Table 1: Event Counts by Cluster

Event	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
CL	32,759	7,003	14,970	10,158	39,359	104,249
ER	86,410	16,975	24,709	13,927	19,410	161,431
HO	95,398	23,884	23,695	19,344	29,154	191,475
NP	18,700	6,704	3,797	3,770	6,141	39,112
PO	290,212	38,676	297,999	126,677	30,940	784,504
RX	450,456	655,085	444,585	1,020,559	104,473	2,675,158

SUPPLEMENTARY MATERIALS

Supplemental Material A

This supplemental material contains basic patient summaries for each of the clusters in Tables 1, 2, and 3. Table 1 gives the number of events for each type and cluster. Table 2 gives the total exposure (in years) across all patients in each cluster for each of the control covariates. We present exposure here instead of patient counts because the primary survival model input is exposure, not patient count. Table 3 gives the total number of patients in each cluster stratified by state and urbanicity.

Supplemental Material B

In this supplemental material, we present an example of a virtual patient to demonstrate of how the patient level data is transformed into the model inputs. Consider patient A who enters the system 30 days into the measurement period, visits the emergency room on days

Table 2: Length of Time Under Observation (in years) by Cluster and Control Variable

Var. Family	Var. Value	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Time	Total	930,365	253,665	220,521	143,076	37,538
Age Group	4-5	395,286	107,361	94,376	63,915	14,701
	6-14	513,240	139,965	121,934	76,421	21,991
	15-17	21,839	6,339	4,211	2,740	846
Race	White	393,511	118,066	92,889	71,209	16,700
	Black	433,568	107,513	101,917	53,516	15,710
	Other	103,286	28,086	25,715	18,351	5,128
Health	Healthy	492,493	119,387	103,052	55,877	14,287
Status	Minor	162,819	45,748	30,631	23,521	3,913
	Chronic	268,704	86,037	85,828	61,389	18,787
	Severe	6,349	2,494	1,010	2,290	551
Medicaid	Blind/Disabled	85,004	24,888	18,832	17,970	5,458
Eligibility	Foster Care	21,263	6,390	5,054	4,855	1,067
	Other	824,098	222,387	196,635	120,251	31,013

Table 3: Patient Counts by Cluster

Var. Family	Var. Value	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Count	Total	237,693	68,640	64,357	44,003	11,707
State	GA	54,196	7,220	15,337	6,755	4,018
	LA	60,892	15,911	9,013	5,500	874
	MS	16,283	9,528	2,155	3,375	954
	MN	1,404	5,396	1,606	1,826	808
	NC	44,757	14,720	21,558	14,616	2,842
	TN	46,162	16,314	14,689	11,930	2,210
Urbanicity	Urban	173,644	45,331	48,435	31,142	8,730
	Suburban	49,403	16,376	13,412	9,511	2,058
	Rural	14,647	6,931	2,510	3,350	919

90, 125, and 270 and fills a prescription for an inhaler on days 155 and 300. Suppose that the patient's Medicaid enrollment lapses between days 100 and 115. Let the length of the measurement period be 365 days. Table 4 provides a data summary for patient A. Table 5 demonstrates how the input table would be shaped for our estimation algorithm with last event type as a covariate in the study. This example demonstrates how our algorithm handles censoring, multivariate survival data, and time varying covariates.

One property of the exponential baseline assumption is that we can simply subtract the start time from the stop time due to the memoryless property of the exponential and use the same interarrival time across all event types but with different event indicator values. This allows for extreme computational efficiency that reduces to simple matrix algebra when estimating the coefficients $\vec{\boldsymbol{\beta}}$. For other distributions, such as the Weibull or log-logistic, we would not be able to simply subtract the start times from the stop times in order

Table 4: Sample Data Summary

Start Time	Stop Time	Event Type	Health Status
30	90	ER	Chronically Ill
90	100	0	Chronically Ill
115	125	ER	Chronically Ill
125	155	RX	Chronically Ill
155	270	ER	Chronically Ill
270	300	RX	Chronically Ill
300	365	0	Chronically Ill

Table 5: Sample Input for Estimation Algorithm

τ	$\delta_{\mathrm{ER}}(\tau)$	$\delta_{\mathrm{PO}}(\tau)$	$\delta_{\mathrm{RX}}(\tau)$	$D_{r,1}(\tau)$: Last Event	$D_{r,2}(\tau)$: Health Status
60	1	0	0	0	0
10	0	0	0	ER	0
10	1	0	0	ER	0
30	0	0	1	ER	0
115	1	0	0	RX	0
30	0	0	1	ER	0
65	0	0	0	RX	0

to get the length of the time period until event or censoring, thus greatly increasing the complexity of the algorithm. Furthermore, as shown in Appendix C, with the exponential distribution we can derive simply the provider transition networks.

Supplemental Material C

In this supplement, we describe the initialization of the algorithm in order to produce the results in Section 4. We use a random initialization with $K = 3, 4, \ldots, 9$ clusters. The random initialization begins with a random (hard) clustering assignment for each patient to a cluster k, $k \in \{1, \ldots, K\}$. That is, $\boldsymbol{Z}_r^{(1)}$ has only one entry with value one and all others are zero. The vector $\vec{\boldsymbol{b}}$ begins with all values set to 0, such that $\pi_{rk} = 1/K$ for all k and for all r. The EM algorithm then proceeds through the iterations to find the maximum likelihood given this initialization. We repeat this with 10 different random initializations for each value of K.

Figure 1 displays the resulting likelihood of each of the initializations. In this article we present the model outputs from the initialization that produces the highest likelihood.

Supplemental Material D

In this supplement, we provide the estimated model parameters. Table 6 contains the raw proportional hazards parameters from the algorithm.

The multinomial logistic regression model parameters are given in Table 20. Tables 8 through 12 contain the average interarrival times in years between events for the baseline group for Clusters 1-5. This is equal to $\exp(-\beta_{0ks}\beta_{\text{s,Event}})$.

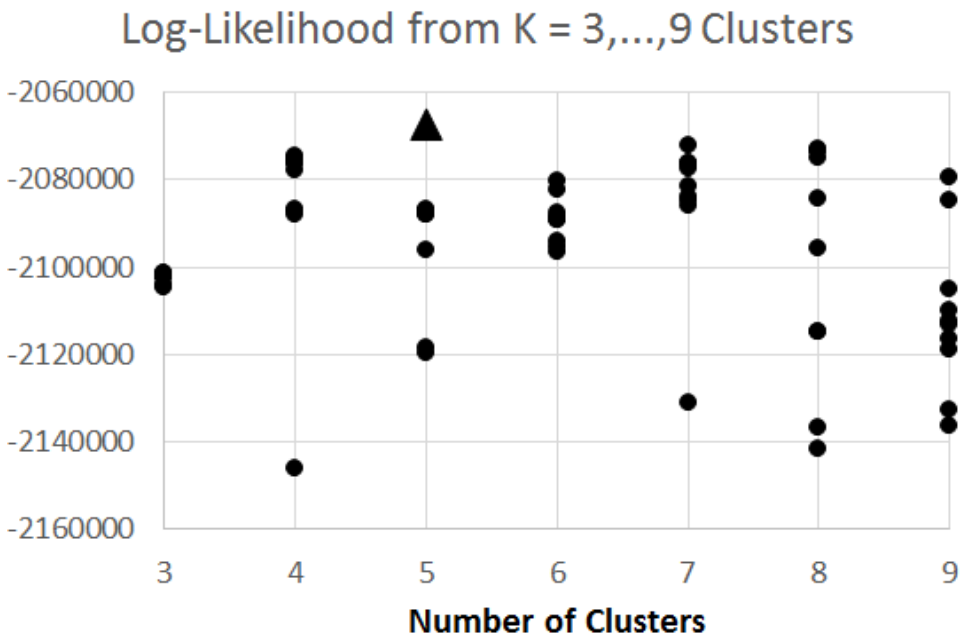

Figure 1: The resulting log-likelihood plotted from $N = 10$ different initializations for $K = 3, \ldots, 9$ clusters. We chose the model that resulted in the highest likelihood after convergence, with $K = 5$ clusters denoted by ▲.

Table 6: Proportional Hazards Coefficients

Var. Family	Var. Value	CL	ER	HO	NP	PO	RX
Baseline	$k = 1$	0.04	0.17	0.16	0.42	0.03	0.53
	$k = 2$	0.04	0.13	0.16	1.84	0.03	0.21
	$k = 3$	0.07	0.19	0.17	1.49	0.02	1.62
	$k = 4$	0.08	0.17	0.21	4.29	0.03	0.95
	$k = 5$	0.58	0.68	0.78	1.90	0.13	0.91
Medicaid	Blind/Disabled	1.25	0.85	0.91	1.05	0.71	0.74
Eligibility	Foster Care	0.99	0.49	0.67	1.14	0.75	0.79
Health	Healthy	0.31	0.10	0.11	0.12	0.12	0.47
Condition	Minor Ilness	1.25	0.96	1.22	1.54	1.31	1.28
	Severe Illness	4.74	4.03	6.79	2.02	2.75	3.99
Race/	Black	1.26	2.06	1.43	0.78	1.05	1.02
Ethnicity	Other	1.23	1.36	1.34	0.86	1.03	1.15
Age Group	6-14	0.64	0.90	0.95	0.98	0.88	0.86
	15-18	0.92	1.56	1.53	0.93	1.24	1.22
Previous	CL	13.96	0.69	0.91	1.62	1.07	1.53
Event	ER	1.39	2.70	3.64	1.68	1.45	1.37
	HO	1.26	1.14	3.77	1.79	0.94	1.11
	NP	1.43	0.81	1.01	2.79	2.51	2.22
	PO	1.00	0.73	0.70	2.14	25.40	1.17
	RX	1.67	0.76	0.85	2.22	1.38	2.20

Table 7: Multinomial Logistic Regression Coefficients

Var. Family	Var. Value	$k = 1$	$k = 2$	$k = 3$	$k = 4$
	Baseline	-0.11	1.79	0.56	-0.79
State	LA	-0.99	-0.70	-1.35	-2.34
	MS	-0.90	-1.37	-2.05	-1.60
	MN	-1.04	-1.00	-2.01	-1.36
	NC	0.10	-0.86	-0.34	-1.01
	TN	-0.22	-0.95	-0.83	-1.39
Urbanicity	Suburban	-0.12	-0.05	-0.10	-0.25
	Rural	-0.17	-0.27	-0.66	-0.18
Access	Travel	0.16	0.55	0.55	0.57

Table 8: Average Interarrival Times (in years): Cluster 1

Previous Event	CL	ER	HO	NP	PO	RX
CL	1.71	8.28	6.83	31.57	1.24	1.46
ER	17.12	2.12	1.70	23.23	1.38	1.41
HO	18.96	5.04	1.64	35.83	1.71	1.33
NP	23.73	7.86	8.90	1.33	1.63	1.11
PO	14.30	7.51	7.30	24.56	0.86	1.07
RX	16.67	7.11	6.11	13.42	1.63	1.11

Table 9: Average Interarrival Times (in years): Cluster 2

Previous Event	CL	ER	HO	NP	PO	RX
CL	2.05	11.28	6.98	31.60	3.15	0.33
ER	20.52	2.89	1.74	23.25	3.51	0.32
HO	22.72	6.86	1.68	35.86	4.33	0.30
NP	28.45	10.71	9.09	4.13	1.33	0.25
PO	17.14	10.23	7.46	24.47	2.19	0.24
RX	19.98	9.69	6.25	13.43	2.17	0.19

Table 10: Average Interarrival Times (in years): Cluster 3

Previous Event	CL	ER	HO	NP	PO	RX
CL	1.00	7.43	6.37	44.27	0.40	0.41
ER	10.00	1.91	1.59	32.58	0.45	0.40
HO	11.08	4.52	1.53	50.24	0.55	0.38
NP	13.88	7.06	8.30	1.86	0.53	0.32
PO	8.36	6.74	6.81	34.29	0.28	0.30
RX	9.75	6.38	5.70	18.81	0.28	0.24

Table 11: Average Interarrival Times (in years): Cluster 4

Previous Event	CL	ER	HO	NP	PO	RX
CL	0.91	8.26	5.12	36.75	0.69	0.14
ER	9.17	2.12	1.28	27.04	0.77	0.14
HO	10.16	5.03	1.23	0.95	41.70	0.13
NP	12.72	7.85	6.68	1.54	0.91	0.11
PO	7.66	7.50	5.48	28.47	0.48	0.10
RX	8.92	7.10	4.59	15.62	0.48	0.08

Table 12: Average Interarrival Times (in years): Cluster 5

Previous Event	CL	ER	HO	NP	PO	RX
CL	0.12	2.12	1.42	0.32	6.99	0.72
ER	1.23	0.54	0.35	5.14	0.80	0.31
HO	1.36	1.29	0.34	7.93	0.99	0.29
NP	1.70	2.01	1.85	0.29	0.95	0.25
PO	1.03	1.92	1.52	0.24	5.41	0.50
RX	1.20	1.82	1.27	0.19	2.97	0.50

Supplemental Material E

While our paper has focused on the practical significance of the effects of the covariates on patient utilization we present a summary of the statistical significance of the parameter estimates in this appendix. We calculate the estimates of the variance of the parameters by first finding the estimate for the Fisher information. The Fisher information is

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\middle|\theta\right],$$

where θ is a model parameter and X is a random variable. In our case, we do not know the parameters $\vec{\boldsymbol{\beta}}$ and $\vec{\boldsymbol{b}}$, and thus we must estimate the Fisher information. We have already calculated the 2nd derivative of the log-likelihood function with respect to β_{ps}, β_{0ks}, and b_{jk} in Section 3.3.3 from the main text. Then the estimates for the Fisher information are

$$\widehat{I(\beta_{ps})} = I(\widehat{\beta}_{ps}) = \sum_{r,k,l_r}\left[\tau_{l_r}\widehat{Z}_{rk}D_{rp}^2(\tau_{l_r})\exp\left\{\widehat{\boldsymbol{\beta}}_{\boldsymbol{ks}}{}^{\intercal}\boldsymbol{D_r}(\tau_{l_r})\right\}\right],$$

$$\widehat{I(\beta_{0ks})} = I(\widehat{\beta}_{0ks}) = \sum_{r,l_r}\left[\widehat{Z}_{rk}\tau_{l_r}\exp\left\{\widehat{\boldsymbol{\beta}}_{\boldsymbol{ks}}{}^{\intercal}\boldsymbol{D_r}(\tau_{l_r})\right\}\right], \text{and}$$

$$\widehat{I(b_{jk})} = I(\widehat{b}_{jk}) = \sum_{r=1}^{R}\widehat{\pi}_{rk}(1-\widehat{\pi}_{rk})E_{rj}^2.$$

We use Wald's test statistic in order to calculate the p-value of the parameters,

$$\frac{\hat{\theta}-0}{\widehat{V(\theta)}},$$

where $\widehat{V(\theta)} = 1/I(\widehat{\theta})$, and $\theta = \{\beta_{ps}, b_{jk}\}$. That is, we assume that the control and explanatory covariates have no effect on the baseline rate of events or cluster membership, respectively. There are only 6 combinations of event types and control covariates for which the p-value is greater than 0.001: *Race: Other*, NP (p-value = 0.036); *Minor Illness*, ER

Table 13: 99% Confidence Intervals for Baseline Rates by Event and Cluster

Event	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
CL	(0.041, 0.043)	(0.034, 0.036)	(0.070, 0.074)	(0.076, 0.081)	(0.572, 0.599)
ER	(0.172, 0.177)	(0.125, 0.132)	(0.190, 0.199)	(0.170, 0.180)	(0.666, 0.701)
HO	(0.159, 0.164)	(0.154, 0.162)	(0.169, 0.177)	(0.210, 0.220)	(0.759, 0.793)
NP	(0.029, 0.031)	(0.028, 0.031)	(0.020, 0.022)	(0.024, 0.027)	(0.128, 0.141)
PO	(0.523, 0.532)	(0.204, 0.211)	(1.611, 1.637)	(0.935, 0.955)	(0.890, 0.922)
RX	(0.419, 0.424)	(1.834, 1.853)	(1.477, 1.494)	(4.266, 4.307)	(1.884, 1.923)

(p-value = 0.001); *Prior Event: CL*, NP (p-value = 0.099); *Prior Event: HO*, NP (p-value = 0.027), *Prior Event: NP*, CL (p-value = 0.921), and *Foster Care*, CL (p-value = 0.777). All of the multinomial logistic regression parameter estimates are significant at a $\alpha = 0.001$ critical value. For the baseline rates, we provide 99% confidence intervals in Table 13. From this table we can see that the following cluster pairs and event types have statistically insignificant differences at the $\alpha = 0.01$ confidence level: Clusters 1 and 2: HO and NP; Clusters 1 and 4: ER. These findings suggest that the practical interpretations provided in the main body of the paper are also statistically significant with few exceptions.

Supplemental Material F

In this supplemental material, we provide the provider transition networks for the other covariate familes: age group, race/ethnicity, and Medicaid eligibility categorization in Figures 2, 7, and 4, respectively. In each case, we consider the baseline group for the other covariates.

The *Age* networks do not show that *Age 6-14* patients show higher probability connec-

tions into HO, but less variation otherwise while and more reliance on RX while *Age 15-17* shows greater variation in provider types with more transitions leading to CL, ER, HO, and PO. The *Race* networks have the same pattern with greater variations for *Black* and *Other* groups in Cluster 1, 3, and 5. Clusters 2 and 4 non-white patients utilize more HO and PO, respectively. Finally, it appears that for the baseline group *Blind/Disabled* and *Foster Care* patients have less variation than those in *Other* and have stronger transitions leading to RX. These plots show that except for the *Medicaid Eligibility* variable the baseline group is less variational than others indicating a possibility that white children, age 4-5 are better managed in asthma care.

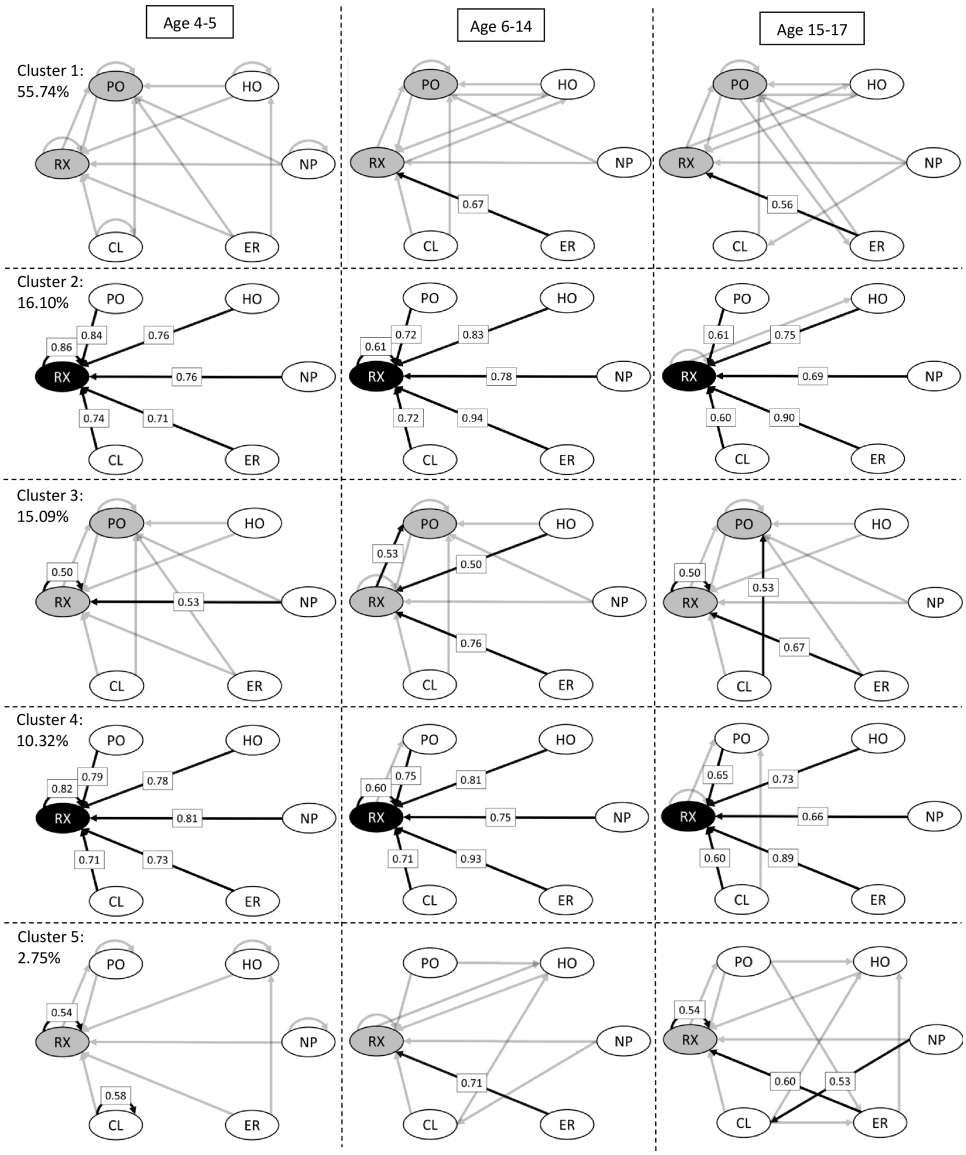

Figure 2: Provider networks for *Age* subgroups induced from the proportional hazards coefficients. The following rules were used in setting the grayscale of the coefficients and nodes: $< 0.2 \rightarrow$ not shown/white, $[0.2, 0.5) \rightarrow$ gray, and $\geq 0.5 \rightarrow$ black.

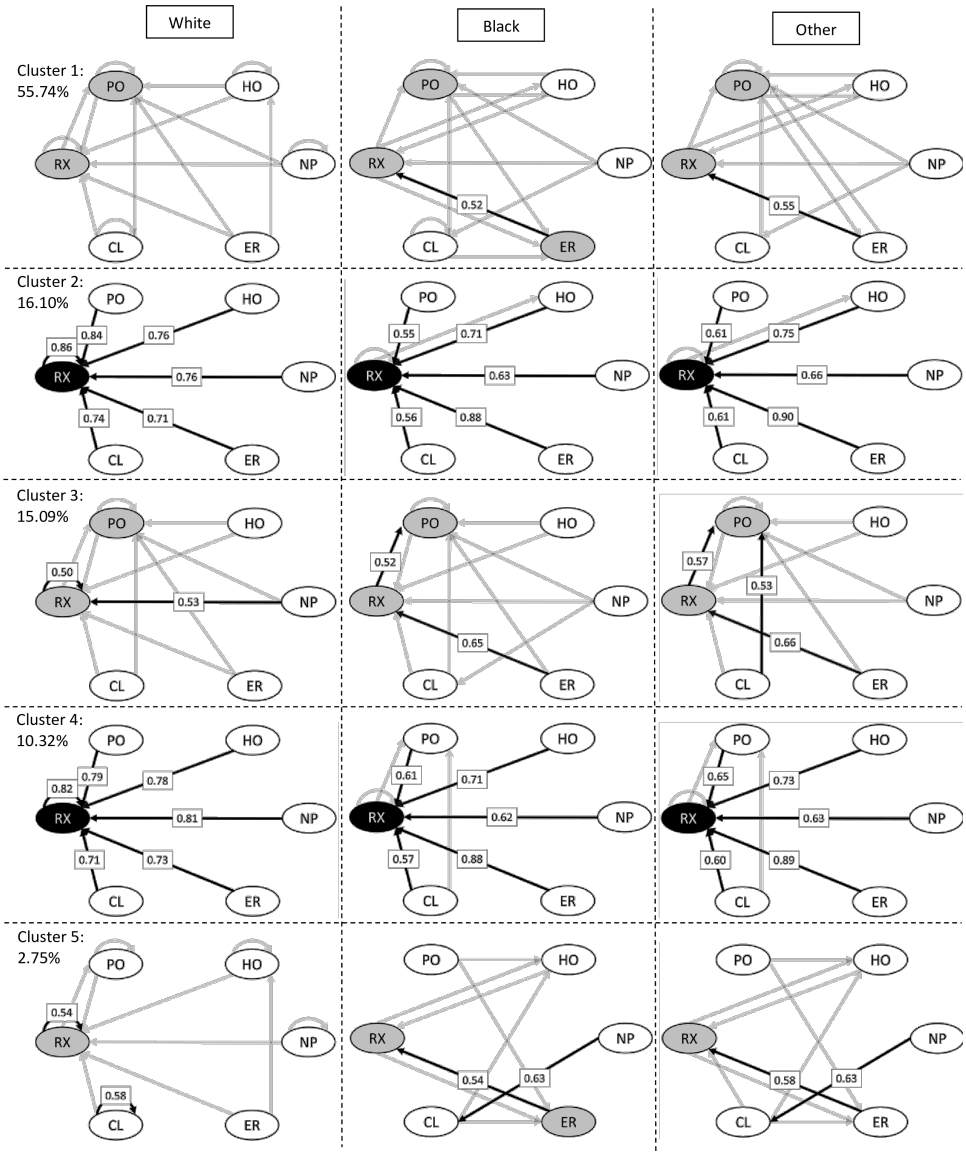

Figure 3: Provider networks for *Race* induced from the proportional hazards coefficients. The following rules were used in setting the grayscale of the coefficients and nodes: $< 0.2 \rightarrow$ not shown/white, $[0.2, 0.5) \rightarrow$ gray, and $\geq 0.5 \rightarrow$ black.

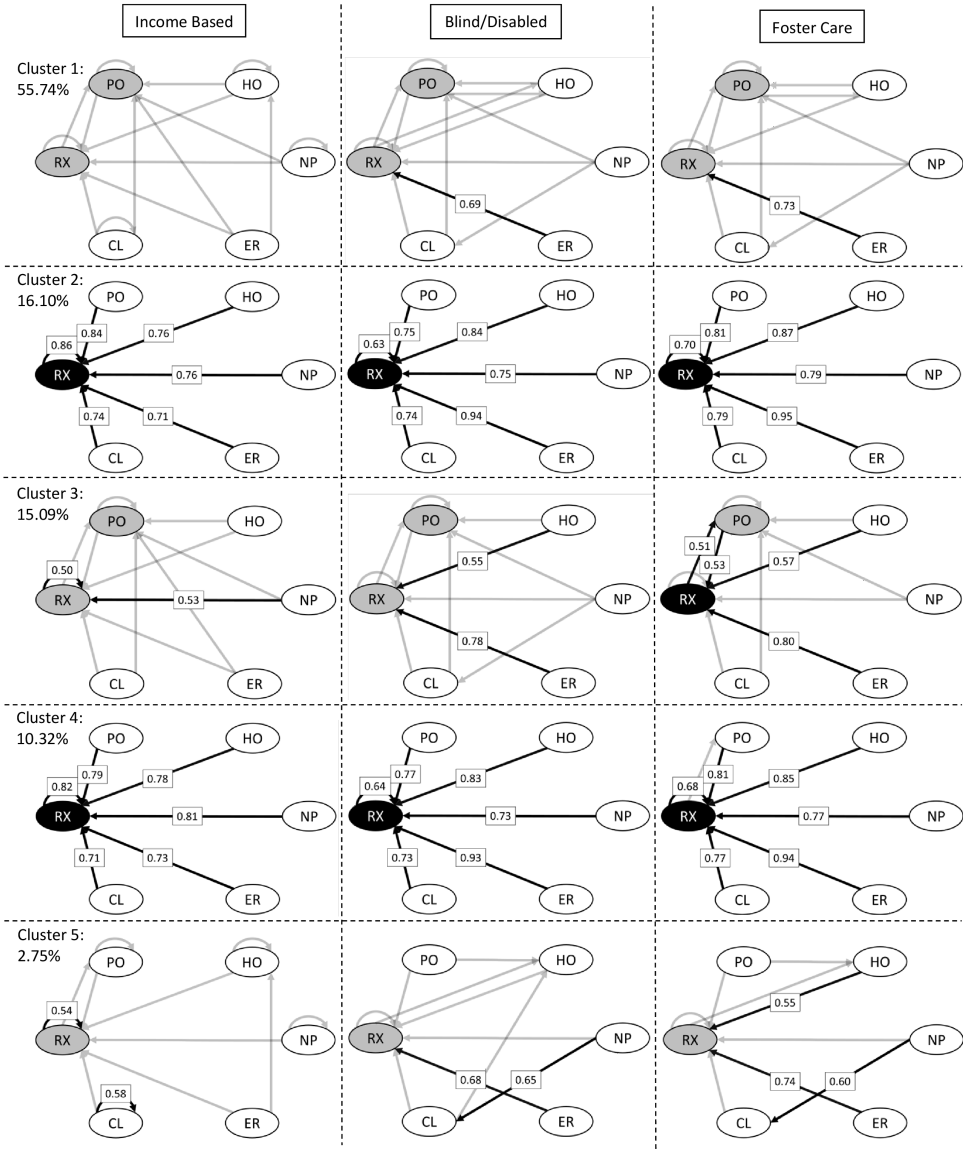

Figure 4: Provider networks for *Medicaid Eligbility* subgroups induced from the proportional hazards coefficients. The following rules were used in setting the grayscale of the coefficients and nodes: $< 0.2 \rightarrow$ not shown/white, $[0.2, 0.5) \rightarrow$ gray, and $\geq 0.5 \rightarrow$ black.

Supplemental Material G

In this supplemental material we present how we use stratified sampling methods to quantify the uncertainties in the estimated regression coefficients and to assess the robustness of the model. Supplemental material E provides detailed derivations on how statistical significance was analyzed using Fisher Information. In this appendix, we use a sub-sampling approach to obtain a sample from the empirical distribution of the estimated model parameters and draw inference based on the empirical confidence intervals.

For our study, drawing random sub-samples from the entire population is not sensible, since there are large heterogeneity across different patient sub-populations. Therefore, we stratify the study population on control variables including Medicaid eligibility, health condition, race and Ethnicity, and age group, and draw 20% of individuals from each sub-population. We then fit the model for each sub sampled data. We iterate the process for 100 times, and construct a 90% empirical confidence interval based on the 5^{th} and 95^{th} quantile. The statistical significance of each covariate and the skewness of the estimates can be analyzed using the derived confidence intervals.

There are two types of model parameters: cluster-independent and cluster-dependent. Cluster-independent variables are invariant of clusters, including all the parameters estimated in the proportional hazard model, excluding the baselines. Since these parameters are exponents, a statistically significant parameter has both its 5^{th} and 95^{th} quantile greater than one or less than one. Cluster-dependent parameters depend on the cluster membership, including the estimates from the multinomial logistic regression and the baselines from the proportional hazard model. The estimation of such variables across different sub samples are more challenging to analyze since the cluster membership will change for each sub-sample. Therefore, an additional membership matching algorithm is implemented be-

fore the inference step. The uncertainty quantification of the multinomial logistic regression is more difficult to compare since the baseline cluster is different across sub-samples. We will demonstrate the inference on such parameters in supplemental material H when we simplify the problem into the two-clusters case. Table 14, 15 displays the mean estimation of the parameters. Table 16,17 displays the 5^{th} and 95^{th} quantile of the estimation as empirical confidence intervals.

Table 14: Underlying True Estimation of Proportional Hazards Coefficients for CL, ER, HO and the Mean Esimation from Multiple Subsamples

		CL		ER		HO	
Var. Family	Var. Value	True	Mean	True	Mean	True	Mean
Age Group	Age 6-14	0.66	0.64	0.91	0.90	0.96	0.95
	Age 15-17	0.95	0.92	1.55	1.56	1.51	1.53
Race/	Black	1.23	1.26	2.06	2.06	1.42	1.43
Ethnicity	Other	1.25	1.23	1.37	1.36	1.35	1.34
Health	Minor Illness	1.21	1.25	0.95	0.96	1.21	1.22
Condition	Chronic	3.25	3.26	10.33	10.34	9.42	9.41
	Severe Illness	5.18	4.74	4.09	4.03	6.77	6.79
Previous	CL	18.81	13.96	0.80	0.69	1.11	0.91
Event	ER	1.66	1.39	2.83	2.70	3.90	3.64
	HO	1.64	1.26	1.24	1.14	4.15	3.77
	RX	1.41	1.43	0.79	0.81	0.98	1.01
	NP	1.23	1.00	0.79	0.73	0.80	0.70
	PO	1.40	1.67	0.72	0.76	0.82	0.85
Medicaid	Blind/Disabled	1.28	1.25	0.86	0.85	0.91	0.91
Eligibility	Foster care	1.00	0.99	0.49	0.49	0.67	0.67

Table 15: Underlying True Estimation of Proportional Hazards Coefficients for NP, PO, RX and the Mean Esimation from Multiple Subsamples

		NP		PO		RX	
Var. Family	Var. Value	True	Mean	True	Mean	True	Mean
Age Group	Age 6-14	0.98	0.98	0.91	0.88	0.86	0.86
	Age 15-17	0.93	0.93	1.29	1.24	1.19	1.22
Race/	Black	0.78	0.78	1.03	1.05	1.05	1.02
Ethnicity	Other	0.87	0.86	1.02	1.03	1.19	1.15
Health	Minor Illness	1.53	1.54	1.32	1.31	1.28	1.28
Condition	Chronic	2.10	2.13	8.45	8.20	8.33	8.26
	Severe Illness	2.02	2.02	2.85	2.75	4.01	3.99
Previous	CL	1.64	1.62	1.25	1.07	1.34	1.53
Event	ER	1.64	1.68	1.60	1.45	1.36	1.37
	HO	1.73	1.79	1.11	0.94	1.13	1.11
	RX	2.76	2.79	2.59	2.51	2.36	2.22
	NP	2.12	2.14	24.92	25.40	0.93	1.17
	PO	2.29	2.22	1.08	1.38	1.99	2.20
Medicaid	Blind/Disabled	1.04	1.05	0.76	0.71	0.71	0.74
Eligibility	Foster care	1.14	1.14	0.76	0.75	0.78	0.79

Table 16: 5^{th} and 95^{th} Quantile Estimation of the Proportional Hazards Coefficients for CL, ER, HO

		CL		ER		HO	
Var. Family	Var. Value	5^{th}	95^{th}	5^{th}	95^{th}	5^{th}	95^{th}
Age Group	Age 6-14	0.63	0.71	0.89	0.93	0.94	0.98
	Age 15-17	0.8	1.17	1.49	1.63	1.43	1.6
Race/	Black	1.14	1.31	2	2.11	1.37	1.46
Ethnicity	Other	1.13	1.37	1.31	1.43	1.29	1.4
Health	Minor Illness	1.12	1.31	0.91	0.99	1.16	1.26
Condition	Chronic	3.07	3.46	10.06	10.58	9.2	9.66
	Severe Illness	4.08	6.59	3.58	4.6	5.98	7.56
Previous	CL	14.18	25.32	0.7	0.91	0.93	1.29
Event	ER	1.39	1.97	2.63	2.95	3.63	4.09
	HO	1.3	1.88	1.13	1.3	3.86	4.33
	RX	1.17	1.55	0.74	0.82	0.92	1.02
	NP	0.93	1.56	0.69	0.86	0.68	0.89
	PO	1.2	1.92	0.67	0.77	0.75	0.88
Medicaid	Blind/Disabled	1.14	1.45	0.82	0.9	0.87	0.96
Eligibility	Foster care	0.81	1.21	0.45	0.53	0.63	0.72

Table 17: 5^{th} and 95^{th} Quantile Estimation of the Proportional Hazards Coefficients for NP, PO, RX

		NP		PO		RX	
Var. Family	Var. Value	5^{th}	95^{th}	5^{th}	95^{th}	5^{th}	95^{th}
Age Group	Age 6-14	0.96	1	0.85	0.95	0.83	0.89
	Age 15-17	0.88	1	1.13	1.42	1.12	1.3
Race/	Black	0.76	0.8	0.98	1.08	1.02	1.08
Ethnicity	Other	0.85	0.9	0.94	1.1	1.16	1.22
Health	Minor Illness	1.51	1.6	1.24	1.42	1.25	1.31
Condition	Chronic	2.06	2.1	7.96	8.95	7.99	8.57
	Severe Illness	1.93	2.1	2.15	3.68	3.66	4.36
Previous	CL	1.54	1.7	1.02	1.47	1.19	1.63
Event	ER	1.57	1.7	1.41	1.77	1.26	1.48
	HO	1.66	1.8	0.95	1.26	1.05	1.22
	RX	2.73	2.8	2.31	2.81	2.2	2.48
	NP	2.03	2.2	22.1	29.4	0.8	1.16
	PO	2.19	2.4	0.91	1.37	1.84	2.23
Medicaid	Blind/Disabled	1.01	1.1	0.67	0.83	0.67	0.77
Eligibility	Foster care	1.09	1.2	0.64	0.86	0.73	0.83

Supplemental Material H

In this supplemenal material, we evaluate whether the model parameters are identifiable and quantify the bias in the estimates using a simulation study. We first fit the model with the original data to obtain the proportional hazards coefficients, which will give us estimates of the inter-arrival time between events conditioned on the control variables. Since the proposed approach uses an exponential proportional hazards model, we simulate the rate between events from exponential distributions using the derived rate estimates, while keeping other covariates fixed. We iterate such process for 100 iterations, and obtain the mean estimates as well as the 90% confidence interval. We simplified the problem into **two clusters** case in order to better estimate the cluster-dependent covariates. Table 18,19 gives the mean parameter estimation as well as the length between 5^{th} and 95^{th} quantile of the estimation. The lengths of the 90% confidence intervals are mostly close to zero. Table 20 displays the mean estimation of the logistic regression as well as the upper and lower bound of the 90% confidence interval.

Table 18: Mean parameter estimation of the proportional hazard coefficients and the length of the 90% confidence interval for CL, ER, HO

		CL		ER		HO	
Var. Family	Var. Value	Mean	Length	Mean	Length	Mean	Length
Baseline	k=1	0.002	0.000	0.002	0.000	0.003	0.000
	k=2	0.002	0.000	0.002	0.000	0.002	0.000
Age Group	Age 6-14	0.682	0.016	0.919	0.011	0.987	0.013
	Age 15-17	1.017	0.048	1.656	0.058	1.627	0.065
Race/	Black	1.135	0.025	2.039	0.021	1.408	0.015
Ethnicity	Other	1.197	0.043	1.333	0.019	1.316	0.023
Health	Minor Illness	1.099	0.052	0.925	0.018	1.167	0.026
Condition	Chronic	4.011	0.111	11.118	0.151	10.208	0.141
	Severe Illness	8.156	0.769	4.552	0.272	7.687	0.564
Previous	CL	30.082	1.408	1.019	0.028	1.412	0.042
Event	ER	1.668	0.062	3.170	0.065	4.209	0.100
	HO	1.636	0.060	1.366	0.028	4.559	0.087
	RX	1.696	0.052	0.834	0.011	1.110	0.017
	NP	1.404	0.061	0.891	0.032	0.881	0.032
	PO	1.799	0.053	0.852	0.013	0.902	0.013
Medicaid	Blind/Disabled	1.457	0.040	0.926	0.016	1.002	0.021
Eligibility	Foster care	1.015	0.066	0.503	0.022	0.692	0.030

Table 19: Mean parameter estimation of the proportional hazard coefficients and the length of the 90% confidence interval for NP, PO, RX

		NP		PO		RX	
Var. Family	Var. Value	Mean	Length	Mean	Length	Mean	Length
Baseline	k=1	0.303	0.012	0.000	0.000	0.030	0.001
	k=2	0.027	0.001	0.001	0.000	0.009	0.000
Age Group	Age 6-14	0.931	0.020	0.927	0.016	0.862	0.013
	Age 15-17	0.812	0.031	1.394	0.051	1.134	0.032
Race/	Black	0.588	0.011	1.046	0.015	0.900	0.011
Ethnicity	Other	0.746	0.022	1.027	0.026	1.075	0.018
Health	Minor Illness	1.604	0.027	1.296	0.044	1.234	0.028
Condition	Chronic	2.249	0.028	9.323	0.157	8.038	0.095
	Severe Illness	2.734	0.189	3.291	0.226	4.037	0.201
Previous	CL	2.908	0.125	1.437	0.048	2.212	0.090
Event	ER	2.221	0.066	1.529	0.045	1.644	0.051
	HO	2.470	0.093	1.028	0.026	1.317	0.049
	RX	3.493	0.104	2.883	0.065	2.035	0.059
	NP	4.133	0.236	27.615	1.156	1.515	0.070
	PO	2.329	0.062	1.375	0.028	2.864	0.066
Medicaid	Blind/Disabled	1.226	0.047	0.775	0.017	0.790	0.017
Eligibility	Foster care	1.389	0.088	0.783	0.042	0.854	0.039

Table 20: Mean parameter estimation of the logistic regression, along with the 5^{th} and 95^{th} quantile of the parameter estimation.

Var. Family	Var. Value	Mean	5^{th}	95^{th}
	Baseline	-0.55	-0.58	-0.52
State	LA	0.60	0.59	0.61
	MS	0.75	0.73	0.76
	MN	0.11	0.10	0.13
	NC	0.68	0.67	0.69
	TN	0.77	0.76	0.78
Urbanicity	Suburban	0.07	0.06	0.07
	Rural	0.17	0.16	0.18
Access	Travel	-0.27	-0.29	-0.26

Supplemental Material I

One of the key motivations of using a parametric exponential survival model is its computational scalability and efficiency. In supplemental material G and H, we demonstrate that the proposed estimation approach provides stable and unbiased estimation, which is an indication of the robustness of the proposed method under different input data. In addition, in this Appendix, we motivate the exponential distribution assumption by specifically investigating the distribution of the inter-arrival time between two events for each subpopulation based on the control variables. The distributions for each subpopulation are approximately Exponential distributed, with small error margin. We compare the empirical histogram with the fitted Exponential distribution. We display the comparison result for two largest subpopulations in Figure 5 and Figure 6.

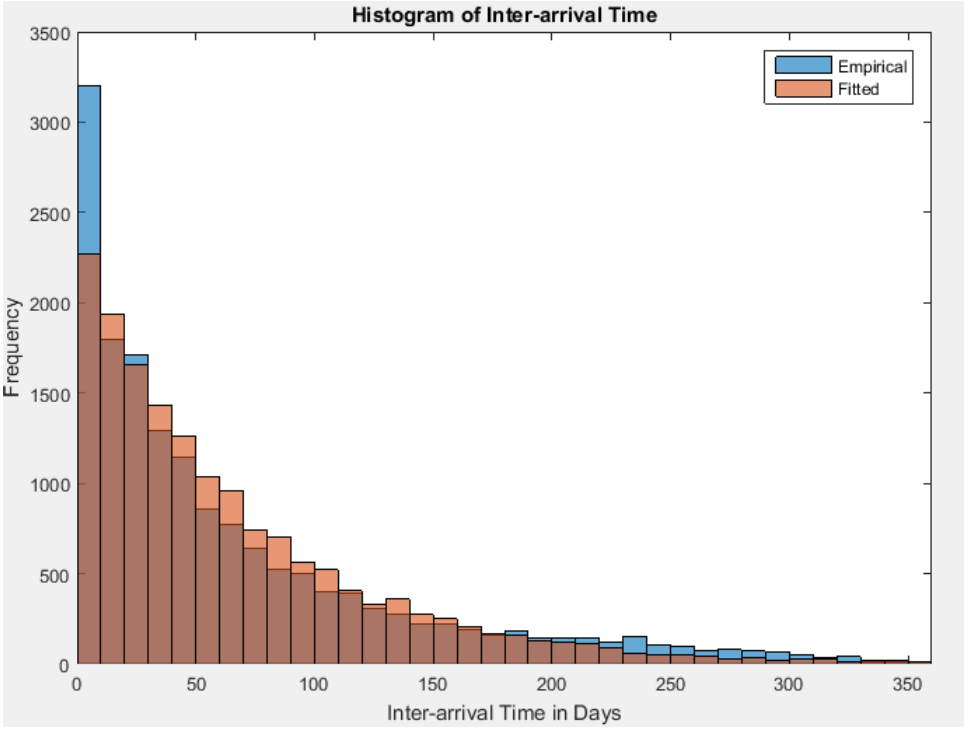

Figure 5: Histograms of the inter-arrival time between a PO event and a RX event for age 4-5, black, minor ill patients, who are not eligible for Medicaid for blindness/disability or foster care. The blue histogram shows the empirical distribution of the inter-arrival time. The brown histogram is the theoretical pdf of the fitted Exponential distribution. The fitted exponential distribution has a rate of 83.0, with a 90% confidence interval of [82.4, 83.6].

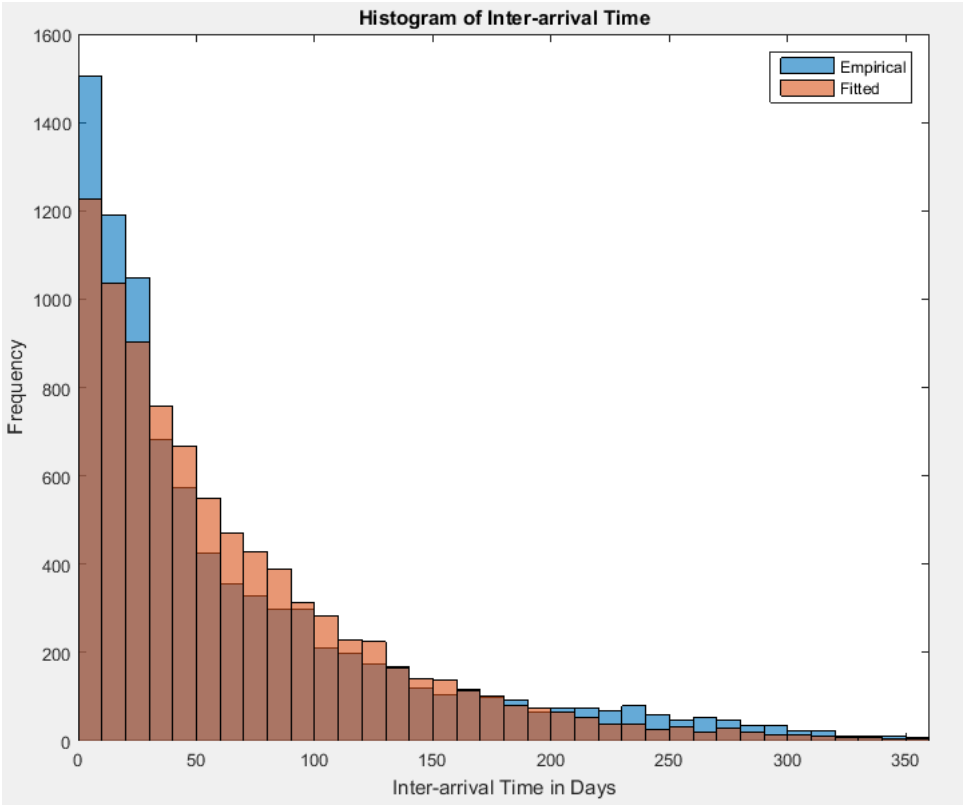

Figure 6: Histograms of the inter-arrival time between a RX event and a PO event for age 4-5, black, minor ill patients, who are not eligible for Medicaid for blindness/disability or foster care. The blue histogram shows the empirical distribution of the inter-arrival time. The brown histogram is the theoretical pdf of the fitted Exponential distribution. The fitted exponential distribution has a rate of 66.4, with a 90% confidence interval of [65.0, 67.8].

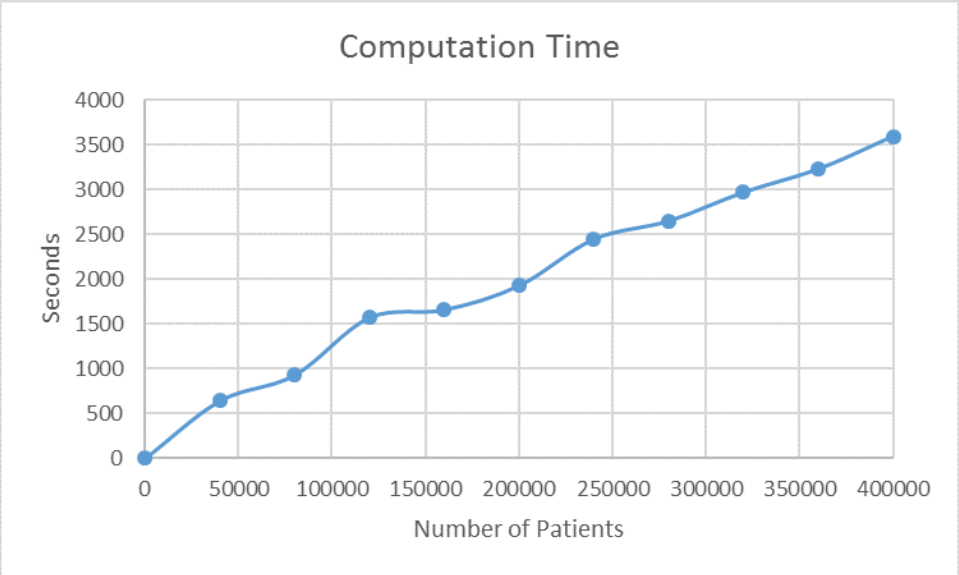

Figure 7: Computational time with different problem sizes.

Supplemental Material J

The proposed algorithm for parameter estimation is computationally attractive. We can estimate the model within 2-3 hours for a dataset of more than 400,000 individuals with about 6 million inter-arrival times and 5 clusters. To demonstrate the scalability of the algorithm, we fit the model with sub-samples of the population starting with 10%, with an increment of 10%, under 2 clusters case, as is shown in Figure 7. The computation is approximately linear with the number of samples.

Supplemental Material H

Reproducibility of research should be part of any research paper published. We developed a synthetic dataset that replicates our sample data to some extent. Due to heavy regula-

tion and strict guidelines from CMS regarding identifiability of patient level data, we are not able to replicate real dataset, but instead use the distribution and counts of events calculated from the real data as the input of the synthetic dataset generation. The data files and MATLAB codes are provided as supplemental materials.

Synthetic Data Generation:

We first summarize the real data by providing some basic statistics as follows:

- *Controlled Variables.csv*: We record the population for each subpopulation classified by the control variables. The populations are used as weights which dictates the number of patients we sample from each subpopulation.
- *Survival Observations.csv*: For each subpopulation descripted by the control variables, we record the number of visits to each event, which are then converted into probabilities used to generate synthetic events. Average inter arrival time are calculated from data for each subpopulation. We simulate the inter arrival time by sampling from an exponential distribution. Note that ideally inter arrival time should not only depend on subpopulation, but also event types. However, dude to identifiability issue, we are not able to extract at that resolution.
- *Explanatory Variables.csv:* We record the population for each subpopulation classified by explanatory variables such as state, urbanicity and travel distance.
- *DataGenerator.m:* This MATLAB file generates the necessary variables required by the main program. User can input the desired number of patients at the first line.

MATLAB Codes

Here we list a brief explanation of each major MATLAB functions used in the main program. The outputs include the cluster parameter Z and pi, coefficient estimate β and b for each EM iteration.

- *Overall_Script.m*: Organizes the estimation step and maximization step functions.
- *Initial_PH_Est_Step.m*: Performs estimation of the initial proportional hazard coefficients under the assumption that K=1.
- *PH_Est_Step_K_5,m*: Takes as input the parameter values estimated from Initial_PH_Est_Step and finds the updated values for the proportional hazards coefficients given current cluster parameters, Z, assuming K=5.
- *Multinom_Est_Step.m*: Takes the current cluster parameters, Z, and updates the multinomial logistic regression coefficients.
- *Estep.m*: Performs the expectation step of the EM algorithm given current values of the multinomial and proportional hazards parameters.