A Measure of Explained Variation for Event History Data

Janez Stare¹*, Maja Pohar Perme¹ and Robin Henderson²

1. Department of Biostatistics and Medical Informatics, University of Ljubljana,
   Vrazov trg 2, SI-1000 Ljubljana, Slovenia

2. Mathematics & Statistics, Newcastle University, UK
   * email: janez.stare@mf.uni-lj.si

SUMMARY: There is no shortage of proposed measures of prognostic value of survival models in the statistical literature. They come under different names, including explained variation, correlation, explained randomness and information gain, but their goal is common: to define something analogous to the coefficient of determination $R^2$ in linear regression. None however have been uniformly accepted, none have been extended to general event history data, including recurrent events, and many cannot incorporate time-varying effects or covariates. We present here a measure specifically tailored for use with general dynamic event history regression models. The measure is applicable and interpretable in discrete or continuous time, with tied data or otherwise, with time-varying, time-fixed or dynamic covariates, with time-varying or time-constant effects, with single or multiple event times, with parametric or semi-parametric models, and under general independent censoring/observation. For single-event survival data with neither censoring nor time-dependency it reduces to the concordance index. We give expressions for its population value and the variance of the estimator and explore its use in simulations and applications.

KEY WORDS: C-index; dynamic models; explained variation; rank correlation; recurrent events.
1. Introduction

Despite some criticism, mostly unwarranted, of the coefficient of determination $R^2$ in linear regression, the search for an analogous measure in survival analysis has been pursued for almost 30 years (eg Harrell et al 1982, Korn and Simon 1990, Schenper and Stare 1996, Rosthøj and Keiding 2004). Many indices have been proposed yet none has been uniformly accepted. Only a few are incorporated in statistical packages, sometimes with simplicity of calculation being a selection criterion placed ahead of statistical properties.

In this paper we propose a new measure of the variation explained by a proposed model for event time data. The phrase “variation explained” is understood to mean any measure which quantifies the gain in prognostic power obtained by using a statistical model which differentiates between study subjects. Our justification for adding a new entry to the list of existing measures is our wish to satisfy the following criteria.

(1) Applicability to general event history data, not only survival.

(2) Facility to incorporate time-varying and/or dynamic covariates (Fosen et al 2006) and/or time-dependent effects.

(3) Model-free interpretation on a well-understood scale, to allow comparison between non-nested models.

(4) Applicability to both parametric and semi-parametric models.

(5) Consistency under general independent censoring mechanisms, including intermittent missingness and delayed entry or truncation.

Together these criteria preclude all existing indices. For example, Brier-score and ROC-related measures (eg Graf et al 1999, Schenper and Henderson 2000, Heagerty et al 2000, Heagerty and Zheng 2005) work well for survival data but do not extend to general event-history. Others, particularly likelihood or information-based (eg Kent 1983, Kent and O’Quigley 1988, Nagelkerke 1991, O’Quigley et al 2005 or Royston 2006) are not transportable be-
tween models. Still others are defined specifically only for proportional hazards models (e.g. O’Quigley and Flandre 1994, Royston and Sauerbrei, 2004, Gönen and Heller 2005).

Our proposal is motivated by two earlier papers. The first, Harrell et al (1982), introduced the C-index based on concordance between pairs of subjects. A pair is concordant if the observed failure order corresponds to the order of estimated hazards determined at time zero. The index is the proportion of pairs which are concordant out of all pairs where a decision can be made. The second motivating paper, Schemper and Kaider (1997), extended usage of the Spearman correlation to censored data through an argument based on ordering of observations and model-based predictions. Censoring is particularly problematic for the C-index and neither of the suggestions can either extend beyond single event survival or incorporate time-dependent covariates or effects. Our solution is to rank model-based intensity estimates at each distinct event time and, with careful treatment of missing or censored data, summarise the extent to which the model matches the data.

In the following short section we give an informal motivation for our suggestion. The proposal is stated formally and for the general case in Section 3. In Section 4 we consider the sampling variance and large sample value of the measure. Simulations to investigate small sample performance and the effect of missing data are described in Section 5. Three distinct types of application are presented in Section 6 before our closing discussion in Section 7.

2. Informal motivation and description

Suppose we have single-event survival data on \( n \) individuals with no censoring and no tied observations. We incorporate covariate information into a model which allows us to estimate for each individual the hazard at any time. Now consider the event time \( t_i \) for person \( i \), and assume there are \( n_i \) subjects still at risk at that time. If we know only that there is one event from amongst this group, then our point prediction for who experiences the event would be the individual with the highest estimated hazard. Or, if we only use the hazards
to rank the subjects $1, 2, \ldots, n$ from highest to lowest risk, our choice would be the person with rank one. The further from one the observed rank $r_{i,\text{model}}$ for person $i$, the worse is our prediction. Under a null model without covariates or exogenous information, subjects are not distinguished and all ranks are the same and equal to $r_{i,\text{null}} = (n_i + 1)/2$. By contrast, under perfect information the person who had rank $r_{i,\text{perfect}} = 1$ would experience the event. The difference $r_{i,\text{null}} - r_{i,\text{perfect}}$, the discrepancy between the null and the perfect model, is what needs to be explained. The difference $r_{i,\text{null}} - r_{i,\text{model}}$ is what the model indeed explains. Summing over individuals gives us

$$R_E = \frac{\sum_i (r_{i,\text{null}} - r_{i,\text{model}})}{\sum_i (r_{i,\text{null}} - r_{i,\text{perfect}})}$$

where the notation $R_E$ is meant to allude to ranks explained. Note that our null model, which assigns equal ranks to all subjects in each risk set, is defined independently of the model being considered for the data. This makes it possible to compare the ranks explained by non-nested models, which is often a problem with other suggested measures.

3. Explained risk ranking

We now consider the general case. Let $\tilde{N}_i(t)$ be an underlying event counting process for subject $i = 1, 2, \ldots, n$ and let $N_i(t)$ be the associated observed counting process. Define $\mathcal{F}_{t-}$ to be the history of events, censoring, and (possibly time-varying) covariates prior to time $t$. The intensity for $N_i(t)$ is

$$\lambda_i(t) = \lambda_i(t|\mathcal{F}_{t-}) = Y_i(t)\alpha_i(t|\mathcal{F}_{t-}),$$

where $Y_i(t)$ is a predictable observation indicator, taking the value one when the individual is under observation, and zero otherwise. Note that this is not the usual at-risk indicator for right-censored survival data, which is defined to fall to zero on occurrence of an event. In our set-up it is the intensity $\alpha_i(t|\mathcal{F}_{t-})$ which is zero whenever jumps in $\tilde{N}_i(t)$ are impossible, in survival data for instance if $\tilde{N}_i(t-) = 1$, and we use $Y_i(t)$ only for observation. Let
$G_i(t|\mathcal{F}_{t-}) = P(Y_i(t) = 1|\mathcal{F}_{t-}) > 0$ for $t \leq \tau$, the maximum follow-up time. Let $\hat{G}_i(t|\mathcal{F}_{t-})$ be an associated estimator. Throughout we assume independent censoring as defined by Aalen et al (2008, Section 2.2.8), which is essentially that conditional upon $\mathcal{F}_{t-}$ there is no further information in knowing whether or not a subject is under observation at $t$.

Define $\mathcal{R}_t$ to be the risk set at time $t$, consisting of all individuals with positive observation intensity $\lambda_i(t)$. As summarised in Section 1, our proposal is based on intensity ranking in risk sets at event times. Specifically, we consider ranks under an assumed model ($M$) for $\alpha_i(t|\mathcal{F}_{t-})$, under a null model (0) allocating common intensity for all, and under a perfect model ($P$) in which subjects with events always have highest estimated intensities. We use weighting by inverse probability of observation to deal with incomplete data in two ways: once when calculating the ranks within risk sets and once when averaging over time. For brevity we write $\hat{\alpha}_it = \hat{\alpha}_i(t|\mathcal{F}_{t-})$ and $\hat{G}_it = \hat{G}_i(t|\mathcal{F}_{t-})$ and we adopt the reasonable assumption that $\hat{G}_it > 0$ for $i \in \mathcal{R}_t$. Then under $M$ the adjusted rank in $\mathcal{R}_t$ of subject $i$ is

$$r_i^{(M)}(t) = \sum_{j \in \mathcal{R}_t} \frac{I(\hat{\alpha}_{jt} > \hat{\alpha}_{it})}{\hat{G}_{jt}} + \frac{1}{2} \left( 1 + \sum_{j \in \mathcal{R}_t} \frac{I(\hat{\alpha}_{jt} = \hat{\alpha}_{it})}{\hat{G}_{jt}} \right).$$

Similarly the adjusted ranks under the null and perfect models are

$$r_i^{(0)}(t) = \frac{1}{2} \left( 1 + \sum_{j \in \mathcal{R}_t} \frac{1}{\hat{G}_{jt}} \right)$$

and

$$r_i^{(P)}(t) = \frac{1}{2} \left( 1 + \sum_{j \in \mathcal{R}_t} \frac{\Delta N_j(t)}{\hat{G}_{jt}} \right)$$

respectively, where $\Delta N_j(t) = 1$ if $N_j$ jumps at $t$. To see why the weighting is used, suppose for illustration that the subject with highest $\hat{\alpha}_it$ is the only one to have the event at $t$, and that for this person $\hat{G}_it = 0.5$. Then without weighting, the naive rank would be one. With weighting, the rank is 1.5 to reflect our expectation that under complete observation there
would be one other subject with the same underlying intensity. This argument can of course be formalised to show that under independent censoring we obtain consistent estimators of properties of the fully observed population. If there is no censoring ($\hat{G}_{it} = 1$ for all $i$) then (3)–(5) reduce to the ranked intensity estimates under the assumed, null and perfect models as outlined in the previous section but with proper allowance for ties. If censoring does not depend on covariates or previous events ($\hat{G}_{it} = \hat{G}_t$ for all $i$) then similar interpretations are obtained based on an effective sample size of $\sum_{j \in \mathcal{A}_i}(1/\hat{G}_t)$.

Our proposal for an overall measure is

$$R_E = \frac{\sum_i \int_0^\tau \hat{w}_i(t) \left( r_i^{(0)}(t) - r_i^{(M)}(t) \right) dN_i(t)}{\sum_i \int_0^\tau \hat{w}_i(t) \left( r_i^{(0)}(t) - r_i^{(P)}(t) \right) dN_i(t)}$$

where $\hat{w}(t) = 1/\hat{G}_{it}$. Here the weighting is used to compensate for missing event times, with more weight to individuals and time periods with high probability of having unobserved events. If required we can integrate over a sub-interval $(t_1, t_2)$ instead of the full follow-up period $(0, \tau)$ and so obtain a local measure of agreement between model and observed risk ranks. The interpretation as “explained variation” follows from the simple fact that

$$\left( r_i^{(0)}(t) - r_i^{(P)}(t) \right) = \left( r_i^{(0)}(t) - r_i^{(M)}(t) \right) + \left( r_i^{(M)}(t) - r_i^{(P)}(t) \right).$$

Thus the total variation at $t$ (left side) is partitioned into an explained part (first term on right) and an unexplained part (second term on right). We can also rearrange to write $R_E$ in the alternative form

$$R_E = 1 - \frac{\sum_i \int_0^\tau \hat{w}_i(t) \left( r_i^{(M)}(t) - r_i^{(P)}(t) \right) dN_i(t)}{\sum_i \int_0^\tau \hat{w}_i(t) \left( r_i^{(0)}(t) - r_i^{(P)}(t) \right) dN_i(t)}.$$

In the special case of single-event uncensored survival data with time-fixed covariates and effects it is possible to show that $R_E$ becomes equivalent to the C-index introduced by Harrell et al (1982). More generally the measure is applicable and interpretable in discrete or continuous time, with tied data or otherwise, with time-varying, time-fixed or dynamic co-
biometrics, with time-varying or time-constant effects, with single or multiple event times, with
parametric or semi-parametric models and under general independent censoring/observation.

If the model discriminates perfectly between individuals then $R_E = +1$; if there is no
discrimination then $R_E = 0$; and if the model-based ranks are fully discordant with the
outcomes in all risk sets then $R_E = -1$. We will see an example of negative local $R_E$ in
Section 6. As an aside we note that for grouped data in which individuals share covariates,
no model can fully discriminate between subjects and any measure of "explained variation"
should not reach the maximum value achievable for the ungrouped case. Appendix B shows
for instance that if covariates group subjects into $k$ groups, with group $j$ including proportion
$\theta_j$ of the sample ($j = 1, 2, \ldots, k; \sum_j \theta_j = 1$), then for uncensored survival data without ties
$R_E$ is bounded above by

$$\frac{n}{n-1} \left(1 - \sum_{j=1}^k \theta_j^2 \right). \tag{7}$$

4. Properties

For this section we will assume events occur in continuous time with no ties.

4.1 Variance

We will write $R_E = Q_1(\tau)/Q_2(\tau)$ where

$$Q_1(t) = \frac{1}{n} \sum_i \int_0^t \hat{w}_i(u) \left( r_i^{(0)}(u) - r_i^{(M)}(u) \right) dN_i(u), \tag{8}$$

and

$$Q_2(t) = \frac{1}{n} \sum_i \int_0^t \hat{w}_i(u) \left( r_i^{(0)}(u) - r_i^{(P)}(u) \right) dN_i(u). \tag{9}$$

Given the usual assumption that $N_i(t)$ is adapted to the history $\mathcal{F}_t$, the ranks of all
subjects in the risk set are predictable. Further, from the Doob-Meyer decomposition a
martingale process can be associated with $N_i(t)$. The predictable variances of $Q_1$ and $Q_2$ are
therefore (Aalen et al, 2008, p56)

\[
< Q_1 > (t) = \frac{1}{n^2} \int_0^t \sum_i \hat{w}_i^2(u) \left( r_i^{(0)}(u) - r_i^{(M)}(u) \right)^2 \lambda_i(u| \mathcal{F}_{u-}) du,
\]

and

\[
< Q_2 > (t) = \frac{1}{n^2} \int_0^t \sum_i \hat{w}_i^2(u) \left( r_i^{(0)}(u) - r_i^{(P)}(u) \right)^2 \lambda_i(u| \mathcal{F}_{u-}) du,
\]

and the covariance is

\[
< Q_1, Q_2 > (t) = \frac{1}{n^2} \int_0^t \sum_i \hat{w}_i^2(u) \left( r_i^{(0)}(u) - r_i^{(M)}(u) \right) \left( r_i^{(0)}(u) - r_i^{(P)}(u) \right) \lambda_i(u| \mathcal{F}_{u-}) du.
\]

We are interested in \( \text{var}(Q_1/Q_2) \) and will use the delta method approximation for the variance of any function \( f(Q_1, Q_2) \):

\[
\text{var}(f(Q_1, Q_2)) \approx \text{var}(Q_1) \left( \frac{\partial f}{\partial Q_1} \right)^2 + 2 \text{cov}(Q_1, Q_2) \left( \frac{\partial f}{\partial Q_1} \right) \left( \frac{\partial f}{\partial Q_2} \right) + \text{var}(Q_2) \left( \frac{\partial f}{\partial Q_2} \right)^2.
\]

When \( f(Q_1, Q_2) = Q_1/Q_2 \) we have

\[
\text{var} R_E \approx \frac{\text{var}(Q_1)}{Q_2^2} - \frac{2Q_1 \text{cov}(Q_1, Q_2)}{Q_2^3} + \frac{Q_1^2 \text{var}(Q_2)}{Q_2^4}.
\]

Substituting estimated predictable variations and covariation leads to our proposed variance estimator:

\[
\text{var}(R_E) = \frac{1}{n^2Q_2(\tau)^2} \int_0^\tau \sum_i \hat{w}_i^2(t) \left( r_i^{(0)}(t) - r_i^{(M)}(t) \right)^2 \lambda_i(t| \mathcal{F}_{t-}) dt - \frac{2Q_1(\tau)}{n^2Q_2(\tau)^3} \int_0^\tau \sum_i \hat{w}_i^2(t) \left( r_i^{(0)}(t) - r_i^{(M)}(t) \right) \left( r_i^{(0)}(t) - r_i^{(P)}(t) \right) \lambda_i(t| \mathcal{F}_{t-}) dt + \frac{Q_1(\tau)^2}{n^2Q_2(\tau)^4} \int_0^\tau \sum_i \hat{w}_i^2(t) \left( r_i^{(0)}(t) - r_i^{(P)}(t) \right)^2 \lambda_i(t| \mathcal{F}_{t-}) dt.
\]

(10)

4.2 The population value of \( R_E \)

The population value \( \mathcal{R}_E \) of \( R_E \) is explored in Appendix A. A general expression is

\[
\mathcal{R}_E = 2 \frac{E \left\{ \int_0^\tau S(t)H_i(u)\alpha(t| \mathcal{F}_{t-}, u) dt \right\}}{E \left\{ \int_0^\tau S(t)\alpha(t| \mathcal{F}_{t-}, u) dt \right\}} - 1,
\]

where \( H_i(u) \) denotes the distribution function of linear predictors \( u \) at time \( t \) and \( S(t) \) is the marginal probability of a subject having non-zero intensity at \( t \) for the true counting process \( \hat{N}(t) \). We have written \( \alpha(t| \mathcal{F}_{t-}, u) \) as a reminder that the intensity depends on the prognostic index appearing in \( H_i(u) \). Further explanation is provided in Appendix A.
In general the expression is quite involved but we note that it is independent, as required, of the censoring/observation process \( G \) before \( \tau \). Useful simplification is possible for the case of single event survival data with time-constant linear predictors \( \beta x \) and unlimited follow-up. In that case the population value simplifies to

\[
R_E = 4 \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^u \alpha(t \mid u) S(t \mid v) S(t \mid u) h(u) h(v) dv du dt - 1,
\]

where \( \alpha(t \mid u) \) is the hazard, \( h(\cdot) \) denotes the density function of the linear predictor, and the integrations with respect to \( u \) and \( v \) are over the range of linear predictor values. We can use this expression to derive population values for specific models. For example, for the Cox model, where \( \alpha(t \mid \beta x) = \alpha_0(t) \exp(\beta x) \) and the covariates are constant in time, we get

\[
R_E = 2 \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{e^u - e^v}{e^u + e^v} h(u) h(v) dv du,
\]

which is equal (up to a linear transformation) to the result obtained by Gönen and Heller (2005) for the population value of the C index under no censoring (Harrell et al 1982).

In the case of one uniformly distributed covariate, \( X \sim U[0, 1] \), with a positive effect \( \beta \), this gives

\[
R_E = 1 - \frac{4}{\beta} \int_0^1 \log \frac{2e^{\beta x}}{1 + e^{\beta x}} dx,
\]

while the population value for the case of a binary covariate \( X \), that takes values \( x_1 \) and \( x_2 \) with probabilities \( \theta \) and \( 1 - \theta \), respectively, equals

\[
R_E = 2\theta(1 - \theta) \frac{e^{\beta x_2} - e^{\beta x_1}}{e^{\beta x_2} + e^{\beta x_1}}.
\]

As \( \beta \to \infty \), (13) converges to 1 and (14) converges to 0.5 for \( \theta = 0.5 \).

5. Simulations

In this section we use simulations to explore \( R_E \) as an estimator of the population value \( R_E \), the performance of the variance estimator (10) and the effect of censoring. We use a
A simple setting of fitting a Cox model to single event survival data generated with true hazard \( \alpha_i(t|\mathcal{F}_{t^-}) \equiv \exp(\beta X) \) and one covariate \( X \), chosen as one of:

- uniform: \( X \sim U[0, 3.46] \) (with unit variance);
- asymmetric: a mixture of \( U[0, 1.24] \) with probability 9/10 and \( U[0, 5 \times 1.24] \) with probability 1/10, (also unit variance); or,
- binary, \( X \sim Ber(0.5) \).

We begin by exploring the accuracy of the large-sample population values derived at (13) and (14) for the Cox model with uniform and binary covariates respectively. Figure 1 compares the population \( R_E \) for various \( \beta \) with the mean value of the sample version \( R_E \) over 1000 simulations at sample size \( n = 250 \), generated without censoring. The lines in the plot correspond to \( \pm 1.96 \) standard deviations, calculated from both the empirical variance of the 1000 repetitions and the square root of the average variance calculated using (10) at each simulation. For modest or large \( \beta \) the theoretical population value \( R_E \) matches very well the simulation mean of \( R_E \). For very small \( \beta \) however, the sample mean overestimates the theoretical quantity. The empirical and theoretical standard deviations are generally close, though (10) overestimates very slightly.

[Figure 1 about here.]

The bias at small \( \beta \) is perhaps expected, given that the same data are used for both model fitting and calculation of \( R_E \). Even if the true regression coefficient is zero our sample estimate will not be, meaning that there will be a suggestion of explained variation whatever measure is adopted. To provide reassurance, we performed further simulations at different sample sizes. At \( n=20 \) the mean estimated \( R_E \) from 1000 simulations with a uniformly distributed covariate was 0.115 (SE=0.113). When sample size was increased to \( n=100 \) the simulated mean\((R_E)\) was 0.048 (0.047), at \( n=500 \) it was 0.021 (0.022) and at \( n = 1000 \) it was 0.014 (0.015). Similar results were obtained for the other covariate distributions considered. We
are thus confident that the bias at small $\beta$ will become negligible as sample size increases. We note also that even in small samples the bias is of comparable magnitude to the standard error, and hence unlikely to lead us to a spurious conclusion of significant explained variation.

We now turn to the effect on $R_E$ of censoring. We distinguish two aspects of censoring: loss of all information after the maximum follow-up time $\tau$, and random censoring before that time. All measures of explained variation for survival data are expected to be affected by changes in $\tau$ but random censoring should affect only efficiency (Schemper and Henderson, 2000). Some remarks on model-based extrapolation of $R_E$ beyond $\tau$ are deferred to the closing discussion.

Figure 2 illustrates the change in $R_E$ as $\tau$ decreases so the amount of censoring increases, assuming no additional random censoring before $\tau$. The plot shows the mean value of $R_E$ over 1000 simulations at sample size 250, together with ±1.96 standard deviations, again calculated from both the empirical variance of the 1000 repetitions and the square root of the average variance calculated using (10). For these simulations we took $\beta = 1$. The theoretical standard deviation matches the empirical throughout, and there is evidence that $R_E$ increases as $\tau$ decreases, particularly for the asymmetric covariate case. This is because the model predicts very well while individuals with high values are still in the risk sets, and worse later on. With early censoring therefore we expect $R_E$ to be high.

[Figure 2 about here.]

In Figure 3 we explore the effect of random censoring before $\tau$, again with $\beta = 1$. Censoring times are exponentially distributed with rates determined to give various proportions of censored cases. In simulating we first generated uncensored data sets, $D_0$ say, and then thinned a copy by random censoring to give $D_1$ say. The largest event times in $D_0$ were then artificially censored to make $D_0$ have the same maximum event time as $D_1$. In this way the effect of random censorship is separated from the effect of changing $\tau$: estimates from both
$D_0$ and $D_1$ are obtained over the same follow-up interval. For each data set we calculated $R_E$ and its estimated variance. Figure 3 shows that the mean $R_E$ is unaffected by random censoring, as required, but the variance increases, as expected. Our variance estimator (10) continues to perform well and properly captures the additional variability.

[Figure 3 about here.]

6. Applications

6.1 Comparing survival data sets

Single-event survival data on three different types of leukaemia were described by Shimakura (2003). The types are acute myeloid leukaemia (AML), acute lymphblastic leukaemia (ALL) and chronic granulocytic leukaemia (CGL). All data were taken at the same time from the UK North West Regional Leukaemia Register. Four covariates are available for each data set: age, sex, white blood cell count and Townsend deprivation index. Censoring rates were 16% (AML), 27% (ALL) and 31% (CGL). There are many more cases of AML than the other types. Sample sizes, likelihood ratio test statistics $D$ for no covariate effects, and our $R_E$ and estimated standard errors are:

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>$D$</th>
<th>df</th>
<th>$R_E$</th>
<th>Std err</th>
</tr>
</thead>
<tbody>
<tr>
<td>AML</td>
<td>1043</td>
<td>263.4</td>
<td>4</td>
<td>0.362</td>
<td>0.011</td>
</tr>
<tr>
<td>ALL</td>
<td>206</td>
<td>65.6</td>
<td>4</td>
<td>0.417</td>
<td>0.043</td>
</tr>
<tr>
<td>CGL</td>
<td>233</td>
<td>32.5</td>
<td>4</td>
<td>0.262</td>
<td>0.043</td>
</tr>
</tbody>
</table>

We take from this the following. Covariate effects are highly statistically significant in all cases, particularly so for AML. This is because of its larger sample size rather than any greater predictive power. The covariates explain about the same rank variation for both acute forms of leukaemia but significantly less for the chronic form.
6.2 Local $R_E$

As a second example, we use data from the clinical trial studied by Stablein & Koutrouvelis (1985), which presents a frequently cited example of non-proportional hazards. The study included 90 patients with locally non-resectable gastric cancer, half of which were assigned to treatment with a combination of chemotherapy and radiation therapy and the other half to chemotherapy alone. The maximum follow-up of the study is 8 years, only 8 individuals were censored (all after the last event time). A Cox model fit gives a non-significant treatment effect $\beta = 0.105$ (SE=0.223) and $R_E = 0.123$ (0.061).

Figure 4 presents Kaplan-Meier curves for the two treatment groups and value of $R_E$ considered as a function of follow-up time $\tau$ (middle plot). As the $R_E$ value is a quotient of two cumulative sums, its value remains high despite the worse predictive power of the model in the later times. In fact, the model predicts worse than chance after the two hazards cross and this is clearly seen from the right plot of Figure 4, which shows the local $R_E$ over moving intervals $(t_1, t_2)$ each capturing 5 successive event times.

[Figure 4 about here.]

If the model allows the effect to change after 20 months, the estimated coefficient is highly significant before the change point ($\beta = -1.241$, SE=0.524) and borderline significant (in the opposite direction) later ($\beta = 0.496$, SE=0.256). The $R_E$ value for this model is 0.175 (0.061).

6.3 Recurrent-event infant diarrhoea data

Our final application concerns a more complex event-history data set and analysis. Borgan et al. (2007) describe an analysis of infant diarrhoea data collected from 926 children over 15 months in Salvador, Brazil. The data record for each day whether or not the child had diarrhoea and thus provide point-process data in discrete time. A variety of socio-economic covariates are available but a complication for analysis is that there are late
entrants, intermittent missingness and dropout. These are properly allowed for in the analyses described by Borgan et al., who concentrate also on dynamic methods by which information in the history $\mathcal{F}_{t-}$ other than baseline covariates is incorporated in a model for the intensity $\alpha_i(t|\mathcal{F}_{t-})$, as described by Aalen et al. (2004) and Fosen et al. (2006). We will follow Borgan et al. and consider both incidence and prevalence analyses. An incidence event occurs on a day when a new episode of diarrhoea begins, an episode being a sequence of days with diarrhoea until there have been at least three consecutive clear days. A prevalence event occurs on a day when the child has diarrhoea. Children in the data set had from 0 to 19 incidence events over 456 days of study, and from 0 to 182 prevalence events. Further information is provided in Borgan et al. (2007).

We consider a variety of models for incidence (I) and prevalence (P). Models I1/P1 are null in that they assume no covariate effects on the intensity of events. They are included for completeness in the residual plot to be described. Models I2/P2 assume a Cox proportional intensity model with time-fixed covariate effects. Models I3/P3 assume an Aalen additive intensity model (Aalen et al. 2008, Section 4.2) with time-fixed covariate effects. Models I4/P4 are more usual Aalen models in which covariate effects are allowed to change with time. Models I5/P5 include as a dynamic covariate the child-specific rate of previous diarrhoea estimates. Model P6 is appropriate only for the prevalence analysis as it additionally includes information on whether the child had diarrhoea on the previous days, up to lag four. The Aalen fits with and without dynamic covariates are described in Borgan et al. (2007): generally there are highly significant covariate effects, on the whole fairly constant over time.

Figure 5 and Table 1 provide diagnostic information. Figure 5 shows the standard deviation of standardised martingale residuals after fitting the models to the diarrhoea data: Borgan et al. used similar plots to compare the models they used. If the model is correct and there
is negligible estimation error then the empirical values in the plot should be close to one. Table 1 gives our $R_E$ and the associated estimated standard error.

For the incidence analysis, the impression from Figure 5 is that a good fit is obtained from $I_5$ but otherwise there is little to choose between models. From Table 1 we conclude that $I_5$ explains most rank variation, at a high level of 0.681, but the Aalen model with time varying covariates also performs reasonably well, as it should do given that covariate effects are estimated anew at each event time and so there are very many parameters. The two fixed-effect models $I_2$ and $I_3$ are similar, and each explains a significant amount of rank variation compared with no model at all, though much less than models $I_4$ and $I_5$. Note that Figure 5 can be used to assess whether a model fits well or not, but does not attempt to rank or otherwise discriminate between ill-fitting models. Conclusions from the prevalence analyses are similar except that only model $P_6$, which includes the lag effects, could be taken as reasonable in Figure 5. Knowing whether or not a child has diarrhoea one day is naturally highly predictive for the next day, leading to an increase from 0.650 to 0.862 in our $R_E$.

7. Discussion

We have proposed a measure or index which compares model-based ranks of subjects in risk sets with observed outcomes and then averages over time. This is in part similar to an informal suggestion of Hans van Houwelingen (personal communication) who has advocated for survival data plotting against time the ranks of events in risk sets. By contrasting with null and perfect models we are able to obtain a well defined scale, and through the introduction of weighted ranks and weighted averaging we are able to preserve consistency under random censoring. The measure is, however, affected if there are regions of the time axis which can
never be observed however large the sample, and specifically by a finite maximum follow-up time \( \tau \). A simple fix for our measure is available, but first we want to make this point: a measure should be consistent under censoring which occurs \emph{before} any maximum follow-up, but we can make a measure consistent under censoring after that time \emph{only} if we assume that covariates and effects remain constant or stationary thereafter. Any measure claiming consistency (e.g., Kent and O’Quigley 1988, Korn and Simon 1990 or Gönen and Heller 2005) makes such an assumption.

Intuitively, for survival data it is easy to understand why censoring after the last failure time will affect any measure of explained variation or predictive power except those that assume fixed covariates and effects. Such a measure gives an estimate of the overall performance of our model. Depending on the distribution of the covariates and changing of effects in time, the model will usually predict better in some time intervals, and worse in others. Cutting off at some point will therefore affect the overall measure.

If there are no time-varying covariates and it can be assumed that the coefficients \( \beta \) are constant in time, a consistent estimate of \( R_E \) can be obtained by replacing the integral in (18) with the sum over the estimated terms. For the Cox model, we use (12) to get

\[
\tilde{R}_E = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{i} \frac{e^{\hat{\beta}x_i} - e^{\hat{\beta}x_j}}{e^{\hat{\beta}x_i} + e^{\hat{\beta}x_j}}.
\] (15)

Consider now the case where we have censoring after the last event time (denote the number of these individuals by \( n_c \)) and we have calculated the sample version of \( R_E \) using (6). In order to obtain an estimate of our measure independent of the censoring after the last event (assuming that everything remains fixed after that time), we calculate the sum in (15) over the individuals still at risk (replace \( n \) by \( n_c \)) and denote this value by \( \tilde{R}_E^* \).

As the value of the denominator is fixed with the number of individuals at risk, and equals \( n_c(n_c - 1)/4 \), we can calculate our measure as
\[
R'_E = \frac{\sum_i \int_0^T \hat{w}_i(t) \left( r_i^{(0)}(t) - r_i^{(M)}(t) \right) dN_i(t) + \tilde{R}'_E * n_c(n_c - 1)/4}{\sum_i \int_0^T \hat{w}_i(t) \left( r_i^{(0)}(t) - r_i^{(P)}(t) \right) dN_i(t) + n_c(n_c - 1)/4}.
\]

As well as deriving the population value, in Section 4 we obtained an estimator of the variance of \(R_E\), which performed well in the simulations. It is relatively unusual to attach an uncertainty estimator to a measure of explained variation, though our first application on comparison of fits to different survival data sets indicated the usefulness. In simulations and the applications we used the observed values \(Q_1(t)\) and \(Q_2(t)\) in calculating the estimated variances but another option is to use estimated expected values \(q_1(t)\) and \(q_2(t)\) in which increments in the observed counting processes are replaced by estimated intensities. Our experience is that this makes little difference to performance.

There has been considerable discussion about the usefulness of \(R^2\) in linear regression models (eg Barrett 1974, Healy 1984, Scott and Wild 1991 or Kvåseth 1985). One of the major complaints was that \(R^2\) depends on the distribution of the covariates. This is true, but we find this property not only natural, but desirable. But this is not the place to discuss this, we only want to add that in survival studies we have another variable, the survival time, whose measurements may affect the estimate of explained variation. Two studies, with equal distribution of the covariates, may give different estimates of explained variation if the observation periods are different. As we have indicated, our measure can be corrected for this (under some assumptions), but we believe that it would be a good practice if such measures were reported for different time values. Such practice would make it possible to compare studies with different observation periods. Another attractive option is to plot our local measure as a function of time, for a given window length, as illustrated on the gastric cancer data in Section 6.

In Section 6 we also explored the use of \(R_E\) in comparing model fits to recurrent event data. We are not aware of any other measure of explained variation which could be used in this
way. We are currently exploring how $R_E$ can be adapted for multi-state models, including competing risks. In this area there is again no existing measure which can be used.

An R function for calculating the value of the $R_E$ measure for the Cox, Aalen and parametric models is available at

http://ibmi.mf.uni-lj.si/ibmi-english/biostat-center/programje/Remeasure.r

References


We assume the fitted model is correct and that consistent estimators are available, including for the censoring time distribution $G$. For simplicity and without loss, here we do not distinguish estimated from true parameter values.

Recall that $Y_i(t)$ is an observation indicator for subject $i$ at time $t$, with $E[Y_i(t)|\mathcal{F}_{t-}] = G_i(t|\mathcal{F}_{t-})$. Let $I_i(t)$ be an indicator of subject $i$ having non-zero $\alpha_i(t|\mathcal{F}_{t-})$ in (2), i.e., of being at "risk" at $t$ of an event in the true counting process $\tilde{N}_i(t)$. Let $E[I_i(t)|\mathcal{F}_{t-}] = S_i(t|\mathcal{F}_{t-})$. Note that for single-event survival data $S_i(t|\mathcal{F}_{t-}) = S(t|x_i)$ is just the usual survival function.

Assume that the rankings are to be made at each event time on the basis of continuous
linear predictors \( \{u_{jt}; j \in \mathcal{I}_t\} \) with no ties. Define
\[
r^{(M)}(t, u) = \frac{1}{n} \sum_{j \in \mathcal{I}_t} \frac{I(u_{jt} > u)}{G_{jt}} = \frac{1}{n} \sum_{j=1}^{n} \frac{Y_j(t)I_j(t)I(u_{jt} > u)}{G_{jt}}
\]
and
\[
r^{(0)}(t, u) = \frac{1}{2n} \sum_{j \in \mathcal{I}_t} \frac{1}{G_{jt}} = \frac{1}{2n} \sum_{j=1}^{n} \frac{Y_j(t)I_j(t)}{G_{jt}}.
\]

Then as \( n \) increases \( r^{(M)}(t, u) \) converges to \( S(t)\{1 - H_t(u)\} \), where \( S(t) \) is the marginal of \( S_i(t|\mathcal{F}_{t-}) \) and \( H_t(u) \) is the distribution function of linear predictors amongst subjects at risk of events in \( \tilde{N} \) at \( t \). Similarly \( r^{(0)}(t, u) \) converges with \( n \) to \( S(t)/2 \). Note that if prognostic indices \( u \) are constant in time (at their baseline values) with density \( h(u) \) then
\[
S(t)H_t(u) = S(t) \int_0^u h(v|I(t) = 1) dv
= S(t) \int_0^u S(t|v)h(v) dv
= \int_0^u S(t|v)h(v) dv.
\] (17)

We are interested in the population value \( \mathcal{R}_E \) say of the explained rank measure \( R_E \) defined at (6). Dividing both numerator and denominator of (6) by \( n^2 \), noting that the terms in \( \left( r^{(0)}_i(t) - r^{(M)}_i(t) \right) \) and \( \left( r^{(0)}_i(t) - r^{(P)}_i(t) \right) \) which are not in \( r^{(M)}(t, u) \) and \( r^{(0)}(t, u) \) are negligible, that the weights \( w(t) \) correctly handle any random censoring, and that we have assumed between subject independence, we find
\[
\mathcal{R}_E = \frac{2E \left\{ \int_0^\tau \left\{ S(t)H_t(u) - S(t)/2 \right\} \alpha(t|\mathcal{F}_{t-}, u) dt \right\}}{E \left\{ \int_0^\tau S(t)\alpha(t|\mathcal{F}_{t-}, u) dt \right\}} - 1.
\] (18)

where we have written \( \alpha(t|\mathcal{F}_{t-}, u) \) as a reminder that the intensity depends on the prognostic index appearing in \( H_t(u) \).

For single-event survival data with unlimited follow-up (\( \tau = \infty \)), time-constant covariates and effects, and \( \alpha(t|\mathcal{F}_{t-}, u) = \alpha(t|u)S(t|u) \) then the denominator is just 1/2 and (11) follows from (18) and (17).
APPENDIX B

Maximum value of $R_E$.

We will consider uncensored survival data without ties and with unlimited follow-up. So (6) can be written

$$R_E = 1 - \frac{\sum_i \left( \int_0^\infty r_i^{(M)}(t) - r_i^{(P)}(t) \right) dN_i(t)}{\sum_i \int_0^\infty (r_i^{(0)}(t) - r_i^{(P)}(t)) dN_i(t)}.$$ 

In a perfect model the integrand in the numerator is equal to zero. But suppose the covariates group individuals into $k$ groups of size $d_j$, $\sum_{j=1}^k d_j = n$. Then the best any model can do is assign to subjects with events the same rank as all subjects in that group. For a group of size $d_j$, this is $(d_j + 1)/2$ when all subjects are still at risk, falling by 0.5 with every individual lost from follow-up. Adopting the convention that $r_i^{(M)}(t) = r_i^{(0)}(t)$ when only one group is at risk, and noting that the perfect rank predictions (5) become $r_i^{(P)}(t) = 1$ at event times when there are no ties, the contribution of group $j$ to the numerator of $R_E$ is at most

$$\int_0^\infty \sum_{i=1}^{d_j} (r_i^{(M)}(t) - 1)dN_i(t) = \sum_{i=1}^{d_j} \left( i + \frac{1}{2} - 1 \right) = \frac{1}{2} \sum_{i=1}^{d_j} (i - 1) = \frac{1}{2} \left\{ d_j(d_j + 1) - d_j \right\} = \frac{d_j^2 - d_j}{4}.$$ 

Summing over all $k$ groups, we get

$$\sum_{i} \left( \int_0^\infty r_i^{(M)}(t) - r_i^{(P)}(t) \right) dN_i(t) = \frac{1}{4} \sum_{j=1}^k (d_j^2 - d_j) = \frac{1}{4} \left( \sum_{j=1}^k d_j^2 - n \right).$$
Using the fact that the sum of the terms in the denominator of $R_E$ is $n(n - 1)/4$, we get

$$
R_E = 1 - \frac{\frac{1}{4} \left( \sum_{j=1}^{k} d_j^2 - n \right)}{\frac{1}{4} n(n - 1)}.
$$

Writing $n\theta_j$ instead of $d_j$ and rearranging the terms gives formula (7). If there are no groups then $k = n$, each $\theta_j = 1$ and (7) is one as required.
Figure 1. $R_E$ population value (○) for Cox proportional hazards, together with the mean sample estimate (●) from 1000 simulations at $n = 250$ and ±1.96 standard deviations based on the theoretical (grey) and empirical (black) variances.
**Figure 2.** $R_E$ for survival data with respect to the proportion of individuals censored at follow-up time $\tau$. Circles represent the mean value of $R_E$ based on 1000 simulation runs, intervals are ±1.96 standard deviations based on theoretical (grey) and empirical (black) variance estimators.
**Figure 3.** $R_E$ calculated on uncensored data $D_0(\bullet)$ and copies $D_1(\circ)$ with additional random censoring. Intervals are ±1.96 standard deviations based on theoretical (grey) and empirical (black) variance estimators.
Figure 4. Gastric cancer data: survival curves with respect to treatment (left), smoothed (loess) $R_E(\tau)$ against follow-up $\tau$ (middle), and smoothed local $R_E(t_1, t_2)$ (right), with intervals selected to include 5 event times. The higher initial value of $R_E(t_1, t_2)$ than $R_E(\tau)$ is an artefact of the smoothing.
Figure 5. Standard deviations of standardised martingale residuals for various model fits to diarrhoea data. Left plot: incidence analysis. Right plot: prevalence analyses.
Table 1

$R_E$ and associated standard error for a variety of model fits to the diarrhoea data

<table>
<thead>
<tr>
<th>Incidence</th>
<th>$R_E$</th>
<th>Std err</th>
<th>Prevalence</th>
<th>$R_E$</th>
<th>Std err</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I1</td>
<td>0.000</td>
<td>0.000</td>
<td>Model P1</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Model I2</td>
<td>0.238</td>
<td>0.013</td>
<td>Model P2</td>
<td>0.350</td>
<td>0.008</td>
</tr>
<tr>
<td>Model I3</td>
<td>0.238</td>
<td>0.014</td>
<td>Model P3</td>
<td>0.346</td>
<td>0.008</td>
</tr>
<tr>
<td>Model I4</td>
<td>0.597</td>
<td>0.010</td>
<td>Model P4</td>
<td>0.590</td>
<td>0.006</td>
</tr>
<tr>
<td>Model I5</td>
<td>0.681</td>
<td>0.010</td>
<td>Model P5</td>
<td>0.650</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Model P6</td>
<td>0.862</td>
<td>0.005</td>
</tr>
</tbody>
</table>