The current duration (backward recurrence time) approach to estimating the distribution of time to pregnancy

Niels Keiding
Department of Biostatistics
University of Copenhagen
Denmark

Joint work with

Rémy Slama and his group (Grenoble, Paris)

Oluf H. Hansen and Ditte Nørbo Sørensen (Copenhagen)

and many others, as you will see.

Overview

Time-to-pregnancy: why?

Prospective design

    Giving up as independent censoring?
    Variants of the prospective design

Retrospective design

    Artefacts from the retrospective conditioning

Current duration design

    Theory
        Nonparametric estimation
        Parametric models

    Covariates

        Accelerated Failure Time Regression for Current Duration design

    Observatoire epidemiologique de la fertilité en France

        Censoring problems
        Fit of parametric models

Other uses of the current duration design

Time to pregnancy

The time from a couple decides they want to become pregnant ("initiation") until they succeed. This is regarded as one of the most precise indicators of biological fecundity.

Survival analysis

Right censoring at *t*: We know that *TTP>t* but not the exact value

Left truncation (delayed entry) at *t*:
         we only include the woman from *t*:
         if her $TTP \leq t$ we would not know about her

Right truncation at *t*: we only include her if her $TTP \leq t$.
         Otherwise we would not know about her

Note important distinction between censoring and truncation

Designs in Time To Pregnancy (*TTP*) studies

Prospective: Follow-up   *from*   initiation

                                        to   pregnancy   "event"

                                        or   end of study   right censoring

                                        or   give up   right censoring*

Conceptually simple

* is "give up" really independent censoring?

Practically very difficult

Prospective sampling: example

Danish study of 430 couples, no children, just starting attempts to conceive.

Followed from initiation for six cycles or until pregnant

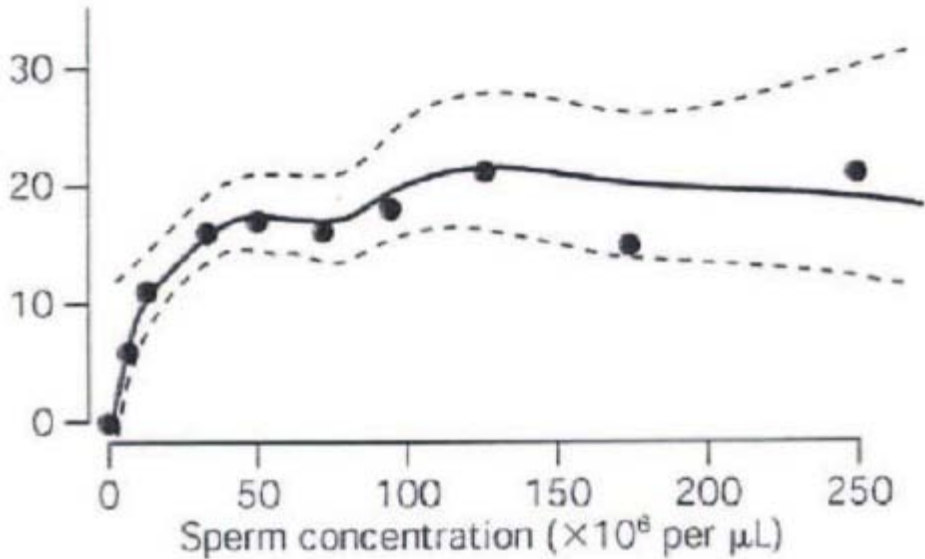|		
	22 terminating before six cycles	*right censoring*
At enrolment	demography, medical	*time-fixed covariates*
	occupational, lifestyle	X
	semen sample	
	blood sample	
Each cycle t	semen sample	*time-dependent covariates*
	female diary: bleeding intercourse	X_t
	female urine samples	

Probability of pregnancy in a menstrual cycle

Sperm concentration ($\times 10^6$ per μL)

Prospective sampling: variants

Historically prospective: "describe your first/most recent attempt to get pregnant"

Delayed entry: identify from cross-sectional sample those currently trying. Follow them up: *prevalent cohort study*.

Example from France: later

Designs in TTP studies

Retrospective: Ask at maternity clinic about how long it took

Practically easy

Conditional on pregnancy

Sterile unrepresented

Low-fecund under-represented

Truncation problems when studying calendar time trends

Retrospective sampling: example

TK Jensen, T Scheike, N Keiding, I Schaumburg, P Grandjean (2000). Selection bias in determining the age dependence of waiting time to pregnancy. *Am J Epidemiol* **152**, 565-572.

40,666 interviews at Odense University Hospital 1972-87

Gestational week 20   Condition on $TTP < \infty$
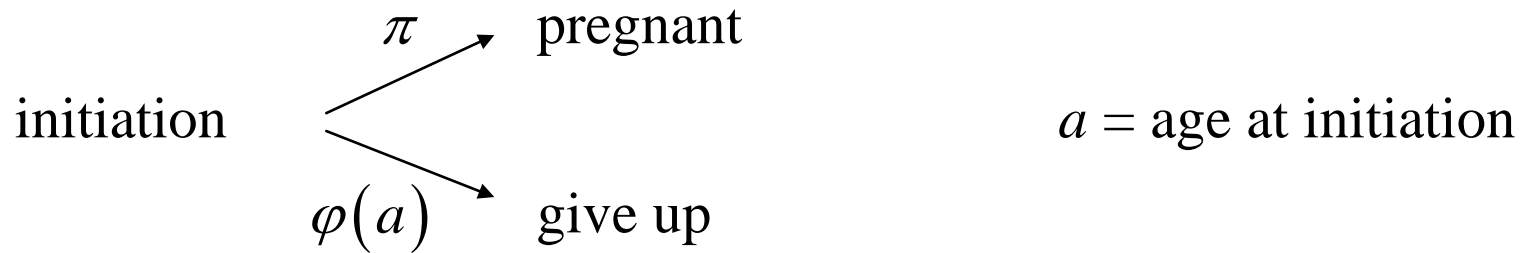
Right truncation

Consider only $TTP \leq 13$ because

1. $TTP > 13$ less reliable
2. medical intervention becomes more common after cycle 13

Right censoring

The retrospective design *conditions* on not having given up

Simplified situation

$$\text{initiation} \quad \begin{array}{c} \xrightarrow{\pi} \quad \text{pregnant} \\ \\ \xrightarrow{\varphi(a)} \quad \text{give up} \end{array} \qquad a = \text{age at initiation}$$

Let T = time to pregnancy, U = time to giving up. We observe conditional distribution of T given $T < U$ (we do *not* observe U!).

Joint distribution (T, U): $f(t, u) = \pi\varphi\, \mathrm{e}^{-\pi t - \varphi u}$

Conditional distribution $(T, U \mid T < U)$: $(\pi + \varphi)\varphi\, \mathrm{e}^{-\pi t - \varphi u} \qquad t < u$

$$0 \qquad\qquad\qquad t > u$$

Observed distribution $T \mid T < U$: $(\pi + \varphi)\mathrm{e}^{-(\pi + \varphi)t}$

so if φ increases with a it will look as if *TTP* decreases with a.

Fecundability odds ratio (FR) among women in different age categories with different parity after control for the age of the father, the cycle at which the women became pregnant, occupation of both parents, municipality, number of previous spontaneous or induced abortions, body mass index before pregnancy, duration and regularity of menstrual cycle, smoking status of the women during pregnancy, use of oral contraceptives as last method of birth control, and the recording secretary, Odense University Hospital, Denmark, 1972-1987.

Age of women (years)	All women			Nulliparous women			Multiparous women		
	No. of cycles	FR mother	95% CI*	No. of cycles	FR mother	95% CI	No. of cycles	FR mother	95% CI
<22	18,180	1		16,432	1		1,748	1	
22–23	12,097	1.11	1.01, 1.22	9,944	1.11	1.00, 1.23	2,153	1.27	1.00, 1.61
24–25	10,691	1.20	1.09, 1.33	7,895	1.19	1.06, 1.33	2,796	1.45	1.15, 1.83
26–27	7,995	1.18	1.06, 1.32	4,685	1.21	1.05, 1.39	3,310	1.37	1.08, 1.73
28–29	4,831	1.23	1.07, 1.40	2,202	1.26	1.06, 1.51	2,629	1.40	1.09, 1.79
30–31	2,778	1.20	1.03, 1.41	987	1.33	1.05, 1.68	1,791	1.36	1.04, 1.79
32–33	1,447	1.36	1.12, 1.65	384	1.34	0.95, 1.89	1,063	1.62	1.21, 2.18
34–35	582	1.39	1.06, 1.82	117	1.58	0.90, 2.77	465	1.60	1.11, 2.30
36–37	188	2.34	1.56, 3.52	38	1.85	0.82, 4.17	150	3.14	1.88, 5.24
≥38	169	3.14	1.95, 5.04	45	1.64	0.67, 4.01	124	5.75	3.08, 10.72

* CI, confidence interval.

Included: Birth during study period

Not included: Birth outside of study period

TTP+gestation

Calendar time

jan. 1972

jan. 1988
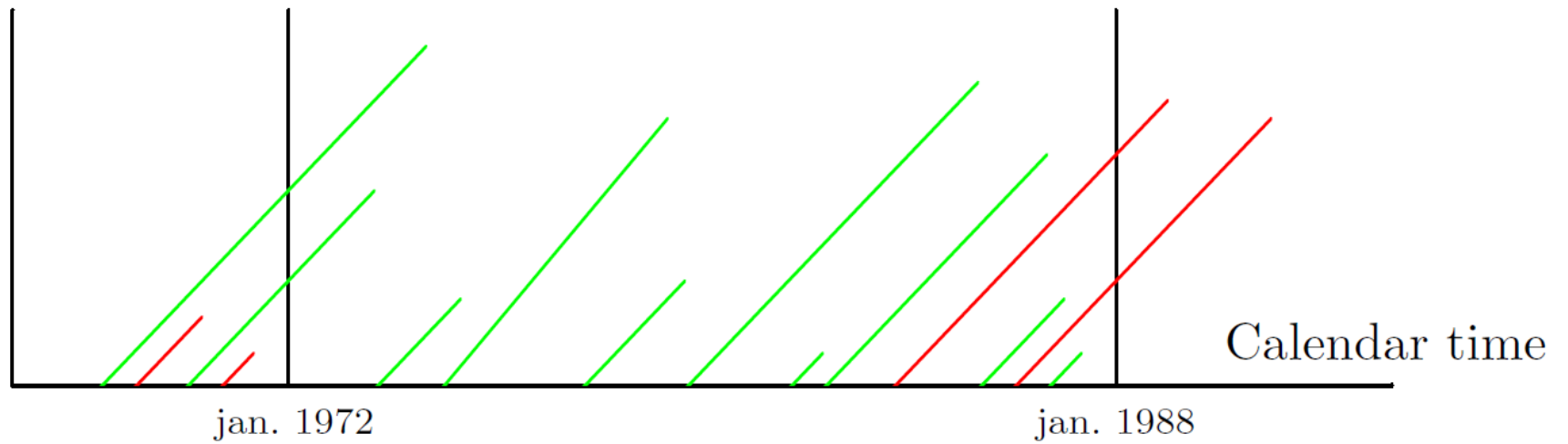
Overrepresentation
of long TTP:
left truncation

Overrepresentation
of short TTP:
right truncation

Current duration design

C.S. Weinberg & B.C. Gladen (1986). *Biometrics* **42**, 547-560.

N. Keiding, K. Kvist, H. Hartvig, M. Tvede, S. Juul (2002). *Biostatistics* **3**, 565-578.

Ask a cross-sectional sample of women

Are you currently attempting to become pregnant?

If yes, for how long have you attempted?

Theory

T time to pregnancy

U time to discontinuation without pregnancy

     (death, lost interest, partner left, become too old)

V time to end of follow-up

Y current duration

Prospective: T, possibly right-censored at $U \wedge V$

                        (very hard to observe $T = 0$)

Retrospective: $T \mid T < U$

Current duration: $Y =$ Backward recurrence time of $T \wedge U$

                        in conditional distribution given $T > 0$.

Prevalent cohort: T, left-truncated at $Y \wedge V$,

                        possibly right-censored at $U \wedge V$.

Backward recurrence time of $T \wedge U$

$X = T \wedge U$ time to (recognized) pregnancy or discontinuation of attempt density $f(x)$ (so assume no mass at 0),

survival function $S(x) = \int_x^\infty f(a)\,da$

mean $\mu_X = \int_x^\infty S(x)\,dx$ assumed finite.

Initiation according to Poisson process in calendar time t with intensity $\beta(t)$. Cross-sectional sample at $t_0 \Rightarrow$ density $g(y, t_0)$ of observed current duration $Y = X \wedge V$ is proportional to $\beta(t_0 - y) S(y)$.

Under *steady state* $\beta(t) = \beta$,

$g(y, t_0) = g(y) = S(y)/\mu_X$ *decreasing*, $\quad 0 < g(0) < \infty$.

$\qquad\qquad$ Estimate $\hat{g}(y)$, then $\hat{S}(x) = \hat{g}(x)/\hat{g}(0)$.

Nonparametric estimation of decreasing density

g decreasing, $0 < g(0+) < \infty$

$\hat{g}(y)$ isotonic estimator (pool-adjacent-violators algorithm)

Grenander	*Skand.Akt.Tidskr.* 1956	complete data
Denby & Vardi	*Technometrics* 1986	right-censored data
Woodroofe & Sun	*Statistica Sinica* 1993	show that

$P \lim \hat{g}(0+) > g(0+)$ **inconsistency**; suggest penalized NPMLE

Sun & Woodroofe	*J.Statist.Plann.Inf.* 1996	adaptive penalized NPMLE
van Es et al.	*J.Statist.Plann.Inf.* 2000	regression model
Kulikov & Lopuhaä	*Ann. Statist.* 2006	NPMLE behaviour near 0
Pal	*Scand. J. Statist.* 2009	as.distr. of penalized LR
Balabsdaoui et al.	*Univ. Washington Tech. Rep. 2009*	inconsistency at 0 indicates that Grenander estimator does not know whether $g(0+)<\infty$

Penalized NPMLE (Sun and Woodroofe)

NPMLE on transformed data points

$$\alpha + \gamma y_k \qquad k = 1, \cdots, n$$

$$\gamma = \min_{k=1,\cdots,n} \left(1 - \frac{\alpha k/n}{\alpha + \gamma \ y_k} \right)$$

$$\alpha = 0.649 \beta^{-1/3} n^{-2/3}$$

$$\beta = -\frac{1}{2} g(0) g'(0)$$

$$= \frac{1}{2} \lambda^3 \text{ if } Y \text{ exponential with intensity } \lambda$$

Alternative approach to $\hat{g}(0+):\hat{g}(\varepsilon)$

Hans van Houwelingen (pers. comm. 2001)

Kulikov & Lopuhaä 2006

For any $\varepsilon > 0,\ \hat{g}(\varepsilon)$ is consistent. So choose ε small

($\varepsilon \to 0$ as $n \to \infty$ whatever that means in practice).

Parametric models: Pareto distribution

Current duration Y has density

$$g(y) = \frac{\lambda\mu}{(1+\mu y)^{\lambda+1}}$$

and survival function

$$S_g(y) = \frac{1}{(1+\mu y)^{\lambda}}$$

$\Rightarrow TTP$ has survival function

$$S(x) = \frac{g(x)}{g(0)} = \frac{1}{(1+\mu y)^{\lambda+1}}$$

So if Y Pareto (λ, μ) then X Pareto $(\lambda+1, \mu)$

Note: μ is scale parameter, λ is shape parameter.

Pareto as mixture of exponentials

If $Y|\theta$ is exponential with survival function $e^{-\theta y}$ and θ is gamma (λ, μ) then Y is Pareto (λ, μ).

Continuous time version of Weinberg & Gladen (1986) who worked in discrete time with a beta-mixture of geometric distributions.

Parametric models: Generalized gamma model

V.T. Farewell & R.L. Prentice (1977). *Technometrics* **19**, 69-75.

Yamaguchi (2003). *Sociological Methodology* **33**, 81-110.

Assume that the underlying density of

$$W = \frac{1}{\sigma}\left(\log X - \theta\right)$$

is log-gamma with shape parameter λ, i.e. density

$$h(w) \quad = \quad \frac{|\lambda|}{\Gamma\left(\lambda^{-2}\right)}\left(\lambda^{-2}\right)^{\lambda^{-2}} \exp\left[\lambda^{-2}\left(\lambda w - e^{\lambda w}\right)\right] \ , \quad \lambda \neq 0$$

$$= \quad \frac{1}{\sqrt{2\pi}}\exp\left(-w^2/2\right) \ , \qquad\qquad \lambda = 0.$$

Note: θ is a scale parameter for X; λ and σ are shape parameters.

Illustrations of the inconsistency of $\hat{g}(0+)$

Generalized gamma distribution with realistic parameter values (see later). 10 simulated distributions with $n = 1000$ in each.

Shown are

    True density

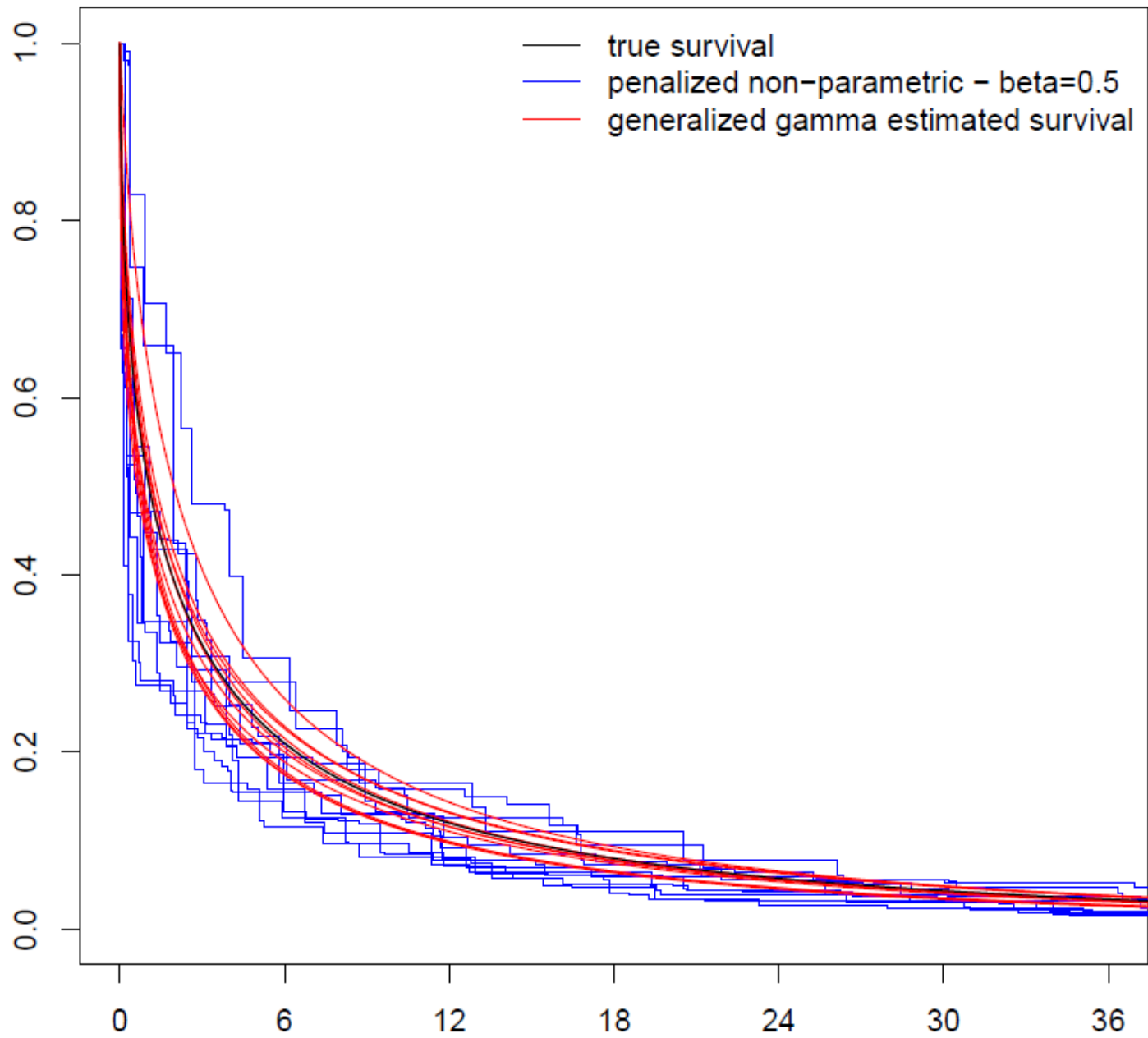   Estimated density (parametric MLE) for each replication

1. NPMLE

2. NPMLE conditioned on $T \geq 1$

3. Penalized NPMLE according to Sun & Woodroofe (1996)

4. Penalized NPMLE with Sun & Woodroofe's β arbitrarily fixed at 0.5

Conditioned on T ≥ 1

——	true survival
——	non-parametric
——	generalized gamma estimated survival
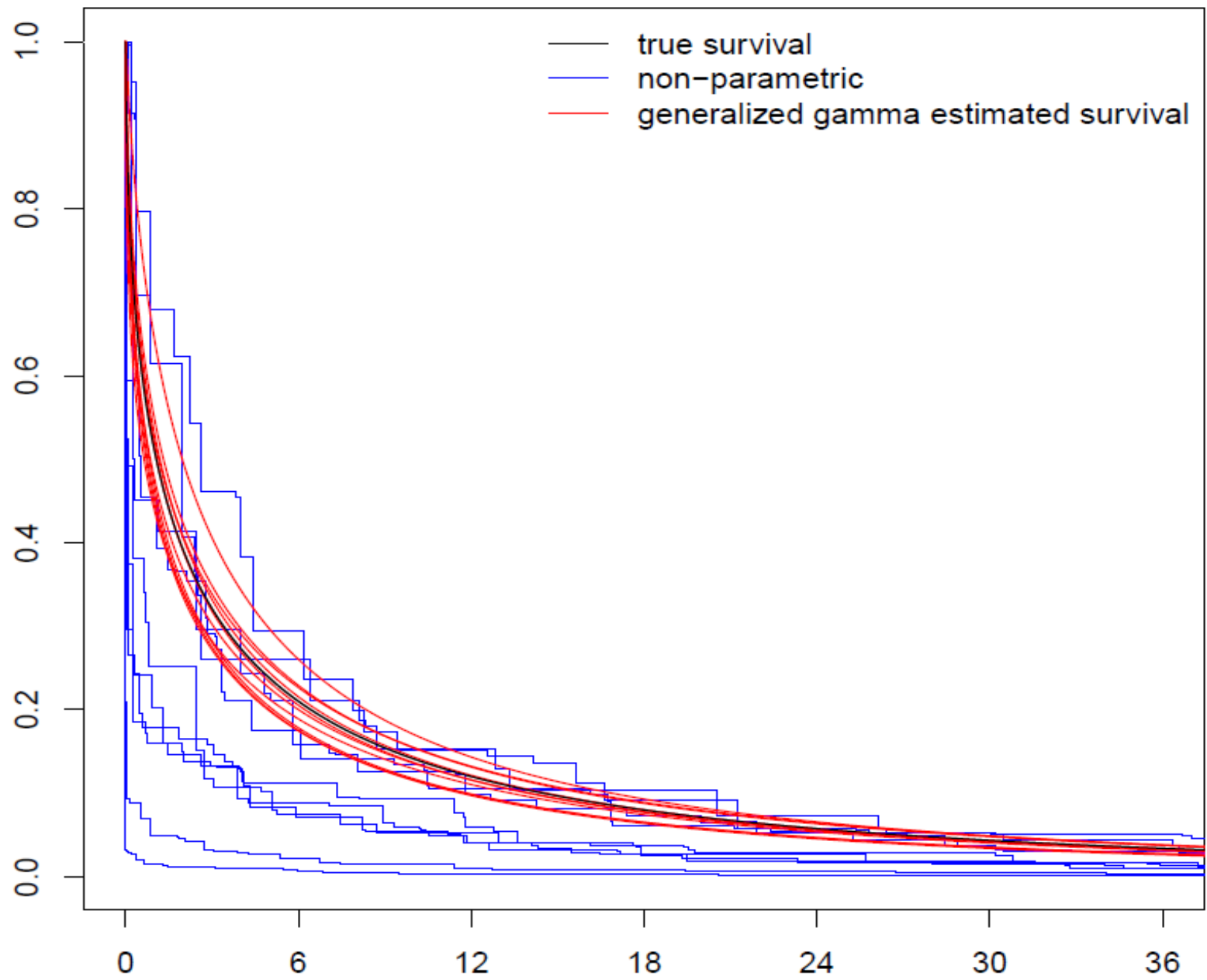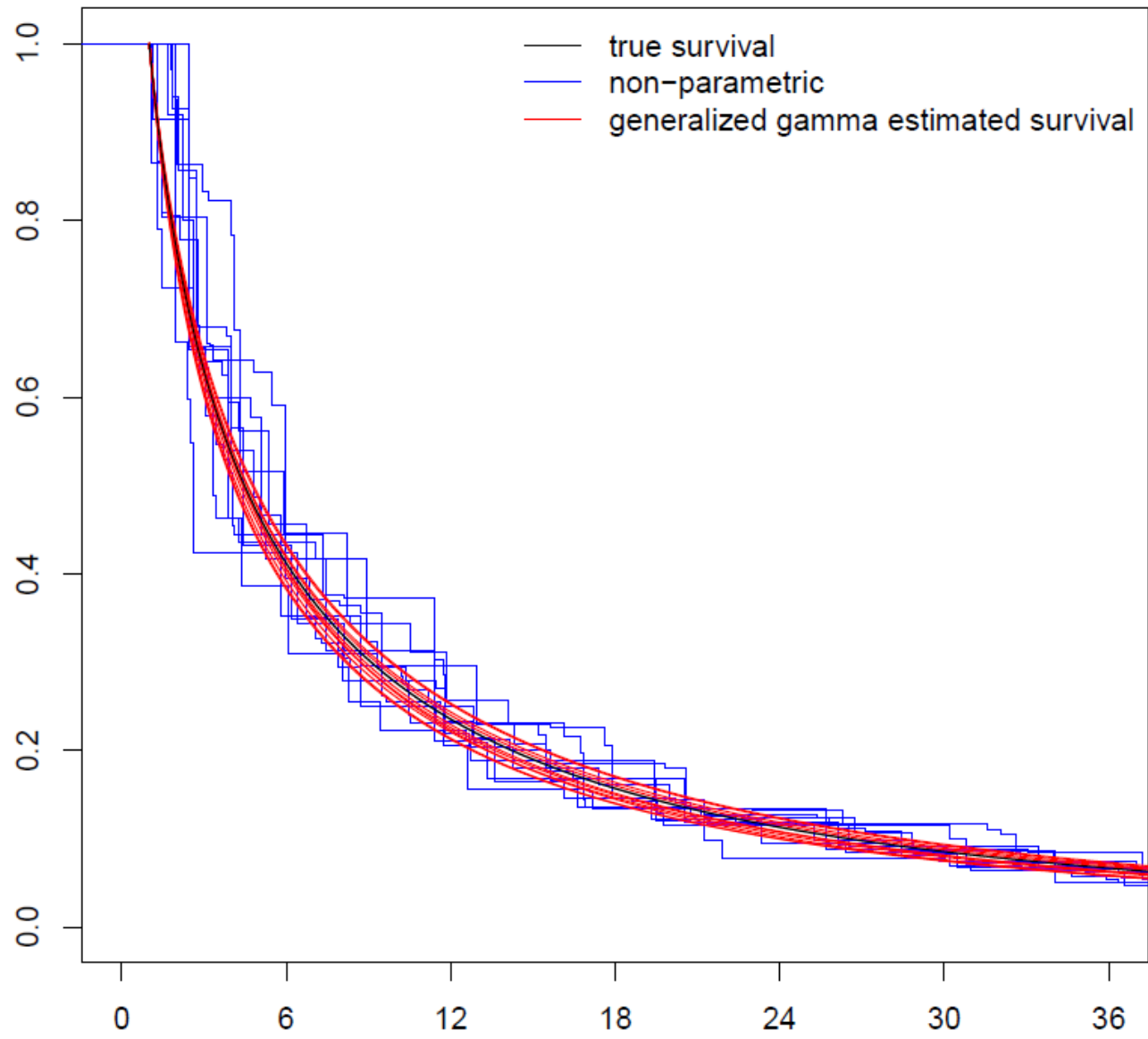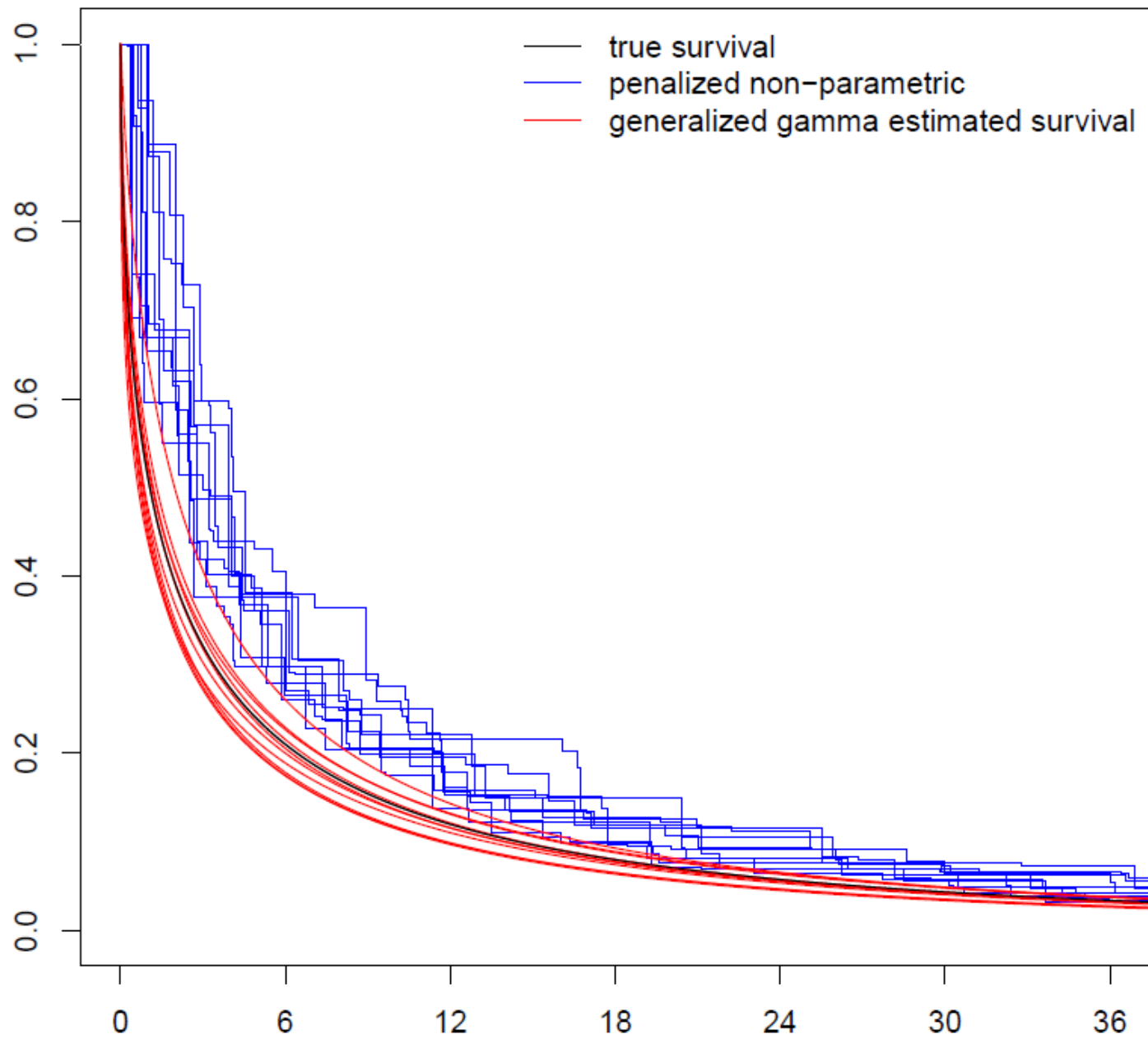
Current duration with covariates

$X = T \wedge U$ density $f(x|z)$

Assume (X, Z) independent of V. As above density $g(y|z)$ of current durations

$$g(y|z) = \frac{S(y|z)}{E(X|z)}$$

so that $S(x|z)$ may be estimated by $\hat{g}(x|z)/\hat{g}(0|z)$

Current duration and unobserved heterogeneity

Let $h(z)$ distribution of Z in population,

i.e. $f(x,z) = f(x|z)h(z)$ joint density of (X,Z).

Recall $g(y|z) = S(y|z)/E(X|z)$.

Density of (X,Z) in current duration sample is

$$x f(x,z)/E(X) \qquad\qquad length-biased$$

$\Rightarrow$ density of Z in sample: $h(z)E(X|z)/E(X)$

$\Rightarrow$ (marginal) density of observed current durations

$$g(y) = \int g(y|z)\frac{h(z)E(X|z)}{E(X)}\,dz = \frac{1}{E(X)}\int S(y|z)h(z)\,dz$$

so that $\hat{g}(x)/\hat{g}(0)$ estimates $\int S(x/z)h(z)dz = S$ also under unobserved heterogeneity.

Regression models for current duration data

Yamaguchi (2003). *Sociological Methodology* **33**, 81-110.
Keiding, Fine, Hansen, Slama (2011). *Stat. Prob. Letters* **81**, 724-729.

$X = T \wedge U$ time to (registered) pregnancy or discontinuation of attempt

$Y = X \wedge V$ observed current duration

z_i covariate for couple i, assume *stationarity* of covariates and of initiating process

$X \mid z$ density f, survival function S, $y \mid z$ density g

$$S(x \mid z) = g(x \mid z) \big/ g(0 \mid z), \; g(y \mid z) = S(y \mid z) \big/ E(X \mid z)$$

Assume $Y \mid z$ *accelerated failure time:*

$$P(Y > y \mid z) = S_0\left(y e^{\beta z}\right)$$

S_0 baseline survival function with density g_0 (parametric or wider class of nonincreasing g_0).

Accelerated failure time model for current duration data

$$Y|z \ AFT: \quad P(Y > y|z) = S_0\left(y \ e^{\beta z}\right)$$

$$g(y|z) = g_0\left(y \ e^{\beta z}\right)e^{\beta z}$$

$$\Rightarrow \ P(X > x|z) = S(x|z) = \frac{g(x|z)}{g(0|z)} = \frac{g_0\left(x \ e^{\beta z}\right)}{g_0(0)}$$

which is again an *AFT* model with *the same regression coefficients* β and a new baseline survival function $g_0(\cdot)/g_0(0)$. So effects of covariates directly estimable from observed current durations.

Examples: Pareto distribution, $\mu = e^{\beta z}$.

Generalized gamma distribution, $\theta = \beta z$.

Semiparametric inference

R. Mukherjee (2006). On accelerated failure time models for forward and backward recurrence times. Ph.D. thesis, University of Wisconsin – Madison.

P.J. Mokveld (2007). The accelerated failure time model under cross sectional sampling schemes. Ph.D. thesis, University of Amsterdam.

Observatoire Epidémiologique de la fertilité en France

Field test of *current duration* design and *prevalent cohort* design.

Feasibility study:

Slama, Ducot, Carstensen, Lorente, de la Rochebrochard, Leridon, Keiding, Bouyer (2006) *Epidemiology* **17**, 440-449.

Women were eligible if they

- were aged 18 to 44 at the interview, and

- currently had a male partner, and

- declared not to be pregnant and had not given birth to a live (or stillborn) baby in the last 3 months before the interview, and

- declared to be trying to become pregnant and/or did not use any birth control

Interview

Eligible women were asked

- for how long they had been trying to become pregnant (and/or not been using birth control) – ie. current duration of TTP

- menstrual cycle

- frequency of sexual intercourse

- tobacco consumption

- contraception and reproductive problems

- ...

Results

Inclusion (first interview)

- 64 224 contacts

- 56 864 contacts without direct refusal

- 997 eligible women – "trying to become pregnant" (GROUP A)

- 13 885 not trying to become pregnant, but eligible according to the other inclusion criteria (GROUP B)

- ... of these 4067 were selected for follow-up

Follow-up will allow prevalent cohort study.

Follow-up

Follow-up concerns the two groups of women

- women trying to become pregnant at first interview (GROUP A)

- women <u>not</u> trying to become pregnant at first interview, but were eligible according to the other inclusion criteria (GROUP B)

GROUP B (n=4067)
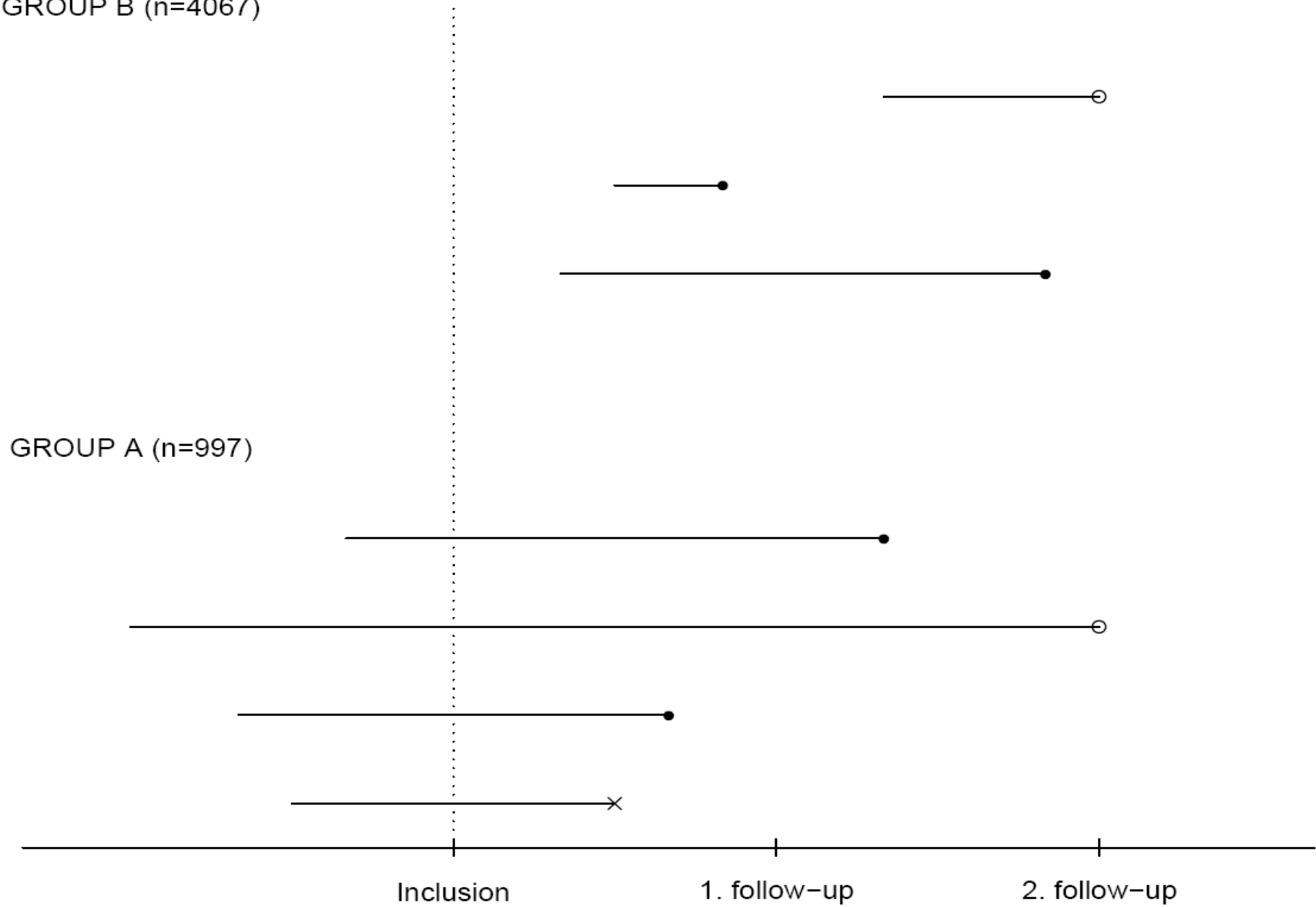
GROUP A (n=997)

Inclusion   1. follow-up   2. follow-up

Censoring problems I: giving up

Basic competing risks situation

$$
\text{initiation} \quad
\begin{cases}
\nearrow & \text{pregnant} \\
\searrow & \text{give up}
\end{cases}
$$
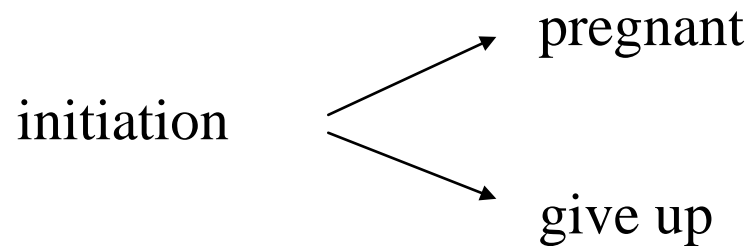
Current duration approach will necessarily concern the minimum of time to pregnancy T and time to giving up U. Note that if T and U are independent exponential with hazards π and φ then the distribution of $T \wedge U$ is exponential $(\pi + \varphi)$. So

$$T \wedge U \text{ has same distribution as } T \mid T < U$$

and the discussion from the retrospective approach applies again.

Censoring problems II:
unrealistically long current durations

The target in the telephone interview was

Current Duration of Unprotected Intercourse (CDUI)

and there are serious problems with these for long periods. This is of course particularly sensitive with parametric models.

Long times-to-pregnancy (*TTP*) are in any case hard to interpret.

Decision: Report only *TTP* distributions over [0, 36] months.

Use either
  - CDUI artificially censored at 36 months
or
  - CDUI truncated at 36 months (see next slide why that is reasonable)

Artificially truncating current duration at y_0.

$g(y)$ density of current duration Y

$S(x) = g(y)/g(0)$ survival function of $X = T \wedge U$

$g_0(y) = g(y)/G(y_0) =$ density of Y truncated at y_0 , $y < y_0$

$$\frac{g_0(x)}{g_0(0)} = \frac{g(x)/G(y_0)}{g(0)/G(y_0)} = S(x) \ , \qquad x < y_0$$

so the artificially truncated distribution of current durations will lead to correctly estimated distribution of *TTP* before the truncation point.

Censoring problems III: fertility treatment
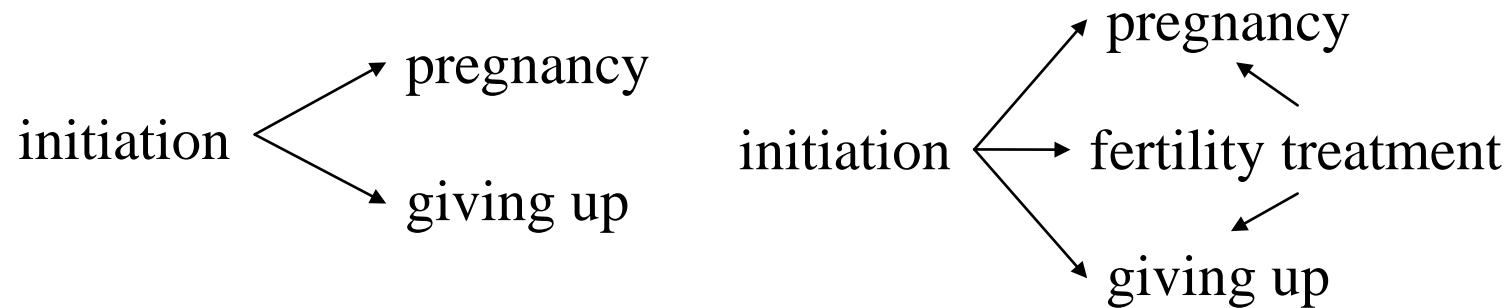
Two possibilities for current duration analysis
- include couples only until they start fertility treatment
  – corresponds to studying $T \wedge U \wedge F$, F = time to fertility treatment.
  Of the 997 women still trying, 745 had not yet started fertility treatment.

- include all couples still trying (target $T \wedge U$) – corresponds to regarding fertility treatment as an integrated component of life today and focusing on the marginal TTP in our society as it is.
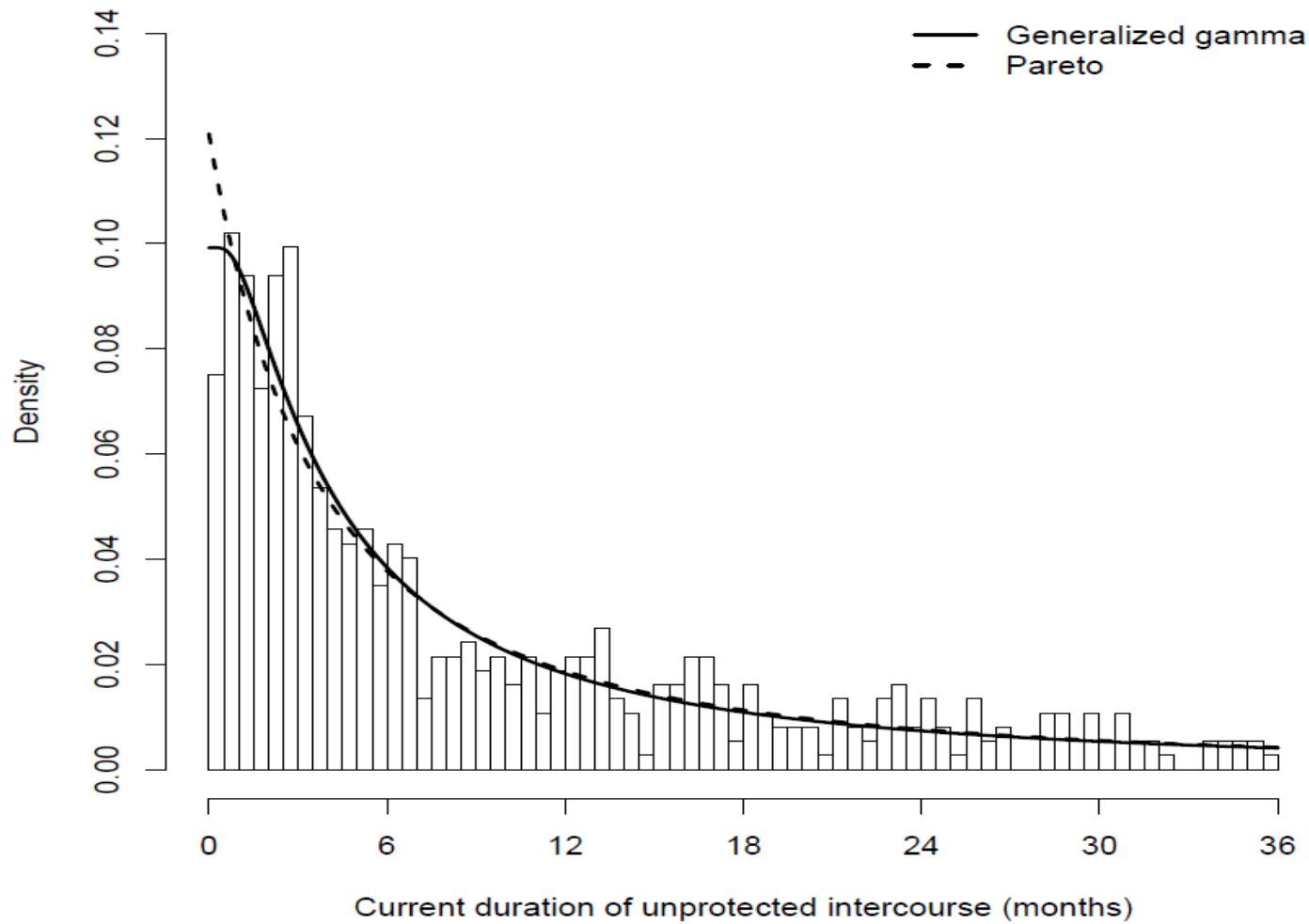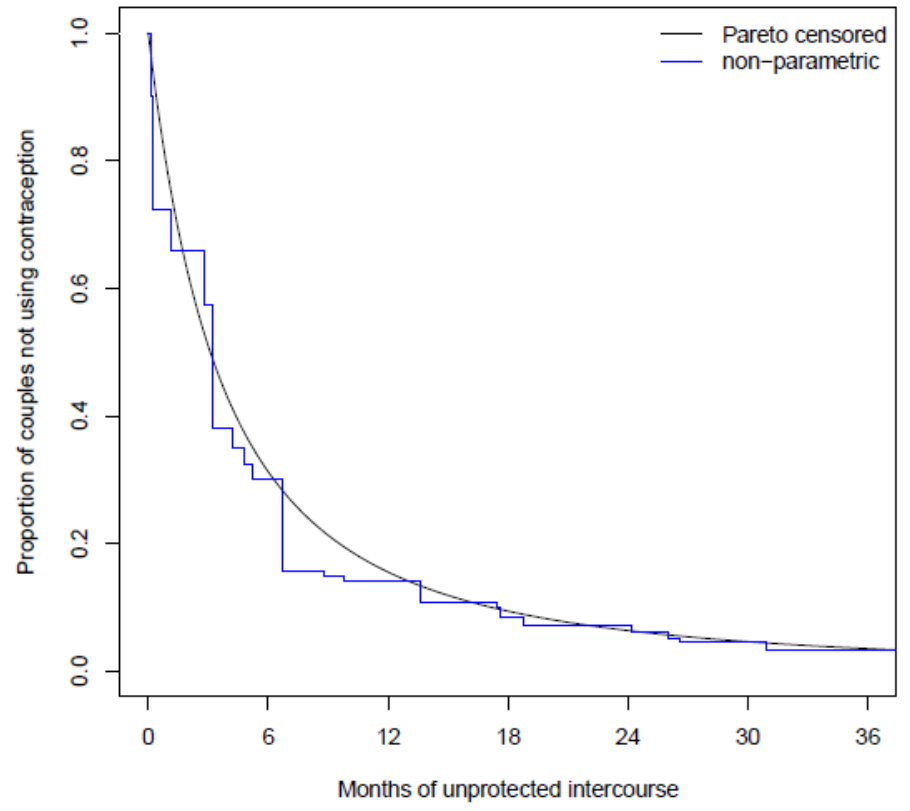
Observed current durations of 745 couples still trying and not having sought fertility treatment, with fitted generalized gamma and Pareto distributions.
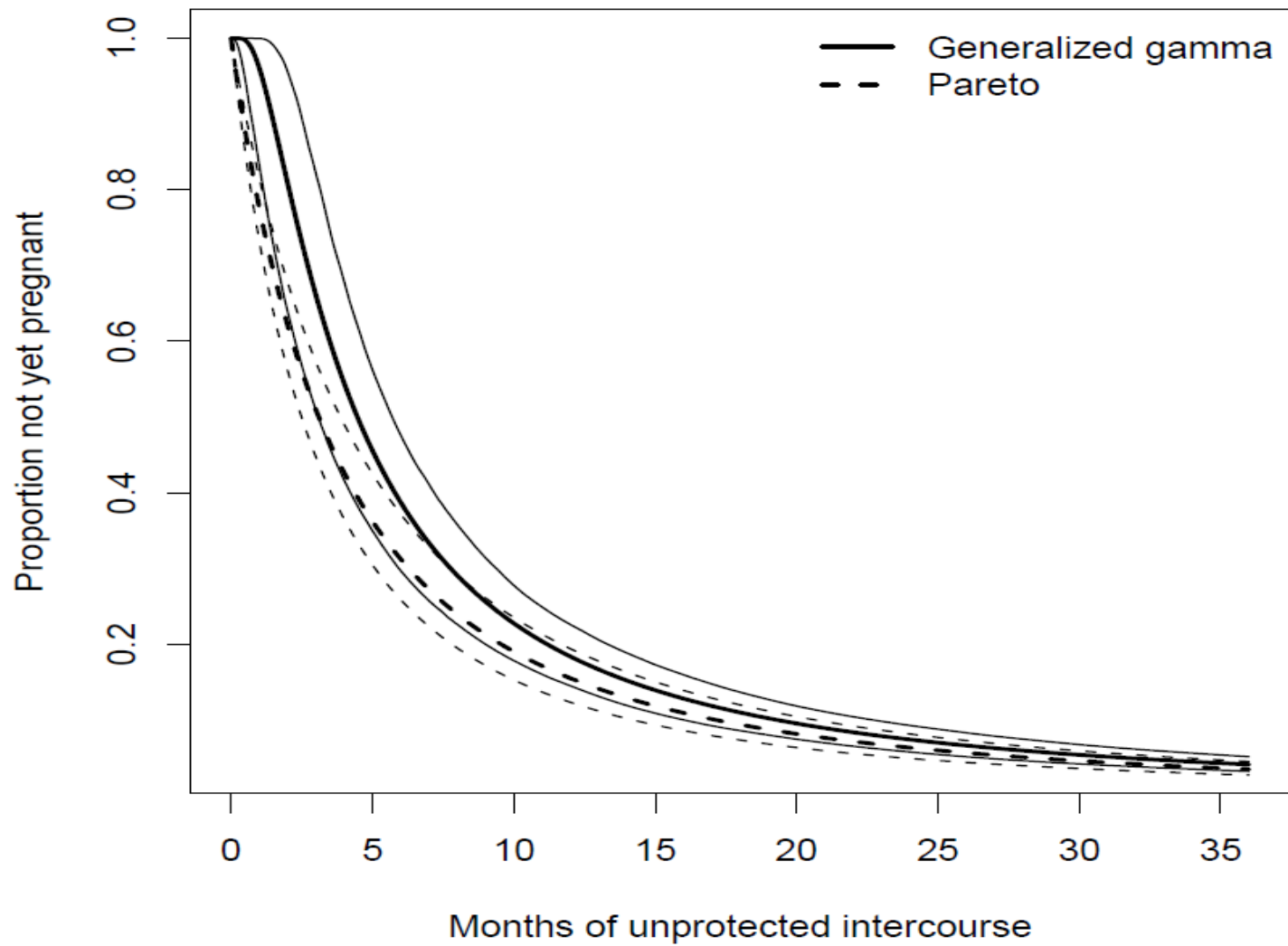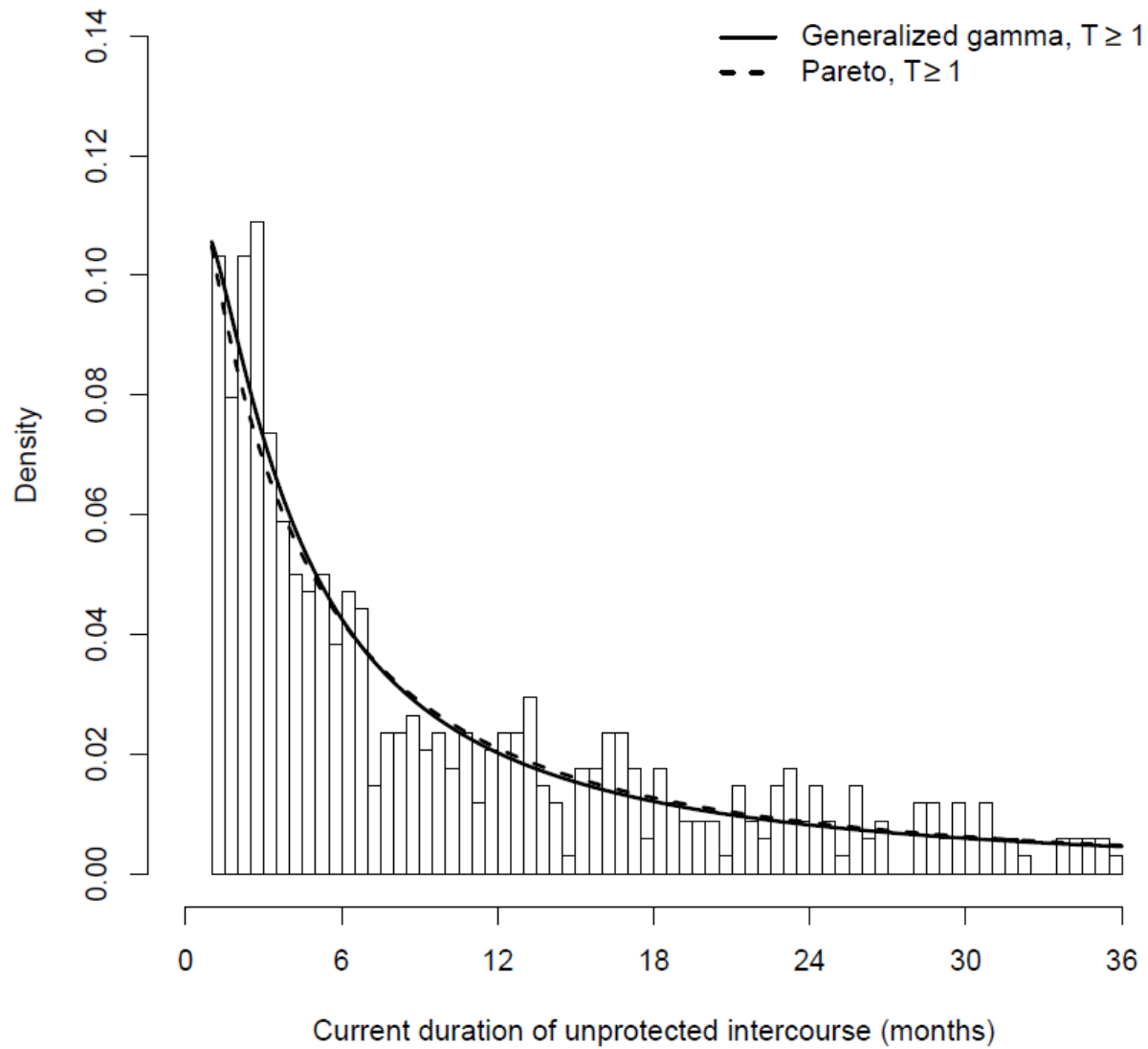
Estimated survival functions based on generalized gamma and Pareto distributions with pointwise 95% bootstrap confidence intervals.
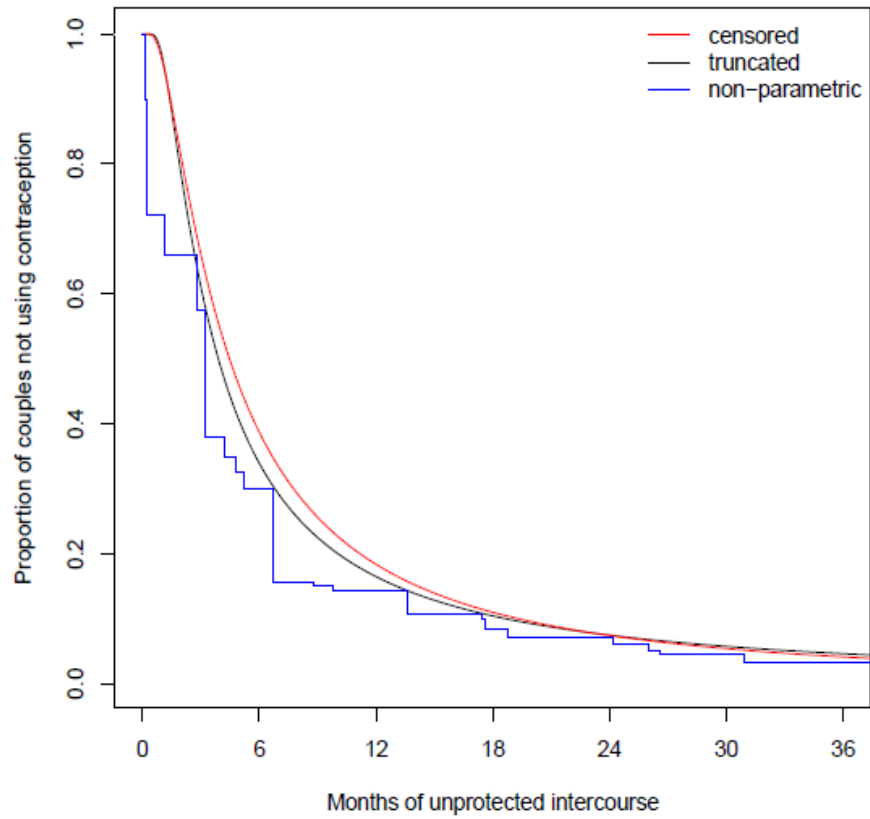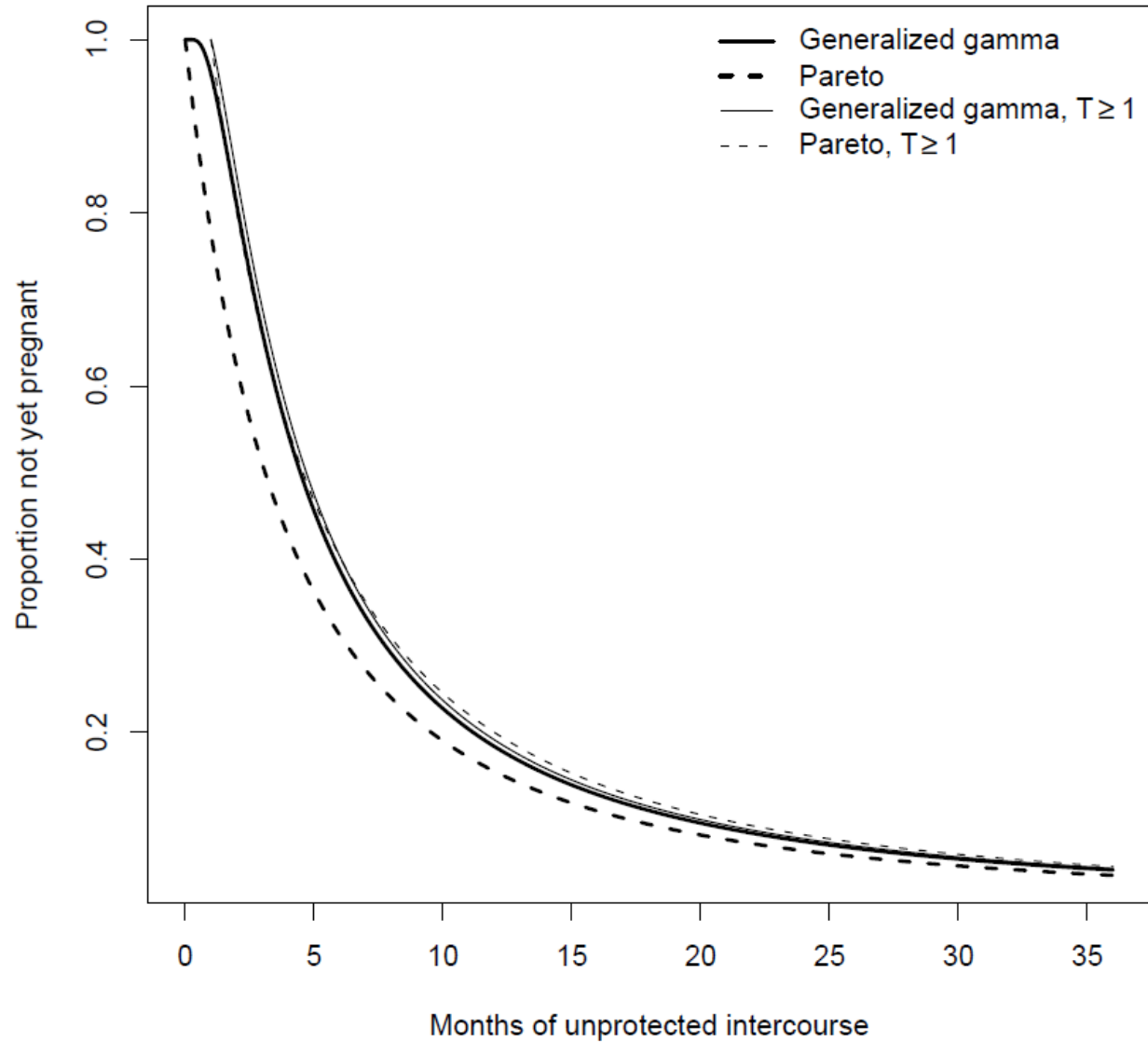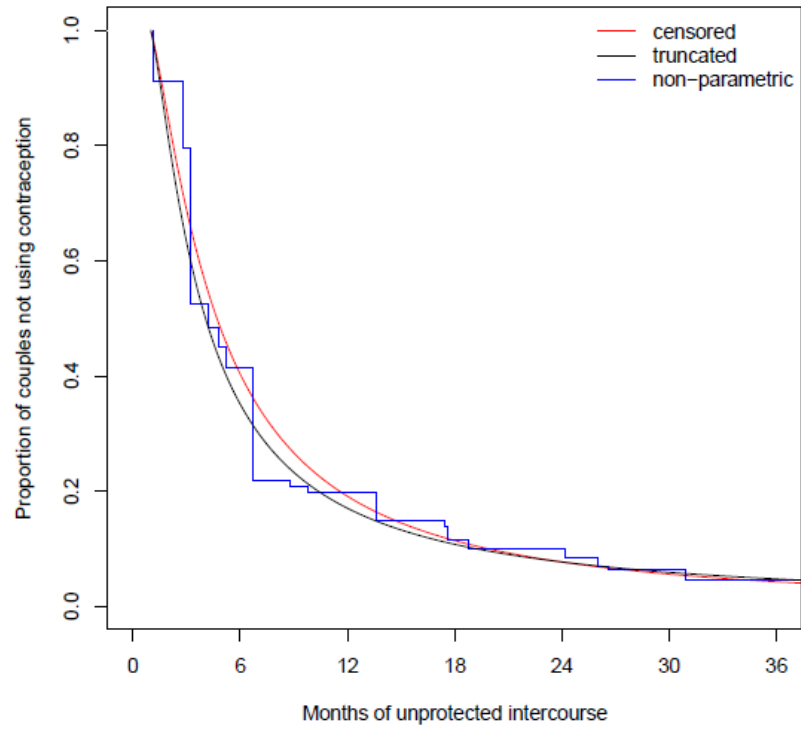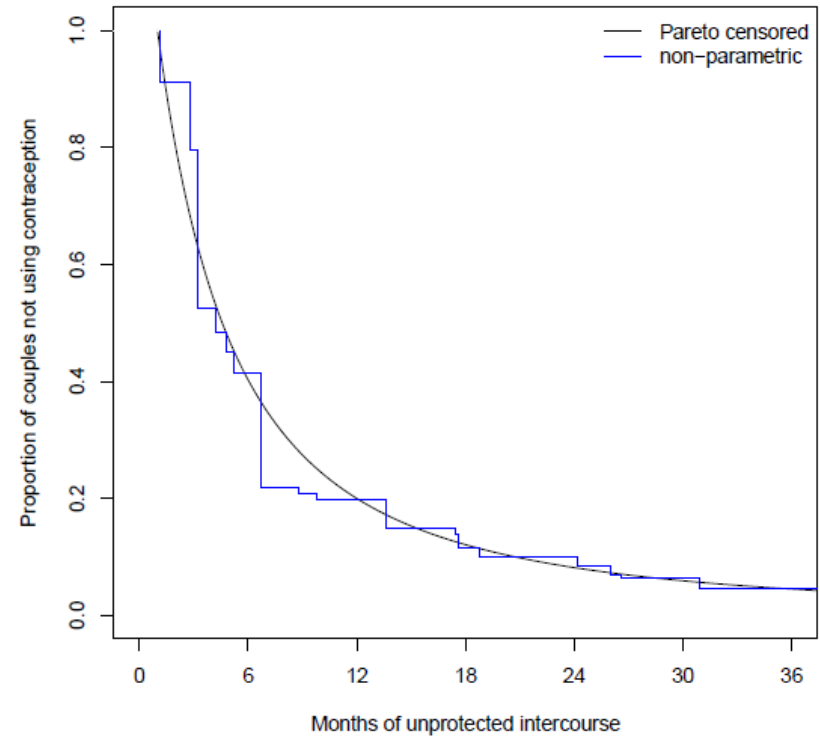
Generalized gamma

Pareto

Observed current durations and fitted densities given $T \geq 1$ month.

Comparison of fits based on all observations and based on $T \geq 1$ month.

Generalized gamma conditional on $T \geq 1$ Pareto conditional on $T \geq 1$

Months in waiting

Proportion still waiting at	TTP or giving up	TTP or giving up or entering fertility treatment
6 months	0.412	0.387
12 months	0.219	0.184
24 months	0.099	0.074
36 months	0.058	0.041
First quartile	2.189	2.372
Median	4.620	4.467
Third quartile	10.521	9.196
Exp. value	12.038	10.079

Based on fitted generalized gamma distribution censored at 36 months.

AFT fit: comparison of generalized gamma, Pareto and OLS

Accelerated failure time regression of current duration of unprotected intercourse (CDUI) on frequency of sexual intercourse using generalized gamma distribution, Pareto distribution and ordinary least square in log (CDUI)

Frequency of sexual intercourse	No.	Generalized gamma			Pareto			OLS		
		Time ratio	95% CI (time ratio)		Time ratio	95% CI (time ratio)		Time ratio	95% CI (time ratio)	
<1 per month	31	2.59	1.23	5.48	1.03	0.41	2.64	1.08	0.54	2.14
1-3 per month	143	1.57	1.07	2.31	1.79	1.20	2.78	1.72	1.17	2.53
1-2 per week	333	1.23	0.91	1.67	1.24	0.91	1.70	1.28	0.94	1.73
≥3 per week	221	1.00			1.00			1.00		
	728									

Other uses of current duration

McLauglin et al. (2010).

Childhood adversities and adult psychiatric disorders in the National Comorbidity Survey Replication II. *Arch. Gen. Psychiatry* **67**, 124-132.

Based on current duration of psychiatric disorders as registered at a cross-sectional survey.

Yamaguchi (2003).

Accelerated failure-time mover-stayer regression models for the analysis of last-episode data. *Sociological methodology* **33**, 81-110.

Migration patterns: based on current duration at the present location.

Ali et al. (2001).

Analysis of incomplete durations with application to contraceptive use. *JRSS A* **164**, 549-563.

Allison (1985).

Survival analysis of backward recurrence times. *JASA* **80**, 315-322.

Includes the 'open birth intervals' of demography.

Conclusion

Current duration approach has proved feasible.

Delicate estimation problems.

We look forward to the embedded prevalent cohort study for confirmation.