HAL
archives-ouvertes.fr

A note on including time-dependent covariate in regression model for competing risks data.

Aurelien Latouche, Raphaël Porcher, Sylvie Chevret

▶ **To cite this version:**

Aurelien Latouche, Raphaël Porcher, Sylvie Chevret. A note on including time-dependent covariate in regression model for competing risks data.. Biom J, 2005, 47, pp.807-14. inserm-00000106

HAL Id: inserm-00000106
https://www.hal.inserm.fr/inserm-00000106

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Author Manuscript Inserm

Hal-Inserm Author manuscript

Hal-Inserm Author manuscript

A note on including time-dependent covariate in regression model for competing risks data

A. Latouche* R. Porcher S. Chevret

April 19, 2005

Abstract

Recently, regression analysis of the cumulative incidence function has gained interest in competing risks data analysis, through the model proposed by Fine and Gray (JASA 1999;94:496–509). In this note, we point out that inclusion of time-dependent covariates in this model can lead to serious bias. We illustrate the problems arising in such a context, using bone marrow transplant data as a working example and numerical simulations. Practical advices are given, preventing the misuse of this model.

1 Introduction

In longitudinal cohort studies, competing risks failure time data are commonly encountered. For instance, after allogeneic bone-marrow transplantation (aBMT) for leukemic patients in complete remission, death in remission competes with relapse. This will be our working example. To isolate the effect of covariates on these risks, several regression models can be used. Actually, regression analysis of competing risks failure time can be performed either by modelling the cause specific hazard function or the cumulative incidence function (also known as the subdistribution function). The former approach is commonly used in this setting (Rosenberg et al., 2004; Cornelissen et al., 2001). However, the instantaneous risk of specific failure cause is sometimes of less interest than the overall probability of this specific failure cause. In our working example, actually, the overall probability of death in remission, often referred as "treatment related mortality" appears more interesting than the instantaneous risk of dying in remission. Otherwise, the overall probability of relapse could also be of interest to quantify outcomes in the population of transplanted patients.

Such a probability of failure could be formulated as either the marginal distribution (of the specific failure cause), that is the probability of this failure cause in a population where only this failure cause acts, or the cumulative incidence function, i.e., the overall probability of the specific cause of failure in the presence of the competing failure causes. However, the marginal distribution is not identifiable from available data without additional assumptions, such as independence between competing failure causes. Therefore, cumulative incidence functions may appear more relevant than marginal probabilities (Pepe and Mori, 1993; Korn, 1992; Gaynor et al., 1993).

To assess the effect of a covariate on the cumulative incidence of a competing risk, Fine and Gray (1999) proposed a regression model. It has been recently used to model clinical data in cancer (Colleoni et al., 2000; Robson et al., 2004) or hematology (Rocha et al., 2001, 2002). It allows to estimate the effect of constant (time-fixed) covariates on the subdistribution hazard of specific failure causes. Time-by-covariate interaction is handled by this model, but most of time-dependent covariates such as "one time jumps" (taking 0 value unless the outcome of interest is observed, and 1 thereafter) do not belong to such a formulation. For instance, in the context of our working example, patients with leukemia frequently develop after aBMT acute graft versus host disease (aGvHD) wherin the transplanted immune cells attack the host tissues. Some evidence exists to consider that occurrence of aGvHD modifies patients' outcome as it increases risk of mortality but decreases risk of relapse. One could be interested in estimating the effect of such a time-dependent covariate (taking zero values unless the aGvHD is observed and 1 thereafter) on the occurrence of failures of interest (death or relapse).

We show that inclusion of such a time-dependent covariate is not relevant when modelling the subdistribution hazards. This article should be considered as a guideline for preventing the misuse of the model in this setting. In Section 2, we present the Fine and Gray regression model. A real data example is proposed in Section 3. In Section

*Corresponding author, aurelien.latouche@paris7.jussieu.fr

Hal-Inserm Author manuscript

4, we present a Monte Carlo simulation study to assess the bias in estimating the effect of a time-dependent covariate using the Fine and Gray model. Concluding remarks are presented in Section 5.

2 Models

Let T be the failure time, ε the cause of failure, where $\varepsilon = 1$ denotes the cause of interest and $\varepsilon = 2$ the competing cause (considering, without loss of generality, a single competing failure cause), and $F_i = \Pr(T \leq t, \varepsilon = i)$ the cumulative incidence function of failure from the cause i $(= 1, 2)$. Gray (1988) defined the subdistribution hazard for cause i as:

$$\lambda_i(t) = \lim_{dt \to 0} \frac{1}{dt} \Pr\left\{t \leq T \leq t + dt, \varepsilon = i | T \geq t \cup (T \leq t \cap \varepsilon \neq i)\right\},$$

by contrast to the cause-specific hazard:

$$\alpha_i(t) = \lim_{dt \to 0} \frac{1}{dt} \Pr\left\{t \leq T \leq t + dt, \varepsilon = i | T \geq t\right\}.$$

Similarly to the Cox model for the cause-specific hazard, $\alpha_i(t; X(t)) = \alpha_{i0}(t) \exp\{b_i X(t)\}$, where $\alpha_{i0}(t)$ is a non specified baseline hazard function, and b_i is the regression parameter, Fine and Gray (1999) proposed a regression model for the subdistribution hazard: $\lambda_i(t; X(t)) = \lambda_{i0}(t) \exp\{\beta_i X(t)\}$. By construction, the subdistribution hazard is explicitely related to the cumulative incidence function of failure from cause i, by $\lambda_i(t) = -d \log\{1 - F_i(t)\}/dt$, while the relation between the cause-specific hazard and the cumulative incidence function is less straightforward, and involves the cause-specific hazard of failure from other causes.

For inference in this model from $j = 1, \ldots, N$ individuals, the risk set at time t expresses as $\mathcal{R}(t) = \{j : (t \leq T_j) \cup (T_j \leq t \cap \varepsilon_j \neq i)\}$. This includes individuals who have not failed from any cause by t, like in the Cox model for the cause specific hazard (with risk set at time t defined by $\{j : t \leq T_j)\}$), and, in addition, those who have previously failed from the competing cause before t.

Let T be the time of failure of the individual, and Z be the time to occurrence of any event of interest. Suppose that we wish to estimate the effect of $X(t) = 1_{\{Z \leq t\}}$ on the subdistribution hazard $\lambda_1(t)$ at a particular time τ. If $T \leq \tau$ and the cause of failure is not that of interest ($\varepsilon = 2$), the risk set includes all the individuals who have not experienced any failure, and those who have previously failed from the competing cause. Moreover, in the case of an absorbing competing cause of failure such as death, the covariate value of a patient who dies cannot be observed anymore while the patient is still considered to be at risk until the maximum failure time from cause 1 of the cohort. Estimation based on the Fine and Gray model requires to know, for patients who failed from the competing cause, the entire path (history) of the covariate as well as futur values (if they merely exist). This is illustrated in Figure 1, in absence of censoring.

[Figure 1 about here.]

Let "*non-identifiable path*" denote further those observations, in opposition to "*identifiable path*" where the occurrence of the competing cause of failure does not avoid the observation of $X(t)$. Before going any further, and for illustration purpose, we will consider the *last observation carried forward* (LOCF) approach for handling unknown $X(t)$. It is known that this method introduces bias in inference even under missing completely at random (MCAR) and missing at random (MAR) settings (Cook et al., 2004). However, the point is that we cannot recover information whatever estimation technique is used to handle unknown $X(t)$. The bias overtly came from a non identifiability problem as attempts are made to condititon on the future.

3 A clinical example

We illustrated estimation of the effect of such a time-dependent covariate on a specific failure cause on real data. Data consist in a sample of 180 children with acute leukemia who underwent aBMT between 1994 and 1998 (Rocha et al., 2001). Of these 180 patients, 34 developed aGvHD followed by either relapse for 6 patients or death in remission for 22. Among the 146 patients who did not experience aGvHD, there were 60 relapses and 22 deaths in remission (Figure 2). No patient was loss to follow-up. We were concerned by estimating the effect of aGvHD on the occurrence of relapse ($\varepsilon = 1$).

Hal-Inserm Author manuscript

Hal-Inserm Author manuscript

[Figure 2 about here.]

Estimation of β_1 was carried out using the `survival` package of `R` with competing failure observations censored at their follow-up time (difference between the reference date and the entry date), as censoring only results from administrative loss to follow-up. Estimation of b_1 was performed by using a standard Cox model, where deaths in remission were censored at the time of death. The time-dependent covariate, aGvHD, was considered as a one-time jump, taking the value 0 unless aGvHD is observed. Of note, for the Fine and Gray model, the last value of the jump was carried over forward after the competing failure time of death in remission.

The estimated effect of aGvHD on the hazard of relapse, with death in remission defining the competing cause of failure, was statistically significant, with $\hat{\beta}_1 = -0.975$ ($SE = 0.429$, $p = 0.023$). By contrast, the effect of aGvHD on the cause-specific hazard of relapse was not, with $\hat{b}_1 = -0.404$ ($SE = 0.43$, $p = 0.322$).

In the next Section, a simulation study will exhibit the fact that the former estimate have no sense as we are obviously in a "*non-identifiable path*" setting.

4 Simulation

We conducted a simulation study to numerically illustrate problems arising when using the Fine and Gray model to estimate the effect of a time-dependent covariate on the subdistribution hazard of failure. Specifically, we were interested in examining the bias in estimating β_1 when the competing cause of failure is either non absorbing or absorbing for the covariate process. For the time dependent covariate, we considered a one jump process as defined by $X(t) = 1_{\{Z \leq t\}}$, where Z is the time to occurrence of some event that could be related to the outcome. We attempted to mimic the data example exposed above.

All simulations were based on 1,000 independent realizations with reasonable sample sizes of 250. For simplicity, we supposed the absence of right censoring. The occurrence of jump in the covariate process was generated from a Bernoulli distribution with parameter $q = 0.6$.

Then, the time to occurrence of aGvHD, Z, was chosen to reach a probability near 1 of aGvHD at time 100 (days), as aGvHD is defined only within the first 100 days post-transplant, with a shape similar to that observed on real data sample. Thus, the individual times Z were generated from a random variable $40 \times W$, where W has Weibull distribution with shape parameter of 2 and scale parameter of 1.

Generating failure times was complicated by the presence of the time-dependent covariate. It was based on inversion of the cumulative subdistribution hazard functions, adapting the method proposed by Leemis et al. (1990) in the case of survival data. Lifetime data from the cause of interest were generated as described by Fine and Gray (1999). Details of the failure times generation is presented in the Appendix.

We simulated two types of covariate paths: (i) identifiable paths, when the covariate process of patients who experienced the competing failure cause can still be observed, and (ii) non identifiable paths, when the time-dependent covariate $X(t)$ cannot be observed after occurrence of the competing failure cause. In this latter case we used the value of $X(t)$ at the time of failure throughout the risk sets.

Parameters (β_1, β_2) were set at $(0.5, 0.5)$ in (i), and at $(-0.5, 0.5)$ in (ii). Let r_1 be the rate of failures from cause of interest and r_2 from the competing cause. From the 1,000 simulations, we computed the mean estimate of β_1 ($E(\hat{\beta}_1)$) and of the proportion γ of patients who experienced the competing cause of failure before any jump of $X(t)$, for values of p ranging from 0.1 to 1 and values of $K = r_2/r_1$ ranging from 0.1 to 2.

We begin by presenting simulation results from model with identifiable paths. Figure 3 displays the mean estimate of $\hat{\beta}_1$ against K (Figure 3a) and p (Figure 3c). Whatever the value of K and of p, $E(\hat{\beta}_1)$ was close to its nominal value. This exemplifies the ability of the model to estimate the regression coefficient when the entire covariate path is known.

[Figure 3 about here.]

Figure 4 displays simulations results when the occurrence of the competing cause of failure avoids the observation of the jump process (non identifiable paths). Contrarily to the previous observable case, $\hat{\beta}_1$ was systematically biased, with bias increasing with K (Figure 4a). Interestingly, the shape of the estimated β_1 against K was very similar to that of γ (Figure 4b). Next, for $K = 1$, we computed $E(\hat{\beta}_1)$ for values of parameter p from 0.1 to 1 (Figure 4c). It appears that the estimates $\hat{\beta}_1$ are biased, except in the case of $p = 1$, *i.e.*, when all individuals fail from the cause of interest. In this case, γ is obviously null, as shown on Figure 4d. When p is close to zero, $F_1(t) \approx 0$, and the model is "ill-posed", so that computing $\hat{\beta}_1$ does not make any sense. Similar shapes were observed for values of $r_1 = 0.005, 0.01, 0.02$,

Hal-Inserm Author manuscript

with an increase in the bias of $E(\hat{\beta}_1)$ as r_1 increases (or equivalently an increase in γ as shown on Figure 4b). Of note, a linear decrease of γ with p was observed (Figure 4d), whereas such pattern was not found between $E(\hat{\beta}_1)$ and p.

[Figure 4 about here.]

Moreover in our simulation setting, one can show that: $\gamma = (1-p)\{q + (1-q) \times \mathbf{C}\}$, where $\mathbf{C}$ is the probability of jump after failure, conditional on failure from competing cause, and is therefore independent of p and q. As a result, γ is indeed a decreasing linear function of p as shown in Figures 3d and 4d.

5 Discussion

In this paper, we showed, on the basis of a working example and a simulation study, that the Fine and Gray model is not appropriate for estimating the effect of any time-dependent covariate unless the entire covariate path is observable. Otherwise, *i.e.*, in the case of so-called "internal" time-dependent covariate using the terminology of Kalbfleisch and Prentice (1980), the use of the Fine and Gray model can lead to a serious bias in estimate, even in the simple studied case of a one time jump process, which is actually often observed in clinical epidemiology data.

To replace unobservable values of the time-dependent covariate, the simple LOCF imputation, i.e using (and keeping) the value of the covariate at the time of failure for patients who developped the competing cause of failure, is not advisable, as shown by our simulation results. Moreover, alternative modelling approaches to impute values for unobservable covariates appear somewhat useless in this context, as they could not recover inexistent information.

Since the Fine and Gray model can only be used if the entire path of the time-dependent covariate is known, this obviously prohibits the introduction of any time-dependent covariate in the model when death is a competing cause of failure. For instance, in our working example, no valid estimation of the effect of aGvHD on the subdistribution hazard of relapse could be obtained, due to deaths in remission. For non fatal competing events, the Fine and Gray model should also not be used, unless checking carefully that the observation period does not end with the occurrence of the competing event.

Our main concern was to prevent the misuse of the Fine and Gray model with time-dependent explanatory variables. Our simulation studies also provide a better understanding of the structure of the "unnatural" risk set of the Fine and Gray model, pointing out that competing failures stay in the risk set until censoring time.

If one could reasonably think of allowing the covariate to influence the subdistributuion hazard only up to the first competing event, this would entail to modify the definition of the risk-set, that is to say to modify the model itself. Nonetheless, this gives direction to further developpments of new models with "weighted influence" of covariates. To cope with estimation of the effect of time-dependent covariates, other statistical models should thus be proposed. Multistate models with cause-specific transition rate have already been used (Andersen et al., 2002; Hougaard, 1999). Further work is needeed to estimate time-dependent transition (non-homogeneous markov process) in this setting.

Appendix

Briefly, the subdistribution of failure times from the cause of interest ($\varepsilon = 1$) is given by $F_1(t, X(t)) = 1 - [1 - p\{1 - \exp(-r_1 t)\}]^{\exp(\beta_1 X(t))}$, which is a unit exponential mixture with mass $1-p$ at $+\infty$, where p is the proportion of failures from the cause of interest, and uses the proportional subdistribution hazards model to obtain the subdistribution for nonzero covariate values. Let $\psi(X(t))$ be the link function relating the covariate process to the subdistribution hazard function, and $\Psi(.)$ the cumulative link function *i.e.* $\Psi(t) = \int_0^t \psi(X(u))du$. Let $\Lambda_1(.)$ be the cumulative subdistribution hazard function, $\Lambda_1(t) = \int_0^t \lambda_1(u)du$. As a result, $\Lambda_1(t) = -\log S_{10}(t)$ for $t \leq \tau$ and $\Lambda_1(t) = -\log S_{10}(\tau) + \exp(\beta_1) \times \{\log S_{10}(\tau) - \log S_{10}(t)\}$ otherwise, where $S_{10}(t) = 1 - F_1(t, X(t) = 0)$. Failure times from the cause of interest were thus generated through, $t \leftarrow \Psi^{-1}[\Lambda_1^{-1}\{-\log(1-u)\}]$, where u is taken from a uniform distribution on $[0, 1]$ and $\psi(X(t)) = \exp\{\beta_1 X(t)\}$.

Since the subdistribution for the competing failure cause was considered exponentially distributed with rate r_2, we directly used the non-modified algorithm of Leemis et al. (1990) to generate corresponding competing failure times, with the link function $\psi(X(t)) = \exp\{\beta_2 X(t)\}$.

Simulation codes are available upon request to the corresponding author.

Hal-Inserm Author manuscript

Hal-Inserm Author manuscript

References

Andersen, P. K., Abildstrom, S., and Rosthoj, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research*, 11(2):203–215.

Colleoni, M., O'Neill, A., Goldhirsch, A., Gelber, R., Bonetti, M., Thrlimann, B., Price, K., Castiglione-Gertsch, M., Coates, A., Lindtner, J., Collins, J., Senn, H., Cavalli, F., Forbes, J., Gudgeon, A., Simoncini, E., Cortes-Funes, H., Veronesi, A., Fey, M., and Rudenstam, C. (2000). Identifying breast cancer patients at high risk for bone metastases. *Journal of Clinical Oncology*, 18(23):3925–3935.

Cook, R., Zeng, L., and Yi, G. (2004). Marginal analysis of incomplete longitudinal binary data: A cautionary note on LOCF imputation. *Biometrics*, 60:820–828.

Cornelissen, J., Carston, M., Kollman, C., King, R., Dekker, A., Lowenberg, B., and Anasetti, C. (2001). Unrelated marrow transplantation for adult patients with poor-risk acute lymphoblastic leukemia: strong graft-versus-leukemia effect and risk factors determining outcome. *Blood*, 97(6):1572–1577.

Fine, J. and Gray, R. (1999). A proportional hazards model for subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

Gaynor, J., Feuer, E., Tan, C., Wu, D., Straus, D., Clarkson, B., Brennan, M., and Little, C. (1993). On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *Journal of the American Statistical Association*, 88:602–607.

Gray, R. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 116:1141–1154.

Hougaard, P. (1999). Multi-state models: A review. *Lifetime Data Analysis*, 5:239–264.

Kalbfleisch, J. and Prentice, R. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York.

Korn, E.L.and Dorey, F. (1992). Applications of crude incidence curves. *Statistics in Medicine*, 11(6):813–829.

Leemis, L., Shih, L., and Reynertson, K. (1990). Variate generation for the accelerated life and proportional hazards models with time dependent covariates. *Statistics and Probability Letters*, 10(1):335–339.

Pepe, M. and Mori, M. (1993). Kaplan-meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine*, 12(8):737–751.

Robson, M., Chappuis, P., Satagopan, J., Wong, N., Boyd, J., Goffin, J., Hudis, C., Roberge, D., Norton, L., Begin, L., Offit, K., and Foulkes, W. (2004). A combined analysis of outcome following breast cancer: differences in survival based on BRCA1/BRCA2 mutation status and administration of adjuvant treatment. *Breast Cancer Research*, 6(1).

Rocha, V., Cornish, J., Sievers, E., Filipovich, A., Locatelli, F., Peters, C., Remberger, M., Michel, G., Arcese, W., Dallorso, S., Tiedemann, K., Busca, A., Chan, K., Kato, S., Ortega, J., Vowels, M., Zander, A., Souillet, G., Oakill, A., Woolfrey, A., Pay, A., Green, A., Garnier, F., Ionescu, I., Wernet, P., Sirchia, G., Rubinstein, P., Chevret, S., and Gluckman, E. (2001). Comparison of outcomes of unrelated bone marrow and umbilical cord blood transplants in children with acute leukemia. *Blood*, 97:2962 – 2971.

Rocha, V., Franco, R., Porcher, R., Bittencourt, H., Silva, W. A., Latouche, A., Devergie, A., Esperou, H., Ribaud, P., Socie, G., Zago, M., and Gluckman, E. (2002). Host defense and inflammatory gene polymorphisms are associated with outcomes after HLA-identical sibling bone marrow transplantation. *Blood*, 100(12):3908–3918.

Rosenberg, P., Huang, Y., and Alter, B. (2004). Individualized risks of first adverse events in patients with fanconi anemia. *Blood*.

Hal-Inserm Author manuscript

Hal-Inserm Author manuscript

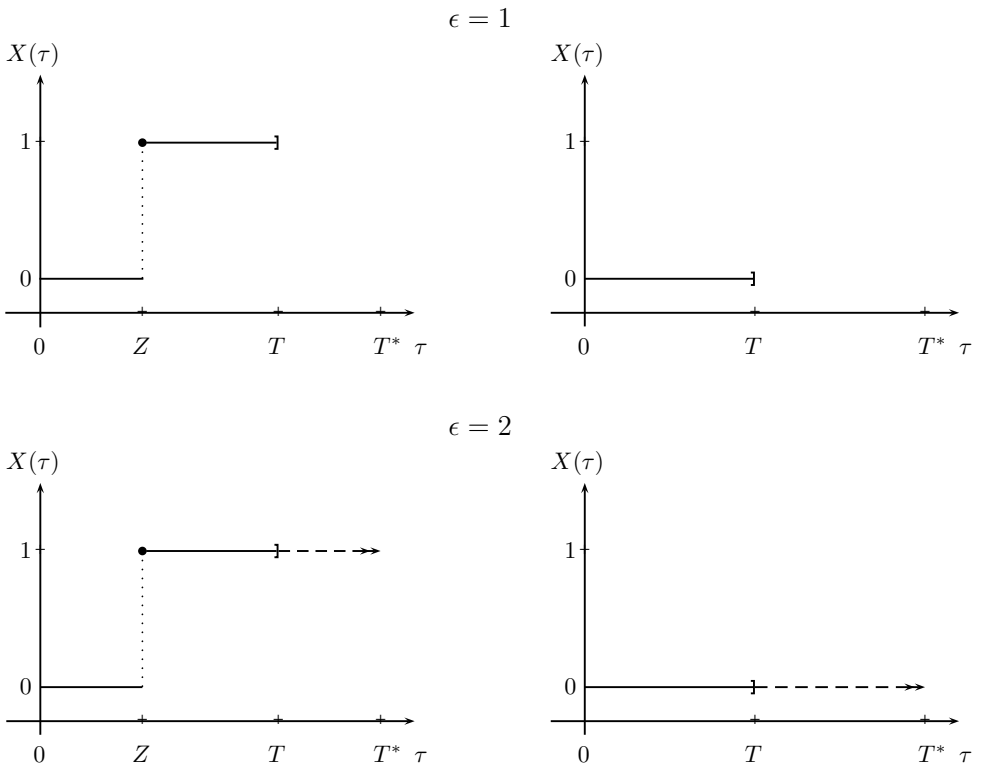

Figure 1: Illustration of the covariate path, $X(\tau)$ overtime τ according to the experienced events. T denotes the failure time and Z denotes the time to occurrence of the event of interest. Upper plots concern patients who failed from the cause of interest ($\varepsilon = 1$) while lower plots concern patients who failed from the competing failure cause ($\varepsilon = 2$). Left plots concerns patients who experienced the event of interest, and right plots concern patients who did not. T^* denotes the maximal failure time among the sample.

Hal-Inserm Author manuscript

Hal-Inserm Author manuscript

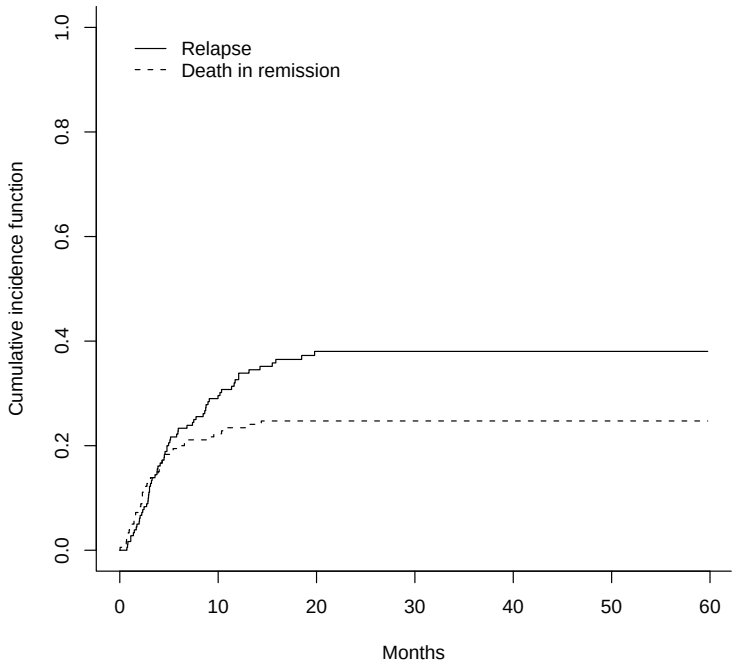

Figure 2: Estimated cumulative incidences of relapse and death in remission

Hal-Inserm Author manuscript

Hal-Inserm Author manuscript

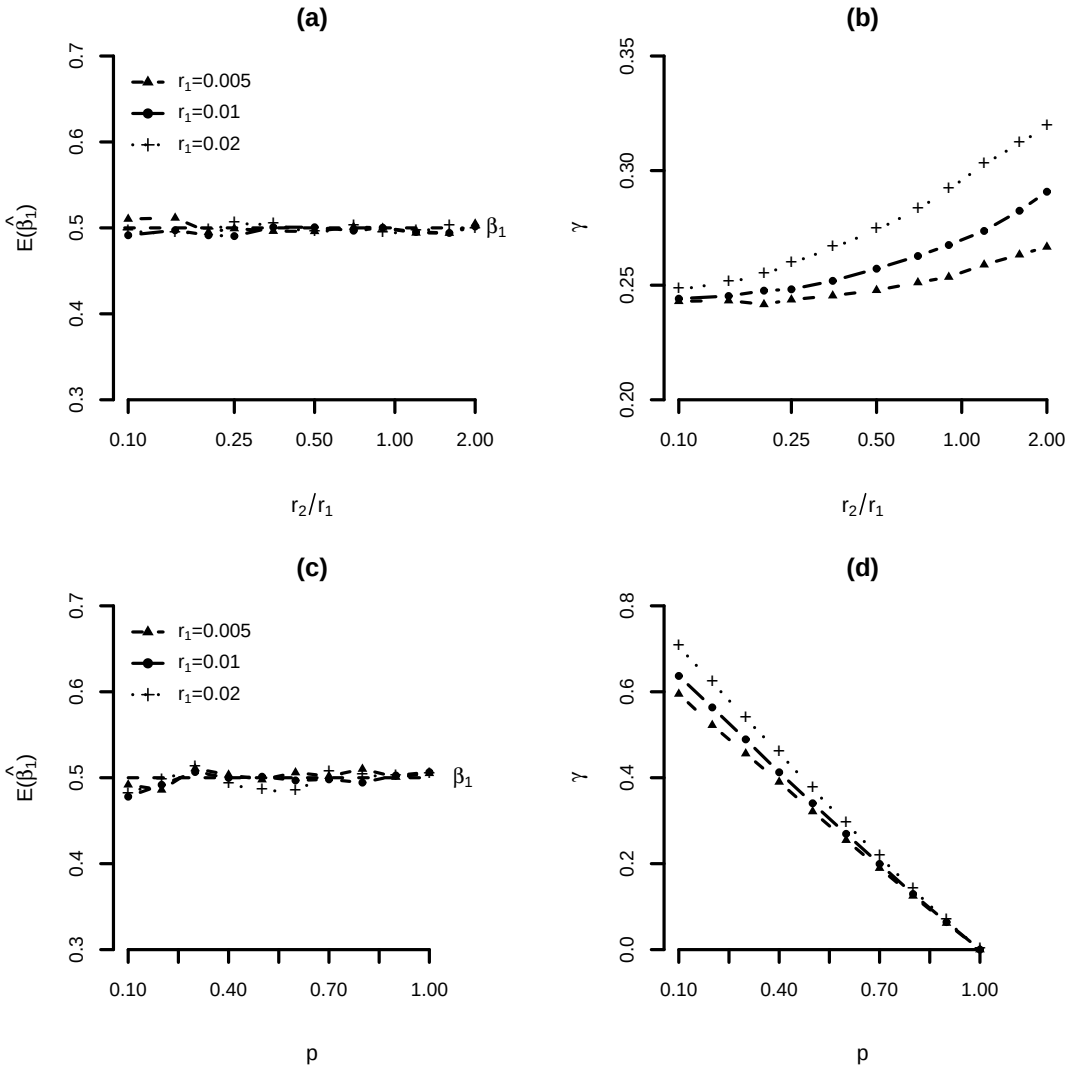

Figure 3: Simulation results in the case of identifiable path: Mean value of $\hat{\beta}_1$ (a) and the proportion γ of patients who experience the competing failure cause before any jump (b) against the ratio r_2/r_1 of the rates of failures from cause 2 and 1. Mean value of $\hat{\beta}_1$ (c) and γ (d) against the proportion p of failure from cause 1.

Hal-Inserm Author manuscript

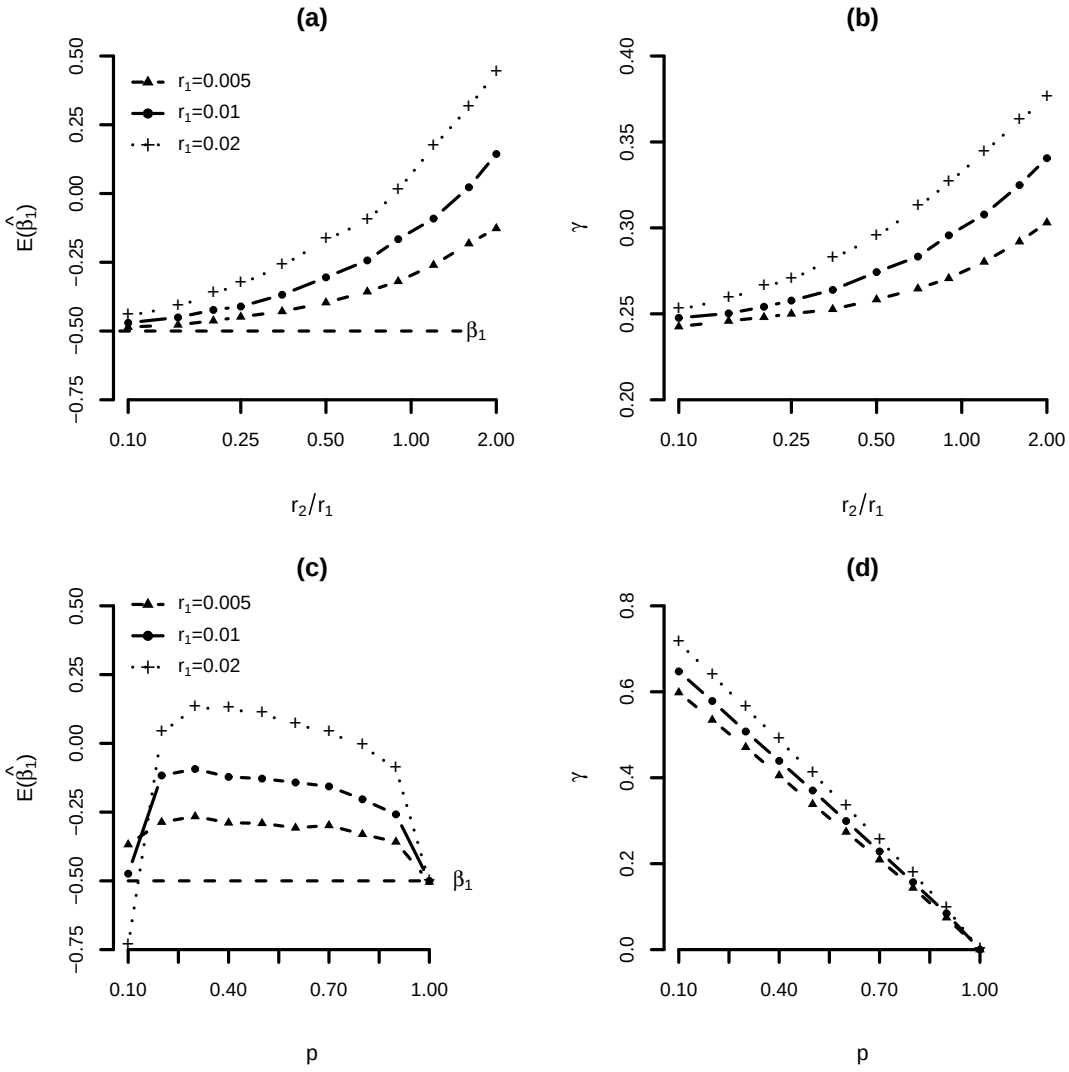

Figure 4: Simulation results in the case of non-identifiable path: Mean value of $\hat{\beta}_1$ (a) and the proportion γ of patients who experience the competing failure cause before any jump (b) against the ratio r_2/r_1 of the rates of failures from cause 2 and 1. Mean value of $\hat{\beta}_1$ (c) and γ (d) against the proportion p of failure from cause 1.

Hal-Inserm Author manuscript