Cluster Analysis to Identify Predictors of Change in Activities of Daily Living in Bladder Cancer Patients

Mojgan Golzy, PhD[1] and Katie Murray, DO[2]

University of Missouri-Columbia

[1]Department of Health Management and Informatics
[2]Department of Surgery-Urology Division

School of Medicine

October 5, 2019

Significance

- In the US, an estimated 80,470 new cases of bladder cancer will be diagnosed in 2019, with an estimated 17,670 deaths from this disease.
- The treatments for BC may include repeat operations with transurethral resection of tumors and for muscle invasive cancer it often involves chemotherapy, radiation therapy and/or aggressive surgeries including radical cystectomy.
- Understanding how treatment affect a patient's daily life and quality of life is key to informed decision-making.
- One important measure of physical function is activities of daily living (ADLs): self-care activities such as bathing, eating, dressing, walking, getting up the chair and using toilet.

Previous Studies

- Smith AB. et al. (2018), Impact of bladder cancer on health-related quality of life, BJUI Int. 2018; 121:549-557.
- Murray KS, et al. (2018) Functional status in Patients Requiring Nursing Home Stay After Radical Cystectomy. J Urol. 2018 Nov;121:39-43.
- Winters BR. et al. (2017), Health Related Quality of Life Following Radical Cystectomy: Comparative Analysis from the Medicare Health Outcomes Survey, J Urol. 2018 Mar;199(3):669-675.
- Fung C. et al. (2014) Impact of Bladder Cancer on Health Related Quality of Life in 1,476 Older Americans: A Cross-Sectional Study, J Urol. 2014 Sep;192(3):690-5.

Objectives

- To identify clusters of bladder cancer patients based on their baseline level of functionality.
- Determine the association of clusters of BC patients and patient-centered outcomes, patient reported falls, mental and physical quality-of-life measures and survival outcome.
- Identify factors associated with "decreased functional status" in two subpopulations: 1) newly diagnosed BC patients (reduction in ADL from pre to post diagnose); and 2) patients with a diagnosis of BC at least 5 years prior (reduction in ADL in the past two years).

Statistical Method

- Cluster analysis: Identification of groups of related observations that are cohesive and separated from other groups.
- Some examples are: Gene expression profiling; The segmentation of images; Patient subgroup identification; and Text classification
- Clustering is subjective and depends on the selection of features and clustering algorithm.

Data Resource and Selected Features

- The SEER-MHOS data resource links data from the Surveillance, Epidemiology and End Results (SEER) program of cancer registries and data from the Medicare Health Outcomes Survey (MHOS).
- Data includes fifteen cohorts of MHOS data (baseline and a two years follow-up surveys), clinical, demographic and cause of death information.
- The selected features for the clustering are: Baseline ADL measures and comorbidities.

Algorithms

- We implemented three clustering algorithms.
  - Standard K-means algorithm (Duda & Hart, 1973; Hartigan & Wong, 1979);
  - Spherical K-means Clustering (Dhillon & Modha, MR 2001);
  - Poisson-Kernel Based Clustering (Golzy & Markatou, JCGS 2019).

Standard K-Means Clustering (Duda & Hart, 1973)

- The standard K-means clustering is an iterative algorithm; given a set of N observations $\mathcal{X}$, where each observation is a d-dimensional real vector, k-means clustering partitions the N observations into $K(\leq N)$ sets $\mathcal{X}_1, \cdots, \mathcal{X}_K$ such that it minimizes the within-cluster sum of squares.
- The objective is to find:

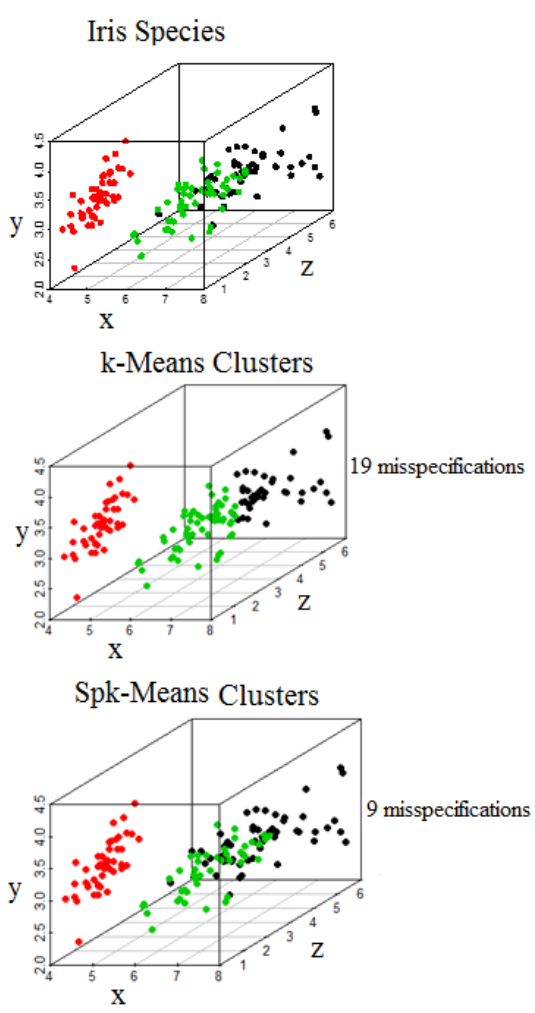
$$\arg\min_{\mathcal{X}} \sum_{i=1}^{K} \sum_{\mathbf{x} \in \mathcal{X}_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

where $\boldsymbol{\mu}_i$ is the mean of points in $\mathcal{X}_i$.

Spherical K-Means clustering (Dhillon & Modha, MR 2001)

- Spherical K-means uses the cosine similarity $\mathbf{x}.\boldsymbol{\mu} = \mathbf{x}^T\boldsymbol{\mu}$ (which is equal to cosine of the angle between $\mathbf{x}$ and $\boldsymbol{\mu}$) instead of Euclidean distance $\|\mathbf{x} - \boldsymbol{\mu}\|^2$, as a measure of dissimilarity.

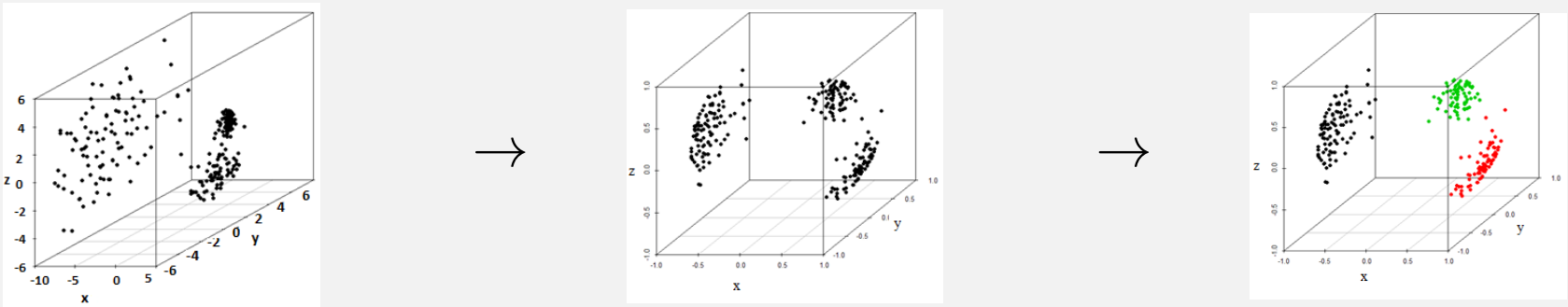
- Spherical K-means algorithm is preferred to standard K-means for clustering of the document vectors or any type of high-dimensional observations on the unit sphere.



Iris flower data set

Spherical Data

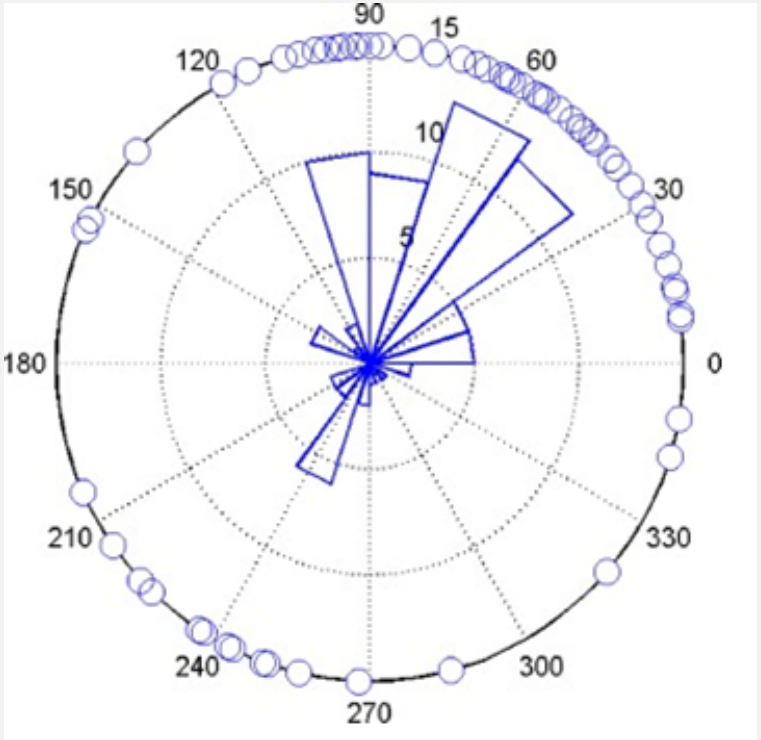
- As the number of categorical variables increases the sparsity and the dimension of the data will increase.
- Projecting the data vectors on to the high-dimensional sphere and using clustering on the sphere has shown superior performance for high dimensional and sparse data set.



- This kind of data are called spherical or directional data.

Probabilistic Model-Based Clustering

Assume that the data comes from a mixture of several probabilities.
$f(\mathbf{x}|\Theta) = \sum_{j=1}^{K} \alpha_j f_j(\mathbf{x}|\boldsymbol{\theta}_j)$, where K is the number of clusters, α_j's are the mixture proportions that are non-negative and sum to one, and $\Theta = (\alpha_1, \cdots, \alpha_M, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$.



Direction of movements of turtles

Poisson-Kernel Based Clustering

- A d-dimensional unit random vector $\mathbf{x}$ is said to have a *d-variate Poisson kernel-based distribution* on S^{d-1} if its density is given by

$$f(\mathbf{x}|\rho,\boldsymbol{\mu}) = \frac{1-\rho^2}{\omega_d \left\|\mathbf{x}-\rho\boldsymbol{\mu}\right\|^d}, \tag{1}$$

  where $\left\|\boldsymbol{\mu}\right\| = 1$, $0 < \rho < 1$ and $\omega_d = \frac{2\pi^{d/2}}{\Gamma(d/2)}$, is the surface area of the unit sphere in $\mathbb{R}^d$.

- Poisson kernel-based clustering assumes that the data come from a mixture of several Poisson kernel-based distributions.

$$f(\mathbf{x}|\Theta) = \sum_{j=1}^{K} \alpha_j f(\mathbf{x}|\rho_j,\boldsymbol{\mu}_j) = \sum_{j=1}^{K} \frac{\alpha_j(1-\rho_j^2)}{\omega_d \left\|\mathbf{x}-\rho_j\boldsymbol{\mu}_j\right\|^d}, \tag{2}$$

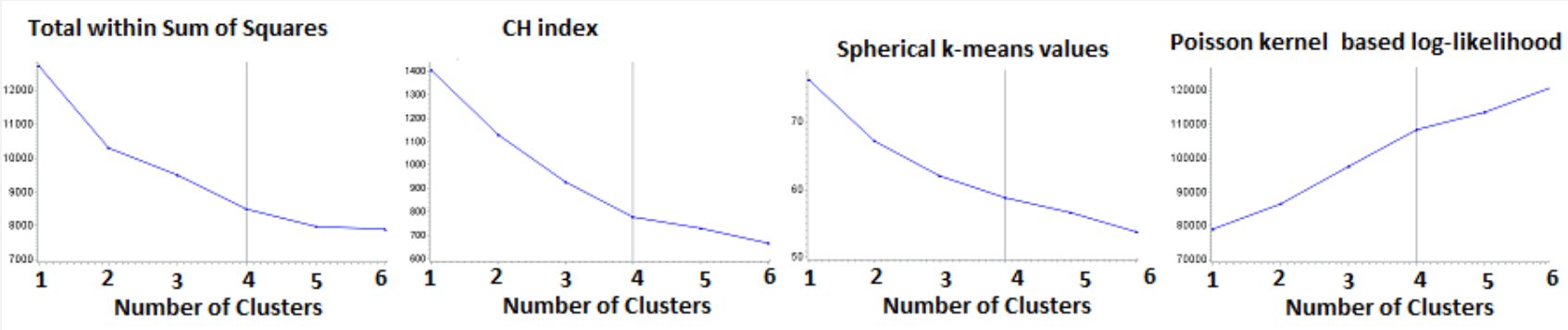
  where K is the number of clusters, and α_j's are the mixture proportions.

Implementation of the Methods

- The functions kmeans and skmeans in R software was used to perform standard and spherical k-means clustering, respectively.
- A modified EM algorithm presented in Golzy & Markatou (2016, 2019) was used to perform Poisson kernel-based clustering.
- We included all patients with both completed baseline and 2 years follow up surveys in the same cohort (N=4721).
- The selected features for the clustering are: Baseline ADL measures (difficulty bathing, dressing, eating, getting in and out of chair, walking, using toilet) and comorbidities (CVD, Muscular disease, any other cancer, COPD, GI, Stroke, Urinary problem and Diabetes).

Optimal Number of Clusters

- The Calinski-Harabasz index, total within sums of square of k-means, Spherical kmeans values and the log-likelihood of Poisson kernel-based was used to determine the optimal number of clusters.



- The clustering experiments suggest the present of four coherent clusters of BC patients.

Identification of Clusters

Table 1: Percentage of ADL difficulties in each class for each method

		Standard K-means				Spherical K-means				Poisson Kernel-Based			
	Clusters →	1	2	3	4	1	2	3	4	1	2	3	4
Activity	Level ↓ \N →	1740	1242	778	787	1230	1169	1289	859	173	168	2497	1709
Bathing	I am unable	0%	0%	0%	9%	0%	0%	0%	8%	0%	0%	0%	4%
	have difficulty	1%	1%	10%	42%	1%	3%	5%	39%	0%	0%	1%	24%
Dressing	I am unable	0%	0%	0%	6%	0%	0%	0%	5%	0%	0%	0%	3%
	have difficulty	1%	1%	3%	41%	0%	2%	4%	33%	0%	0%	1%	20%
Eating	I am unable	0%	0%	0%	4%	0%	0%	0%	3%	0%	0%	0%	2%
	have difficulty	1%	1%	2%	14%	0%	2%	1%	12%	0%	0%	1%	7%
Getting	I am unable	0%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	2%
off chair	have difficulty	6%	5%	10%	90%	4%	7%	22%	62%	0%	0%	6%	47%
Walking	I am unable	0%	0%	0%	10%	0%	0%	1%	7%	0%	0%	0%	4%
	have difficulty	11%	14%	41%	88%	6%	16%	31%	85%	0%	0%	9%	68%
Using	I am unable	0%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	2%
Toilet	have difficulty	1%	1%	4%	31%	0%	2%	3%	26%	0%	0%	1%	16%
General	Excellent	8%	5%	0%	0%	17%	0%	0%	0%	29%	0%	6%	0%
Health	very good	38%	32%	0%	1%	83%	0%	3%	0%	71%	0%	38%	0%
	Good	54%	63%	0%	38%	0%	79%	81%	6%	0%	100%	56%	26%
	Fair	0%	0%	86%	47%	0%	21%	16%	69%	0%	0%	0%	61%
	Poor	0%	0%	14%	13%	0%	0%	0%	25%	0%	0%	0%	12%

Table 2: Percentages of comorbidities in each class for each method

	Standard K-means				Spherical K-means				Poisson Kernel-Based			
Clusters →	1	2	3	4	1	2	3	4	1	2	3	4
Urination Problem	4%	6%	10%	19%	3%	8%	7%	18%	0%	0%	6%	14%
Stroke	10%	13%	28%	22%	9%	11%	18%	30%	0%	0%	13%	24%
COPD	11%	10%	16%	25%	9%	11%	15%	23%	0%	0%	11%	20%
GI	2%	4%	6%	7%	2%	2%	5%	9%	0%	0%	3%	6%
Diabetes	17%	17%	27%	30%	11%	21%	24%	32%	0%	0%	19%	29%
CVD	61%	69%	81%	80%	55%	69%	75%	84%	0%	100%	65%	82%
Muscular Disease	42%	48%	56%	85%	37%	0%	100%	80%	0%	0%	48%	72%
Any other caner	0%	100%	50%	45%	36%	43%	48%	49%	0%	0%	47%	48%

- Possion kernel-based clustering (PKB) provides more homogenous subgroups with respect to ADL measures and the comorbidities than the other two methods.
- Poisson kernel based distribution clustering is more robust with respect to noisy data points than other clustering algorithms.

Characterization of Clusters

Considering the Poisson kernel-based clustering (PKB):

- The individuals in the first class have no difficulty in any ADL measures and no comorbidities.
- The individuals in the second class have no difficulty in any ADL measures and all have CVD comorbidity.
- The individuals in the third class have some comorbidities. More specificity 65% have CVD and 48% have muscular diseases.
- The individuals in class 4 have worst outcome with respect to ADL measures and comorbidities.

Characterization of Clusters (Cont.)

Table 3: Mean and standard deviation of the continuous variables by cluster

	Poisson Kernel-Based clusters				
	1	**2**	**3**	**4**	**P-value**
Age	78.9 (6.6)	77.9 (7.1)	78.5 (6.8)	78.4 (6.8)	0.730
BMI	26.1 (4.9)	27.2 (4.7)	27 (4.9)	26.8 (4.8	0.270
PCS12	37.4 (10.5)	35.6 (11.4)	38.1 (11.8)	38.2 (11.7)	0.770
MCS12	51.1 (10.3)	50.9 (11.1)	51.3 (10.6)	51.3 (10.8)	0.950

- There was no statistically significant difference in Age, BMI, PCS12 and MCS12 among clusters.

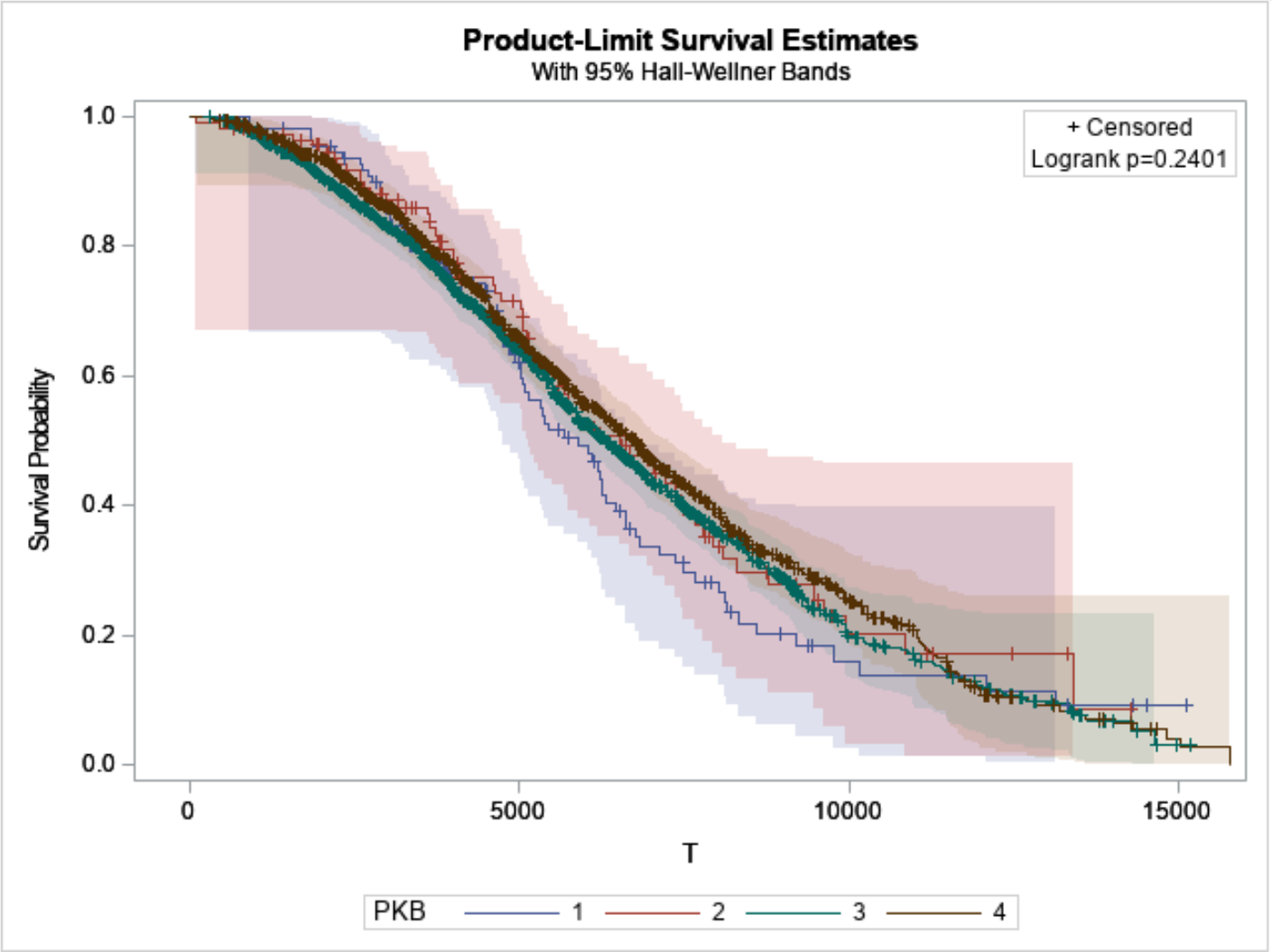
Characterization of Clusters (Cont.)

Table 4: Percentages of the categorical variables by cluster

		Poisson Kernel-Based clusters				
		1	**2**	**3**	**4**	**P-value**
Outcomes	Fall rate	12.8%	12.3%	15.5%	16.5%	0.037
	Mortality rate	68.8%	56.1%	57.4%	56.4%	0.094
	Dead in 6 months	2.8%	1.8%	3.3%	2.8%	0.740
	Dead in 12 months	6.4%	2.6%	7.4%	7.2%	0.287
Treatment	Big surgery	7.3%	10.5%	5.4%	6.4%	0.005
	Small surgery	23.9%	21.9%	19.4%	21.2%	
	No surgery	68.8%	67.5%	75.2%	72.5%	

Note: Fall information was collected from the follow-up surveys.

Survival Curves



- Group 1 had a shorter survival time but not significantly.

Stages of Cancer

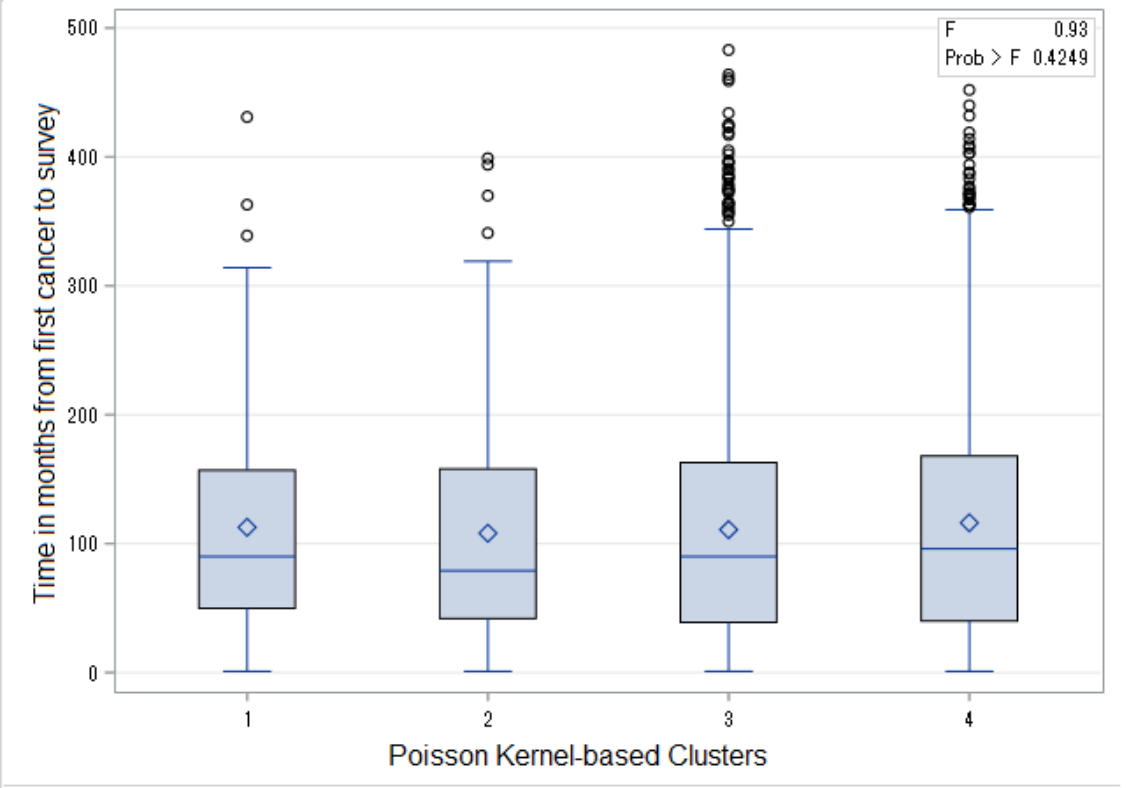
Table 5: Distribution of the Stages of BC cancer

	1	2	3	4	P-value
Stage 0	17.91	20.83	22.45	23.29	0.0027
Stage 1	64.18	55.56	59.41	59.41	
Stage 2	5.97	19.44	9.26	9.9	
Stage 3	5.97	4.17	5.64	4.46	
Stage 4	5.97	0	3.25	2.93	

- Group 1 had a higher rate of stage 3 and 4 of BC cancer patients.

Time from First Cancer to Survey

Figure 3: Box plot of the Time from first cancer to survey in months



- There was no significant difference in time from first cancer to survey among clusters (p-value=0.42)

Analysis of the Reduction in ADL Outcomes

Table 6: Distribution of reduction of ADL measures in each PKBD class

		Poisson Kernel-Based			
		1	2	3	4
Reduction in ADL activities	Bathing	0%	0%	0%	27%
Pre vs post BC	dressing	0%	0%	0%	21%
(N=475)	eating	0%	0%	0%	9%
	getting up chair	0%	0%	3%	34%
	Walking	0%	0%	5%	44%
	Using Toilet	0%	0%	1%	17%
Reduction in ADL activities	Bathing	0%	0%	1%	23%
More advanced BC patients	dressing	0%	0%	1%	18%
(N=1718)	eating	0%	0%	1%	8%
	getting up chair	0%	0%	4%	32%
	Walking	0%	0%	5%	41%
	Using Toilet	0%	0%	1%	16%

- Group 4 had the highest rate of reduction of ADL measures in two months interval.

Predictors of the Reduction in ADL Outcomes

Table 7: Odds Ratio and the corresponding 95% confidence limit for significant predictors of reduction in ADL from the logistic regression models

Outcome=Reduction in ADL	Effect	Odds Ratio	95% CL for Odds Ratio		P-value
Bathing	COPD Yes vs No	1.34	1.04	1.72	0.0224
	Stroke Yes vs No	2.21	1.67	2.93	<.0001
	Urinary problem Yes vs No	1.45	1.12	1.89	0.0053
Dressing	CVD Yes vs No	0.72	0.54	0.96	0.0267
	Muscular disease Yes vs No	1.59	1.20	2.10	0.001
	GI Yes vs No	1.64	1.06	2.52	0.0246
	Stroke Yes vs No	1.98	1.46	2.70	<.0001
	Urinary problem Yes vs No	1.55	1.18	2.04	0.0017
Eating	CVD Yes vs No	0.52	0.36	0.75	0.0004
	GI Yes vs No	1.98	1.15	3.41	0.0132
	Stroke Yes vs No	1.55	1.00	2.40	0.0494
Getting up chair	Muscular disease Yes vs No	1.85	1.51	2.27	<.0001
	Stroke Yes vs No	1.76	1.36	2.28	<.0001
	Urinary problem Yes vs No	1.36	1.09	1.71	0.0066
Walking	No surgery vs Big surgery	1.28	0.83	1.98	0.2603
	No surgery vs small surgery	1.34	1.04	1.71	0.0226
	Muscular disease Yes vs No	1.52	1.26	1.82	<.0001
	Stroke Yes vs No	2.03	1.58	2.61	<.0001
Using Toilet	Muscular disease Yes vs No	1.39	1.04	1.86	0.0247
	GI Yes vs No	1.92	1.24	2.97	0.0036
	Stroke Yes vs No	1.89	1.37	2.61	0.0001
	Urinary problem Yes vs No	1.79	1.35	2.38	<.0001

Conclusions

In our cohort:

- COPD was associated with reduction in bathing activity.
- Stroke was associated with reduction in all ADL activities.
- Gastrointestinal disorders (GI) were associated with reduction in dressing and using toilet activities.
- Muscular diseases were associated with reduction in all dressing, getting up the chair, walking and using toilet activities.
- Urinary issue was associated with reduction in bathing, dressing, getting up the chair and using toilet activities.
- Diabetes and any other cancer were not associated with reduction in any ADL.

Conclusions (Cont.)

- Cluster 1 with no ADL dificulties and no comorbidities had significantly higher rate of stage 3 and 4 of BC cancer and non-significant higher mortality rate.
- They had significantly higher rate of surgeries.
- They had no reduction in ADL measures in a two years interval.

We conclude that, the stage of cancer and the type of surgery are stronger predictors of mortality outcome than ADL measures.

Conclusions (Cont.)

- The individuals in the cluster 4 had the worst ADL measures and comorbidity rate;
- have significantly higher rate of Fall outcome;
- They had highest rate of reduction in ADL measures in a two years interval.

We conclude that, the baseline ADL measure and comorbidities are predictors for Fall outcome and reduction in ADL in two years in BC patients.

Contact Information

- Mojgan Golzy, PhD
  Biostatistics and Research design unit
  Department of Health Management and Informatics
  University of Missouri-Columbia
  School of Medicine
  golzym@health.missouri.edu

- Katie Murray, MO
  Department of Surgery-Urology Division
  University of Missouri-Columbia
  School of Medicine
  murraykat@health.missouri.edu


Thank you!