

# Rejoinder for “Meta-analysis for Surrogacy: Accelerated Failure Time Models and Semicompeting Risks Modeling”

Debashis Ghosh,<sup>1,\*</sup> Jeremy M. G. Taylor,<sup>2,\*\*</sup> and Daniel J. Sargent<sup>3,\*\*\*</sup>

<sup>1</sup>Departments of Statistics and Public Health Sciences, Penn State University, University Park, Pennsylvania 16802, U.S.A.

<sup>2</sup>Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

<sup>3</sup>Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota 55905, U.S.A.

\**email:* ghoshd@psu.edu

\*\**email:* jmgt@umich.edu

\*\*\**email:* sargent.daniel@mayo.edu

## 1. Introduction

We would first like to express our appreciation to coeditor David Zucker and the Associate Editor for organizing this discussion. We also thank the discussants for their comments on our article. They have raised many excellent points, and in our response, we only deal with a subset of them.

Geert Molenberghs (M) and John O’Quigley and Philippe Flandre (OF) accurately describe the methodology in our article as joint regression and association modeling of the surrogate and true endpoints in which a constraint is placed on the type of data that are used (the “wedge” region). As OF noted, this constraint leads to the multistate model of Fix and Neyman (1951). This data structure complicates the standard estimation procedures that were developed by Burzykowski, Molenberghs, and Buyse (2005, Ch. 11). However, much of the model formulation is very similar to what was described there. The constraints in our approach can be viewed as a different model for the error distribution. Our focus is not on predictions, as advocated by Edward Korn (K), partly because it is very hard with censored data to estimate the intercept parameter in a linear model well without making strong assumptions (Ying, Jung, and Wei, 1995). K is suspicious of the standard errors in our semicompeting risks analysis, but our application of the methodology to data from Ghosh (2009) yielded essentially identical answers to those reported there (data not shown). An implication of the artificial censoring strategy we propose here is that we are throwing away information on recurrences. Consequently, the standard errors for the treatment effects on the surrogate endpoint will increase in our approach relative to approaches that do not throw away that information (e.g., the analyses in Table 1 of K’s discussion). An implication of the semicompeting risks approach will be that the magnitude of the treatment effect on the surrogate endpoint will be less than or equal to that on the true endpoint because of the wedge constraint.

Vance Berger, Grant Izmirlian, and Diana Knoll (BIK) and K criticize us with respect to composite endpoints. There are two issues here. The first is whether or not composite endpoints should be used for assessing treatment effects in clinical trials. BIK and K strongly advocate for composite endpoints such as disease-free survival in oncology trials. Since disease-free survival is arguably a meaningful clinical endpoint, we agree with BIK and K’s point if the goal is simply to understand the treatment effect. However, a second goal is attempting to understand the association between the surrogate endpoint with the true endpoint. As we discussed in the article, this is problematic if the surrogate endpoint is a composite endpoint that uses information on the true endpoint. In the context of the motivating colorectal cancer example, we are arguing that recurrence and death are separate processes. One can interpret our modeling strategy as a model for the process that gave rise to the data, rather than a model for the observed data. In modeling the biology, in this context, it is useful to recognize that recurrence is not a spontaneous event. It occurs because the cancer is regrowing and reaches a size where it is detected. From this perspective, there is some rationale for considering when the cancer would have grown to such a size to be detected had not the patient died from something else. K says patients might die from their disease without having progression or having it observed. That is context dependent and pretty rare in the cancer clinical trials we analyze.

OF and K advocate the use of proportional hazards (PH) models in their discussions. Since we were focusing on estimation using the wedge constraint, PH models were not available to us. The recent work of Xu, Kalbfleisch, and Tai (2010), discussed by OF, allows for proportional hazards models for  $S$  and  $T$  in the semicompeting risks setting. The type of  $R^2$  that OF describe comes from a comparison of models for  $T|Z$  and  $T|S, Z$ . It is not at all straightforward to calculate this

**Table 1**  
*R*<sup>2</sup> values for recurrence in the colorectal cancer data

Study	<i>R</i> <sup>2</sup>
C01	0.51
C02	0.38
C03	0.55
C04	0.52
C05	0.48
C06	0.42
C07	0.38
INT-0035	0.52
NCCTG 784852	0.57
NCCTG 874651	0.54
NCCTG 894651	0.56
NCCTG 914653	0.53

*Note:* The method of Nagelkerke (1991) was used to calculate *R*<sup>2</sup>. In particular, two PH models were compared. Both had age (log transformed), stage, and treatment as covariates; one included recurrence as a time-dependent covariate, the other did not. The baseline hazard function was modeled using a Weibull distribution.

quantity here because of two reasons. The first is that including *S* as a covariate, in conjunction with the constraint that  $S < T$ , will complicate estimation. Second, provided one could develop a valid method for estimation in the model for  $T | S, Z$  with  $S < T$ , calculating an *R*<sup>2</sup>-type measure poses its own issues. Guidance for constructing such measures would come from previous proposals to create likelihood ratio-type statistics from estimating equations (e.g., Li, 1993).

OF were interested in the *R*<sup>2</sup> values for our example. We show them in Table 1 for the colorectal cancer data in which parametric Weibull PH models are fit, along with adjustment for stage, age (log transformed), and treatment. The method of Nagelkerke (1991) for calculating *R*<sup>2</sup> was used. The values range between 0.38 and 0.56, compared with the *R*<sup>2</sup> value of 0.69 that OF obtained in their example. The question remains of how to set guidelines for using the *R*<sup>2</sup> value in deciding whether to use the surrogate marker.

M makes a push for performing sensitivity analyses in our modeling procedures. We agree this is an important task and area for future research. He also asks about the potential for causal interpretations of the parameters that we have estimated. Using the structural modeling framework of Pearl (2001), we (Ghosh, Elliott, and Taylor, 2010) have recently shown that the relative effect (i.e., ratio of the two regression coefficients) can be interpreted as a causal parameter in the linear case. There has been recent work on framing the surrogacy problem in the potential outcomes framework (Gilbert and Hudgens, 2008; Li, Taylor, and Elliott, 2010). Attempting to incorporate the semicompeting risks data structure into the potential outcomes framework is more challenging. Suppose we define the potential outcomes  $\{S_i^*(1), S_i^*(0), T_i^*(1), T_i^*(0)\}$ ,  $i = 1, \dots, n$ , where  $\{S_i^*(Z), T_i^*(Z)\}$  denotes the joint potential outcome for time to the surrogate and true endpoints, respectively, for the *i*th individual if assigned treatment *Z*,  $Z = 0/1$ . Then causal estimands are defined to be within-individual contrasts in *T*<sup>\*</sup> and *S*<sup>\*</sup>. Frangakis and Rubin (2002) defined the concept of principal stratification, in which within-individual contrasts for *T*<sup>\*</sup> are considered conditional

on *S*<sup>\*</sup>. The problem with the semicompeting risks approach is that *S*<sup>\*</sup> might not be well defined if the person experiences the true endpoint but not the surrogate endpoint. This has been referred to by Zhang and Rubin (2003) as “truncation by death.” While the potential outcomes framework might not allow for well-conceptualized causal estimands with semicompeting risks data, this is not the only model for causality that exists in the literature. In particular, econometricians work with so-called structural selection models (Abbing and van den Berg, 2003), and such a modeling framework might allow for better incorporation of semicompeting risks data. Of course, “causal estimand” has a different meaning using these models relative to the potential outcomes framework. This research is currently under investigation.

We broadly agree with much of what the discussants proposed regarding trial-level meta-analysis. K advocates prediction, but as noted before, that is not straightforward in our modeling scheme with censoring present. In our example, we combined treatment arms despite well-documented evidence of heterogeneity in the different groups. We note that of 14,246 initial subjects, there are 56 and 48 subjects with time to recurrence and time to death equaling zero, all of which are censored, so this represents a very small percentage of observations.

The lively discussion of our article has led us to consider a compromise between the association framework proposed here with another view of surrogates, termed auxiliary variables, that might lead to greater consensus. If *T* is missing, auxiliary variable methods would impute the value of *T* based on the value of *S*. In this way, the composite endpoint of disease-free survival (DFS) can be viewed as an imputation strategy by replacing missing values of *T* with *S*. From the perspective of auxiliary variables this is clearly biased, but in this setting this might be reasonable for two reasons. First, *S* and *T* are highly correlated so we might expect *S* to be a good prediction of *T*. Second, DFS as an endpoint has a clinically meaningful interpretation. Thinking of surrogate markers as auxiliary information would seem to be a strategy that could keep the discussants such as K and BIK happy because we still use the real endpoint if it is available but would allow for information in *S* to be utilized. If *S* was only weakly related to *T* then there would be little gain in efficiency. By contrast, if *S* was strongly related to *T* then there are potential gains in efficiency.

In closing, we would like to stress that if the goal is to identify surrogate endpoints that occur before the true endpoint so that trials can be done more quickly, then this will necessitate accepting a greater level of uncertainty. There are aspects based on the semicompeting risks framework that allow for this, but by no means is this the only type of methodology available. The question then becomes how much are you willing to lean on the knowledge from biology and data from other trials to help control this uncertainty. How to do this is where the role of statistics is crucial.

#### ACKNOWLEDGEMENTS

The authors would like to thank Anna Conlon for the *R*<sup>2</sup> calculations in Table 1. This research is supported by National Institutes of Health grants CA129102, CA25224, and CA12027.

## REFERENCES

- Abbing, J. H. and van den Berg, G. J. (2003). The nonparametric identification of the treatment effects in duration models. *Econometrica* **71**, 1491–1517.
- Burzykowski, T., Molenberghs, G., and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints*. New York: Springer-Verlag.
- Fix, E. and Neyman, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology* **23**, 205–241.
- Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.
- Ghosh, D. (2009). On assessing surrogacy in a single-trial setting using a semi-competing risks paradigm. *Biometrics* **65**, 521–529.
- Ghosh, D., Elliott, M. R., and Taylor, J. M. (2010). Links between surrogate endpoints and endogeneity. *Statistics in Medicine* **29**, 2869–2879.
- Gilbert, P. B. and Hudgens, M. G. (2008). Evaluating causal effect predictiveness of candidate surrogate end-points. *Biometrics* **65**, 1223–1232.
- Li, B. (1993). A deviance function for the quasi likelihood method. *Biometrika* **80**, 741–753.
- Li, Y., Taylor, J. M. G., and Elliott, M. R. (2010). A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* **66**, 523–531.
- Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692.
- Pearl, J. (2001). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.
- Xu, J., Kalbfleisch, J. D., and Tai, B. (2010). Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics* **66**, 716–725.
- Ying, Z., Jung, S. H., and Wei, L. J. (1995). Survival analysis with median regression models. *Journal of the American Statistical Association* **90**, 178–184.
- Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *Journal of Educational and Behavioral Statistics* **28**, 353–368.