have constrained these scores in such a way that a nominal score of approximately 90 represents roughly a 50:50 chance of survival over 12 months.

The basis of the score is illustrated in Panel 1. There is a virtually linear relationship between the derived scores and the probits. The calibration of our scoring system is illustrated in Appendix 1. Even though we have excluded terms that are delay dependent (we considered that practitioners would form a view about the patient prognosis on admission, before any elapsed delay), the calibration is still excellent. If investments are made to shorten delay time to surgery, for example, this scoring system can illustrate the implications. It enables us to make calculations of how the probability of death within 12 months changes as a result.
Our main objectives were to assess the determinants of survival 12 months after repair of hip fracture to assess the independent effects of delay to surgery and to devise a clinical prediction score. Although a number of studies from within and outside the UK have examined mortality after treatment for hip fracture, it is useful to put our own results in context and to describe how they affirm or extend previous findings.

First, although we share the rigor and discipline of prospective data collection to which studies like the East Anglian audit have subscribed [10], our sample size was substantially larger, and our follow-up was longer than in many others in the field [11–15]. Studies with different durations of follow-up are not strictly comparable because the prognostic factors for early versus late mortality may be inherently different. Studies may also differ according to the proportion of all patients presenting who end up having surgery. In our study, the proportion not receiving surgery was 4.2%, and we confined our prognostic model to the survivors 12 months after repair of hip fracture to assess the independent effects of delay to surgery and to devise a clinical prediction score.

**4. Discussion**

of delay and hence to calculate the number of additional patients who need to have their surgery within 24 hours to yield one additional survivor at 1 year (a sort of “number needed to treat”). Table 6 illustrates the results. The dividend is greatest for patients with scores in the middle range.
had this observation been attributable to bias as a further explored in future data from this ongoing cohort other categories) runs counter to other studies and will be survival among the single and divorced (compared with all score) was found in the East Anglian audit [10]. The better activities of daily living (measured here using the Barthel score) among those who, pre-injury, were less able to perform

<table>
<thead>
<tr>
<th></th>
<th>Odds ratio</th>
<th>Confidence intervals</th>
<th>Wald</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex (female versus male)</td>
<td>0.39</td>
<td>0.29–0.52</td>
<td>39.96</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>ASA (1) versus (4)</td>
<td>0.10</td>
<td>0.03–0.35</td>
<td>13.06</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>ASA (2) versus (4)</td>
<td>0.39</td>
<td>0.25–0.63</td>
<td>15.42</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>ASA (3) versus (4)</td>
<td>0.47</td>
<td>0.32–0.70</td>
<td>14.01</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Mental score (0–10)</td>
<td>0.92</td>
<td>0.88–0.96</td>
<td>14.22</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Barthel score (0–20)</td>
<td>0.94</td>
<td>0.91–0.98</td>
<td>10.20</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age (per yr)</td>
<td>1.08</td>
<td>1.04–1.12</td>
<td>16.47</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Operation delayb</td>
<td>7.98</td>
<td>2.17–29.3</td>
<td>9.79</td>
<td>.002</td>
</tr>
<tr>
<td>Interaction (delay × age)c</td>
<td>0.98</td>
<td>0.96–0.99</td>
<td>7.24</td>
<td>.007</td>
</tr>
<tr>
<td>Marital statusc</td>
<td>0.56</td>
<td>0.39–0.80</td>
<td>10.23</td>
<td>.001</td>
</tr>
</tbody>
</table>

Abbreviation: ASA, American Society of Anesthesiology.

a Fully adjusted model.
b Delay in days, with square root transformation.
c (Single or divorced) versus (all other categories).

Table 4
Logistic regression modela for outcome at 12 months for combined data set

Table 5
Validation and calibration of prognostic models

<table>
<thead>
<tr>
<th>Data used for prediction</th>
<th>Data predicted</th>
<th>c statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>All</td>
<td>0.769</td>
</tr>
<tr>
<td>Set 1</td>
<td>Set 1</td>
<td>0.747</td>
</tr>
<tr>
<td>Set 1</td>
<td>Set 2</td>
<td>0.791</td>
</tr>
<tr>
<td>Set 2</td>
<td>Set 1</td>
<td>0.728</td>
</tr>
<tr>
<td>Set 2</td>
<td>Set 2</td>
<td>0.800</td>
</tr>
<tr>
<td>Both sets</td>
<td>Same sets</td>
<td>0.773</td>
</tr>
<tr>
<td>Both sets</td>
<td>Opposite sets</td>
<td>0.759</td>
</tr>
</tbody>
</table>

The real impact of operative delay is rather harder to determine, and findings in previous studies have been mixed [14,21]. Although we attempted multivariate adjustment in the logistic regression analysis, longer delays were more common among patients with higher ASA scores (data not shown). The real impact of delay is confounded if the sicker patients, with an inherently higher risk of mortality, take longer to prepare for the shorter-term risks of surgery. If this were the entire explanation for our findings, then what we observed as the effect of delay might have been different (i.e., greater) for those dying early in the follow-up period (the sickest of the sick) compared with those dying later. However, the odds ratio for the effect of delay increased when our analysis was restricted to deaths after 30 days (data not shown), and the average delay to surgical intervention for those dying within 30 days was 4.8 days versus 5.7 days for those dying later. These observations are less compatible with a “confounding by intention” explanation for the delay effect.

Our analysis showed evidence of a significant interaction between the effects of delay and age. The degree of moderation of the multiplicative effects of age and delay was 0.98. However, we have excluded terms that are delay dependent from our prognostic score because we considered that practitioners would form a view about the patient’s prognosis on admission, before any elapsed delay. The calibration of the model is still excellent. By the same token, the NNTs in Table 6 are based on the main effect of delay without interaction, and these estimates are therefore more reliable in the middle of the age distribution. Because there are seven main effects in the model, there are 21 possible bi-variate interaction terms. Thus, given the chance of a type I error, observing a significant interaction should be viewed as hypothesis generating rather than hypothesis testing.

Given the magnitude of the possible effect of delay, however, any attempt by a fracture unit to improve efficiency and reduce delays needs to be targeted to those most likely to benefit. The derivation of our prognostic score places the remainder. In the studies cited above, the proportion of their cohorts not receiving surgery varied from 0% to 13.9%. Patients not operated upon represent a high-risk group (in our study, the mortality exceeded 60% at 12 months).

Second, we have gone beyond subjective assessments by using, where possible, validated measures of pre-operative function and by excluding patients for whom surgery was not offered or with unknown dates of injury. Fifteen years ago, orthopedic surgeons from our region derived a simple grading system to guide prognosis after hip fracture, but that was based on the surgeon’s pre-operative subjective judgment of the importance of factors such as the patient’s medical state and social circumstances [16].

Third, we have attempted some validation of our prognostic model and its translation into a more clinically friendly “score.”

The poorer outcome among men has been reported previously [10,12], as has the worse outcome with declining mental function [17,18]. Similarly, the higher mortality among those who, pre-injury, were less able to perform activities of daily living (measured here using the Barthel score) was found in the East Anglian audit [10]. The better survival among the single and divorced (compared with all other categories) runs counter to other studies and will be further explored in future data from this ongoing cohort study. Had this observation been attributable to bias as a consequence of how patients were selected for surgery, we would have expected to see relatively more high-mortality single and divorced subjects among those excluded from surgery. This was not the case (data not shown). Residual confounding remains a possible explanation. When comparing disparate findings in the literature for the effects of any specific variable, it is important to bear in mind how the choice of the statistical method affects the magnitude of the observed effect. Employing a Cox’s proportional hazards function, the method of choice when there is variable follow-up would permit the estimation of (instantaneous) relative risks, and these would tend to be somewhat lower than the equivalent relative odds for a reasonably common outcome (as in our study) that occurs in 20% of the cohort. Logistic regression has frequently been the approach in previous studies [10,15,19,20], probably because the chances of survival or death at 12 months (or a given period) are more readily interpretable than instantaneous hazards.
The number of patients whose delay needs to be shortened to \(<24\) hr to yield an additional survivor (NNT) is illustrated by producing an unconstrained goodness-of-fit test for all patients whose scores rounded to the nearest 10 were in the range 50 to 120. The probabilities from the probit transformed model are applied, and expectations for each of the eight groups formed can be calculated. An approximate chi-squared test is then constructed to test the goodness of fit.

### Table 6

<table>
<thead>
<tr>
<th>Score</th>
<th>Probability of survival if treated within 24 hr (NNT)</th>
<th>With delay 1 to 5 d (NNT)</th>
<th>With delay (&gt;5) days (NNT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.9634</td>
<td>0.9276 (28)</td>
<td>0.8783 (12)</td>
</tr>
<tr>
<td>60</td>
<td>0.9154</td>
<td>0.8512 (16)</td>
<td>0.7734 (7)</td>
</tr>
<tr>
<td>70</td>
<td>0.8311</td>
<td>0.7340 (10)</td>
<td>0.6306 (5)</td>
</tr>
<tr>
<td>80</td>
<td>0.7060</td>
<td>0.5825 (8)</td>
<td>0.4668 (4)</td>
</tr>
<tr>
<td>90</td>
<td>0.5497</td>
<td>0.4175 (8)</td>
<td>0.3085 (4)</td>
</tr>
<tr>
<td>100</td>
<td>0.3853</td>
<td>0.2660 (8)</td>
<td>0.1797 (5)</td>
</tr>
<tr>
<td>110</td>
<td>0.2394</td>
<td>0.1488 (11)</td>
<td>0.0912 (7)</td>
</tr>
<tr>
<td>120</td>
<td>0.1303</td>
<td>0.0724 (17)</td>
<td>0.0401 (11)</td>
</tr>
</tbody>
</table>

### Acknowledgments

We thank the patients who took part and their respective clinicians. We are grateful to Angela Dolan, Martin McAnespie, and Patricia Durkin for supervising the follow-up procedures and data entry. This study was supported by the Eastern, Northern and Southern Health and Social Services Boards; the Regional Multi-Disciplinary Audit Committee; Johnson and Johnson; Howemedica UK Ltd; and the BCH and RGH Trusts.

### Appendix 1

#### Calibration of the scoring method

This is illustrated by producing an unconstrained goodness-of-fit test for all patients whose scores rounded to the nearest 10 were in the range 50 to 120. The probabilities from the probit transformed model are applied, and expectations for each of the eight groups formed can be calculated. An approximate chi-squared test is then constructed to test the goodness of fit.
<table>
<thead>
<tr>
<th>Score (Group)</th>
<th>Dead</th>
<th>Not dead</th>
<th>Expected dead</th>
<th>Expected not dead</th>
<th>Chi square(d)</th>
<th>Chi square (not d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>12</td>
<td>157</td>
<td>8.08</td>
<td>160.92</td>
<td>1.902</td>
<td>0.095</td>
</tr>
<tr>
<td>60</td>
<td>22</td>
<td>219</td>
<td>21.98</td>
<td>219.02</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>70</td>
<td>49</td>
<td>302</td>
<td>54.86</td>
<td>296.14</td>
<td>0.626</td>
<td>0.116</td>
</tr>
<tr>
<td>80</td>
<td>90</td>
<td>270</td>
<td>90.90</td>
<td>269.10</td>
<td>0.009</td>
<td>0.003</td>
</tr>
<tr>
<td>90</td>
<td>96</td>
<td>190</td>
<td>105.65</td>
<td>180.35</td>
<td>0.881</td>
<td>0.516</td>
</tr>
<tr>
<td>100</td>
<td>70</td>
<td>75</td>
<td>72.50</td>
<td>72.50</td>
<td>0.086</td>
<td>0.086</td>
</tr>
<tr>
<td>110</td>
<td>42</td>
<td>21</td>
<td>39.63</td>
<td>23.37</td>
<td>0.142</td>
<td>0.240</td>
</tr>
<tr>
<td>120</td>
<td>9</td>
<td>7</td>
<td>11.96</td>
<td>4.04</td>
<td>0.733</td>
<td>2.169</td>
</tr>
<tr>
<td>Total</td>
<td>390</td>
<td>1241</td>
<td>405.56</td>
<td>1225.44</td>
<td>4.379</td>
<td>3.226</td>
</tr>
</tbody>
</table>

Total 7.605 on 8 degrees of freedom because totals were not constrained

This illustrates that the scoring system appears to be well calibrated

References