CLINICAL ETHICS

# Individual survival time prediction using statistical models

R Henderson, N Keiding

J Med Ethics 2005;31:703–706. doi: 10.1136/jme.2005.012427

Doctors' survival predictions for terminally ill patients have been shown to be inaccurate and there has been an argument for less guesswork and more use of carefully constructed statistical indices. As statisticians, the authors are less confident in the predictive value of statistical models and indices for individual survival times. This paper discusses and illustrates a variety of measures which can be used to summarise predictive information available from a statistical model. The authors argue that models and statistical indices can be useful at the group or population level, but that human survival is so uncertain that even the best statistical analysis cannot provide single-number predictions of real use for individual patients.

Clinicians cannot avoid facing requests from patients and relatives for individual prediction of residual lifetime after the diagnosis of a potentially terminal condition. Christakis and Lamont[1] and Glare *et al*[2] studied the accuracy of clinical predictions of survival (CPS) and found poor agreement with actual survival (AS), with a clear tendency in the optimistic direction: longer predicted than actual life times. In a comment on Christakis and Lamont[1] Parkes (having worked on these matters for over 30 years) argued that the use of carefully developed statistical indices should improve this situation considerably.[2]

The main point of the present contribution is to emphasise that in all realistic scenarios we can imagine, the intrinsic statistical variations in life times are so large that predictions based on statistical models and indices are of little use for individual patients. This applies even when the prognostic model is known to be true and there is no statistical uncertainty in parameter estimation. The inaccuracy of CPS reported by Christakis and Lamont[1] and Glare *et al*[3] is not much worse than that which would be observed if the theoretically *best possible* predictions based on statistical models were to be used instead, at least for the survival patterns with which we have experience.

Although this may be comforting to the clinician faced with unrealistic demands for precision from concerned patients and relatives, there are several contexts where such inherent variability will have consequences that need careful consideration. Among these are how to formulate public health prevention campaigns (where the intervention is necessarily at the individual level); how to handle rigid requirements for limited lifetimes of terminally ill patients in programmes for hospice care or care leave of their relatives; and compensation claim situations where an actually realised residual lifetime after a suboptimal treatment needs to be compared with an individual prediction under optimal treatment.

## ILLUSTRATIVE DATA

To illustrate, we will use data from a study into the accuracy of survival time prediction for patients diagnosed with non-small cell lung cancer, described by Muers *et al*[4] and discussed by Henderson *et al*.[5] We concentrate here on a subset of 272 patients for whom complete information was available on the following risk factors: age, sex, activity score, anorexia, hoarseness, and metastases. Some 17% of patients were still alive at follow up and so gave censored AS, and the remainder all died within 30 months of diagnosis. We used imputation when assessing predictive accuracy for patients with censored lifetimes.

A summary of the effects of the risk factors under a standard Cox proportional hazards model is given in table 1. These results can be used to construct a *prognostic index* (PI), which is a single-number summary of the combined effects of a patient's risk factors and is a common method of describing the risk for an individual. Usually the PI is a linear combination of the risk factors, with the estimated regression coefficients as weights. For a 70 year old male patient with activity score 3, anorexia, hoarseness, but no metastases, the coefficients in table 1 could be combined to give the PI.

Sometimes the coefficients may be simplified and/or PI values scaled for easier interpretation. For the Cox model, after subtracting the median PI the exponentiated prognostic index gives the *relative* risk of each patient in comparison with a baseline ''typical'' patient. For the lung cancer data the median PI is 1.117 and the relative risk for the patient above is 2.53. For the data as a whole, five patients (all with activity score 3 or 4) had relative risks in the range 4–8 and the remainder had values between 0.3 and 4.

The statistical model can also be used to produce a survival curve for each individual patient. Figure 1 shows these for patients classified as being low, median, and high risk, defined as those with the 10%, median, and 90% highest PI values, respectively (the shaded regions in the plot will be discussed later). Overall, the high statistical significance of the risk factors, the wide range of relative risks, and the discrimination shown in figure 1 suggest that the statistical model could have good predictive power. This is examined in the following sections.

## POINT PREDICTIONS

A point prediction is a single valued forecast for survival time. After omission of cases which could not be classified because of censored AS, 49% of clinicians' predictions for the lung cancer data fell into Parkes' definition of ''serious error'', which is prediction either less than half survival time or prediction more than twice survival time. Predictions were optimistic, namely more than twice lifetime, for 32% of patients while 17% of predictions were pessimistic, less than half of lifetime. Although poor, this performance is slightly better than that reported by Christakis and Lamont[1] in a study of predicted residual lifespan of hospice patients, where some 65% of predictions were in Parkes' error category, and there was again a tendency to be too optimistic.

**Abbreviations:** AS, actual survival; CPS, clinical prediction of survival; PI, prognostic index.

**Table 1** Summary of proportional hazards analysis of lung cancer data

| | Coefficient | Standard error | p Value |
|---|---|---|---|
| Age | 0.010 | 0.008 | 0.022 |
| Male | −0.685 | 0.180 | 0.000 |
| Activity (0–4) | 0.346 | 0.082 | 0.000 |
| Anorexia | 0.314 | 0.145 | 0.031 |
| Hoarseness | 0.680 | 0.210 | 0.001 |
| Metastases | 0.417 | 0.220 | 0.059 |

PI = 0.010×70 − 0.685×1 + 0.346×3 + 0.314×1 + 0.680×1 + 0.417×0 = 2.047 (for a 70 year old male patient with activity score 3, anorexia, hoarseness, but no metastases).

Given the poor performance of clinicians in predicting lifetime, we analysed these data with a view to finding a statistical model which would yield objective predictions based on individual risk factors, starting with the standard Cox proportional hazards model, summarised above. Despite highly statistically significant effects of these risk factors, point predictions obtained from the model were also poor: 52% fell into the Parkes' serious error definition. There was less bias however, with roughly equal numbers of optimistic and pessimistic predictions—28% and 24% respectively.

We also considered a variety of alternatives to the Cox proportional hazards model, exploiting the extensive armoury of statistical models now available. In terms of prediction, the best model we could find included clinician's prediction as an additional risk factor and so allowed subjective information in the CPS to be exploited. Details are omitted except to report that 47% of predictions were still in the serious error category.

Parkes' definition of serious error gives a generous range of predicted values deemed to be accurate when compared with AS. Even so, about half of statistical predictions were in error and there was no real improvement on the serious error rate for CPS. In analyses of various other data, not reported here, the error rates for statistical predictions were also typically 50%–60%. This poor performance is no surprise: assuming for the sake of the argument that the statistical model is completely true so that no estimation uncertainty blurs the picture, it can be shown mathematically that the expected *best* serious error rate is usually around 50% for the shapes of survival curves usually seen in practice.[5]

## PREDICTIVE INTERVALS

Predictive *intervals* can be obtained from survival curves, to give for each patient a range of outcomes within which AS will lie with a specified probability, akin to a confidence interval. Interval estimates accurately quantify the uncertainty in prognosis but our experience is that the intervals are often so wide as to be of little practical use.

Table 2 shows 95% and 80% predictive intervals for patients with survival curves which correspond to those in figure 1. There is considerable uncertainty in prediction even for the patient with very high risk and poor prognosis.

## CATEGORY PREDICTIONS

It is also interesting to explore the use of broad categorical predictions such as short, medium, or long term survival. Definition and interpretation of such vague terms will of course depend upon the disease and population characteristics of interest, because what is considered a long survival time with one disease might be relatively short for another.

To illustrate, for the lung cancer data we defined survival times to be short if death occurred within four months, to be long if the patient survived at least a year, and to be medium otherwise. There were 32%, 38%, and 30% AS in the short, medium, and long categories respectively.

**Table 2** Specimen prediction intervals

| Patient | 95% interval | 80% interval |
|---|---|---|
| Low risk | 15 days to >29 months | 65 days to >29 months |
| Median risk | 7 days to >29 months | 34 days to 22 months |
| High risk | 3 days to 15 months | 14 days to 9 months |

Using the standard Cox proportional hazards statistical model, we tried to retrospectively predict survival by choosing for each patient the categorised time interval with the highest probability, obtained from the patient-specific survival curves like those in figure 1. This gave 28%, 39%, and 33% of predictions in each category, which are comparable to the corresponding proportions of AS.

In assessing the accuracy of the predicted categories we gave benefit of doubt when AS was near a boundary by defining fuzzy zones between the groups, where predictions in either neighbouring category could be considered reasonable. For the four month short/medium boundary we took 3–5 months as the fuzzy zone, and for the 12 month medium/long boundary we took 10–14 months. These are the shaded areas in figure 1. If, for instance, AS was 11 months then we considered predictions of either medium or long term survival to be accurate. About 25% of outcomes fell into the fuzzy zones and could contribute to two categories.

Table 3 shows the results. Categorical predictions were accurate under our definition for 56%–67% of cases. The table also gives results for clinical predictions of survival, obtained by choosing as prediction category the interval which included the CPS. Clinician predictions were good for 60%–76% of patients. Overall, the proportion of accurate predictions was 64% for clinicians and 61% for the statistical modelling approach.

Prediction is thus poor for a significant proportion of patients even for these broad categories with fuzzy boundaries. The reason is that for the majority of patients the most likely outcome category is still rarely *very* likely and there is significant probability of AS falling into one of the other groups. Figure 2 shows for each patient the estimated probability of falling into the short, medium, and long

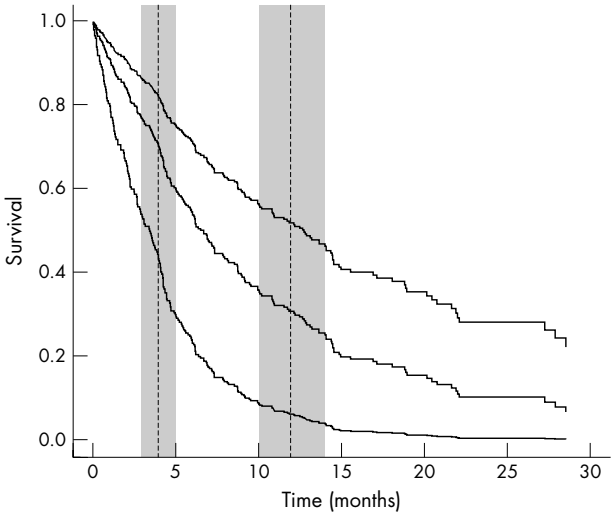


**Figure 1** Lung cancer data. Survival curves for low (10%), median, and high (90%) risk patients. Vertical dashed lines divide the scale into short/medium/long survival times and the shaded regions define fuzzy zones between them.

**Table 3** Outcome distribution for each prediction category

| Prediction | Number | Outcome AS | | |
|---|---|---|---|---|
| | | Short % | Medium % | Long % |
| Statistical model | | | | |
| Short | 76 | 67 | 42 | 11 |
| Medium | 105 | 40 | 61 | 32 |
| Long | 91 | 26 | 41 | 56 |
| Clinician | | | | |
| Short | 76 | 72 | 38 | 11 |
| Medium | 144 | 33 | 60 | 37 |
| Long | 52 | 23 | 35 | 62 |

Note that for each row the sum of the percentages exceeds 100 because some outcomes can be in two categories.

**Table 4** Probability that a low risk patient dies before a high risk patient as a function of relative risk ratio $\theta$

| θ | Probability (low risk dies first) |
|---|---|
| 1 | 50% |
| 1.5 | 40% |
| 2 | 33% |
| 3 | 25% |
| 4 | 20% |

survival time groups as defined here. The most likely category has probability over 0.75 for only five patients while for 76% of patients it is less than 0.5, meaning there is higher chance of being outside the predicted range than inside it.

## RELATIVE RISKS

Our argument is that statistical indices provide poor discriminatory power at the individual level. Another way to illustrate this is to consider two patients, one with low risk and one with high risk. Assume their relative risks differ by a proportionality factor $\theta>1$. Then, the probability that the high risk patient will live longer than the low risk patient can be shown to be $1/(1+\theta)$, or, equivalently the rate ratio $\theta$ is equal to the odds that the high risk patient dies before the low risk patient. Table 4 shows characteristic values of the rate ratio and corresponding probability of the low risk patient outliving the high risk one. To give these values some perspective, for the patients corresponding to figure 1 we have: low PI, relative risk (RR) = 0.56; medium PI, RR = 1.0; high PI, RR = 2.35.

The rate ratio for the very high risk patient in comparison with the very low risk person is $\theta = 2.35/0.56 = 4.2$. Even for this quite extreme example the high risk patient has non-negligible probability of 19% of outliving the low risk one.

## DISCUSSION

Neither clinicians nor statisticians were able to produce reliable point or category estimates of survival for the cancer data. Although we have used just one example to illustrate, we believe that poor predictive accuracy is inherent for realistic survival time patterns. Clinical predictions can be statistically significantly correlated with outcome[3] and statistical models may show highly statistically significant covariate effects but neither in itself guarantees accuracy.

The picture changes when we consider population characteristics, because here of course a carefully constructed statistical model can be extremely valuable in predicting survival probabilities as well as for estimating the effects of treatment or demographic characteristics. Altman and Royston[6] point out however that "the distinction between what is achievable at the group and the individual levels is not well understood". Table 5 attempts to survey the varying roles of the individual and the population viewpoints across several uses of predicting lifetimes. Prognostic indices or palliative scores can be useful in assigning patients to risk groups and from some viewpoints—insurers perhaps—all that is necessary is to know the proportion of each group who will survive any given time. A difference between groups of, say, 10% in one year survival probability can then be hugely important. For the individual patient however, our view is that such a between-group difference is small compared with the variability in residual lifetimes, even between patients with identical characteristics.

What advice then should be given by clinicians faced with a request for information from a potentially terminally ill patient? As argued more generally by Hollnagel[6] it is important to inform patients about individual uncertainty while at the same time conveying population based knowledge and experience. For residual lifetimes this means avoiding use of a single quantity to characterise a probability distribution, whether a point or categorical prediction, prognostic index, relative risk, or probability of surviving a given time. Prediction intervals such as those given in table 2 are often too wide to be of use in forecasting survival time. Another possibility is to give three equiprobable time intervals and paraphrase Hollnagel's technique for communicating information in clear and appropriate language. For the median risk patient of table 2 this would be: "If a group of 90 people like you are followed, research indicates that 30 will die within four months, 30 will die between four and

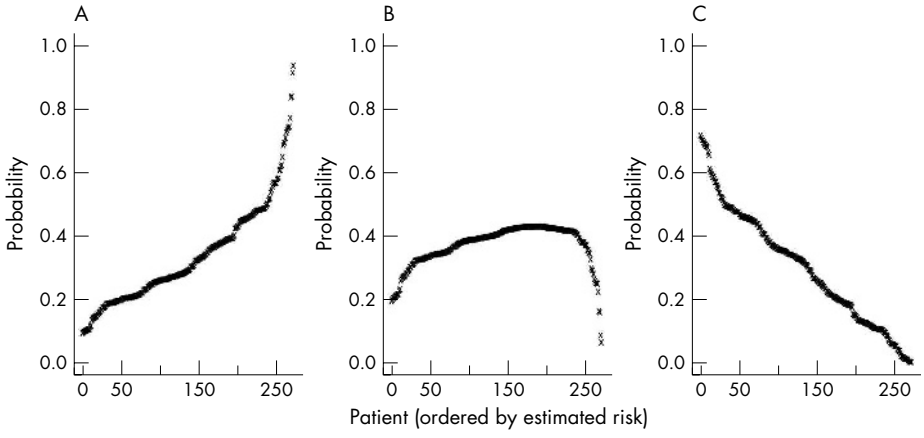


**Figure 2** Estimated probabilities of falling in (A) short, (B) medium, and (C) long categories, ordered by risk.

**Table 5** Individual and population prediction of residual life length

| | Individual | Population |
|---|---|---|
| Life insurance | | ✓ |
| Health economy | | ✓ |
| Hospice, care leave | Visitation | Capacity |
| Public health prevention | Intervention | Effect |
| Individual prevention | ✓ | |
| Clinical: treatment | Patient | Doctor |
| Clinical: counselling | ✓ | |
| Compensation | ✓ | |

11 months, and 30 will live more than 11 months. I do not know which group you will belong to.''

Communicating this information effectively would seem to provide a good compromise between providing the patient with accurate information and avoiding spurious impressions of precision associated with single-number forecasts.

## ACKNOWLEDGEMENTS

We thank Margaret Jones for providing the lung cancer data. An earlier version of this paper has appeared in Danish: Henderson R, Keiding N. Forudsigelse af individuelle levetider ved hjaelp af statistiske modeller. *Ugeskr Laeger* 2005;**167**:1174–7.

## Authors' affiliations

**R Henderson,** Department of Mathematics & Statistics, Newcastle University, Newcastle NE1 7RU, UK
**N Keiding,** Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark; N.Keiding@biostat.ku.dk

Competing interests: the authors have no conflicts of interest to declare.

Correspondence to: Professor N Keiding, Department of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, entr. B, PO Box 2099, DK-1014, Copenhagen, Denmark; N.Keiding@biostat.ku.dk

Received 31 March 2005
Accepted for publication 4 April 2005

## REFERENCES

1 **Christakis NA**, Lamont EB. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *BMJ* 2000;**320**:469–73.
2 **Parkes CM**. Prognoses should be based on proved indicators not intuition. *BMJ* 2000;**320**:473.
3 **Glare P**, Virik K, Jones M, *et al*. A systematic review of physicians' survival predictions in terminally ill cancer patients. *BMJ* 2003;**327**:196–8.
4 **Muers MF**, Shevlin P, Brown J. Prognosis in lung cancer: physicians' opinions compared with outcome and a predictive model. *Thorax* 1996;**51**:894–902.
5 **Henderson R**, Jones M, Stare J. Accuracy of point predictions in survival analysis. *Statist Med* 2001;**20**:3083–96.
6 **Altman DG**, Royston P. What do we mean by validating a prognostic model? *Statist Med* 2000;**19**:453–73.
7 **Hollnagel H**. Explaining risk factors to patients during a general practice consultation: conveying group-based epidemiological knowledge to individual patients. *Scand J Primary Health Care* 1999;**17**:3–5.