Identifying Risk Groups Associated with Colorectal Cancer

Jie Chen[1], Hongxing He[1], Huidong Jin[1], Damien McAullay[1], Graham Williams[1,2], and Chris Kelman[3]

[1] CSIRO Mathematical and Information Sciences,
GPO Box 664, Canberra ACT 2601, Australia
Firstname.Lastname@csiro.au
[2] Current address: Australian Taxation Office,
51 Allara Street, Canberra ACT 2601, Australia
Graham.Williams@togaware.com
[3] National Centre for Epidemiology and Population Health,
The Australian National University, Canberra ACT 0200, Australia
Chris.Kelman@anu.edu.au

Abstract. In this paper, we explore data mining techniques for the task of identifying and describing risk groups for colorectal cancer (CRC) from population based administrative health data. Association rule discovery, association classification and scalable clustering analysis are applied to the colorectal cancer patients' profiles in contrast to background patients' profiles. These data mining methods enable us to identify the most common characteristics of the colorectal cancer patients. The knowledge discovered by data mining methods which are quite different from traditional survey approaches. Although it is heuristic, the data mining methods may identify risk groups for further epidemiological study, such as older patients living near health facilities yet seldom utilising those facilities, and with respiratory and circulatory diseases.

1 Introduction

Colorectal cancer (cancer of the colon or rectum, abbreviated as CRC) is the second leading cause of cancer-related deaths in the United States for both men and women combined. The disease surpasses both breast and prostate cancer in mortality, and is second only to lung cancer in cause of cancer deaths. Despite the fact that it is highly preventable, approximately 146,940 new cases of colorectal cancer was diagnosed in 2004 and more than 56,000 people will die from the disease in USA [1]. An almost equal number of men and women are diagnosed each year.

In Australia, colorectal cancer is the third most common cause of death from cancer in women (after breast and lung cancer) and men (after lung and prostate cancer). The exact cause of colorectal cancer is unknown, in fact it is thought that there is not one single cause. It is more likely that a number of factors, some known and many unknown, may work together to trigger the development of

G.J. Williams and S.J. Simoff (Eds.): Data Mining, LNAI 3755, pp. 260–272, 2006.
© Springer-Verlag Berlin Heidelberg 2006

colorectal cancer. Previous studies have identified risk factors which may increase a person's risk of developing colorectal cancer. The following factors are widely accepted risks:

- **Age.** Increasing age is considered a major risk factor for developing colorectal cancer. Colorectal cancer is rare in people under 40. The risk increases after the age of 40, rising sharply and progressively after the age of 50.
- **Dietary factors.** It is estimated that rates of colorectal cancer could be reduced in western populations by up to 35% through changes to the food we eat. A diet that is high in fat and low in fibre and vegetables has been linked with an increased risk of colorectal cancer. There has also been an association between heavily browned or charred meat and colorectal cancer. Excessive alcohol intake and a diet low in calcium have also been implicated.
- **Behavioural and lifestyle factors.** An inactive lifestyle, obesity and smoking have been associated with an increased risk of developing colorectal cancer.
- **Regional factors.** People in western countries have a higher incidence of colorectal cancer than people in Asian or African countries. This may be partly due to differences in diet.

This paper aims at studying the relationship between CRC prevalence and various attributes of the patients. These attributes include demographics, medical service history etc. We use large administrative data sets instead of data from survey. The advantages are large coverage of population, low cost and less selection bias. In our case, our dataset covers the medical records of more than one million people. The disadvantage is that we can not design what information we get from individual patients as in designed survey data. Since the transaction data is collected for administrative purposes only, some important information regarding patients' diet and lifestyle is missing. Therefore some well known risk factors can not be verified using administrative data. Nevertheless, it is a good practice to discover unexpected and interesting relationships from this dataset. Since data to be explored is large, and most traditional methods dealing with small samples are not working well, therefore we employ various data mining techniques in our analysis. Exploratory tools of data mining on large dataset may be able to find some factors previously unnoticed. Furthermore, we may identify a few groups of patients with some common characteristics who are at risk of colorectal cancer.

Exploratory health data mining is a rewarding but highly challenging area [14,3]. Recently there have been a few data mining projects initiated for the surveillance and analysis of colorectal cancer patients. A Bayesian framework to extract recurrence, the key outcome for measuring treatment effectiveness for colorectal cancer patients, has been built in [13]. Logistic regression [11] and survival analysis [15] have been applied to identify recurrences and to model the prognosis of colorectal cancer patients. Different from these studies, this paper aims at identifying and describing risk groups rather than single risk factors efficiently. This paper applies various data mining techniques on linked administrative

health dataset QLDS. In [2,4], adverse drug reaction has been successfully identified from the same dataset using association and classification algorithms.

The rest of the paper is organised as follows. Section 2 describes the dataset and features selected for the mining process. Sections 3, 4 and 5 describe the methods and mined results for association rule discovery, scalable cluster analysis and association classification analysis respectively. Section 6 discusses the advantages and limitations of the methods. Section 7 concludes the paper.

2 Data Preparation

2.1 QLDS

We use the Queensland Linked Data Set (QLDS) [17] for this exploratory data mining study. The Queensland Linked Data Set (QLDS) was made available to CSIRO under an agreement between Queensland Health and the Commonwealth Department of Health and Ageing (DoHA). The data set contains de-identified and confidentially linked patient level hospital separation data (1 July 1995 to 30 June 1999), Medicare claims data and Pharmaceutical Benefits Scheme (PBS) data (both 1 January 1995 to 31 December 1999). All data were de-identified, and actual dates of service were removed, so that time sequences are indicated by time from first admission. This process provided strong privacy protection, consistent with the requirement of the relevant Federal and State legislations.

The QLDS is based on the collection of patients hospitalised in Queensland between 1 July 1995 to 30 June 1999, with linked PBS and MBS data. Because the linkage relied on a valid Medicare number, around 30% of hospital records (those without a valid Medicare Number associated) were discarded. The QLDS therefore contains 3,087,454 hospital records, corresponding to 1,176,294 individuals, which represents about 35% of the Queensland population. The issues of selection bias and data quality of the QLDS are discussed in the report [18].

2.2 Population Selection

A patient is flagged as a CRC patient if they have ever had a hospital separation between July 1995 and June 1999 with a diagnosis indicating CRC. The ICD9 (The International Classification of Diseases, 9th Revision) codes included are those beginning with 153 (for malignant neoplasm colon) or 154 (malignant neoplasm of rectum/anus). All ten diagnosis flags in hospital separation data are considered. There are 8,104 such patients. In our analysis, the CRC patients are classified into class 1, all the other patients are classified into class 0.

2.3 Feature Selection

Table 1 lists the features selected for the study. The postcode is based on the patients' MBS records. Those patients who do not have any MBS records or their postcode does not fall in Queensland have the field value "NO" as a missing

Table 1. Features selected for the study

Feature	Description	Data Type
Linkid	Encrypted link id	ID variableSymbol
Gender	m:male, f:female	Binary
Age	Age at 1995	Integer
Age Group	Discrete age group	00-43,44-53,54-63,64-73,74-00
Postcode	Postcode of Patient	categorical
Aria Continuous	Access to health facility	Continuous
Aria Discrete	Access to health facility	HA, A, MA,R, VR
Seifa Continuous	Postcode's average household income[1]	Continuous
Seifa Discrete	Postcode's average household income	High, Medium, Low
Consultation Continuous	Number of physician consultations	Continuous
Consultation Discrete	Number of physician consultations	High, Medium, Low
Diagnostic Continuous	Number of diagnostic items in MBS	Continuous
Diagnostic Discrete	Number of diagnostic items in MBS	High, Medium, Low
Procedure continuous	Number of procedure items in MBS	Continuous
diabetes	diabetes flag	0,1
mental	mental flag	0,1
circulatory	circulatory flag	0,1
heart	heart flag	0,1
respiratory	respiratory flag	0,1
asthma	asthma flag	0,1
musculoskeletal	musculoskeletal flag	0,1
Class	0:non-crc 1:crc	Binary

value. For the patients who have more than one MBS record, the majority value is used to decide the value of postcode. The Seifa (Social Economic Index for Areas) data are mapped from postcodes according to the 1996 Australian Census data. The Aria (Accessibility/Remoteness Index of Australia) data are derived from postcodes to reflect the accessibility to health care facilities.

Consultation and diagnosis record the average number of physician consultations and diagnostic items per year respectively. This is calculated for the period prior to the first CRC hospital event for CRC patients and for the entire five years (1996–1999) for non-CRC patients. Consultation is discretised to Low ($c < 4.8$), Medium ($4.8 \leq c < 9.2$) and High($c \geq 9.2$). Diagnostic is discretised to Low ($d < 2.0$), Medium ($2.0 \leq d < 5.43$) and High($d \geq 5.43$). These cutoff values are chosen based on results from running an association algorithm (Magnum Opus). The discretised Seifa values are Low ($s \leq 856.86$), Medium($856.86 < s \leq 1032.15$) and High($s > 1032.15$) so that the population of Queensland has 25% belonging to High, 50% to Medium and 25% to Low.

3 Association Rule Discovery

3.1 Method

The aim of association analysis is to discover the association between available variables and the colorectal cancer prevalence. Magnum Opus was first applied to the whole population in the QLDS. Magnum Opus is an ease to use association rule discovery tool with excellent flexibility. It finds rules from both transaction

[1] Year 2000 survey result from Australian Bureau of Statistics.

data and attribute-value data efficiently [16]. It can discretise the numeric attributes automatically.

3.2 Feature Selection

The selected features used for analysis are listed as follows:

- Gender: m, f
- Age: numeric 3
- AriaDis: categorical
- Seifa: numeric 3
- Consultation: numeric 3
- Diagnostic: numeric 3
- Procedure: numeric 3
- Seven Diagnosis flags: 0,1
- Class: 0, 1

All numeric features are discretised into three sub-ranges, each of which contains approximately the same number of cases.

3.3 Results for All Patients

A rule has two parts: a Left Hand Side (LHS) and a Right Hand Side (RHS). The strength of a rule is the proportion of examples covered by the LHS of the rule that are also covered by the RHS. The lift of a rule is the strength divided by the RHS coverage proportion. This indicates how much more frequent the RHS is than normal if the LHS occurs. Table 2 shows a part of the association rules sorted by lift in descending order. Magnum Opus took 21.12 seconds to generate sorted best 100 association rules. Our observations are as follows.

Table 2. Part of the rules identified by Magnum Opus on all patients

Rule No	LHS	RHS(Class 1)	Lift
1	Gender=m Age $>$ 50 AriaDis=HA Consultation $<$ 5.80	806	4.75
2	Age $>$ 50 AriaDis=HA Consultation $<$ 5.80	1299	4.61
3	Age $>$ 50 Consultation $<$ 5.80 2.40 $\leq$ Diagnostic $\leq$ 6.40	767	4.55

- Rule 1 is interesting, covering about 10% of the 8,104 CRC patients. This group of patients include males aged above 50, with high accessibility to health facilities and having consultation counts less than 5.8. Patients identified by the rule are 4.75 times more likely to have CRC than the general population.

– Rule 2 is a more general rule covering 16% of the CRC population and retaining a lift of 4.61. Rule 3 is similar.
– These rules all suggest that older patients (50+) with accessibility to health care facilities but low utilisation rates are more than four times more likely than the general population to develop colorectal cancer.

3.4 Results for Patients Over 44

Since most CRC patients are older than 40, we selected patients over 44 years of age to form a new dataset for analysis. Magnum Opus was applied to this dataset with selected features and parameters for discretisation as above. Table 3 shows a part of the association rules sorted by lift in descending order. Magnum Opus took 4.57 seconds to generate sorted best 100 association rules. Our observations are as follows.

– Patients aged between 55 and 68 with circulatory disease and a low utilisation of consultations are more than twice as likely as the general population to have colorectal cancer.

Table 3. Part of the rules identified by Magnum Opus on all patients over 44

Rule No	LHS	RHS(Class 1)	Lift
1	$55 \leq$ Age ≤ 68 Consultation < 8.00 circulatory=1 heart=0	565	2.82
2	AriaDis=HA Consultation < 8.00 respiratory=1 asthma=0	486	2.55
3	$55 \leq$ Age ≤ 68 Consultation < 8.00 circulatory=1	714	2.41
4	AriaDis=HA Consultation < 8.00 respiratory=1	572	2.35
5	Consultation < 8.00 heart=0 respiratory=1	711	2.28
6	Consultation < 8.00 respiratory=1 asthma=0	726	2.22
7	AriaDis=HA Consultation < 8.00 circulatory=1 heart=0	760	2.20
8	Gender=m $55 \leq$ Age ≤ 68 Consultation < 8.00 musculoskeletal=0	906	2.15
9	$55 \leq$ Age ≤ 68 AriaDis=HA Consultation < 8.00 musculoskeletal=0	863	2.12
10	Consultation < 8.00 circulatory=1 heart=0 musculoskeletal=0	1052	2.10

– Patients with circulatory disease and a low utilisation of consultations living in regions highly accessible to health care facilities are more than twice as likely as the general population to have colorectal cancer.

4 Scalable Cluster Analysis

4.1 Method

Clustering is one of the most widely used techniques in data mining. It is used to reveal patterns in data that can be extremely useful to data analysts. The task of clustering is to partition a data set into clusters in such a way that the data records within each cluster are more similar among themselves than data records in other clusters [5,8]. A scalable clustering system, the computational time of which grows linearly or sub-linearly with the number of data records, bridges the gap between the limited computational resources and large databases [9,7].

We employed a scalable clustering algorithm, BIRCH [19], to identify the groups of patients who are more likely to suffer from CRC. First we normalised each continuous attribute into the interval [0,1]. Then BIRCH with default setting was used to generate 100 clusters based on these continuous attributes. After that, CRC patients within each cluster was used to identify high risk clusters in comparison with the whole data set. For example, the lift is defined as the proportion of CRC patients covered by a cluster divided by the proportion of non-CRC patients covered by this cluster. It roughly indicates to what degree this cluster of people are more likely to suffer from CRC than the whole population. The clusters that have less than 200 patients are left out since they are too small compared with the whole data set.

4.2 Feature Selection

The selected features are listed as following.

– Age: numeric
– AriaCon: numeric
– Seifa: numeric
– Consultation: numeric
– Diagnostic: numeric
– Class: 0, 1

4.3 Clusters for All Patients

We first applied BIRCH to generate 100 clusters for all the 1,176,294 patients. It took about 8.19 seconds in total and about 52,608 patients were not clustered and viewed as outliers.

Table 4 lists typical clusters with high proportions of CRC patients. The clusters are listed in descending order with respect to their lift. The clusters with lift less than 2.0 are omitted from the table. Each row indicates an interesting

Table 4. Typical clusters with high risk for CRC patients identified from all the 1,176,294 patients

Cluster ID	Age	Aria-Con	Seifa	Consu-ltation	Diag-nostic	Class 1	Coverage		Lift
							Cardinality	%	
0	81.7	0.12	859.6	11.9	5.7	45	1623	0.144	4.17
82	71.7	4.23	967.2	13.2	8.5	236	9485	0.844	3.73
98	71.5	4.97	1014.6	11.8	7.2	38	1696	0.151	3.35
12	79.4	0.55	965.7	16.0	8.5	821	38625	3.437	3.18
43	65.4	0.02	1048.5	43.8	62.1	62	2934	0.261	3.16
63	77.5	2.84	963.8	13.7	7.9	377	18120	1.613	3.11
46	78.2	0.13	1164.3	15.1	9.0	65	3146	0.280	3.09
39	78.0	5.92	950.0	11.9	6.7	68	3298	0.293	3.08
72	67.7	0.32	969.3	15.1	9.5	1293	62716	5.581	3.08
83	64.7	11.46	943.8	8.2	4.2	15	728	0.065	3.08
37	67.9	2.74	882.5	10.8	6.3	58	2820	0.251	3.07
55	62.1	7.89	922.1	9.8	6.1	23	1128	0.100	3.04
86	78.2	0.06	1049.2	16.2	9.3	559	27477	2.445	3.04
6	78.4	10.64	923.3	9.3	3.5	19	948	0.084	2.99
95	65.6	0.11	1045.6	14.4	9.1	750	37627	3.349	2.97
53	80.8	3.54	897.7	11.9	5.3	45	2345	0.209	2.86
79	60.5	3.09	995.7	11.4	7.8	266	14073	1.252	2.82
47	61.1	2.64	941.2	11.3	7.6	424	22693	2.020	2.79
76	64.7	4.15	907.8	10.4	6.3	63	3396	0.302	2.76
97	60.9	10.17	1020.7	6.5	3.9	12	655	0.058	2.73
58	72.9	1.70	1020.4	12.9	7.8	54	3011	0.268	2.67
75	64.7	5.79	953.2	10.4	7.0	97	5566	0.495	2.59
78	55.9	0.36	869.4	11.6	7.2	72	4145	0.369	2.59
13	78.0	8.12	921.3	9.8	4.2	7	411	0.037	2.53
21	60.8	0.16	1168.3	12.3	8.6	102	6417	0.571	2.36
74	62.4	10.47	892.7	8.7	5.0	24	1521	0.135	2.35
66	57.9	1.76	1035.7	10.7	6.9	43	3191	0.284	2.00

cluster described by a cluster centroid. For example, as listed in the first row of Table 4, Cluster 0 has a centroid of Age: 81.7, Aria: 0.12, Seifa: 859.6, Consultations: 11.9, and Diagnostics: 5.7. There are 1,623 patients in the cluster, and 45 CRC patients. The lift is 4.17, i.e., the patients within the cluster are 4.17 times more likely to suffer from CRC. It indicates that this cluster of patients are more likely to suffer from CRC, compared with the whole data set. Similar interesting results can be found in Table 4.

4.4 Clusters for Patients Over 44

We also conducted cluster analysis on the patients over 44 years of age. BIRCH took about 2.39 seconds to generate 100 clusters from the 453,645 patients and generated 27,955 outliers.

Table 5 lists some typical clusters with high proportions of CRC patients from these old patients. They are sorted by lift in descending order, while those with lift less than 1.30 are omitted. A typical example is Cluster 31 as listed in Table 5. Its cluster centre is Age: 65.1, Aria: 5.41, Seifa: 896.3, Consultations: 41.3, and Diagnostics: 54.0. There are 284 patients in the cluster, and 11 CRC patients. The lift is 2.50, this cluster of patients are significantly different from other patients over 44 years of age. Similar results can be observed from other clusters.

Table 5. Typical clusters with high risky of CRC patients on 453,645 patients elder than 44

Cluster ID	Age	Aria-Con	Seifa	Consu-ltation	Diag-nostic	Class 1	Coverage		Lift
							Cardinality	%	
31	65.1	5.42	896.3	41.3	54.0	11	284	0.063	2.50
88	62.9	7.17	886.8	26.6	35.0	11	296	0.065	2.39
29	75.2	8.40	922.0	37.7	49.3	7	216	0.048	2.08
30	81.1	3.64	861.0	25.8	33.9	13	423	0.093	1.97
65	69.1	11.03	938.3	33.8	44.1	12	416	0.092	1.84
71	68.2	4.98	1016.0	31.1	40.6	34	1295	0.285	1.67
2	73.3	0.31	1199.2	14.5	19.0	26	996	0.220	1.66
76	72.9	4.74	907.5	31.2	40.7	34	1338	0.295	1.62
37	65.3	0.14	887.2	28.0	36.6	152	6040	1.331	1.60
81	68.0	3.62	981.7	38.2	49.9	126	5133	1.132	1.56
67	74.5	3.72	981.9	40.5	52.8	85	3487	0.769	1.55
4	82.0	5.93	952.1	26.3	34.3	43	1774	0.391	1.54
13	79.0	0.06	866.1	35.7	46.7	49	2061	0.454	1.51
26	81.9	3.59	991.3	35.5	46.4	82	3540	0.780	1.47
15	66.1	0.31	1197.0	31.9	41.7	20	867	0.191	1.46
86	73.6	0.09	1133.6	37.3	48.7	22	985	0.217	1.42
16	71.5	0.39	966.4	41.0	53.5	355	15943	3.514	1.41
79	76.1	0.66	945.5	37.7	49.2	284	12816	2.825	1.41
90	80.8	4.10	933.4	33.6	43.9	39	1766	0.389	1.40
78	75.9	0.10	1044.1	28.3	36.9	420	19294	4.253	1.38
39	80.6	0.31	1198.2	30.5	39.8	11	514	0.113	1.36
44	65.6	2.58	945.7	35.2	45.9	212	9921	2.187	1.35
46	77.3	2.69	944.4	36.0	46.9	98	4628	1.020	1.34
94	61.3	3.12	997.6	32.7	42.6	160	7569	1.668	1.34
34	69.1	0.09	1049.0	34.5	45.0	424	20059	4.422	1.34
82	71.0	2.82	940.2	32.5	42.4	188	9016	1.987	1.32
54	58.4	0.34	869.3	20.7	27.1	40	1942	0.428	1.30
51	73.5	1.75	1032.9	33.5	43.7	34	1654	0.365	1.30
27	66.3	3.24	887.3	37.4	48.8	24	1168	0.257	1.30
74	81.7	10.53	903.4	32.3	42.3	6	292	0.064	1.30

5 Association Classification

5.1 Method

The association classification algorithm developed in [10] generates the optimal class association rule set. The experimental results in [10] show that the optimal class rule set achieves a very high classification accuracy.

However, our dataset has very unbalanced classes. Our main interest is in finding rules (or cohorts) which lead to higher occurrences of colorectal cancer patients than the average occurrence. As a result, the original algorithm has been modified to increase classification accuracy of class 1 patients. The modification is that, instead of using the minimum global support as a criterion for rules to be included, local support is introduced to find the rules describing the small class (class 1). *Local Support* is defined by Equation 1.

$$lsup(A \rightarrow c) = \frac{sup(A \rightarrow c)}{sup(c)} \tag{1}$$

Here $sup(c)$ and $sup(A \rightarrow c)$ represent the support (or proportion or relative frequency) of class c in the whole population and the support of pattern A in

class c respectively. The algorithm will identify rules which give high "lift" values for class 1. Lift is defined in Equation 2.

$$lift(A \rightarrow c) = \frac{lsup(A \rightarrow c)}{sup(A)} \tag{2}$$

5.2 Results for All Patients

Example rules identified are listed in Table 6. Features selected are similar to Section 3 except that some features are discretised to categorical variables. Rule 1 identifies patients with the following characteristics:

- Aged between 64 and 73.
- Living in areas highly accessible to medical facilities.
- Having small number of doctor's consultations.
- No heart and musculoskeletal diseases.

Table 6. Part of the rules identified by association classification algorithm for all patients

Rule No	Rule	Class 1	Lift
1	Age = 64-73 Aria = HA Consultation = Low heart = 0 musculoskeletal = 0	273	6.74
3	Age = 54-63 Aria = HA Consultation = Low heart = 0 musculoskeletal = 0	317	6.28
10	Age = 64-73 Consultation = Medium diabetes = 0 circulatory = 1 heart = 0	255	6.19
62	Gender = m Aria = HA mental = 0 circulatory = 1 heart = 0 respiratory = 1 asthma = 0	260	5.05
66	Gender = m Age = 64-73 heart = 0 respiratory = 1 asthma = 0	269	4.97
79	Age = 74-00 circulatory = 1 heart = 0 respiratory = 1 musculoskeletal = 0	278	4.89
91	Consultation = Low circulatory = 1 respiratory = 1 asthma = 0	283	4.77

There are a total of 273 CRC patients in this group. The lift of the group is 6.74. It implies that the individuals who have these characteristics are 6.74 times more likely to have CRC than general population. Rule 62 indicates that for males, living in highly accessible area with circulatory and respiratory diseases, but no heart and asthma diseases, the likelihood of CRC is 5.05. Rules 62, 79 and 91 all suggest that CRC is correlated with circulatory and respiratory diseases.

5.3 Results for Patients Over 44

Results for patients over 44 are shown in Table 7.

Table 7. Part of the rules identified by association classification algorithm for patients older than 44

Rule No	Rule	Class 1	Lift
1	Age Group = 64-73 Consultation = Low heart = 0 musculoskeletal = 0	475	2.62
2	Consultation = Low heart = 0 Respiratory = 1	414	2.48
4	Aria = HA Consultation = Low circulatory = 1 heart = 0	385	2.51
6	Age = 54-63 Consultation = Low musculoskeletal = 0	568	2.35

6 Discussion

The results obtained by the three data mining techniques are consistent with aggregation results based on CRC and non CRC patients profiles [6]. Most of the interesting results are agreeable in terms of high lift value, especially for the results by using association rule and association classification techniques. For instance, Rules 5 and 7 in Table 3 agree with Rules 2 and 4 in Table 7. The results from scalable clustering analysis are not as expressive as those from the former two techniques, but it can efficiently draw a big picture about the characteristics of CRC patients against whole population.

Logistic regression in R has been tried on the dataset [12]. Risk factors highlighted include Age, "Gender = m", "AriaDis" except for "AriaDis=R" and "AriaDis=VR", "Consultation", "Diagnostic", and seven diagnosis flags (mental flag was not highlighted for patients over 44). Nonetheless, it is not trival to identify risk groups through the statistical method.

Our current feature selection is based on domain knowledge and the limitations of each techinque. For instances, the scalable clustering algorithm can only handle continuous features. It will be interesting to apply automatic feature selection methods to this health data mining problem since there are a large number of variables in our data to be explored.

7 Conclusion

Three different data mining techniques have been used to explore possible risk groups for colorectal cancer. The analysis was performed on two populations in this study. The first population comprises the population who have developed colorectal cancer during the period of study. The second population consists of patients who have not developed colorectal cancer. The analysis explored the main differences between the two populations to identify risk groups for colorectal cancer. Each technique has been applied to the two datasets with demographic and socio-economical variables and variables extracted from patients' health care history.

These heuristic results from data mining explorations may help health care professionals in identifying areas for further study of the causes and preventative factors of colorectal cancer. Typical risk groups identified for colorectal cancer have the following potential characteristics:

- Older patients.
- People living near health facilities yet seldom utilising those facilities.
- Patients with respiratory and circulatory diseases.

As mentioned before, limitation of the data (in particular the lack of lifestyle factors including diet, physical exercise, smoking, and drinking) severely limits the scope of detailed analyses. The study is not intended to identify the most important factors leading to colorectal cancer. Rather it can only explore through the variables included in the data sets.

Acknowledgements

The authors acknowledge the Australian Government Department of Health and Ageing and the Queensland Department of Health for providing data for this research. The authors also would like to thank their colleagues, Ross Sparks, Jisheng Cui and Lifang Gu, as well as Jiuyong Li of University of South Queensland and the anonymous reviewers for their comments and suggestions.

References

1. Colorectal cancer: The importance of prevention and early detection. Division of Cancer Prevention and Control, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, 2004.
2. J. Chen, H. He, G. Williams, and H. Jin. Temporal sequence associations for rare events. In *Proceedings of PAKDD04, Lecture Notes in Computer Science (LNAI 3056)*, pages 235–239, Sydney, Australia, May 2004.
3. K. J. Cios and G. W. Moore. Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26(1-2):1–24, 2002.

4. L. Gu, J. Li, H. He, G. Williams, S. Hawkins, and C. Kelman. Association rule discovery with unbalanced class. In *Proceedings of AI03, Lecture Notes in Artificial Intelligence*, pages 221–232, Perth, Western Australia, December 2003.
5. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2001.
6. H. He, J. Chen, H. Jin, S. Hawkins, G. Williams, D. McAullay, R. Sparks, J. Cui, and C. Kelman. QLDS: Colorectal cancer data mining analysis. Technical Report 04/92, CSIRO Mathematical and Information Sciences, Canberra, 2004.
7. H.-D. Jin, K.-S. Leung, M.-L. Wong, and Z.-B. Xu. Scalable model-based cluster analysis using clustering features. *Pattern Recognition*, 38(5):637–649, May 2005.
8. H.-D. Jin, W. Shum, K.-S. Leung, and M.-L. Wong. Expanding self-organizing map for data visualization and cluster analysis. *Information Sciences*, 163:157–173, Jun. 2004.
9. H.-D. Jin, M.-L. Wong, and K.-S. Leung. Scalable model-based clustering by working on data summaries. In *Proceedings of Third IEEE International Conference on Data Mining (ICDM 2003)*, pages 91–98, Melbourne, Florida, USA, Nov. 2003.
10. J. Li, H. Shen, and R. Topor. Mining the optimal class association rule set. *Knowledge-Based Systems*, 15(7):399–405, 2002.
11. D. McClisha, L. Penberthyb, and A. Pughc. Using medicare claims to identify second primary cancers and recurrences in order to supplement a cancer registry. *Journal of Clinical Epidemiology*, 56:760–767, 2003.
12. R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-00-3.
13. R. B. Rao, S. Sandilya, R. S. Niculescu, C. Germond, and H. Rao. Clinical and financial outcomes analysis with existing hospital patient records. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416 – 425, 2003.
14. J. Roddick, P. Fule, and W. Graco. Exploratory medical knowledge discovery : Experiences and issues. *SIGKDD Exploration*, 5(1):94–99, 2003.
15. A. E. Smith and S. S. Anand. Patient survival estimation with multiple attributes: adaptation of coxs regression to give an individuals point prediction. In *Proceedings of European Conference in Artificial Intelligence in Intelligent Datamining in Medicine & Pharmacology*, pages 51–54, Berlin, 2000.
16. G. I. Webb. Efficient search for association rules. In *Proceedings of SIGKDD'00*, pages 99–107, 2000.
17. G. Williams, D. Vickers, R. Baxter, S. Hawkins, C. Kelman, R. Solon, H. He, and L. Gu. The Queensland Linked Data Set. Technical Report CMIS 02/21, CSIRO, Canberra, 2002.
18. G. Williams, D. Vickers, C. Rainsford, L. Gu, H. He, R. Baxter, and S. Hawkins. Bias in the Queensland Linked Data Set. Technical Report 02/117, CSIRO Mathematical and Information Sciences, Canberra, 2002.
19. T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182, 1997.