Evaluation of Treatment Effect with Paired Failure Times in a Single-Arm Phase II Trial in Oncology

Matthieu Texier,1 Federico Rotolo,1,2 Michel Ducreux,3 Olivier Bouché,4 Jean-Pierre Pignon,1,2 and Stefan Michiels1,2

1Biostatistics and Epidemiology Unit, Gustave Roussy, Université Paris-Saclay, 94805 Villejuif, France
2CESP INSERM U1018, Paris-Sud University, 94805 Villejuif, France
3Gastrointestinal Oncology Unit, Gustave Roussy, Université Paris-Saclay, 94805 Villejuif, France
4Gastrointestinal Unit, University Hospital, Reims, France

Correspondence should be addressed to Stefan Michiels; stefan.michiels@gustaveroussy.fr

Received 2 May 2017; Accepted 28 November 2017; Published 11 January 2018

Academic Editor: Lev Klebanov

Copyright © 2018 Matthieu Texier et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In early phase clinical trials of cytotoxic drugs in oncology, the efficacy is typically evaluated based on the tumor shrinkage. However, this criterion is not always appropriate for more recent cytostatic agents, and alternative endpoints have been proposed. The growth modulation index (GMI), defined as the ratio between the times to progression in two successive treatment lines, has been proposed for a single-arm phase II trials. The treatment effect is evaluated by estimating the rate of patients having a GMI superior to a given threshold. To estimate this rate, we investigated a parametric method based on the distribution of the times to progression and a nonparametric one based on a midrank estimator. Through simulations, we studied their operating characteristics and the impact of different design parameters (censoring, dependence, and distribution) on them. In these simulations, the nonparametric estimator slightly underestimated the rate and had slightly overconservative confidence intervals in some cases. Conversely, the parametric estimator overestimated the rate and had anticonservative confidence intervals in some cases. The nonparametric method appeared to be more robust to censoring than the parametric one. In conclusion, we recommend the nonparametric method, but the parametric method can be used as a supplementary tool.

1. Introduction

In oncology, if a new treatment is found to be acceptably safe in a phase I clinical trial, it can be tested in a phase II trial to look for evidence of efficacy. The type of response or benefit to evaluate depends on the goals of the treatment; in advanced cancer trials, the most used endpoints are related to the change of the size of the lesion or its disappearance. Historically, the tumor shrinkage was the primary endpoint in phase II trials for cytotoxic cancer drugs. Since the 90s, cytostatic drugs, which are supposed to modulate the tumor growth without causing immediate shrinkage, are being developed. Thus, Von Hoff [1] and Mick et al. [2] advocated for rather evaluating the time to progression (TTP) as the primary endpoint in a one-stage design. Since patients being offered phase II studies of new agents have typically failed a previous regimen, then all first progressions are observed and TTP before experimental treatment, say $TTP_1$, is known for all the patients enrolled. Conversely, the TTP after the experimental agent, $TTP_2$, may or may not be censored at the time of the analysis. As the TTP is highly variable across patients and the degree of correlation between the paired failure times is a key feature, Von Hoff [1] proposed to evaluate the growth modulation index (GMI = $TTP_2/TTP_1$) instead, so that each patient serves as his/her own historical control. Von Hoff [1] assumed a null ratio value of 1 and that the GMI needs to be greater than 1.33 for a new regimen to be considered effective in delaying progression. Mick et al. [2] argued that because patients enter a new treatment line after a new progression, the prognosis is expected to be poorer than at the previous treatment line. Thus, because of the natural history of the disease, one expects that in general $TTP_2$ is shorter than $TTP_1$, which would indicate a null ratio value.
smaller than 1 and that a GMI superior to 1 is enough for considering a new regimen as effective.

Some authors have started employing the GMI as primary endpoint. At the time of writing (April 2017), there were a total of ten oncology trials registered in the European Union Clinical Trial Register and eleven oncology trials registered on the https://www.clinicaltrials.gov website as using GMI. For example, Von Hoff et al. [3] used the GMI to measure the activity of a targeted therapy selected by molecular profiling in patients having failed all effective treatments. Eighteen out of 66 patients (27%) had a progression-free survival (PFS) ratio superior to 1.33 (95% confidence interval: [17%; 38%]). Several others published trials [4–6] used a GMI-based approach to assess the activity of second-line treatments, but the estimation did not account for patients with censored times to progression. Only a recent secondary analysis of the SHIVA trial estimated the PFS ratio by Kaplan–Meier curves [7].

Before the GMI can be regularly used as primary endpoint in phase II studies, we need appropriate statistical methods and detailed knowledge of its statistical characteristics. In the present paper, we present methods to estimate the proportion of patients having a GMI greater than a given threshold by handling censored observations, we explore their operating characteristics via simulations and we show an application on a real data set. Such a motivating study in advanced colorectal cancer is presented in Section 2. Section 3 summarizes the statistical methodology to estimate the probability that the GMI is higher than a given threshold. Section 4 presents a simulation study to investigate parameters which could influence the performance of the estimators. Finally, in Section 5, we apply the presented methods to real data. Section 6 discusses the findings.

2. Motivating Example

The FFCD 2000-05 trial [8, 9] was a randomized trial conducted by the French Federation of Digestive Oncology, which included 410 patients with advanced colorectal cancer. It was a phase III trial comparing a sequential (S) arm to a combination (C) arm. Patients in arm S were treated with 5-fluorouracil and leucovorin (LV5FU2) in first line, then with FOLFOX (LV5FU2 + oxaliplatin) in second line, and then with FOLFIRI (LV5FU2 + irinotecan) in third line. Patients in arm C were treated directly with FOLFOX in first line and then with FOLFIRI (LV5FU2 + irinotecan) in third line. The times to progression in the first, second, and third treatment lines were recorded with FOLFIRI (LV5FU2 + irinotecan) in third line. Patients in arm C were treated directly with FOLFOX in first line, then with FOLFIRI (LV5FU2 + irinotecan) in third line. Patients in arm S were treated with 5-

The FFCD 2000-05 trial [8, 9] was a randomized trial conducted by the French Federation of Digestive Oncology, which included 410 patients with advanced colorectal cancer. It was a phase III trial comparing a sequential (S) arm to a combination (C) arm. Patients in arm S were treated with 5-fluorouracil and leucovorin (LV5FU2) in first line, then with FOLFOX (LV5FU2 + oxaliplatin) in second line, and then with FOLFIRI (LV5FU2 + irinotecan) in third line. Patients in arm C were treated directly with FOLFOX in first line and then with FOLFIRI (LV5FU2 + irinotecan) in second line. The times to progression in the first, second, and third treatment lines were recorded with FOLFIRI (LV5FU2 + irinotecan) in third line. Patients in arm C were treated directly with FOLFOX in first line, then with FOLFIRI (LV5FU2 + irinotecan) in third line. Patients in arm S were treated with 5-

the frailty term is a random variable with probability density

\[
\theta = \frac{1}{\theta} - 1
\]

\[
\Gamma(1/\theta)
\]

\[
f(u) = \frac{\theta^{-1/\theta} u^{1-\theta} \exp(-u/\theta)}{\Gamma(1/\theta)}, \tag{2}
\]

for patient \(i \in \{1, \ldots, n\}\) at treatment line \(j \in \{1, 2\}\), and where \(h_0(t)\) is the treatment line-specific baseline hazard function, \(u_i\) the frailty term for the patient \(i, x_j\) the vector of his/her covariates in the \(j\)th treatment line, and \(\beta_j\) the vector of regression coefficients. In a gamma frailty model, the frailty term is a random variable with probability density function:

\[
\begin{align*}
h_{ij}(t | u_i) = h_0(t) u_i \exp(x_i^T \beta_j), \tag{1}
\end{align*}
\]
where $\Gamma(\cdot)$ is the gamma function. This distribution corresponds to a gamma distribution with mean and variance equal to $1$ and $\theta$. Shared frailty models allow estimating the intrapatient dependence via Kendall’s $\tau$, which is a rank correlation measure of the concordance between time pairs. In the case of a gamma frailty model, Kendall’s $\tau$ is equal to $\theta/(\theta+2)$ and can thus be estimated by plugging in the estimate of $\theta$. Different distributions can be assumed for the baseline argument is true and $0$ otherwise. The maximum rank is the rank of $R_{ji}$ among the $2n$ pooled $L_{ji}$’s:

$$
\max_{ji} : L_{j(1)} \leq L_{j(2)} \leq \cdots \leq L_{j(\text{max})} \leq R_{ji} \leq L_{j(\text{max}+1)} \\
\leq \cdots \leq L_{j(2n)}.
$$

(5)

Now, the midrank $M_{ji}$ is the midpoint of the minimum and the maximum rank:

$$
M_{ji} = \frac{\min_{ji} + \max_{ji}}{2}.
$$

(6)

To estimate $\hat{S}_{GMI}(\delta)$, we replace $\text{TTP}_{ji}$ with $\text{TTP}_{ji}' = \delta\text{TTP}_{ji}$ and compute the midranks $M_{ji}'$ of $\text{TTP}_{ji}'$ and $M_{2j}$ of $\text{TTP}_{2j}$ to obtain the $n$ pairs of midranks $(M_{ji}'; M_{2j})$. Finally, the estimate of the probability of interest is as follows:

$$
\hat{S}_{GMI}(\delta) = \frac{1}{n} \sum_{j=1}^{n} I(M_{2j} \geq M_{ji}') ,
$$

(7)

with $I(\cdot)$ being the indicator function which takes value $1$ if its argument is true and $0$ otherwise.

3.2.2. Parametric Method. In this approach, a parametric probability distribution is assumed for the GMI, so that the probability of interest can be easily derived as a function of the estimated distribution parameters. Let us assume that, conditionally on a frailty term $u_i$, TTP and TTP have Weibull marginal distributions $W(a; b, u_i)$ and $W(a; b, u_i)$ with a common shape parameter $a$:

$$
f_j \left( x; a, b_j \mid u_i \right) = a (u_i b_j)^{-a} x^{a-1} \exp \left\{ - \frac{x}{(u_i b_j)^a} \right\}.
$$

(8)

Then, Owen [15] showed that the ratio $\text{TTP}_2/\text{TTP}_1$ follows a log-logistic distribution,

$$
f(\delta; a, \kappa) = \kappa a^a \delta^{-a} \left( 1 + (\delta \kappa)^a \right)^{-2} , \quad \delta \geq 0 ,
$$

(9)

with $\kappa = b_2/b_1$, which does no longer depend on the shared frailty $u_i$.

By using this distribution, we can obtain maximum likelihood estimates of the distribution parameters and directly derive the probability of interest from the survival function:

$$
S(\delta; \hat{a}, \hat{\kappa}) = \left( 1 + (\delta \hat{\kappa})^a \right)^{-1}.
$$

(10)

Parameter estimates were computed using the survreg function in the R package survival.

R code of the two methods is available for download on https://github.com/Oncostat/TTPratio.

4. Simulation Study

4.1. Simulation Design. We designed a simulation study to evaluate the influence of the design parameters on the two estimators of $S_{GMI}(\delta)$. We varied (i) the dependence between the two successive times to progression via Kendall’s $\tau$, (ii) the shape $a$ of the distribution of TTP, (iii) the relative effect $e$ of the second-line treatment as compared to the first-line treatment, and (iv) the censoring rate $r$ for TTP.

4.1.1. Data Generation. First, for given values of the parameters of interest, we generated a frailty term $u_i$ for each patient using random values from a gamma distribution with density given in Section 3. Due to the link between $\tau$ and $\theta$, for a given $\tau$, we could fix $\theta = 2\tau/(1 - \tau)$. Three values of $\tau$ were used in our simulation: 0.1, 0.2, and 0.3.

Then, we generated times to first and second progressions from Weibull distribution with density:

$$ f_j(x; a, b_j | u_i) = a (u_i b_j)^a x^{a-1} \exp \left\{ \frac{-x}{(u_i b_j)} \right\}, \quad j = 1, 2. \tag{II} $$

For the shape parameter $a$, common to the two distributions, we considered three values: 0.5, 1, and 2. A shape of $a = 0.5$ represents a metastatic disease with a median of TTP$_2$ greater than 15 months, whereas a shape of $a = 2$ corresponds to a more aggressive disease (median of TTP$_1$ close to 6 months).

The scale parameter was different for the two distributions: $b_2 = b_1 e$, where $e$ is the median of TTP$_2$/TTP$_1$. We considered three values for $e$: 0.77, representing inactivity of the second-line treatment; 1, representing an equivalence of the two treatments; and 1.33, representing efficacy according to the definition of Von Hoff [1].

Independent and noninformative censoring was introduced by taking the minimum between TTP$_2$ and a random uniform variable. Desired censoring rates (10% and 40%) were obtained by controlling the support of the uniform distribution.

We performed 10,000 simulations for each of the 54 scenarios defined by $a$, $e$, $\tau$, and a censoring rate. The statistical properties of the parametric and nonparametric estimators were evaluated in terms of the mean bias, the average standard error, and the empirical standard error, the latter being defined as the standard deviation of the 10,000 estimates.

4.2. Results. The results of the simulations are summarized in Figure 2 (see Supplementary Tables A1–A6 for detailed results). The nonparametric method underestimated the probability of interest in 51/54 scenarios, but the mean bias was low in general, ranging across scenarios from −0.062 to 0.001 (median: −0.006). On the contrary, the parametric method always overestimated the probability of interest, but the mean bias was low as well, ranging across scenarios from 0.009 to 0.082 (median: 0.028). With a censoring rate of 10% and considering all scenarios, the nonparametric estimator was slightly less biased than the parametric estimator (median absolute bias: 0.003 versus 0.014): the absolute bias of the nonparametric estimator was at most 0.011 and the bias of the parametric estimator was at most 0.018. The bias of the parametric estimator increased with increasing censoring rate; across all scenarios with censoring rate of 40%, its median absolute bias was 0.069. The nonparametric estimator was more robust to censoring with a median absolute bias of 0.018 for 40% of censoring.

Both estimators were robust to changes in dependence, shape parameter $a$, and treatment effect $e$. Considering all scenarios, the average (over the 10,000 replicates) of the estimated standard error (ASE) via the nonparametric method was greater than or equal to the empirical standard error (ESE). This suggests that the nonparametric confidence intervals are more conservative than their nominal level. For the parametric estimator, on the contrary, when we considered second-line treatment inactivity (median GMI = 0.77) and 40% of censoring, the ASE was smaller than the ESE. This means that parametric confidence intervals are too liberal under the null hypothesis.

5. Application to the FFCD 2000-05 Trial

In this section, we illustrate the presented methodology to the data of the FFCD2000-05 trial (see Section 2 and Figure 1). As discussed previously, we will consider situations 1 and 4 only, in which the same couple of treatments are contrasted. The ratio TTP$_2$/TTP$_1$ could be evaluated on 129 patients in situation 1. The ratio TTP$_3$/TTP$_2$ could be evaluated on 92 patients in situation 4. A total of 15 patients (12%) had their TTP$_2$ censored in situation 1 and 13 patients (14%) had their TTP$_3$ censored in situation 4.

5.1. Dependence between TTP1 and TTP2. As discussed in Section 3, we estimated Kendall’s $\tau$ by modeling the risks of progression via shared frailty models. Weibull distributions were assumed for the baseline hazard functions. The use of a gamma distribution for the frailty term was justified by a preliminary study comparing the Akaike Information Criterion (AIC) of the model with gamma and inverse Gaussian frailty distributions. The positive stable frailty distribution was considered too, but it was also discarded due to the lack of numerical convergence. In all four situations, the model with gamma distribution had the smallest AIC.

In situation 1, the estimated Kendall’s $\tau$ was 0.195, a relatively low correlation. In situation 4, that is FOLFIRI versus FOLFIRI again, but after a first line with LV5FU2, the estimated Kendall’s $\tau$ was slightly higher: 0.225. Even weaker dependence was estimated for situation 3 ($\tau = 0.152$) and situation 2 ($\tau = 0.142$). Overall, these values fell in between the first and second values of $\tau$ considered in our simulations: 0.1 and 0.2.

5.2. Estimation of $\hat{S}_{GMI}(\delta)$. To apply the parametric estimation method for $\hat{S}_{GMI}(1)$ described in Section 3, we assumed Weibull distributions of times to progression with common shape parameter. This assumption was needed in order to assume a log-logistic distribution for their GMI. Thus, we fitted the Kaplan-Meier estimates of the GMI and compared them to the maximum likelihood log-logistic survival curves to informally check the appropriateness of the parametric assumption. Figure 3 shows, for situations 1 and 4, the Kaplan-Meier estimates of the GMI with the estimated log-logistic survival curves. This distribution seems to fit quite well to the data.

In situation 1, the estimated probability that the GMI $\geq 1$ was $\hat{S}_{GMI}(1) = 0.21$ with the parametric estimator (95% Confidence interval: [0.14; 0.29]) and $\hat{S}_{GMI}(1) = 0.24$ with the
Figure 2: Probability $\hat{S}_{\text{GMI}}(\delta = 1)$ of GMI being greater than 1 estimated in the simulation study via the parametric (black) and nonparametric (red) methods. Normally approximate 95% confidence intervals using the empirical standard error.
nonparametric estimator (95% CI: [0.17; 0.31]). In situation 4, comparing the same two treatments after an LV5FU2 line, the estimated probability was 0.24 (95% CI: [0.15; 0.33]) with the parametric estimator and 0.27 (95% CI: [0.18; 0.36]) with the nonparametric estimator. These results suggest that the sequence “FOLFOX in first line/FOLFIRI in second line” leads to a shortened time to progression: FOLFIRI’s activity in second line seems inferior to FOLFOX’s activity in first line.

Table 1 shows the different estimations for the other situations, too. The activity of FOLFOX in second line seems to be comparable to the activity of LV5FU2 in first line for patients in the arm S.

6. Discussion

The growth modulation index (GMI) is more and more used to evaluate the treatment effect in single-arm phase II trials. An increasing number of clinical trials employ the GMI and the European Medicine Agency (EMA), in its “Guideline on Evaluation of Anticancer Medicinal Products in Man,” admits its utilization for a comparison between two successive therapies [16]. By choosing an adequate threshold $\delta$ (0.77, 1, or 1.33), the estimated probability of interest $S_{GMI}(\delta)$ is a practical measure of the proportion of patients for whom two successive lines of treatment are ineffective, equivalent, or effective.

In this article, we evaluated two ways to estimate $S_{GMI}(\delta)$ and we investigated how the design parameters had an impact on these estimators. The censoring rate had an impact on the parametric and nonparametric estimators, respectively. In our simulations, the nonparametric method was more robust to high censoring rates, but the average bias was small in any case. Thus, the use of this method in phase II studies could represent substantial time savings for the analysis when the disease in question progresses slowly over time. Von Hoff [1] showed the key role of dependence between the paired times to progression, but in our study this parameter did not have a noticeable impact neither on the bias nor on the
empirical standard error. The few published clinical trials that used the GMI as a criterion of activity reported a rather low correlation of the paired time to progression. However, in some of them, such a low correlation may be due to the heterogeneity of the first-line treatment (different nature of chemotherapy) or to the localization of the tumor. In Penel et al. [17], for instance, the analysis did not account for the heterogeneity of the subtypes of sarcoma. Further studies are needed to detect the influence of cancer localization on the different design parameters. To date, it is not well known in which cancer types the intrapatient correlation is the strongest.

There are practical limitations to the use of GMI in a phase II study. The collection of PFS or TTP measurements for each patient has to be very precise and homogeneous between patients and, if the case, between centers. The frequency of the follow-up evaluations affects the estimation of TTP and PFS [18]. This issue should be considered carefully in the design and the conduct of a trial employing this endpoint.

In clinical practice, patients can interrupt the first-line treatment for many reasons such like toxicity occurrence. In that case, they can enter the second line without a progression, causing TTP₁ being censored. For these patients, TTP₂/TTP₁ is left censored (the GMI is unknown but an upper bound is known) and inferential methods can be adapted to that situation. Nevertheless, if both TTP₁ and TTP₂ are censored, neither an upper nor a lower bound is known and the observation is noninformative. However, one could argue that phase II studies using GMI as the primary endpoint should enroll only patients who have failed previous treatment and thus exclude cases where TTP₁ is censored. A third approach would be to consider also treatment interruptions due to toxicity as events in a treatment-failure perspective. Eventually, the most appropriate approach will depend on clinical considerations about whether the new treatment is intended for patients recurring only, or for any interruptions of the previous treatment, whatever the cause.

In our simulations, nonparametric and parametric methods, when biased, had biases in opposite directions. We recommend using the nonparametric method to estimate the proportion of patients having a GMI superior to a threshold because it is more conservative. Nevertheless, the parametric method can more easily deal with interval censoring, which is an inherent issue with progression-free survival data [19]. Consequently, the parametric method can be used as a supplementary tool.

Conflicts of Interest
The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments
The authors thank the FFCD-2000-05 trial investigators for their participation.

Supplementary Materials
Online-only Supplementary Materials detail the numerical results for each scenario of the simulation study. Mean bias, average standard error, and empirical standard error of nonparametric and parametric estimation are presented in Tables A1–A6. Table A1: bias, average standard error, and empirical standard error of the nonparametric estimator of the probability in $S_{\text{GMI}}(\delta = 1)$ with equivalent treatments (median (GMI) = 1). Table A2: bias, average standard error, and empirical standard error of the parametric estimator of the probability in $S_{\text{GMI}}(\delta = 1)$ with equivalent treatments (median (GMI) = 1). Table A3: bias, average standard error, and empirical standard error of the nonparametric estimator of the probability in $S_{\text{GMI}}(\delta = 1)$ with an inactive second-line treatment (median (GMI) = 0.77). Table A4: bias, average standard error, and empirical standard error of the parametric estimator of the probability in $S_{\text{GMI}}(\delta = 1)$ with an inactive second-line treatment (median (GMI) = 0.77). Table A5: bias, average standard error, and empirical standard error of the nonparametric estimator of the probability in $S_{\text{GMI}}(\delta = 1)$ with an active second-line treatment (median (GMI) = 1.33). Table A6: bias, average standard error, and empirical standard error of the parametric estimator of the probability in $S_{\text{GMI}}(\delta = 1)$ with an active second-line treatment (median (GMI) = 1.33). (Supplementary Materials)

References


Submit your manuscripts at www.hindawi.com