Web-Appendix for the manuscript "A Semi-parametric Bayesian Dynamic Hurdle Model with an Application to the Health and Retirement Study", by Kiranmoy Das, Bhuvanesh Pareek, Sarah Brown and Pulak Ghosh.

1 Some Additional Plots from the HRS Data Analysis

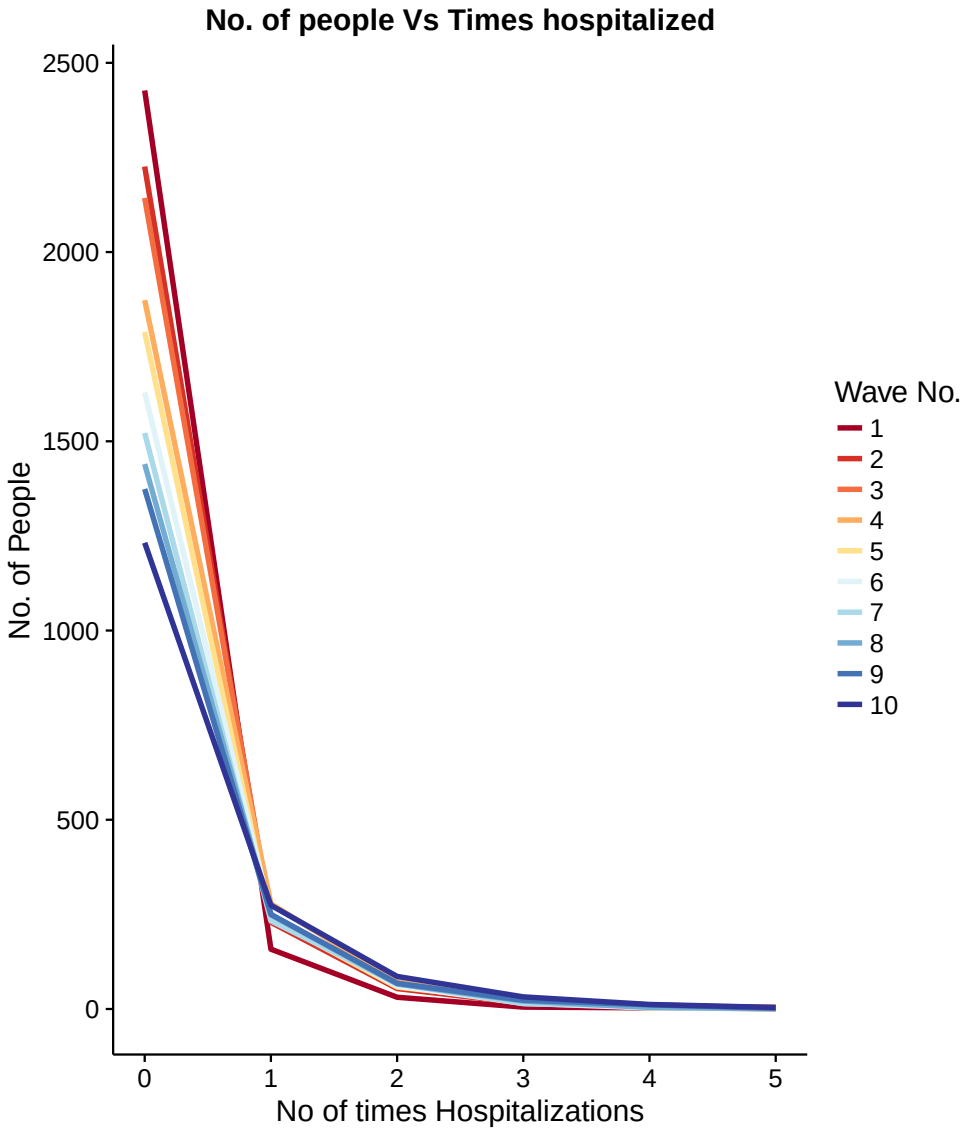

Figure S.1: Plot showing the distribution of individuals across the number of hospitalizations for different waves.

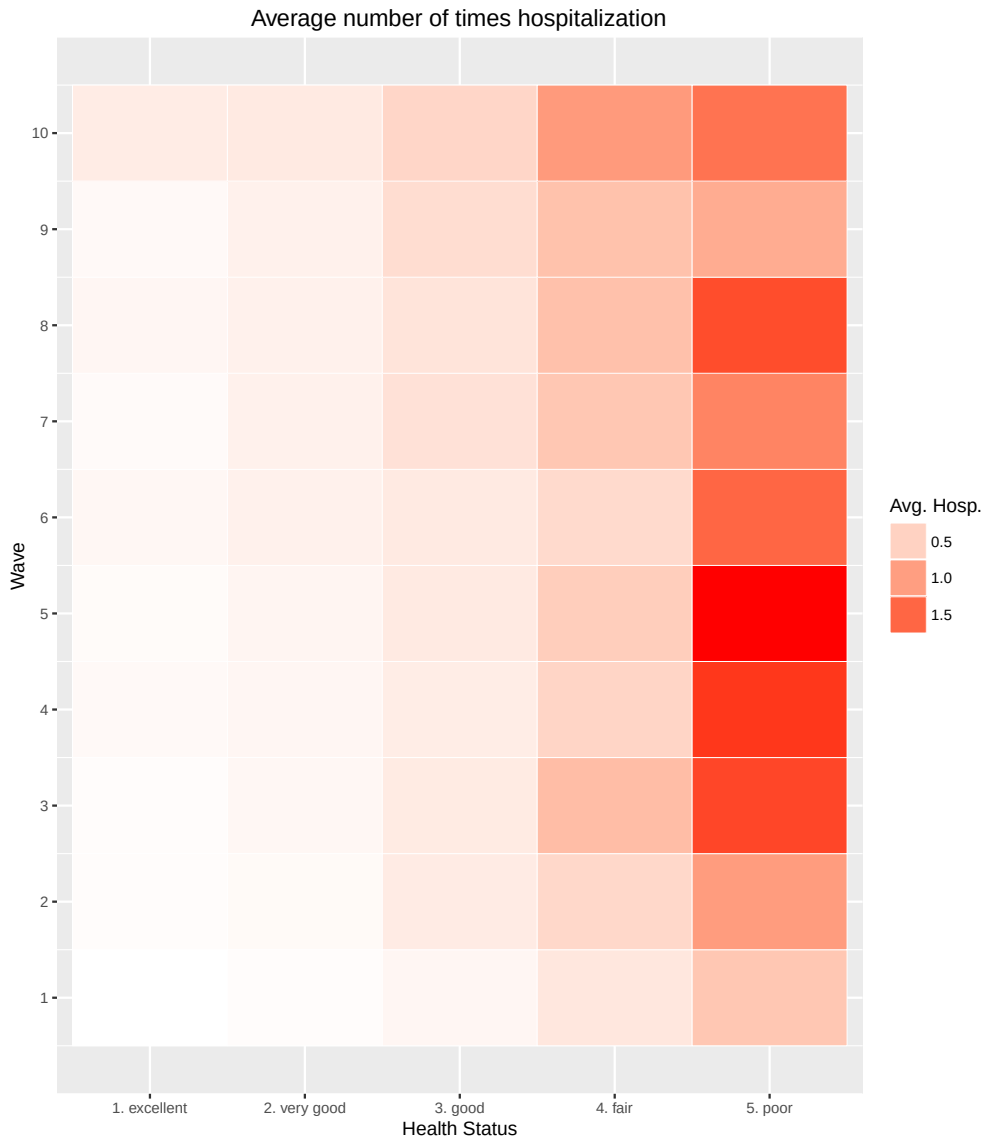

Figure S.2: Heat map displaying the average number of hospitalizations across different waves and self reported health status.

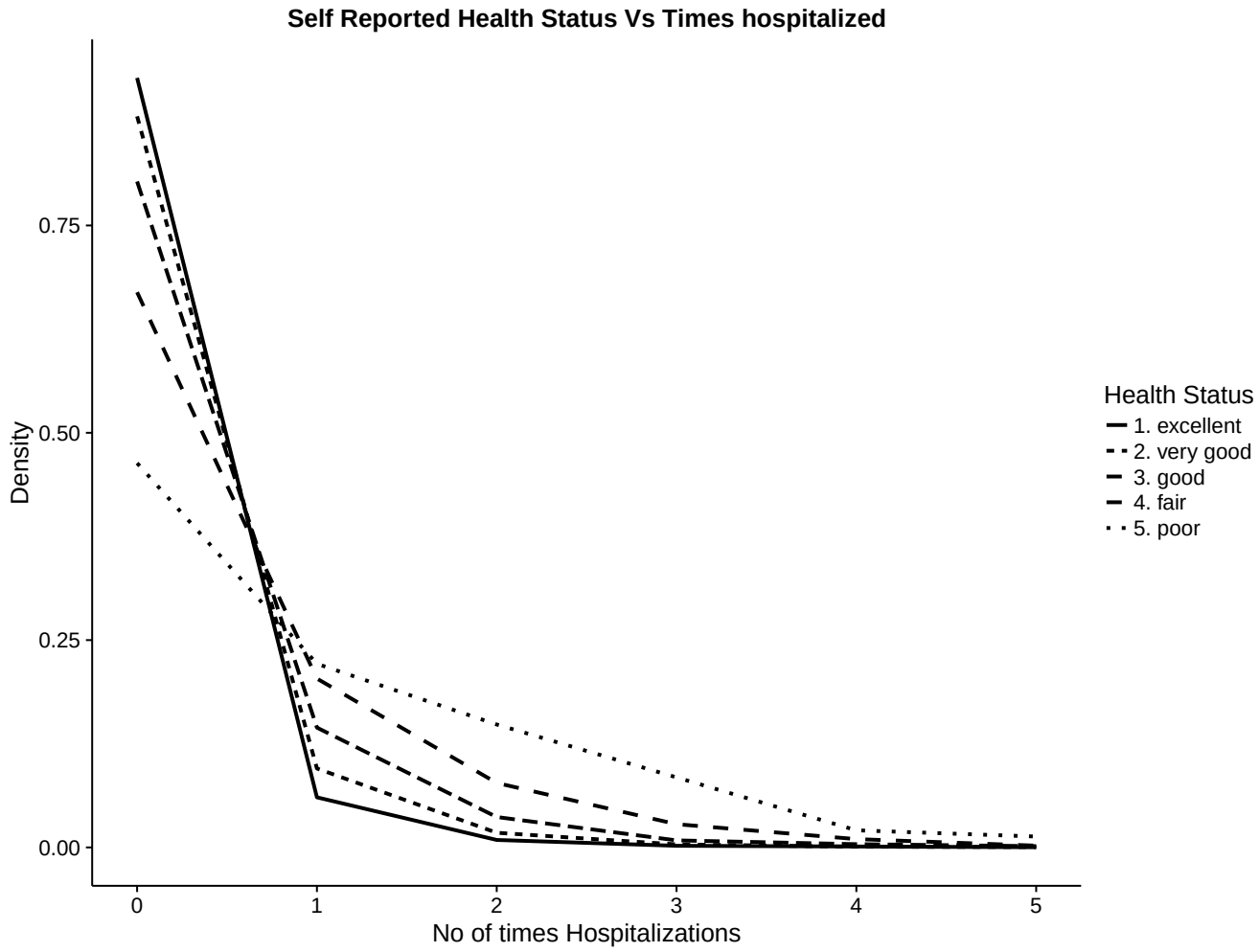

Figure S.3: Figure showing the distribution of individuals across the number of hospitalizations, for each self-reported health category.

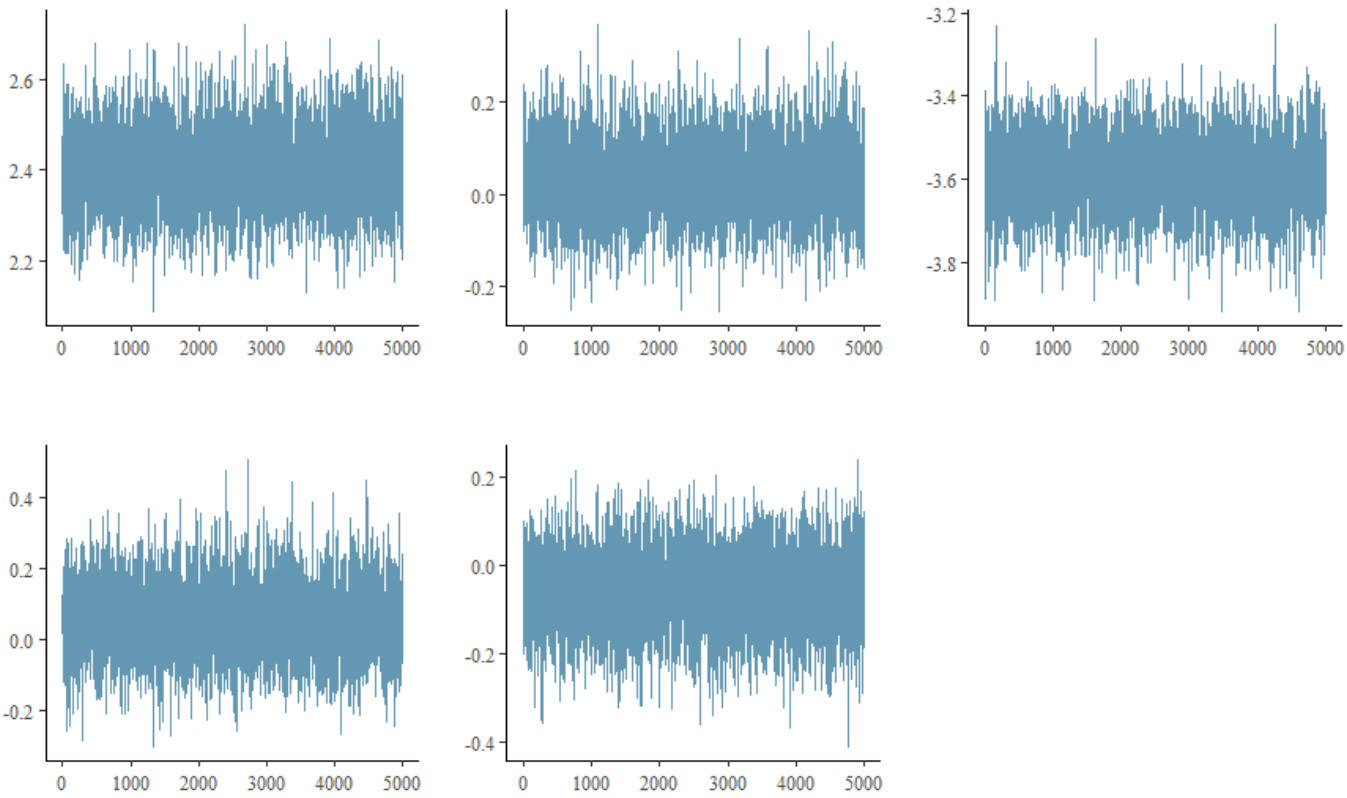

Figure S.4: Trace plots for some of the time-invariant regression coefficients in the HRS data analysis.

2 Simulation Study

We investigate the operating characteristics of the dynamic hurdle model and the MSBP prior through a simulation study. We consider a zero-inflated longitudinal count response, two covariates with time-varying effects and 10 covariates with time-invariant effects on the response.

We simulate data on 100 individuals belonging to 4 related groups (with size 30, 20, 20 and 30, respectively) at 10 different time points. Our response is a count variable and we consider two continuous predictors with time-varying effects on the response; and ten predictors with time-invariant effects. Among the predictors with time-invariant effects, there are eight continuous predictors, and the remaining two are categorical in nature.

Let $\mathbf{x}$ be the set of all covariates; thus $\mathbf{x} = [x_1, x_2, \ldots, x_{12}]^T$. Again, for each predictor x_i, we have measurements for $T = 10$ time points. Hence, $x_i = [x_i(1), x_i(2), \ldots, x_i(10)]^T$.

We simulate the predictors x_is, for $i = 1, 2, \ldots, 10$, from a multivariate normal density with mean= $\boldsymbol{\mu}_1 = [1, 3, 5, 4, 5, 6, 3.5, 5.5, 6, 3.8]^T$, and covariance matrix=$\Sigma$, which is the first order auto-regressive structure with $\rho_1 = 0.65$ and $\sigma^2 = 3.6$. The predictors x_{11} and x_{12} are generated at each time point from Bernoulli distributions with p=0.46 and 0.55, respectively.

Next, we generate p_{irt}, the probabilities of non-zero response. We first generate the iid samples of (η_i, b_i) from a bivariate normal density with mean vector=0 and covariance matrix= $\begin{bmatrix} 10 & 5.01 \\ - & 8.6 \end{bmatrix}$. Then the probabilities p_{irt} are generated from the following probit model: $p_{irt} = \Phi(x_i^T \boldsymbol{\delta} + \eta_i)$, with $\boldsymbol{\delta} = [0.04, 1.4, 2.5, 0.005, 3.6, 6.3, -2.56, 0.02, -0.003, 4.36, 1.1, -3.9]^T$. For each individual i, at each time point t, we sample from a uniform (0,1) distribution; and assign a zero value with probability=1-p_{irt}.

Next, for each individual i we find the time corresponding to the first non-zero response. If t is that time, then we generate T_i, the exact time of the first hospital visit for individual i. We generate T_i from a Weibull (1,5) distribution truncated below at $t - 1$, and truncated

above by t. Next, we sample Y_{irt} for each individual at each time point t as follows.

For the i-th individual, if t is the time for the first non-zero response, then we sample Y_{irt} from the distribution given in equation (2) of the main text, and if t is the time after the first non-zero response, then we sample Y_{irt} using the model given in equation (5) of the main text, with $k = 0, 1, \ldots, 5$. We consider x_1 and x_2 as the predictors with time-varying effects on the response, while the other 10 predictors are treated as the covariates with time-invariant effects on the response. Thus, we have $J = 2$ and $J' = 10$.

Next, we specify the parameter values for the model in equation (5) of the main text. For our simulation, we take $\beta_{j'kr} = \beta_{j'}$, and consider $\beta = [2.3, 4.5, 3.9, 0.04, 6.1, 3.2, 0.003, -1.65, -0.06, 1.4]$. For the time-varying part, we consider g_j=2, for $j = 1, 2$ and two knots at time 3 and 7, thus $S_j = 2$. Define $\mathbf{b}_{jkr} = [c_{jkr1}, c_{jkr2}]^T$. We show the parameter values for our simulation study for different choices of k and j in Tables S.1- S.4. Note that the set of parameters for both $\mathbf{b}$ and $\mathbf{c}$ are shared across the groups (r) for different values of k. In general, we consider the parameter values across the groups (r) and different counts (k) to be somewhat similar but not exactly the same. The values for the model parameters are based on the results from the HRS data analysis.

After generating the data, we fit three different proportional odds models to the simulated data: (i) a model with a completely different set of parameters for each group (the group-specific model); (ii) a model with exactly the same set of parameters for all groups (the common model); and (iii) a model with the proposed MSBP prior in the parameter set, where some of the parameters are similar across the groups. For each of these models, we consider two different specifications for modeling $G(Y_{irt}|Y_{irt} > 0)$: (i) a dynamic hurdle as discussed in Section 2.2 of the main text; and (ii) where $G(Y_{irt}|Y_{irt} > 0)$ is modeled at all the time points using the proportional odds model given in equation (6) of the main text.

We simulate 100 replicates of the data, and for each replicate we run the MCMC algorithm for 65,000 iterations, discard the first 5,000 ("burn-in") and thin the chains by saving every

10-th iteration. The model parameters are estimated by the respective posterior means. We compute the LPML values across all replications and Table S.5 shows the average LPML values for the different competing models. We note that the model with the MSBP prior as the shrinkage prior along with a dynamic hurdle gives the best fit (i.e. the largest LPML value). The common model (where all coefficients are exactly same for all groups) with a proportional odds model gives the worst fit. This justifies the importance of our proposed modeling approach.

Table S.6 shows the average estimated bias, average width of the 95% credible intervals (CIs) and the estimated coverage probabilities (based on 100 replications) of a randomly selected subset of the model parameters for the three competing models with larger LPML values (in Table S.5). We note that the proposed MSBP prior with the dynamic hurdle results in the smallest bias and shortest CI with a comparable coverage probability. This again illustrates the practical usefulness of our proposed model.

Table S.1: Model parameter values for the simulation study for $\mathbf{b}_{jkr}$; j=1, across groups for different k.

k	Group 1	Group 2	Group 3	Group 4
0	(1.74,0.86,1.31)	(1.71,0.82,1.24)	(1.80,0.91,1.31)	(1.75,0.86,1.34)
1	(1.69,0.85,1.42)	(1.80,0.73,1.26)	(1.63,0.81,1.19)	(1.77,0.81,1.34)
2	(1.82,0.93,1.24)	(1.77,0.79,1.14)	(1.69,0.76,1.24)	(1.72,0.88,1.30)
3	(1.77,0.88,1.37)	(1.76,0.81,1.37)	(1.76, 1.02,1.41)	(1.80,0.81,1.38)
4	(1.65,0.78,1.30)	(1.69,0.86,1.42)	(1.79,0.87,1.26)	(1.71,0.87,1.36)
5	(1.73,0.84,1.28)	(1.75,0.84,1.28)	(1.92,0.93,1.33)	(1.80,0.88,1.33)

Table S.2: Model parameter values for the simulation study for $\mathbf{b}_{jkr}$; j=2, across groups for different k.

k	Group 1	Group 2	Group 3	Group 4
0	(1.84,0.89,1.26)	(1.73,0.87,1.34)	(1.80,0.91,1.31)	(1.85,0.91,1.24)
1	(1.68,0.86,1.44)	(1.85,0.76,1.27)	(1.69,0.89,1.10)	(1.71,0.82,1.33)
2	(1.77,0.91,1.34)	(1.75,0.81,1.24)	(1.66,0.86,1.32)	(1.75,0.83,1.31)
3	(1.71,0.82,1.33)	(1.74,0.95,1.36)	(1.77, 1.08,1.49)	(1.80,0.86,1.34)
4	(1.75,0.78,1.30)	(1.63,0.84,1.44)	(1.69,0.87,1.31)	(1.74,0.88,1.32)
5	(1.73,0.86,1.23)	(1.77,0.88,1.38)	(1.82,0.83,1.23)	(1.80,0.78,1.43)

Table S.3: Model parameter values for the simulation study for $\mathbf{c}_{jkr}$; j=1, across groups for different k.

k	Group 1	Group 2	Group 3	Group 4
0	(0.74,0.96)	(0.77,0.98)	(0.79,0.95)	(0.81,0.95)
1	(0.79,0.92)	(0.72,0.95)	(0.71,0.91)	(0.76,0.92)
2	(0.81,1.01)	(0.79,0.91)	(0.76,0.94)	(0.77,0.88)
3	(0.77,0.94)	(0.70,0.93)	(0.72,0.97)	(0.79,0.89)
4	(0.74,0.95)	(0.86,0.99)	(0.83,0.90)	(0.74,0.84)
5	(0.75,0.99)	(0.67,0.91)	(0.73,0.99)	(0.76,0.95)

Table S.4: Model parameter values for the simulation study for $\mathbf{c}_{jkr}$; j=2, across groups for different k.

k	Group 1	Group 2	Group 3	Group 4
0	(0.64,0.75)	(0.67,0.73)	(0.59,0.71)	(0.61,0.75)
1	(0.69,0.72)	(0.62,0.75)	(0.61,0.71)	(0.66,0.72)
2	(0.71,0.81)	(0.68,0.71)	(0.56,0.70)	(0.67,0.78)
3	(0.67,0.74)	(0.60,0.73)	(0.62,0.77)	(0.69,0.79)
4	(0.64,0.75)	(0.76,0.75)	(0.63,0.70)	(0.64,0.74)
5	(0.65,0.79)	(0.57,0.71)	(0.63,0.79)	(0.66,0.75)

Table S.5: Bayesian model selection: LPML values for different competing models in the simulation study.

Model specification	LPML value
MSBP prior with dynamic hurdle	**-336.9**
Group-specific model with dynamic hurdle	-397.2
Common model with dynamic hurdle	-539.8
MSBP prior with non-dynamic hurdle	-401.4
Group-specific model with non-dynamic hurdle	-463.8
Common model with non-dynamic hurdle	-639.5

Table S.6: Estimated Bias, average width of confidence interval, and the estimated coverage probability for some of the model parameters in the simulation study.

	Group specific + dynamic hurdle		MSBP + non-dynamic hurdle		MSBP + dynamic hurdle	
Parameter	Bias	C.I. width (Cov.Prob)	Bias	C.I. width (Cov.Prob)	Bias	C.I. width (Cov.Prob)
b_{1210}	0.38	0.87(0.96)	0.35	0.92(0.96)	0.11	0.26(0.95)
b_{2321}	0.33	0.69(0.95)	0.37	0.74(0.96)	0.09	0.24(0.95)
b_{1532}	0.29	0.58(0.95)	0.26	0.55(0.95)	0.10	0.31(0.95)
c_{1212}	0.37	0.73(0.95)	0.34	0.68(0.95)	0.08	0.25(0.95)
c_{2321}	0.35	0.84(0.97)	0.42	0.91(0.97)	0.11	0.28(0.96)
c_{2542}	0.41	0.93(0.96)	0.39	0.86(0.95)	0.13	0.23(0.95)
β_{131}	0.39	0.59(0.96)	0.33	0.64(0.96)	0.06	0.29(0.95)
β_{554}	0.28	0.63(0.96)	0.35	0.58(0.95)	0.08	0.25(0.94)
β_{823}	0.36	0.65(0.96)	0.34	0.67(0.96)	0.10	0.33(0.95)