Supplementary Material for "Using smoothing splines to resolve the curvature identifiability problem in age-period-cohort models with unequal intervals"

Connor Gascoigne (c.gascoigne@bath.ac.uk)[1] and Theresa Smith[1]

[1]Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK

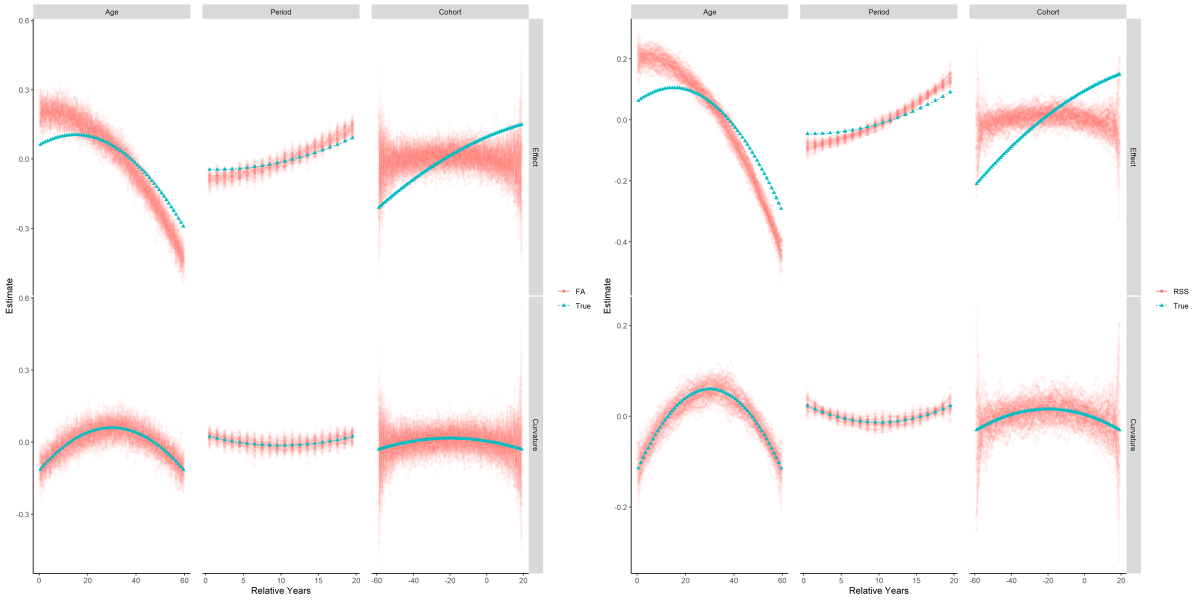
S1 Additional material for equal interval simulation

Here we include further results to supplement the binomial results for equal intervals displayed in the main paper. A more in depth look at the individual simulations for the binomial distribution is presented. Furthermore, results from the simulations for data generated under the Gaussian and Poisson distributions are included.

To show the structural link identification problem lies in the data, rather than in the choice of model fit, an additional simulation for the binomial distribution is included. The additional simulation study is for the full re-parameterised APC models fit to data generated where only two of the three temporal effects are influential. The re-parameterised APC model fit will have the cohort linear trend dropped and the temporal terms that influence the data are age and period.

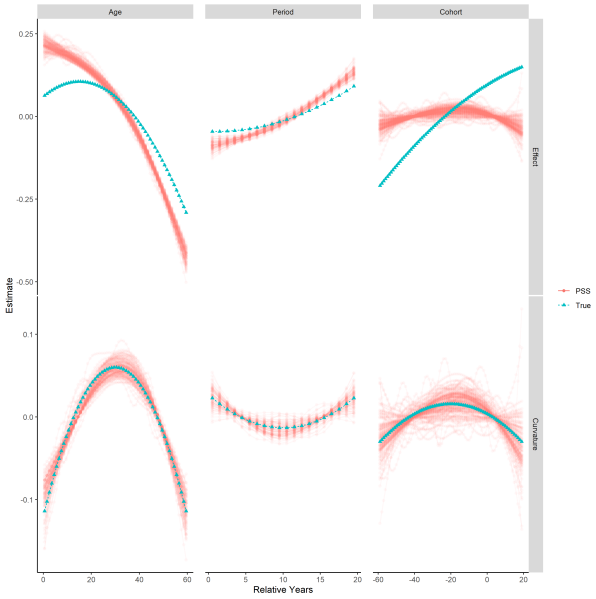
S1.1 Individual simulations plot for the binomial distribution

Figure S1 presents the estimated functions from each simulation for the FA, RSS and PSS models for binomial data generated with all three effects present. In both the estimated effects and curvature plots, the FA and RSS models have a greater variability than the PSS model across all three temporal trends. The variability is much larger in the FA and RSS models for the youngest and oldest cohorts than for the PSS model. The earlier and later cohorts are seen the least throughout the data, hence suffer from greater variability; larger variability will incur a larger penalty in the PSS model.



(a) Factor model.

(b) Regression smoothing spline model.



(c) Penalised smoothing spline model.

Figure S1: Individual simulation plots for equal interval binomial data

S1.2 Binomial simulation study for data generated with only two temporal effects present

Figure S2 shows the simulation study results for equal interval, binomial data generated with only age and period effects present. Full APC models are fit to the data. In each model, cohort is the linear slope dropped in the re-parameterisation.

The *ad-hoc* choice of what linear trend to drop is forcing that trend to be zero. When all three effects are present in the data generation, this is rarely the right choice as the true effect of the dropped trend is often not zero. In this simulation, we know cohort does not influence the data generated; therefore, the cohort linear trend is zero and the *ad-hoc* choice is correct. Th estimated effects correctly estimate the cohort linear trend and are shown to be the same as the true effects.

For the identifiable curvatures, the results for the FA and RSS models for the cohort have a relatively large bias and MSE in comparison to the PSS model. Each of the models is estimating a cohort curvature that is not present in the data; therefore, it is over-fitting the cohort curvature. The penalty term in the PSS model penalises the over-fitting hence why the cohort curvature bias and MSE is smaller in the PSS model than in the other two.

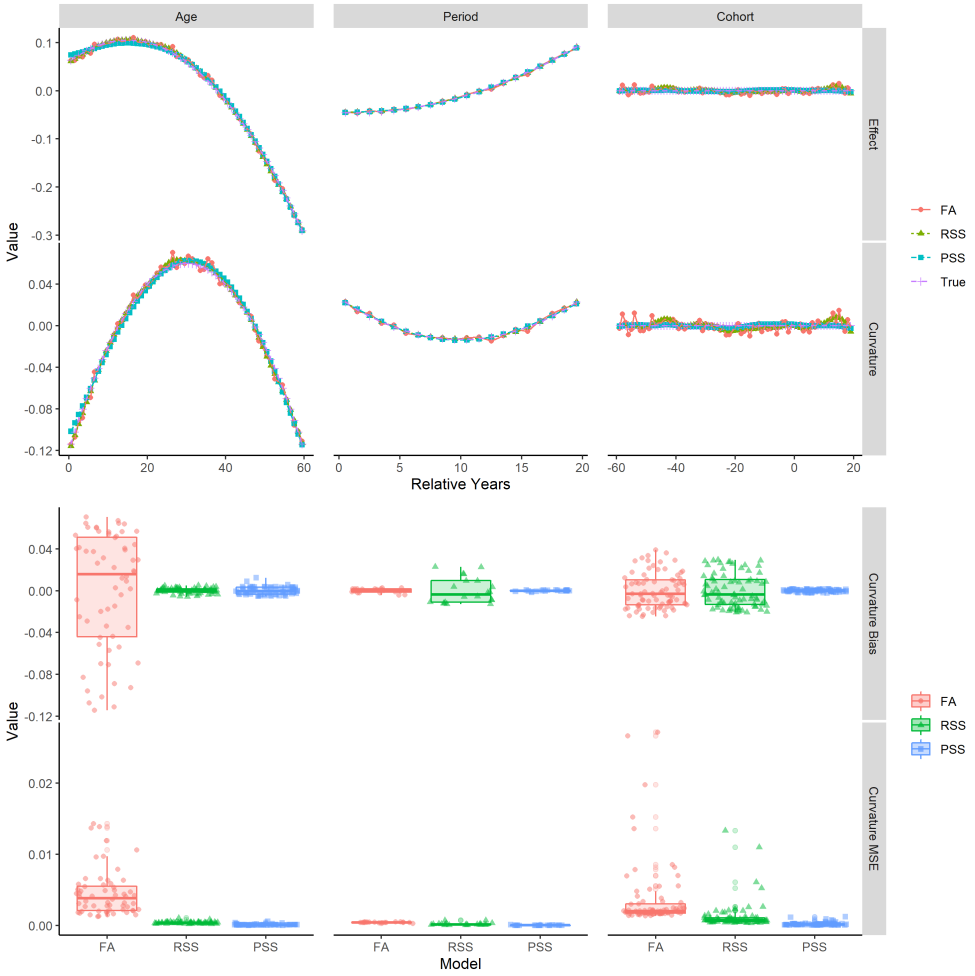


Figure S2: Simulation study results for equal interval binomial data generated when only age and period effects are present.

S1.3 Gaussian simulation studies

Figure S3 shows the simulation study for equal interval, Gaussian data where all three effects are present. The results for Figure S3 are similar to the results seen in the main body for the binomial distribution. The structural link identification issue is displayed by the model estimates for the temporal effects being different to the true effects. Furthermore, the curvatures are identifiable and this is reflected in the model estimates of them matching the truth. The bias and MSE box-plots show the PSS model performs in line with the current literature.

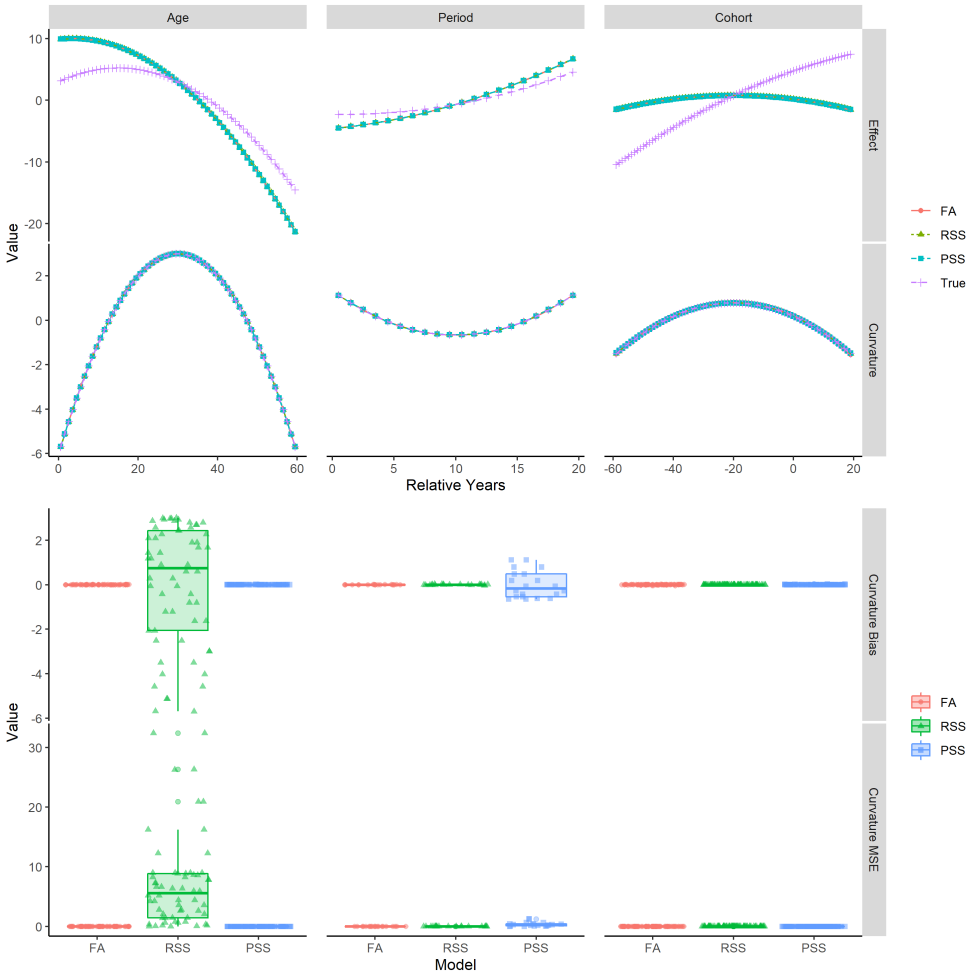


Figure S3: Simulation study results for equal interval, $M = 1$, Gaussian data generated when all three temporal effects are present.

S1.4 Poisson simulation studies

Figure S4 shows the simulation study for equal interval, Poisson data where all three effects are present. The interpretation of the results from the Poisson simulation study is the same as those from the binomial and Gaussian distributions.

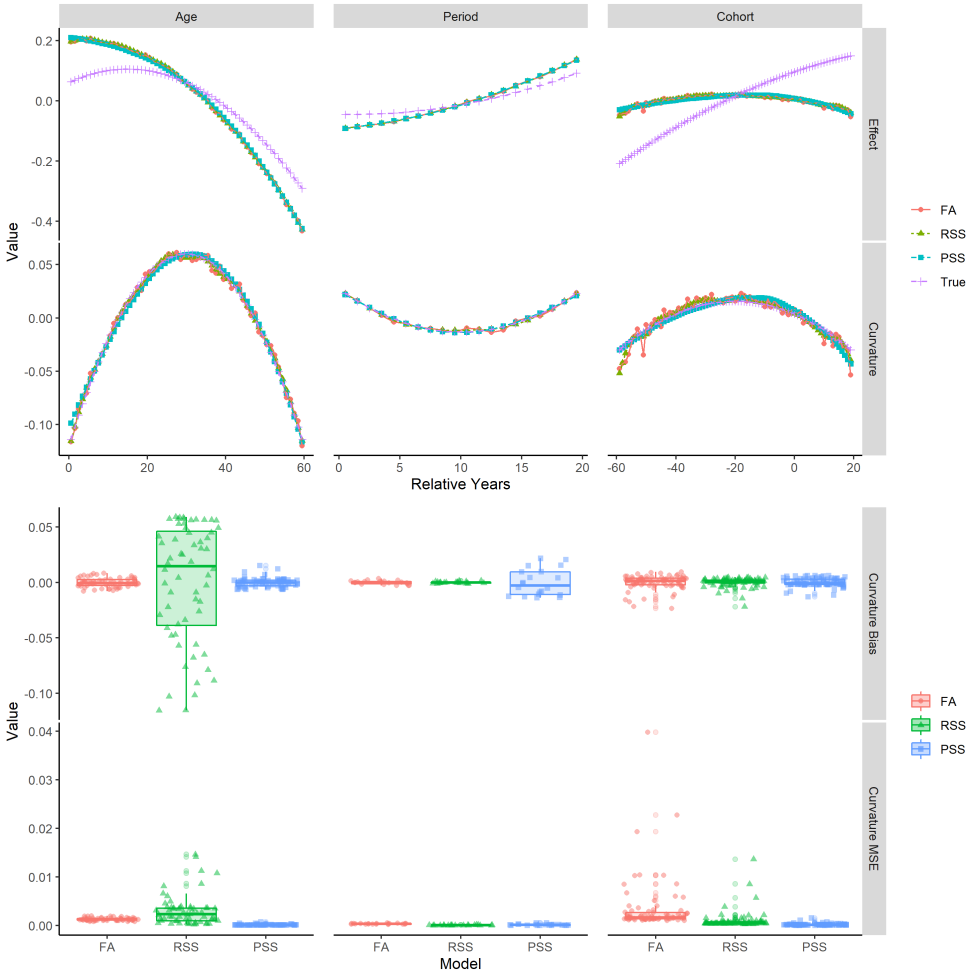


Figure S4: Simulation study results for equal interval, $M = 1$, Poisson data generated when all three temporal effects are present.

S2 Additional material for unequal interval simulation

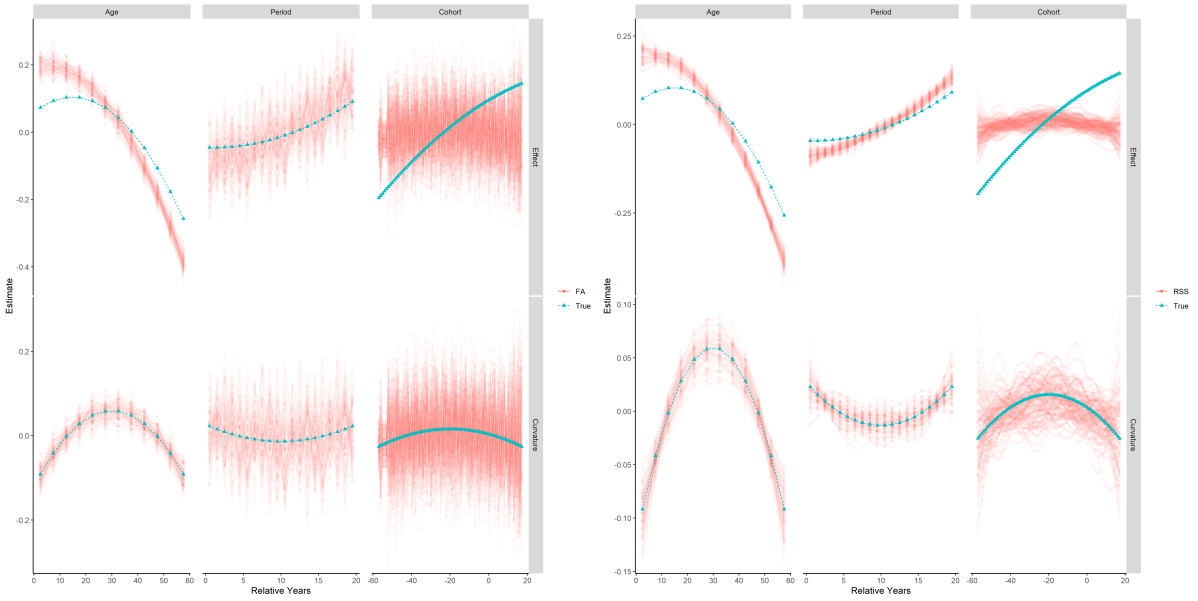
Here we include further results to supplement the binomial results for unequal intervals displayed in the main paper. As with the additional results for the equal intervals simulation we have: individual simulations for the binomial distribution; an additional simulation study for the binomial distribution when the data is generated with only two temporal effects; and results from the the Gaussian and Poisson distributions for data generated with all three temporal effects.

S2.1 Individual simulations plot for the binomial distribution

Figure S5 presents the estimated functions from each simulation for the FA, RSS and PSS models for unequal interval, binomial data generated with all three effects present.

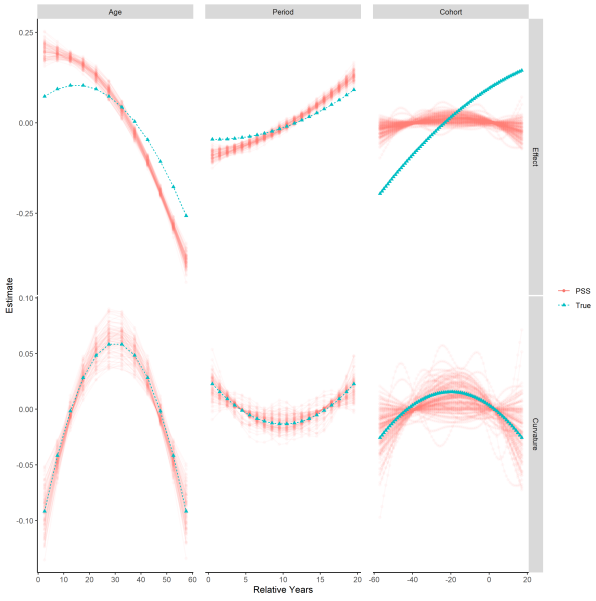
Figure S5a shows how the FA model is unsuitable for data that comes in unequal intervals. The large variability throughout each of the estimated curvature functions highlights how the previously identifiable terms are no longer identifiable. Furthermore, the age curvatures not suffering the same issues as the period and cohort curvatures shows how the additional identification from fitting a model to unequally aggregated data only affects the period and cohort terms.

The RSS model has a larger variation for its simulations than the PSS model as well as suffering from greater variability in the tails of each function, most notably for cohort. This shows how the penalty term being applied in the PSS model is working to reduce the additional identification in the curvature estimates for each individual simulation.



(a) Factor model.

(b) Regression smoothing spline model.



(c) Penalised smoothing spline model.

Figure S5: Individual simulation plots for equal interval binomial data

S2.2 Binomial simulation study for data generated with only two temporal effects present

Figure S6 presents the results from the simulation study where only age and period are influence the data generation. As with the equal intervals, the structural link identification problem is no longer present in the data hence why the estimated effects appear to be identifiable at first look.

Looking closer, the FA model estimates are exhibiting the periodic pattern in the period and cohort estimated effects due to the additional identification issues.[1] The periodic pattern in the FA model becomes more pronounced in the curvature estimates. A smaller bias for the PSS period and cohort curvature box-plots in comparison to the RSS period and cohort curvature bias boxplots shows the PSS model is performing better than the RSS model. The additional identification problems in the estimated curvature functions are being alleviated by the PSS model; the penalty is resolving the additional identification issues in the estimates whereas the use of smoothing functions alone is not.

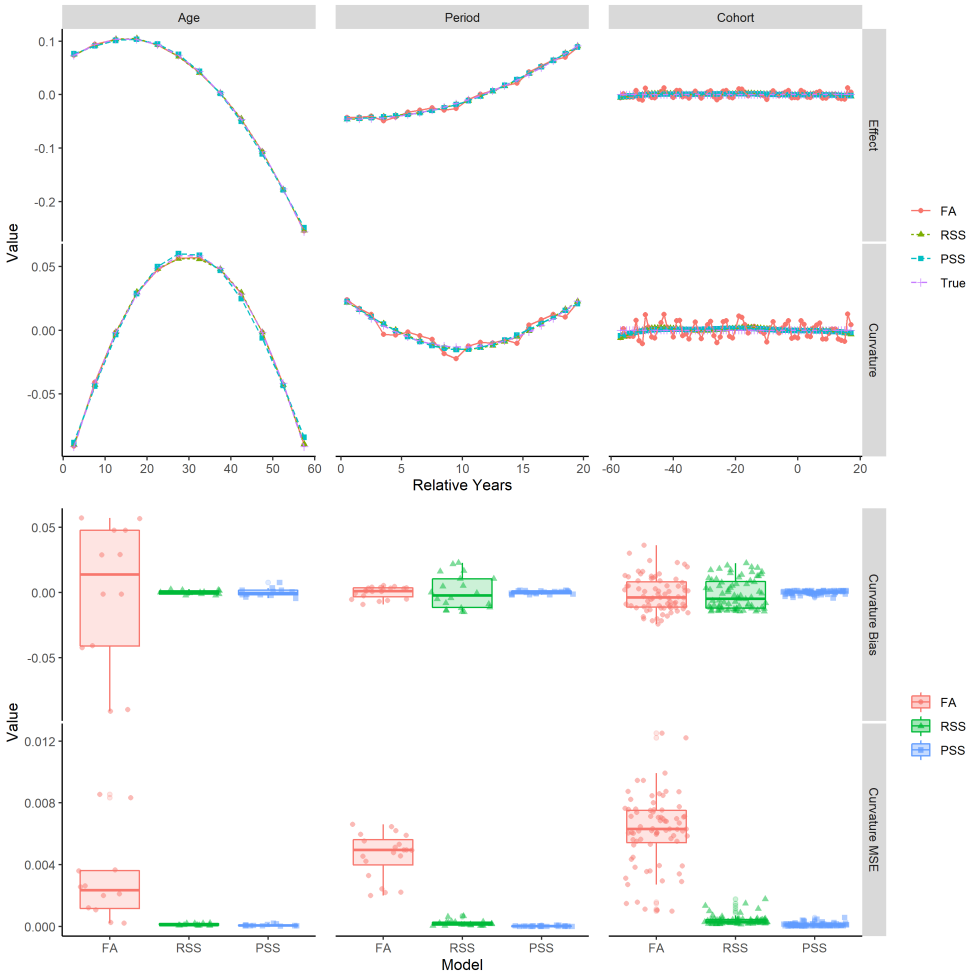


Figure S6: Simulation study results for unequal interval binomial data generated when only age and period effects are present.

S2.3 Gaussian simulation studies

Figure S7 presents the simulation study for unequal interval, normal data generated with all three temporal effects present.

The results from Figure S7 show the FA model displaying the cyclic pattern of the identification issues due to fitting an APC model the unequally aggregated APC data. As with the binomial results, the RSS and PSS results are hard to distinguish between but this is due to the choice of basis function used to approximate the true function. If the chosen basis function is not a good approximation to the true function, the estimates will not display the cyclic pattern of the added identification and the difference between the methods is hard to tell. When the basis is more informative (spans a larger space of the true function), the strengths of the PSS model become apparent; the PSS model gives the user confidence the results they have is not as influenced by the choice in basis as the RSS result. The penalisation of the estimates in the PSS model will always alleviate the added identification no matter the basis, this is not the case for the RSS model.

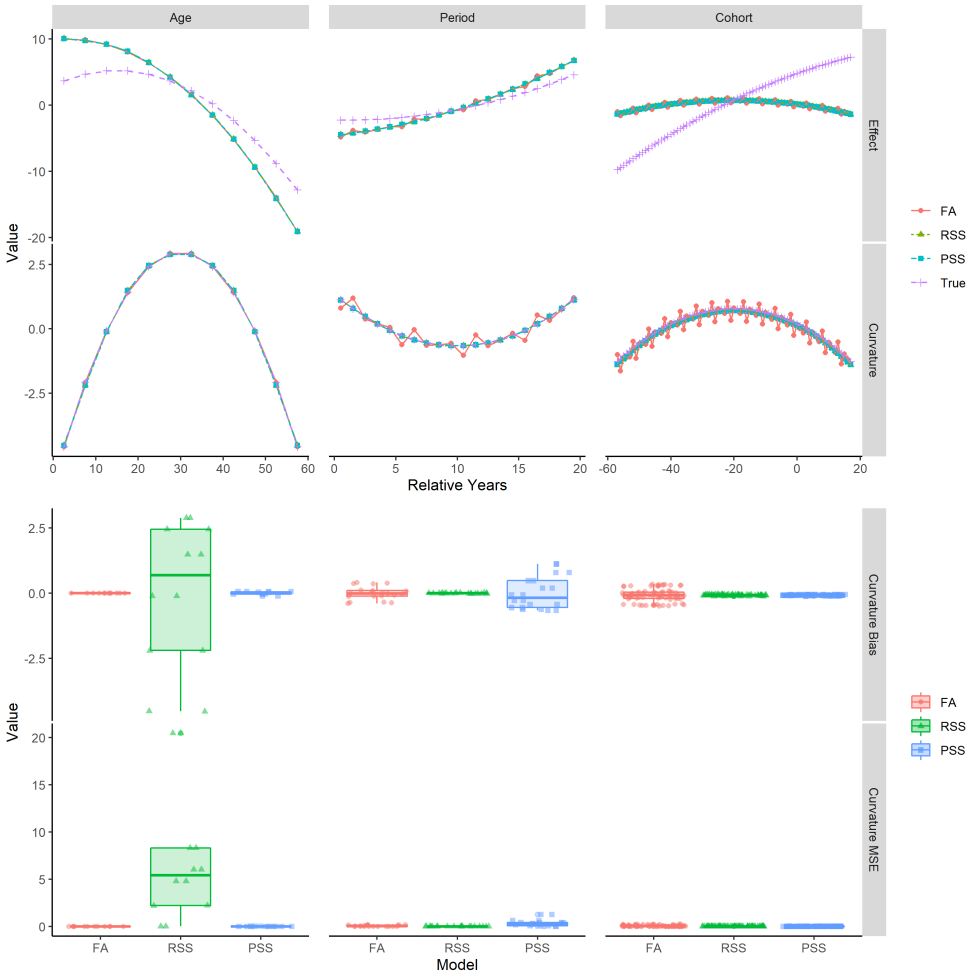


Figure S7: Simulation study results for unequal interval, $M = 5$, Gaussian data generated when all three temporal effects are present.

S2.4 Poisson simulation studies

Figure S8 shows the simulation study for unequal interval, Poisson data with all three temporal effects present. The interpretation of the results from the Poisson simulation studies is the same as those from the binomial and Gaussian distributions.

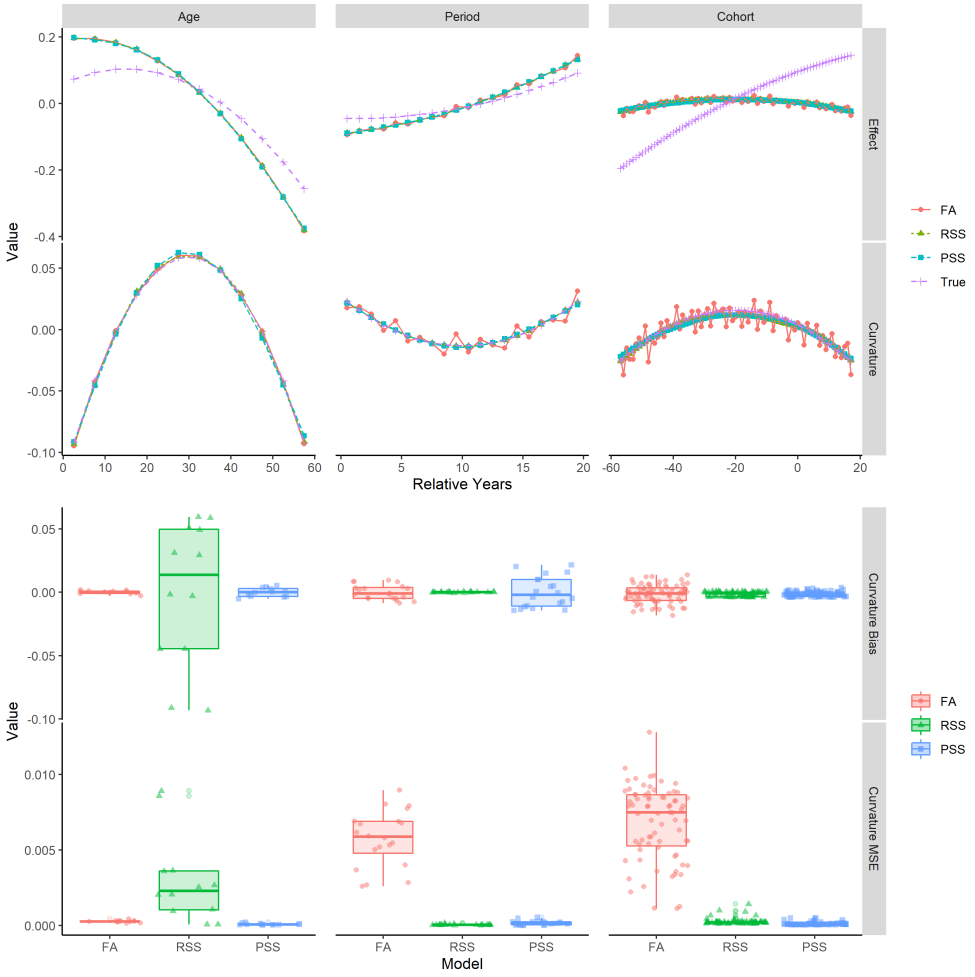


Figure S8: Simulation study results for unequal interval, $M = 5$, Poisson data generated when all three temporal effects are present.

S3 Application

Figure S9 shows the heat map of single-year log all cause mortality for ages 0-99 and years 1925-2015 in the UK. A gradient of dark blue means the mortality rate is low and red, mortality rate is high. For a fixed period and in the absence of cohort effects, the apparent age effects are seen along the y-axis. For the first year (1926), the age effect is the high mortality rate during infant years which decreases during childhood and adolescence and increases during adulthood and old age. For a fixed age and in the absence of cohort effects, the apparent period effects are seen along the x-axis. For the first year of life, the period effect is the gradual decreasing from high mortality rate in the earlier periods to the low mortality rate in more recent periods for the same age. The cohort effects are gradual, long-term changes that are due to both age and period and can be see looking along the bottom left to top right diagonal. For example, the cohort effect can be seen in the diagonal frontiers of blue/turquoise from 1945-1965 for ages 10-30 and the yellow/orange from 1980-2015 for ages 70-80.

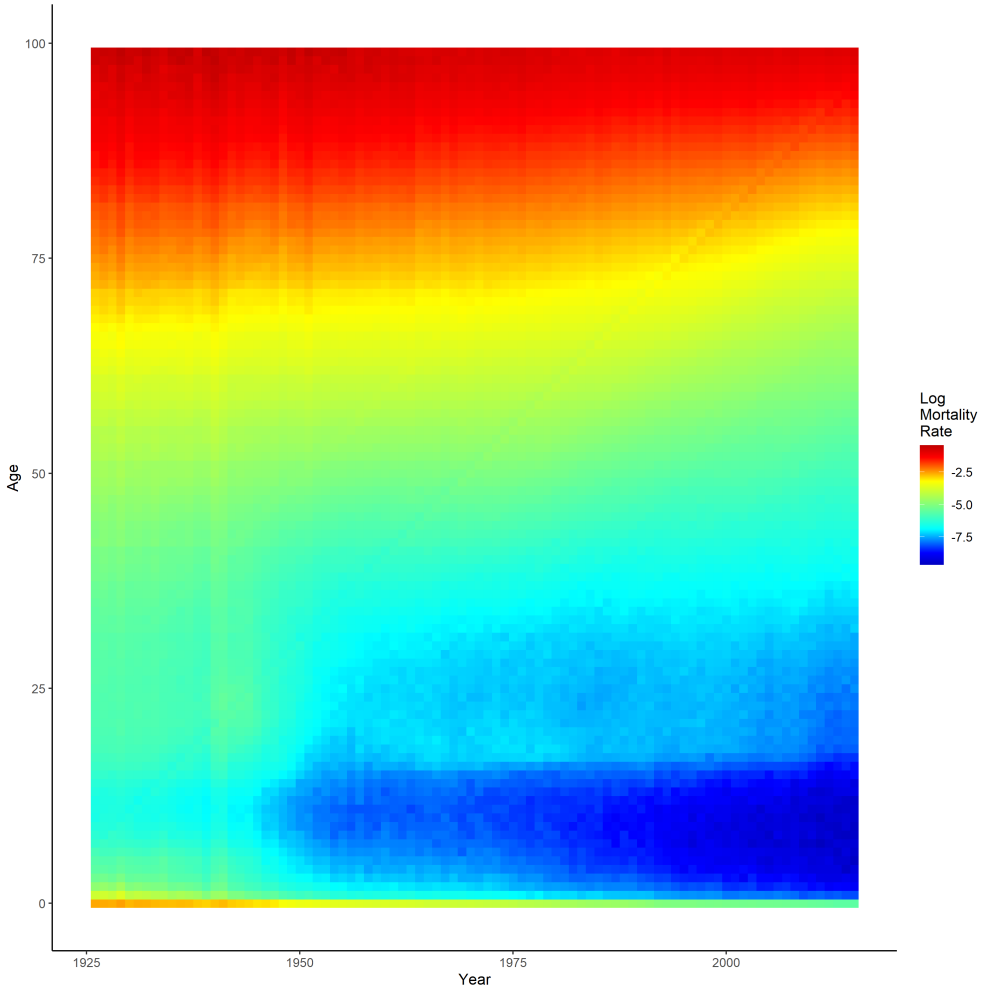


Figure S9: Single-year log all cause mortality for years 1926-2015 and ages 0-99 in the United Kingdom taken from the Human Mortality Database.

References

[1] Theodore R. Holford. Approaches to fitting age-period-cohort models with unequal intervals. *Statistics in Medicine*, 25:977–993, 2006.