Journal of the Korean Institute of Industrial Engineers
Vol. 42, No. 2, pp. 129-137, April 2016.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2016.42.2.129

<Application Research>

생존분석을 이용한 맞춤형 대장내시경 검진주기 추천

구자연[1] · 김은선[2] · 김성범[1†]

[1]고려대학교 산업경영공학과 / [2]고려대학교 의과대학 내과학교실

Recommendation of Personalized Surveillance Interval of Colonoscopy via Survival Analysis

Jayeon Gu[1] · Eun Sun Kim[2] · Seoung Bum Kim[1]

[1]Department of Industrial Management Engineering, Korea University
[2]Gastroenterology, Korea University College of Medicine

A colonoscopy is important because it detects the presence of polyps in the colon that can lead to colon cancer. How often one needs to repeat a colonoscopy may depend on various factors. The main purpose of this study is to determine personalized surveillance interval of colonoscopy based on characteristics of patients including their clinical information. The clustering analysis using a partitioning around medoids algorithm was conducted on 625 patients who had a medical examination at Korea University Anam Hospital and found several subgroups of patients. For each cluster, we then performed survival analysis that provides the probability of having polyps according to the number of days until next visit. The results of survival analysis indicated that different survival distributions exist among different patients' groups. We believe that the procedure proposed in this study can provide the patients with personalized medical information about how often they need to repeat a colonoscopy.

Keywords: Surveillance Interval, Survival Analysis, Patients Clustering, Kaplan-Meier Estimator, Log-Rank Test, Decision Tree, Colonoscopy

1. 서 론

IT 기술의 발달로 데이터의 양이 폭증하고 있다. 특히 헬스케어 분야에서는 건강검진자료, 질병자료, EMR(Electronic Medical Record), 유전체 분석 데이터 등 데이터가 급증하고 있다. 이런 헬스케어 데이터는 효과적인 의사결정을 목적으로 여러 가지 데이터마이닝 기법을 적용하여 연구되고 있다(Jo and Kim, 2011).

예를 들어, Goldman *et al.*(1988)은 흉통으로 응급실에 내원한 환자 중 특정 투약과 시술이 필요한 심근경색증 환자를 구분해내기 위해 의사결정나무 기법을 활용하여 9개 주요 생리적 지표와 증상을 규명하였다. 이 모델을 4,770명 코호트에 적용하여 전문 의사 판단보다 더 높은 민감도와 특이도를 보여 환자의 불필요한 입원율을 낮출 수 있음을 보여주었다. 그리고 캐나다의 온타리오 공과대학 병원에서는 인큐베이터 내의 미숙아로부터 얻은 심장 박동 및 호흡 데이터를 분석하여 병원균의 감염을 예측하고 감염징후를 조기 발견하여 의사 전달이 어려운 미숙아를 위한 진단 및 치료 시스템을 구축하였다(Bhambhri, 2011).

그 외에도, 피부 병변 이미지 데이터 분석을 통하여 피부암을 진단하거나(Burroni*et al.*, 2004), 유방 초음파검사 결과로부터 악성과 양성 종양 감별 진단하고(Joo, 2004), 운동신경 활동전위 모양과 흥분율을 통하여 신경근육계 질환을 진단한 사례가 있다(Christodoulou and Pattichis, 1999). 유전자 및 단백질 데이터로부터 얻은 혈청단백질 패턴에 의한 전립선암 조기 진단(Banez *et al.*, 2003), 유전자 표현형 패턴의 침투성 차이를 이용한 유방암

† 연락저자 : 김성범 교수, 02841 서울시 성북구 안암로 145 고려대학교 산업경영공학과, Tel : 02-3290-3397, Fax : 02-929-5888, E-mail : sbkim1@korea.ac.kr
2015년 8월 28일 접수; 2015년 11월 11일 수정본 접수; 2015년 11월 23일 게재 확정.

하위분류 감별(Perou *et al*., 2000), 중환자실 생리지표 측정과 관찰 데이터를 통하여 욕창발생 위험도를 조기에 감지하고 경고(Cho and Chung, 2011)를 가능하게 하기도 하였다.

위와 같이 헬스케어 분야에서 데이터마이닝의 활용은 의학 발전을 이끄는 필수적인 분석 기술로 자리잡았다. 특히, 의학 전분야에 걸쳐서 진단이나 예후를 예측하는 형태의 연구가 진행되었으며 그중에서도 다양한 종류의 검진 주기를 타당하게 설정하는 것에 대한 연구 또한 필요성이 증가하였다. 한국인 사망원인 1위에 해당하는 암(South Korea Statistics, 2013)은 비교적 간단한 주기적인 검진으로 발견할 수 있으며 대부분 조기에 발견할수록 완치율이 높아진다(National Cancer Center, 2011).

특히 대장암은 80% 이상이 선종-암화 과정을 통해서 5~10년에 걸쳐 암으로 진행되기 때문에 전암성 병변인 선종을 잘 찾아내고 제거하면 대장암의 발생을 줄일 수 있다(Ries *et al*., 2007). 또한, 미국 SEER(Surveillance, Epidemiology, and End Results) 프로그램에서 개발한 암의 병기분류에 따른 우리나라 대장암의 5년 생존율은 국한단계에서는 94.5%, 국소 단계는 80.4% 그리고 원격 단계는 18.6%이다(Jung *et al*., 2015). 이와 같이 대장암의 경우 진행 정도에 따라 생존율이 큰 차이를 보이므로, 증상이 나타나기 전 조기 발견 및 치료가 매우 중요하다.

대장내시경은 대장암 진단뿐 아니라, 전암 병변인 선종을 제거하여 대장암 발생과 대장암 관련 사망을 줄이는 가장 효과적인 방법으로 알려졌다(Thiis-Evensen *et al*., 1999). 또한 미국의 한 국가 연구로부터 대장내시경 검사를 통해 대장암 발생을 76~90% 정도 감소시켰다는 결과가 발표 되었다(Winawer, 1993).

그러므로 환자의 특징에 따라 맞춤형 대장내시경 검진주기를 추천하는 것은 대장암 발병률을 낮추는 데 큰 역할을 할 수 있으며 제한된 의료자원을 효율적으로 사용하기 위하여 반드시 필요하다.

따라서 본 연구에서는 환자의 건강검진 데이터를 기반으로 환자의 특징에 맞는 맞춤형 대장내시경 검진주기를 제안하고자 한다. 이를 위하여 시간에 따른 용종 진단확률을 체계적이고 과학적으로 추정할 수 있는 생존 분석을 이용하였다.

생존분석은 사건이 일어날 때까지의 시간을 대상으로 분석하는 통계적 방법으로써 사건의 발생 여부가 불확실한 중도절단된 자료를 포함하여 분석할 수 있는 특징을 가지고 있다(Hosmer *et al*., 2011).

이러한 특징으로 생존분석은 다양한 분야에서 활용되고 있는데 공학분야에서는 산업 생산품의 수명을 시간에 따라 관찰함으로써 제품의 품질과 생산성을 향상시키는데 이용되었다. 예를 들어, 디스플레이 제조 공정에서 카세트 반송시간을 생존분석 방법으로 예측하는 모형을 설계하여 반송지연을 줄임으로써 설비 비가동과 생산성도 개선 가능하게 하였다(Han, 2014). 경제분야에서는 기업의 생존함수를 추정 및 비교하고 생존함수에 영향을 끼치는 요인 크기를 추정하거나, 기업의 도산예측 모형에 적용하였다(Lee, 2010). 미국 석유정제산업에서 탈규제 이후 기업규모, 연령, 규제보조금, 기술사용 등이 기업의 생존기간에 어떤 영향을 주는지 연구하였다(Chen, 2002).

또한, 생존분석은 치료 종류에 따른 환자의 예후를 비교평가하고 그 치료 결과에 미치는 관련 인자를 규명하기 위하여 의학과 임상분야에서 널리 쓰이고 있다(Hosmer *etal*., 2011). 그 예로, Strober *et al*.(1997)는 생존분석을 이용하여 장기간에 걸친 청소년 거식증환자의 정상체중 회복 기간과 폭식의 예측인자를 확인하였다. 그리고 어린이 방광 요관 환자에 대하여 로봇 복강경을 통한 수뇨관 재이식 수술과 개방 재이식 수술의 시간에 따른 임상 및 방사선학적 성공률이 다르지 않음을 보여주기도 하였다(Patil, 2008).

본 논문의 구성은 다음과 같다. 제 2장에서는 본 연구에서 제안하는 대장내시경 검진주기 추천 방법론에 사용한 이론에 대하여 서술하였고, 제 3장에서는 실제 건강검진 데이터에 제안방법론을 적용한 결과를 제시하였다. 제 4장에서는 본 연구의 결론 및 기대효과와 함께 본 연구의 한계점에 대하여 논의하고 향후 연구 방향을 모색하였다.

2. 대장내시경 검진주기 추천 제안 방법론

본 연구에서 제안하는 대장내시경 검진주기 추천 방법은 <Figure 1>과 같이 5단계로 이루어진다. 환자의 건강검진데이터를 기반으로 환자들을 군집화하고, 각 군집별 시간에 따른 대장 용종 진단 확률을 생존분석을 통하여 추정한다. 이때, 로그순위검정법의 통계량을 이용하여 대장 용종 진단에 대한 생존함수 분포가 유사한 환자들을 재군집화하여 해당 군집 별 대장내시경 검진 시점을 추천한다. 마지막으로 의사결정나무를 이용하여 각 군집의 특징을 파악하였다.

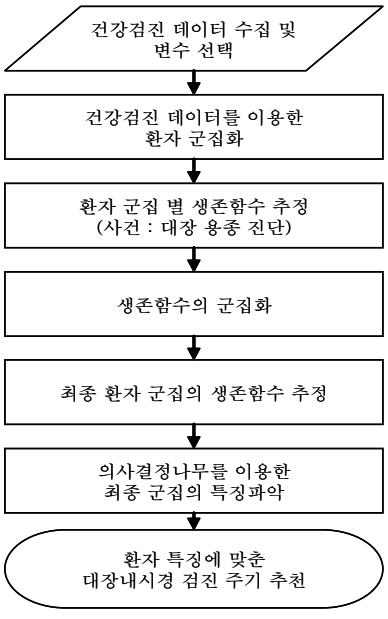

Figure 1. Process of the Recommendation of Personalized Surveillance Interval of Colonoscopy

2.1 군집분석

군집분석은 군집 내 분산은 최소화하고 군집간 분산은 최대화하여 가장 유사한 관측치끼리 그룹화하는 방법이다(Gorden, 1999). k-medoids 군집 방법은 대표적인 군집 방법 중 하나이며, k개 군집의 대표객체들(medoids)을 정하고 객체와 그가 속하는 군집의 대표 객체와의 거리의 총합을 최소로 하는 방법이다(Kaufman and Rousseeuw, 2009). 여기서 군집의 대표 객체란 그 군집에 속하는 객체 중 다른 객체와의 거리의 합이 최소가 되는 객체를 의미한다.

k-medoids 군집 방법은 객체의 위치 정보없이 객체간 거리 정보만을 이용하여 군집이 가능하다는 특징을 가지고 있다. 그러므로 수치형과 명목형 변수를 모두 가지고 있는 데이터에서 객체 사이의 거리를 정의할 수만 있다면, 혼합형 데이터 또한 군집 가능하다. 뿐만 아니라 k-medoids 방법은 이상치에 덜 민감하다는 장점이 있다(Hartigan, 1975). 환자 건강검진 데이터에는 수치형뿐만 아니라 명목형 변수도 존재하므로, 두 가지 형태의 데이터가 혼합된 경우거리 계산이 가능한 가우어 유사도를 이용(Gower, 1971)하여 k-medoids 군집 방법을 적용하였다.

본 연구에서는 k-medoids 군집 방법 중 성능이 좋다고 알려진 PAM(Partitioning Around Medoids) 알고리즘을 이용하였다(Kaufmanand Rousseeuw, 2009).

2.2 생존분석

생존분석은사건-시간분석이라고도 불리는데 이는 특정 사건이 일어날 때까지의 시간을 대상으로 분석하는 통계 방법이기 때문이다. 즉, 생존분석에서 생존기간이란 어떤 사건이 일어날 때까지의 시간을 의미한다.

*T*를 생존시간이라고 할 때, *t*라는 시점에서 생존함수인 *S*(*t*)는 *t*시점까지 사건이 발생하지 않고 생존할 확률의 추정량이며 다음 식 (1)과 같이 표현된다.

$$S(t) = \Pr(T > t) \tag{1}$$

$S(t)$는 누적생존율이라고도 불린다. 이를 시간에 따라 그린 그래프가 생존곡선이며 <Figure 2>와 같은 형태로 나타난다(Hosmer *et al*., 2011). 예를 들어, <Figure 2>가 "대장 용종의 진단"이라는 사건에 대한 누적생존율이라고 가정하면, 1년 후 대장 용종이 진단되지 않을 확률은 80%이다. 즉, 1년 후 대장 용종으로 진단될 확률은 20%라고 해석할 수 있다. 생존곡선이 급한 경사를 보이면 생존율이 낮고, 완만하고 평평하면 생존율이 높고 오래 생존한다는 것을 의미한다(Ziegel, 1997).

본 연구에서는 생존함수를 추정하기 위하여, 카플란-마이어법을 활용하였다(Kaplan and Meier, 1958). 이는 특정한 분포를 가정하지 않고 생존기간을 추정하는 비모수적 방법으로 생존시간에 대한 자료의 분포를 가정할 필요가 없다. 또한, 카플란 - 마이어법을 이용하여 두 집단간의 생존율 차이 유무를 통계적으로 판단할 수 있다.

카플란-마이어법은 관찰된 생존시간을 관찰기간이 짧은 순서에서 긴 순서로 다시 배열한 후 생존율을 계산하며, 이는 다음 식 (2)로 정의된다.

$$\hat{S}(t) = \prod_{i=1}^{t} \frac{(n_i - d_i)}{n_i} \tag{2}$$

여기서 n_i는 t_i시점에 사건 발생 위험이 있는 대상의 수, d_i는 t_i시점에 사건이 발생한 대상의 수를 나타낸다. 즉, t에서 추정된 생존함수 $\hat{S}(t)$는 t보다 작거나 같은 시점에 모든 대상자의 생존시간을 곱하여 추정된다(Kaplan and Meier, 1958). $\hat{S}(t)$의 분산은 테일러 전개에 기초한 델타방법(delta method)으로부터 유도된 Greenwood 공식과 log-log 변환을 이용하여 다음과 같이 식 (3)과 같이 추정한다(Kalbfleischand Prentice, 2011).

$$\widehat{Var}\{\ln[-\ln(\hat{S}(t))]\} = \frac{1}{[\ln(\hat{S}(t))]^2} \sum_{t_1 \le t} \frac{d_i}{n_i(n_i - d_i)} \tag{3}$$

이를 통해, 생존율 및 생존기간의 신뢰구간 또한 추정 가능하게 한다.

본 연구에서 사건은 용종의 진단이며, 생존기간은 용종 진단까지 걸리는 시간을 의미한다.

2.3 로그 순위 검정법

로그 순위 검정법은 독립된 두 군 또는 여러 군의 생존확률을 총괄적으로 비교하는데 많이 쓰이는 비모수적 검정법이다. 이는 관찰대상 개개인을 관찰 순서대로 배열하고 두 군에서 사건이 발생한 시점에서 관찰된 사건의 수와 기대 사건 수의 차이를 이용하여 계산된다(Hosmer *et al*., 2011). 두 군의 생존확률 S_1과 S_2를 비교하는 로그 순위 검정법에 대한 귀무가설과 대립가설은 식 (4)과 같다.

$$\mathrm{H}_0 : S_1(t) = S_2(t) \ vs \ \mathrm{H}_1 : S_1(t) \neq S_2(t) \tag{4}$$

위 가설 검정의 검정통계량 Q는 식 (5)로 정의된다.

$$Q = \frac{\left[\sum_{i=1}^{m}(d_{1i} - \widehat{e_{1i}})\right]^2}{\sum_{i=1}^{m} \widehat{v_{1i}}} \tag{5}$$

여기에서 각 변수는 다음 <Table 1>과 같다. *i*번째 사건이 발생하였을 때, d_{1i}는 비교 집단(Group 1)에서 발생한 사건수를 의미한다. 그리고 $\hat{e}_{1i}$는 비교 집단의 기대 사건 수이며, 식 (6)와 같이 정의된다.

$$\hat{e}_{1i} = \frac{n_{1i} \cdot d_i}{n_i} \tag{6}$$

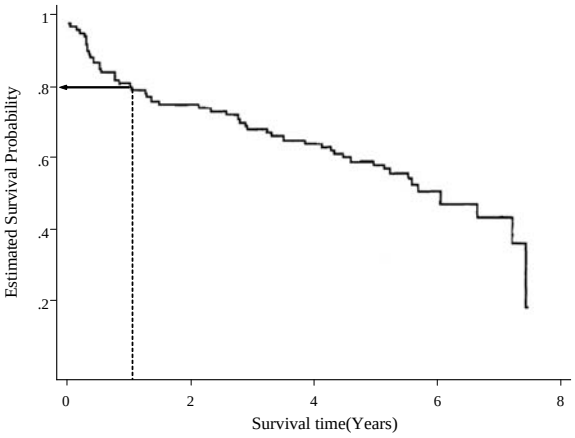

Figure 2. Example of the Survival Function

$\widehat{v_{1i}}$는 d_{1i}의 분산이고, d_{1i}는 초기하분포를 따른다고 가정하여, 식 (7)과 같이 계산된다.

$$\widehat{v_{1i}} = \frac{n_{1i} \cdot n_{0i} \cdot d_i \cdot (n_i - d_i)}{n_i^2 (n_i - 1)} \tag{7}$$

로그 순위 검정법의 검정통계량 Q는 자유도가 1인 카이제곱분포를 따르므로, 해당 검정법의 p-value는 아래 식 (8)과같이 계산되며, 이는 두 집단의 생존함수의 차이가 통계적으로 유의한지 여부를 알 수 있는지표가 된다(Mantel, 1996). 일반적으로 p-value가 0.05보다 작을 때, 두 집단의 생존함수는 차이가 유의 하다고 판단할 수 있다(Hosmer *et al*., 2011).

$$p-\text{value} = \Pr(\chi^2(1) > Q), \;\; Q \sim \chi^2(1) \tag{8}$$

본 연구에서는 해당 검정 결과의 p-value를 용종 진단에 대한 생존분포가 유사한 환자 군집들을 하나의 군집으로 다시 묶을지를 판단하는데 사용한다.

Table 1. Table Used for Test of Equality of Survival Function in Two Groups at Observed Survived Time t_i

Event/Group	1	0	Total
Die(Event = 1)	d_{1i}	d_{0i}	d_i
Not Die(Event = 0)	$n_{1i} - d_{1i}$	$n_{0i} - d_{0i}$	$n_i - d_i$
At Risk	n_{1i}	n_{0i}	n_i

2.4 의사결정나무

의사결정나무는 의사결정규칙을 도표화하여 관심대상이나 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법이다. 분석의 과정이 나무구조에 의해서 룰 형식으로 표현되기 때문에 분석과정을 쉽게 이해하고 설명할 수 있는 장점을 가지고 있다(Curram and Mingers, 1994). 따라서 분석의 정확도보다 분석 과정에 대한 설명이 필요한 경우에 보다 유용하게 사용되고 있다(Choi, 1998). 그뿐만 아니라 독립변수에 수치형과 명목형이 섞여 있는 경우에도 별도의 변환 작업 없이 사용할 수 있다는 장점이 있다(Hastie *et al*., 2001).

본 연구에서는 CART 알고리즘을 사용하였다. 이는 의사결정트리 방법론 중 가장 잘 알려진 방법론 가운데 하나이며(Berry and Linoff, 1997), 지니 지수 또는 분산의 감소량을 사용하여 나무의 가지를 이진 분리한다(Breiman *et al*., 1984).

3. 건강검진 데이터를 이용한 분석 결과

3.1 데이터 설명

2004년부터 2009년 사이 고려대학교 안암병원에 2회 이상 내원하여 건강검진을 받은 총 824명의 환자 데이터를 분석에 사용하였다. 먼저 결측치가 총 데이터의 50% 이상인 변수는 제거하고, 남은 변수 중 결측치가 존재하는 환자들은 삭제하여 총 625명의 데이터로 분석을 시행하였다. 변수는 환자의 기초정보와 대장내시경 결과를 포함한 종합건강검진검사로부터 수집된 데이터로 구성되었으며 수치형 변수는 25개, 명목형 변수는 7개이다.

<Table 2>는 본 연구에서 사용한 데이터의 통계 분석 결과이다. 환자들의 기초정보는 나이, 혈압, BMI, 키, 몸무게 등의 변수가 있다. 환자들의 평균 나이는 52세이며, 남성이 426명(68%), 여성이 199명(32%)의 비율로 구성되었다. 건강검진데이터는 채뇨와 채혈을 통해 얻을 수 있는 기본적인 데이터로 헤모글로빈, 혈침, 중성구, 임파구, 호산구, 콜레스테롤, 중성지방, 공복시혈당, 혈중요소질소, 크레아티닌, 인, 칼슘, C-반응성단백, CA19-9과 인슐린 수치가 있다. 그리고 대장내시경 결과와 관련된 변수는 1차 검진 이후 2차 검진까지의 경과 시간을 나타내는 검진 주기, 1차 검진에 과증식성 용종, 선종성 용종 및 용종의 진단 여부와 2차 검진에서의 용종의 진단 여부가 있다. 1차 건강검진에서 대장내시경으로 용종이 1개 이상 진단된 환자는 176명(28%)이며, 이때 진단된 대부분의 용종은 제거되었다. 그 후, 2차 검진에서 용종이 진단된 환자는 190명(30%)이다.

3.2 환자 군집화

환자의 1차 건강검진데이터를 이용하여, k-medoids 군집화를 진행하였다.

k-medoids 군집기법을 적용하기 위해서는 객체 간 유사도를 정의해야 하는데, 본 연구에서는 가우어 유사도를 사용하였다(Gower, 1971). 객체 i와 객체 j의 공통변수를 $x = (x_{i1}, x_{i2}, \cdots, x_{in})$로 정의 할 때, 객체 i와 객체 j의 가우어유사도 $S(i, j)$는 식 (9)와 같이 정의된다.

Table 2. Description of Comprehensive Medical Test Data

Numerical Variable		Categorical Variable	
	mean±SD		count
Age(나이), yr	51.65±8.75	Gender(성별)	
Blood pressure-contraction(혈압-수축기), mmHg	119.11±11.08	Female	199
Blood pressure-relaxation(혈압-이완기), mmHg	72.64±9.29	Male	426
BMI(체질량지수), kg/m^2	24.34±2.85	Stress(스트레스)	
Height(키), cm	165.42±8.01	Need interview	44
Weight(체중), kg	66.8±10.63	Normal	581
Hemoglobin(헤모글로빈), g/dl	14.75±1.27	ABO Typing(혈액형)	
Erythrocytesedimentation rate(혈침), mm/hr	6.42±6.24	A	183
Neutrophil(중성구), %	53.58±8.49	AB	91
Lymphocyte(임파구), %	35.52±7.77	B	185
Eosinophil(호산구), %	3.39±2.66	O	166
Total cholesterol(총콜레스테롤), mg/dl	188.27±33.36	Hyperplastic Polyps_1st(1차 검진의 과증식성 용종)	
HDL-cholesterol(고비중콜레스테롤), mg/dl	53.53±13.86	0	562
LDL-cholesterol(저비중콜레스테롤), mg/dl	118.44±29.3	≥ 1	63
Triglyceride(중성지방), mg/dl	137.02±94.94	Adenoma_1st(1차 검진의 선종성 용종)	
Fasting blood glucose(공복혈당), mg/dl	94.14±13.3	0	414
BUN(혈중요소질소), mg/dl	13.1±3.28	≥ 1	211
Creatinine(크레아티닌), mg/dl	0.96±0.17	Polyps_1st(1차 검진의 용종 진단)	
P(인), mg/dl	3.47±0.66	0	449
CA(칼슘), mg/dl	9.39±0.45	≥ 1	176
C-reactive protein(C-반응성단백), mg/dl	1.43±3.37	Polyps_2nd(2차 검진의 용종 진단)	
CA19-9(암항원 19-9), U/ml	12.52±6.24	0	435
Insulin(인슐린), ulU/ml	8.81±3.41	≥ 1	190
Test interval(검진 주기), day	775.79±382.92		

$$S(i,j)=\frac{\sum_{k=1}^{n}\delta_{ijk}d_{ijk}}{\sum_{k=1}^{n}\delta_{ijk}} \tag{9}$$

여기서 δ_{ijk} 는 객체 i와 객체 j의 k번째 변수의 값이 비교 가능한 경우(= 1)를 정의하기 위한 변수이다. 즉 비교가 불가능한 경우, 0으로 계산된다. d_{ijk}는 변수의 특징에 따라 다르게 정의되는데 명목형 변수일 경우, 식 (10)과 같이 정의된다.

$$d_{ijk}=\begin{cases}0, & \text{if } x_{ik}=x_{jk},\\ 1, & otherwise\end{cases} \tag{10}$$

반면, 수치형 변수일 경우 식 (11)과 같이 정의하며, R_k는 k번째 변수의 범위이다.

$$d_{ijk}=\frac{|x_{ik}-x_{jk}|}{R_k} \tag{11}$$

다음으로 k-medoids 군집기법을 적용하기 위해서는 사용자가 군집의 개수인 k 또한 사전에 결정해야 하는데, 본 연구에서는 최적의 군집 개수를 정하는 방법의 하나인 실루엣 통계량 기법을 이용하여 군집의 개수를 정하였다(Rousseeuw, 1987). 실루엣통계량은 개체들이 군집에 얼마나 잘 군집화 되었는지를 측정해주는 값으로 -1에서 1사이의 값으로 나타나며, 큰 값이 나올수록 군집화가 잘 되었다고 볼 수 있다. 본 연구에서는 군집의 개수 k를 2부터 차례대로 변경하여 군집화를 실행하고, 최대의 실루엣 값을 가지는 k를 군집의 개수로 정하였다. <Figure 3>은 k의 변화에 따른 실루엣 통계량 값을 보여준다. 군집수가 2개일 때 실루엣 통계량이 가장 높으며 군집 수가 13개일 때, 두 번째로 높음을 알 수 있다. 비록 실루엣 통계량 값이 k = 2일 때 가장 높게 나왔으나 환자의 군집을 2개로 정했을 경우 너무 단순화될 수 있다고 판단하여 군집 수를 13개로 결정하였다.

<Figure 4>는 환자 건강검진 데이터를 k-medoids 군집화 기법을 적용하여 13개의 군집으로 분류한 결과를 다차원 척도법(Multiscale dimensional scaling)을 적용하여 시각화한 그림이다. 다차원 척도법은 고차원의 데이터를 축소하여 사람이 인식할 수 있는 2차원 또는 3차원 공간으로 시각화하는 위상 보

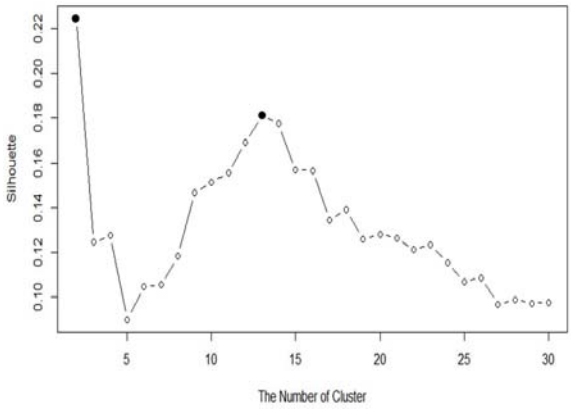

Figure 3. Silhouette Plot to Determine the Number of Clusters

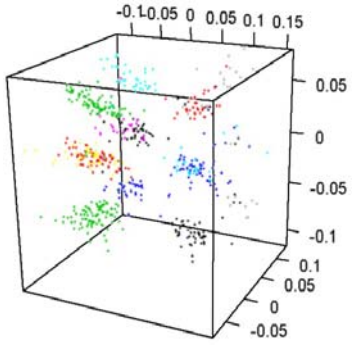

Figure 4. 3D MDS Plot of Patients Clustering

존 알고리즘 중 하나이다(Bender*et al.*, 2000). 하나의 점은 한 명의 환자를 의미하며, 점 사이의 거리는 환자간의 유사성을 의미한다(Borg and Groenen, 2005). 점의 색깔은 군집화 결과를 나타낸다. 실제 거리상 가까운 점들끼리 같은 색상을 띄고 있는 것을 보아, 유사한 정보를 가진 환자들끼리 같은 군집에 속해 있는 것을 확인할 수 있다.

3.3 각 군집 별 용종 진단에 대한 생존분석 결과

<Figure 5>는 카플란-마이어법을 적용하여 13개 군집마다 용종을 진단되기까지의 기간을 추정한 결과를 보여주고 있다. <Figure 5>의 *x*축은 1차 건강검진 이후 경과 시간이고, *y*축은 (1-누적생존확률) 즉, 용종 진단 확률을 의미한다. 예를 들어, 1차 검진 이후 500일이 지났을 때 가장 높은 용종 진단 확률을 보이는 군집 6에 속하는 환자는 1차 건강검진 이후 500일이 지난 시점에 용종이 진단될 확률은 40%라고 해석할 수 있다. 또한, 13개의 그래프는 서로 비슷한 확률분포를 가지는 군집들이 존재함을 알 수 있다.

3.4 생존함수 분포를 이용한 환자의 군집화

생존함수 분포가 비슷한 군집을 재군집하기 위하여, 환자 군

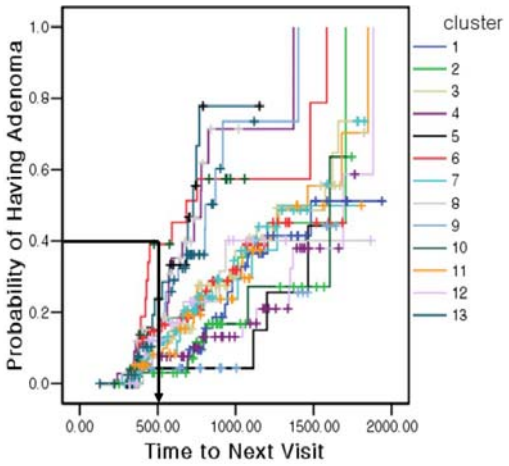

Figure 5. 1-Survival Functions of 13 Patients Cluster

집 간의 거리를 군집간 로그 순위 검정법의 1-*p*-value로 정의하였는데 이는 두 군집의 생존함수의 차이가 클수록 로그 순위 검정법의 1-*p*-value가 크다고 해석할 수 있기 때문이다. 따라서 본 연구에서는 로그순위 검정법의 1-*p*-value를 기반으로 k-medoids 군집화 기법을 이용하여 환자 재군집화를 수행하였다.

군집의 개수는 실루엣 통계량이 최대값이 되는 $k = 3$으로 결정하였다. 환자 재군집화를 통하여 얻은 3개의 군집에 속하는 환자의 데이터를 통합하여 생존함수를 다시 추정한 결과를 <Figure 6>에 나타내었다. <Figure 6>에서, 가장 급한 경사로 용종진단 확률이 증가하는 SF_3 군집은 다른 군집에 비하여 빨리 용종이 진단되고, 가장 완만한 경사를 보이는 SF_1 군집은 다른 군집에 비하여 느리게 용종이 진단되는 것을 알 수 있다.

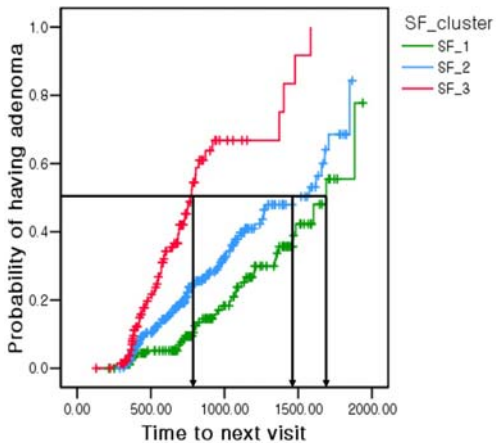

Figure 6. 1-Survival Functions of New Cluster

또한, <Table 3>에서 재군집화 된 각 군집의 용종 진단 확률이 25%, 50% 그리고 75%인 시점을 추정한 결과를 나타내었다.

13개의 군집을 재군집화 한 범주로 나누어 생존함수를 그린 결과는 <Figure 7>에서 보여주고 있다. 재군집화를 통하여 생성된 군집에 속하는 생존함수들은 서로 유사한 생존분포를 가지는 것을 확인할 수 있다.

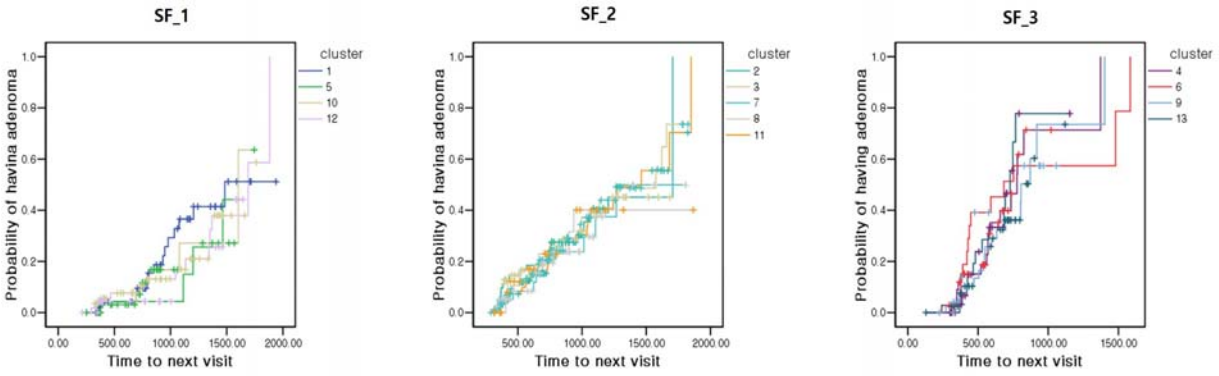

Figure 7. Clustering Results Based on Distribution of Survival Functions

Table 3. Quantile of Having Adenoma by Final Cluster

Cluster	n	Quantile			Risk
		25%	50%	75%	
SF_1	169	1,114	1,690	1,883	Low
SF_2	317	793	1,462	1,849	Middle
SF_3	139	546	767	1,372	High

3.5 환자 군집 별 검진 주기 제안

본 연구에서는 용종 진단 확률이 50%인 시점을 다음 대장내시경 검진일로 제안하고자 한다. 최종적으로 얻은 총 3개의 환자 군집에 대한 추천 주기는 <Table 4>에 보다 이해가 쉬운 월 단위로 표시하였다. 같다. 추후 새로운 환자가 대장내시경을 포함한 종합건강검진을 시행할 때, 어떤 군집에 속할지 확인 후, <Table 4>를 참고하여 검진 주기를 적용할 수 있을 것이다. 예를 들어, SF_1 군집에 해당하는 환자는 55.4개월(4년 7개월)을 2차 검진 주기로 추천하며 이에 대한 95% 신뢰구간은 50.4개월(4년 2개월)에서 60.4개월(5년)이다. SF_1 저위험군에 속하는 군집의 추천 주기는 4년 7개월~ 5년으로 의료계에서 50세 이상에게 일반적으로 권고하는 대장내시경 검진주기인 5~10년(Lee *et al*., 2002)과 어느 정도 부합한다. 각 군집의 특징은 의사결정나무를 통하여 확인할 수 있다.

Table 4. Recommendation of Surveillance Interval of Colonoscopy

Cluster	Interval Suggestions (month)	Confidence Interval (month)
SF_1	55.4	(50.4, 60.4)
SF_2	48.0	(43.2, 52.7)
SF_3	25.1	(24.1, 26.2)

3.6 환자 군집의 특징 파악

<Figure 8>은 이전 단계에서 최종 생성한 3개 환자 군집의 특징을 파악하기 위하여, 건강검진 데이터 정보를독립변수로 환자군집화 결과를 종속변수로 활용하여 의사결정나무 모델을 구축한 결과이다. CART 소프트웨어(http://www.salford-systems.com/products/cart)를 이용하여 결과를 생성하였으며 이를 직접 그림으로 표현하였다. 결과를 통하여 성별, 1차 검진 시 선종성 용종의 진단 여부, LDL(저밀도 콜레스테롤) 변수가 군집을 분류하는 기준으로 중요하게 작용함을 알 수 있었다.

의사결정나무의 첫 번째 노드는 성별에 의하여 나뉜다. 남성 환자는 고위험군(SF_3)과 중위험군(SF_2)으로만 구성되어, 저위험군(SF_1)에 속하는 환자는 없다고 해석할 수 있다. 반면, 여성 환자 대부분은 저위험군과 중위험군에 속하는 데, 이는 남성 환자가 여성 환자보다 더 짧은 대장내시경 검진 주기를 가져야 한다는 의미를 내포한다.

그리고 남성 환자는 1차 검진 시 선종이 진단을 받을 시 고위험군으로, 진단받지 않은 경우 중위험군으로 99% 이상 정확히 분류된다. 반면, 여성 환자는 선종의 진단 여부에 따라 중위험군과 저위험군으로 구분되긴 하지만 남성 환자만큼 높은 정확도에 의하여 구분되지 않는다. 여성 환자도 남성 환자와 비슷하게 1차 검진 시 선종을 진단받으면 중위험군, 진단받지 않으면 저위험군으로 분류되어 1차 검진 시 선종을 진단받을 경우 더 빠른 검진 주기를 추천받아야 한다. 하지만 남성과 달리 1차 검진 시 선종을 진단을 받은 경우에도, LDL(저밀도 콜레스테롤)이 112mg/dl보다 낮으면 1차 검진 시 선종을 진단받지 않은 환자와 동일한 검진 주기를 추천받아도 무방하다고 해석할 수 있다.

4. 결 론

이제까지 대장내시경의 검진 주기는 환자의 성별이나 나이 혹은 신체 상태와 무관하게 일관적으로 제시되어 왔다.본 연구에서는 대장내시경을 포함한 종합건강검진 결과 데이터를 기반으로, 환자의 특성에 맞추어 맞춤형 대장내시경 검진 주기를 제안할 수 있는 방법론을 소개하였다.

종합건강검진 결과 데이터를 이용하여 환자를 군집화한 후,

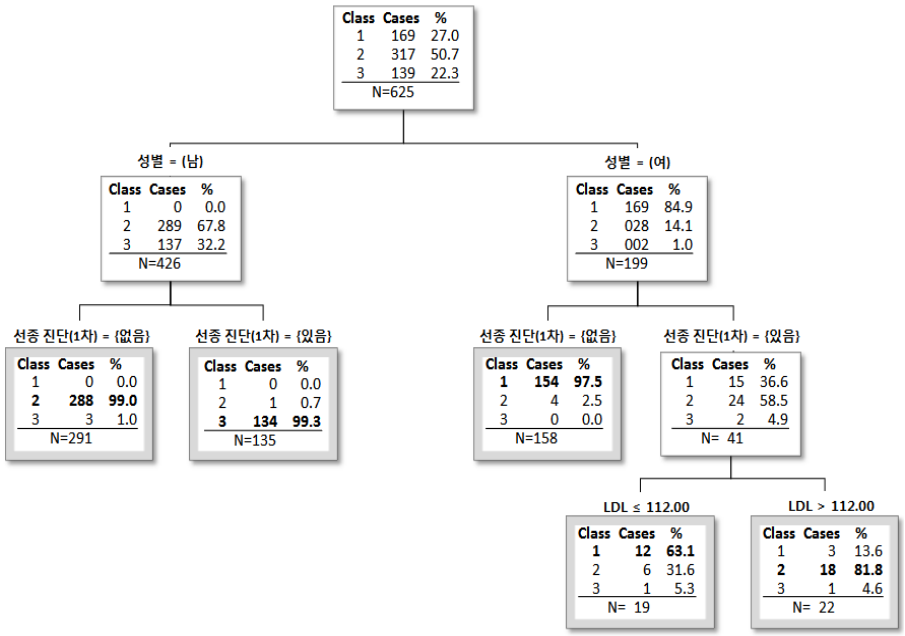

Figure 8. Decision Tree to Find Out the Characteristics of Final Clusters

군집 별 용종의 진단 사건에 대한 생존함수 분포가 유사한 군집들끼리 재군집화를 실시하여, 궁극적으로 시간에 따른 용종의 진단 확률이 유사한 환자들을 군집화 할 수 있었다. 최종 환자 군집을 이용하여, 용종의 진단까지 걸리는 시간을 추정하였으며 이 정보는 추후 새로운 환자의 대장내시경 검진주기를 신체 정보에 따라 추천하는데 사용할 수 있을 것이다. 본 연구에서는 용종 진단 확률이 50%인 시점을 다음 건강검진일로 제안하였지만, 이는 사용자가 원하는 용종 진단 확률에 맞게 검진 주기를 계산할 수 있다.

또한 본 연구에서는 대장내시경에 대한 검진주기에 초점을 맞추었지만 검진주기의 결정이 필요한 다른 분야에도 제안 방법론의 적용이 가능할 것이다.

한편, 본 연구에서는 약 5년간의 데이터를 사용하였으므로 용종에 대한 추적기간이 최대 5년이다. 용종 진단에 대한 대부분의 연구가 최대 추적 기간이 5년이기는 하지만(Hong *et al*., 2012), 용종 진단에 대한 저위험군의 경우 5년보다 더 긴 기간을 추적해야 더 의미 있는 분석이 가능할 것으로 사료된다.또한, 본 연구에서 대장내시경 검사 결과에 관련된 변수는 용종의 진단 여부, 과증식성 용종의 유무, 선종의 유무 3가지만 사용하였지만 용종의 모양, 크기 그리고 개수 정보도 대장내시경 검진 주기 제안 시 중요한 변수로 작용한다고 알려졌으므로(Lieberman *et al*., 2012) 해당 변수들을 추가한 분석이 진행되어야 할 것으로 보인다.

참고문헌

Bhambhri, A. (2011), *Smarter Analytics for Big Data*, IBM.

Banez, L. L., Prasanna, P., Sun, L., Ali, A., Zou, Z., Adam, B. L., and Srivastava, S. (2003), Diagnostic potential of serum proteomic patterns in prostate cancer, *The Journal of urology*, **170**(2), 442-446.

Bender, M., Klein, R., Disch, A., and Ebert, A. (2000), A functional framework for web-based information visualization systems, *Visualization and Computer Graphics, IEEE Transactions*, **6**(1), 8-23.

Berry, M. J. and Linoff, G. (1997), *Data mining techniques : for marketing, sales, and customer support*, John Wiley and Sons, Inc.

Borg, I. and Groenen, P. J. (2005), *Modern multidimensional scaling : Theory and applications*, Springer Science and Business Media.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and regression trees*, CRC press.

Burroni, M., Corona, R., Dell'Eva, G., Sera, F., Bono, R., Puddu, P., and Rubegni, P. (2004), Melanoma computer-aided diagnosis reliability and feasibility study, *Clinical cancer research*, **10**(6), 1881-1886.

Chen, M.Y. (2002), Survival duration of plants : evidence from the US petroleum refining industry, *International Journal of Industrial Organization*, **20**(4), 517-555.

Cho, I. S. and Chung, E. (2011), Predictive bayesian network model using electronic patient records for prevention of hospital-acquired pressure ulcers, *Journal of Korean Academy of Nursing*, **41**(3), 423-431.

Choi, J., Han, S., Kang, H., and Kim, E. (1998), Data mining decision tree analysis using answer tree, *SPSS Academy*, 17-23.

Christodoulou, C. and Pattichis, C. S. (1999), Unsupervised pattern recognition for the classification of EMG signals, *Biomedical Engineering, IEEE Transactions*, **46**(2), 169-178.

Curram, S. P. and Mingers, J. (1994), Neural networks, decision tree induction and discriminant analysis : An empirical comparison, *Journal of the Operational Research Society*, **45**(4), 440-450.

Goldman, L., Cook, E. F., Brand, D. A., Lee, T. H., Rouan, G. W., Weisberg, M. C., and Jakubowski, R. (1988), A computer protocol to predict myocardial infarction in emergency department patients with chest pain, *New England Journal of Medicine*, **318**(13), 797-803.

Gorden, A. D. (1999), *Classification*, Chapman and Hall/CRC.

Gower, J. C. (1971), A general coefficient of similarity and some of its

properties, *Biometrics*, **27**(4), 857-871.
Hastie, T., Friedman, J., and Tibshirani, R. (2001), *The elements of statistical learning*, Springer.
Han, P. and Baek, J. G. (2014), Prediction model on delivery time in display FAB using survival analysis, *Journal of the Korea Institute of Institute of Industrial Engineers*, **40**(3), 283-290.
Hartigan, J. A. (1975), *Clustering algorithms*, John Wiley and Sons, Inc.
Hong, S. N., Yang, D. H., Kim, Y. H., Hong, S. P., Shin, S. J., Kim, S. E., and Yang, S. K. (2012), Korean guidelines for post-polypectomycolonoscopic surveillance, *The Korean Journal of Gastroenterology*, **59**(2), 99-117.
Hosmer Jr, D. W., Lemeshow, S., and May, S. (2011), *Applied survivalanalysis : regression modeling of time to event data*, Wiley.com.
Jo, I. and Kim, J. (2011), Trend research-based clinical decision support systems based on Electronic Health Records, *Communications of the Korean Institute of Information Scientists and Engineers*, **29**(2), 92-100.
Joo, S., Yang, Y. S., Moon, W. K., and Kim, H. C. (2004), Computer-aided diagnosis of solid breast nodules : use of an artificial neural network based on multiple sonographic features, *Medical Imaging, IEEE Transactions*, **23**(10), 1292-1300.
Jung, K. W., Won, Y. J., Kong, H. J., Oh, C. M., Cho, H., Lee, D. H., and Lee, K. H. (2015), Cancer statistics in Korea : incidence, mortality, survival, and prevalence in 2012, *Cancer research and treatment : official journal of Korean Cancer Association*, **47**(2), 127.
Kalbfleisch, J. D. and Prentice, R. L. (2011), *The statistical analysis of failure time data*, John Wiley and Sons.
Kaplan, E. L. and Meier, P. (1958), Nonparametric estimation from incomplete observations, *Journal of the American statistical association*, **53**(282), 457-481.
Kaufman, L. and Rousseeuw, P. J. (2009), *Finding groups in data : an introduction to cluster analysis*, John Wiley and Sons.
Lee, B. and Jung, S. (2002) Korean National Guidelines on Screening and Surveillance for Early Detection of Colorectal Cancers (KSCP and NCC), *Korean Society of Gastointestinal Endoscopy*, **45**(8), 981-891.
Lee, Y. (2010), Study on Prediction Model of insolvent companies using survival analysis techniques Guarantee, *Korean Market Economy Research*, **39**(3), 1-24.
Lieberman, D. A., Rex, D. K., Winawer, S. J., Giardiello, F. M., Johnson, D. A., and Levin, T. R. (2012), Guidelines for colonoscopy surveillance after screening and polypectomy : a consensus update by the US Multi-Society Task Force on Colorectal Cancer, *Gastroenterology*, **143**(3), 844-857.
Mantel, N. (1966), Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer chemotherapy reports*, **50**(3), 163-170.
National Cancer Center (2011), National Cancer Control Project, Available at : http://www.ncc.re.kr/manage/manage12_00.jsp.
Patil, N. N., Mottrie, A., Sundaram, B., and Patel, V. R. (2008), Robotic-assisted laparoscopic ureteral reimplantation with psoas hitch: a multi-institutional, multinational evaluation, *Urology*, **72**(1), 47-50.
Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., and Botstein, D. (2000), Molecular portraits of human breast tumours, *Nature*, **406**(6797), 747-752.
Ries, L. A. G., Melbert, D., and Krapcho, M. (2007), SEER Cancer Statistics Review, 1975~2004, *Bethesda, MD: National Cancer Institute, based on November 2006 SEER data submission, posted to the SEER Web site*.
Rousseeuw, P. J. (1987), Silhouettes : a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics*, **20**, 53-65.
South Korea Statistics (2013), *Year mortality statistics*.
Strober, M., Freeman, R., and Morrell, W. (1997), The long-term course of severe anorexia nervosa in adolescents : Survival analysis of recovery, relapse, and outcome predictors over 10~5 years in a prospective study, *International Journal of Eating Disorders*, **22**(4), 339-360.
Thiis-Evensen, E., Hoff, G. S., Sauar, J., Langmark, F., Majak, B. M., and Vatn, M. H. (1999), Population-based surveillance by colonoscopy : effect on the incidence of colorectal cancer : Telemark Polyp Study I, *Scandinavian journal of gastroenterology*, **34**(4), 414-420.
Winawer, S. J., Zauber, A. G., Ho, M. N., O'Brien, M. J., Gottlieb, L. S., Sternberg, S. S., and Stewart, E. T. (1993), Prevention of colorectal cancer by colonoscopic polypectomy, *New England Journal of Medicine*, **329**(27), 1977-1981.
Ziegel, E. R. (1997), Survival analysis using the SAS system, *Technometrics*, **39**(3), 344.