Predicting Survival in Cost-Effectiveness Analyses Based on Clinical Trials

Ulf-G Gerdtham*and Niklas Zethraeus†
Stockholm School of Economics

SSE/EFI Working Paper Series in Economics and Finance
No. 442

May 2001

Abstract

This paper deals with the question how to model health effects after the cessation of a randomised controlled trial (RCT). Using clinical trial data on severe congestive heart failure patients we illustrate how survival beyond the cessation of a RCT can be predicted based on parametric survival models. In the analysis we compare the predicted survival and the resulting incremental cost-effectiveness ratio (ICER) of different survival models with the actual survival/ICER. Our main finding is that the results are highly sensitive to the choice of survival model and that extensive sensitivity analysis in the CE analysis is required. We also show that adding the true survival after the end of the clinical study will underestimate the true variability.

Keywords: cost-effectiveness analysis, modelling, confidence intervals.

JEL-classification: I10, I12, I19.

*Centre for Health Economics, Stockholm School of Economics, P.O. Box 6501, SE-113 83 Stockholm. Tel: +46-8-7369283. Fax: +46-8-302115. e-mail: heug@hhs.se.

†Corresponding author: Centre for Health Economics, Stockholm School of Economics, P.O. Box 6501, SE-113 83 Stockholm. Tel: +46-8-7369640. Fax: +46-8-302115. e-mail: henz@hhs.se.

1 Introduction

Several economic evaluations of new pharmaceuticals are being conducted alongside clinical trials where individual patient cost and health effect data are available [1,2]. An advantage of using a controlled clinical trial as the base for the economic evaluation is that the results from the clinical study are of high internal validity showing whether a new therapy has an effect or not. A drawback however is that the clinical study is of relatively low external validity, i.e. it may not reflect the costs and health effects for patients in routine care when the drug is out on the market.

A common and a very important methodological question in economic evaluation studies based on clinical trials is how to predict health effects after the end of the cessation of the clinical trial. E.g. if the economic evaluation is based on a clinical survival study the goal of modelling is to obtain an accurate estimate of the survival gain after the end of the clinical study to be used in the economic evaluation. In general the modelled gain in survival after the end of the clinical trial is much greater than the observed gain stated in the clinical trial [3]. Thus the way of modelling can have major consequences for the cost-effectiveness (CE) of assessed therapies.

In analysing the effect of a new treatment programme in health economic studies, one would ideally like to carry out a study where patients are randomly allocated to "treatment" or "no treatment" alternatives in order to study the causal effect of the programme on survival in clinical practice. Usually such data are not available at the time for the implementation of the new technology. For example, for a newly registred drug often the only available

data comes from a randomized controlled trial. In randomised controlled trials (RCTs), patients are commonly followed-up for a limited amount of time which implies that one must model the survival after the end of the clinical study. Such a modelling can be made either based on information within the clinical study or information external to the clinical study. In Jönsson et al. [1,4] external information is used and patients being alive at the end of follow up is assumed to live according to actuarial data from national statistics. In these studies the assumed expected survival time of 10 years is reached by adding a constant survival time of 10 years for all patients being alive at the end of the clinical trial. To account for that health effects will be discounted, this is fulfilled by assuming that 5% of surviving patients after the cessation of the clinical study dies each year over a period of 20 years [1,4].

Methods for predicting survival based on observed patient data in a clinical trial can be characterised as parametric, semi-parametric and non-parametric. For example, in Jönsson et al. [4] a Weibull model was used in a sensitivity analysis to predict the expected conditional survival while Raikou et al. [2] used a simulation model. However there are a number of models that can be used for the purpose of predicting survial and *a priori* there is no reason to prefer one model to the other. To evaluate the accuracy of the model predictions one would ideally compare the predictions with the actual survival. However by definition this is not possible since the reason for modelling survival is the lack of survival data beyond the cessation of the clinical trial.

When individual patient cost and health effect data become available, this opens up for the possibility of analysing uncertainty due to sampling

variability. The ICER has been the main focus of interest recently in the health economic literature and several methods have been presented dealing with the problem of computing confidence interval for the ICER. It has been shown that the Fieller's and non-parametric bootstrap methods represent most accurately the variablity of the ICER estimator [5]. An alternative presentation of the CE results is provided by the CE acceptability curve. The CE acceptability curve produces the minimum significance level at which the new therapy can be said to be cost-effective for different marginal willingness to pay. When observed data in a clinical study is combined with modelling after the cessation of the clinical trial it is unclear whether modelling satisfactory represents the actual variability. There is an obvious risk that statistical inference based on such combined data will generate a false picture of the 'true' variability and yield incorrect statistical information. For example if survival after the end of the clinical trial is modelled by adding a constant survival time, there is a risk that the actual variability in health effects are underestimated.

The aim of this paper is to illustrate how survival beyond the cessation of a clinical trial can be predicted based on different parametric survival models. In the analysis we discriminate between the different suvival models by the commonly used Likelihood Ratio (LR) test. We also compare the predicted survival/ICER based on the different models with the actual survival/ICER. All the predictions are based on observed patient information contained within a clinical trial. The aim is further to investigate whether "pseudo" P-values and confidence intervals obtained from the combination of individual patient data and econometric modelling represents the "true"

P-values and confidence intervals.

2 Methods

Assume that a CE study is carried out alongside a clinical study where a new therapy is compared with an existing standard therapy. Costs and health effects (measured in survival) are collected from a clinical trial with a mean follow up time equal to X. After the end of the clinical study it is assumed that all patients convert to the new therapy. The new therapy is thus assessed against the old therapy for a mean follow up equal to X. The clinical trial showed a positive and significant effect on mortality, which means that there will be a gain in life years within the follow up time of the clinical trial. At the end of the clinical trial there are more patients alive in the active treatment group compared with placebo and the question is whether there will be further gains in survival after the cessation of the clinical trial when all patients convert to the active treatment. How can information within the clinical trial be used to predict the expected remaining survival for patients being alive at the end of the follow up? This question is investigated by using data according to below.

2.1 Data

The data are based on the Consensus trial [6] and a follow up study [7]. The initial clinical study [6] compared enalapril with standard therapy in the treatment of congestive heart failure patients. The mean follow up time in

the clinical study was 0.515 years and after study completion all patients were offered enalapril therapy [7]. In the placebo group 67% began with enalapril, while 88% continued with enalapril in the enalapril group. Individual patient data on life years were available in the clinical study and whether they were alive or not at the end of follow up. At the end of the clinical study 77 (out of 127) patients were alive in the enalapril group while 58 (out of 126) were alive in the placebo group. No comprehensive cost data were collected in the clinical study. To be able to calculate ICERs individual cost data are randomly selected from a study comparing bisoprolol with standard therapy in heart failure [3]. Individual cost and health effect data are thus available within the follow up time in the clinical study. The health economic question is whether it is good value for money to implement the enalapril therapy (treatment 1) added to the standard therapy instead of using the standard therapy (treatment 0).

In this paper different modelling assumptions are made for patients being alive at the end of the follow up. The outcome of the modelling is then compared with the actual outcome presented in a 10-year follow up study [7]. At the 10 year follow up 5 patients were still alive in the enalapril treatment group while 1 patient was lost to follow up in the placebo group. In our calculations we assume that the follow up study contains information of the total survival after the end of the clinical study and that patients still alive at the 10 year follow up dies immediately after that point in time. The conditional mean survival time for patients alive at the end of follow up in the clinical study were 941 and 774 days in the enalapril and placebo group respectively. This difference can partly be explained by the fact that a higher

share of the surviving patients in the enalapril group used enalapril also after the cessation of the clinical trial.

2.2 Modelling

Our data on survival time has some characteristics that are important in selecting an estimation method. One characteristic is that the survival time distribution usually is skewed in some way, which violates the ordinary least squares assumption of normally distributed error terms. The survival time is also by definition positive, while a normally distributed variable can take both positive and negative values. Another characteristic is that a certain proportion of individuals have not reached the end-point of interest, i.e. some individuals are still alive at the cessation of the clinical study, which means that such individuals are right censored. This calls for the application of duration data models that incorporate the above characteristics [8]. The random variable T is assumed to have a density function $f(t)$ reflecting the probability of survival time having length t, and a distribution function $F(t) = \int_0^t f(s)ds = \Pr(T \leq t)$ which defines the survival function $S(t) = \Pr(T \geq t) = 1 - F(t)$. The survival function shows the probability that the individual survives for at least t periods. From the survival function one can define the hazard function $\lambda(t) = f(t)/S(t)$, which shows the mortality rate at time t conditional on surviving to time t (see Kiefer [9] and Lancaster [10] for surveys of duration models).

In the results section we estimate four common parametric survival models, i.e. exponential, Weibull, lognormal and generalized gamma models [8].

These models are distributions for a non-negative random variable, with hazard functions that display different behaviours; for example, the hazard function for the exponential distribution is constant while the hazards for the Weibull distribution are monotonically increasing or decreasing depending on the shape parameter p [8]. The hazard function of the generalized gamma model is very flexible allowing for a large number of shapes and the exponential, Weibull, and lognormal models are special cases of the generalized gamma model. Thus these models can be tested as null hypotheses against the alternative generalized gamma model by use of a LR $\chi^2 \sim$ test.

2.3 Methods for Assessing Uncertainty

2.3.1 A 95% upper bound confidence limit for the ICER

To compute an upper bound confidence limit for the ICER a non-parametric bootstrap method and Fieller's method are used.

Non-parametric bootstrap approach The bootstrap procedure is based on the following steps [5,11]:

1. Given a chosen model with predicted conditional expected survival for patients being alive at the end of the clinical trial, resample (with replacement) n_0 cost and effect pairs from the control sample. Also, resample (with replacement) n_1 cost and effect pairs from the new treatment sample. Calculate the bootstrap cost and effect averages estimate $\widehat{\mu}_{C_0}^{*b} = \overline{C}_0^{*b}$, $\widehat{\mu}_{E_0}^{*b} = \overline{E}_0^{*b}$, $\widehat{\mu}_{C_1}^{*b} = \overline{C}_1^{*b}$ and $\widehat{\mu}_{E_1}^{*b} = \overline{E}_1^{*b}$, respectively.

2. Calculate the bootstrap ICER estimate $\widehat{icer}^{*b} = \widehat{\mu}^{*b}_{\Delta C}/\widehat{\mu}^{*b}_{\Delta E}$

3. Repeat step (1) and (2) B times. Following the recommendations by Efron and Tibshirani [14], the number of bootstrap replications is set to $B = 1000$ in the empirical application. This size of B is recommended in order to make the variability of the boundaries of the confidence intervals constructed from the bootstrap "acceptably" low.

The B bootstrap replicates of the $ICER$ statistic $\widehat{icer}^{*b}$, $b = 1, ..., B$, are then arranged in increasing order. 5% of the bootstrap replicates in the upper tail is then cut away. The remaining upper value is defined as the 95% upper bound bootstrap percentile confidence limit for the ICER.

Fieller's approach As shown by, for example, Briggs and Fenn [5], the Fieller's confidence limits for $ICER = R$ are found by solving the second order equation:

$$\frac{\widehat{\mu}^2_{\Delta C} + R^2\widehat{\mu}^2_{\Delta E} - 2R\widehat{\mu}^2_{\Delta E}\widehat{\mu}_{\Delta C}}{R^2\widehat{Var}\left(\widehat{\mu}_{\Delta E}\right) + \widehat{Var}\left(\widehat{\mu}_{\Delta C}\right) - 2R\widehat{Cov}\left(\widehat{\mu}_{\Delta E}, \widehat{\mu}_{\Delta C}\right)} = z^2_{a/2}, \tag{1}$$

where z_β denotes the β quantile from the standard normal density function, defined by $\Phi\left(z_\beta\right) = \beta$. For $\beta = 0.9$ the upper limit is a 95% upper bound Fieller's confidence limit for the ICER.

2.3.2 A CE acceptability curve

An informative presentation of the CE results is provided by the acceptability curve (see e.g. Gray et al. [12]). CE acceptability curves can be defined in two

completely equivalent ways, either in terms of the ICER estimator (as first proposed by van Hout et al. [13]) or the NB estimator, as discussed by Briggs and Fenn [5] and formally shown by Löthgren and Zethraeus [14]. A monetary net benefit (NB) measure is defined as $NB(\lambda) = \lambda\mu_{\Delta E} - \mu_{\Delta C}$, where λ is the maximum price society is willing to pay for one more unit of health effects [15]. Based on the net benefit the decision rule is that the new treatment should replace the control treatment if $NB(\lambda) > 0$. The CE acceptability curve is given by $CE_{acc}(\lambda) = \Phi\left(NB(\lambda)/\sigma_{\widehat{NB}(\lambda)}\right)$ and can be estimated by $\widehat{CE}_{acc}(\lambda) = \Phi\left(\widehat{nb}(\lambda)/\widehat{\sigma}_{\widehat{NB}(\lambda)}\right)$ (using the sample estimates). One minus the acceptability estimate corresponds to the minimum significance level at which the null hypothesis (the new treatment is not cost-effective) can be rejected. Thus, a CE acceptability curve is simply the mirror-image of a P-value curve.

3 Results

Table 1 shows the actual and expected survival (discounted and undiscounted) and costs with and without treatment based on different model alternatives. According to the LR $\chi^2 \sim$ test the Weibull and exponential model are rejected against the gamma model at the 1% level of significance but not the lognormal model. However, the model that best predicts the actual survival (discounted or undiscounted) in the enalapril and placebo group is the Weibull model. The exponential model systematically underestimates the survival while the gamma and log normal models overestimates the survival. The model that best predicts the difference in survival is the lognormal and

gamma models. The reason for this is that the lognormal and gamma models overestimate the survival in the treatment and placebo group to about the same extent. The Weibull model on the other hand overestimates the survival in the placebo group and underestimates the survival in the enalapril group. Thus when comparing the ICERs (discounted or undiscounted) derived from the models with the 'true' one the lognormal and gamma models are the most accurate ones (Table 2).

Note that we assume equal expected survival in the two treatment groups in our modelling exercises. However, the actual observed data shows that the conditional expected survival time is higher in the enalapril group (941) compared with the placebo group (774). Thus even if we predicted the correct conditional expected survival in the enalapril group and used that for the placebo group the true difference in survival would be understated. Models that 'correctly' overestimates the conditional expected survival in the enalapril group gives survival differences equal to the true one.

In Figure 1a and 1b the survival curves based on actual data and models are presented for the enalapril and placebo groups. Note that the survival curve based on the observed data is above the modelled survival curves in the beginning of the prediction period (for the first 1500 days) and therafter comes close the modelled survival curves.

To account for the variability in the patient cost and health effect data confidence interval can be computed. In Table 2 the upper 95% confidence limit based on a non-parametric bootstrap percentile method and the Fieller's methods are presented. In all the modelling alternatives a constant survival is added to all the patients being alive at the cessation of the clinical study.

In the modelled (true) 941 and 774 days were added in the therapy and placebo group respectively. This corresponds to the actual expected survival in the two treatment groups observed in the follow up study [7]. In all the other model alternatives (gamma, lognormal, weibull, and exponential) the conditional expeced survival is obtained based on the specified model as described above. These models predict the conditional expected survival based on the patient data in the enalapril group, which is used for all surviving patients at the end of the clinical trial. In the gamma model the predicted conditional expected survival was 1669 days. The corresponding figures for the lognormal, Weibull and exponential models are 1642 days, 901 days, and 546 days, respectively.

Ceteris paribus adding the true conditional expected survival for patients being alive in the two treatment groups (Modelled (true)) will underestimate the true variability, which is reflected by a lower upper confidence limit (47 081) compared to the 'true' one of 64 881. The other model alternatives either over or underestimates the 'true' 95% confidence limit. The gamma and lognormal models produce upper limits that slightly underestimate the 'true' one while the Weibull and exponential models overestimates the true 95% upper confidence limit.

These results are confirmed in Figure 2 that shows CE acceptability curves based on the actual data and based on different modelling alternatives. Ceteris paribus adding the true conditional expected survival for patients being alive in the two treatment groups underestimates the true variability which is reflected by the CE acceptability curve (Modelled (true)) always being above the true acceptability curve. The interpretation is that the new intervention

is significanly cost-effective at a lower level compared to the actual true case. The other model alternatives either over or underestimates the true CE acceptability curve. The gamma and lognormal models are just over the true acceptability curve while the Weibull and exponential models underestimates the true actual curve.

Based on the true data for all prices above 60 000 the null hypothesis that the new intervention is not cost-effective can be rejected at the 5% significance level. The price based on the Modelled (true) is 47 000. The corresponding price based on the gamma and lognormal models are close to the true value and are estimated at 59 000 and 60 000 respectively. The corresponding price based on the Weibull and exponential models are 87 000 and 114 000 respectively. Alternatively given a price of 100 000 the true significance level is 0.02 which can be compared with Modelled (true) of 0.00. The gamma and lognormal models produce the same significance level of 0.02 while the Weibull model ($P = 0.04$) and the exponential model ($P = 0.07$) overestimates the significance levels.

4 Summary and Conclusion

This paper investigates the question of predicting survival in CE studies based on the information contained in a RCT with a given follow up. In the paper we predict survival using different parametric survival models (generalized gamma, lognormal, Weibull, exponential), and which are tested against each other by use of the likelihood ratio test. Furthermore the model predictions are compared with the observed true survival.

The Weibull model predicts the actual survival in the enalapril group most accurately. However this was not confirmed by the statistical tests. The used likelihood ratio test rejected the Weibull model (and the exponential) but not the lognormal model against the gamma model. Thus for this particular data the gamma and lognormal models would be selected based on the statistical tests but the Weibull model generated the most accurate survival predictions. The observed difference in survival between the two therapies are most accurately predicted by the log-normal and gamma models. The reason for this is that these models overstate the conditional survival in the enalapril group that reduces the underestimation of the actual difference in survival that results from the Weibull model. A model that exactly predicts the conditional actual survival in the enalapril group will underestimate the true difference in survival.

Our conclusion is that statistical tests discriminating between models used to predict survival out of sample should be complemented by an extensive sensitivity analysis since it is not obvious that the model that performed best in the statistical tests generates the best survival predictions.

Ceteris paribus adding a 'true' conditional expected survival for all patients being alive at the end of a clinical trial underestimates the true variability which is reflected by a lower upper confidence limits compared to the 'true' one. This will also produce CE acceptability curves that underestimates the minimum significance level at which the null hypothesis that a new therapy is not cost-effective can be rejected. Also the analysis of uncertainy requires extensive sensitivity analysis with respect to the chosen model.

To assess our model predictions we ideally need health effect data from an extended RCT, i.e. a study that continues also after the cessation of the initial RCT and where all patients are using enalapril in a controlled setting. Our data are based on an open setting where the two patient groups are offerred the new therapy. In what way does our data differ from the ideal data? The data in this study reflect actual clinical practice (effectiveness) and not a controlled situation (efficacy). It is a lower share of patients in the initial placebo group that after the cessation of the clinical trial begins with enalapril compared with the share of patients that continues with enalapril in the enalapril group. In a controlled setting no difference in the share of patients on the active therapy is expected. However, patients in this study are in severe disease states and are probably in frequent contact with the health care, which means that the clinical practice is rather close to the controlled enviroment.

Instead of using a RCT as a base for the economic evaluation a controlled trial conducted under more naturalistic circumstances can be used as a base for the health economic evaluation. Such a study would be characterised with both a high internal and external validity. However at the time for the registration of a new chemical entity the 'only' available information is usually the one offerred by the RCT. The appropriateness of using a RCT should be assessed and discussed and can vary depending on the patient group under study.

Acknowledgements

We are grateful to John Kjekshus for providing us with data from the consensus study and from a 10-year follow up. We are also grateful to Magnus Johannesson for comments on a previous version of the paper.

References

1. Jönsson B, Cook JR, Pederswen TR. The cost-effectiveness of lipid lowering in patients with diabetes: results from the 4S trial. *Diabetologia* 1999;42:1293-1301.

2. Raikou K, Gray A, Briggs A, Stevens R, Cull C, McGuire A, Fenn P, Stratton I, Holman R, Turner R, for the UK prospective diabetes study group, Cost effectiveness analysis of improved bloodpressure control in hypertensive patients with type 2 diabetes: UKPDS 40. *British Medical Journal* 1998;317:720-726.

3. Ekman M, Zethraeus N, Jönsson B. The cost-effectiveness of bisoprolol in the treatment of chronic congestive heart failure: an analysis from CIBIS-II trial data. Forthcoming in *Pharmacoeconomics*.

4. Jönsson et al. Cost-effectiveness of cholesterol lowering. Results from the Scandinavian Simvastatin Survival Study (4S). *European Heart Journal* 1996;17:1001-1007.

5. Briggs A and Fenn P. Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane. *Health Economics* 1998;7:723-40.

6. The Consensus trial study group. Effects of enalapril on mortality in severe congestive heart failure. Results of the cooperative north scandinavian enalapril survival study (consensus). *The New England Journal of Medicine* 1987;316:1429-1435.

7. Swedberg K, Kjekshus J, Snapinn S, for the consensus investigators. Long term survival in severe heart failure patients treated with enalapril. *European Heart Journal* 1999;20:136-139.

8. Greene WH. (2000). Econometric Analysis, 4th Ed, Prentice Hall International, Inc, New York.

9. Kiefer NM. Economic Duration Data and Hazard Functions. *Journal of Economic Literature* 1988;26:646-679.

10. Lancaster, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.

11. Efron B, and Tibshirani RJ. (1993). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability, No. 57, Chapman and Hall, New York.

12. Gray A, Raikou M, McGuire A, Fenn P, Stevens R, et al. Cost-effectiveness of an intensive blood glucose control policy in patients with type 2 diabetes: economic analysis alongside randomised controlled trial (UKPDS 41). *British Medical Journal* 2000;320:1373-1378.

13. van Hout BA, Al MJ, Gordon GS, Rutten FFH. Costs, effects and C/E-ratios alongside a clinical trial. *Health Economics* 1994;3:309-319.

14. Löthgren M, Zethraeus N. Definition, interpretation and calculation of cost-effectivness acceptability curves. *Health Economics* 2000;9:623-630.

15. Tambour M, Zethraeus N. and Johannesson M. A Note on Confidence Intervals in Cost-Effectiveness Analysis. *International Journal of Technology Assessment in Health Care* 1998;14:467-471.