Brilleman, S., Crowther, M., May, M. T., Gompels, M., & Abrams, K. (2016). Joint longitudinal hurdle and time-to-event models: an application related to viral load and duration of the first treatment regimen in HIV patients initiating therapy. Statistics in Medicine, 35(20), 3583-3594. DOI: 10.1002/sim.6948

Supplementary material for "Joint longitudinal hurdle and time-to-event models: an application related to detectable viral load and treatment switching in HIV"

1. Model implementation in Stan

The models were estimated using Stan [1]. Stan provides a similar modelling language to WinBUGS [2] or JAGS [3], but uses a Hamiltonian Monte Carlo algorithm which is intended to provide faster convergence of MCMC chains and lower autocorrelation between subsequent MCMC samples than Gibbs sampling. Additionally Stan includes several preferred prior distributions, for example the Cauchy distribution (for regression coefficients) [4] or half-Cauchy distribution (for scale parameters) [5]. Stan also includes the "LKJ" correlation matrix distribution [6, 7] which provides a desirable alternative to use of the inverse-Wishart prior distribution for covariance matrices.

In general we used "weakly informative" prior distributions [5] intended to provide support to parameter values which are plausible. For the regression coefficients ($\boldsymbol{\delta_1}, \boldsymbol{\beta_2}, \boldsymbol{\gamma}, \alpha$) we used Cauchy priors with location 0 and scale 2, and for the regression coefficients for the spline terms in the Bernoulli part ($\boldsymbol{\beta_1}$) we used Cauchy priors with location 0 and scale 20. For the gamma rate parameter (ω) we used a half-Cauchy prior with location 0 and scale 2 whilst for the shape parameter for the Weibull baseline hazard we used a half-Cauchy prior with location 0 and scale 5. For the random effects distribution we avoided the need to specify a prior distribution for the unstructured variance-covariance matrix by instead using the "separation strategy" [8] to decompose $\boldsymbol{\Sigma}$ into a corresponding correlation matrix $\mathbf{R}$ and a vector of standard deviations. For the correlation matrix we assigned an "LKJ" correlation distribution [6] prior with shape parameter 1, which corresponds to a uniform prior over all possible correlation matrices [7], whilst for the standard deviations for each of the random effects we specified half-Cauchy priors with location 0 and scale 5.

We ran four MCMC chains in parallel with each chain consisting of a discarded "warm up" period of 2,500 iterations followed by a "sampling" period of 7,500 iterations which were used for inference.

The chains were initialised using randomly generated values, intended to be disparate across chains. Convergence was assessed using the potential scale reduction statistic [7, 9] as well as visually assessing trace plots of the MCMC samples. Figure S1 shows MCMC trace plots (four chains, each with 7,500 sample iterations following a discarded warm up period of 2,500 iterations) for the estimated subject specific means from the hurdle model, $\mu_i(t)$, for 8 selected patients calculated at their event or censoring time (i.e., $t = T_i$).

The Stan model code for fitting the joint longitudinal and time-to-event model is located in the attached plain text file entitled "stan_hurdlejm.stan". This is the Stan model file only. To fit the model to a specific dataset one of the available Stan interfaces must be used (http://mc-stan.org/). For the analysis presented in our manuscript we used the RStan 2.6.0 interface within R Version 3.1.3 [10]. Therefore, all pre-processing of the data and post-processing of the MCMC samples was done using R Version 3.1.3.

2. Observed marker data and estimated trajectories

Figure S2 shows the observed outcome data and the estimated marker trajectories for 8 selected patients in the dataset. The estimates are based on the expected value for the hurdle model (both the Bernoulli and Gamma parts combined), given the estimated subject-specific random effects.

3. Results from joint models using alternative association structures

Here we present the fixed effect parameter estimates from three joint models which used alternative association structures to the main joint model presented in the paper. These alternative joint models were described in Section 4.6 of the main manuscript. Table S1 contains the parameter estimates from the joint model with a null association structure (equivalent to fitting the longitudinal hurdle model and the time-to-event model separately). Table S2 contains the parameter estimates from the joint model with an association structure based on the linear predictor from the Bernoulli part only

(equivalent to a joint model based only on binary marker data). Table S3 contains the parameter estimates from the joint model with a quadratic association structure based on the expected outcome from the hurdle model.

4. Results from sensitivity analyses

The main purpose of the sensitivity analyses were to investigate alternative choices for the number and location of knot points for the (fixed) restricted cubic splines in the Bernoulli part of the hurdle model. The fixed spline terms and the random linear slope term which were used in the Bernoulli part of the hurdle model provided the main flexibility for modelling changes in the marker outcome over time (recall that the gamma part of the hurdle model did not allow for any changes over time due to the limitations of the dataset) and therefore we were interested in how these assumptions influenced the estimated association parameter ($\propto$). We considered the following sets of knot locations for the fixed splines: $\{0, 3, 6, 60\}, \{0, 12, 36, 60\}$ and $\{0, 12, 60\}$ months and also considered a model which only included a linear slope in the Bernoulli part. We also calculated an estimate of the Watanabe-Akaike information criterion (WAIC) for each of the models. Smaller WAIC provides an indication of a better fitting model.

Table S4 provides the fixed effect parameter estimates (and 95% credible intervals) and WAIC for each of the models which were estimated as part of the sensitivity analyses. Models with a knot point during the first 12 months (main model and model A) appeared to fit the data better and resulted in a smaller estimated association parameter (closer to zero). Models without a knot point during the first 12 months (models B, C and D) fit the data worse based on WAIC and had a larger estimated association parameter. This may be because the models without a knot point in the first 12 months had less flexibility to capture changes around the time most patients achieved viral suppression (approximately 6 months after initiating ART), thereby providing a less adequate fit to the longitudinal data and potentially leading to biased estimates of the association parameter. On the other hand, the

estimated association parameter was similar across models which did allow for a knot point during the first 12 months.

References

1. Stan: A C++ Library for Probability and Sampling, Version 2.6.0, 2014.

2. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput* 2000; **10**: 325-337.

3. Plummer M. JAGS Version 3.4.0 User Manual. 2013.

4. Gelman A, Jakulin A, Pittau MG, Su YS. A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *Annals of Applied Statistics* 2008; **2**: 1360-1383.

5. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; **1**: 515-533.

6. Lewandowski D, Kurowicka D, Joe H. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* 2009; **100**: 1989-2001.

7. Stan Development Team. Stan Modelling Language User's Guide and Reference Manual, Version 2.6.2. 2015.

8. Barnard J, McCulloch R, Meng XL. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* 2000; **10**: 1281-1311.

9. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**: 457-472.

10. R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria, 2015.

Figure S1. MCMC trace plots (four chains, each with 7,500 sample iterations following a discarded warm up period of 2,500 iterations) for the estimated subject specific means from the hurdle model, $\mu_i(t)$, for 8 selected patients calculated at their event or censoring time (i.e., $t = T_i$).

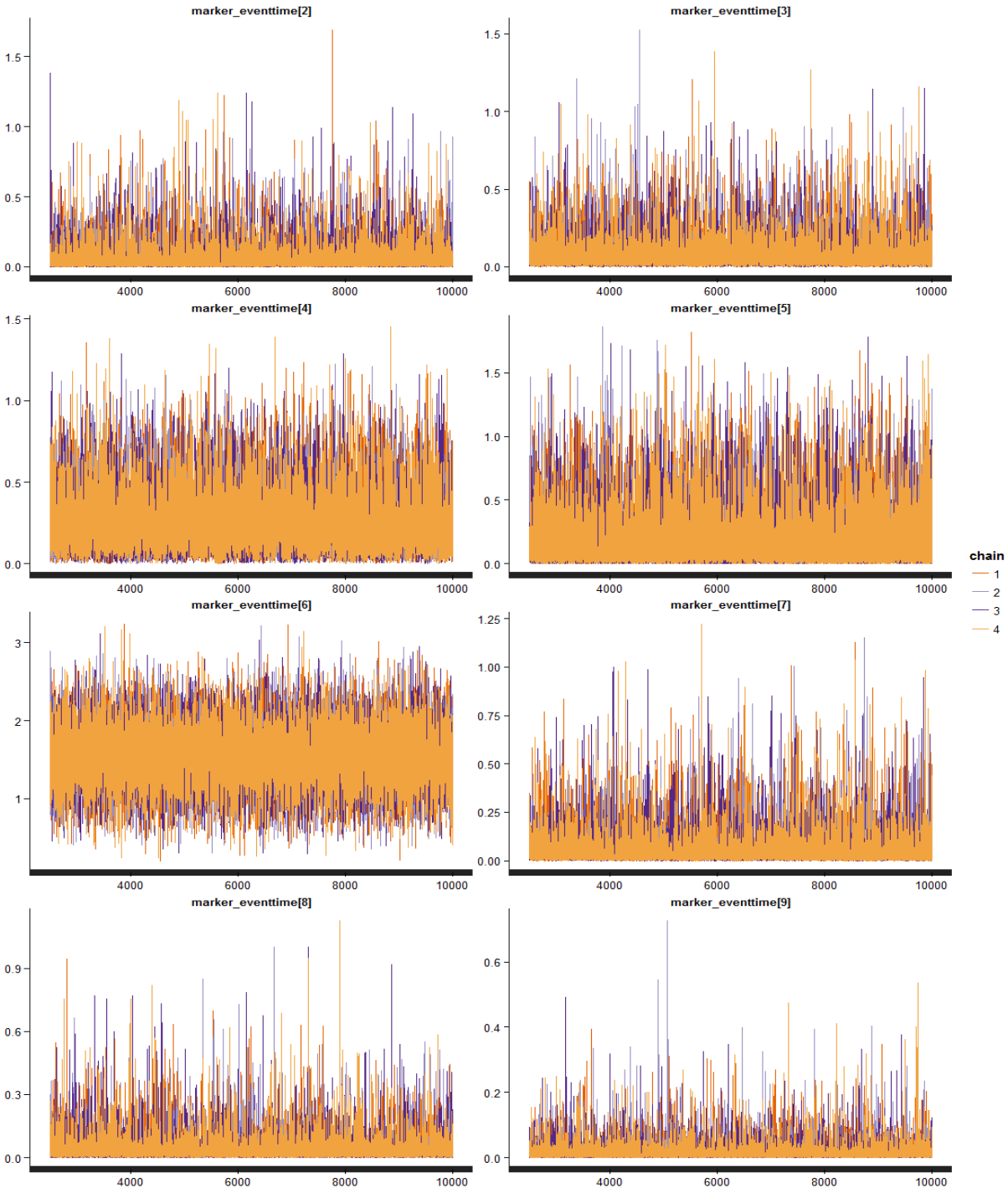

Figure S2. Observed marker data and estimated trajectories under the model for 8 selected patients. The vertical axis shows $\log_{10}$ viral load "excess" which is the amount by which a patient's $\log_{10}$ viral load exceeds the detection limit of 1.7 $\log_{10}$ copies/ml: this is the expected value for the hurdle model (both the Bernoulli and gamma parts combined), given the estimated subject-specific random effects $\boldsymbol{b_i}$. The horizontal axis shows time (in years) since the patient began their first antiretroviral treatment regimen.

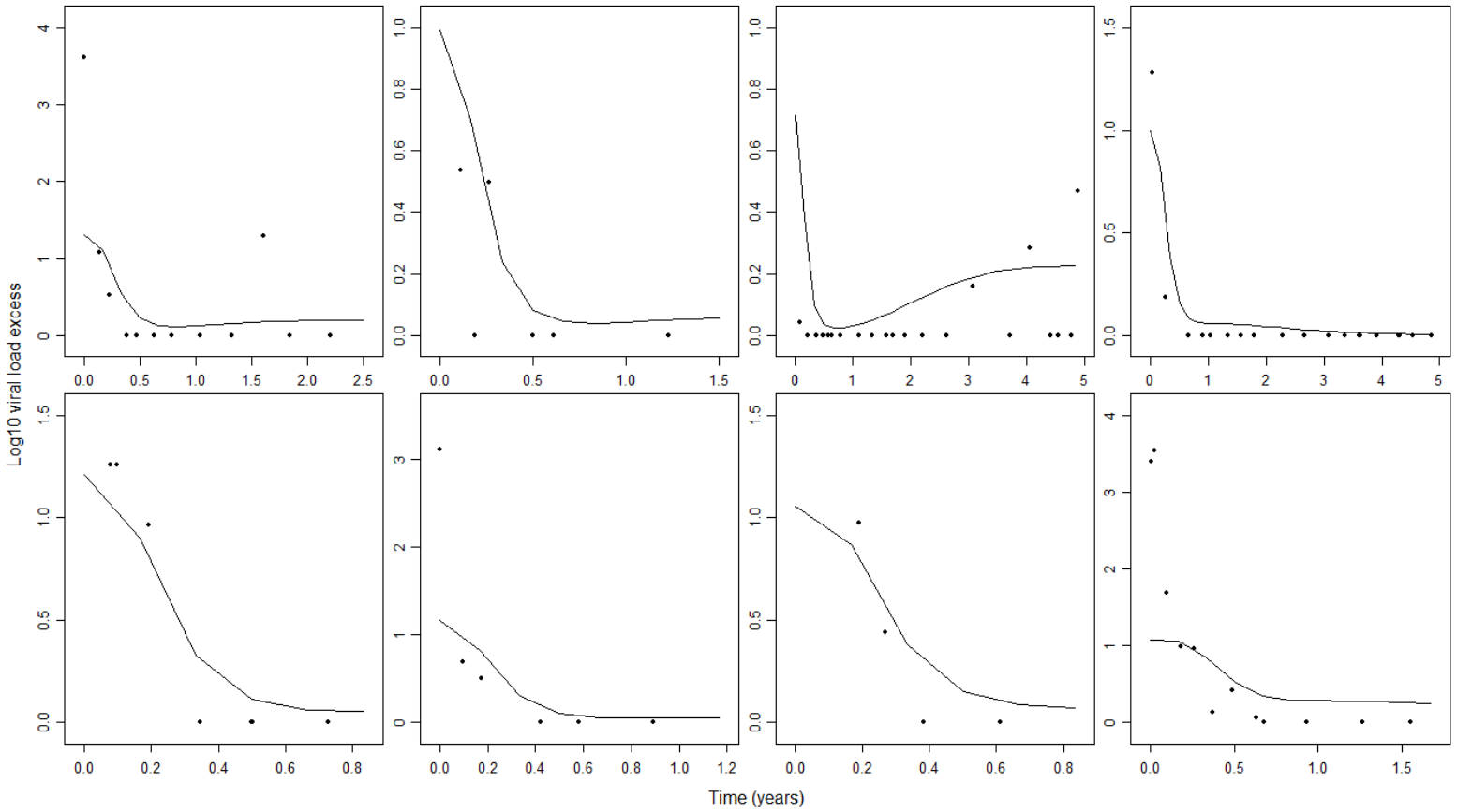

Table S1. Fixed effects parameter estimates from the joint longitudinal hurdle and time-to-event model with a null association structure (equivalent to fitting the longitudinal hurdle model and the time-to-event model separately).

	Estimated coefficient (95% CrI)	Odds ratio or hazard ratio (95% CrI)
Marker model: Bernoulli part		
Constant	4.64 (3.83 to 5.50)	
Gender (ref: male)		
Female	-1.10 (-1.61 to -0.61)	0.33 (0.20 to 0.54)
Age (ref: 18-29 years)		
30-39 years	0.13 (-0.47 to 0.74)	1.14 (0.63 to 2.10)
40-49 years	0.05 (-0.63 to 0.72)	1.05 (0.53 to 2.06)
50+ years	0.17 (-0.63 to 0.97)	1.18 (0.53 to 2.63)
Start period (ref: 2010-12)		
2000-03	0.06 (-0.63 to 0.75)	1.06 (0.53 to 2.12)
2004-06	-0.78 (-1.46 to -0.13)	0.46 (0.23 to 0.88)
2007-09	-0.61 (-1.22 to -0.02)	0.54 (0.29 to 0.98)
RCS terms		
RCS1	-19.90 (-22.07 to -17.87)	
RCS2	-37.21 (-42.10 to -32.60)	
RCS3	9.51 (8.28 to 10.81)	
Marker model: gamma part		
Constant	0.01 (-0.06 to 0.08)	
Gamma rate parameter	1.17 (1.05 to 1.30)	
Time-to-event model		
Constant	-0.69 (-1.02 to -0.36)	
Gender (ref: male)		
Female	0.28 (0.07 to 0.49)	1.32 (1.07 to 1.64)
Age (ref: 18-29 years)		
30-39 years	-0.32 (-0.58 to -0.05)	0.73 (0.56 to 0.95)
40-49 years	-0.32 (-0.61 to -0.01)	0.73 (0.54 to 0.99)
50+ years	-0.02 (-0.39 to 0.35)	0.98 (0.68 to 1.42)
Start period (ref: 2010-12)		
2000-03	-1.08 (-1.45 to -0.71)	0.34 (0.23 to 0.49)
2004-06	-0.33 (-0.65 to -0.01)	0.72 (0.52 to 0.99)
2007-09	0.04 (-0.24 to 0.34)	1.04 (0.78 to 1.40)
Association parameter	Na	
Weibull shape parameter	1.05 (0.97 to 1.14)	

Abbreviations. CrI: Bayesian credible interval. ref: reference category. RCS: restricted cubic spline. Na: not applicable.

Table S2. Fixed effects parameter estimates from the joint longitudinal hurdle and time-to-event model with an association structure based on the linear predictor from the Bernoulli part only (equivalent to a joint model based only on binary marker data).

	Estimated coefficient (95% CrI)	Odds ratio or hazard ratio (95% CrI)
Marker model: Bernoulli part		
Constant	4.49 (3.67 to 5.35)	
Gender (ref: male)		
Female	-1.08 (-1.58 to -0.58)	0.34 (0.21 to 0.56)
Age (ref: 18-29 years)		
30-39 years	0.14 (-0.46 to 0.73)	1.15 (0.63 to 2.09)
40-49 years	0.08 (-0.60 to 0.77)	1.08 (0.55 to 2.15)
50+ years	0.21 (-0.59 to 1.00)	1.23 (0.56 to 2.73)
Start period (ref: 2010-12)		
2000-03	0.01 (-0.68 to 0.71)	1.01 (0.51 to 2.03)
2004-06	-0.76 (-1.42 to -0.13)	0.47 (0.24 to 0.88)
2007-09	-0.64 (-1.23 to -0.04)	0.53 (0.29 to 0.96)
RCS terms		
RCS1	-19.18 (-21.47 to -17.06)	
RCS2	-35.49 (-40.64 to -30.65)	
RCS3	9.06 (7.77 to 10.43)	
Marker model: gamma part		
Constant	0.01 (-0.06 to 0.08)	
Gamma rate parameter	1.17 (1.05 to 1.29)	
Time-to-event model		
Constant	-0.61 (-0.96 to -0.27)	
Gender (ref: male)		
Female	0.36 (0.13 to 0.58)	1.43 (1.13 to 1.79)
Age (ref: 18-29 years)		
30-39 years	-0.33 (-0.60 to -0.06)	0.72 (0.55 to 0.94)
40-49 years	-0.33 (-0.62 to -0.03)	0.72 (0.54 to 0.97)
50+ years	-0.02 (-0.39 to 0.34)	0.98 (0.67 to 1.41)
Start period (ref: 2010-12)		
2000-03	-1.13 (-1.51 to -0.75)	0.32 (0.22 to 0.47)
2004-06	-0.30 (-0.62 to 0.03)	0.74 (0.54 to 1.03)
2007-09	0.07 (-0.22 to 0.37)	1.07 (0.80 to 1.45)
Association parameter	0.06 (0.00 to 0.13)	1.07 (1.00 to 1.14)
Weibull shape parameter	1.18 (1.02 to 1.34)	

Abbreviations. CrI: Bayesian credible interval. ref: reference category. RCS: restricted cubic spline.

Table S3. Fixed effects parameter estimates from the joint longitudinal hurdle and time-to-event model with a quadratic association structure based on the expected outcome from the hurdle model.

	Estimated coefficient (95% CrI)	Odds ratio or hazard ratio (95% CrI)
Marker model: Bernoulli part		
Constant	4.56 (3.74 to 5.41)	
Gender (ref: male)		
Female	-1.08 (-1.59 to -0.58)	0.34 (0.20 to 0.56)
Age (ref: 18-29 years)		
30-39 years	0.13 (-0.46 to 0.74)	1.14 (0.63 to 2.09)
40-49 years	0.07 (-0.60 to 0.75)	1.08 (0.55 to 2.13)
50+ years	0.19 (-0.61 to 0.99)	1.21 (0.54 to 2.70)
Start period (ref: 2010-12)		
2000-03	0.02 (-0.68 to 0.71)	1.02 (0.51 to 2.03)
2004-06	-0.77 (-1.44 to -0.12)	0.46 (0.24 to 0.89)
2007-09	-0.62 (-1.22 to -0.03)	0.54 (0.30 to 0.97)
RCS terms		
RCS1	-19.47 (-21.69 to -17.39)	
RCS2	-36.28 (-41.27 to -31.59)	
RCS3	9.27 (8.01 to 10.59)	
Marker model: gamma part		
Constant	0.00 (-0.07 to 0.07)	
Gamma rate parameter	1.17 (1.05 to 1.30)	
Time-to-event model		
Constant	-0.92 (-1.32 to -0.54)	
Gender (ref: male)		
Female	0.31 (0.10 to 0.53)	1.37 (1.10 to 1.69)
Age (ref: 18-29 years)		
30-39 years	-0.32 (-0.58 to -0.05)	0.73 (0.56 to 0.95)
40-49 years	-0.31 (-0.61 to -0.01)	0.73 (0.55 to 0.99)
50+ years	0.00 (-0.37 to 0.36)	1.00 (0.69 to 1.44)
Start period (ref: 2010-12)		
2000-03	-1.13 (-1.51 to -0.75)	0.32 (0.22 to 0.47)
2004-06	-0.31 (-0.63 to 0.01)	0.73 (0.53 to 1.01)
2007-09	0.05 (-0.24 to 0.34)	1.05 (0.78 to 1.41)
Association parameters		
Linear term (α_1)	0.81 (-0.13 to 1.81)	2.24 (0.88 to 6.10)
Quadratic term (α_2)	-0.23 (-0.91 to 0.28)	0.79 (0.40 to 1.32)
Weibull shape parameter	1.16 (1.03 to 1.29)	

Abbreviations. CrI: Bayesian credible interval. ref: reference category. RCS: restricted cubic spline.

Table S4. Results from sensitivity analyses investigating alternative choices for the number and location of knot points for the (fixed) restricted cubic splines in the Bernoulli part of the hurdle model. Shown in the table are fixed effect parameter estimates (and 95% credible intervals) for the various models as well as the Watanabe-Akaike information criterion (WAIC), for which a smaller WAIC indicates a better fitting model. The 'main model' is the model presented in the manuscript and is based on fixed splines with 3 degrees of freedom (d.f.) and knots placed at {0, 3, 12, 60} months. Models A and B are also based on fixed splines with 3 d.f. but knots placed at {0, 3, 6, 60} and {0, 12, 36, 60} months respectively. Model C is based on fixed splines with 2 d.f. and knots placed at {0, 12, 60} months. Model D is based on a model with a fixed linear slope term (no use of restricted cubic splines). All models include a random intercept and random linear slope term in the Bernoulli part.

	Main model	Model A	Model B	Model C	Model D
WAIC	6520.2	6483.5	6684.9	6685.5	6683.3
Marker model: Bernoulli part					
Constant	4.58 (3.77 to 5.44)	5.72 (4.74 to 6.69)	2.58 (2.00 to 3.20)	2.55 (1.97 to 3.17)	2.67 (2.04 to 3.33)
Gender (ref: male)					
Female	-1.09 (-1.60 to -0.59)	-1.13 (-1.68 to -0.61)	-0.68 (-1.06 to -0.32)	-0.64 (-1.02 to -0.27)	-0.66 (-1.06 to -0.28)
Age (ref: 18-29 years)					
30-39 years	0.15 (-0.46 to 0.76)	0.17 (-0.47 to 0.82)	0.11 (-0.35 to 0.58)	0.12 (-0.35 to 0.58)	0.14 (-0.35 to 0.63)
40-49 years	0.08 (-0.60 to 0.77)	0.10 (-0.59 to 0.82)	0.16 (-0.36 to 0.68)	0.13 (-0.39 to 0.64)	0.13 (-0.41 to 0.68)
50+ years	0.20 (-0.62 to 1.02)	0.16 (-0.68 to 1.02)	0.35 (-0.26 to 0.98)	0.40 (-0.22 to 1.03)	0.45 (-0.19 to 1.10)
Start period (ref: 2010-12)					
2000-03	0.02 (-0.67 to 0.72)	0.13 (-0.61 to 0.84)	-0.38 (-0.93 to 0.16)	-0.54 (-1.09 to 0.00)	-0.81 (-1.38 to -0.24)
2004-06	-0.78 (-1.45 to -0.13)	-0.75 (-1.43 to -0.08)	-0.56 (-1.06 to -0.06)	-0.62 (-1.13 to -0.13)	-0.72 (-1.24 to -0.20)
2007-09	-0.63 (-1.24 to -0.03)	-0.63 (-1.26 to 0.00)	-0.57 (-1.04 to -0.12)	-0.61 (-1.07 to -0.15)	-0.63 (-1.11 to -0.16)
Linear slope	Na	Na	Na	Na	-9.46 (-11.10 to -7.93)
RCS terms (omitted as not comparable between models)					

Marker model: gamma part					
Constant	0.00 (-0.07 to 0.07)	0.00 (-0.07 to 0.08)	0.00 (-0.07 to 0.07)	0.00 (-0.07 to 0.07)	0.01 (-0.06 to 0.08)
Gamma shape parameter	1.18 (1.06 to 1.30)	1.17 (1.05 to 1.30)	1.17 (1.05 to 1.29)	1.17 (1.05 to 1.30)	1.18 (1.06 to 1.30)
Time-to-event model					
Constant	-0.87 (-1.25 to -0.52)	-0.88 (-1.25 to -0.53)	-1.03 (-1.41 to -0.66)	-1.00 (-1.38 to -0.63)	-0.93 (-1.31 to -0.57)
Gender (ref: male)					
Female	0.31 (0.09 to 0.52)	0.31 (0.09 to 0.52)	0.33 (0.11 to 0.54)	0.32 (0.10 to 0.53)	0.31 (0.10 to 0.53)
Age (ref: 18-29 years)					
30-39 years	-0.30 (-0.57 to -0.04)	-0.30 (-0.57 to -0.03)	-0.29 (-0.55 to -0.02)	-0.29 (-0.55 to -0.02)	-0.30 (-0.56 to -0.03)
40-49 years	-0.29 (-0.59 to 0.00)	-0.29 (-0.60 to 0.01)	-0.28 (-0.58 to 0.02)	-0.28 (-0.58 to 0.02)	-0.29 (-0.58 to 0.01)
50+ years	0.02 (-0.35 to 0.38)	0.02 (-0.36 to 0.39)	0.05 (-0.33 to 0.41)	0.04 (-0.33 to 0.41)	0.03 (-0.35 to 0.40)
Start period (ref: 2010-12)					
2000-03	-1.13 (-1.51 to -0.75)	-1.13 (-1.51 to -0.76)	-1.19 (-1.57 to -0.81)	-1.17 (-1.56 to -0.80)	-1.15 (-1.53 to -0.78)
2004-06	-0.32 (-0.64 to 0.00)	-0.32 (-0.63 to 0.00)	-0.31 (-0.63 to 0.01)	-0.32 (-0.64 to 0.01)	-0.32 (-0.64 to 0.00)
2007-09	0.04 (-0.26 to 0.34)	0.04 (-0.25 to 0.32)	0.03 (-0.26 to 0.32)	0.03 (-0.26 to 0.33)	0.03 (-0.26 to 0.32)
Association parameter	0.44 (0.06 to 0.78)	0.45 (0.07 to 0.79)	0.77 (0.43 to 1.11)	0.70 (0.35 to 1.04)	0.56 (0.20 to 0.91)
Weibull shape parameter	1.14 (1.02 to 1.25)	1.14 (1.02 to 1.25)	1.21 (1.09 to 1.33)	1.20 (1.08 to 1.31)	1.17 (1.05 to 1.28)

Abbreviations. RCS: restricted cubic splines. Na: not applicable