KMDI
REPORTS

KNOWLEDGE
MEDIA
DESIGN
INSTITUTE

Predicting Emergency Department Visits Based on Cancer Patient Types

Authors:
Mahsa Rouzbahman, Mark Chignell
&Lisa Barbera

KMD-2016-2

Predicting Emergency Department Visits Based on Cancer Patient Types

Mahsa Rouzbahman, Mark Chignell**,Lisa Barbera***,*

**PhD Candidate, Department of Mechanical and Industrial Engineering, University of Toronto. **Professor, Department of Mechanical and Industrial Engineering, University of Toronto. ***MD, MPA, Associate Professor, Department of Radiation Oncology, University of Toronto, Institute for Clinical Evaluative Sciences, Toronto, Canada.*

Abstract

This study evaluated the predictive ability of patient types (clusters of similar patients) in identifying cancer patients at high risk for ED visits within one year (365 days) following their index date. A descriptive and retrospective cohort study of 254,552 unique cancer patients with only one primary cancer type was done using linked administrative sources of health care data. Three outcomes were investigated in this study. First, the time of ED visit following an index date was predicted using multiple linear regression. Second, those patients who visited an ED within seven days of their index date were detected using logistic regression. In addition to predicting emergency department visit, vital status of patients was also predicted using logistic regression. We implemented the linear/logistic regression once on unclustered raw data. Then cluster analysis was done before the prediction step and the results of these two analyses were compared to each other. Clustering was found to contribute to a modest improvement in prediction accuracy for all three outcome variables. In addition, for the first outcome variable, a privacy preserving analysis was carried out using summarized clustered data (mean, standard deviation and correlation). The results are discussed in terms of the ability of summarized patient type data to provide clinical support tools while also respecting the privacy of patients.

Keywords: Cancer, Emergency Visit, Mortality, Clustering, Regression Analysis, Prediction, ICES.

1 Introduction

Since physicians often use case based reasoning, clinical decision support tools based on patients who are similar to the current patient (Chan, Chan, Cheng, & Mak, 2010) might be useful both in supporting differential diagnosis, and in making predictions of outcomes, such as the probability of emergency department visits. Visits to the emergency department (ED) can be long and uncomfortable for patients in general, but especially for those with cancer. ED visits made near the end of life may be indicative of poor quality in cancer care. Barbera et al. (2013) demonstrated that worsening symptoms can contribute to ED visits (Barbera, et al., 2013). The most common reasons for cancer patients to visit the ED during the final six months of life and the final two weeks of life include abdominal pain, lung cancer, dyspnea, pneumonia, malaise and fatigue, and pleural effusion (Barbera, Taylor, & Dudgeon, 2010). Since ED visits occur quite frequently near the end of life, better and more comprehensive evaluations of the reasons for such visits are needed as a step towards development of interventions for avoiding unnecessary ED visits (Barbera, Taylor, & Dudgeon, 2010) (Meldon, et al., 2003).

Identifying patients with a high probability of hospital admission or subsequent return to the emergency department (ED) might facilitate the development of interventions to improve the admission process and patient care, and decrease overcrowding in hospitals (LaMantia, et al., 2010). Alternatively, it could inform the development of additional ambulatory resources to mitigate the risk of an ED visit.

The main goal of this study is to validate the predictive ability of patient types (clusters of similar patients) in predicting useful outcomes about cancer patients while preserving patient data confidentiality. In particular, next emergency department visit and vital status of patients were chosen as outcomes of interest based on the need for better planning of cancer patients. This study seeks to replicate previous findings that clusters of similar patients can contribute to more accurate predictions without using individual medical records and violating the privacy of patients (Rouzbahman, Jovicic, & Chignell, 2016).

2 Methodology

A retrospective study was performed using administrative sources of Ontario health care data. We predicted when patients with cancer diagnosis might visit an Emergency Department after their index date (which is the date of a cancer clinic visit where standardized symptom screening is also completed and which is followed by an ED visit). We also predicted mortality (vital status) of patients with cancer diagnosis based on their last symptom scores.

2.1 Sources of Data

About 95% of all incident cases of cancer in Ontario are extractable from the Ontario Cancer Registry which is a comprehensive, population based registry (Clarke, Marrett, & Kreiger, 1991) (Robles, Marrett, Clarke, & Risch, 1988). In this study, 9 different datasets were used, which are listed in table 1.

Table 1. List of Datasets Used in the Analysis

Name of the Dataset	Description of Dataset
ESAS	Edmonton Symptom Assessment System
CIHI	Canadian Institute for Health Information
OHIP	Ontario Health Insurance Plan
NACRS	National Ambulatory Care Reporting System
ODB	Ontario Drug Benefit
NDFP	New Drug Funding Program
HCD	Home Care Database
OCR	Ontario Cancer Registry
RPDB	Registered Persons Database

All visits to the emergency department are captured in The National Ambulatory Care Reporting System. All residents of Ontario who are eligible for the Ontario Health Insurance Plan are included in The Registered Persons Database and we can find the residents' demographics from this database (Iron, Zagorski, & Sykora, 2008). For all the patient variables we linked the datasets mentioned above to the cohort dataset using a common unique identifier.

2.2 Inclusion criteria

We created a cohort dataset containing 254,552 unique adult (> 17 years old) cancer patients whose diagnosis dates were between Jan 1, 2007 and Mar 31, 2011, and who were diagnosed with only one primary cancer type (patients with multiple primary cancers were excluded). We considered Jan 1, 2007 as start date for inclusion and subjects were followed from Jan 2007 to the end of study or death, whichever came first. The observation window terminated at Dec 31, 2013. Index date is the date of a cancer clinic visit where standardized symptom screening is completed and which is also followed by an ED visit.

2.3 Outcome

We were interested in predicting the next ED visit after the index date for cancer patients. We created this outcome variable with two forms: a continuous variable (time between index date and ED date) and a binary variable (visit an ED within 7 days after the index date or not). Date of registration in the emergency department was used to create the outcome variable from the NACRS dataset. The outcome variables were defined as follows:

Outcome 1 (ED time): Time between index date and ED date (365/90)

We started with each index date and looked forward to find the closest ED visit. If there were multiple dates associated with a unique ED, we considered the closest date to the ED as the index date. Then we calculated the time between index date and ED visit (Figure 1). Each patient might have multiple index and ED dates and as a result multiple records for the outcome variable (Figure 1-A). Table 2 shows the number of patients versus number of records for the outcome variable. The majority of patients (around 70%) had only 1 record for this outcome variable. As a result, we chose to focus on patients who had only a single ED visit, and patients with multiple records for this criterion were excluded from the study (Figure 1-B). Figure 1 shows how we created the first outcome variable. Figures 2 and 3 show the histogram and box plot of the first outcome variable. We limited the maximum time between index date and subsequent ED visit to 1 year, or to 3 months, using those two time windows separately. This was done because it is harder to attribute association over long periods where there is more opportunity for the outcome variable to be influenced by extraneous variables. Thus the two outcomes used in the study were:

ED time 365: Time between index date and the next ED visit, only 1 record, maximum time being 365 days

ED time 90: Time between index date and the next ED visit, only 1 record, maximum time being 90 days

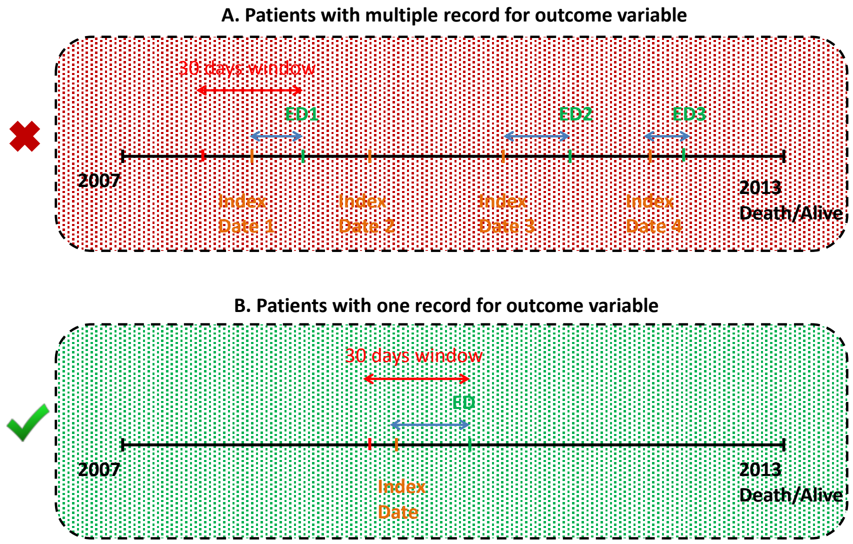


Figure 1: Creating ED visit variable

Table 2. Number of patients versus number of ED visits

Number of ED visits after an assessment date	Number of patients	% of patients
1	31573	69.61
2	9020	19.88
3	2965	6.53
4	1129	2.48
5	381	0.84
6	156	0.34
7	68	0.14
8	36	0.07
9	12	0.02
>= 10	16	0.03

Figure 2. Frequency histogram of time between index date and ED date

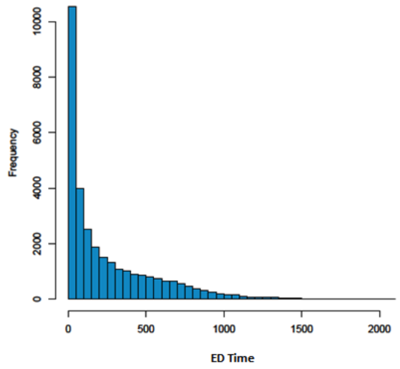
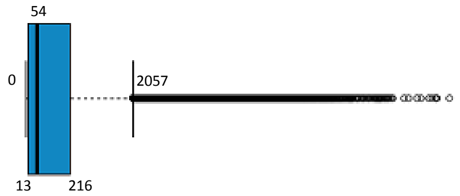


Figure 3. Box plot of time between index date and ED visit



Outcome 2 (ED within 7d): visit an ED within 7 days after the index date or not

We determined if patients had an ED visit within 7 days of the index date. The same methodology was followed and only patients with one record and a maximum time of 1 year between ED and index dates were considered.

Outcome 3: Vital Status

We also considered vital status of patients as a third outcome variable in this study. This outcome variable reported vital status of patients as of December 31, 2013.

2.4 Patient Variables

The dataset constructed for this study contained baseline characteristics such as cancer type (Breast, Central Nervous System, Gastrointestinal, Genitourinary, Gynecologic, Hematology, Head and Neck, Lung, Other, Primary Unknown, Sarcoma and Skin), vital status, sex, cancer stage (stage: 0, stage: I, stage: II, stage: III, stage: IV), income quintile, age at diagnosis and Charlson comorbidity, rural/urban and geographical region (in this case the LHIN, or local health integration network within the Province of Ontario).

Charlson comorbidity is a method of classifying comorbidities based on International Classification of Diseases (ICD) diagnosis codes of patients. A weight is assigned to each comorbidity category and a single comorbidity score is calculated based on the sum of the weights for each patient. The higher the score, the more probability the predicted outcome will contribute to a higher resource use or mortality (Deyo, Cherkin, & Ciol, 1992).

We used data from the Registered Persons Database to determine vital status of patients. Sex of patients, and cancer stage, were recorded from the Ontario Cancer Registry. For type of cancer, we used the International Classification of Diseases (ICD-10) codes of the Ontario Cancer Registry, which were grouped and coded as follows: head and neck 140–149, 160, 161; breast 174; lung 162; prostate 185; ovary 183; colorectal 153, 154; central nervous system 191; lymphoma or leukemia 200–208; other genitourinary or gynecological 179–182, 184, 186–189; melanoma or sarcoma 170–172; non-melanoma skin cancer 173; other gastrointestinal 150–152, 155–159; metastases 196–199; other 163–165, 190, 192–194.

The Edmonton Symptom Assessment System (ESAS) is one of the key assessment tools used to help in the assessment of nine common symptoms experienced by cancer patients. The descriptors for those symptoms are Tired, Nausea, Depression, Anxious, Drowsy, Pain, Appetite, Wellbeing and ShortBreath (Bruera, Kuehn, Miller, Selmser, & Macmillan, 1991). The ESAS dataset contained symptom scores, date of assessment and a unique identifier, which was used to link data from the disparate datasets into the master data sheet used for the analyses reported in this study.

In addition to using the variables described above, we also derived additional variables.

A margin of 30 days before the ED visit date was defined as the time window of interest. We then created all the services/events for patients in that time window (figure 4).

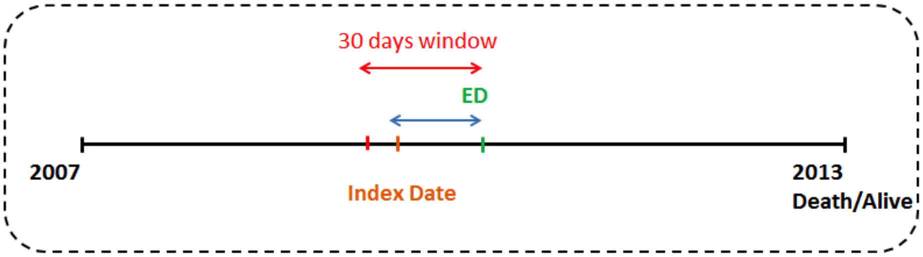


Figure 4. Creating predictors based on one month time window before ED date

The CIHI dataset contains detailed demographic, administrative and clinical data for both hospital admissions and day surgeries. Admission date, discharge date, length of stay, diagnosis and intervention codes are some of the important variables among others in this dataset. From the CIHI dataset, we considered those patients who were admitted/discharged to/from the hospital within 30 days prior to the ED visit. This information was captured as a binary variable that showed hospital admission for patients (1=admitted/discharged, 0=else). To calculate total number of admitted days, we considered admission and discharge dates for each patient. If both “admission date” and “discharge date” occurred prior to the margin of 30 days (i.e., prior to the time window start date), the number of admitted days for that patient was defined as 0. If both “admission date” and “discharge date” occurred within 30 days prior to the ED visit, the number of admitted days for that patient was defined as the LOS (length of stay) in hospital. If only “discharge date” was within 30 days, and the “admission date” was outside this time window, the number of admitted days was defined to be equal to (discharge date – time window start date). Each patient might have multiple hospital visits. To calculate the total number of admitted days (inpatient), we calculated the sum of all admitted days for all hospital visits and kept the diagnosis associated with the longest stay.

Diagnosis codes (ICD 10) were available from the inpatient, same day surgery and emergency department datasets. Among multiple diagnosis codes associated to each hospital admission/ED visit, only the primary diagnosis was considered for each patient. The first 3-digits of each diagnosis code were considered, which reduced the number of codes under consideration from 1937 to 675. We then carried out a more general categorization of the ICD10 codes to reduce the number of diagnostic categories further. 24 categories of ICD10 codes were identified as being relevant to our study, as shown in Table 3:

Table 3: Grouping ICD 10 codes

ICD10 Code	Description	ICD10 Code	Description
800-998	Morphology	L00-L99	Diseases of the skin and subcutaneous tissue
A00-B99	Certain infectious and parasitic diseases	M00-M99	Diseases of the musculoskeletal system and connective tissue
C00-D48	Neoplasms	N00-N99	Diseases of the genitourinary system
D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	O00-O99	Pregnancy, childbirth and the puerperium
E00-E90	Endocrine, nutritional and metabolic diseases	P00-P96	Certain conditions originating in the perinatal period
F00-F99	Mental and behavioural disorders	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
G00-G99	Diseases of the nervous system	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
H00-H59	Diseases of the eye and adnexa	S00-T98	Injury, poisoning and certain other consequences of external causes
H60-H95	Diseases of the ear and mastoid process	U00-U49	Provisional codes for temporary assignment by WHO of new diseases of uncertain etiology
I00-I99	Diseases of the circulatory system	U50-U99	Provisional codes for research and temporary assignments in Canada
J00-J99	Diseases of the respiratory system	V01-Y98	External causes of morbidity and mortality
K00-K93	Diseases of the digestive system	Z00-Z99	Factors influencing health status and contact with health services

Triage level, diagnosis and intervention codes and length of stay are some of the other variables in NACRS dataset that were used in the analysis.

In the OHIP dataset, we calculated total number of OHIP codes that each patient had in the one month time window prior to the ED visit. To reduce the number of OHIP fee codes, we created 4 main categories which are: house calls codes, palliative care codes, A codes (assessment/consultation codes - office visits) and W codes (long term care visits).

Ontario's Community Care Access Centres (CCACs) were established by the Ministry of Health and Long-Term Care to arrange government-funded home and community services and long-term care homes. The HCD dataset contains all services provided by CCACs (Institute for Clinical Evaluative Sciences (ICES): Home Care Database (HCD) [Canada] [administrative database], 2011). Using service date, we calculated total number of services that patients received in the 30 days preceding the ED visit. We also calculated total number of services before the start of the time window under consideration. There were 19 different categories for type of service and we grouped them into three main categories of “nursing visit”, “personal support worker (PSW) visit” and “other”. Table 4 shows how we created the three main categories out of 19 different types of services.

Table 4: Creating home care variables

Service	Nursing visit	Personal visit	Other
1 = NURSING - VISIT	×		
2 = NURSING - SHIFT (HOUR)	×		
3 = RESPIRATORY SERVICES			×
4 = NUTRITION/DIETETIC			×
5 = PHYSIOTHERAPY			×
6 = OCCUPATIONAL THERAPY			×
7 = SPEECH LANGUAGE THERAPY			×
8 = SOCIAL WORK			×
9 = PSYCHOLOGY			×
10 = CASE MANAGEMENT			×
11 = PERSONAL SERVICES (HOUR)		×	
12 = HOMEMAKING SERVICES (HOUR)		×	
13 = COMBINED PS AND HM SERVICES (HOUR)		×	
14 = PLACEMENT SERVICES			×
15 = RESPITE			×
99 = OTHER			×

To provide equal access to high-quality intravenous (IV) cancer drugs, the New Drug Funding Program (NDFP) was created to cover the cost of many newer, and expensive, injectable cancer drugs prescribed in hospitals/cancer centres (New Drug Funding Program Database (NDFPDB), 1995) (New Drug Funding Program (NDFP) & Evidence Building Program (EBP), Approved Drugs and Eligibility Criteria, 2015). The NDFP dataset captures all the newer chemotherapy drugs and treatment dates for patients with a cancer diagnosis. Using treatment date in the NDFP dataset, two sets of treatment-related variables were created. The first set counted all treatments prescribed for patients in the 30 days before the ED visit, while the second set described all the treatments before the beginning of the time window. Both sets included approximately 30 different categories.

The ODB database captures prescription claims data in the Claims History Database for patients older than 65. Drug identification number (DIN), service date and estimated number of days supplied are recorded in this dataset (Ontario Drug Benefit (ODB), 1992). In the ODB dataset, based on the service date variable, a margin of six months before ED visit was considered and we counted all the drugs prescribed for patients in 6 months before the ED visit. We also determined the drugs before the margin date. The ODB dataset contains a total of 865 different drugs. We then classified the drugs and grouped them into a final set of 42 groups based on their medical properties.

2.5 Analysis

To replace missing values, we used the median value for symptom scores such as tired score (155 missing records) and for demographic variables such as income quantile (57 missing records), rural/urban (5 missing records) etc. To take into the account the differences in variable scales, we standardized some predictor variables for parts of the analysis including k-means cluster analysis and nearest neighbors matching (in order to make the Euclidean distances, used in these analyses, insensitive to differences in scale between the variables). The final dataset for adult patients was randomly split into two sets, a training set containing 80% of the cases, and a test set containing the remaining 20% of the cases.

K-means clustering was carried out on the patients in the training data set, varying the number of clusters between 2 and 50. The Hartigan-Wong algorithm was applied in the analysis (Morissette & Chartier, 2013). After each clustering of the training set, regression models were fitted for each cluster within the clustering solution. Since the first outcome was a continuous variable (ED time), we used linear regression in its analysis. However, for the second and third outcome variables (ED within 7d and vital status) logistic regression was used since the criterion in those cases was dichotomous. Each test case was then matched to the closest training cluster based on minimum Euclidean distance between each case in the test set and the centroid of each cluster (Nearest neighbour matching) and values of the outcome variable were then predicted based on the corresponding regression equation.

In order to smooth out the effects of random variations in the clustering (due to different initial clusters selected randomly during each run of the k-means process) and regression processes, the complete procedure was run 50 times each for 49 partition sizes (k=2 to 50) and the mean values of the outcome variables of the 50 clustering solutions for each partition size were calculated. In order to evaluate the prediction process, for "ED time", mean absolute error (MAE) was calculated, whereas for "ED within 7d" and vital status accuracy measures for a dichotomous criterion, such as sensitivity, specificity, accuracy, AUC (area under the curve) and d-prime were considered. In order to simulate situations that guard against re-identification, in the analysis of ED time we utilized clusters that had at least 100 patients in them, for each solution. The cluster solution with the minimum mean absolute error (MAE) for the first outcome and the maximum accuracy measures for the second and third outcomes were determined.

The accuracy measures were defined as follows (Sayad, 2010):

- Accuracy: the proportion of the total number of predictions that were correct (equation 2).
- Sensitivity or Recall: the proportion of actual positive cases which are correctly identified (hits).
- Specificity: the proportion of actual negative cases which are correctly identified (correct rejections).
- D-prime: is the standardized difference between the means of the Signal Present and Signal Absent distributions (i.e., z[hits] – z[false alarms] as shown in equation 3).

Equation 1 shows the error measure used in the linear regression analyses. Equations 2 and 3 show the accuracy measures used in the logistic regressions.

$$(1)\ MAE = \frac{\sum_{i=1}^{n} |p_i - a_i|}{n}$$

$$(2)\ Accuracy = \frac{(true\ positives + true\ negatives)}{(total\ population)}$$

$$(3)\ d - prime = \frac{(\mu_s - \mu_N)}{\sqrt{\frac{1}{2}({\sigma^2}_S + {\sigma^2}_N)}} = Z(hit\ rate) - Z(false\ alarm)$$

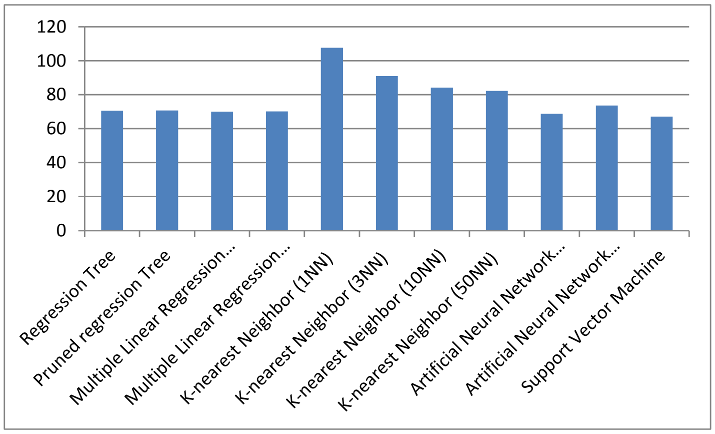
3 Results

3.1 First outcome: time between index date and ED date (ED time)

3.1.1 Comparison of different prediction methods (no cluster analysis)

We first present the results of prediction of time between index date and ED date (ED Time) on the unclustered, raw data for the test sets using different prediction methods, to allow comparison with the analysis using clustered data. Different prediction methods were compared with each other using mean absolute error (MAE). Multiple linear regression, artificial neural network (size=1) and support vector machines produced the lowest error values. However, when data are summarized (to preserve confidentiality) as means, standard deviations, and correlations, linear regression analysis is feasible, but not the other two methods. As a result, linear regression was utilized in the remainder of the analysis to simulate a privacy preserving analysis that utilized summarized data only. Figure 5 shows the error measure MAE for different prediction methods. It can be seen that regression analysis did well compare to the other methods, demonstrating that it is a competitive method for this data set._

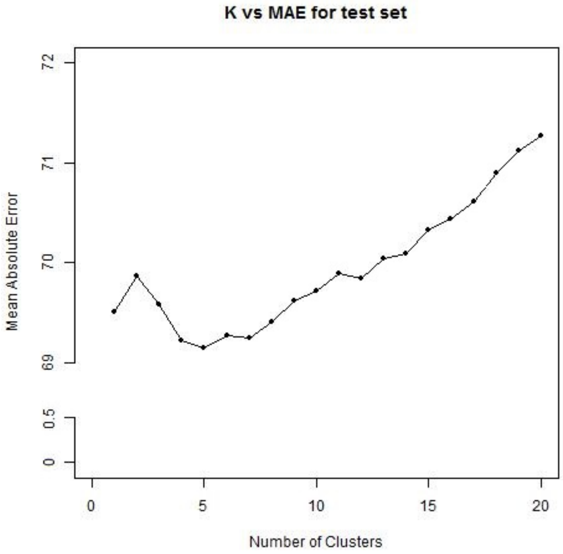
Figure 5: Mean absolute error (MAE) for different prediction methods



3.1.2 All variables in cluster/regression analysis

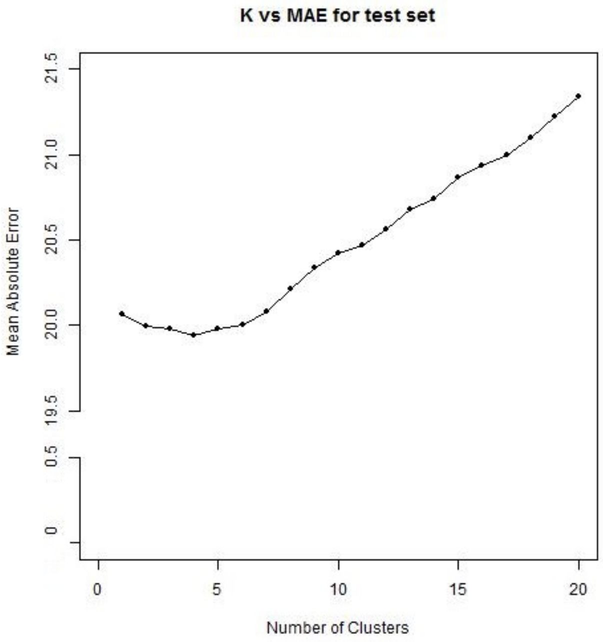
Using cluster analysis and linear regression, we predicted the time between index date and ED date (ED Time) using all the variables. First, we limited the outcome variable to a maximum of 1 year (365 days). The parameter k, representing Number of clusters, was changed from 2 to 20 (k=2 to 20) and the whole process was repeated 50 times (N=50). The five cluster solution (k=5) was detected as the best solution with a minimum mean absolute error of 69.14. Compared to the condition where cluster analysis wasn't used before regression analysis (MAE=69.508), clustering with k=5 showed a 0.36 day improvement in terms of mean absolute error (MAE). Figure 6 shows the relationship between the number of partitions and the resulting MAE for ED time 365.

Figure 6. Relationship between number of partitions and MAE for the test dataset (ED time 365)



We then limited the outcome variable to a maximum of three months (90 days). Since a relatively small number of clusters was found to be optimal in the earlier analysis we also limited the range of clusters consider (k=2 to 20) and the whole process was repeated 50 times (N=50). A four cluster solution (k=4) was detected as the best solution with a minimum mean absolute error of 19.94 days. Comparing to the condition when we don't apply cluster analysis before regression analysis, the four cluster solution showed a 0.12 day (around 2.5 hours) improvement in terms of amount of error. Figure 7 shows the relationship between the number of partitions and the resulting MAE for ED time 90.

Figure 7. Relationship between number of partitions and MAE for the test dataset (ED time 90)

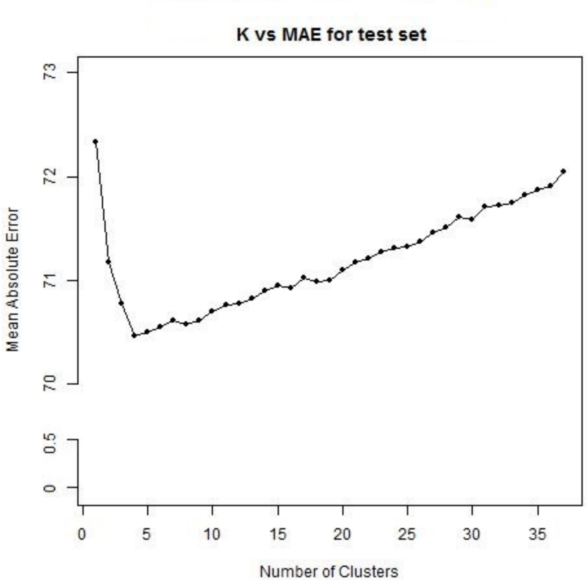


3.1.3 Correlated predictor variables in both cluster and regression analysis (correlation >= 0.1)

The predictive relationships in the data as assessed by correlations of predictor variables with the outcome were generally low. In order to reduce the chances of overfitting due to error variance we then limited the predictor variables considered in the clustering and in the regression models to those with a correlation of at least 0.1 with the outcome. The bivariate correlation matrices between the outcome variable and all the predictor variables were calculated in each cluster and the predictor variables were limited to only those variables that had correlations higher than 0.1 with the outcome. This pruning left 30 of the predictor variables that had a correlation larger than 0.1 with the outcome.

We limited the outcome to a maximum of one year. The number of clusters varied between two and 35 (k=2 to 35) and the whole process was repeated 50 times (N=50). A four cluster solution (k=4) was detected as the best solution with a minimum mean absolute error of 70.473. Using cluster analysis prior to regression analysis (unclustered data, MAE = 72.334), k=4 showed a 1.86 day improvement in MAE compared to when no clustering was used. Figure 8 shows the relationship between the number of partitions and the resulting MAE for “ED time 365”.

Figure 8. Relationship between number of partitions and MAE for the test dataset (ED time 365 and correlation >=0.1)



3.2 Second Target: visit an ED within 7 days after the index date (ED within 7d)

Using cluster analysis and logistic regression, we predicted whether or not patients visited an ED within seven days of their index date. We limited the outcome to a maximum of one year (365 days). The bivariate correlation matrices between the outcome variable and all the predictor variables was calculated and the predictor variables were limited to only those variables that had correlations higher than 0.1 with the outcome. Thirty of the predictor variables had correlations greater than 0.1 with the outcome and we used only those variables in the clustering, and logistic regression analysis. Number of clusters varied between two and 20 (k=2 to 20) and the whole process was repeated 50 times (N=50). A four cluster solution (k=4) was chosen as the best solution with a maximum accuracy of 62.07%, a maximum sensitivity of 73.84%, a maximum specificity of 60.26%, a maximum AUC of 72.05% and a maximum d-prime of 0.7692.

Compared to the condition when we don't apply cluster analysis before logistic regression, k=4 had a 1.8% improvement in accuracy, a 1.3% improvement in sensitivity, a 1.9% improvement in specificity, a 0.66% improvement in AUC and an 0.077 improvement in d-prime. Figure 9 shows the relationship between the number of partitions and the accuracy measures for "ED within 7d". Table 5 compares the best cluster solution (logistic regression analysis based on four clusters) with the solution where logistic regression analysis was carried out without prior cluster analysis.

Figure 9. Relationship between number of partitions and accuracy measures for the test set (ED within 7d and correlation >=0.1)

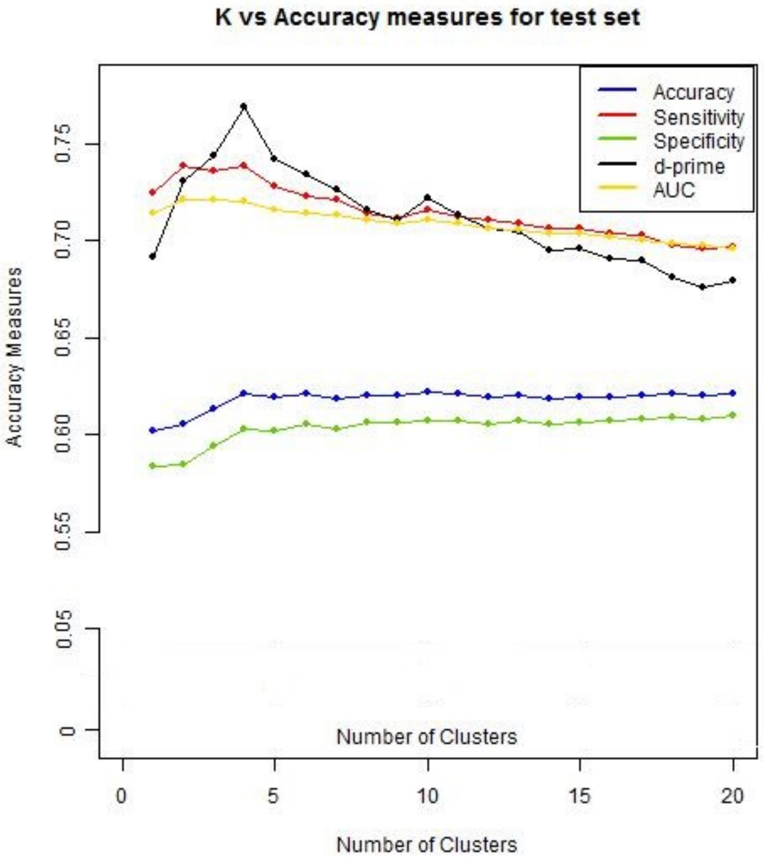


Table 5. Comparison of best cluster solution and unclustered data for "ED within 7d"

K (Cluster Number)	Accuracy	Sensitivity	Specificity	d-prime	AUC
K=1	0.6022	0.7249	0.5834	0.6913	0.7139
K=4	0.6207	0.7384	0.6026	0.7692	0.7205
Maximum Value	0.6218	0.7388	0.6094	0.7692	0.7214
Difference between k=1 and k=4	0.0184	0.0135	0.0192	0.0779	0.0066

3.3 Third Target: Vital Status

For the third outcome variable (vital status), the predictor variables used in the cluster analysis and logistic regression analysis were individual symptom scores, demographics (Sex, Age, Income quantile, Rural, Charlson, LHIN), cancer type and cancer stage.

For patients with multiple possible index dates, we chose the index date that was closest to the date of death for patients who died during the period of study or that was closest to the end of the study (Dec, 31, 2013) for those who survived to the end of the study period.

Using cluster analysis and logistic regression, we predicted vital status of patients. Number of clusters was varied from 2 to 50 (k=2 to 50) and the whole process was repeated 50 times (N=50). A five cluster solution (k=5) was detected as the best solution, with maximum accuracy of 73.8%, sensitivity of 83.1%, specificity of 78.8%, AUC of 86.2% and maximum d-prime of 1.60.

Compared to the condition when we didn't apply cluster analysis before logistic regression, the five cluster solution led to a 0.4% improvement in accuracy, a 0.2% improvement in sensitivity, a 0.3% improvement in specificity, a 0.3% improvement in AUC and a 0.02 improvement in d-prime. Table 6 shows the comparison of best cluster solution to unclustered data for prediction of vital status. This finding that gains due to prior clustering for logistic regression analysis tend to be low, and lower than the benefits of clustering for linear regression analysis, is a general finding that we have found in a number of analyses that we have carried out, both on this data, and on MIMIC II data (Rouzbahman, Jovicic, & Chignell, 2016).

Table 6. Comparison of best cluster solution to unclustered data for vital status

Death/alive status	K=1	K=5	Difference
Accuracy	0.7339	0.7380	0.4%
Sensitivity	0.8286	0.8311	0.2%
Specificity	0.7855	0.7888	0.3%
AUC	0.8597	0.8629	0.3%

4 Discussion

Based on the results of this study, we conclude that summarized patient types not only facilitates non-confidential data dissemination, but also contribute to more accurate predictions. Prediction of ED visit and vital status for cancer patients based on the obtained patient types resulted in higher accuracy/less error. This greater accuracy is useful because determining the next ED visit for cancer patients can assist health care professionals in better planning of patients' further assessment and better allocation of resources.

In this study, an unsupervised method was applied to find similar patients. K-means cluster analysis was carried out, varying number of partitions between 2 and 50 and with the use of different random input conditions to provide more stable estimates of mean accuracy. The same method was used previously in other studies of ICU data (Chignell, Rouzbahman, Kealey, Samavi, & Sieminowski, 2013) (Rouzbahman, Jovicic, & Chignell, 2016).

For prediction of first outcome variable in the adult cancer patients, with ED time (time between index date and ED date) as the outcome, and with clustering applied before regression (using all the predictor variables), MAE (mean absolute error) improved slightly versus no clustering. In the case of a maximum one year for the outcome (ED 365), a five cluster solution reduced the MAE by 0.36 days. However, when a maximum of only three months was used for the outcome (ED time 90), the best-fitting four cluster solution reduced the MAE by only 0.12 days. When we applied as predictors only those variables having a correlation greater than 0.1 with the target variable, using a maximum of one year for the outcome, a four cluster solution reduced the MAE by 1.86 days. These results indicate that regression analysis using a small number of clusters (where analyses are based on predictors that show some relation with the outcome) boosts prediction performance, and that clustering performance is better when the analysis is restricted to predictors with larger correlations with the outcome variable (in this study a correlation cutoff of 0.1 was used). Since regression analysis can be carried out on summarized data, this result suggest that privacy preserving analyses can be carried out on summarized data, and that the results obtained using those analyses should be competitive with analyses carried out on the raw (case level) data.

We also applied clustering before regression for prediction of the second outcome variable in the adult cancer patients, ED within 7d (visit an ED within 7 days after the index date or not). In that case, accuracy measures were improved moderately versus no clustering. In the case of a maximum of one year for the outcome, a four cluster solution increased the accuracy by 1.8%, the sensitivity by 1.3%, the specificity by 1.9% and the AUC by 0.66%. Additionally, the D-prime measure increased by 0.08.

5 Conclusions

This study utilized datasets which contain detailed clinical, demographic and administrative data for patients with cancer diagnosis across Ontario. While such administrative data are incomplete and can be improved (Juurlink D, et al., 2006), this study tried to use the most relevant features, currently available, for ED visit prediction. In this study, variables and outcomes were created using medical expertise relating to management of cancer patients. The cohort dataset was created with help of a physician and the methodology applied in linking the 6 mentioned datasets to the cohort dataset was also generated by clinicians. All the preprocessing work such as grouping the diagnosis codes and medications, and replacing the missing values, were done after consultation with clinicians.

Applying analytical tools on electronic health records can assist researchers and health care professionals in taking advantages of valuable knowledge hidden in patients' records. This knowledge can then be used to facilitate service delivery and assessment planning for patients in general, and in particular for the cancer patients in this study who were suffering pain and severe symptoms. If we can predict when patients might visit an emergency department, strategies can be developed to decrease ED visits. Predicting next emergency department visit and vital status are only two of the outcomes possible and the same method can be applied to make predictions of other outcomes of interest.

Based on the results of this study, we can conclude that it is possible to automatically generate meaningful clusters (patient types) in health data repositories (in this case for cancer patients), and to make relatively accurate predictions based on summaries of those clusters. We have previously demonstrated the feasibility of our approach in the context of ICU patients (Rouzbahman et al., 2016) and the results of the present study study show that the earlier results generalize to the cancer data considered here.

Clusters of similar patients are beneficial when case level data is confidential or unavailable, but where non-confidential summaries of patient types - means, standard deviations and correlations amongst a set of variables – are available for further usage. One of the major findings of our work in this study and the earlier one using ICU data (Rouzbahman et al., 2016) is that summarized data can be used in place of case level (raw) data with little if any loss of predictive power.

This study utilized a variety of Ontario health data sets in order to investigate how severe symptom scores contribute to ED visits. We were able to show a benefit to using cluster-boosted regression (Rouzbahman et al., 2016) even though correlations between the predictors and the outcomes were generally low (and typically no more than 0.3). At present it is typically very different to get access to health data because of privacy concerns. We hope that the present results will show that summarized data can support predictive analytics while respecting the privacy of patients.

Detecting patients that have high probability of ED visit and understanding why they visit the ED near the end of life is an interesting research question by physicians and health care providers. This study presented a methodology to find patients with high risk of ED visits based on patient types. This study also showed how summarized patient clusters can contribute to more accurate predictions. In this study, cluster-boosted regression (regression on summarized, clustered data) was competitive with, and even better than, linear/logistic regression on raw data, for all the three outcomes. Our method reduced MAE by 1.86 days for time between index date and ED date and increased the accuracy measures by 1.9% for "ED within 7d" and by 0.4% for vital status prediction.

Bibliography

Barbera, L., Atzema, C., Sutradhar, R., Seow, H., Howell, D., Husain, A., et al. (2013). Do patient-reported symptoms predict emergency department visits in cancer patients? A population-based analysis. *Annals of emergency medicine , 61* (4), 427-437.

Barbera, L., Taylor, C., & Dudgeon, D. (2010). Why do patients with cancer visit the emergency department near the end of life? *Canadian Medical Association Journal , 82* (6), 563-568.

Bruera, E., Kuehn, N., Miller, J. M., Selmser, P., & Macmillan, K. (1991). The Edmonton Symptom Assessment System (ESAS): a simple method for the assessment of palliative care patients. *Journal of palliative care* .

Chan, L., Chan, T., Cheng, L. F., & Mak, W. S. (2010). Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy. *IEEE International Conference on* (pp. 467-470). Bioinformatics and Biomedicine Workshops (BIBMW).

Chignell, M., Rouzbahman, M., Kealey, R., Samavi, R., & Sieminowski, T. (2013). Nonconfidential Patient Types in Emergency Clinical Decision Support. *IEEE Security and Privacy , 11* (6), 12-18.

Clarke, A. E., Marrett, L. D., & Kreiger, N. (1991). Appendix 3 (c) Cancer registration in Ontario: a computer approach. Cancer Registration: Principles and Methods. *IARC Publication , 95*, 246-257.

Deyo, A. R., Cherkin, C. D., & Ciol, A. M. (1992). Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *Journal of clinical epidemiology , 45* (6), 613-619.

Haim, B., Nutman, A., Shoseyov, D., Shalom, M., Peled, R., Kivity, S., et al. (2002). Prediction of emergency department visits for respiratory symptoms using an artificial neural network. *122* (5), 1627-1632.

Hu, Z., Jin, B., Shin, Y. A., Zhu, C., Zhao, Y., Hao, S., et al. (2015). Real-Time Web-Based Assessment of Total Population Risk of Future Emergency Department Utilization: Statewide Prospective Active Case Finding Study. *Interactive journal of medical research , 4* (1).

Institute for Clinical Evaluative Sciences (ICES): Home Care Database (HCD) [Canada] [administrative database]. (2011). (Ministry of Health and Long-Term Care, Government of Ontario;) From http://ophid.scholarsportal.info/details/view.html?q=IDs&uri=/phirn/hcd_PHIRN_E.xml

Iron, K., Zagorski, M. B., & Sykora, K. (2008). *Living and dying in Ontario: an opportunity for improved health information.* ICES investigative report [Internet], Institute for Clinical Evaluative Sciences, Toronto.

Juurlink D, D., Preyra C, C., Croxford R, R., Chong A, A., Austin, P., Tu, J., et al. (2006, June). *Canadian institute for health information discharge abstract database: a validation study.*

LaMantia, A. M., Platts-Mills, F. T., Biese, K., Khandelwal, C., Forbach, C., Cairns, B. C., et al. (2010). Predicting hospital admission and returns to the emergency department for elderly patients. *Academic Emergency Medicine , 17* (3), 252-259.

McCusker, J., Bellavance, F., Cardin, S., Belzile, E., & Verdon, J. (2000). Prediction of hospital utilization among elderly patients during the 6 months after an emergency department visit. *Annals of emergency medicine , 36* (5), 438-445.

Meldon, W. S., Mion, C. L., Palmer, M. R., Drew, L. B., Connor, T. J., Lewicki, J. L., et al. (2003). A Brief Risk-stratification Tool to Predict Repeat Emergency Department Visits and Hospitalizationsin Older Patients Discharged from the Emergency Department. *Academic Emergency Medicine , 10* (3), 224-232.

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology , 9* (1), 15-24.

New Drug Funding Program (NDFP) & Evidence Building Program (EBP), Approved Drugs and Eligibility Criteria. (2015, 07). From https://www.cancercare.on.ca/toolbox/drugs/ndfp/

New Drug Funding Program Database (NDFPDB). (1995). From http://ophid.scholarsportal.info/details/view.html?q=NDFP&field=&date-gt=1871&date-lt=2011&uri=/phirn/ndfpdb_PHIRN_e.xml

Ontario Drug Benefit (ODB). (1992). From http://ophid.scholarsportal.info/details/view.html?q=he&uri=/phirn/odb_PHIRN_e.xml

Robles, C. S., Marrett, D. L., Clarke, E. A., & Risch, A. H. (1988). An application of capture-recapture methods to the estimation of completeness of cancer registration. *Journal of clinical epidemiology , 41* (5), 495-501.

Rouzbahman, M., Jovicic, A., & Chignell, M. (2015). Can Cluster-Boosted Regression Improve Prediction: Death and Length of Stay in the ICU. *Submitted to Journal of Behavioural Health Informatics (under revision)* .

Sayad, S. (2010). *An Introduction to Data Mining*. (Data Mining Group, University of Toronto) From http://www.saedsayad.com/data_mining_map.htm