Research Article

Statistics in Medicine

Received 30 November 2012, Accepted 8 July 2014 Published online 3 August 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6280

Correcting for the dependent competing risk of treatment using inverse probability of censoring weighting and copulas in the estimation of natural conception chances

N. van Geloven,[a*†] R. B. Geskus,[b] B. W. Mol[c‡] and A. H. Zwinderman[b]

When estimating the probability of natural conception from observational data on couples with an unfulfilled child wish, the start of assisted reproductive therapy (ART) is a competing event that cannot be assumed to be independent of natural conception. In clinical practice, interest lies in the probability of natural conception in the absence of ART, as this probability determines the need for therapy. We thus want to estimate the marginal cumulative pregnancy distribution. Without assumptions on the dependence structure between the two competing events, this marginal distribution is not identifiable. We first use inverse probability of censoring weighting assuming that the factors influencing the choice to start ART are known. Then, we parameterize the event distributions for conception and for start of ART and use copulas to account for the dependency between both events. By using these two ways of correcting for the dependent risk of treatment, we obtain a plausible estimation region for the cumulative pregnancy curve and for the prognostic effect of tubal tests.

Keywords: dependent competing risk; inverse probability of censoring weights; copula

1. Introduction

A common problem encountered when estimating a survival function of time-to-event T is the occurrence of some eventuality that prevents T from being observed. In many situations, the censoring cause can be considered to be independent of the event of interest, for instance, in case of the end of study censoring, patients moving to another location for reasons unrelated to their health status or accidental deaths [1]. In such situations, the common techniques used for estimation of the marginal survival distributions such as the Kaplan–Meier method and the Cox proportional hazards model that rely on the assumption of all censoring being non-informative for the event of interest are valid. In other situations, however, such an independence assumption may not be valid, for instance, when patients are lost to follow-up because of reasons related to therapy success (an ineffective treatment may lead to patients becoming too sick to follow up) or when censoring occurs because of a competing event related to the event of primary interest. Analysis methods for competing risks usually estimate and base regression on the cumulative incidence function and the cause-specific hazard function as these quantities are directly estimable from the data [2–4]. These two quantities, however, express event incidence in the presence of the competing risk(s) and do not give information on the marginal distributions, that is, the event incidence in case the competing

[a] *Clinical Research Unit, Academic Medical Center, Amsterdam, The Netherlands*
[b] *Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherlands*
[c] *Center for Reproductive Medicine, Department of Obstetrics and Gynaecology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands*
**Correspondence to: N. van Geloven, Clinical Research Unit, Academic Medical Center, Amsterdam, The Netherlands.*
[†] *E-mail: n.vangeloven@amc.nl*
[‡] *Current address: School of Paediatrics and Reproductive Health, University of Adelaide, 5000 SA Australia.*

risk(s) would not exist. A marginal survival function of the event of interest is not identifiable in case of dependent competing risks [5, 6], unless one is willing to make assumptions about the dependence structure between the competing events [7]. Several methods have been proposed to enable the estimation of the marginal distributions using inverse probability of censoring weighting (IPCW) techniques [8–10], copula functions [7, 11], dependent frailty terms [12, 13] or other specific dependence assumptions [14, 15].

When estimating the probability of natural conception from observational time-to-event data on couples with an unfulfilled child wish, the start of assisted reproductive therapy (ART) precludes the observation of this event and can be considered a competing risk [16]. The quantity of main interest is the probability of natural conception in the absence of ART, as this quantity determines the need for therapy. We thus want to estimate the marginal cumulative pregnancy distribution. Observational data available from such patients usually contain follow-up information from both untreated patients and patients that start ART. As couples with worse natural pregnancy prognosis are expected to opt for treatment earlier than couples with better prognosis, the competing risk of treatment cannot be considered to be independent of the event of interest. In this situation where the competing risk is an (additional) treatment that is started when the event of interest (pregnancy, or in general recovery) fails to occur, the amount of censored data can be large. We will here focus on a cohort of 5360 couples aiming for natural pregnancy of whom by the end of the 3-year follow-up period 3012 (56%) had started treatment.

After introducing the dataset and notation, we first use IPCW to correct for the dependent risk of start of treatment, assuming that the observed patient characteristics contain all necessary information on the decision to start treatment. Second, we use copulas to model the dependency between the events, assuming parametric distributions for both events. In all analyses, we aim to estimate the marginal cumulative pregnancy curve and the prognostic ability of tubal patency tests for natural conception.

2. The OFO cohort and notation

The Oriënterend Fertiliteits Onderzoek; basic fertitlity workup (OFO) study was an observational cohort study that followed 7860 couples with unfulfilled child wish presenting at 38 centers in the Netherlands between January 2002 and February 2004 from their first workup in a fertility clinic until the occurrence of a natural pregnancy [17]. Time to pregnancy was considered censored at the moment ART treatment, that is, intrauterine insemination (IUI) or in vitro fertilization (IVF), had been started or at the last date of contact. We here include 3 years of follow-up from 5360 couples with unexplained cause of their childlessness [16]. During this time, 1005 patients conceived naturally, 3012 started treatment (2290 IUI and 722 IVF), and the remaining 1343 were still on expectant management at the end of their follow-up. We will denote the time to natural pregnancy for each patient i, $i = 1, \ldots, n$, as T_i and the time to censoring as C_i^1 for IUI treatment, C_i^2 for IVF treatment and C_i^3 for censoring due to other reasons (mostly end of study), each maximized at 3 years (C^{max}). The observed data consist of the follow-up time $X_i = \min\left(T_i, C_i^1, C_i^2, C_i^3\right)$, the pregnancy indicator $\delta_i = I\left(X_i = T_i\right)$ and similar indicators for the competing events. Prognostic factors available were center, female age, duration of child wish, previous pregnancy, semen count, body mass index (BMI), cycle length and tubal test results: chlamydia antibody titer (CAT) test, hysterosalpingography (HSG) and diagnostic laparoscopy (DLS). With the exception of center, these are considered the most important predictors for natural pregnancy [18].

Figure 1 shows the cumulative pregnancy distribution as calculated by the Kaplan–Meier method assuming all censoring types $\left(C^1, C^2 \text{ and } C^3\right)$ were non-informative. We also added to this plot the Peterson upper and lower bounds [19], representing the cumulative pregnancy curves under the extreme scenarios where couples that started treatment are assumed to never conceive naturally or, to the contrary, to conceive naturally immediately after the start of treatment:

$$\text{Peterson lower bound: } P\left(T > C_{max} | X = C^1 \text{ or } X = C^2\right) = 1$$
$$\text{Peterson upper bound: } P\left(T < X + \epsilon | X = C^1 \text{ or } X = C^2\right) = 1,$$

with ϵ small. The upper bound was estimated with the Kaplan–Meier estimator for any event (both pregnancy and treatment), and the lower bound was estimated by the cumulative incidence function for pregnancy [20]. Anticipating on the IPCW analyses, we excluded 335 couples from centers with low treatment numbers, leaving 5025 patients for the analyses. The figure shows a wide range of pregnancy curves theoretically consistent with the data. Given the expectation that couples who started treatment

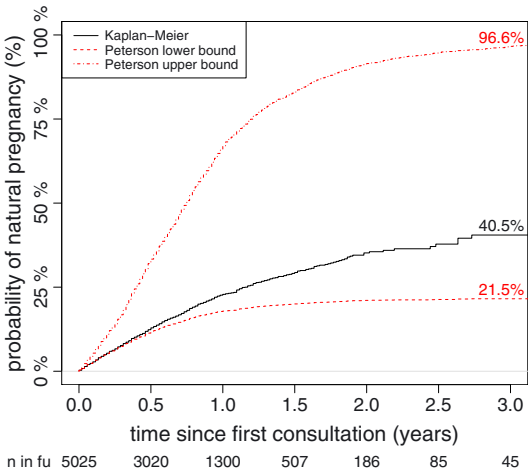

Figure 1. Kaplan–Meier cumulative pregnancy curve with Peterson bounds.

had worse natural pregnancy prognosis than couples who remained in follow-up longer, we expect the true cumulative pregnancy curve to lie in between the Kaplan–Meier curve and the lower bound curve. To gauge a more narrow estimate for the curve, we have to make assumptions on the dependency between natural pregnancy prognosis and the start of treatment.

3. Correction using inverse probability of censoring weighting

Inverse probability of censoring weighting is a method that can be used to estimate marginal survival in the presence of dependent competing risks [8, 10, 21, 22]. By assigning patients' contributions to the risk sets a weight that is inversely proportional to the estimated conditional probability of not yet experiencing the competing event, a pseudopopulation is created, which would have been observed if the competing risk did not exist [21]. The ability of the IPCW method to create this pseudopopulation depends on whether the assumptions of conditional exchangeability and correct model specification are met. The assumption of conditional exchangeability in our situation entails that conditional on the measured common predictors for natural pregnancy and start of treatment, couples who started treatment had the same natural pregnancy prognosis as couples who did not. This assumption holds under three conditions. First, all common predictors are appropriately measured and accounted for in the analysis. Second, there are a sufficient (nonzero) number of participants under follow-up at all relevant times for every combination of values observed for the common predictors (positivity). Third, the common predictors cannot be deterministic or nearly deterministic in relation to both natural pregnancy and start of treatment [21]. The assumption of correct model specification requires appropriate functional forms of the common predictors in the (weighted) model for natural pregnancy as well as in the censoring model used for construction of the weights [9]. In the following description of our IPCW analysis, we will describe the attempts we made to fulfill these assumptions.

In our dataset, we hypothesized that all recorded predictors for natural pregnancy (female age, duration of child wish, previous pregnancy, semen count, BMI, cycle length and tubal test results) could have influenced the decision to start treatment and should therefore be included in the censoring models. For two of the three tubal tests (HSG and DLS), the timing of the performance of the test was known; these two factors were used as time-dependent covariates. We used a missing indicator (i.e., a separate test outcome category 'missing') in the period before test performance, as when a test is not performed, the underlying disease status remains unknown and cannot therefore influence the treatment decision. Not all patients received all tests: 80% received the CAT, 53% received the HSG, and 32% received the DLS. Patients that did not undergo one of the tubal tests were given a missing indicator during the whole follow-up period. All other patient characteristics were known at the first visit to the fertility clinic ($t = 0$) and thus used as time-fixed covariates. Furthermore, we hypothesized that the center in which a couple was included may have influenced the choice to start treatment. Variability in fertility management between gynaecologists and the availability of IVF facilities in the particular center may have caused differences in the timing from first consultation until treatment start. In general, one only wants to

involve factors in the inverse probability of censoring weights that are predictive for both the competing event and the event of interest. There is no hypothesized causal relation between center and natural pregnancy as couples did not undergo any active treatment in the centers but were on expectant management. The centers may, however, represent unmeasured factors related to outcome (e.g., selection caused by variation in the diagnostic work-up protocol or diagnostic equipment) [23]. We therefore considered center as a potential predictive factor for the pregnancy outcome too.

As the patient factors may have different prognostic values for the start of IUI than for the start of IVF, we computed stabilized weights for both of these competing events separately:

$$sw_i^1(t) = \frac{P\left(C_i^1 > t\right)}{P\left(C_i^1 > t|\bar{v}_i(t)\right)}$$
$$sw_i^2(t) = \frac{P\left(C_i^2 > t\right)}{P\left(C_i^2 > t|\bar{v}_i(t)\right)},$$

with $\bar{v}_i(t)$ the observed prognostic factors for couple i, the bar referring to the history of all recorded values of these factors between time 0 and time t.

Weights were estimated by fitting a Cox model for each treatment (using the *ipwtm* function from the ipw package in R [24]) and combining both resulting weights [10, 25]:

$$sw_i(t) = sw_i^1(t) * sw_i^2(t).$$

We then calculated a time-dependent weighted Kaplan–Meier curve with natural pregnancy as the event of interest in which a couple was assigned its estimated weight $sw_i(t)$ at each time point where an end of follow-up was measured in the dataset.

In a first model, all factors were assumed to be linearly associated with the (log of the) censoring hazards. All previously described covariates were used in the models, without selection. The following factors showed a significant association with the start of IUI: female age, duration of child wish, previous pregnancy, semen count, BMI, HSG, DLS and center. For IVF, these were female age, duration of child wish, CAT, HSG, DLS and center. Full estimation results are shown in webappendix A (see supporting information).

As estimated weights with the mean far from one or with very extreme values are indicative for non-positivity or misspecification of the weight model [9], we assessed the weight distribution resulting from this censoring model. Weights turned out to be highly skewed (model 1 in Table I). In an attempt to improve the model specification, we reformulated the censoring models by using natural cubic splines ($df = 3$) for the numerical predictors age, BMI, cycle length, duration of child wish and semen count, thus relaxing the linearity assumption. With this adjusted specification semen count showed additional significant association with start of treatment in the IVF model. The weight distribution improved somewhat, but serious skewness indicative for positivity problems remained (model 2, Table I). Indeed, when making a cross tabulation of all used predictive factors (cutting numerical predictors at the mean), we observed many empty cells, especially at longer follow-up times where the risk sets were smaller. Further inspection of the weights showed that the high weights ($\geq$ 200) were given to six couples, five of whom belonged to the same center. In this center, only five out of 434 couples remained in the risk set beyond 2 years of follow-up. To minimize the influence of these outliers, we progressively truncated the weights (models 3 to 7 in Table I). Model 8 shows the results of a conventional analysis in which no adjustment is made, comparable with a plain Kaplan–Meier analysis. Table I also shows the estimated 3 years of natural pregnancy chances. Confidence intervals were obtained with the bootstrap technique. We made 200 bootstrapped sets of couples. For each of these sets, we used the weights obtained in the original analysis to obtain a new weighted estimate of the 3-year pregnancy chance. A more time-consuming bootstrap procedure in which all steps (also the weight estimation) were repeated was performed for one of the scenarios and gave a similar confidence interval.

The estimated 3 years of pregnancy chances from models 1 to 7 that used IPCW adjustment were all lower than the estimated 41% from the unadjusted analysis. This confirms our expectation of couples starting treatment having worse pregnancy prognosis than couples who stay in follow-up. In the model with the most extreme weights (model 1), the 3-year pregnancy estimate was as low as 33%. Skewed weights, however, are indicative for violation of the correct model specification or positivity assumptions. Formally, stabilized weights ought to be centered at one [26]. Therefore, after progressively truncating

Table I. Three years of natural pregnancy chances estimated by time-dependent weighted Kaplan–Meier analyses.

Model	Functional form weight model	Truncation percentiles	Weight distribution Mean(*SD*)	Weight distribution min/max	Pregnancy at 3 years Estimate (%)	Pregnancy at 3 years 95% CI
1	Linear effects	0, 100	2.16(87.0)	0.07/14,670	33	(29, 36)
2	Splines	0, 100	1.29(9.96)	0.05/1101	33	(29, 37)
3	Splines	0.1, 99.9	1.12(3.57)	0.08/284	34	(30, 39)
4	Splines	0.2, 99.8	1.06(1.65)	0.12/80.3	36	(31, 40)
5	Splines	0.5, 99.5	1.01(0.61)	0.18/15.4	38	(33, 43)
6	Splines	0.7, 99.3	1.00(0.47)	0.20/9.35	39	(34, 44)
7	Splines	1, 99	0.99(0.38)	0.24/5.98	39	(34, 44)
8	No weight model	None	1.00(0.00)	1.00/1.00	41	(36, 45)

In each model, different weights were used, either based on a different choice of the functional form of the numerical covariates in the Cox model that was used to calculate the weights or based on a different truncation rule of the weights.

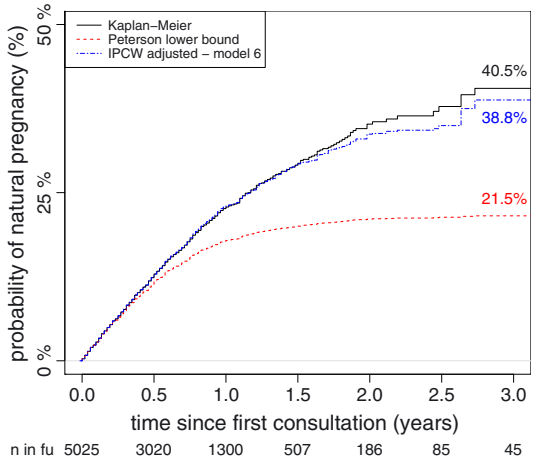

Figure 2. Inverse probability of censoring weighting (IPCW)-adjusted cumulative pregnancy curve.

the weights, we considered the first model with weight mean one and not too extreme minimum and maximum weights (model 6) as the most trustworthy model. The 3-year pregnancy estimate in this model (39%) was very similar to neighboring models 5 and 7 and only slightly lower than that in the analysis assuming independent censoring. Figure 2 shows the estimated cumulative pregnancy curve over the 3 years of follow-up using model 6. The figure shows that the adjusted and unadjusted curves only start to diverge from about 1.5 years of follow-up onwards, when the number of couples that is censored becomes large.

As competing risks may also influence the estimation of the effects of covariates, we studied the influence of the IPCW on estimates of the prognostic ability, the hazard ratios (HRs), of the three tubal patency tests (CAT, HSG and DLS). We assessed the HRs of a positive test result compared with a negative test result given that the test was performed by using a missing indicator for patients that did not receive the test under study. In the prognostic model, a one-sided as well as a two-sided occlusion of the tubes was considered a positive test result for the HSG and DLS. As for these two also the timing of the tests was known, these were used as time-dependent covariates with a missing indicator in the period before the test was performed.

The results of the analyses are shown in Table II. For the CAT and HSG tests, the adjusted models in general estimated slightly weaker prognostic ability of tubal tests than the unadjusted models. This may have been caused by a selection of test-negative patients with relatively good prognosis in later follow-up. This selection may have led to an overestimation of the prognostic ability of the diagnostic tests in the unadjusted model. This mechanism has been described before when an unadjusted analysis was compared with a competing risks analysis using the Fine and Gray model [16]. However, model 6, which as explained earlier was considered the most trustworthy model because of not too extreme weights

Table II. Hazard ratios for the three tubal tests estimated by time-dependent weighted Cox analyses.

Model	Functional form weight model	Truncation percentiles	HR CAT		HR HSG		HR DLS	
			Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
1	Linear effects	0, 100	0.72	(0.60, 0.86)	0.76	(0.56, 1.02)	0.75	(0.47, 1.21)
2	Splines	0, 100	0.72	(0.60, 0.86)	0.79	(0.59, 1.08)	0.72	(0.45, 1.16)
3	Splines	0.1, 99.9	0.71	(0.59, 0.85)	0.77	(0.57, 1.04)	0.72	(0.45, 1.16)
4	Splines	0.2, 99.8	0.70	(0.59, 0.84)	0.76	(0.56, 1.02)	0.72	(0.44, 1.16)
5	Splines	0.5, 99.5	0.69	(0.58, 0.82)	0.74	(0.55, 0.99)	0.71	(0.44, 1.15)
6	Splines	0.7, 99.3	0.68	(0.57, 0.81)	0.73	(0.55, 0.99)	0.72	(0.44, 1.16)
7	Splines	1, 99	0.67	(0.57, 0.80)	0.73	(0.55, 0.98)	0.73	(0.45, 1.18)
8	No weight model	None	0.68	(0.58, 0.80)	0.73	(0.55, 0.97)	0.80	(0.50, 1.26)

In each model, different weights were used, either based on a different choice of the functional form of the numerical covariates in the Cox model that was used to calculate the weights or based on a different truncation rule of the weights. HR, hazard ratio; CAT, chlamydia antibody titer; HSG, hysterosalpingography; DLS, diagnostic laparoscopy.

centered at one, did not show any proof of such a mechanism. The point estimates of the HRs for the CAT and the HSG are the same in this model as in the unadjusted models. Actually, the width and huge overlap of the confidence intervals in all models prevent any strong conclusions. The IPCW analyses do therefore not give definite proof of bias in the estimates of the prognostic ability of the CAT and HSG from the unadjusted model.

The results from the DLS test showed a different trend. The adjusted models, including model 6, estimated slightly stronger prognostic ability of the DLS than the unadjusted model. This is not in line with previous analyses comparing an unadjusted analysis with the Fine and Gray model [16]. The IPCW analyses suggest that the unadjusted model underestimates the prognostic ability of the DLS, although the bias was small compared with the width of the confidence interval.

4. Correction using copulas

Copulas provide a way to model the dependence between two competing time-to-event variables by giving the dependence structure that relates the marginal event distributions to their bivariate joint distribution. In particular, according to Sklar's theorem, when T and C are competing events with continuous marginal survival distributions F and G, then there exists a unique copula C_θ such that the joint survival function $S(T, C)$ for each t, c can be calculated as

$$S(t, c) = P(T > t, C > c) = C_\theta(F(t), G(c)),$$

where C_θ is a continuous bivariate distribution function on the unit square with uniform marginals that includes one or more dependence parameter(s) θ [6, 11]. Zheng and Klein [7] proved that given the observable data in a competing risk setting and when the copula is fully known, the marginal distribution functions are uniquely determined and can be estimated. This means that when the dependency between the competing risks is specified in the copula function, the marginal distribution functions are identifiable. Zheng and Klein advised to choose a copula type and vary the dependence parameter(s) over a plausible range to obtain bounds on the marginal distribution function of interest. Their approach has been extended to models incorporating the estimation of covariate effects on the marginal distributions [27, 28]. As in general the copula type and the association parameter(s) are not known, this approach will often lead to wide bounds, in ultimo as wide as the Peterson bounds presented in Section 2. If one is willing to parameterize the marginal distributions, then the parameters of both the copula and the marginal distributions can be estimated from competing risk data by standard maximum likelihood techniques [7, 11]. Here, we choose this latter approach. The copula approach is appealing as we have flexibility in choosing functional forms of the marginal distribution functions, many copula functions have been proposed and their properties well studied [29] and the degree of dependence between the competing risks can be expressed through commonly used association measures such as Spearman's rho [27].

When we parameterize the time to pregnancy, T_i, with marginal survival function F, density function f and hazard function h_f, and the time to the start of ART, $C_i = \min\left(C_i^1, C_i^2\right)$, with marginal survival

Table III. Dependence structures of the Plackett and Frank copulas.

Copula type	Dependence function			
Frank	Copula	$C_\theta(F(t),G(c))$	=	$-\frac{1}{\theta}\log\left(1+\frac{(e^{-\theta F(t)}-1)(e^{-\theta G(c)}-1)}{e^{-\theta}-1}\right)$
	Crude densities	$f^*(t)$	=	$h_f(t)F(t)\frac{e^{-\theta F(t)}(e^{-\theta G(t)}-1)}{(e^{-\theta}-1)e^{-\theta S(t)}}$
		$g^*(c)$	=	$h_g(c)G(c)\frac{e^{-\theta G(c)}(e^{-\theta F(c)}-1)}{(e^{-\theta}-1)e^{-\theta S(c)}}$
	Spearman's rho	ρ	=	$1-\frac{12}{\theta}\left(\frac{2}{\theta^2}\int_{x=0}^{\theta}\frac{x^2}{e^x-1}dx+\frac{2}{3}\theta-\frac{1}{\theta}\int_{x=0}^{\theta}\frac{x}{e^x-1}dx+\frac{\theta}{2}\right)$
Plackett	Copula	$C_\theta(F(t),G(c))$	=	$\frac{1}{2(\theta-1)}(1+(\theta-1)(F(t)+G(c))-$
				$\sqrt{[1+(\theta-1)(F(t)+G(c))]^2-4F(t)G(c)\theta(\theta-1)}\Big)$
	Crude densities	$f^*(t)$	=	$h_f(t)F(t)\left(\frac{1}{2}-\frac{1+(\theta-1)F(t)+(\theta-1)G(c)-2G(c)\theta}{\left(2\sqrt{(1+(\theta-1)(F(x)+G(c)))^2-4F(x)G(c)\theta(\theta-1)}\right)}\right)$
		$g^*(c)$	=	$h_g(c)G(c)\left(\frac{1}{2}-\frac{1+(\theta-1)G(t)+(\theta-1)F(c)-2F(c)\theta}{\left(2\sqrt{(1+(\theta-1)(G(x)+F(c)))^2-4G(x)F(c)\theta(\theta-1)}\right)}\right)$
	Spearman's rho	ρ	=	$\frac{\theta+1}{\theta-1}-\frac{2\theta\log(\theta)}{(\theta-1)^2}$

function G, density function g and hazard function h_g, and we use the indicator $\xi_i = I\left(X_i = C_i\right)$ for observed censoring due to ART, the likelihood of the observed data $\left\{x_i, \delta_i, \xi_i\right\}$, $i = 1, \ldots, n$, is

$$L = \prod_{i=1}^{n}\left[f^*\left(x_i\right)\right]^{\delta_i}\left[g^*\left(x_i\right)\right]^{\xi_i}\left[S\left(x_i\right)\right]^{1-\delta_i-\xi_i}, \tag{1}$$

with $S(x_i) = C_\theta(F(x_i), G(x_i))$ the joint survival function and f^* and g^* the so-called crude density functions [11]:

$$f^*\left(x_i\right) = \frac{d}{dT}\Pr\left(X < x_i, T < C\right) = -\frac{d}{dT_i}S\left(T_i, C_i\right)|_{T_i=C_i=x_i},$$

$$g^*\left(x_i\right) = \frac{d}{dC}\Pr\left(X < x_i, C < T\right) = -\frac{d}{dC_i}S\left(T_i, C_i\right)|_{T_i=C_i=x_i}.$$

We chose to use the Plackett and Frank copulas because with these copulas, the association between the two marginal distributions can be expressed through a Spearman's correlation coefficient (ρ) and by varying the dependence parameter, both copulas are capable of covering the full range of dependence (from $\rho = -1$ to $\rho = 1$) between the pregnancy and treatment processes. Table III summarizes the dependence structure of these two copula types.

For the parameterizations of the marginal survival distributions F and G, we used four common distributions: Weibull, Gompertz, loglogistic and lognormal. For each of the copula types and each of the combinations of parameterizations ($2 \times 2^4 = 32$ scenarios in total), we optimized the log of the likelihood (1) using the *optim* function in R. Webappendix B (see supporting information) shows the model fit, estimated Spearman correlation coefficient and the estimated cumulative pregnancy chance at 3 years of follow-up for all scenarios. In each of the scenarios, the correlation between the time to pregnancy and the time to treatment was estimated to be negative, and consequently, all estimated 3-year cumulative pregnancy rates were below the 41% estimated previously assuming independence between these events. This again confirmed our hypothesis that the conventional unadjusted models overestimate the pregnancy chances at later time points. The estimated adjusted levels showed to be robust with respect to the chosen copula but varied greatly between the different parameterizations (range 24–41%). Based on the deviance (−2 log likelihood) score, the best performing model is the one using Frank copula with the pregnancy distribution following a Gompertz and the treatment distribution following a loglogistic

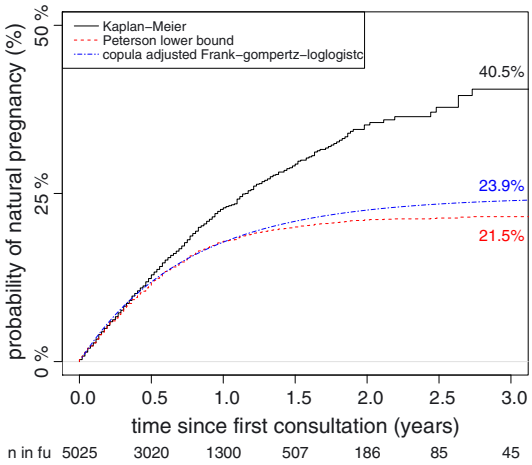

Figure 3. Copula-adjusted cumulative pregnancy curve.

shape. This scenario estimates a strong correlation between the events ($\rho = -0.83$) and a fairly low 3-year pregnancy rate: 24%.

Figure 3 shows the estimated cumulative pregnancy curve over the 3 years of follow-up using this scenario. However, there are some concerns when picking out this scenario as giving the 'best' estimates. There is no clear correlation between the deviances and the pregnancy estimates over the scenarios. Competing models that vary only slightly on their deviance score can vary substantially on their 3-year pregnancy estimate. For instance, the scenario Frank–Weibull–loglogistic has the same deviance as the scenario Plackett–lognormal–loglogistic, but their estimates for the pregnancy chance differ by 7% points. Also, the estimated Gompertz distributions have a decreasing hazard leading to an improper marginal survival function, that is, $\lim_{t\to\infty} F(t) > 0$. The other parameterizations do not allow such limit behavior of the pregnancy distribution. A 'flattening off' of the cumulative pregnancy curve has been observed before in (part of) this dataset [30]. Therefore, other parameterizations that allow for such asymptotic behavior may also give good fit. Given that in all 32 copula scenarios the pregnancy rate was estimated lower than that in the unadjusted analysis, we conclude that an unadjusted analysis assuming independence between the events will very likely overestimate the pregnancy rate. As we could not make out the 'best' scenario, however, we cannot say exactly by what amount.

To assess the impact of the copula adjustment on the estimates of covariate effects, we assessed the prognostic ability of the three tubal patency tests (CAT, HSG and DLS) on pregnancy. To this end we included these covariates in the marginal distributions of both the pregnancy and the treatment event. For the Weibull and the Gompertz distributions, we expressed the prognostic ability of the tubal tests as HRs. These were estimated by modeling the log of the scale parameter of the distributions as a linear function of the tubal test. For the lognormal and the loglogistic distribution, we expressed the prognostic ability of the tests as time ratios (TRs). The mean of the log time and the log of the scale parameter for the lognormal and loglogistic distributions, respectively, were modeled as linear functions of the tubal test result.

The results of optimizing the log of the likelihood (1), with the tubal test effects added to the density and survival functions, are shown in webappendix C (see supporting information). The HRs reported in the scenarios using a Weibull or a Gompertz shape of the pregnancy distribution were compared with the HRs calculated previously in Cox models assuming independence of the events (CAT 0.68, HSG 0.73 and DLS 0.80). The TRs reported for the scenarios using a loglogistic or a lognormal shape of the pregnancy distribution could not be compared with HRs. We compared these with TRs estimated with parametric models assuming independence of the events using a loglogistic/lognormal pregnancy distribution (CAT 0.62/0.57, HSG 0.73/0.73 and DLS 0.79/0.81). In the HR models, we in general observed the same trend as we did in the IPCW-adjusted model: the adjusted models estimated slightly weaker prognostic ability for the CAT and HSG tests than the unadjusted models. In the TR models, the differences between the CAT and HSG effects in the models assuming dependence and the models assuming independence were more diffuse. For the DLS, the majority of dependent scenarios estimated a stronger effect than the independent models. Scenarios with the best deviance score, however, estimated similar prognostic strength of the DLS as the unadjusted model. For all comparisons, the differences between the unadjusted

and adjusted models were small in comparison with the confidence intervals. We therefore conclude that the dependence assumption does influence the estimation of the prognostic tubal tests somewhat, but we cannot obtain a clear direction nor size of the effect.

5. Discussion

In a dataset consisting of over 5000 couples with unfulfilled child wish of whom by the end of the 3-year follow-up period more than half had started assisted treatment, we assessed the consequences of adjusting for the dependence between the pregnancy event and the competing treatment event with an IPCW analysis and by using copulas. Both approaches make quite different assumptions on the dependence structure between the events. The IPCW method relies on the measured prognostic factors and correct model specification. In our dataset, the most important known predictors of natural pregnancy were present [18], and therewith, we believe that we captured the wanted predictors for both pregnancy and treatment. However, only two (HSG and DLS) were collected in a time-dependent manner, limiting the discriminating capacity of the censoring weights over time. Also, there may exist unknown predictors of pregnancy that influenced the treatment decisions. The copula method necessitates a parameterization of the event distributions and of the dependence structure to allow the estimation of the marginal distributions.

Despite their different approaches, both methods consistently showed that the two events are negatively correlated and that analyses ignoring this dependency give too optimistic pregnancy chances at later follow-up times. This result confirmed our hypothesis. The adjustment methods were unfortunately not able to relieve the identifiability problem as both methods rely on ultimately untestable assumptions and the adjusted estimates were not stable enough to give an exact adjusted estimate. We can, however, use the results of both methods to obtain a plausible estimation region that is far more narrow than the Peterson bounds. For the 3-year pregnancy rate, this implies a range of plausible chances ranging from 24% to 41%, compared with 22% to 97% resulting from the Peterson bounds. The range of 24% to 41% may, however, still be too wide for use in clinical practice. At earlier time points, where less treatment censoring has occurred, bounds will be narrower.

Adjusting for the dependent competing risk of treatment also influenced the estimation of covariates. For the CAT and HSG tests, both methods showed small and diffuse effects of the dependency adjustment. For the DLS test, the results were more consistent. As this test is considered as the most accurate test for tubal pathology, the fact that the prognostic effect of this test remained or was more pronounced in the adjusted models is understandable from a clinical point of view. It was, however, not anticipated based on our a priori assumptions. These results illustrate that the effect of a dependent competing risk on covariates cannot be foreseen and should always be assessed in formal sensitivity analyses [27]. When covariates are available, we advise to use both IPCW adjustment and copula techniques as their joint results give a more complete picture of the influence of a dependent competing risk.

If one would have more information about the hypothetical natural pregnancy chance of couples who started assisted treatment, estimates could be improved. In classical competing risk settings where the competing risk is death due to a certain cause, obtaining such 'after competing event' information is impossible. In our application, however, there may be data available that contain the desired information. For instance, data from couples who have made the decision to start treatment but who were placed on a waiting list and therefore could be followed for natural pregnancy for a longer period [31] could be used. Also data from couples who ceased assisted treatment and returned on an expectant management strategy might be of use, although selection may have occurred in such a group. Incorporating these kinds of 'after competing event' data and using improper pregnancy distributions that account for part of the population that will never conceive naturally [32] are considered promising extensions for future work.

References

1. Moeschberger ML, Klein JP. Statistical methods for dependent competing risks. *Lifetime Data Analysis* 1995; **1**:195–204.
2. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* 2007; **26**:2389–2430.
3. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 1999; **94**:496–509.
4. Klein JP, Andersen PK. Regression modelling of competing risk data based on pseudovalues of the cumulative incidence function. *Biometrics* 2005; **61**:223–229.

5. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences* 1975; **72**:20–22.
6. Carriere JF. Removing cancer when it is correlated with other causes of death. *Biometrical Journal* 1995; **35**:339–350.
7. Zheng M, Klein P. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* 1995; **82**:127–138.
8. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log rank tests. *Biometrics* 2000; **56**:779–788.
9. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 2008; **168**:656–664.
10. Matsuyama Y, Yamaguchi T. Estimation of the marginal survival time in the presence of dependent competing risks using inverse probability of censoring weighted (IPCW) methods. *Pharmaceutical Statistics* 2008; **7**:202–214.
11. Escarela G, Carriere JF. Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research* 2003; **12**:333–349.
12. Di Serio C. The protective impact of a covariate on competing failures with an example from a bone marrow transplantation study. *Lifetime Data Analysis* 1997; **3**:99–122.
13. Huang X, Wolfe RA. A frailty model for informative censoring. *Biometrics* 2002; **58**:510–520.
14. Larson MG, Dinse GE. A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society Series C* 1985; **34**:201–211.
15. Siannis F, Copas J, Lu G. Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics* 2005; **6**:77–91.
16. Van Geloven N, Broeze KA, Bossuyt PM, Zwinderman AH, Mol BW. Treatment should be considered a competing risk when predicting natural conception in subfertile women. *Human Reproduction* 2012; **27**:889–95.
17. Van der Steeg JW, Steures P, Eijkemans MJ, Habbema JD, Hompes PG, Broekmans FJ, van Dessel HJ, Bossuyt PM, van der Veen F, Mol BW. Pregnancy is predictable: a large-scale prospective external validation of the prediction of spontaneous pregnancy in subfertile. *Human Reproduction* 2007; **22**:536–542.
18. Hunault CC, Habbema JD, Eijkemans MJ, Collins JA, Evers JL, te Velde ER. Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three previous models. *Human Reproduction* 2004; **19**:2019–2026.
19. Peterson AV. Bounds for a joint distribution function with fixed sub-distribution functions: applications to competing risks. *Proceedings of the National Academy of Sciences* 1976; **73**:11–13.
20. Dignam JJ, Weissfeld LA, Anderson SJ. Methods for bounding the marginal survival distribution. *Statistics in Medicine* 1995; **14**:1985–1998.
21. Howe CJ, Cole SR, Chmiel JS, Muñoz A. Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. *American Journal of Epidemiology* 2011; **173**:569–577.
22. Grunkemeier GL, Jin R, Eijkemans MJ, Takkenberg JJM. Actual and actuarial probabilities of competing risks: apples and lemons. *Annals of Thoracic Surgery* 2007; **83**:1586–1592.
23. Tjon-Kon-Fat RI, Lar DN, Steyerberg EW, Broekmans FJ, Hompes P, Mol BWJ, Steures P, Bossuyt PMM, Van der Veen F, van der Steeg JW, Eijkemans MJC. Inter-clinic variation in the chances of natural conception of subfertile couples. *Human Reproduction* 2013; **28**:1391–1397.
24. van der Wal WM, Geskus RB. ipw: An R package for inverse probability weighting. *Journal of Statistical Software* 2011; **43**:1–23.
25. Yoshida M, Matsuyama Y, Ohashi Y. Estimation of treatment efect adjusting for dependent censoring using the IPCW method: an application to a large primary prevention study for coronary events (MEGA study). *Clinical Trials* 2007; **4**:318–328.
26. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health* 2006; **60**:578–586.
27. Huang X, Zhang N. Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. *Biometrics* 2008; **64**:1090–1099.
28. Chen YH. Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *Journal of the Royal Statistical Society B* 2010; **72**:235–251.
29. Nelsen RB. *An Introduction to Copulas*. Springer: New York, 1999.
30. Van Geloven N, Van der Veen F, Bossuyt PMM, Hompes PG, Zwinderman AH, Mol BW. Can we distinguish between infertility and subfertility when predicting natural conception in couples with an unfulfilled child wish? *Human Reproduction* 2013; **28**:658–665.
31. Eijkemans MJ, Lintsen AM, Hunault CC, Bouwmans CA, Hakkaart L, Braat DD, Habbema JD. Pregnancy chances on an IVF/ICSI waiting list: a national prospective cohort study. *Human Reproduction* 2008; **23**:1627–1632.
32. Li Y, Tiwari RC, Guha S. Mixture cure survival models with dependent censoring. *Journal of the Royal Statistical Society B* 2007; **69**:285–306.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site.