SAMPLE SIZE ESTIMATION FOR CANCER PROGRESSION MODELS

Christian Netzer* and Jörg Rahnenführer*

Abstract

Human tumours are often associated with the accumulation of chromosomal alterations in the cancer cells. The identification of characteristic pathogenic routes improves prediction of survival times and optimal therapy choice. The simplest model assumes independent alterations. Then progression is measured by the count statistic, the total number of alterations. An advanced model is the oncogenetic trees mixture model. An oncogenetic tree allows both independent and sequential relationships between alterations, and the mixture model divides the patients into groups with different progression paths. Progression along such a model can be quantified univariately by the GPS (genetic progression score). On real cancer data, the GPS was shown to discriminate better than the count statistic between patient subgroups with different survival prognosis. Here, in a simulation study, we evaluate the necessary numbers of patients for detecting true relationships between genetic progression and survival time. We generate survival times correlated with count statistic and GPS, respectively. If the simple model is the correct one, misspecification with the advanced model requires about 20% larger sample size, independent from the number of events. In contrast, misspecification with the simple model leads with increasing numbers of events from 20% to 70% larger sample size. Additionally, if the true data-generating model is the mixture model, the absolute numbers are more than twice as large, thus favouring the advanced modelling approach especially in situations with limited model knowledge.

Key Words

Genetic progression models, sample size, survival analysis, simulation study

1. Introduction

Cancer pathogenesis and progression in tumours is characterized by the accumulation of genetic changes in the cancer cells. Frequently observed alterations are gains or losses of parts of chromosomes or mutations of specific genes. Tumour progression of a single patient can be estimated by quantifying the state of the tumour in a progression model. The state is typically related to the remaining time of the respective patient until death or tumour recurrence. Here, we provide a study for estimating and comparing the required sample size (of tumour probes) for detecting significant relationships between different cancer progression scores and the survival times of the corresponding patients.

When modelling disease progression, the genetic alterations are typically assumed to be irreversible. In general, the crucial alterations are not independent of each other such that simple counting of genetic alterations is not sufficient for accurately quantifying disease progression. Oncogenetic trees constitute a more flexible modelling approach than the basic independence model [1]. In an oncogenetic tree model, every genetic event has exactly one precursor event. This allows modelling both independent and sequential relationships. A probabilistic framework [1] provides convenient estimation of the tree topology. The topology consists of order relationships between the events and conditional probabilities for the occurrence of successive events. The latter are probabilities for the occurence of successor events given the precurser events have happend.

However, this model suffers from the drawback that single observations not fitting the general overall model influence the model building process too strong. *Oncogenetic trees mixture models* are mixtures of single oncogenetic trees. They can be used to estimate different disease progression paths for different subgroups of the patient cohort under consideration [2]. This biostatistical model for genetic tumour progression has been evaluated statistically and clinically in many ways over the recent years [3]–[6]. The advantage of the mixture model is twofold. First, a likelihood approach can be applied, as a noise component pools all patients that are not in line with the main disease progression mechanisms. Second, the model offers a high level of interpretability, also in terms of subsequent analysis of potential biological explanations of ordered genetic alterations.

Furthermore, the probabilistic framework of the oncogenetic trees mixture model allows for the introduction of a genetic progression score (called GPS in the following) that quantifies tumour progression univariately according to the state of a patient along the disease progression paths. In various detailed analyses using Cox regression models, it was shown that for patients with prostate cancer and for patients with different types of brain cancer, a higher

* Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany; e-mail: {netzer, rahnenfuehrer}@statistik.tu-dortmund.de

Recommended by Dr. R. K. M. Karuturi
(DOI: 10.2316/J.2012.210-1029)

GPS is significantly correlated with shorter survival time or time to tumour recurrence [3], [7]. Major issues for evaluating the efficiency of the mixture model are assessment of model stability and of the true prediction quality of the GPS regarding survival times. Previous simulation studies showed that the topology of the mixture model cannot always be reliably estimated for small or moderate sample sizes [5]. Furthermore, the complex model can lead to overfitting due to the larger number of parameters in comparison with a simple independence model. This raises the question in which situations simple counting of genetic events is better suited for quantifying disease progression than using the more complex GPS.

Here, we provide guidance for a qualified model selection, *i.e.*, for deciding if the independence model or the mixture model is better suited for identifying a true relationship between genetic progression and survival time. We first simulate genetic measurements from the independence model and the tree mixture model and calculate the progression scores for the simulated data. Then we draw survival times related to either the GPS or the count statistic and calculate necessary sample sizes of patients for re-identifying the specified relationships. The main goal is to compare the two approaches with respect to the number of patients needed to detect an association between the progression score and the survival. Especially of interest is the situation in which one model agrees with the truth and the other is misspecified. The question is how much the required sample size increases due to this misspecification.

In Section 2, we introduce the progression models and progression scores in more detail and present a short description of the Cox regression framework [8] for modelling our favoured dependencies between genetic measurements and survival times. In Section 3, we introduce our simulation study for computing the sample sizes needed to obtain a required power. Particularly, in Section 4.2, we analyse in detail the impact of the decision between simple and complex progression scores.

2. Methods

We first briefly introduce oncogenetic trees, mixture models of oncogenetic trees, and the derived progression score GPS. Then we explain how to relate genetic measurements to survival times with the Cox proportional hazard regression model.

2.1 Oncogenetic Trees Mixture Models

An oncogenetic tree $\mathcal{T}$ is a probabilistic model for describing dependencies between several binary random variables $X_r, X_1, \ldots, X_l$ (with value 0 or 1), where 1 indicates occurrence of a genetic event, *e.g.*, of a mutation. The tree $\mathcal{T} = (V, E, r, p)$ is a directed and weighted acyclic graph with root r. $V = \{r, 1, \ldots, l\}$ denotes the set of vertices (events) and $E \subset \left[V^2\right]$ denotes the set of edges. The edge weights p correspond to the conditional probabilities $p(e) = Pr(X_v = 1 \mid X_u = 1)$, where u is a precursor event of v and $p(e)$ is the probability that v occurs after u has been observed. For the initial null event we set $Pr(X_r = 1) = 1$.

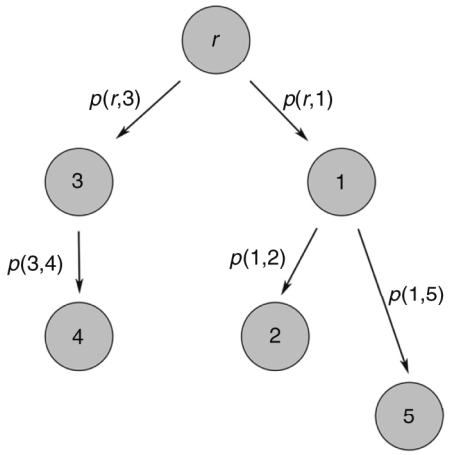

Figure 1. Oncogenetic tree with $l = 5$ events and conditional edge probabilities $p(e)$.

It can be interpreted as starting point of the disease. Figure 1 depicts a basic example of an oncogenetic tree with $l = 5$ events.

The genetic pattern of a patient is described by a binary vector x of length l, where 1 indicates occurrence of the respective event. The tree $\mathcal{T}$ is a star if and only if all events are independent. Only in this case it holds $Pr(x) > 0 \;\; \forall x \in \{0,1\}^l$.

Such a tree model induces a probability distribution $P(X)$ on the set $\Omega = \{0,1\}^l$ of all possible 2^l genetic patterns. The probability that the tree $\mathcal{T}$ generates pattern x is given by

$$Pr(x \mid \mathcal{T}) = \prod_{e \in E'} p(e) \cdot \prod_{e \in S \times (V \setminus S)} \big(1 - p(e)\big) \qquad (1)$$

Here $S \subseteq V$ denotes the vertices in the tree that belong to the events that have occurred according to pattern x. $E' \subseteq E$ is the subset of edges such that S is the set of all vertices that can be reached from the origin in the partial tree $\mathcal{T} = (S, E', r, p)$. For all patterns x for which such an edge subset E' does not exist, it holds $Pr(x \mid \mathcal{T}) = 0$ as these patterns cannot be generated from $\mathcal{T}$.

Given a sample of N patients each associated with a genetic pattern x, the data can be summarized in the sample matrix $X_N = (x_{ij})_{\substack{1 \le i \le N \\ 1 \le j \le l}}$. In many applications, every patient is observed only once. For example, in cancer tumours genetic events can often only be measured once due to surgery that removes the tumour tissue. For this type of cross-sectional data, Edmonds' branching algorithm [9] can be used to estimate an adequate compatible $\mathcal{T}$. It can be shown that this algorithm generates the true tree asymptotically with probability 1 [1].

In a single oncogenetic tree model $\mathcal{T}$, only patterns that are in line with tree topology have positive likelihood. For a data sample with at least one pattern with probability 0, the joint likelihood would also be 0. Alternatively, observed patterns would have to be excluded inappropriately from the data. This drawback was overcome by the introduction of mixture models of oncogenetic trees [2]. An oncogenetic

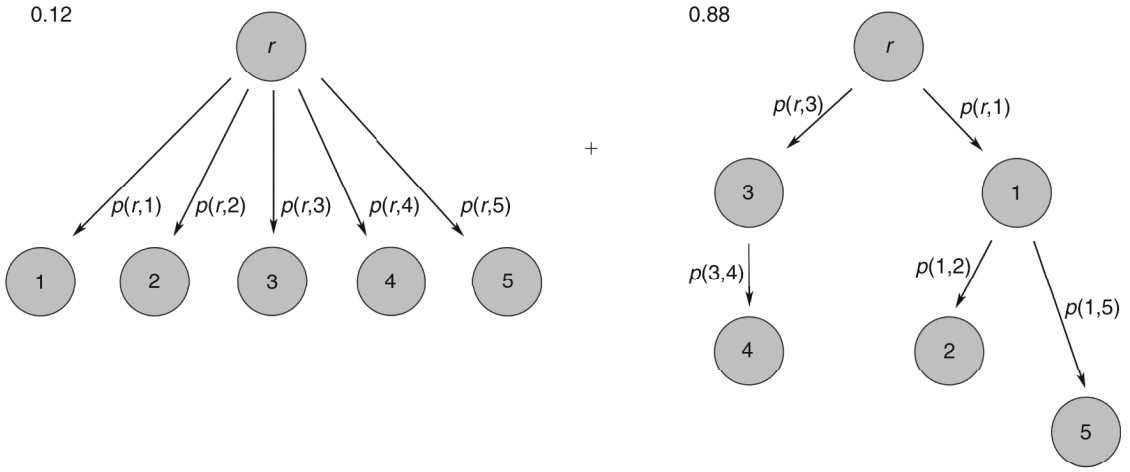

Figure 2. Oncogenetic mixture model $\mathcal{M} = 0.12 \cdot \mathcal{T}_1 + 0.88 \cdot \mathcal{T}_2$ with $l = 5$ events and conditional edge probabilities $p(e)$.

trees mixture model is defined by

$$\mathcal{M} = \sum_{k=1}^{K} \alpha_k \mathcal{T}_k, \quad \left(\alpha_k \in [0,1] \text{ and } \sum_{k=1}^{K} \alpha_k = 1 \right) \tag{2}$$

where $\mathcal{T}_k$, $k \in 1, \ldots, K$ are single oncogenetic trees with vertex set V. In addition, the first tree $\mathcal{T}_1$ is restricted to have star topology, *i.e.*, all events are assumed to be independent; see Fig. 2 for an example of a mixture model with one trivial component. The weights $\alpha_1, \ldots, \alpha_K$ are called mixture parameters and denote the probabilities with which the respective tree components generate patterns. The parameter α_1 can be interpreted as noise fraction in the data. For $\mathcal{T}_1$ and consequently for the full model $\mathcal{M}$, it holds $Pr(x) > 0 \; \forall x \in \Omega$.

The probability that pattern x is generated by a mixture model $\mathcal{M}$ is given by

$$Pr(x \mid \mathcal{M}) = \sum_{k=1}^{K} \alpha_k Pr(x \mid \mathcal{T}_k) \tag{3}$$

where the probabilities $Pr(x \mid \mathcal{T}_k)$ are given by (1).

For fitting an oncogenetic mixture model given a sample matrix $X_N = (x_{ij})_{\substack{1 \le i \le N \\ 1 \le j \le l}}$, Edmonds' branching algorithm cannot be directly applied as the assignment of samples to tree components is not *a priori* given. With an Expectation-Maximization(EM)-like learning algorithm introduced in [2], the optimal assignment of samples to estimated tree components can be achieved.

2.2 Genetic Progression Scores

The quantification of disease progression as a univariate measure depends on the underlying progression model. In the case of the independence model, one obvious choice is to simply count the events observed. Formally, for the ith patient ($i = 1, \ldots, N$) with genetic pattern x_i, the so-called count statistic is defined by

$$\text{count}(x_i) = \sum_{j=1}^{l} x_{ij} \tag{4}$$

If all events in the vertex set V are independent and $Pr(X_u) = Pr(X_v) \; \forall u, v \in V$, the count statistic is sufficient for the sample X_N.

For the oncogenetic trees mixture model, a more sophisticated progression measure was introduced in [3]. The so-called GPS measures how many consecutive steps of the tree model already have occurred according to the respective pattern. More precisely, if a pattern x_i is in line with one of the non-trivial components of a mixture model $\mathcal{M}$, the GPS tells how far the patient has advanced regarding the tree, *i.e.*, in terms of cancer progression how far the disease has progressed.

Mathematically, the GPS is calculated by first transforming the conditional probabilities on the tree edges into waiting times. The GPS of a pattern x_i is the expected waiting time until this pattern is observed. We briefly recapitulate this approach. Let the waiting time W_v be the time between occurrence of an event v and its precursor event u. Assume that W_v is exponentially distributed with parameter λ_v. Let W_S be the sampling time of a tumour, *i.e.*, the time between onset and discovery of the disease, and assume that W_S is also exponentially distributed with parameter λ_S. The conditional probability $p(u, v)$ for the occurrence of v given occurrence of u can then easily be calculated as

$$p(u, v) = Pr(W_v < W_S) = \frac{\lambda_v}{\lambda_v + \lambda_S} \tag{5}$$

The parameter λ_S of the sampling time cannot be estimated from the data and is chosen such that

$$E(W_S) = \frac{1}{\lambda_S} = 1 \tag{6}$$

This choice can be interpreted as a scaling of the progression times (*e.g.*, the tumour age) to a mean value of 1. With (5) and (6), the conditional probabilities p can be converted into the parameters of the exponential distribution by

$$E(W_v) = \frac{1}{\lambda_v} = \frac{1 - p(u,v)}{p(u,v)} \frac{1}{\lambda_S} = \frac{1 - p(u,v)}{p(u,v)} \tag{7}$$

The expected value of the waiting time W_v, conditional on its precursor event, can thus be directly calculated for every event in V and estimated by plugging in the estimates for the probabilities p.

For a genetic pattern x_i, the GPS is defined as the waiting time from the root node r until all events in x_i have been observed:

$$\mathrm{GPS}(x_i) = E_{\mathcal{M}}\left(W(x_i)\right) = \sum_{k=1}^{K} Pr(\mathcal{T}_k \mid x_i) \cdot E_{\mathcal{T}_k}\left(W(x_i)\right) \tag{8}$$

In general, this waiting time is a combination of maxima and sums of exponential distributions and cannot be calculated explicitly. It is approximated by simulating data from the mixture model with corresponding exponential distributions on the tree edges and then averaging all times at which a specific pattern x_i was observed. For a reliable estimate, at least 10^6 simulation runs are required [3]. Confidence intervals for GPS values can be obtained with compute intensive bootstrap calculations [5]. All these methods are available in the R package Rtreemix [10].

2.3 Cox Proportional Hazards Model

We are interested in evaluating the significance of a relationship between genetic progression scores and survival times. The standard established model for this task is the Cox proportional hazards regression model, where the hazard ratio of two patients with different covariate vectors is modelled to be constant over time. Here, the covariate of interest is the progression score, in our case either the count statistic or the GPS. Confounding variables like age or tumour size can be added as additional covariates in the model. In the following, we briefly introduce the standard Cox approach; for details see [11].

Let T be a nonnegative continuous random variable that describes the time until occurrence of an event, *i.e.*, in this context the time until death or tumour recurrence. Its distribution can be characterized by the survival time $S(t) = 1 - F(t) = P(T > t)$, where $S(t)$ is the probability that the event of interest happens after time point t. Survival times are often (right-)censored such that it is only known that the survival time exceeds a random time point t_c. Thus survival times are often modelled within a hazard rate framework. The hazard rate $\lambda(t)$ describes the instantaneous failure rate at time point t, given the event has not been observed at t, and is related to the survival function by

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left[t \le T < t + \Delta t \mid T \ge t\right]}{\Delta t} = \frac{f(t)}{S(t)} \tag{9}$$

The Cox model [8] is a semiparametric regression model for describing the influence of a covariate vector $Z = (Z_1, \ldots, Z_p)$ on the hazard rate. In its basic version, the hazard rate is modelled as a combination of a baseline hazard rate and a time-independent factor containing the covariates:

$$h\left(t \mid Z\right) = h_0(t) \exp\left(\beta' Z\right) = h_0(t) \exp\left(\sum_{k=1}^{p} \beta_k Z_k\right) \tag{10}$$

where $\beta = (\beta_1, \ldots, \beta_p)$ has to be estimated from the data and β_k describes the strength of the influence of covariate k on the hazard rate. For two patients with different covariate vectors Z and Z^*, the hazard ratio is given by

$$\frac{h(t \mid Z)}{h(t \mid Z^*)} = \frac{h_0(t) \exp\left[\sum_{k=1}^{p} \beta_k Z_k\right]}{h_0(t) \exp\left[\sum_{k=1}^{p} \beta_k Z_k^*\right]} = \exp\left[\sum_{k=1}^{p} \beta_k (Z_k - Z_k^*)\right] \tag{11}$$

and thus independent of the time. The Cox model allows simultaneous integration of continuous and categorical covariates. For the special case of a univariate covariate Z, we define the hazard ratio HR as the relative risk change when the covariate is changed by 1 unit:

$$\mathrm{HR} = \frac{h\left(t \mid Z + 1\right)}{h\left(t \mid Z\right)} = \exp(\beta) \tag{12}$$

One standard approach for testing the global null hypothesis $H_0 : \beta = \beta_0$ is the Wald test (cf. [11], p. 254). The test statistic is given with

$$\chi^2_{Wald} = (\hat{\beta} - \beta_0)' I(\hat{\beta})(\hat{\beta} - \beta_0)$$

where $\hat{\beta}$ and $I(\hat{\beta})$ are the maximum-likelihood estimates for β and the Fisher information matrix. Under the null hypothesis H_0, the test statistic is asymptotically χ^2-distributed with p degrees of freedom, see [11] for details.

3. Simulation Study for Sample Size Estimation

In recent studies it was demonstrated that the GPS is a highly significant prognostic marker for cancer survival [3], [7]. The GPS is a significant covariate in a Cox model even conditioned on adjustment for established clinical markers such as, the Gleason score for prostate cancer. However, in recent new analyses regarding gliomas (a malign type of brain cancer) with less than 40 uncensored cases in the patient cohort, the sample size was too low for confirming the relevance of the GPS as a prognostic factor [6] (data not shown).

These findings raise the question how many patients are required in such a study to detect a (true) relationship between GPS and survival time? In Section 3.1, we introduce a new simulation study answering this question. Next, with the scenarios described in Section 3.2, we want to analyse the effect of sample size if the true data-generating progression model is the independence model and the progression score is the count statistic. Finally, we compare the numbers of required samples when the model is misspecified, *i.e.*, when either the mixture model is the true model and the count statistic is calculated from the data as progression measure or the independence model is the true model and the GPS is calculated from the data.

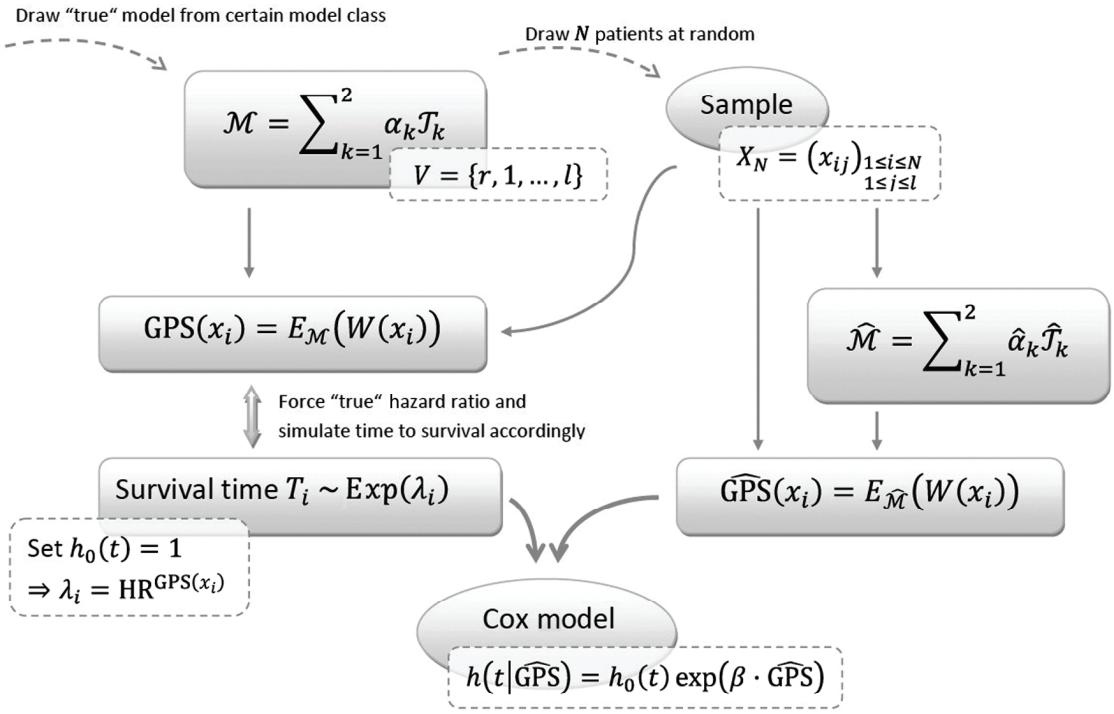

Figure 3. Schematic description of the simulation setup. Running 1,000 iterations, for a fixed number of events l, sample size N and hazard ratio HR, we estimate the power by the mean number of rejected hypothesis.

3.1 Simulation Setup for Evaluating the Performance of the GPS

We analyse and compare two progression models, the oncogenetic trees mixture model $\mathcal{M} = \sum_{i=1}^{2} \alpha_k \mathcal{T}_k$ and the independence model, see the following section. The latter can be described as a star $\mathcal{T}_{star}$ with unequal edge weights. In both cases, the genetic events are represented by a vertex set V. We assume that the number of genetic events is given and fixed. Thus, in case of data sets with originally too large numbers of events, we suppose that a certain type of event (feature) selection has already been applied.

For assessing the relevance of the progression scores regarding survival prediction, we plug in the estimated GPS or count statistic, respectively, in a Cox model as single covariate. The significance is then obtained via the p-value of the Wald test as described in Section 2.3.

We assume that the survival time T_i of patient i is exponentially distributed with parameter λ_i. Thus the hazard rate λ_i of patient i is constant over time. For modelling the relationship between progression score and survival, we use the Cox model (10) and assume, w.l.o.g., $h_0(t) \equiv 1$. Let x_i be the genetic pattern of patient i. Then the hazard ratio HR (see (12)) connects the survival parameter λ_i with the progression score in the following way:

$$\begin{aligned} \lambda_i = h(t \mid \mathrm{GPS}(x_i)) &\overset{(10)}{=} h_0(t) \cdot \exp\big(\beta \cdot \mathrm{GPS}(x_i)\big) \\ &\overset{(12)}{=} 1 \cdot \exp\big(\ln(\mathrm{HR}) \cdot \mathrm{GPS}(x_i)\big) \\ &= \mathrm{HR}^{\mathrm{GPS}(x_i)} \end{aligned} \tag{13}$$

Our basic default simulation setup is visualized in Fig. 3 and defined as follows:

1. Draw at random a mixture model $\mathcal{M} = \sum_{i=1}^{2} \alpha_k \mathcal{T}_k$ with $l+1 = |V|$ events from a specified model class (see details below).
2. Draw a sample of size N from the model.
3. Compute for each genetic pattern x_i its genetic progression score $\mathrm{GPS}(x_i) = E_{\mathcal{M}}(W(x_i))$.
4. Simulate for each patient i a corresponding survival time from an exponential distribution with parameter $\lambda_i = \mathrm{HR}^{\mathrm{GPS}(x_i)}$, see (13).
5. Calculate the estimate $\widehat{\mathcal{M}} = \sum_{i=1}^{2} \hat{\alpha}_k \hat{\mathcal{T}}_k$ from the samples obtained in step 2.
6. Compute the estimated GPS values $\widehat{\mathrm{GPS}}(x_i) = E_{\widehat{\mathcal{M}}}(W(x_i))$.
7. Fit a Cox proportional hazard model with survival times from step 4 as response and estimated GPS values from step 6 as single covariate. Then test the hypothesis $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ with significance level $\alpha = 0.05$.

For each set of parameters $\{\mathrm{HR}, N, l\}$ we apply $1,000$ iterations of this simulation. We then estimate the power $\varphi_{\{\mathrm{HR},N,l\}}$ of detecting the relevance of the GPS by counting the number of rejected null hypothesis. There is no analytical way of determining the sample sizes. Although several methods exist for sample size calculations for Cox regression models with single continous covariates [12], in our special case the GPS always has to be simulated. For any combination N, l and for $\mathrm{HR} = 1$, the expected value for rejecting the null hypothesis $\beta = 0$ is equal to α.

The model space from which random models are drawn in step 1 of the simulation setup is characterized by the

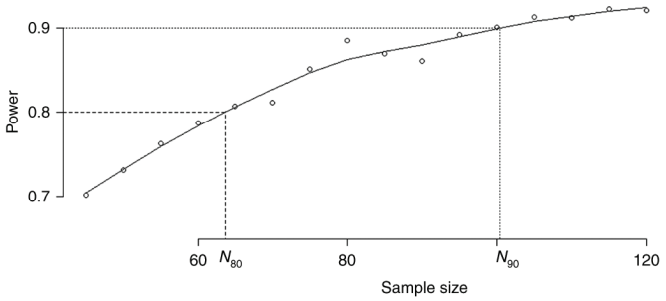

Figure 4. Reading out the required sample sizes N_{80} and N_{90} to achieve 80% and 90% power, respectively.

number of events l, the mixture parameters α_1 and α_2 and the conditional edge probabilities $p(u, v)$. To evaluate the influence of the number of genetic events in the model, we choose several values of l varying from 4 to 15. With fixed l, the topology of a single tree can be generated by randomly choosing a Prüfer code [13] and translating it in the corresponding tree. The mixture parameters α_1 and α_2 are set to 5% and 95%, respectively, corresponding to a noise level of 5% which is common in statistical analyses. Conditional edge weights $p(u, v)$ are drawn independently from a uniform distribution in the interval [0.2, 0.8], following [14]. These parameter choices are in agreement with those in previous publications [5], [14]. Evaluations with modifications of these parameters did not change the overall conclusions [6] (data not shown).

We want to determine the sample size required to detect a certain relationship between progression score and survival time. The strength of the true relationship is determined by the hazard ratio (12). We select critical interesting power levels 80% and 90% and denote the corresponding sample sizes with N_{80} and N_{90}, respectively. Because one cannot directly compute these sample sizes, we estimate them from simulated power curves as follows. We approximate the power curve by smoothing the power estimates obtained on a grid around the expected sample size values, see Fig. 4. In this plot, each point corresponds to the percentage of rejected null hypotheses of 1,000 simulations following the setup depicted in Section 3.1. Smoothing is performed with the *smoothing.spline* function in R where the smoothing parameter is chosen automatically by cross-validation. Visual inspection confirmed adequate approximations in all cases.

For all computations of the simulation study, the free available statistic software R was used [15]. Functions for generating mixture models and for computing GPS values were provided by the R package *Rtreemix* [10] in version 1.8.0. Cox proportional hazard models and Wald tests were computed with the R package *survival* [16] in version 2.35-8.

3.2 Simulation Setup for Comparing Mixture Model and Independence Model

In this section, we introduce the setup for comparing a trees mixture model with corresponding GPS and an independence model with corresponding count statistic. In all cases, we are interested in comparing sample sizes needed to detect an association between progression score and survival time.

In the first scenario, the true model is a mixture model, in agreement with step 1 of the simulation algorithm presented in Section 3.1. Here, in addition to running the standard algorithm, we modify steps 5–7. We skip steps 5 and 6 and replace in step 7 the GPS as covariate in the Cox model with the count statistic (4), *i.e.*, the number of events that has occured. For a selected set of hazard ratios and model classes, we calculate the sample size N_{90}^{count} needed to achieve a power of 90% for the Wald test to confirm a relationship between the number of events and survival. Here, N_{90}^{count} can be compared with N_{90}, the corresponding number when fitting the correct mixture model. Obviously, fitting the correct model and using the GPS will yield smaller sample sizes than simply using the count statistic. The question is, how relevant this reduction in sample size is?

In the second scenario, we assume that the true model is not the mixture model but the independence model. In this case, in step 1 of the simulation algorithm (Fig. 3), instead of drawing at random a mixture model $\mathcal{M} = \sum_{i=1}^{2} \alpha_k \mathcal{T}_k$ we draw a star model $\mathcal{T}_{star}$. Here, each edge in the progression model is starting from the root, where edge weights are allowed to be unequal. Accordingly, in this setup, in step 4 of the algorithm, we replace the GPS with the count statistic. Thus, here we link the survival time of the patients to the count statistic. Then we fit once a mixture model with GPS as progression score and once an independence model with count statistic as score to the simulated data and again evaluate the required sample sizes.

With the simulation scenarios described above, we are able to compare the performance of the GPS and the count statistic. After running these simlations, we receive the required sample sizes for the following four cases: In the first two cases the underlying true model is the trees mixture model and the derived GPS is linked with the survival times. Once the GPS is calculated from simulated data and plugged into the Cox model, and once the count statistic is used. In the other two cases, the underlying true model is the independence model and the number of events is linked to survival. These comparisons also allow us to analyse the performance of the progression scores when the underlying models are misspecified.

4. Results

In this section, we present the estimated sample sizes needed for detecting a given relationship between the progressions scores and the survival times, for each case of the different simulation scenarios described in Section 3. First, in Section 4.1, the estimated sample sizes in case of underlying trees mixture models and the GPS as predicting variable are shown. This refers to the default simulation scenario from Section 3.1. Next, in Section 4.2, we consider other scenarios as described in Section 3.2. Here, we compare estimated sample sizes when either the GPS or the count statistic is used for predicting survival. Particularly, we point out the changes in required sample size in the case

of model misspecification. In this section we determine sample sizes needed to achieve at least a power of 90%.

Note that all sample sizes are estimated on the basis of fully uncensored survival samples. For a pre-specified proportion of censored observations, the estimators of the sample sizes can easily be adapted. In this case, the estimators simply have to be divided by the expected proportion of uncensored observations [17].

4.1 Sample Size Estimators for Mixture Models

Table 1 shows estimated sample sizes for several combinations of hazard ratio HR and number of events l in the trees mixture model. For example, for $l=6$ events and a true hazard ratio of HR $=1.4$, to confirm a significant relationship between GPS and survival time with a power of 80% and 90%, around 89 and 126 patients are needed, respectively. Repeating the simulation with identical parameter sets provides a variance of the estimates. We found that for a hazard ratio of 1.6, the sample sizes N_{80} and N_{90} vary at most by 2. For the smallest considered hazard ratio of 1.2, the largest observed difference between two simulation was 8. This relatively small variance can also be deduced from the fitted power curve in Fig. 4.

Table 1
Required Sample Sizes N_{80} (N_{90}) to Achieve a Power of 80% (90%), Dependent on the True Hazard Ratio and the Number l of Events in the Model

l	Hazard Ratio HR				
	1.8	1.6	1.4	1.2	1.1
4	46 (75)	64 (99)	106 (160)	290 (432)	1025 (1489)
5	43 (60)	56 (83)	94 (138)	256 (403)	899 (1354)
6	39 (57)	54 (78)	89 (126)	266 (397)	935 (1324)
7	41 (56)	53 (78)	92 (127)	259 (373)	917 (1345)
8	41 (58)	55 (78)	93 (130)	270 (389)	929 (1362)
9	43 (59)	60 (80)	92 (135)	282 (400)	934 (1382)
10	42 (59)	59 (82)	98 (138)	283 (412)	949 (1395)
∅	42 (61)	57 (83)	95 (136)	272 (401)	941 (1379)

In Fig. 5 we illustrate the influence of the number of events on the required sample size, for a fixed hazard ratio of 1.6. It is striking that with $l=4$ events, the required sample size is relatively high, in models with $l=6,7,8$ events the GPS has the best performance, and from 8 events onwards sample size increases again. However, the larger the hazard ratio is, the less pronounced this effect can be observed. At large, the higher the model complexity is (especially from $l=8$ onwards) the more samples are needed to obtain 80% and 90% power. The contrary behaviour for extremely small numbers of events ($l=4$) can be explained as follows. The second component $\mathcal{T}_2$ from a randomly generated model with $l=4$ events is often similar to a star. As a result, for small sample sizes, the estimated topologies differ considerably from the true ones. For larger sample sizes, the underlying true topology can be reconstructed more accurately [5].

The hazard ratio HR is a more critical parameter than the number of events. For example, when HR decreases from 1.4 to 1.2, the sample size increases approximately by a factor of 3, from around 100 to around 300. Figure 6 shows the required sample sizes depending on the hazard ratio, for models generated with $l=6$ events. The sample size decreases almost exponentially with increasing hazard

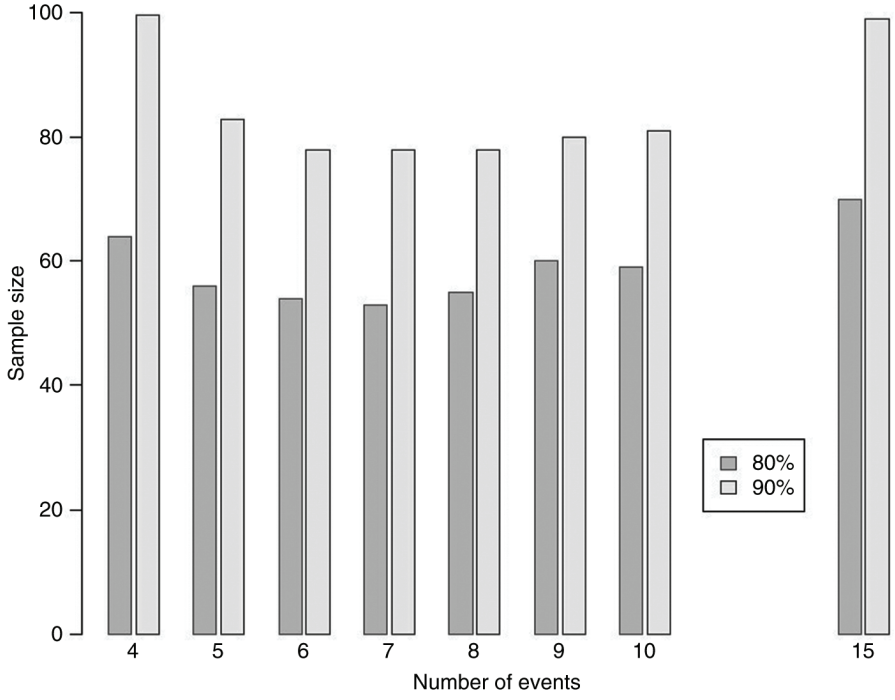

Figure 5. Required sample sizes for a hazard ratio of 1.6 and for varying numbers of events l.

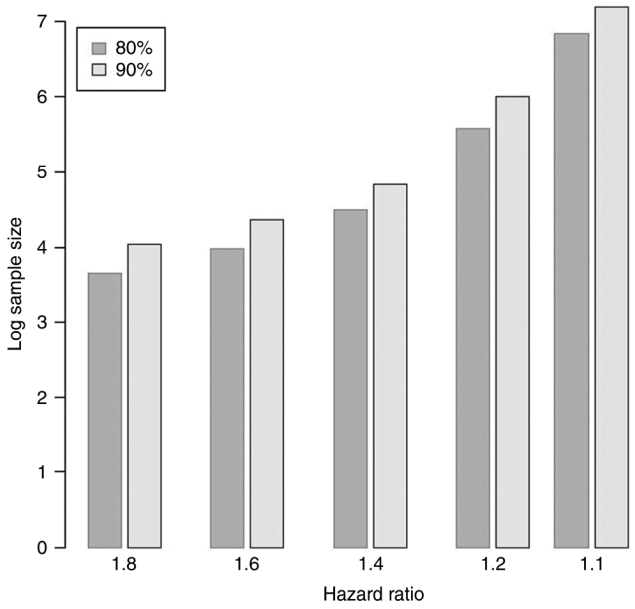

Figure 6. Required sample sizes for $l = 6$ events and for varying hazard ratio HR (log scale).

ratio. Roughly, to achieve a power of 90% instead of 80%, the sample size has to be multiplied by 1.4.

4.2 Comparison between Mixture Model and Independence Model

Table 2 shows the estimated sample sizes N_{90}^{count} and N_{90}. In this section, we consider only a power of 90%. The numbers in Table 2 can be interpreted in the following way. If the genetic events are generated from an oncogenetic trees mixture model with $l = 6$ events and survival is linked with GPS, then we need a sample size of approximately 126 to detect a hazard ratio of 1.4, when fitting the correct model. Fitting the wrong model, which in this case means just calculating the count statistic for every sample, leads to an increase in sample size of 26% to 159. Conversely, if the true model is the independence model and survival is linked to the number of events, then 17% more samples are required (96 instead of 82) when using the GPS instead of the count score.

In Table 2 it is striking that comparing trees mixture models with $l = 6$ and $l = 10$ events, the increase in required sample size is far less pronounced for the GPS compared to the count statistic. For the GPS, sample size increases by 3–9%, for the count statistic by 19–24%. Comparing models with $l = 10$ and $l = 15$ events, the increase in sample size is about 24–32% for the GPS and 40–49% for the count statistic. This results in a benefit of sample sizes of about 40% and 60%, respectively, when using the GPS instead of the count statistic. Thus, the more complex the generating mixture model is, the more sample size can be saved by using the GPS.

If the independence model is the generating model, we observe a different behaviour. As expected, in independence models, larger numbers of events are beneficial for discriminating between progression states. In this case, more events reflect more information. For a large number of events, the variance of the observed count statistic of a specific patient is small, relative to the interval of possibly observable values. However, in contrast to the case of an underlying mixture model, the relative difference in sample size between the modelling approaches is not increasing when the number of events increases. Here, the differences in power for the count statistic and for the GPS are stable; the advantage of the count statistic varies around 20%. As a matter of fact, the count statistic does not benefit as much from more events in the model as the GPS did in case of a true underlying mixture model. Furthermore, in all cases, the absolute numbers of required sample size are considerably smaller than for the trees mixture model.

For a relatively large hazard ratio of HR = 1.6, the absolute differences in sample size are rather small. However, for smaller values (*e.g.*, HR = 1.2), the required samples sizes differ in a three-digit range, when the underlying true model is the mixture model. In cancer studies, such small hazard ratios are often observed and clinically relevant. In such cases the advantage of the GPS is clearly important. For the independence model, see right side of Table 2, the advantage of the count statistic is not as distinct, and in general the absolute numbers are also considerably smaller than for the trees mixture model. Thus, we suggest to favour the advanced GPS-based modelling approach, especially in situations with limited model knowledge. The advanced model is clearly superior for true tree models and only slightly inferior for true independence models.

Table 2
Sample Size N_{90} and N_{90}^{count} (in brackets) to Reach a Power of 90%

	True Model					
l	Mixture of Trees/GPS			Independence/Count		
HR	1.2	1.4	1.6	1.2	1.4	1.6
6	397 (485)	126 (159)	78 (94)	312 (256)	96 (82)	55 (46)
10	412 (576)	138 (195)	82 (117)	181 (159)	62 (53)	37 (31)
15	516 (809)	182 (291)	102 (171)	124 (107)	46 (38)	31 (23)

5. Conclusion

In this paper, we presented a simulation study for evaluating the required sample sizes for detecting clinically relevant differences between patient cohorts with different cancer progression states. Progression states were determined by first fitting progression models and then deriving univariate progression scores from these models. Clinical relevance was determined by testing the significance of these scores in univariate Cox models via Wald tests. More precisely, we compared two different scenarios of cancer progression. The independence model assumes independence between events and is an extremely simple approach. The oncogenetic trees mixture model is far more complex. It was investigated and applied for cancer progression before and shown to represent well disease progression for various cancer types.

It turned out that misspecification of the model results in a larger increase of sample size if the data are generated from the complex model, compared to the case when they are generated from the simple model. For models based on 6 genetic events, about 20% more samples are needed if both data are generated from the simple model and progression scores are fitted with the complex model, and vice versa. However, for models with larger numbers of events, the complex model has a distinct advantage. For 10 genetic events, the saving in sample size is slightly above 40% for the true complex model and under 20% for the true simple model. The more events we have in the model, the larger is the difference in sample size. For a model based on 15 events, the saving is up to 67% for the GPS and around 20% for the simple model.

Another important observation is that the required sample size is always smaller if the simple model is the true data-generating model. Thus the relative advantage when fitting the correct simple model translates into a smaller advantage in terms of absolute numbers. A common guideline for model selection is to use the less complex model in case of limited model knowledge, to avoid overfitting of the model to the data at hand. In our scenario, this statement does not apply. In general, the absolute numbers of required sample size obtained in our simulations favour the oncogenetic trees mixture model with corresponding GPS, even in case of little model knowledge, as the power loss due to misspecification is considerably larger when fitting the simple model. Thus, a general recommendation for modelling disease progression in practical oncogenetic studies is to prefer a mixture trees model over a simple tree model.

It is important to note that the model complexity depends on the specific shape of the model. For example, for an oncogenetic trees mixture model with a non-uniform star and a tree, the effective dimension of the model can already be different from the number of parameters. For a detailed discussion of the complexity of these models, see [14]. However, in our simulations, this was not relevant as the true models were random samples from the space of all possible models.

Several alternative approaches for modelling cancer progression exist, *e.g.*, conjunctive Bayesian networks (CBN) [18]. In these models, more than one precursor event is allowed, and a fraction of data deviating from the CBN can be explained by observation errors. Here, we restrict to tree models and mixtures of oncogenetic trees. However, future research regarding competing modelling approaches can be helpful. A comprehensive overview of such disease progression models can be found in [19].

Acknowledgements

This work was supported by a grant from the German Ministry of Science and Education (BMBF) as part of the German National Genome Research Network (NGFNplus) program (01GS08104 to Jörg Rahnenführer).

References

[1] R. Desper, F. Jiang, O.P. Kallioniemi, H. Moch, C.H. Papadimitriou, and A.A. Schöffer, Inferring tree models for oncogenesis from comparative genome hybridization data, *Journal of Computational Biology*, *6*(1), 1999, 37–51.

[2] N. Beerenwinkel, J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer, Learning multiple evolutionary pathways from cross-sectional data, *Journal of Computational Biology*, *12*(6), 2005, 584–598.

[3] J. Rahnenführer, N. Beerenwinkel, W.A. Schulz, C. Hartmann, A. von Deimling, B. Wullich, and T. Lengauer, Estimating cancer survival and clinical outcome based on genetic tumor progression scores, *Bioinformatics*, *21*(10), 2005, 2438–2446.

[4] L. Tolosi, *Analysis of array CGH data for the estimation of genetic tumor progression*, Master's Thesis, Saarland University, 2006.

[5] J. Bogojeska, T. Lengauer, and J. Rahnenführer, Stability analysis of mixtures of mutagenetic trees, *BMC Bioinformatics*, *9*, 2008, 165.

[6] C. Netzer, *Statistische Analyse der Signifikanz des genetischen progressions Scores für Überlebenszeiten von Hirntumorpatienten*, Master's Thesis, TU Dortmund University, 2008.

[7] R. Ketter, S. Urbschat, W. Henn, W. Feiden, N. Beerenwinkel, T. Lengauer, W.-I. Steudel, K.D. Zang, and J. Rahnenführer, Application of oncogenetic trees mixtures as a biostatistical model of the clonal cytogenetic evolution of meningiomas, *International Journal of Cancer*, *121*(7), 2007, 1473–1480.

[8] D. Cox, Regression models and life tables, *Journal of Royal Statistical Society*, *34*(2), 1972, 187–220.

[9] J. Edmonds, Optimum branchings, *Journal of Research National Bureau Standards*, *71B*, 1967, 233–240.

[10] J. Bogojeska, *Rtreemix: Mutagenetic trees mixture models*, 2009, R package version 1.8.0.

[11] J.P. Klein and M.L. Moeschberger, *Survival analysis: Techniques for censored and truncated data*, 2nd ed. (New York: Springer, 2003).

[12] E. Vittinghoff, S. Sen, and C.E. McCulloch, Sample size calculations for evaluating mediation, *Statistics in Medicine*, *28*(4), 2009, 541–557.

[13] H. Prüfer, Neuer Beweis eines Satzes über Permutationen, *Architectural Mathematical Physics*, *27*, 1918, 742–744.

[14] J. Yin, N. Beerenwinkel, J. Rahnenführer, and T. Lengauer, Model selection for mixtures of mutagenetic trees, Genetics *Statistical Applications in Genetics and Molecular Biology*, *5*, Article 17, 2006.

[15] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2009, ISBN: 3-900051-07-0.

[16] T. Therneau and original R port by T. Lumley, *Survival: Survival analysis, including penalised likelihood*, 2009, R package version 2.35-8.

[17] D.A. Schoenfeld, Sample-size formula for the proportional-hazards regression model, *Biometrics*, *39*(2), 1983, 499–503.

[18] M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel, Quantifying cancer progression with conjunctive Bayesian networks, *Bioinformatics*, *25*(21), 2009, 2809–2815.

[19] K. Hainke, J. Rahnenführer, and R. Fried, Disease progression models: A review and comparison, Technical report, TU Dortmund University, Department of Statistics, 2011.

Biographies

Christian Netzer, born in 1981, received his diploma in statistics with a specialization in biometrics in 2008. He is a Ph.D. student at the Department of Statistics at TU Dortmund University, Germany. His research deals with statistical modelling of drug response in lung cancer patients.

Jörg Rahnenführer, born in 1971, is a professor for statistical methods in genetics and chemometrics at the Department of Statistics at TU Dortmund University, Germany. His research group works on statistical methods in bioinformatics and medicine. Currently, his main research interests are the analysis of high dimensional genetic data and survival data.