Note on the Role of the Placebo Group in The Short-Term and Long Term Hazard Ratio Model

Philippe Flandre, John O'Quigley

Note on the Role of the Placebo Group in The Short-Term and Long Term Hazard Ratio Model

Philippe Flandre a, and John O'Quigley b

December 9, 2021

We read with interest the paper, "Improving testing and description of treatment effect in clinical trials with survival outcomes" by Yang [1] and related works published by Yang and Prentice [2, 3]. Alternatives to absence of treatment effect on survival are most commonly framed within the proportional hazards model structure. When broader alternatives of a non-proportional hazards nature are of interest it is known that the high efficiency of the log-rank test can be lost. One test that focuses on this issue is the test of Yang and Prentice [3], based on the short-term and long-term hazard ratio model. The authors argue that the test is able to detect departures from a null hypothesis of no effect against quite broad alternatives. We recall the model on which this test is based and the test itself. Theoretical work and simulations appear to show the test to work well. However, it is important to underline that the test can only be used when the placebo group is clearly defined. Failing to correctly identify which group is the placebo group can have serious consequences and lead to erroneous conclusions. The results may not be interpretable. In situations where we wish to compare Treatment A with Treatment B, neither of which can be considered to correspond unambiguously to the placebo group, our advice would be to avoid this test altogether. We give two examples before concluding that the validity of this test hinges crucially upon a clear and sharp definition of the placebo group.

1 Short term and long term hazard ratio model

Consider a sample of n observations for which, for the subject i, $i \in \{1, \ldots, n\}$, the observed survival is denoted X_i. The true survival, T_i may not be observed due to a censoring variable, C_i and we have: $X_i = \min(T_i, C_i)$. These different outcomes can be distinguished and we have; $\delta_i = I_{T_i \leq C_i}$, where I is an indicator function. It is common to broaden the setting to include covariables but we limit our discussion to the 2-sample problem via the use of the group indicator variable Z_i, taking the value 1 if the ith subject is from the first group and is zero otherwise. The data can be summarized as: $\{(X_i, \delta_i, Z_i)\,,\ i = 1, \ldots, n\}$, where, from a sampling viewpoint, the n triplets are taken to be independent and identically distributed copies of the random variable (X, δ, Z), with $X = \min(T, C)$ and $\delta = I_{T \leq C}$. The hazard, or instantaneous risk, function is defined by, $\lambda(t) = \lim_{h \downarrow 0} P\left(T_i \in [t, t+h[\mid T_i \geq t\right)h, \qquad t \geq 0.$ and the survivorship function is $S(t) = \exp\left(-\int_0^t \lambda(s)ds\right), \quad t \geq 0$. Our problem concerns the com-

parison of two groups with hazards, $\lambda_T(t)$ and $\lambda_P(t)$ and survivorship functions, $S_T(t)$ and $S_P(t)$ respectively. Yang and Prentice [2] suggested modelling any potential differences between these two hazard rates by,

$$\lambda_T(t) = \theta_1\theta_2\theta_1 + (\theta_2 - \theta_1)S_P(t)\lambda_P(t), \quad 0 \leq t \leq \mathcal{T}, \tag{1.1}$$

where θ_2 and θ_1 are two positive parameters. Special cases of this model arise under different specifications for the parameters, in particular we obtain proportional hazards when $\theta_2 = \theta_1$ and the proportional odds model when $\theta_1 = 1$. The idea of the model is to increase the flexibility of all the special cases, in particular the proportional hazards formulation, and covers such cases as crossing hazards for example. The model's parameters are such that $\theta_1 = \lim_{t\downarrow 0} \lambda_T(t)/\lambda_P(t)$ and $\theta_2 = \lim_{t\uparrow\mathcal{T}} \lambda_T(t)/\lambda_P(t)$. In this way we can argue that θ_1 represents the relative risk in the short term while θ_2 would correspond to the relative risk in the long term. Estimation is carried out by working with the partial likelihood that corresponds to the model. The authors show how to introduce covariates into the model but, here, we limit ourselves to the 2-group problem only.

Recall that the weighted log-rank test is based on the statistic LW_n where;

$$LW_n = \sum_{j=1}^{k_n} W_n(t_j)\mathcal{Z}(t_j) - \mathcal{E}_0(Z \mid t_j)\sqrt{\sum_{i=1}^{k_n} W_n(t_i)^2\mathcal{V}_0(Z \mid t_i)}, \tag{1.2}$$

and in which W_n is a positive weight function defined on the interval $[0, 1]$ and, for technical reasons $\mathcal{F}_t^*$-predictable. The choice of weights allows us to obtain tests with greater power for specific departures from a null hypothesis of absence of effect. When the true effects, $\beta(t)$, have a particular form, the power of our test will be maximized by a choice of weights that reflect this form. The proposal of Yang and Prentice [3] is based on an adaptive strategy that reduces to a test close to the log-rank test when the risks show themselves to be proportional. The test leans on the class of models proposed by these same authors [2] and defined by Equation 1.1. We would hope for the test to have more generality than either the log-rank test or a test based on a proportional odds model since both of these fit into this model structure as special cases. Their proposal is to use both the structure of the above weighted log-rank test and that of the above model by taking as weights $LW_{1,n}$ and $LW_{2,n}$ having respective weights $W_n^1(t) = \hat{\lambda}_T(t)/\hat{\lambda}_P(t)$ and $W_n^2(t) = \hat{\lambda}_P(t)/\hat{\lambda}_T(t)$, where the estimators of the hazards in the groups receiving placebo, $\hat{\lambda}_P$ or the treatment, $\hat{\lambda}_T$ are based on the model (1.1).

The difficulty with Equation 1.1 becomes quickly apparent upon inspection. If we use the equation to express placebo in terms of treatment, rather than treatment in terms of placebo, then we find ourselves with an entirely different model. This is unlike anything we are used to seeing in say linear regression, logistic regression or Cox regression where, if we change the group definition from (0,1) to (1,0), then, while the sign of the regression coefficient may change, its magnitude will remain the same. Any statistical test will be unchanged. As we will see in the examples of the following section, not only can such a change in coding lead to changes in the results, these changes can be very large. Indeed, there are examples where the method will lead to a conclusion of a good fit with one coding but a very poor fit with another. All interpretation is lost if we allow

for a change of group coding. For this reason the placebo group needs to be unambiguously defined and it needs to be systematically attributed the coding zero.

2 Practical examples of changes in coding

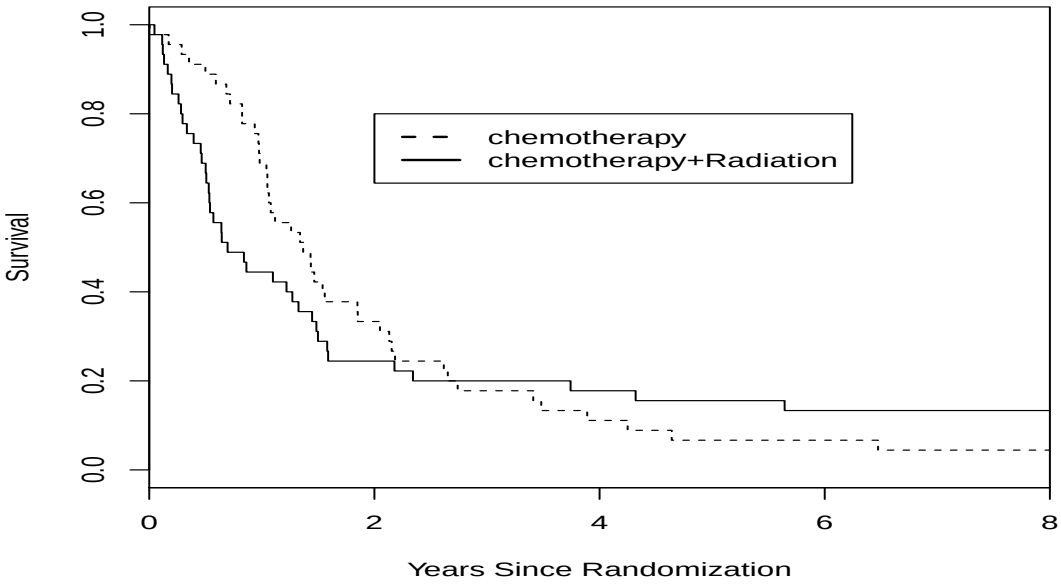

Figure 1: Kaplan-Meier curves of the trial conducted by the Gastrointestinal Tumor Study Group comparing chemotherapy alone with chemotherapy and radiation therapy in the treatment of locally unresectable gastric cancer.

Our first illustration comes from the paper of Yang [1] and corresponds to the second example from that paper. The Gastrointestinal Tumor Study Group conducted a trial comparing chemotherapy alone with chemotherapy and radiation therapy in the treatment of locally unresectable gastric cancer [4]. Each treatment arm had 45 patients, of which 2 observations from the chemotherapy group and 6 observation from the combination group were censored. These data are also provided publicly with the YPmodel package. The coding for the chemotherapy group was represented by $Z = 0$ whereas for the coding for the chemotherapy and radiotherapy group was given by $Z = 1$. The data were analyzed by Yang [1] and the log-rank test has a p value of 0.64, a value which remains unchanged by a change in coding. The adaptative weighted log-rank test (LRAD) gives a p value of 0.035 using the resampling method with 1 million repetitions. The residual based goodness-of-fit test gives a $p-$value of 0.10 while the contrast based goodness-of-fit results in a $p-$value of 0.60 [5]. Now, if we reverse the coding so that $Z = 1$ for chemotherapy alone while $Z = 0$ for chemotherapy and radiotherapy, then we find a $p-$value of 0.08 for the LRAD test, in contrast to the 0.035 found with the original coding, indicating no sig-

nificant difference between the two randomized groups (Table). As concerns the goodness-of-fit tests, we find, under the new coding, that the residual based goodness-of-fit test gives a $p-$value of 0.64 while the contrast based goodness-of-fit results in a $p-$value of 0.35.

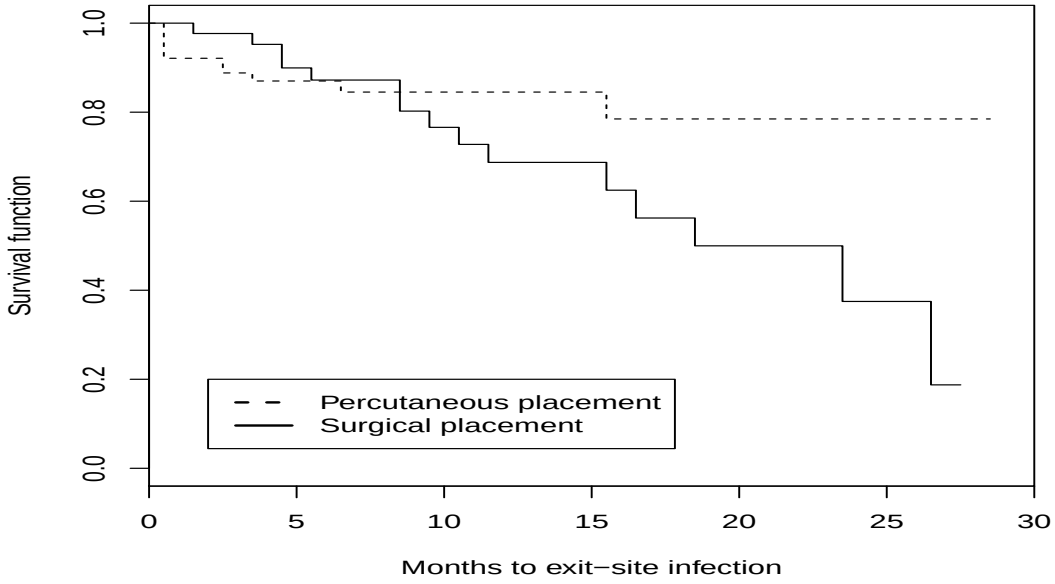

Figure 2: Kaplan-Meier curves for the kidney dialysis trial comparing surgically placed catheter with percutaneous placement of catheter.

Our second illustration comes from a trial on kidney dialysis. The details can be found in Lin and Xu [6]. The purpose of this example is to show that in a situation where there is no clearly defined placebo group then we are not able to interpret the results. This trial was designed to assess the time to first exit-site infection (in months) in patients with renal insufficiency. There are two treatment groups. Forty-three patients utilized a surgically placed catheter (Group I), and 76 patients utilized a percutaneous placement of their catheter (Group II). Catheter failure was the primary reason for censoring. Briefly, cutaneous exit-site infection was defined as a painful cutaneous exit site and positive cultures, or peritonitis. We are interested in testing if there is a statistically significant difference in the time to cutaneous exit-site infection between patients whose catheter was placed surgically and the patients who had their catheters placed percutaneously. As just mentioned, it is not at all clear that either of these groups could be unambiguously taken to be the placebo group.

The log-rank test leads to $p = 0.11$, a value which, once again, is unaltered by any change in coding. In the first instance, for patients who utilized a surgically placed catheter the coding $Z = 1$ was used whereas we took $Z = 0$ for patients who utilized a percutaneous placement of their catheter. In this case, the LRAD is not statistically significant with $p = 0.14$. The residual-based goodness-of-fit test produced the value $p = 0.009$ indicating a strong departure from the proportional hazards assumption. If we reverse the coding so that now $Z = 0$

Table 1: Analysis using the short-term and long-term model in two examples. Gastric data: coding 1/0 chemotherapy group ($Z = 0$) and chemotherapy + radiotherapy group ($Z = 1$); inverse for coding 0/1. Kidney dialysis data: coding 1/0 surgically placed catheter ($Z = 1$) and percutaneous placement of catheter ($Z = 0$); inverse for coding 0/1.

		p values	
Study	coding	LRAD	goodness of fit test
Gastric	1/0	0.04	0.10
	0/1	0.08	0.64
Kidney Dialysis	1/0	0.14	0.01
	0/1	0.04	0.41

for patients who utilized a surgically placed catheter and $Z = 1$ for patients who utilized a percutaneous placement of their catheter we find very different results. Under the reversed coding, the LRAD test indicates a significant difference between the two groups with $p = 0.04$. Unlike the goodness-of-test obtained under the orginal coding, under the new coding the residual-based test does not indicate a departure from the proportional hazards assumption with $p = 0.41$. In this case both conclusions - (1) there exists a difference between the treatments and (2), the proportional hazards assumption appears reasonable, are turned on their head by a simple change of coding. Since we do not know which group ought be considered the placebo group, we are not able to interpret the results of the tests.

References

[1] Yang S, Improving testing and description of treatment effect in clinical trials with survival outcomes. *Statistics in Medicine.* 2019; **38**:530-544.

[2] Yang S, Prentice RL. Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika.* 2005; **92**:1-17.

[3] Yang S, Prentice RL. Improved logrank-type tests for survival data using adaptive weights. *Biometrics.* 2010; **66**:33-38.

[4] Gastrointestinal Tumor Study Group; A Comparison of Combination Chemotherapy and Combined Modality Therapy for Locally Advanced Gastric Carcinoma. *Cancer.* 1982; **49**:1771-1777.

[5] Yang S, Zhao Y. Checking the Short-Term and Long-Term Hazard Ratio Model for Survival Data. *Scandinavian Journal of statistics.* 2012; **39**:554-567.

[6] Lin X, Xu Q. A new method for the comparison of survival distributions. *Pharmaceuticals statistics.* 2010; **9**:67-76.