보건학 석사 학위논문

은닉마코프모델을 이용한 당뇨병의 생애 위험도 추정

Lifetime Risk Estimation of Diabetes Mellitus using Hidden Markov Model

2014년 2월

서울대학교 보건대학원

보건학과 바이오인포매틱스 전공

윤 재 문

은닉마코프모델을 이용한 당뇨병의 생애 위험도 추정

Lifetime Risk Estimation of Diabetes Mellitus using Hidden Markov Model

지도교수 손 현 석

이 논문을 윤재문 석사학위논문으로 제출함

2013년 10월

서울대학교 보건대학원

보건학과 바이오인포매틱스 전공

윤 재 문

윤재문의 석사학위 논문을 인준함

2013년 12월

위 원 장 정 해 원 (인)

부 위 원 장 조 성 일 (인)

위 원 손 현 석 (인)

국 문 초 록

높은 유병률과 사망에 대한 기여로 인하여 당뇨병의 중요도는 점차 증가하고 있으며, 이에 대한 대중의 관심 또한 늘어날 것으로 기대하고 있다. 때문에 당뇨병에 대한 예측 모형과 평생 당뇨병 위험도는 일반인과 의료인에게 유용한 정보를 제공할 수 있을 것이다. 국외에는 당뇨병의 평생 위험도를 추정한 연구가 보고 되어 있으나 개인별 위험요인을 충분히 반영하지 못하였으며, 국내에서는 이와 같은 연구가 이루어 진 적이 없다. 따라서, 본 연구는 개인별 위험요인에 따라 당뇨병의 평생 위험도를 추정하는 모형을 제시하고, 제시된 예측 모형의 타당도를 평가하고자 수행 하였다. 당뇨병 위험도 추정을 위해 은닉마코프모델(HMM)을 기반으로 한 모형을 이용하였다. 생물학적 고혈당 여부를 상태열로, 당뇨병의 인지 여부를 관측열로 가정하였고 사망률은 통계청의 자료로부터 추출하였다. 국민건강영양조사 5기 자료를 1:1로 무작위 분할하여 “training set”과 “validation set”으로 이용하였다. 당뇨병의 유병률과 위험요인에 대한 자료들은 “training set”으로부터 계산하였다. HMM의 기초 모수들은 사망률, 당뇨병의 유병률, 당뇨병 인지확률로부터 계산하였다. 기초 모수들은 위험요인에 따른 오즈비에 따라 보정하여 최종 모형을 구성하였다. “validation set”에 본 모형을 적용하여 개인별 위험도를 추정하고 타당도를 평가하였다. 첫 번째로, 성별 및 연령에 따라 표준 위험도를 갖고 있는 대상자의 위험도를 추정하였다. 추정된 기대여명은 기존의 산출 방법에 비해 과소평가되었으나, 그 차이는 초기 연령과 무관하게 일정한 수준이었다. 두 번째로, 개인별 위험요인에 따른 위험도 평가를 시행하였다. 위험도에 따른 예측 유병률은 한국인 당뇨병 예측 점수보다 더 높은 예측력을 보였다. 당뇨병의 인지기간 및 비인지기간의 기대값은 가용한 자료가 부재하여 타당성 평가를

수행하지 못하였다. 몇 가지 제한점이 있음에도 불구하고, 본 연구의 예측 모형은 새로운 형태의 결과값을 포함하고 있었고, 기존의 예측도구에 비해서도 좋은 결과를 제시 하였으며, 새로운 위험요인에 대해서도 쉽게 확장이 가능하다는 장점을 보여 주었다. 이 모형은 일반인들에게 당뇨병 위험인자를 교정하는 동기를 부여하는 데에 도움을 주고, 전문가들에게는 의학적 결정을 하는 데 있어서 가치 있는 정보를 제공할 수 있을 것이라 기대한다.

주요어: 당뇨병, 평생 위험도, 은닉마코프모델, 개인별 위험도, 타당성 검토

학 번: 2011-22112

목 차

List of Tables

List of Figures

제 1장. 서 론

1.1 당뇨병(Diabetes mellitus)

당뇨병은 내분비 질환의 하나로서 인슐린 분비 장애나 인슐린 저항성으로 인한 고혈당 상태를 말한다(American Diabetes, 2013). 일반적으로 당뇨병은 발병 기전에 따라 분류하고 있으나, 그 분류는 지속적으로 변화하고 있다. 최근의 분류에 따르면, 당뇨병은 크게 1형 당뇨병(T1DM: Type 1 diabetes mellitus), 2형 당뇨병(T2DM: Type 2 diabetes mellitus) 그리고 기타 당뇨병(Other specific type)으로 구분된다(American Diabetes, 2013). T1DM은 췌장의 베타세포 파괴에 의한 인슐린 결핍으로 발생하며, 면역 매개성과 특발성으로 다시 나뉜다. T2DM은 인슐린 분비 작용 결함에 의해 발생한다. 기타 당뇨병에는 베타세포 기능의 유전적 결함, 인슐린 작용의 유전적 결함, 췌장염이나 종양과 같은 췌장 외분비 기능의 장애, 쿠싱 증후군이나 갑상선 기능 이상과 같은 다른 내분비 질환, 간경화, 약물에 의한 유발, 감염, 다운 증후군과 같은 유전적 증후군 등이 포함된다(대한당뇨병학회, 2011). 이중 연령에 따라 증가하며, 전체 당뇨병의 90% 이상을 차지할 것으로 생각되는 것이 T2DM이다. T2DM의 주된 발병기전은 표적 장기의 인슐린 저항성 증가로 생각되며 인종, 연령, 운동, 흡연, 비만과 같은 다양한 요인과 관련되어 있는 것으로 생각된다(Janzon et al., 1983; American Diabetes, 2013). 또한, 당뇨병은 가족력과 밀접하게 관련되어 있는 것으로 알려져 있으며, 국내에서도 이미 HHEX, CDKN2A/B, CDKAL1, KCNQ1, SLC30A8와 같은 다양한 유전자들이 T2DM과 관련이 되어 있는 것으로 밝혀졌다(Lee et al., 2008).

역학적인 관점에서, 2011년 현재 우리나라의 만 30세 이상 성인의

당뇨병 유병률은 10.5%, (남성 12.6%, 여성 8.5%)이다(보건복지부 & 질병관리본부, 2012). 이것은 미국의 당뇨병 유병률 11.3%(남성 11.8%, 여성 10.8%)과 거의 비슷한 수준이다(CDC, 2011). 전세계적으로 당뇨병 환자의 수는 2008년 347만명으로 성인 남성의 9.8%, 성인 여성의 9.2%에 해당한다(Danaei et al., 2011). 당뇨병은 현재도 지속적으로 증가하고 있어, 2030년에는 전세계적으로 당뇨병 환자수가 552만에 이를 것으로 예상되고 있다(Whiting et al., 2011). 당뇨병의 유병률은 연령에 따라 증가하여, 국내 조사에 따르면 70대의 유병률이 20%를 넘는 것으로 나타났다(Figure 1.1). 당뇨병은 주요 사망 원인 중의 하나로, 2011년 사망 통계(Table 1.1)에 따르면 당뇨병으로 인한 사망률은 남성에서 5위, 여성에서 4위에 해당한다(통계청, 2012). 전 세계적으로는 당뇨병으로 인한 직접 사망은 사망 원인 12위에 해당하지만, 고소득 국가일수록 순위가 높아지는 경향이 있다(WHO, 2008). 당뇨병은 심혈관계 질환, 신장질환, 망막질환, 신경계통 질환, 면역계통 질환 등의 합병증을 통하여 삶의 질을 떨어뜨리고 사망률을 증가시킨다(Klein, 1995; Danaei et al., 2006; American Diabetes, 2013). 또한, 2008년 세계보건기구의 보고서에 따르면, 당뇨병으로 인한 질병 부담은 해가 거듭될수록 증가 하여 2004년 당뇨병으로 인한 질병 부담은 19위지만 2030년에는 10위로 상승할 것으로 예상되었다(WHO, 2008). 실제로 당뇨병은 20세에서 74세 사이에 실명하게 되는 주요 원인이며, 말기 신부전의 주요 원인이기도 하다(CDC, 2011). 이러한 합병증은 당뇨병의 유병기간이 길수록 발생률이 증가하며, 흡연 여부, 혈당과 혈압의 조절 등과도 밀접한 관련이 있다. 인구 10만당 당뇨병으로 인한 사망률은 31.7명 정도이며, 이는 전체 사망자수의 약 4%에 불과하지만, 일반적으로 당뇨병과 밀접한 관련이 있다고 알려져 있는 뇌혈관 질환, 심장질환 등을 포함한다면 당뇨병이 사망에 미치는 영향은 매우 크다고 할 수 있다(통계청, 2012). 따라서,

국내외 역학적 상황과 질병 부담을 고려하였을 때, 당뇨병은 보건학적으로 매우 중요한 질병이며, 향후 그 중요도가 증가할 것으로 짐작할 수 있다.

현재 미국과 한국의 당뇨병 진료 지침은 당뇨병의 진단 방법으로 공복혈당(plasma fasting glucose), 경구당부하검사(oral glucose tolerance test)와 당화혈색소(hemoglobin A1c)를 제안하고 있다(대한당뇨병학회, 2011; American Diabetes, 2013). 경부당부하검사가 장기간 임상적 근거가 가장 많이 축적되어 있으나, 현실적으로는 검사 시간이 오래 걸리며, 재현성(reproducibility)이 낮아 선별검사의 방법으로는 잘 추천이 되지 않는다. 공복혈당 검사는 가격이 저렴하고 비교적 간편하여, 당뇨병의 선별검사로 이용되어 왔으나, 재현성이 높지 않고 내당능장애를 놓칠 수 있다는 제한점이 있다. 반면에 당화혈색소 검사는 저렴하지는 않으나, 검사가 간편하고 재현성도 높아 최근 당뇨병의 진단 기준에 포함이 되었다(Bonora & Tuomilehto, 2011). 하지만, 당뇨병(특히 T2DM)은 초기에는 증상이 잘 나타나지 않아 진단이 늦어지는 경우가 있다. 실제로 2010년 국민건강통계에 따르면, 당뇨병의 인지율은 73.0%(남성 69.0%, 여성 77.6%)에 불과하며, 특히 30대 남성의 인지율은 30.5%에 불과하였다(보건복지부 & 질병관리본부, 2012).

당뇨병의 발병을 인지하지 못함으로써 당뇨병에 대한 치료 및 관리의 시작 시기를 늦추어 합병증의 발생률을 높일 수 있으므로, 대부분의 당뇨병 진료지침은 그 위험도에 따라 다른 당뇨병의 선별검사를 권고하고 있다(대한당뇨병학회, 2011; RACGP, 2012; USPSTF, 2012; American Diabetes, 2013). 미국 당뇨병 학회에 의해 매년 발표되는 당뇨병 진료지침에 따르면, 45세 이상의 연령에서는 3년 마다 당뇨병 선별검사를 시행할 것을 권고하고 있으며, 당뇨병의 위험요인이 있는 경우에는 연령과 무관하게 당뇨병의

선별검사를 시행할 것을 언급하고 있다(American Diabetes, 2013). 국내 진료지침은 기본적으로 미국 당뇨병 학회의 지침을 바탕으로 하고 있으나, 40세 이상의 성인이나 위험인자가 있는 30세 이상의 성인에게 매년 당뇨병 선별검사를 시행할 것을 권고하고 있다(대한당뇨병학회, 2011). 그에 반하여, US preventive services task force (USPSTF)의 지침이나 호주의 Redbook은 위험도 평가를 먼저 시행하고, 고위험군만 선별검사를 시행할 것을 권고하고 있다(RACGP, 2012; USPSTF, 2012). 즉, 전세계적으로 당뇨병의 특성상 선별검사가 필요하다는 점에는 대부분 동의를 하고 있으나, 역학적 특성과 국가간 의료환경의 차이 등을 고려하였을 때에 그 선별검사의 시작 시기가 달라질 수 있다고 정리를 할 수 있다.

당뇨병은 이와 같이, 여러 요인이 복합적으로 작용하여 발생하며 그 위험요인에 따라 선별검사의 시기가 달라질 수 있기 때문에 ARIC, AUSDRISK, “Cambridge Risk Score”, FINDRISC, “Framingham Offspring Study risk score”와 같은 다양한 위험도 평가 도구가 개발되어 있다. 이 중에 가장 대표적이며 타당도가 검증된 도구는 핀란드의 전향적 코호트를 기반으로 한 FINDRISC이다(Silventoinen et al., 2005). FINDRISC는 연령, 체질량지수, 허리둘레, 신체활동, 채소/과일의 섭취, 고혈압약 복용의 여부, 고혈당의 과거력, 당뇨병 가족력의 8개 항목을 점수화 하여 최대 26점 중 15점이면 당뇨병의 위험이 높고, 7점 이상부터 당뇨병의 위험이 증가한다고 정의하였다. 이를 바탕으로 호주에서는 자국민을 대상으로 한 AUSDRISC를 개발하여, 선별검사에 선행하여 위험도를 평가하도록 권고하였다(RACGP, 2012). 국내에서는 코호트를 바탕으로 타당도가 검증된 위험도평가 도구가 없으나, 최근 한 연구에서 국민건강영양조사를 바탕으로 한 위험도평가 도구를 제시한 바가 있으며, 이 도구가 기존 국외의 평가도구에 비해서 한국인에게

타당도가 높음을 비교하여 증명한 바가 있다(Lee et al., 2012).

당뇨병의 치료 및 관리는 크게 생활습관의 개선, 약물 치료, 그리고 합병증 관리로 나눌 수 있다. 특히 인슐린 치료는 T1DM과 T2DM 등 여러 형태의 당뇨병에 모두 효과적인 것으로 알려져 있으나, 현재로서는 주사를 통해서만 투여가 가능하기 때문에 삶의 질의 매우 떨어진다. 또한, 과거에는 철저한 혈당의 관리가 당뇨병 치료의 주요 목표였으나, 최근에는 합병증 관리도 중요한 치료의 목표가 되었다(American Diabetes, 2013). 당뇨병에 있어서 가장 많이 연구가 된 부분은 진단과 치료이지만, 당뇨병의 예방에 대해 관심도 또한 높아지고 있다. 실제로, 대규모 코호트 연구를 통해서 생활습관의 변화가 고위험군의 당뇨병 발생률이 낮아질 수 있음이 증명된 바가 있다(Diabetes Prevention Program Research et al., 2009). 따라서, 당뇨병 위험도의 평가는 단지 선별검사 시행 여부를 결정하는 데에 그치지 않고, 당뇨병의 예방에도 활용이 가능하다고 하겠다.

요약하자면, 당뇨병은 여러 위험요인이 복합적으로 작용하여 발생하는 만성질환으로서, 국내외 역학적 상황과 질병 부담을 고려할 때 보건학적 중요도가 높은 질환이다. 최근에는 당뇨병의 치료 및 관리에서 예방과 조기 발견의 관심도가 커지고 있으며, 타당도 높은 위험도 평가는 당뇨병의 이해와 예방, 조기 발견에 유용하게 활용될 수 있을 것이다.

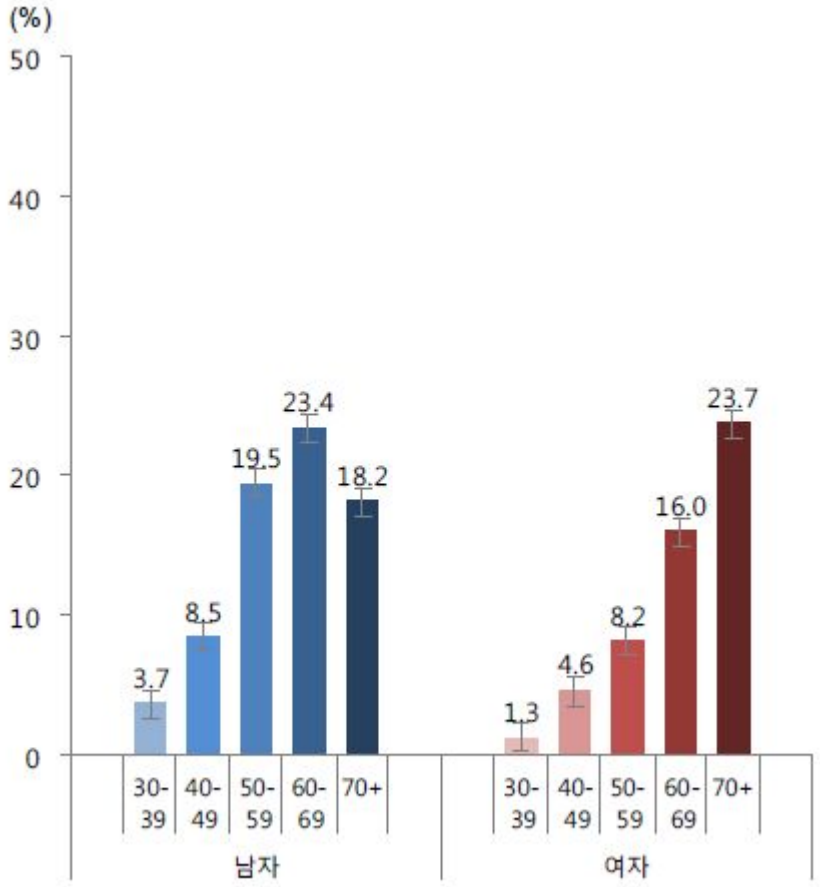

Figure 1.1 Prevalence of diabetes mellitus by gender and age groups in Korea, 2011 (Korean Ministry of Health and Welfare, 2012**)**

Table 1.1 Major cause and mortality in Korea, 2001-2011 (Statistics Korea, 2012**)**

Rank	2001		2010		2011	
	Cause	Mortal-ity[1)]	Cause	Mortal-ity[1)]	Cause	Mortal-ity[1)]
1	Malignancy	122.9	Malignancy	144.4	Malignancy	142.8
2	Stroke	73.7	Stroke	53.2	Stroke	50.7
3	Heart disease	33.9	Heart disease	46.9	Heart disease	49.8
4	Diabetes mellitus	23.8	Suicide	31.2	Suicide	31.7
5	Liver disease	22.2	Diabetes mellitus	20.7	Diabetes mellitus	21.5
6	Traffic accident	20.9	Pneumonia	14.9	Pneumonia	17.2
7	Chronic lower respiratory disease	19.0	Chronic lower respiratory disease	14.2	Chronic lower respiratory disease	13.9
8	Suicide	14.4	Liver disease	13.8	Liver disease	13.5
9	Hypertension	10.2	Traffic accident	13.7	Traffic accident	12.6
10	Respiratory tuberculosis	6.3	Hypertension	9.6	Hypertension	10.1

1) Per 100,000 persons

1.2 질병위험평가 (Health Risk Appraisal)

대개의 의학적 근거들은 개별 위험요인에 대해 위험도를 오즈비(OR; odds ratio) 또는 상대위험도(RR; relative risk)로 표현한다. 하지만, 이러한 개념들을 일반인들에게 전달하는 것은 쉽지 않으며, 여러 위험 요인이 복합적으로 작용하는 다요인질환의 위험도를 직관적으로 받아들이기 위해서는 여러 위험도를 통합하는 과정이 필요하다. 질병위험평가는 이러한 다양한 질병 위험 요인을 수학적으로 결합한 질병 예측 모형이다. 질병위험평가는 근거 자료와 목적에 따라 여러 형태의 통계적 방법과 표현 방식을 취한다. 대표적인 예로 심혈관질환 발생 위험을 평가하는 Framingham Risk Score(Wilson et al., 1998)와 유방암 발생 위험도를 평가하는 Gail model(Gail et al., 1989)은 코호트 자료를 기반으로 하고 있기 때문에 콕스 회귀분석(Cox proportional hazard model)을 따르고 있으며, 5년 또는 10년 질병 위험으로 결과를 표시한다. 국민건강보험공단의 건강나이는 여러 가지 질병들을 모두 포함하고 있으며, 생활습관 개선의 동기 부여가 그 주요 목적이기 때문에 건강나이와 같이 친숙한 형태의 결과로 표현을 하였다(조비룡, 2012). 또한, 선별검사가 필요한 고위험군을 감별하고자 하는 한국인 당뇨병 위험도 평가도구는 미래가 아닌 현재의 당뇨병 상태를 예측하고자 하기 때문에 국민건강영양조사와 같은 단면적 조사자료를 기반으로 개발을 해도 논리적 타당성에 큰 결여가 없으며, 점수화 하여 표현하는 것만으로도 그 목적을 달성할 수 있다(Lee et al., 2012).

여러 가지 위험도를 표현하는 지표 중에, 평가도구의 점수는 타당성이 검증된 기준치가 함께 제공이 되지 않으면 그 위험도를 가늠할 수가 없다. 또한, 기준치가 함께 제공이 된다고 하더라도, 확률적인 개념의 위험도가 아니기 때문에, 원래 개발 목적과 다른

용도로 활용하는 데에는 논리적인 한계점이 있으며, 다른 연구에 직접적인 연계가 어려울 수 있다. 반면에, OR이나 RR은 특정 기준치를 필요로 하지 않고, 일반적으로 역학연구에서 가장 많이 표현하는 위험도이기 때문에 그 활용도에도 장점이 있을 수 있다. 하지만, OR이나 RR도 상대적인 개념의 위험도이기 때문에, 유병률(prevalence)과 발병률(incidence)과 같은 정보를 함께 제공하여야만 절대적인 위험도를 알 수가 있다. 이러한 점들을 고려하였을 때에, 특정 기간의 질병 위험도(5년 발생률 등)는 확률적 지표로서 특정 기준치가 필요하지 않으며, 별도의 정보가 없이 절대적인 위험도를 알 수 있다는 점에서 장점이 있다. 하지만, 이런 특정 기간의 질병 위험도는 대개 코호트 자료를 필요로 하며, 추적기간이 길어질수록 비용과 노력이 많이 투자되어야 하기 때문에, 특정 기간의 질병 위험도는 5년 또는 10년과 같이 제한된 기간의 위험도를 표현하는 경우가 많다(Gail et al., 1989; Wilson et al., 1998). 충분한 기간의 추적 관찰 자료가 있다면, 비슷한 방법으로 평생 질병 위험도를 추정하는 것도 가능하다(Wilkins et al., 2012). 하지만 이 경우에도 수십 년 전 인구집단을 기반으로 하여 현재 인구 집단의 예측 모형을 만들었다는 제한점은 여전히 남아 있다. 따라서, 평생 질병 위험도는 목표 인구 집단을 대표하는 자료를 생명표 등과 수학적으로 연계하여 추정되는 경우가 많다(Fraser & Shavlik, 1999; Narayan et al., 2003; Hopkins et al., 2012). 여기에서 사용되는 수학적 모형은 일반적으로 활용되는 역학적 통계 기법에 비하여 복잡하고 계산량이 많아 바이오인포매틱스와 같은 전산적인 기법이 도움이 될 것이다.

1.3 마코프모델과 은닉마코프모델

마코프모델(Markov model)은 1906년 Andrey Markov가 처음 제안한 뒤, 바이오인포매틱스 분야에서 오래 전부터 개발되어 사용하는 추계적 모형(stochastic model)으로, 이전 시점과 현재 시점간의 상태 전이를 확률적으로 가정하고, 특정 시점에서 어떤 상태에 있을 확률을 구하는 모형이다(Wikipedia, 2013). n차 Markov 가정은 n시점 전의 값에 의존하여 다음 시점의 값이 결정되는 형태이다. 가장 많이 사용되는 1차 마코프모델은 다음과 같은 가정이 필요하다(한학용, 2009).

시간 t에서의 관찰되는 상태: q_t

$P(q_n|q_{n-1}, q_{n-2}, \dots, q_1) = P(q_n|q_{n-1})$

마코프연쇄(Markov chain)는 이러한 가정이 성립하는 시스템의 출력열(output sequence)로 정의되며, 마코프과정(Markov process)이란 연속적인 시간에서 전이 확률에 의해 이루어지는 과정을 의미한다. 이는 다음과 같이 수학적으로 표현 가능하다.

관찰 가능한 상태열: $S_1, S_2, S_3, \dots, S_N$

초기 상태의 분포 벡터: $\pi = (\pi_i),\ \pi_i = P(q_i = S_i), 1 \le i \le N$

전이확률 $a_{ji}=P(q_t=S_j|q_{t-1}=S_i)$

마코프 가정(Markov assumption)에 의하면 $q_1, q_2, \dots, q_n$의 sequence로 관찰될 확률은 결합 확률로 표현이 가능하며, 1차 마코프 가정하에서는 이는 다음과 같이 행렬의 곱으로 표현이 가능하다.

$P(q_1, q_2, \cdots, q_n) = \prod_{i=1}^{n} P(q_i|q_{i-1})$

전이확률 행렬 $A=\begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix}$

시간 T에서 상태의 분포=$A^{T-1}\pi$

여기에서 특정 시점의 상태들은 서로 배타적이기 때문에, 특정 시점의 상태 S_i에서 서로 배타적인 상태 S_j로 가는 전이확률들의 합은 1이 되어야 한다.

$$\sum_{j=1}^{N} a_{ji} = 1, \ \forall \ i$$

마코프모델은 시간의 흐름에 따른 질병 상태의 변화를 모형화 하기에 적합해 비용효과분석에서 널리 사용되는 방법이다(Huang et al., 2009; Neumann et al., 2011; Mortaz et al., 2012). 하지만, 마코프모델은 각 상태의 관찰이 가능하다는 것을 전제로 한다.

반면에 은닉마코프모델(Hidden Markov Model, HMM)은 마코프모델과 유사하지만, 상태를 관찰할 수 없고, 상태로부터 확률적으로 발생하는 사건만 관찰 가능할 때 적용하는 모형이다(한학용, 2009). 마코프모델은 사회현상과 같이 상태를 직접적으로 관찰이 가능한 경우에 적합한 반면에, 특정 시점의 상태를 직접적으로 관찰은 어렵지만, 그 상태로부터 확률적으로 발생하는 관찰 가능한 상태(생체지표나 의무기록 등)가 있을 때 HMM을 적용해 볼 수 있다(Kawamoto et al., 2013; Khader et al., 2013). 하지만, 이것은 절대적인 구분이 아니며, 연구 가설에 따라 차이가 있을 수 있다. 예를 들어, 생체지표를 통해서 특정 질병의 상태를 확률적으로 관찰 가능하다고 가정하면 이것은 HMM이 더 적합하지만, 특정 질병의 상태를 생체지표를 통해 정의하거나, 생체지표의 결과값이 특정 질병의 상태와 항상 일치한다고 가정한다면 HMM이 마코프모델에 비하여 얻는 이익이 없다. 즉, 두 모형의 차이는 특정 상태로부터 확률적으로 관찰되는 값의 유무라고 할 수 있다(Figure 1.2).

HMM을 수학적으로 표현하면 다음과 같다.

상태열(status sequence): $S_1, S_2, S_3, \ldots, S_N$

관측열(observation sequence): $o_1, o_2, o_3, \ldots, o_M$

초기 상태의 분포 벡터: $\pi = (\pi_i),\ \pi_i = P(q_i = S_i), 1 \le i \le N$

시간 t에서의 상태: q_t

시간 t에서의 관측 상태: O_t

전이확률 $a_{ji}=P(q_t=S_j|q_{t-1}=S_i)$

전이확률 행렬 A=$\begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix}$

시간 T에서 상태의 분포=$A^{T-1}\pi$

상태에 따른 관측 확률 $b_{jk}=P(O_t=o_k|q_t=S_j)$

관측 행렬 B=$\begin{bmatrix} b_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ b_{M1} & \cdots & a_{MN} \end{bmatrix}$

시간 T에서 관측값의 분포=$BA^{T-1}\pi$

HMM의 전이확률은 마코프모델에서와 같이 행렬의 같은 열에 있는 값들의 총 합은 1이 되어야 한다. 또한, 관측행렬의 경우에도 관측값들은 서로 배타적인 사건이기 때문에 같은 열의 확률 값의 합은 1이 되어야 한다. 이를 수식으로 표현하면 다음과 같다.

$$\sum_{j=1}^{N} a_{ji} = 1,\ \ \sum_{j=1}^{M} b_{ji} = 1,\ \ \forall\, i$$

시간 t에 관계없이 전이행렬(A)과 관측행렬(B)이 항상 일정하다고 가정한 경우에는 관측열만 주어진다면, Baum-Welch 재추정 알고리즘(한학용, 2009)을 이용하여 해당 관측열이 나타날 확률이 최대가 되도록 전이행렬, 관측행렬, 그리고 초기상태를 계산할 수 있다. 하지만, 장기간에 걸쳐 질병의 상태 변화를 모델링 하는 데에는 이와 같은 가정이 잘 맞지 않는다. 당뇨병의 경우를 예로 든다면, 30세의 발병율과 31세의 발병율은 거의 같다고 가정해도 큰 문제가 없지만, 30세의 발병율과 60세의 발병율이 같다고

가정하기는 어렵기 때문이다. 따라서, 수학적 모형을 실제 질병 발생 형태와 비슷하게 일치시키려면, 매 시점에서 각각 다른 전이행렬을 필요로 한다. 하지만, 이런 경우에는 추정 알고리즘이 알려진 바가 없으며, 추정 알고리즘이 향후 개발이 된다고 하더라도 추정해야 할 모수(parameter)들이 매우 많기 때문에 매우 큰 크기의 “training set”이 필요할 것이다.

HMM의 특징이자 장점은 관측 가능한 상태로부터 관측이 가능하지 않은 상태를 알고리즘을 통해 추정한다는 것이다. HMM의 모수들을 별도의 자료를 통해 계산하여 입력하는 것은 HMM의 가장 큰 장점을 희생하는 것을 의미하지만, 결과를 해석하는 데에 있어서는 부가적인 이익이 있다. 최적화 해를 찾는 알고리즘은 Figure 1.3에서 보는 바와 같이 별도로 모수를 재설정하는 알고리즘이 없으면 지역적 최대점(local maxima)에 도달하였을 때에 수렴하게 된다(Wikepedia, 2013). 하지만, 이 지역적 최대점이 전체적 최대점(global maxima)과 차이가 있을 수 있다. 또한, 이러한 최대점은 단지 주어진 가정하에서 수학적인 최적값을 의미한다. 실제 질병의 상태 변화에는 모형에 반영되지 않은 수많은 변수들이 관련되어 있기 때문에, 수학적인 최적값이 역학적으로 의미가 없거나 경우에 따라서는 사망한 사람이 다시 생존을 하는 것과 같은 설명이 불가능한 결과를 보여줄 수도 있다. 별도의 역학적 자료로 모수를 계산하는 것은 수학적으로 최대점은 아니지만, 전체적 최대점에 더 가까운 모형이 될 가능성을 높여주며, 특정 확률을 0 또는 1과 같이 고정값으로 설정함으로써 역학적으로 설명이 되지 않는 일을 사전에 배제할 수 있다.

매 시점의 전이확률을 계산하는 것은 상태열을 직접적으로 관측이 가능해야 하기 때문에 HMM의 기본 가정과 잘 맞지 않는다. 하지만, 상태열의 관측이 어렵지만 불가능하지는 않은 경우에는

HMM을 적용하는 것이 의미가 있을 수 있다. 질병의 진단에 있어서 표준검사(gold standard)는 시간, 비용, 검사상의 어려움 등의 이유로 대부분 관측이 어렵기 때문에 연구단계에서만 활용이 되고, 실제 임상에서는 표준검사에 확률적인 연관성이 있는 다른 검사 방법을 채택하기 때문이다. 즉, 연구를 통해 얻어진 자료의 관측이 쉬운 지표와 관측이 어려운 지표를 한 모형 안에 포함시킴으로써, 실제 임상에서 관측이 쉬운 지표만을 측정했을 때에 질병 상태 변화를 확률적으로 추정해 볼 수 있겠다. 그리고, 이러한 모형은 전형적인 HMM이라기 보다는 마코프모델에 HMM의 확률적 요소가 추가된 형태라고 하는 것이 맞을 것이다.

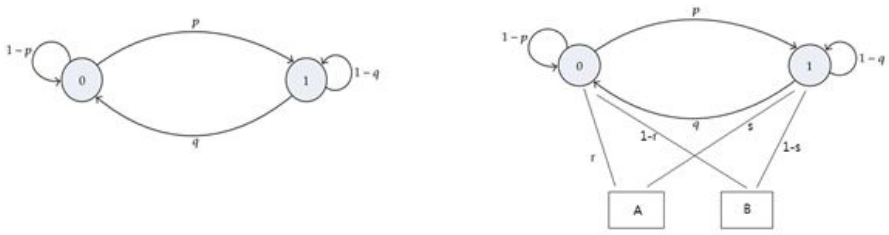

Figure 1.2 Markov model (left) and Hidden Markov model (right). (0, 1: Each status; A, B: observation or output; p, q, 1-p, 1-q: transitional probability; r, s, 1-r, 1-s: observational probability of each status)

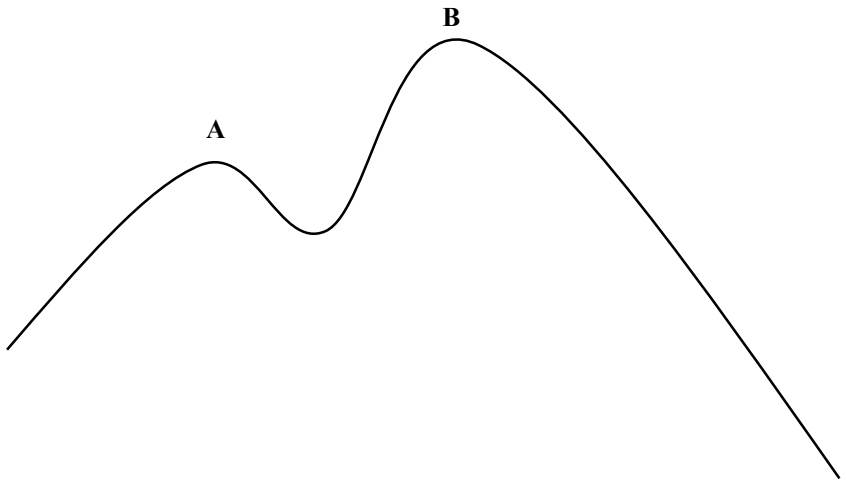

Figure 1.3 Local maxima(A) and global maxima(B).

1.4 연구의 필요성과 목적

일반적으로 희귀 질환은 단일 유전자의 이상과 관련이 되어 있을 것으로 알려져 있으나, 비교적 흔한 질환은 여러 요인이 복합적으로 관련되어 발생할 것으로 생각된다. 또한 여러 요인이 연관되어 있는 질병은 질병의 경과(natural course) 또한 다양한 형태로 나타날 수 있으며, 질병의 진단, 치료, 관리, 예후 등에도 여러 요인들이 지속적으로 연관이 되어 일반적인 방법으로 이해하고 분석하는 것이 쉽지 않다. 따라서, 바이오인포매틱스(Bioinformatics)는 이러한 복잡한 질환들을 이해하고 분석하는 데에 다양하게 활용될 수 있다. 그 대표적인 예가 바로 당뇨병이다. 당뇨병은 여러 유전자가 연관되어 있음이 밝혀졌고(Grant et al., 2006; Todd et al., 2007; Wellcome Trust Case Control, 2007; Concannon et al., 2009), 동물 실험을 통해 일부 신호전달체계를 규명하기도 하였으나(Korc, 2003), 전체적인 관점에서 포괄적으로 분석하기 위해서는 “tree algorithm”이나 “gene network” 같은 전산적 기법들이 도움이 될 수 있다. 특히, 이전 연구에서 당뇨병의 조기 진단에 있어서 “computational intelligence”가 유용하게 사용될 수 있다는 점은 이미 정리된 바가 있으며(Shankaracharya et al., 2010), “network analysis”는 합병증들 사이의 복잡한 관계를 이해하는 데에 도움을 줄 수 있다(Liu et al., 2012). 이전의 리뷰논문을 통해 그 밖의 다양한 영역에서 바이오인포매틱스가 활용되고 있는 예시들을 찾을 수 있으며(Table 1.2)(윤재문 & 손현석, 2012), 이 외에도 수많은 관련 연구들이 진행되고 있을 것이다.

바이오인포매틱스 기법을 이용한 진료도구들은 기존의 평가도구들에 비해서 계산방법이 복잡하고, 많은 양의 정보를 필요로 한다는 점에서 실제 진료실 등에서 활용이 되기에는 어려웠다. 하지만, 최근에는 진료실에서 컴퓨터의 이용이 확산됨에 따라, 다양한 정보를 더욱 손쉽게 접근할 수 있게 되었으며,

자동화된 “clinical decision support system”(CDSS)이 진료환경에 통합되려는 시도가 지속적으로 이루어 지고 있다(GLIDES, 2010; EBMeDS, 2011). CDSS는 여러 가지 정보를 재입력 없이 통합해서 즉각적으로 보여주는 것이 가능하며, 많은 양의 계산이 필요한 수학적 모형도 진료실에서 활용이 가능하게 할 수 있다. 또한, 적절한 CDSS의 활용은 비용효과적인 측면에도 기여를 할 것으로 기대된다(Chi et al., 2010). 평생 질병 위험도와 같이 복잡한 계산이 필요한 모형의 경우는 현재 시점에서의 활용도가 떨어질 수 있으나, 향후 CDSS가 보편화 되면 이와 같은 모형의 개발과 활용이 가속화 될 것을 예상해 볼 수 있다.

앞에서 언급한 바와 같이 당뇨병의 중요도는 점차로 증가하고 있으며, 이에 따라 당뇨병에 대한 대중의 관심도도 늘어날 것으로 기대하고 있다. 질병 예측 모형은 일반인과 의료인에게 모두 유용한 정보를 제공할 수 있으나, 정보를 전달하는 방식에 따라 효과에 차이가 있을 수 있다. 질병 위험도를 표현하는 방법 중 하나인 평생 질병 위험도는 다른 개념들에 비해 일반인들에게 이해가 어렵지 않을 것으로 기대되지만, 평생 질병 위험도를 이상적으로 산출하기 위해서는 장기간의 코호트가 필요하다. 이에 대한 대안으로, 바이오인포매틱스 기법을 이용하여 평생 질병 위험도 추정을 생각해 볼 수 있다. 실제로, 외국에서는 마코프모델을 이용한 평생 당뇨병 평생 위험도를 산출한 예가 있다(Narayan et al., 2003; Narayan et al., 2007). 하지만, 기존 연구에서는 자가기입식 설문을 통하여 당뇨병 유병률과 발생률을 산출하여, 이를 마코프모델의 모수를 산출하는 데에 이용하였다. 이와 같은 모형은 당뇨병에 대한 인지율이 80%에 채 미치지 못하여, 생물학적 당뇨병 상태를 반영하기는 어렵다. 따라서, 생물학적 당뇨병 상태를 직접 모델링하는 형태가 더 타당할 것으로 생각되며, 이러한 형태는

HMM을 통해 표현 가능하다.

이와 더불어 개인별 위험인자에 따른 예측 연구는 없었다. 학술적 또는 보건통계학적 목적으로 질병 위험도를 사용할 경우에는 표준 인구집단의 예측 모형만으로 충분할 수 있다. 하지만, 의학적 결정이나 위험요인 교정과 같은 목적으로 활용하고자 한다면 개인별 위험요인에 따른 위험도 추정을 필요로 한다. 하지만, 개인별 추정은 확률적 요소가 더 많아 평균적 추정에 비해 편차가 클 것으로 생각된다. 따라서, 개인별 위험도 추정 모형은 타당성 검토가 함께 이루어져야 그 가치가 있겠다고 하겠다.

따라서, 본 연구에서는 개인별 위험요인에 따른 평생 당뇨병 위험도를 바이오인포매틱스 기법 중 하나인 HMM을 활용하여 추정하고, 그 타당성을 검토하고자 하였다.

Table 1.2 Computational methods for understanding diabetes mellitus in previous studies (Yoon & Son, 2012**)**

Category	Computational method
Pathology	Ingenuity pathway analysis
	Yet another biological networks analyzer
	Electrostatic distribution
	Distance method and maximum parsimony method
	Maximum parsimony and neighbor joining analyses
Diagnosis	Tree-structured model
	Support vector machine
	Linear discriminant analysis
	Classification and regression tree
	Neural network mixture of experts model
	Graphic user interface
Treatments	Fuzzy inference system
	Artificial neural network
	Core diabetes model
	Lamarckian genetic algorithm
Complications	Hybrid decision support system
	Artificial neural network
	Decision trees
	Multilayer perceptron
	Network topology
	Clustering and artificial neural network
Cost-effectiveness	Markov model
	Markov and Monte Carlo model
	Core diabetes model

제 2 장. 연 구 방 법

2.1 연구 절차

본 연구는 크게 예측 모형의 개발, 위험도 추정, 타당성 평가의 세 단계로 나눌 수 있다. 또한 예측 모형의 개발은 표준 위험도에서의 예측 모형을 먼저 설계한 다음에 위험도에 따라서 모형을 수정하는 과정을 거친다. 그리고 위험도의 추정은 개인별 위험 요인에 따라 추정 결과를 시뮬레이션 해보는 과정이다. 이 과정은 각각의 위험요인에 따른 위험도 선택, 복합 위험도 계산, 추정의 3단계로 다시 나눌 수 있다. 타당도 평가의 과정은 타당도 평가를 위해 준비된 자료를 이용하여 개인별 위험인자에 따른 위험도를 구하고, 실제 당뇨병 여부와 위험도 간의 관계를 기존의 평가도구와 비교하는 과정이다.

본 연구를 위해서 필요한 한국인 당뇨병의 기초자료는 국민건강영양조사 5기(2010~2011년) 결과를 이용하였다(보건복지부 & 질병관리본부, 2012). 국민건강영양조사는 자가기입식 당뇨병 유무와 당뇨병 진단 시기에 대한 정보가 있으며, 동시에 혈액검사(공복혈당 검사 및 당화혈색소) 결과를 포함하고 있어 생물학적인 당뇨병 상태를 반영 하는 것이 가능하다. 이 자료를 각각 “training set”과 “validation set”으로 구분하여 연구에 이용하였다. 그 밖에 사망률, 생명표와 관련된 자료는 통계청 홈페이지(국가통계포털)의 자료를 이용할 계획이다. 위 자료들을 바탕으로 HMM에 따른 당뇨병 질병 모형을 설계하고, 성별 및 연령에 따른 당뇨병 평생 위험도를 추정하였다.

2.2 평생 당뇨병 유병률의 추정

2.2.1 HMM의 정의 및 가정

특정 시점에서의 상태는 다음 세가지 중 하나로 가정하였다.

S_1: 정상

S_2: 당뇨병

S_3: 사망

여기서 당뇨병은 혈액검사에서 공복혈당검사 126mg/dL이상, 또는 당화혈색소 검사 6.5%이상으로 정의하였다(American Diabetes, 2013). 75g 경구당부하검사에 따른 당뇨병 기준은 가용한 자료가 부재하여 진단기준에 포함시키지 않았다. 다만 현재 당뇨병의 약물치료(인슐린 또는 경구혈당강하제)를 받고 있다고, 응답한 경우에는 혈액검사의 수치가 기준치 보다 낮아도 당뇨병으로 정의하였다. 또한, 당뇨병이 한번 발생하면 정상의 상태로 돌아가지 않는다고 가정하였다. 각 연령의 상태에서 관측되는 값은 다음 세가지 중 하나로 가정하였다.

O_1: 정상

O_2: 당뇨병

O_3: 사망

관측값은 특정 시점의 상태에서, 인지되고 있는 상태로 정의하였다. 예를 들어, 현재 혈액검사의 수치는 당뇨병의 기준에 해당이 되지만 당뇨병으로 진단 받은 적이 없는 경우에는 정상으로 정의하였다. 반대로 당뇨병에 대해서 치료받고 있지 않으며 혈액검사는 정상이지만 과거의 시점에서 당뇨병으로 인지된 경험이 있는 경우도 있을 수 있으나, 앞의 모형에서 당뇨병의 정의상 혈액검사의 특이도는 100%이며 당뇨병 상태에서 정상의 상태로 돌아가지 않는다고 가정하였으므로, 과거 당뇨병에 대한 인지를 오류로

간주하여 모형에 포함시키지 않았다. 시간 변수는 1년 단위로 하였으며, 최초의 시점(T_0)은 현재의 연령으로 하였다. T_0의 범위는 20~83세로 정의하였으며, 84세의 상태와 무관하게 남녀 모두 85세에는 사망하는 것으로 가정하였다.

2.2.2 모수(parameter)의 추출

HMM를 계산하기 위해서는 다음과 같은 3가지 모수를 필요로 한다.

초기 상태: $\pi = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

t 시점 $\rightarrow$ (t + 1)시점의 전이행렬: $T_t = \begin{pmatrix} A_t & 0 & 0 \\ B_t & C_t & 0 \\ D_t & E_t & 1 \end{pmatrix}$

t 시점의 관측 행렬: $O_t = \begin{pmatrix} 1 & P_t & 0 \\ 0 & Q_t & 0 \\ 0 & 0 & 1 \end{pmatrix}$

위 행렬에서 보는 바와 같이 초기상태에서는 당뇨병이 없으며, 생존한 상태로 가정하였다.

A_t: t시점에서 정상인 사람이, t+1 시점에서 정상일 확률

B_t: t시점에서 정상인 사람이, t+1 시점에서 당뇨병일 확률

C_t: t시점에서 당뇨병인 사람이, t+1 시점에서 당뇨병일 확률

D_t: t시점에서 정상인 사람이, t+1 시점에서 사망하였을 확률

E_t: t시점에서 당뇨병인 사람이, t+1 시점에서 사망하였을 확률

P_t: t시점에서 당뇨병이지만, 정상으로 인지하고 있을 확률

Q_t: t시점에서 당뇨병이면서, 당뇨병으로 인지하고 있을 확률

여기에서 $A_t+B_t+D_t=1$, $C_t+E_t=1$, $P_t+Q_t=1$ 이 됨은 정의에 따라 자명하다. A_t~E_t는 다음과 같은 방법으로 구하였다. 먼저, 각 성별, 연령별 당뇨병 유병률은 국민건강영양조사 2007~2011년 자료를 통해 구할

수 있다. 하지만, 국민건강영양조사의 유병률은 1세 단위로 유병률을 직접 계산할 경우에는 각 연령별 표본수가 매우 작아 불규칙한 패턴을 보여 모수로 사용하기에 적합하지 않았다. 따라서, “restricted cubic spline”에 따른 로지스틱 회귀분석(logistic regression)을 적용하여 산출하였다(pseudo R^2, 남성 0.10, 여성 0.15). 또한 통계청 사망자료에서 성별, 5세 구간 연령대별 전체 사망률을 구할 수 있으며, 각 구간의 중앙값(20~24의 경우 22세)을 대표 연령으로 두었다. 연령과 사망률은 식(1)의 모형을 따른다고 가정하여(R^2, 남성 0.99, 여성 0.96), 회귀분석을 통하여 각 연령의 사망률을 추정하였다.

$$ln(M_{age}) = \beta \times age + \varepsilon \quad \cdots\cdots (1)$$

당뇨병이 없는 대상자와 비교할 때, 당뇨병이 있는 대상자의 사망위험비를 R이라고 가정하였을 때(Woodward et al., 2003), 연령 t에서의 당뇨병 유병률을 Pr(t)라고 한다면, 식(2)와 식(3)과 같이 표현이 가능하고, 여기에서 Pr(t)와 M_t는 이미 알고 있으므로, Et와 Dt를 계산할 수 있다. 다만, 85세에는 모두 사망한다는 가정에 따라, $D_{84}=E_{84}=1$로 설정하였다.

$$(1 - \mathrm{Pr(t)}) \times \mathrm{D_t} + \mathrm{Pr(t)} \times \mathrm{E_t} = M_t \quad \cdots\cdots (2)$$

$$\mathrm{E_t} = \mathrm{D_t} \times \mathrm{R} \quad \cdots\cdots (3)$$

또한, $C_t=1-E_t$로 계산 가능하다.

B_t는 정의에 따르면, 해당 시점에서의 1년간 당뇨병 발생률과 같다. 당뇨병의 발생률에 대한 직접적인 자료가 있으면 B_t를 산출하는 것이 보다 용이하지만, 본 연구에서는 생물학적 당뇨병 상태를 직접 관찰이 어려운 은닉 상태(hidden status)로 정의하였기 때문에, 인접한 두 시점의 유병률과 사망률을 이용하여 나머지 모수들을 계산하였다. t+1시점의 유병률인 Pr(t+1)은 전이행렬의

정의에 의해서, 식(4)로부터 B_t를 계산할 수 있으며, A_t=1-B_t-D_t로부터 A_t를 계산할 수 있다.

$$Pr(t+1) = \frac{B_t \times (1 - Pr(t)) + C_t \times Pr(t)}{1 - M_t} \quad \cdots\cdots (4)$$

이를 모두 정리하면 식(5)~(9)와 같다.

$$A_t = 1 - \frac{Pr(t+1) - Pr(t) + M_t \times \left(\frac{Pr(t) \times R}{1 - Pr(t) + Pr(t) \times R} - Pr(t+1) \right)}{1 - Pr(t)} - \frac{M_t}{1 - Pr(t) + Pr(t) \times R} \quad (5)$$

$$B_t = \frac{Pr(t+1) - Pr(t) + M_t \times \left(\frac{Pr(t) \times R}{1 - Pr(t) + Pr(t) \times R} - Pr(t+1) \right)}{1 - Pr(t)} \quad \cdots\cdots (6)$$

$$C_t = 1 - \frac{M_t \times R}{1 - Pr(t) + Pr(t) \times R} \quad \cdots\cdots (7)$$

$$D_t = \frac{M_t}{1 - Pr(t) + Pr(t) \times R} \quad \cdots\cdots (8)$$

$$E_t = \frac{M_t \times R}{1 - Pr(t) + Pr(t) \times R} \quad \cdots\cdots (9)$$

Q_t는 국민건강영양조사 2007~2011년 자료에서 연령별로 정의상 당뇨병 상태에 해당되는 사람들 중에서 당뇨병으로 진단받은 적이 있다고 응답한 사람들의 비율로 산출하였다. 하지만, 유병률과 비슷하게, 연령별 표본수가 매우 작아 직접 계산한 값은 모수로 사용하기에 적합하지 않았다. 따라서 연령별로 계산한 값을 기초로 하여 3차 다항식으로 회귀분석을 하여 산출된 예측된 값을 Q_t로 사용하였으며 P_t는 1-Q_t로 계산할 수 있었다.

2.2.3 추정 계산식

연령 m세의 사람의 n세에서의 상태는 식(10)와 같은 확률로 분포한다.

$$\pi_{mn} = \begin{pmatrix} P(S_1) \\ P(S_2) \\ P(S_3) \end{pmatrix} = T_{n-1}T_{n-2}\cdots T_m\pi \quad \cdots\cdots (10)$$

W_m을 식(11)과 같이 정의하였을 때, W_m의 첫 행은 당뇨병 없이 지내는 기간의 기대값, 둘째 행은 당뇨병이 있으면서 지내는 기간의 기대값에 해당한다. 또한, W_m의 첫 행과 둘째 행의 합은 연령 m인 사람의 총 기대여명이 된다.

$$W_m = \pi + \sum_{i=m}^{84} \pi_{mi} \quad \cdots\cdots (11)$$

비슷하게, 관측값의 행렬인 U_{mn}, 그리고 그 합의 행렬인 V_m을 식(12), (13)과 같이 정의할 수 있다.

$$U_{mn} = \begin{pmatrix} P(O_1) \\ P(O_2) \\ P(O_3) \end{pmatrix} = O_{n-1}T_{n-1}T_{n-2}\cdots T_m\pi \quad \cdots\cdots (12)$$

$$V_m = \sum_{i=m}^{84} U_{mi} \quad \cdots\cdots (13)$$

이 때, V_m의 첫 행은 당뇨병 없다고 알고 지내는 기간의 기대값, 둘째 행은 당뇨병이 있음을 인지하고 지내는 기간의 기대값에 해당한다.

현재 당뇨병이 없는 m세의 사람의 당뇨병 평생 유병률(LR_m: lifetime risk)은 사망 직전에 해에 당뇨병의 유병상태인 경우를 모두 더함으로써 산출 가능하다. 따라서, 식(14)와 같이 표현할 수 있다.

$$LR_m = \sum_{i=m}^{84}(0 \quad E_i \quad 0)\pi_{mi} \quad \cdots\cdots (14)$$

2.3 위험 요인에 따른 개인별 당뇨병 위험비 산출

2.3.1 당뇨병 위험 요인의 선정

개인별 당뇨병 위험비를 산출하기 위하여, HMM 모형에 이미 반영되어 있는, 연령과 성별을 제외한 당뇨병 위험 요인들을 문헌 고찰을 통해서 선정하였다. 당뇨병의 위험 요인은 여러 가지가 알려져 있지만, 비교적 쉽게 정보를 얻을 수 있으며, 기존의 국내 연구에서 연관성이 알려진 요인들은 당뇨병의 가족력, 고혈압 유무, 허리둘레, 흡연여부와 음주량이었다(Lee et al., 2012). 공복혈당장애에 따른 당뇨병의 발생 위험은 단면적 조사인 국민건강영양조사로부터는 산출할 수 없었기에 위험요인에서 제외하였다.

2.3.2 위험 요인별 위험비 추출

당뇨병의 가족력, 고혈압 유무, 허리둘레, 흡연여부, 음주량의 위험비를 계산하기 위하여, 국민건강영양조사 5기의 1, 2차 연도 자료를 1:1로 무작위 추출 분할한 “training set”을 이용하였다. 허리둘레는 한국인의 허리둘레 기준인 남성 90cm, 여성 80cm 이상인 경우를 허리둘레 이상(복부 비만)으로 정의하였다(Lee et al., 2007). 체질량 지수는 25.0 미만, 25.0~29.9, 30.0 이상의 세 군으로 나누었다. 음주량은 1일 평균 음주량(단위 잔)으로 산출하였다. 가족력은 부모, 형제자매가 현재까지 당뇨병을 진단 받은 적이 있는 경우로 정의하였다. 다변량 로지스틱 회귀분석을 하여 각 위험요인별로 얻어진 오즈비(OR)를 각 위험요인의 위험비로 사용하였다. 로지스틱 회귀분석은 성별에 따라 나누어 시행하였으며, 연령을 함께 보정하였다.

2.3.3 오즈비의 표준화

HMM은 표준 인구집단을 기준으로 하지만, 위험 요인별 위험비는 저위험군에 대한 오즈비로 계산이 되기 때문에, 이를 일치하기 위해서 표준화 오즈비(sOR)를 구하였다. 특정 위험요인 j에서 위험도에 따라 대상자의 비율을 각각 p_i, 저위험군에 대한 오즈비를 OR_i라고 하였을 때, 식(15), (16)를 통하여 sOR을 구할 수 있다.

$$sOR_i = e^{ln(OR_i) - E(ln(OR))} \quad \cdots\cdots (15)$$

$$E(ln(OR)) = \sum p_i \times ln(OR_i) \quad \cdots\cdots (16)$$

여기에서, p_i는 국민건강영양조사에서 당뇨병이 없는 대상자의 성별, 연령대에 따른 비율을 이용하였다. 특히, 위험요인의 위험도에 따라 2개의 집단으로 구분하고, 특정 성별의 연령대에서 위험요인이 있는 대상자의 비율이 p라고 한다면, 저위험 집단의 오즈비는 식(17)과 같고, 고위험 집단의 오즈비는 식(18)과 같다.

$$sOR_{low} = e^{ln(1.0) - E(ln(OR))} = e^{-p\,ln(OR)} = OR^{-p} \quad \cdots\cdots (17)$$

$$sOR_{high} = e^{ln(OR) - E(ln(OR))} = e^{(1-p)\,ln(OR)} = OR^{1-p} \quad \cdots\cdots (18)$$

2.3.4 복합 위험도의 산출

복합 위험도(OR_{mix})는 개별 위험요인의 OR을 모두 곱하여 계산하였다. 즉 k개의 위험 요인이 있을 때, OR_{mix}는 식(19)과 같이 표현 가능하다.

$$OR_{mix} = \prod_{j=1}^{k} sOR_j \quad \cdots\cdots (19)$$

2.4 개인별 당뇨병 위험도 추정

본 연구에서는 위험요인이 당뇨병의 발생위험에는 영향을 미치나, 발생한 당뇨병의 사망위험은 변하지 않는다고 가정하였다. 위험군은 당뇨병이 표준인구 집단에 비해 많이 발생한다고 가정할 때, 다른 모수들은 동일하나, A_t와 B_t는 조정이 필요하다. 식(20)에 의해서, 식(21)와 같이 풀 수 있다.

$$\frac{B_t'/A_t'}{B_t/A_t} = OR_{mix},\ A_t' + B_t' = A_t + B_t \quad \cdots\cdots (20)$$

$$A_t' = \frac{A_t(A_t+B_t)}{A_t+OR_{mix}\times B_t}, B_t' = OR_{mix} \times \frac{B_t(A_t+B_t)}{A_t+OR_{mix}\times B_t} \quad \cdots\cdots (21)$$

조정된 전이행렬을 통해 2.2.3의 과정을 반복하면, 개인 위험도에 따른 당뇨병 유병기간과 위험도를 구할 수 있다. 위험요인에 따라 당뇨병의 사망위험은 변하지 않는다고 가정하였다.

2.5 개인별 당뇨병 위험도 모델의 타당성 검토 및 민감도 분석

일반적으로 타당성 검토를 위해서는 해당 지표를 측정한 장기간의 코호트 자료를 필요로 한다. 하지만 본 연구에서 S_1, S_2, S_3는 인지되지 않은 당뇨병을 포함하므로 측정이 불가능 한 것은 아니나 측정에 많은 시간과 노력을 필요로 한다. 따라서 본 연구에서는 개발된 모델을 다음과 같이 간접적인 방법으로 타당성 검토를 시행하였다.

2.5.1 표준 위험도 집단의 기대 여명의 비교

통계청에서는 2010~2011년 기준 각 연령별, 성별로 기대여명을 추정하여 제공하고 있다. 통계청에서는 사망자수, 지연신고, 연령 등을 보정하였으며, 최종 상한 연령을 100세로 두고 있다. 본 연구에서는 최종 상한 연령을 85세로 두고 있으며, 추정 방법에도 차이가 있으나, 표준 위험도를 가진 집단의 각 세별 기대여명을 통계청 산출 결과와 비교함으로써 간접적으로 모델의 타당성을 검토하였다.

2.5.2 예측 당뇨병 유병률과 당뇨병 유병 여부의 비교

국민건강영양조사 5기의 1, 2차 년도 자료 중에 “training set”을 제외한 나머지를 “validation set”으로 이용하였다. 국민건강영양조사는 횡적 관찰 자료로서 타당성 검토를 위하여 20세에 당뇨병이 없다고 가정하였다(20세 이전에 이미 당뇨병이 있는 대상자를 제외). 본 연구의 모델에 따라 현재의 연령에서

예측되는 당뇨병의 유병률과 실제 당뇨병의 유병 여부를 이용하여 ROC곡선을 작성하였다(시작연령=20세, 종료연령=현재 연령). 예측 유병률은 식(10)에 따라 식(22)와 같이 계산하였다. 또한 최근 연구에서 국민건강영양조사를 이용하여 개발된 한국인 당뇨병 위험도 예측도구를 비교대상으로 하였다(Lee et al., 2012).

$$\text{Prevalence}_{\text{predicted}} = \frac{P(S_2)}{P(S_1)+P(S_2)} \quad \cdots\cdots (22)$$

2.5.3 민감도 분석

위험요인의 종류를 변화시켜 다양한 모형을 만들어 민감도 분석을 시행하였다. 앞에서 제안한 모형을 “full model”로 하였으며, 연령과 성별만을 통해 추정한 모형을 “basal model”로 정의하였다. “basal model”에 당뇨병의 가족력, 고혈압 유무, 허리둘레, 흡연여부, 체질량지수를 각각 한가지씩 추가한 모형, 그리고 “full model”에 앞의 5가지 위험요인을 각각 한가지씩 제외한 모형을 추가로 정의하였다. 여기에, 한국인 당뇨병 위험도 예측도구를 포함하여 총 13가지 모형에 대해 area under curve와 95% 신뢰구간을 비교하였다. 95% 신뢰구간은 “Delong method”로 불리는 비모수적 비교방법을 이용하였다(Delong et al., 1988).

2.6 기 타

2.6.1 연구 방법의 차용과 독창적 요소

본 연구에서 사용된 모형의 주요 구조와 가정은 마코프모델, HMM의 일반적인 형태(한학용, 2009)와 기존의 평생 당뇨병 위험도 연구에 기술된 내용을 참고하였다(Narayan et al., 2007). 특히, 식 (8), (9)의 D_t, E_t는 기존 연구와 동일한 방법이 사용되었다.

본 연구는 기존의 연구에 비교하였을 때, 3가지 독창적 요소가 있다. 첫 번째는 기존의 연구에서는 마코프모델을 기반으로 해서 전이행렬만으로 모형을 구성하였지만, 본 연구에서는 HMM의 이중 확률적 요소를 추가하여 적용하였다는 점이다. 두 번째는, 개인별 위험 요인에 따른 복합 위험도를 이용하여 전이행렬을 교정하는 논리(logic)가 포함되어 있는 점이다. 마지막으로 기존 연구에는 발생률에 대해서 자료로부터 직접 산출하였지만, 본 연구에서는 발생률을 직접 산출할 수 없는 단면적 조사자료로부터 간접적으로 발생률을 계산하였다는 점이라고 하겠다.

2.6.2 분석에 사용된 패키지

본 자료 분석을 위해서 기초 사망률, 유병률의 추출 및 생성에서는 EXCEL 2010과 STATA 12.1이 사용되었으며, HMM의 모델링과 타당성 검토에는 R 3.0.2이 사용되었다.

제 3장. 연 구 결 과

3.1 모수의 추정 (Estimation of parameters)

3.1.1 전체 사망률

남성의 전체 사망률은 20세에서 최소값(10만명 당 36명), 그리고 84세에서 최대값(10만명 당 8,787명)을 나타내었다(Table 3.1). 그리고 전체적인 양상은 지수함수적으로 증가하는 양상이었다(Figure 3.1). 여성에서도 남성에서와 비슷한 양상이었으나, 각 연령대에 해당하는 사망률은 남성의 1/2 수준이었다(Table 3.2).

3.1.2 당뇨병의 유병률

당뇨병의 유병률은 국민건강영양조사의 자료를 기반으로 한 “restricted cubic spline curve”에 의해 남녀 모두에서 전 구간에 걸쳐 증가하는 형태로 모형화 되었다(Figure 3.2). 남성의 경우에는 20세의 10만명 당 557명에서 시작하여, 84세의 29,091명까지 S자 형태의 곡선을 보여주었으며, 여성의 경우에는 전체적으로 남성보다 낮은 유병률을 보여주었지만(Table 3.1, 3.2), 그래프 양상은 비슷한 S자 형태로 나타났다. 하지만, 남성은 40세에서 60세에 걸쳐 유병률의 상승이 가속화된 반면에, 여성은 50세 이후부터 가파르게 유병률이 상승하는 형태로 나타났다. 이에 따라 남녀간의 당뇨병 유병률 차이는 50대에 많이 벌어지지만 65세에 이르러서는 남녀의 차이가 다시 줄어들었다.

Table 3.1 Data for parameter estimation in men.

Age	Death	Prevalence	Recognition rate	Age	Death	Prevalence	Recognition rate
20	36.46	556.66	4.78	53	616.66	16786.70	65.12
21	39.72	628.68	5.11	54	671.84	17593.82	67.84
22	43.28	709.95	5.46	55	731.96	18360.53	70.27
23	47.15	801.64	5.83	56	797.45	19085.45	72.41
24	51.37	905.07	6.22	57	868.81	19768.31	74.25
25	55.97	1021.70	6.64	58	946.54	20409.74	75.82
26	60.97	1153.18	7.09	59	1031.24	21011.31	77.11
27	66.43	1301.23	7.57	60	1123.51	21575.35	78.15
28	72.37	1467.77	8.08	61	1224.04	22104.44	78.94
29	78.85	1654.85	8.64	62	1333.57	22599.73	79.5
30	85.91	1864.69	9.26	63	1452.89	23062.32	79.88
31	93.59	2099.69	9.94	64	1582.89	23493.62	80.08
32	101.97	2362.40	10.69	65	1724.53	23895.35	80.13
33	111.09	2655.54	11.53	66	1878.83	24269.49	80.03
34	121.03	2981.96	12.48	67	2046.95	24618.25	79.81
35	131.86	3344.63	13.55	68	2230.11	24944.08	79.48
36	143.66	3746.61	14.76	69	2429.65	25249.58	79.04
37	156.51	4191.02	16.15	70	2647.05	25537.52	78.50
38	170.52	4680.98	17.74	71	2883.9	25810.79	77.88
39	185.78	5219.57	19.56	72	3141.95	26072.43	77.18
40	202.4	5809.79	21.64	73	3423.09	26325.56	76.42
41	220.51	6454.02	24.03	74	3729.38	26573.42	75.61
42	240.24	7152.53	26.72	75	4063.07	26819.32	74.77
43	261.74	7903.86	29.72	76	4426.63	27066.07	73.90
44	285.16	8705.01	32.99	77	4822.72	27314.25	73.02
45	310.67	9551.18	36.50	78	5254.24	27563.84	72.11
46	338.47	10435.66	40.19	79	5724.38	27814.84	71.19
47	368.75	11349.74	44.00	80	6236.59	28067.24	70.25
48	401.75	12282.72	47.85	81	6794.63	28321.04	69.29
49	437.7	13221.97	51.65	82	7402.59	28576.22	68.32
50	476.86	14153.13	55.34	83	8064.96	28832.77	67.33
51	519.53	15062.17	58.84	84	8786.6	29090.68	66.32
52	566.02	15941.61	62.11				

1) All cause death per 100,000 persons
2) Prevalence of diabetes mellitus per 100,000 persons
3) Proportion of recognition per 100 persons with diabetes mellitus

Table 3.2 Data for parameter estimation in women.

Age	Death	Prevalence	Recognition rate	Age	Death	Prevalence	Recognition rate
20	18.08	226.85	47.95	53	293.66	7375.94	54.75
21	19.67	263.41	48.21	54	319.55	8242.04	58.14
22	21.41	305.84	48.47	55	347.71	9227.90	61.63
23	23.29	355.09	48.73	56	378.36	10326.85	65.07
24	25.35	412.23	48.99	57	411.71	11524.25	68.31
25	27.58	478.52	49.25	58	448	12795.91	71.25
26	30.01	555.34	49.50	59	487.49	14106.97	73.81
27	32.66	644.00	49.75	60	530.46	15411.88	75.92
28	35.54	745.70	49.98	61	577.22	16662.90	77.56
29	38.67	861.59	50.18	62	628.1	17837.82	78.80
30	42.08	992.66	50.35	63	683.47	18925.00	79.70
31	45.79	1139.64	50.47	64	743.71	19916.79	80.31
32	49.82	1302.89	50.54	65	809.27	20809.42	80.67
33	54.21	1482.26	50.54	66	880.6	21602.56	80.81
34	58.99	1677.00	50.47	67	958.22	22298.96	80.75
35	64.19	1885.58	50.31	68	1042.68	22904.00	80.52
36	69.85	2105.62	50.07	69	1134.59	23425.15	80.13
37	76.01	2333.76	49.72	70	1234.6	23871.77	79.60
38	82.71	2565.68	49.26	71	1343.43	24254.49	78.95
39	90	2796.07	48.69	72	1461.84	24585.12	78.18
40	97.93	3018.78	47.99	73	1590.7	24876.24	77.33
41	106.56	3228.48	47.16	74	1730.91	25141.14	76.40
42	115.95	3426.4	46.28	75	1883.49	25393.55	75.41
43	126.18	3616.97	45.43	76	2049.51	25645.21	74.40
44	137.3	3806.28	44.68	77	2230.16	25898.50	73.36
45	149.4	4002.06	44.11	78	2426.74	26153.41	72.29
46	162.57	4213.69	43.81	79	2640.65	26409.93	71.20
47	176.9	4452.43	43.84	80	2873.41	26668.07	70.08
48	192.49	4731.91	44.28	81	3126.69	26927.80	68.94
49	209.46	5068.90	45.23	82	3402.3	27189.08	67.77
50	227.92	5484.53	46.76	83	3702.19	27451.99	66.58
51	248.01	6001.64	48.92	84	4028.53	27716.47	65.37
52	269.87	6630.26	51.63				

1) All cause death per 100,000 persons
2) Prevalence of diabetes mellitus per 100,000 persons
3) Proportion of recognition per 100 persons with diabetes mellitus

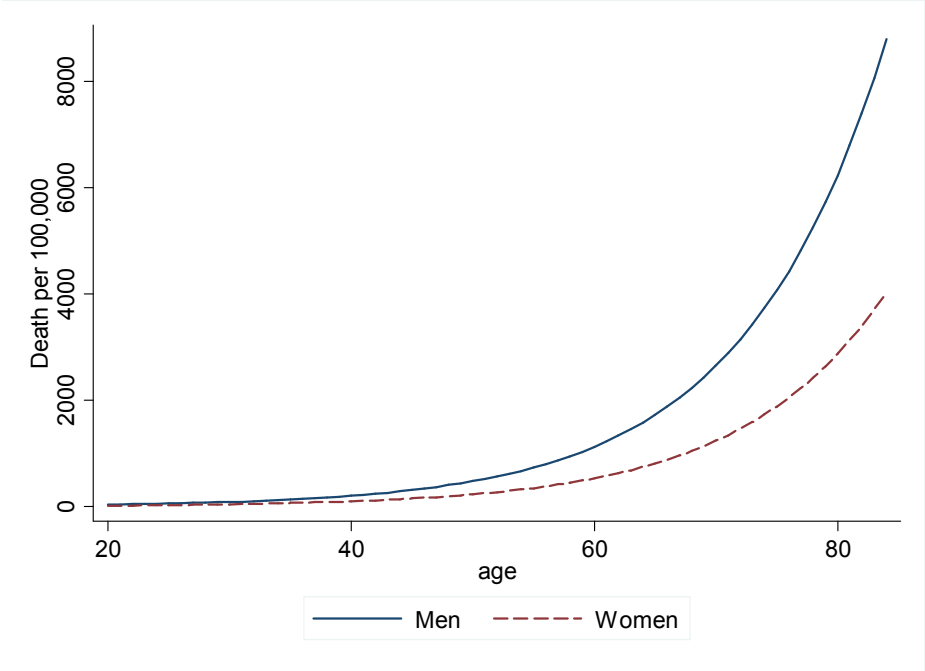

Figure 3.1 All-cause death per 100,000 persons

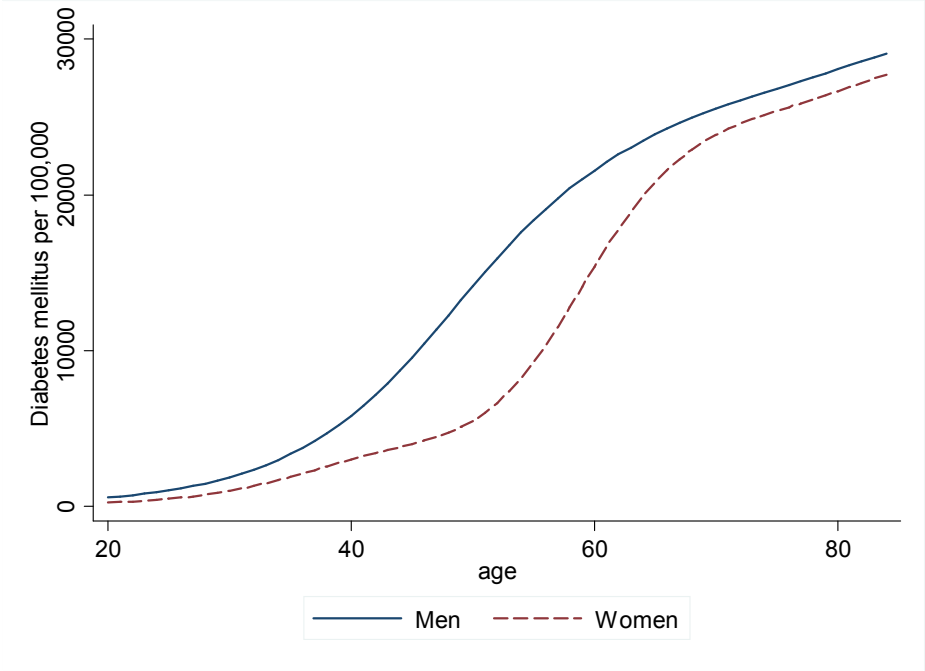

Figure 3.2 Prevalence of diabetes mellitus per 100,000 persons

3.1.3 당뇨병에 대한 인지율

당뇨병에 대한 인지율은 남녀에서 매우 다른 양상으로 나타났다(Figure 3.3). 남성은 20대에 10% 미만의 매우 낮은 당뇨병 인지율을 보여주었으며, 30대에서 20%를 넘지 못하였다. 40세부터 가파르게 증가하여, 49세 정도에 50%를 넘는 인지율율을 보여주었다. 인지율은 60세 까지는 지속적으로 증가하여 80% 가까이 되었으나, 그 이후로 증가가 둔화되다가 65세 이후에는 오히려 감소하는 양상을 보여주었다(Table 3.1). 8

반면에, 여성에서는 20~40세에 걸쳐 꾸준히 50% 수준의 당뇨병 인지율을 보여주었으나, 40대에는 오히려 5% 정도 인지율이 감소하는 양상으로 나타났다. 이에 따라 40세 이전에는 여성에서의 인지율이 남성에 비하여 현저히 높았지만, 50대에는 오히려 남성에 비해 인지율이 약간 낮게 나타났다. 하지만, 65세 이후로는 당뇨병의 인지율이 남성과 매우 유사하게 나타났다.

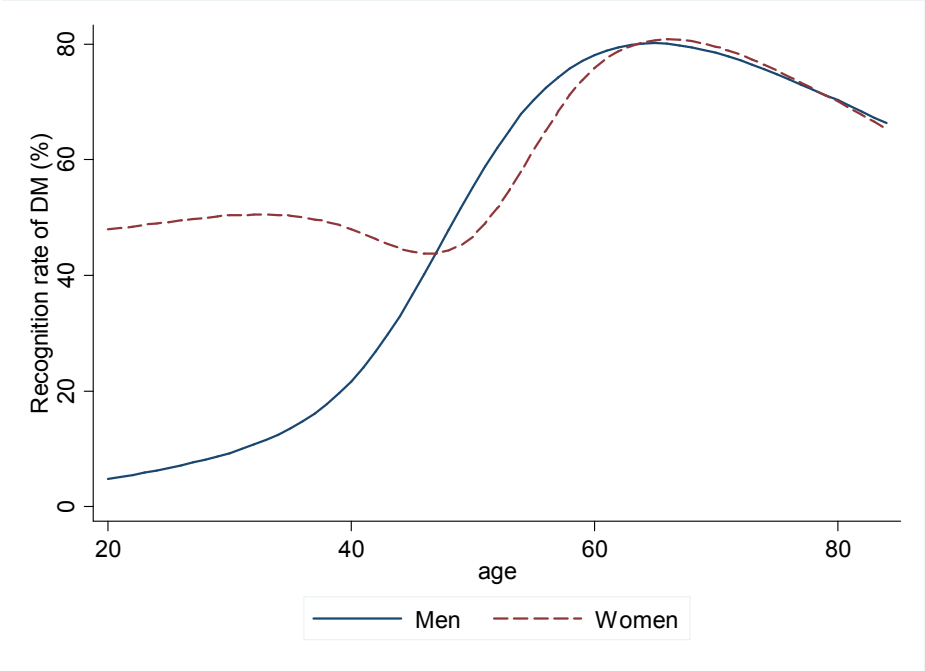

Figure 3.3 Recognition rate in persons with diabetes mellitus

3.1.4 발병율과 기타 모수들

연구 방법(2.2.2절)에서 언급한 바와 같이, 본 연구에서는 연구의 가정에 따라 발병율에 대한 직접적인 자료를 얻는 것이 쉽지 않았기 때문에, 간접적으로 당뇨병의 발생율을 계산하였다(Table 3.3, 3.4의 B_t). 사망률, 유병률 같은 자료들과 달리, 발병율은 직접적인 평활화(smoothing)를 하지 않았지만, 평활화가 된 다른 자료들로부터 산출되었기 때문에 전체적으로 부드러운 곡선의 형태를 보여주었다(Figure 3.4). 이렇게 간접적으로 산출된 당뇨병의 발생율은 남성에서 20세에 연간 0.1% 미만에서 시작하여, 49세에는 연간 1.1%까지 증가하였지만, 이후 60대 후반에는 0.7% 수준으로 다시 감소하였다. 70대 이후에는 지속적으로 증가하는 양상이었다. 여성에서는 33세가 되어야 연간 0.2% 수준이 되었으나, 전체적으로 45세 정도까지는 크게 증가하지 않는 양상이었다. 하지만, 45세 이후부터는 급격히 증가하여 59세에는 1.6% 수준까지 증가하지만, 60세 이후부터는 다시 급격히 감소하는 양상이었다. 60세 이후에 발병률이 가장 낮은 연령은 73세로 연간 약 0.6% 정도 이었다.

그 밖에, 0 또는 1과 같은 자명한 모수를 제외하고 전이행렬의 모수들을 표로 정리한 결과는 Table 3.3와 3.4에서 보는 바와 같다. 관측행렬의 모수들은 0 또는 1과 같은 자명한 모수를 제외하고는, Table 3.1과 3.2의 “인지율”과 “1-인지율”로 간단히 정의되기 때문에 별도로 표로 정리하지는 않았다.

Table 3.3 Parameters of transitional matrix in men with standard risk (trivial parameters are omitted).

Time (age)	At[1)]	Bt[1)]	Ct[1)]	Dt[1)]	Et[1)]
20	0.998911	0.000725	0.999411	0.000363	0.000589
21	0.998785	0.000819	0.999359	0.000396	0.000641
22	0.998644	0.000925	0.999302	0.000431	0.000698
23	0.998486	0.001044	0.99924	0.000469	0.00076
24	0.99831	0.001179	0.999172	0.000511	0.000828
25	0.998113	0.001331	0.999099	0.000556	0.000901
26	0.997893	0.001501	0.999019	0.000605	0.000981
27	0.99765	0.001692	0.998932	0.000659	0.001068
28	0.997379	0.001904	0.998838	0.000717	0.001162
29	0.99708	0.00214	0.998736	0.00078	0.001264
30	0.996748	0.002402	0.998624	0.000849	0.001376
31	0.996383	0.002693	0.998503	0.000924	0.001497
32	0.995981	0.003014	0.998372	0.001005	0.001628
33	0.99554	0.003368	0.99823	0.001093	0.001771
34	0.995056	0.003756	0.998075	0.001188	0.001925
35	0.994528	0.00418	0.997907	0.001292	0.002093
36	0.993953	0.004643	0.997726	0.001404	0.002274
37	0.993329	0.005146	0.997529	0.001525	0.002471
38	0.992654	0.005689	0.997316	0.001657	0.002685
39	0.991927	0.006274	0.997085	0.0018	0.002915
40	0.99115	0.006896	0.996835	0.001954	0.003165
41	0.990344	0.007535	0.996565	0.00212	0.003435
42	0.989525	0.008175	0.996273	0.0023	0.003727
43	0.988706	0.008799	0.995958	0.002495	0.004042
44	0.987906	0.009388	0.995617	0.002706	0.004383
45	0.987145	0.009922	0.995249	0.002933	0.004751
46	0.986444	0.010377	0.99485	0.003179	0.00515
47	0.985827	0.010728	0.994419	0.003445	0.005581
48	0.985318	0.010949	0.993952	0.003733	0.006048
49	0.98494	0.011015	0.993447	0.004045	0.006554
50	0.984693	0.010923	0.992898	0.004384	0.007102
51	0.984505	0.010744	0.992303	0.004752	0.007698
52	0.984343	0.010506	0.991655	0.005151	0.008345

Time (age)	At[1)]	Bt[1)]	Ct[1)]	Dt[1)]	Et[1)]
53	0.984194	0.010221	0.990952	0.005585	0.009048
54	0.98404	0.009902	0.990187	0.006058	0.009813
55	0.983866	0.009563	0.989354	0.006572	0.010646
56	0.983654	0.009216	0.988448	0.007131	0.011552
57	0.983387	0.008874	0.987462	0.00774	0.012538
58	0.983048	0.00855	0.986389	0.008402	0.013612
59	0.982621	0.008256	0.985219	0.009124	0.014781
60	0.982094	0.007996	0.983947	0.00991	0.016053
61	0.981479	0.007756	0.982561	0.010765	0.017439
62	0.980767	0.007536	0.981051	0.011697	0.018949
63	0.979947	0.007342	0.979408	0.012711	0.020592
64	0.979003	0.00718	0.977617	0.013816	0.022383
65	0.977923	0.007057	0.975668	0.01502	0.024332
66	0.976693	0.006976	0.973544	0.016331	0.026456
67	0.975297	0.006945	0.971231	0.017759	0.028769
68	0.973719	0.006967	0.968711	0.019314	0.031289
69	0.971945	0.007047	0.965967	0.021008	0.034033
70	0.969957	0.007191	0.962979	0.022852	0.037021
71	0.967736	0.007403	0.959726	0.024861	0.040274
72	0.965264	0.007689	0.956183	0.027047	0.043817
73	0.96252	0.008052	0.952327	0.029428	0.047673
74	0.959482	0.008499	0.94813	0.032019	0.05187
75	0.956135	0.009028	0.943563	0.034838	0.056437
76	0.952482	0.009613	0.938593	0.037905	0.061407
77	0.948505	0.010253	0.933187	0.041243	0.066813
78	0.944175	0.010952	0.927305	0.044874	0.072695
79	0.93946	0.011716	0.920905	0.048824	0.079095
80	0.934326	0.012552	0.913943	0.053122	0.086057
81	0.928736	0.013467	0.906368	0.057798	0.093632
82	0.922648	0.014467	0.898127	0.062884	0.101873
83	0.916019	0.015562	0.889162	0.068419	0.110839
84	0	0	0	1	1

1) At, Bt, Ct, Dt and Et are parameters of transitional matrix at time t. Formula (5)~(9) in Chapter 2.2.

Table 3.4 Parameters of transitional matrix in women with standard risk (trivial parameters omitted)

Time (age)	At[1)]	Bt[1)]	Ct[1)]	Dt[1)]	Et[1)]
20	0.999453	0.000367	0.999708	0.000181	0.000292
21	0.999378	0.000426	0.999682	0.000196	0.000318
22	0.999292	0.000494	0.999654	0.000214	0.000346
23	0.999194	0.000574	0.999624	0.000232	0.000376
24	0.999081	0.000666	0.99959	0.000253	0.00041
25	0.998953	0.000772	0.999555	0.000275	0.000445
26	0.998809	0.000892	0.999516	0.000299	0.000484
27	0.99865	0.001025	0.999473	0.000325	0.000527
28	0.998477	0.001169	0.999427	0.000354	0.000573
29	0.998292	0.001324	0.999377	0.000385	0.000623
30	0.998095	0.001486	0.999323	0.000418	0.000678
31	0.997892	0.001654	0.999263	0.000455	0.000737
32	0.997685	0.00182	0.999199	0.000494	0.000801
33	0.997482	0.001981	0.99913	0.000537	0.00087
34	0.99729	0.002126	0.999054	0.000584	0.000946
35	0.997117	0.002249	0.998972	0.000634	0.001028
36	0.996973	0.002338	0.998883	0.000689	0.001117
37	0.996867	0.002384	0.998786	0.000749	0.001214
38	0.99681	0.002376	0.998681	0.000814	0.001319
39	0.996811	0.002304	0.998567	0.000885	0.001433
40	0.996861	0.002178	0.998443	0.000961	0.001557
41	0.996891	0.002064	0.998308	0.001045	0.001692
42	0.99687	0.001995	0.998161	0.001135	0.001839
43	0.996777	0.001989	0.998001	0.001234	0.001999
44	0.996595	0.002064	0.997827	0.001341	0.002173
45	0.996305	0.002237	0.997638	0.001458	0.002362
46	0.995886	0.00253	0.997433	0.001584	0.002567
47	0.995311	0.002967	0.997211	0.001721	0.002789
48	0.994545	0.003585	0.996971	0.00187	0.003029
49	0.993536	0.004433	0.99671	0.002031	0.00329
50	0.992262	0.005534	0.996429	0.002204	0.003571
51	0.990849	0.00676	0.996126	0.002391	0.003874
52	0.989337	0.008071	0.995801	0.002592	0.004199

Time (age)	At[1)]	Bt[1)]	Ct[1)]	Dt[1)]	Et[1)]
53	0.98774	0.009452	0.995451	0.002808	0.004549
54	0.986095	0.010865	0.995075	0.00304	0.004925
55	0.984458	0.012253	0.994672	0.003289	0.005328
56	0.982914	0.01353	0.994239	0.003556	0.005761
57	0.981569	0.014588	0.993775	0.003843	0.006225
58	0.980553	0.015296	0.993276	0.004151	0.006724
59	0.980007	0.01551	0.992738	0.004483	0.007262
60	0.979984	0.015174	0.992156	0.004842	0.007844
61	0.980211	0.014558	0.991525	0.005232	0.008475
62	0.98057	0.013774	0.990838	0.005656	0.009162
63	0.981016	0.012867	0.990091	0.006117	0.009909
64	0.9815	0.011881	0.989276	0.00662	0.010724
65	0.981973	0.010859	0.988388	0.007168	0.011612
66	0.982389	0.009845	0.987419	0.007766	0.012581
67	0.982706	0.008876	0.986362	0.008418	0.013638
68	0.982884	0.007986	0.985209	0.00913	0.014791
69	0.982888	0.007205	0.983951	0.009907	0.016049
70	0.982689	0.006557	0.982578	0.010754	0.017422
71	0.982259	0.006063	0.981081	0.011678	0.018919
72	0.981578	0.005737	0.979451	0.012685	0.020549
73	0.980623	0.005596	0.977674	0.013781	0.022326
74	0.979378	0.005648	0.975741	0.014975	0.024259
75	0.977856	0.005872	0.973638	0.016273	0.026362
76	0.976168	0.006148	0.971353	0.017683	0.028647
77	0.974335	0.006449	0.96887	0.019216	0.03113
78	0.972343	0.006775	0.966172	0.020881	0.033828
79	0.970178	0.007131	0.963241	0.022691	0.036759
80	0.967826	0.007517	0.960055	0.024657	0.039945
81	0.965269	0.007937	0.956594	0.026794	0.043406
82	0.962489	0.008396	0.952834	0.029115	0.047166
83	0.959468	0.008895	0.948748	0.031637	0.051252
84	0	0	0	1	1

1) At, Bt, Ct, Dt and Et are parameters of transitional matrix at time t. Formula (5)~(9) in Chapter 2.2.

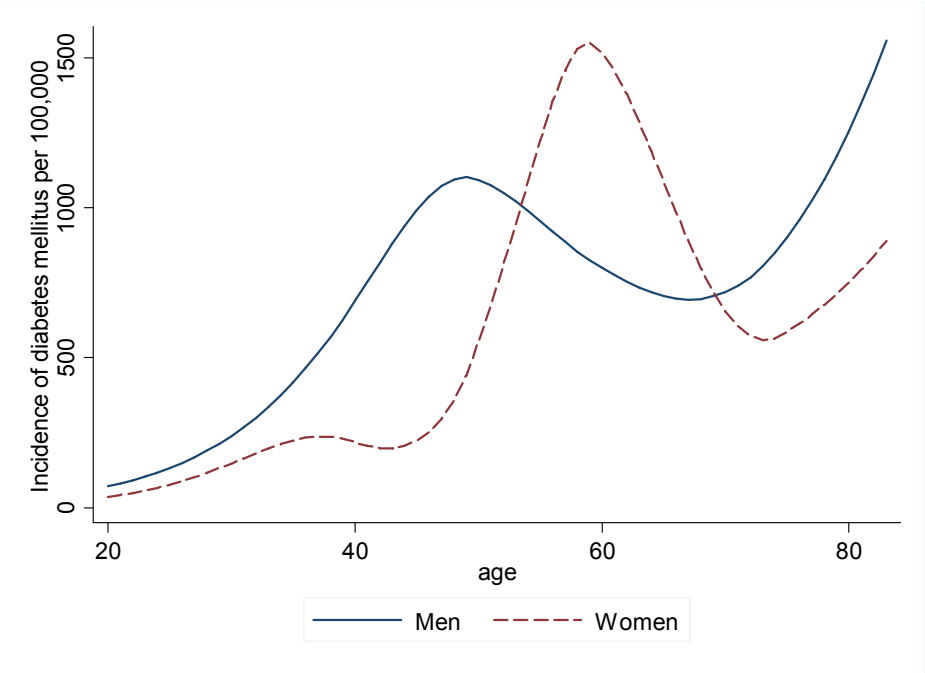

Figure 3.4 Incidence of diabetes mellitus per 100,000 persons.

3.2 개인별 위험요인에 따른 위험비 산출

3.2.1 위험요인의 선정

“Training set”에서 성별을 포함하여, 연령, 고혈압 여부, 허리둘레 이상, 현재 흡연 여부, 당뇨병 가족력, 체질량 지수 그룹과 1일 평균 음주량을 이용한 다변량 로지스틱 회귀분석을 한 결과, 음주량(p-value 0.37)을 제외하고, 모두 유의수준 0.05로 유의하였다. 따라서, 최종적으로 위험요인으로서 성별, 연령, 고혈압, 여부, 허리둘레 이상, 현재 흡연 여부, 당뇨병 가족력과 체질량 지수 그룹을 선정하였다.

3.2.2 위험요인 별 오즈비

본 연구의 Markov chain 및 HMM은 성별에 따라 나뉘어져 구조화되어 있기 때문에 오즈비도 성별에 따라 개별화된 오즈비를 산출하였다(Table 3.5). 성별에 따라 나뉘었을 때에, 고혈압 유무와 연령에 대한 오즈비는 남녀가 거의 유사하였다. 하지만 비만도에 대해서는 남성은 여성에 비해 허리둘레에 의한 영향이 작으며, 체질량 지수에 의한 영향이 큰 것으로 나타났다. 남녀를 구분함에 따라 남성의 허리둘레, 여성의 흡연여부, 그리고 남성의 체질량 지수 일부 구간의 오즈비가 통계적으로 유의하지 않은 것으로 나타났다. 하지만, 표본 크기의 감소로 인하여 검정력이 감소하여 나타난 결과일 수 있으며, 전체적으로 남녀의 양상이 비슷하기 때문에 모형의 일관성을 유지하기 위하여 위험요인에서 제외하지는 않았다.

Table 3.5 Association between diabetes mellitus and risk factors.

	Odds ratio (95% confidential interval)	
	Men	Women
Hypertension		
No	1.00 (Reference)	1.00 (Reference)
Yes	1.82 (1.38 - 2.41)	1.82 (1.33 - 2.51)
Waist circumference		
Normal	1.00 (Reference)	1.00 (Reference)
Abnormal	1.41 (0.99 - 2.01)	2.24 (1.54 - 3.24)
Current smoking		
No	1.00 (Reference)	1.00 (Reference)
Yes	1.34 (1.01 - 1.77)	1.20 (0.63 - 2.29)
Family history of diabetes mellitus		
No	1.00 (Reference)	1.00 (Reference)
Yes	3.99 (2.92 - 5.45)	3.24 (2.35 - 4.45)
Body mass index (kg/m^2)		
~24.9	1.00 (Reference)	1.00 (Reference)
25.0~29.9	1.28 (0.90 - 1.83)	1.52 (1.05 - 2.19)
30.0~	3.04 (1.42 - 6.53)	2.41 (1.35 - 4.28)
Age		
per 1 year	1.06 (1.05 - 1.08)	1.06 (1.05 - 1.07)

3.2.3 위험요인의 분포

남성의 경우, 고혈압의 유병률은 20대에서 7% 정도이나 70세 이상에서는 50%를 넘는다. 복부 비만율은 60대에서 29%로 가장 높았으며 현재 흡연율은 30대에서 58%로 가장 높았다. 체질량 지수는 25이상의 비율이 3~40대에서 가장 높았다(Table 3.6). 여성의 경우, 고혈압의 유병률은 20대에서 1% 미만이나 40대부터 그 비율이 크게 증가하여, 70세 이상에서는 64%에 달했다. 복부 비만율은 남성에서와 같이 60대에서 41%로 가장 높았으며, 현재 흡연율은 20대를 제외하고는 전 연령구간에서 10% 미만이었다. 체질량 지수는 25이상의 비율이 60대에서 37%로 가장 높았다(Table 3.7).

3.2.4 위험요인 별 오즈비의 표준화

위험요인의 분포와 위험요인 별 오즈비를 통해서 표준화 오즈비를 구할 수 있다(Table 3.8, 3.9). 표준화 오즈비는 연령대에 따라 일관된 경향이 없이 다양한 형태로 나타났다. 남녀 공통적으로 고혈압은 연령에 따라 유병률이 증가하기 때문에, 연령대의 증가에 따라 표준화 오즈비는 낮아지는 경향이 있었다. 체질량지수는 연령대에 따라 표준화 오즈비가 크게 차이가 나지는 않았지만, 연령대가 증가할수록 남성은 약간 증가하고, 여성은 약간 감소하는 경향을 보였다. 반면에 가족력의 경우에는 특별한 경향성이 나타나지 않았다. 즉, 표준화 오즈비는 고정된 오즈비와 성별, 연령대별로 다른 위험요인의 분포에 의해 계산되므로 위험요인의 분포와 같은 양상으로 나타났다.

Table 3.6 Proportions of risk factors by age group in men.

	Age group (year old)					
	20-29	30-39	40-49	50-59	60-69	70-84
Hypertension						
No	92.94	87.84	75.33	66.24	56.01	48.28
Yes	7.06	12.16	24.67	33.76	43.99	51.72
Waist circumference						
Normal	86.56	75.31	74.50	78.77	70.71	72.69
Abnormal	13.44	24.69	25.50	21.23	29.29	27.31
Current smoking						
No	51.97	42.27	51.00	54.73	71.47	77.39
Yes	48.03	57.73	49.00	45.27	28.53	22.61
Family history of diabetes mellitus						
No	84.00	79.45	78.77	85.51	91.39	94.81
Yes	16.00	20.55	21.23	14.49	8.61	5.19
Body mass index (kg/m^2)						
~24.9	71.65	58.87	59.47	73.66	62.94	78.16
25.0~29.9	23.62	36.95	36.08	25.83	36.47	21.07
30.0~	4.72	4.18	4.45	0.51	0.59	0.77

Table 3.7 Proportions of risk factors by age group in women.

	Age group (year old)					
	20-29	30-39	40-49	50-59	60-69	70-84
Hypertension						
No	99.47	98.59	89.05	70.1	47.45	35.48
Yes	0.53	1.41	10.95	29.9	52.55	64.52
Waist circumference						
Normal	88.03	88.5	85.82	70.82	58.98	62.54
Abnormal	11.97	11.5	14.18	29.18	41.02	37.46
Current smoking						
No	88.86	93.57	94.68	95.89	97.1	95
Yes	11.14	6.43	5.32	4.11	2.9	5
Family history of diabetes mellitus						
No	83.74	76.01	72.6	81.47	88.69	92.2
Yes	16.26	23.99	27.4	18.53	11.31	7.8
Body mass index (kg/m^2)						
~24.9	82.75	83.18	75.82	66.12	62.53	67.74
25.0~29.9	13.48	13.52	19.96	28.97	34.37	28.45
30.0~	3.77	3.30	4.21	4.91	3.10	3.81

Table 3.8 Standardized odds ratios of risk factors by age group in men.

	Age group (year old)					
	20-29	30-39	40-49	50-59	60-69	70-84
Hypertension						
No	0.96	0.93	0.86	0.82	0.77	0.73
Yes	1.75	1.70	1.57	1.49	1.40	1.34
Waist circumference						
Normal	0.95	0.92	0.92	0.93	0.90	0.91
Abnormal	1.35	1.30	1.29	1.31	1.27	1.28
Current smoking						
No	0.87	0.84	0.87	0.88	0.92	0.94
Yes	1.16	1.13	1.16	1.17	1.23	1.25
Family history of diabetes mellitus						
No	0.80	0.75	0.75	0.82	0.89	0.93
Yes	3.20	3.00	2.97	3.26	3.54	3.71
Body mass index (kg/m^2)						
~24.9	0.89	0.87	0.87	0.93	0.91	0.94
25.0~29.9	1.15	1.12	1.12	1.20	1.16	1.21
30.0~	2.72	2.65	2.65	2.84	2.76	2.86

Table 3.9 Standardized odds ratios of risk factors by age group in women.

	Age group (year old)					
	20-29	30-39	40-49	50-59	60-69	70-84
Hypertension						
No	1.00	0.99	0.94	0.84	0.73	0.68
Yes	1.82	1.81	1.71	1.52	1.33	1.24
Waist circumference						
Normal	0.91	0.91	0.89	0.79	0.72	0.74
Abnormal	2.03	2.04	2.00	1.77	1.61	1.65
Current smoking						
No	0.98	0.99	0.99	0.99	0.99	0.99
Yes	1.17	1.18	1.19	1.19	1.19	1.19
Family history of diabetes mellitus						
No	0.83	0.75	0.72	0.80	0.88	0.91
Yes	2.67	2.44	2.35	2.60	2.83	2.95
Body mass index (kg/m^2)						
~24.9	0.91	0.92	0.89	0.85	0.84	0.86
25.0~29.9	1.39	1.39	1.34	1.29	1.28	1.30
30.0~	2.20	2.21	2.13	2.04	2.03	2.07

3.3 표준 위험 집단의 당뇨병 위험도 추정

3.3.1 질병 상태의 기대 기간

남성의 경우 Table 3.10에서 보는 바와 같이, 현재 당뇨병이 없고 표준 위험도를 갖고 있는 20세의 경우에는 48.74년의 기간 동안 당뇨병이 없는 상태로 지낼 것이 기대가 되며, 당뇨병이 있지만 인지를 못하는 기간은 2.37년, 당뇨병이 있으면서 인지를 하는 기간은 4.63년으로 기대된다. 당뇨병 유병 기간의 기대값은 두 당뇨병 기간의 합인 7.00년이 되고, 기대 여명은 모든 기간의 합인 55.74년이 된다(Figure 3.5). 여성의 경우도 비슷하게 해석 가능하며, 인지하고 있는 당뇨병 기간과 인지하지 못한 당뇨병 기간이 모두 같은 연령의 남성에 비해서 짧은 것을 관찰할 수 있다(Table 3.11, Figure 3.7).

3.3.2. 평생 당뇨병 위험도

Table 3.9에서 보는 바와 같이, 현재 당뇨병이 없고 표준 위험도를 갖고 있는 20세 남성의 경우에는 평생 당뇨병 위험도가 31.6% 정도이며, 당뇨병의 발병률이 높은 기간인 4~50대를 지난 후인 60세의 경우에는 당뇨병의 평생 위험도가 15.3%에 불과하였다. 20세 여성에서는 당뇨병 평생 위험도가 28.8%로 남성에 비하여 약간 낮은 수준이나(Table 3.11), 당뇨병 발생률이 높은 기간은 60세 전후이기 때문에 그 이전인 50세에는 오히려 당뇨병 평생 위험도가 남성에 비해 높은 것을 관찰할 수 있다(Figure 3.6, 3.8).

Table 3.10 Expected durations of disease status and lifetime risk of diabetes mellitus in men with standard risk.

Age	Duration of status (years)			Lifetime risk (%)
	Normal	DM without insight	DM with insight	
20	48.74	2.37	4.63	31.62
25	44.06	2.25	4.56	31.36
30	39.55	2.08	4.43	30.88
35	35.29	1.83	4.19	29.97
40	31.39	1.50	3.79	28.35
45	27.92	1.12	3.14	25.67
50	24.75	0.77	2.33	22.02
55	21.53	0.51	1.58	18.34
60	18.16	0.35	1.04	15.26
65	14.71	0.24	0.68	12.81
70	11.27	0.17	0.43	10.74
75	7.87	0.10	0.24	8.50
80	4.34	0.04	0.08	5.01

Table 3.11 Expected durations of disease status and lifetime risk of diabetes mellitus in women with standard risk.

Age	Duration of status (years)			Lifetime risk (%)
	Normal	DM without insight	DM with insight	
20	53.92	1.78	4.40	28.77
25	49.11	1.73	4.33	28.62
30	44.42	1.64	4.20	28.30
35	39.90	1.50	3.99	27.71
40	35.47	1.34	3.75	26.97
45	31.00	1.24	3.57	26.37
50	26.68	1.11	3.32	25.42
55	22.98	0.87	2.69	22.63
60	19.89	0.55	1.67	17.27
65	16.63	0.29	0.82	11.79
70	12.89	0.15	0.38	8.14
75	8.86	0.08	0.18	5.80
80	4.66	0.02	0.05	3.10

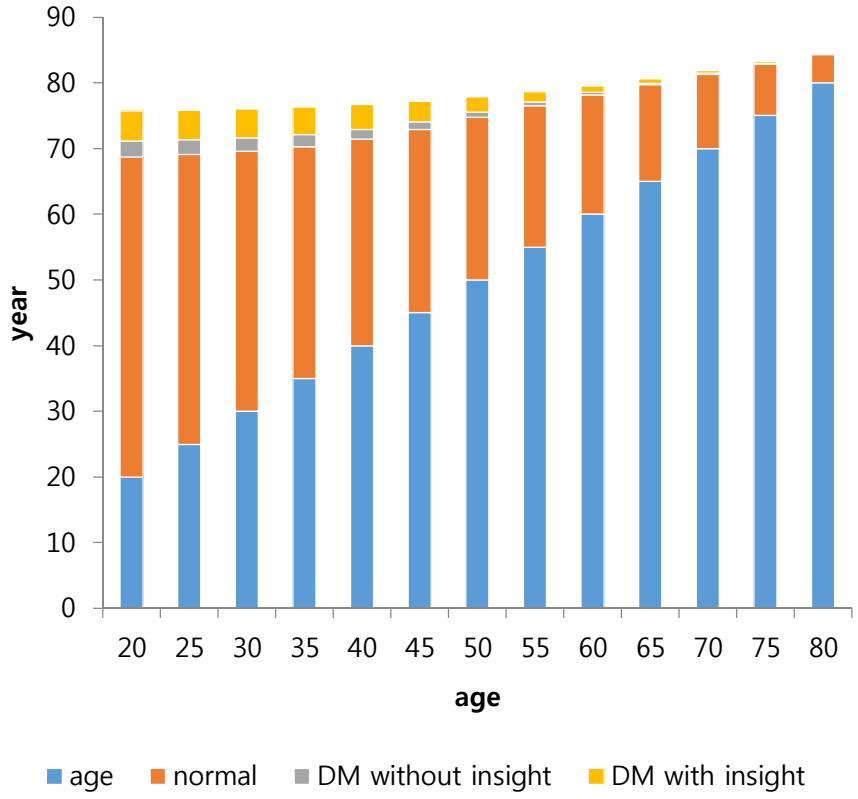

Figure 3.5 Expected duration of disease status by current age in men with standard risk.

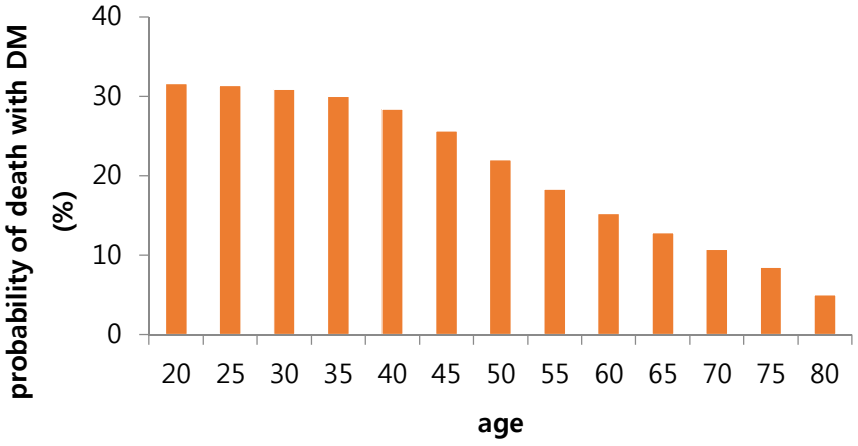

Figure 3.6 Lifetime risk of diabetes mellitus by current age in men with standard risk.

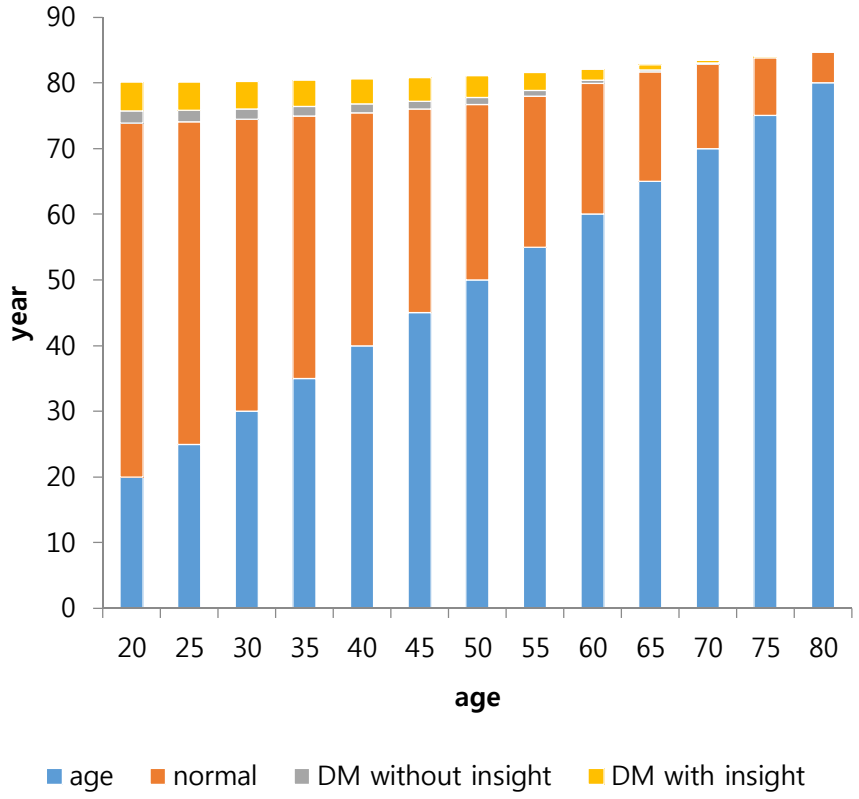

Figure 3.7 Expected duration of disease status by current age in women with standard risk.

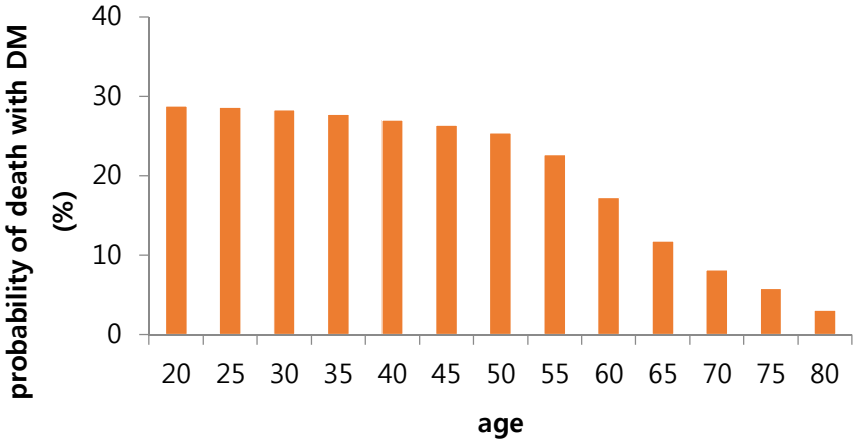

Figure 3.8 Lifetime risk of diabetes mellitus by current age in women with standard risk.

3.3.3 기대 여명의 비교

각 단계의 기대기간을 모두 더함으로써 현재 연령에서의 기대 여명을 구할 수 있다. 본 연구의 기대 여명을 2010, 2011년 통계청의 보고서의 평균값과 비교하였을 때, 남성은 약 2년, 여성은 약 4.5년 정도 짧게 나타났다(Table 3.12).

20세부터 83세까지, 당뇨병이 없으며 평균 위험도를 갖는 대상자의 기대 여명을 본 연구의 방법으로 추정한 결과와 통계청 보고서의 기대 여명의 차이를 그래프로 그린 것은 Figure 3.9와 같다. 남성에서는 20~35세에서는 기대 여명이 2.2~2.3년의 일정한 차이를 보였으나, 65세에 1.7년까지 감소하였으며, 70세 이후에는 급격히 증가하여 83세에는 4.3년에 달하였다. 여성에서는 20~50세에서는 기대 여명이 4.5~4.6년의 일정한 차이를 보였으나, 70세에 4.0년까지 감소하였으며, 이후에는 급격히 증가하여 83세에는 6.1년에 달하였다. 남녀에서 기대 여명의 차이가 감소하는 구간은 발병율(Figure 3.4)이 10만명당 대략적으로 500명이 넘는 구간과 비슷하게 일치하는 것을 관찰할 수 있었다.

Table 3.12 Comparison of lifetime expectancy between Markov chain and conventional model.

Age	Men			Women		
	A (years)	B (years)	A-B (years)	A (years)	B (years)	A-B (years)
20	57.97	55.74	2.22	64.69	60.09	4.59
25	53.12	50.87	2.25	59.79	55.16	4.63
30	48.31	46.05	2.26	54.92	50.26	4.66
35	43.51	41.32	2.19	50.06	45.39	4.67
40	38.78	36.68	2.09	45.23	40.57	4.65
45	34.15	32.19	1.96	40.42	35.81	4.62
50	29.68	27.84	1.84	35.66	31.12	4.54
55	25.39	23.63	1.76	30.96	26.54	4.42
60	21.25	19.55	1.70	26.32	22.1	4.21
65	17.29	15.63	1.66	21.78	17.73	4.05
70	13.6	11.87	1.73	17.44	13.41	4.03
75	10.35	8.21	2.14	13.41	9.11	4.29
80	7.61	4.45	3.16	9.89	4.74	5.15

A; Conventional model as report of Statistics Korea
B; Markov chain of this study

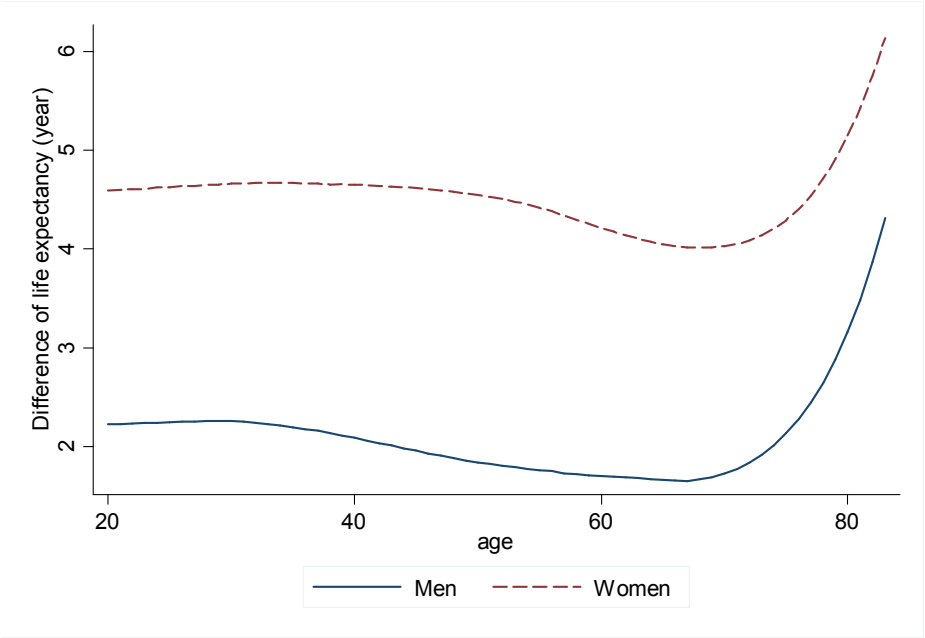

Figure 3.9 Difference of lifetime expectancy between Markov chain and conventional model.

3.4 개인별 위험요인에 따른 위험도 추정

3.4.1 위험도 추정을 위한 표본 선정

개인별 위험요인에 따라 위험도를 추정하기 위하여 “validation set”에서 조사 시점에서 당뇨병이 없는 10명을 임의로 추출하였다(Table 3.13). 추출된 대상은 남녀 각각 5명씩 이었으며, 연령 분포는 30대 2명, 40대 3명, 50대 2명, 70대 2명, 80대 1명 이었다. 연령과 성별을 제외하였을 때, 위험요인의 개수가 가장 많은 대상자는 P5(허리둘레, 체질량지수, 가족력)와 P7(허리둘레, 체질량지수, 고혈압)이었으며 가장 적은 대상자는 P2(없음)와 P6(없음) 이었다.

3.4.2 개인별 위험도 추정

Table 3.13에 명시된 개인별 위험요인을 본 연구에 적용한 결과 개인별 위험도를 추정할 수 있었다(Table 3.14, Figure 3.10, 3.11). 이 중에 당뇨병 유병 기간의 기대값이 가장 긴 대상자는 P5(인지되지 않은 기간 5.98년, 인지된 기간 15.14년)이었으며, 평생 당뇨병 위험도가 가장 높은 대상자도 P5(86.13%)이었다. 위험요인의 개수가 3개로 많았던 P7은 73세의 시점에서 당뇨병이 없었기 때문에 당뇨병의 유병기간의 기대값이 길지 않았으며(인지되지 않은 기간 0.22년, 인지된 기간 0.54년), 평생 당뇨병 위험도도 낮은 수준(16.24%)이었다. 하지만, P6과 비교해 보았을 때, 당뇨병 유병기간의 기대값이 짧음에도 불구하고, 평생 당뇨병 위험도가 높은 것으로 보아, 당뇨병 유병 기간의 기대값과 평생 당뇨병 위험도가 비례하지는 않음이 관찰되었다.

Table 3.13 Sample data for individual risk estimation.

ID	P1	P2	P3	P4	P5
Sex	Female	Female	Male	Male	Female
Age	44	36	46	82	35
Waist circumference (cm)	78.3	69.7	84.3	90	96
Body mass index (kg/m^2)	24.6	21.9	25.2	23.6	26.4
Family history of diabetes mellitus	Yes	No	Yes	No	Yes
Hypertension	No	No	No	Yes	No
Current smoking	No	No	Yes	No	No
ID	P6	P7	P8	P9	P10
Sex	Female	Male	Female	Male	Male
Age	57	73	58	48	77
Waist circumference (cm)	82.6	91.2	84.5	93	69
Body mass index (kg/m^2)	24.8	26.5	25.5	23.8	21.3
Family history of diabetes mellitus	No	No	No	No	No
Hypertension	No	Yes	No	Yes	Yes
Current smoking	No	No	No	No	No

Table 3.14 Expected durations of disease status and lifetime risk of diabetes mellitus according to risk factors assuming no current DM.

ID	Age	Duration of status (years)			Lifetime risk (%)
		Normal	DM without insight	DM with insight	
P1	44	28.86	1.94	5.77	40.80
P2	36	41.04	0.93	2.58	18.19
P3	46	20.54	2.61	7.41	56.72
P4	82	2.75	0.02	0.04	4.01
P5	35	23.27	5.98	15.14	86.13
P6	57	23.42	0.35	1.09	9.95
P7	73	8.87	0.22	0.54	16.24
P8	58	21.98	0.48	1.46	13.85
P9	48	25.5	1.01	2.99	26.33
P10	77	6.49	0.07	0.17	7.310

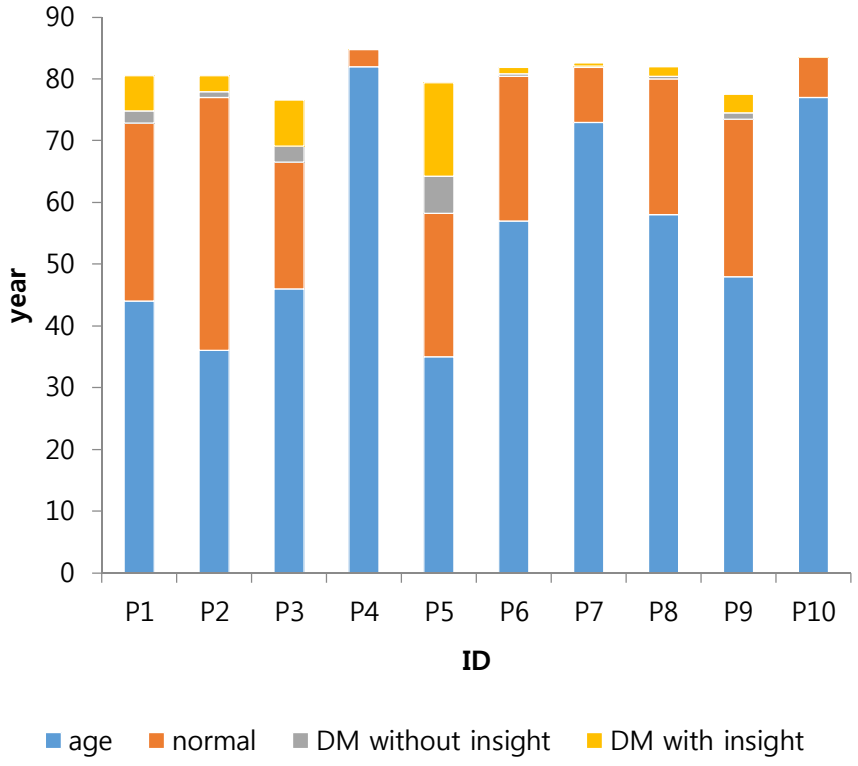

Figure 3.10 Expected duration of disease status according to individual risk factors.

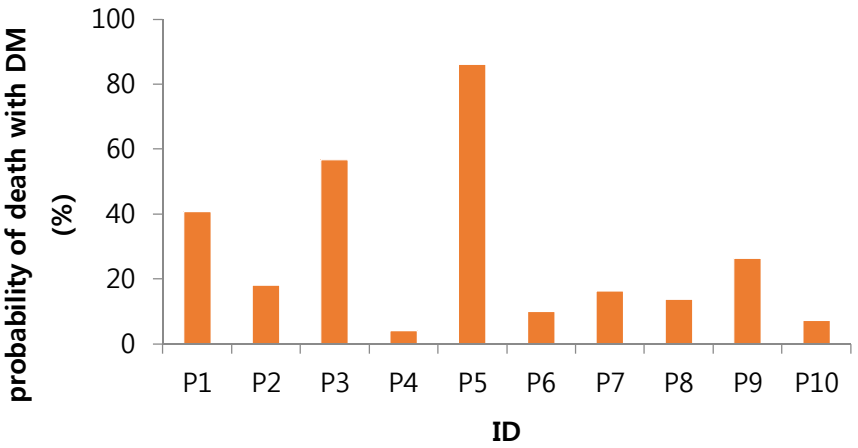

Figure 3.11 Lifetime risk of diabetes mellitus according to individual risk factors.

3.4.3 교정 위험도 추정

한 개인의 위험요인이 교정됨으로써 얻을 수 있는 기대효과를 시뮬레이션하기 위한 대상으로 위험 요인의 개수가 가장 많고 연령이 낮은 P5를 선정하였다. P5는 기초 상태에서 허리둘레가 기준치(85cm) 이상이며, 체질량지수도 2번째 위험군(25.0 ~ 29.9 kg/m^2)에 해당하였다. 당뇨병의 가족력도 위험요인에 해당하지만, 교정이 불가능한 요인으로 간주하여 고려하지 않았다. 따라서, 허리둘레와 체질량지수가 교정되는 것에 따라 추가적으로 3가지의 시나리오를 구성할 수 있었다(Table 3.15).

3가지 시나리오에 대해서 기초 상태에서의 위험도 추정과 같은 방법으로 위험도를 추정하여 Table 3.16과 Figure 3.12의 결과를 얻을 수 있었다. 허리둘레만 교정되었을 때가 체질량지수만 교정되었을 때보다 총 당뇨병 유병기간의 기대값(13.29년 대 17.01년)이 작았으며, 평생 당뇨병 위험도도 낮을 것으로 예측되었다(61% 대 75%). 여러가지 시나리오 중에 가장 평생 당뇨병 위험도가 낮은 것은 체질량지수와 허리둘레 모두 교정된 경우이었다(47%). 또한, 당뇨병 유병기간의 기대값이 작을수록 평생 당뇨병 위험도도 낮아 한 개인(동일 성별, 동일 초기 연령)에서 다양한 위험요인을 비교할 때에는 당뇨병 유병기간과 평생 당뇨병 위험도가 양의 상관관계가 있음을 보여주는 한 예시가 될 수 있었다(Figure 3.13).

Table 3.15 Correcting scenarios of risk factors in P5.

	Base	BMI (only)	WC (only)	BMI and WC (Both)
Sex	Female	Female	Female	Female
Age	35	35	35	35
WC (cm)	96	96	83	83
BMI (kg/m^2)	26.4	23.0	26.4	23.0
Family history of diabetes mellitus	Yes	Yes	Yes	Yes
Hypertension	No	No	No	No
Current smoking	No	No	No	No

Abbreviation; WC, waist circumference; BMI, body mass index

Table 3.16 Expected durations of disease status and lifetime risk of diabetes mellitus according to correcting scenarios.

Corrected risk factor(s)	Duration of status (years)			Lifetime risk (%)
	Normal	DM without insight	DM with insight	
None (Base)	23.27	5.98	15.14	86%
BMI (only)	27.64	4.72	12.29	75%
WC (only)	31.59	3.64	9.65	61%
BMI and WC (Both)	35.31	2.66	7.14	47%

Abbreviation; WC, waist circumference; BMI, body mass index

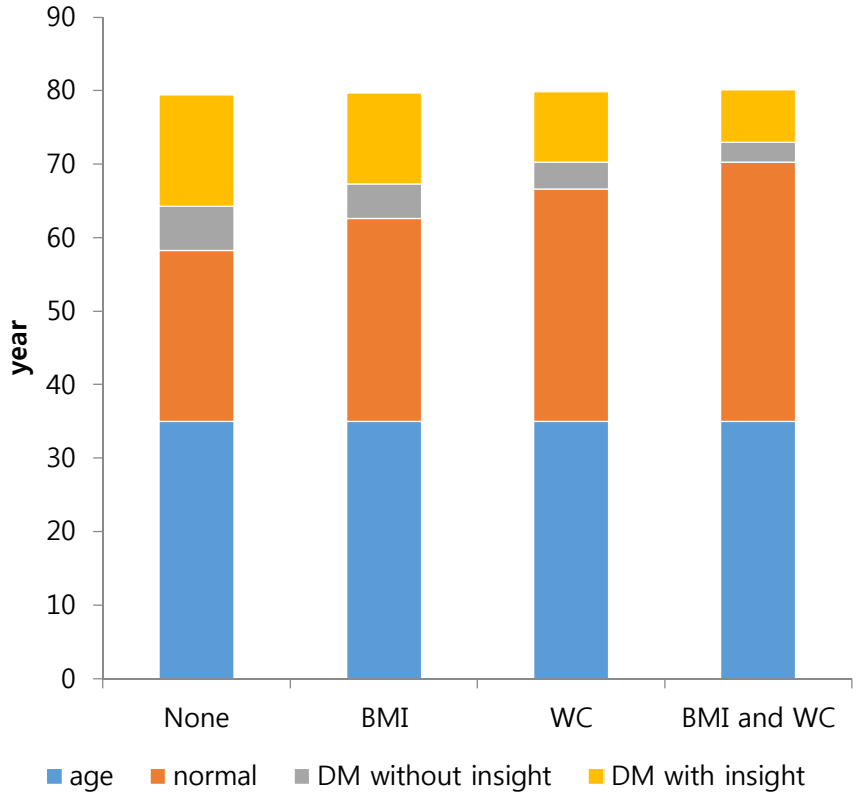

Figure 3.12 Expected duration of disease status according to correcting scenarios.

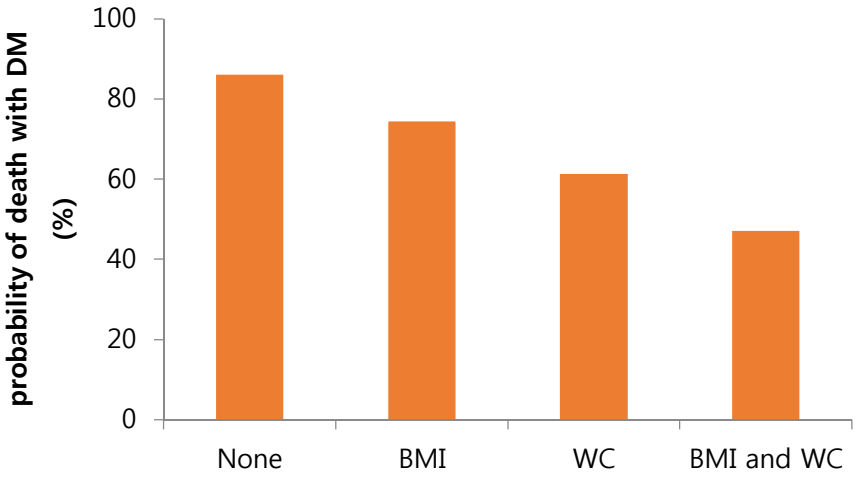

Figure 3.13 Lifetime risk of diabetes mellitus according to correcting scenarios.

3.4.4 개인별 위험요인의 타당성 검토 및 민감도 분석

“validation set”에서 초기 연령을 20세에 당뇨병이 없었다고 가정하고, 현재 연령에서의 예측 당뇨병 유병률과 실제 당뇨 여부를 비교하였을 때, ROC 곡선의 면적은 0.815로 기존의 예측 도구의 0.761보다 높았다(Figure 3.14). 예측 당뇨병 유병률의 절단값(cutoff value)을 변화시켰을 때, 8.0%에서 약 90%의 민감도를 나타내었으며, 20.0%에서 약 80%의 특이도를 보여주었다(Table 3.17). “validation set”의 표본을 예측 당뇨병 유병률을 10% 구간으로 분할한 뒤에, 평균 예측 당뇨병 유병률과 해당 구간의 실제 당뇨병 유병률을 비교한 결과 30% 미만의 구간에서는 비교적 일치하는 결과를 보여주었으나, 30% 이상의 구간에서는 그 차이가 크게 확대되었다(Table 3.18).

위험요인의 종류에 따른 13가지 모형의 민감도 분석에 의하면, 연령과 성별로만 추정한 “basal model”의 예측도가 가장 낮았으며, 여기에 흡연여부를 포함한 모형이 두 번째로 예측도가 낮았다. 이 두 모형을 제외하고 나머지 모형들은 모두 기존의 예측도구에 비해 높은 정확도를 보여주었다. “full model”은 가장 높은 예측력을 보여주었으며, “full model”에서 흡연여부만 제외한 모형이 두 번째로 높은 예측도를 보여주었다. 특히, “full model”과 여기에서 흡연여부, 고혈압 여부, 허리둘레, 체질량 지수를 각각 제외한 모형의 AUC 하한값은 기존 예측도구의 AUC 상한값보다 높았다. 하지만, 가족력을 제외하였을 때에는 예측력이 떨어져, 가족력이 예측에 있어 중요한 위험요인임을 보여 주었다.

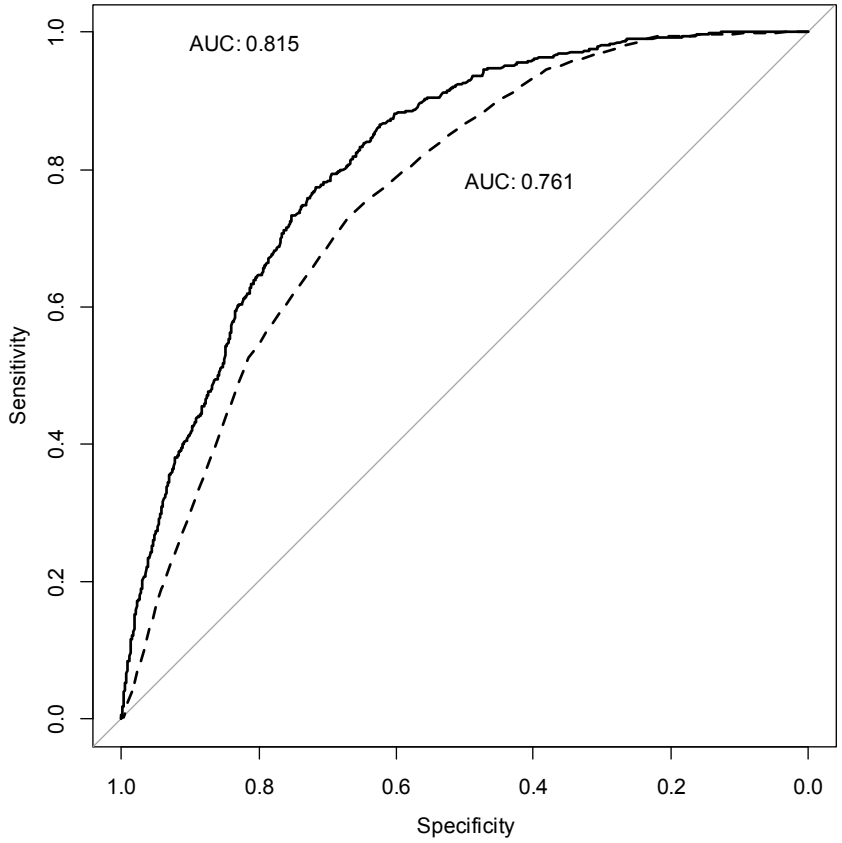

Figure 3.14 ROC curve of expected prevalence comparing with Korean diabetic prediction score (Estimation at current age from 20 years old).

Abbreviation; AUC: Area Under Curve

Table 3.17 Determinant performance of predicted prevalence.

Prevalence (%)	Count of TP	Count of FP	Count of FN	Count of TN	SN (%)	SP (%)	PC (%)
5.00	541	2,464	30	2,049	94.75	45.40	18.00
6.00	531	2,289	40	2,224	92.99	49.28	18.83
7.00	522	2,145	49	2,368	91.42	52.47	19.57
8.00	514	1,988	57	2,525	90.02	55.95	20.54
9.00	506	1,891	65	2,622	88.62	58.10	21.11
10.00	499	1,771	72	2,742	87.39	60.76	21.98
11.00	480	1,627	91	2,886	84.06	63.95	22.78
12.00	471	1,545	100	2,968	82.49	65.77	23.36
13.00	456	1,440	115	3,073	79.86	68.09	24.05
14.00	446	1,345	125	3,168	78.11	70.20	24.90
15.00	438	1,259	133	3,254	76.71	72.10	25.81
16.00	422	1,173	149	3,340	73.91	74.01	26.46
17.00	410	1,100	161	3,413	71.80	75.63	27.15
18.00	399	1,045	172	3,468	69.88	76.84	27.63
19.00	383	971	188	3,542	67.08	78.48	28.29
20.00	369	904	202	3,609	64.62	79.97	28.99
21.00	358	849	213	3,664	62.70	81.19	29.66
22.00	349	805	222	3,708	61.12	82.16	30.24

Abbreviation; TP: true positive, FP: false positive, FN: false negative, TN: true negative, SN: sensitivity, SP: specificity, PC: precision

Table 3.18 Predicted prevalence and actual prevalence of diabetes mellitus.

Categories of predicted prevalence (%)	Count	Mean of age (years)	Mean of predicted prevalence (%)	Actual prevalence (%)
0.00~9.99	2,814	39.62	3.30	2.56
10.00~19.99	997	59.53	14.44	13.04
20.00~29.99	541	62.73	24.69	22.55
30.00~39.99	284	63.17	34.50	24.65
40.00~49.99	221	66.33	45.20	33.48
50.00~100.00	227	65.91	65.95	45.37

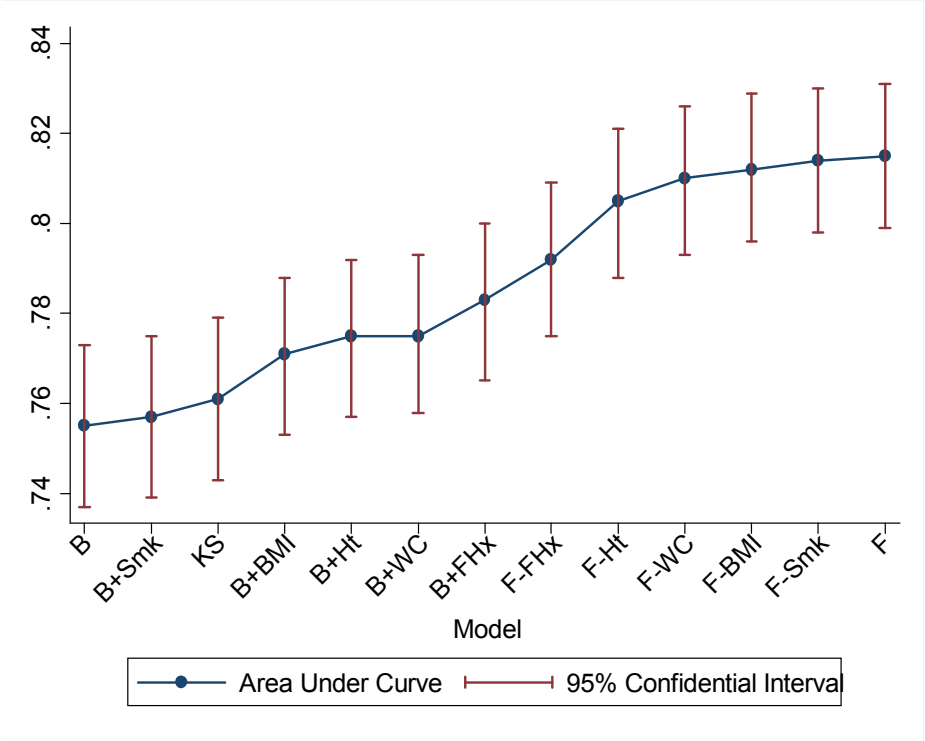

Figure 3.15 Area under curve and 95% confidential intervals of ROC of various models (Estimation at current age from 20 years old). Abbreviation; B: Basal model (age, sex); F: Full model (age, sex, hypertension, waist circumference, smoking, family history, body mass index); KS: Korean diabetic prediction score (age, sex, hypertension, waist circumference, smoking, family history, drinking); Smk: smoking; BMI: body mass index; Ht: hypertension; WC: waist circumference; FHx: family history; +: include a factor to basal model; -: exclude a factor from full model

제 4장. 고 찰 및 결 론

4.1 타당성 검토

통계청 보고서의 기대여명과 본 연구의 기대 여명을 비교하였을 때, 전체적으로 본 연구의 기대여명이 짧게 나타났다(Table 3.12). 그 원인은 본 연구의 상한 연령이 통계청보다 낮기 때문으로 생각된다(100세 대 85세). 또한, 초기 연령이 증가함에 따라 기대여명의 차이가 줄어드는 경향이 있다. 통계청의 기대여명은 당뇨병이 있는 대상자를 포함하고 있는 반면 본 연구의 기대여명은 초기 연령에서 당뇨병이 없다고 가정을 하였다. 초기 연령이 증가할수록 당뇨병 유병률이 높아지기 때문에 통계청의 기대여명은 연령이 증가함에 따라 낮게 추정된다. 이것이 통계청의 기대여명과 본 연구의 기대여명의 차이를 줄이는 방향으로 기여하게 된다. 70세 이상에서는 다시 기대여명의 차이가 일시적으로 증가하는데, 이것은 초기 연령과 상한 연령이 가까워 상한 연령에 의한 영향이 크게 작용한 것으로 해석된다. 이러한 여러 가지 가정과 추정 방법의 차이에도 불구하고, 기대여명은 차이는 비교적 일정하게 유지되었다. 따라서, 본 연구의 모형이 기존의 모형에 크게 벗어나지 않는다고 해석할 수 있다.

일반적으로 여러 가지 위험요인을 모형에 포함할수록, 그리고 “training set”과 “validation set”이 동질성이 있을수록 예측력이 높게 나타날 수 있다. 타당성 검토에서 사용한 한국인 당뇨병 예측도구는 본 연구와 동일하게 국민건강영양조사를 바탕으로 개발되었으며, 위험요인의 대부분이 일치하므로 매우 적합한 비교대상으로 생각할 수 있다. 따라서, 본 연구의 예측 당뇨병 유병률이 기존의 예측 도구에 비해 높은 예측력을 보여주었다는 점은 본 연구의 타당성을

지지해 주는 근거가 된다(Figure 3.13). 본 연구에서 비교 대상이 된 예측 도구는 기존의 다른 예측도구들과 비교하여 우수한 예측력을 보여주었기 때문에(Lee et al., 2012), 그 의미가 더욱 크다고 할 수 있다. 또한 30% 미만의 구간에서 예측 유병률과 실제 유병률이 잘 일치한다는 점도 본 연구의 타당성을 지지하는 요소이다(Table 3.18). 다만, 예측 유병률이 30% 이상의 구간에서는 실제 유병률에 비해서 과대 평가되는 경향이 있었다. 본 연구에서 예측 유병률은 위험인자들이 20세 때부터 발생하여 유지되었다고 가정하였으나, 실제로는 그 이후에 발생하였을 가능성이 높다. 특히, 고위험군의 경우 평가 시점의 평균 연령이 60세 이상인 것으로 미루어 볼 때, 위험인자의 노출기간이 과대평가되었을 것을 고려해야 하며, 특히 위험인자의 개수가 많을수록 노출기간이 과대평가되었을 가능성도 높아진다. 위험 요인이 발생한 시점을 알 수 있는 자료를 통해 비교를 하면, 예측값과 실제 유병률의 차이가 줄어들 것으로 기대된다.

민감도 분석에서는 위험요인을 일부 제외하여도 기존의 예측 도구에 비해 높은 예측도를 보여주어, 본 연구의 예측모형이 우수함을 알 수 있었다. 특히, 본 모형에서 예측력에 있어서 흡연여부는 가장 기여하는 바가 낮으며, 가족력이 기여하는 바가 큼을 민감도 분석을 통해 알 수 있었다. 실제 의료환경에서 활용도를 높이기 위해서는 위험요인의 개수를 줄이는 것이 도움이 되는데, 이 민감도 분석의 결과를 참고로 하여 위험요인 선택의 우선 순위를 정할 수 있을 것으로 기대된다.

당뇨병 유병 기간의 기대값은 확률적인 계산이기 때문에, 특정 집단을 특성을 나타내는 데에는 적합하지만, 개인별로 적용하는 것은 적합하지 않다. 예를 들어 Table 3.13의 P7의 대상자가 8.87년 후에 당뇨병이 발생해서, 0.22년 후에 당뇨병을 진단 받고, 0.54년

후에 사망한다는 해석은 질병에 대한 잘못된 위험 인식을 심어줄 수 있다. P7은 당뇨병이 있는 채로 사망할 확률이 16.24%이기 때문에, 사망하기 전까지 당뇨병이 없을 확률은 83.76%가 되며, 사망하기 전에 당뇨병이 발생한다면 0.76년(0.22년+0.54년)보다는 긴 기간이 될 것이다. 즉, 본 연구에서의 당뇨병 유병기간 기대값은 P7과 같은 특성을 가진 대상자가 매우 많을 때에 당뇨병이 없는 사람을 포함한 평균적인 당뇨병 유병 기간이다.

인종 및 코호트에 따라 당뇨병의 발병률이 다를 수 있으나, 일반적으로 당뇨병의 발병률은 연령에 따라 증가하며 60대에 가장 높고, 이후에는 감소하는 것으로 알려져 있다(Garancini et al., 1996; Chang et al., 2010). 본 연구에서는 발병률을 계산할 수 있는 직접적인 자료를 획득하기 어렵기 때문에, Figure 3.4의 발병률은 단면적 연구를 통해 수집된 자료에서 인접한 연령대가 같은 코호트라고 가정하여 계산한 값이다. 이와 같은 계산을 통해 코호트 효과를 구분해 낼 수가 없기 때문에, 실제 유병률과 다른 양상으로 발병률이 추정된 것으로 생각된다. 따라서, 본 연구에서 제안한 발병률 계산은 수학적 과정으로만 이해하여야 하며, 역학적 의미로 해석하는 것에는 제한이 따른다. 향후 한국인에서의 연령대별 당뇨병 발병률 연구가 이루어 진다면, 이 부분에 대한 보완이 가능할 것이라고 생각한다.

인지여부에 따른 기간의 기대값의 경우에 개인별 해석은 더욱 의미가 없다. 본 연구의 가정에서 당뇨병의 발생은 비가역적인 것으로 모델링 하였으므로, 정상-당뇨병-사망은 순차적으로 일어난다고 볼 수 있으나, 당뇨병에 대한 인지 여부(observation status)는 당뇨병이 있는 상태(hidden status)에서 확률적으로 일어난다. 따라서, 비인지-인지 상태가 순차적으로 나타나지 않고, 여러 번 교차해서 나타날 수도 있다. 한 개인에 대해서는 실제로 이런

상태가 당뇨병의 초기에 나타날 수 있으나, 특정 기간이 지나면 인지 상태로 지속이 될 것이다. 따라서, 한 개인에 대해서 특정 시점의 관측 상태가 실제 유병 여부 및 인지 여부와 일치할 것을 기대하기 어렵다. 그럼에도 불구하고, 기대값은 다른 목적의 활용에서 유용할 수 있다. 예를 들어, 특정 집단의 의료 비용을 예측하거나, 최적화 된 보건 정책 전략을 수립하는 데에 기초 자료로 활용될 수 있다.

4.2 연구의 특징 및 활용방안

질병 위험 평가는 크게 두 가지 형태의 목적을 위해 개발되어 사용된다. 첫 번째로는, Framingham Risk Score(Wilson et al., 1998)나 골다공증 위험 예측도구(Kanis et al., 2008)와 같이 치료방침의 결정에 필요한 정보를 얻고자 하는 목적으로 사용하는 형태이며, 두 번째로는 국민건강보험공단에서 제공하는 건강나이와 같이 일반인들의 이해를 돕고 생활습관 개선을 유도하기 위한 목적으로 사용되는 형태이다(Wagner et al., 1982; 조비룡, 2012). 질병 위험 평가를 개발하는 단계에서 고려해야 할 몇 가지 요소들이 이러한 목적에 따라 중요도가 달라질 수 있다.

우선 위험요인의 선택에 있어서 의학적 결정을 주요 목적으로 하는 경우에는 정확도가 중요시 될 수 있으며, 교정 가능성의 중요도는 작을 것이다. 따라서 대체로 과거력이나 가족력과 같은 위험요인이 포함 되며 최근에는 다수의 유전자형을 포함하고자 하는 시도도 이루어지고 있다(Noble et al., 2011). 이에 반하여, 위험요인의 교정을 주요 목적으로 하는 경우에는 교정이 불가능한 위험요인을 다수 포함하는 것이 큰 의미가 없다. 실제로 질병 위험 평가는 생활습관을 개선하는 것과 같은 긍정적인 효과를 기대할 수도 있으나(Smerecnik et al., 2012), 오히려 과도한 불안과 함께 불필요한 의료행위를 받게 되는 부작용을 낳을 수도 있다(Kye et al., 2012).

두 번째로, 결과 지표의 선택에 있어서 의학적 결정을 주요 목적으로 하는 경우에는 보다 객관적인 형태로 표현하는 것이 좋을 것이다. 하지만, 위험인자 교정을 주 목적으로 하는 경우에는 일반인도 이해하기 쉽고 흥미를 유발할 수 있는 형태의 지표를 선택하는 것이 도움이 될 수 있다. 실제로 건강나이(조비룡, 2012)는

이러한 효과를 극대화 하기 위해 만들어진 추상적인 개념이라고 할 수 있다.

마지막으로 모형의 선택에 있어서, 의학적 결정을 주요 목적으로 하는 경우에는 타당성이 가장 최우선시 되기 때문에 다소 복잡한 수학적 모형이 선택이 되는 경우가 있다. 하지만, “neural network”나 “support vector machine”과 같이 선형적이지 않은 판별 분석을 이용한다면 위험요인의 변화가 질병 위험도에 어떻게 영향을 미치는 지 직관적으로 알기 어렵다는 제한점이 있다. 이에 반해, 위험인자 교정을 주 목적으로 하는 경우에는 위험요인의 변화에 따라 질병 위험도가 쉽게 예측 가능하도록 선형적인 모형을 조합하여 사용하는 것이 좋을 것이다. 경우에 따라 윤리적인 문제에 의해서 근거의 질이 충분히 보장된 연구가 시행되기 어려울 수 있다. 타당성을 강조한다면 이 경우에는 질병 위험도 예측이 의미가 없을 수 있으나, 동기 부여를 강조한다면 차선책으로 가장 근거가 있는 자료를 이용하여 질병 위험도 예측을 하는 것에도 의미가 있겠다. 또한, 동기 부여를 강조한 질병 위험 평가는 경우에 따라 질병 위험이 과도하게 인식되어 정신사회학적인 부작용을 낳는 것을 줄이기 위하여, 타당도를 일부 희생하면서 의도적으로 결과의 편차를 줄이는 방향으로 설계되기도 한다(조비룡, 2012; Park et al., 2013).

본 연구에서 활용된 위험요인은 특별한 검사 없이 신체검진과 설문을 통해서 쉽게 얻을 수 있으며, 대부분 교정 가능한 정보(가족력 제외)로 구성하였다. 이와 같은 구성은 진료실에서의 활용도도 높일 수 있으며, 생활습관 교정의 동기 부여 목적으로도 활용이 가능하다 사료된다. 여기에 다른 연구를 통해 밝혀진 위험요인을 추가 또는 변경한다면 향후 모형의 타당도를 더 높일 수도 있을 것이다. 이 경우 전이행렬을 조정하는 과정이 모듈화

되어 있기 때문에 전체적인 수정 없이 모형을 업데이트 하는 것이 가능하다. 예를 들어 유전자의 “allele type”에 따라 당뇨병의 평생 위험도를 구하고자 한다면, 해당 “allele type”에 따른 표준화된 OR의 정보만 추가함으로써 유전자형이 포함된 예측 모형으로 변형이 가능하다(Noble et al., 2011). 또한 기존의 질병 위험평가는 일부 위험요인만 표준화 함으로써 위험요인의 개수가 늘어남에 따라 과대평가가 되는 예가 있었다(Park et al., 2013). 하지만 본 연구에서는 모든 위험요인에 대해 표준화된 OR을 이용하기 때문에 위험요인의 개수가 늘어나도 위험도가 과대평가 될 가능성이 적은 것이 장점이라고 하겠다. 이와 같이 전체적인 수정 없이 위험요인을 자유롭게 선택할 수 있는 것이 이 연구의 장점이라고 하겠다.

본 연구가 기존의 모형에 비해 다른 점 중에 하나가 당뇨병 유병기간의 기대값을 인지하고 있는 기간과 인지하고 있지 못하는 기간으로 나누어 볼 수 있다는 점이다. 본 모형에서는 당뇨병의 인지 여부는 당뇨병이 있다는 전제하에 발생하는 사건이므로, 당뇨병의 유병기간이 길수록 당뇨병을 인지하지 못하는 기간이 길어지게 된다. 또한, 젊은 연령에서 당뇨병의 인지율이 낮기 때문에, 당뇨병이 젊은 연령에서 발생하였을 수록 당뇨병을 인지하지 못하는 기간이 길어지게 된다. 이것은 고위험군에서는 젊은 연령에서부터 당뇨병의 선별검사가 필요함을 지지하는 결과이다(American Diabetes, 2013). 하지만, 본 연구에서 당뇨병의 인지여부에 대한 타당성 검토는 하지 못하였기 때문에, 임상적 활용은 제한적으로만 고려되어야 할 것이다.

본 연구는 당뇨병의 위험도를 예측하는 데에 있어서 다소 복잡한 수학적 모형으로 설계되었지만, 비교적 쉬운 개념인 평생 당뇨병 유병률의 형태로 결과를 표현함으로써 일반인들의 이해를 도울

것으로 기대된다. 또한, 위험요인 교정에 따른 당뇨병 위험도를 쉽게 알 수 있어 위험도를 낮추기 위한 동기 부여에도 도움이 될 것으로 기대할 수 있다. 이와 더불어, 본 연구의 모형은 일반적으로 쉽게 얻을 수 있는 정보만을 위험요인으로 포함하면서, 그 예측력이 기존의 예측 도구보다 높아 의학적 결정 도구로도 충분히 활용 가능하다. 예를 들면, 개인의 위험요인에 따라 선별검사 시행 여부를 결정하는 데에 유용한 정보를 제공할 수 있다. 타당도를 희생하지 않고 쉽게 이해할 수 있는 형태로 결과를 표현하는 점, 그리고 비선형적 모형이지만 위험요인 교정 후의 예측값을 알 수 있다는 점이 이 연구의 또 다른 장점이 될 수 있다.

4.3 연구의 제한점

모형의 관점에서 볼 때, 본 연구의 가정에 따르면 T시점의 당뇨병 인지여부는 T시점의 유병 여부에만 의존하여 확률적으로 발생하지만, 실제로는 유병 기간과 이전 시점까지의 인지여부도 영향을 받는다. 하지만, 본 연구의 모형은 그와 같은 요소를 반영할 수 없는 한계점을 갖고 있다. 이와 같은 한계점을 보완하기 위하여 HMM이 아닌 마코프모델로 구성하고, 상태를 정상, 비인지 당뇨병 상태, 인지된 당뇨병 상태, 사망의 4개로 가정하는 방법이 있다. 하지만, 이 경우 4×4 형태의 전이행렬이 필요하며, 비인지 상태에서 인지 상태로 갈 확률, 비인지 상태에서의 사망률, 인지 상태에서의 사망률과 같은 추가적인 모수들이 필요하지만 이것을 계산하기 위한 기초 자료의 확보는 현실적으로 기대하기 어렵다.

본 연구에 이용된 자료의 관점에서 볼 때, 당뇨병이 있는 대상자의 사망률과 당뇨병이 없는 대상자의 사망률을 전 연령대에 걸쳐 같은 위해비(hazard ratio)를 이용해 계산하였다. 기존 문헌에 따르면, 연령에 따라 당뇨병에 의한 위해비가 차이가 있을 것으로 기대가 되지만(Woodward et al., 2003), 그 연령대가 세분화 되어 있지 못해 연구에 활용할 수가 없었다. 보다 세분화된 연령대에 따른 위해비가 보고 된다면 본 연구의 모형을 더욱 구체화 할 수 있을 것으로 기대된다. 또한 본 연구에 활용된 국민건강영양조사 자료는 단면적 관찰 자료로서, 여러 가지 가정을 통해 시간적 재구성을 하여 활용하였다. 특히, 발병률의 경우는 직접적으로 산출할 수가 없었기에 인접한 연령의 유병률을 바탕으로 계산하였다. 하지만 시대(period) 효과와 코호트 효과가 있는 경우 계산된 발병률은 실제 발병률과 큰 차이가 있을 수 있다. 대표성이 있는 추적관찰 자료를 이용한다면 이와 같은 문제점의 해결이 가능할 것이다.

4.4 결론 및 발전방향

본 연구는 국민건강영양조사와 통계청 자료를 바탕으로 하여 평생 당뇨병 위험도를 추정하는 것이다. 기존의 연구에서는 당뇨병 진단 여부를 기초 자료로 활용하여 모형화 하였고, 그 위험요인에 대해 세분화 하지 못한 제한점이 있었다. 본 연구에서는 혈당 검사와 같은 생체 지표를 당뇨병 진단에 포함시켜, 생물학적 상태를 반영한 모형을 구조화 한 것이 그 차별점이라고 하겠다. 이와 더불어, 당뇨병 인지 여부를 같이 표현하기 위하여 마코프모델과 HMM을 함께 사용하였다. 또한, 개인별 위험요인을 반영하는 요소를 추가하여, 그 활용성과 확장성을 높이고자 하였다.

이 연구의 결과로 최초로 한국인의 평생 당뇨병 위험도를 추정하였으며, 개인별 위험요인에 따른 위험도의 추정을 가능케 하였다. 본 연구는 평생 당뇨병 위험도를 산출하여 일차적으로는 일반인에게 정보를 제공하고자 하는 데에 그 목적이 있었으나, 기존의 타당도가 검증된 예측도구와 비교하여도 예측력이 높음을 확인할 수 있어, 의학적 결정에도 크게 활용될 수 있을 것으로 기대된다. 하지만 당뇨병 인지 여부에 대해서는 자료의 제한과 모형의 한계로 인하여 타당성을 확신할 수가 없었다는 것이 아쉬운 점이다.

고혈압, 골다공증, 암과 같은 만성질환들은 당뇨병과 같이 비인지 기간이 있을 것으로 추정되나, 그에 대한 연구는 거의 찾을 수 없다. 이와 같은 만성질환들은 당뇨병과 질병의 경과가 비슷하므로, 향후 적절한 기초 자료만 있다면 비슷한 형태로 모형을 개발 하는 것도 기대해 볼 수 있겠다.

참 고 문 헌 (Bibliography)

American Diabetes, A. Diagnosis and classification of diabetes mellitus. *Diabetes care* **36 Suppl 1**, S67-74, doi:10.2337/dc13-S067 (2013).

American Diabetes, A. Standards of medical care in diabetes--2013. *Diabetes care* **36 Suppl 1**, S11-66, doi:10.2337/dc13-S011 (2013).

Bonora, E. & Tuomilehto, J. The pros and cons of diagnosing diabetes with A1C. *Diabetes care* **34 Suppl 2**, S184-190, doi:10.2337/dc11-s216 (2011).

CDC. National Diabetes Fact Sheet. (2011).

Chang, C. H. et al. Type 2 diabetes prevalence and incidence among adults in Taiwan during 1999-2004: a national health insurance data set study. Diabetic medicine : *a journal of the British Diabetic Association* **27**, 636-643, doi:10.1111/j.1464-5491.2010.03007.x (2010).

Chi, C. L., Street, W. N. & Katz, D. A. A decision support system for cost-effective diagnosis. *Artificial intelligence in medicine* **50**, 149-161, doi:10.1016/j.artmed.2010.08.001 (2010).

Concannon, P., Rich, S. S. & Nepom, G. T. Genetics of type 1A diabetes. *The New England journal of medicine* **360**, 1646-1654, doi:10.1056/NEJMra0808284 (2009).

Danaei, G., Finucane, M. M., Lu, Y., Singh, G. M., Cowan, M. J., Paciorek, C. J., Lin, J. K., Farzadfar, F., Khang, Y. H., Stevens, G. A., Rao, M., Ali, M. K., Riley, L. M., Robinson, C. A., Ezzati, M. & Global Burden of Metabolic Risk Factors of Chronic Diseases Collaborating, G. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. *Lancet* **378**, 31-40, doi:10.1016/S0140-6736(11)60679-X (2011).

Danaei, G., Lawes, C. M., Vander Hoorn, S., Murray, C. J. & Ezzati, M. Global and regional mortality from ischaemic heart disease and stroke attributable to higher-than-optimum blood glucose concentration: comparative risk assessment. *Lancet* **368**, 1651-1659, doi:10.1016/S0140-6736(06)69700-6 (2006).

DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845 (1988).

Diabetes Prevention Program Research, G., Knowler, W. C., Fowler, S. E., Hamman, R. F., Christophi, C. A., Hoffman, H. J., Brenneman, A. T., Brown-Friday, J. O., Goldberg, R., Venditti, E. & Nathan, D. M. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *Lancet* **374**, 1677-1686, doi:10.1016/S0140-6736(09)61457-4 (2009).

EBMeDS. *The Evidence-Based Medicine electronic Decision Support*, <http://www.ebmeds.org> (2011).

Fraser, G. E. & Shavlik, D. J. The estimation of lifetime risk and average age at onset of a disease using a multivariate exponential hazard rate model. *Statistics in medicine* **18**, 397-410 (1999).

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C. & Mulvihill, J. J. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* **81**, 1879-1886 (1989).

Garancini, M. P., Gobbi, C., Errera, A., Sergi, A. & Gallus, G. Age-specific incidence and duration of known diabetes. The Cremona Study. *Diabetes care* **19**, 1279-1282 (1996).

GLIDES. *Guidelines into decision support*, <http://gem.med.yale.edu/glides/> (2010).

Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., Styrkarsdottir, U., Magnusson, K. P., Walters, G. B., Palsdottir, E., Jonsdottir, T., Gudmundsdottir, T., Gylfason, A., Saemundsdottir, J., Wilensky, R. L., Reilly, M. P., Rader, D. J., Bagger, Y., Christiansen, C., Gudnason, V., Sigurdsson, G., Thorsteinsdottir, U., Gulcher, J. R., Kong, A. & Stefansson, K. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature genetics* **38**, 320-323, doi:10.1038/ng1732 (2006).

Hopkins, R. B., Pullenayegum, E., Goeree, R., Adachi, J. D., Papaioannou, A., Leslie, W. D., Tarride, J. E. & Thabane, L. Estimation of the lifetime risk of hip fracture for women and men in Canada. *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA* **23**, 921-927, doi:10.1007/s00198-011-1652-8 (2012).

Huang, E. S., Basu, A., O'Grady, M. & Capretta, J. C. Projecting the future diabetes population size and related costs for the U.S. *Diabetes care* **32**, 2225-2229, doi:10.2337/dc09-0459 (2009).

Janzon, L., Berntorp, K., Hanson, M., Lindell, S. E. & Trell, E. Glucose tolerance and smoking: a population study of oral and intravenous glucose tolerance tests in middle-aged men. *Diabetologia* **25**, 86-88 (1983).

Kanis, J. A., Johnell, O., Oden, A., Johansson, H. & McCloskey, E. FRAX and the assessment of fracture probability in men and women from the UK. *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA* **19**, 385-397, doi:10.1007/s00198-007-0543-5 (2008).

Kawamoto, R., Nazir, A., Kameyama, A., Ichinomiya, T., Yamamoto, K., Tamura, S., Yamamoto, M., Hayamizu, S. & Kinosada, Y. Hidden markov model for analyzing time-series health checkup data. *Studies in health technology and informatics* **192**, 491-495 (2013).

Khader, K., Leecaster, M., Greene, T., Samore, M. & Thomas, A. Improved hidden Markov model for nosocomial infections. *Mathematical medicine and biology : a journal of the IMA*, doi:10.1093/imammb/dqt013 (2013).

Klein, R. Hyperglycemia and microvascular and macrovascular disease in diabetes. *Diabetes care* **18**, 258-268 (1995).

Korc, M. Diabetes mellitus in the era of proteomics. *Molecular & cellular proteomics : MCP* **2**, 399-404, doi:10.1074/mcp.R300005-MCP200 (2003).

Kye, S. Y., Park, K., Park, H. G. & Kim, M. H. Psychological impact of health risk appraisal of Korean women at different levels of breast cancer risk: neglected aspect of the web-based cancer risk assessment tool. *Asian Pacific journal of cancer prevention : APJCP* **13**, 437-441 (2012).

Lee, S. Y., Park, H. S., Kim, D. J., Han, J. H., Kim, S. M., Cho, G. J., Kim, D. Y., Kwon, H. S., Kim, S. R., Lee, C. B., Oh, S. J., Park, C. Y. & Yoo, H. J. Appropriate waist circumference cutoff points for central obesity in Korean adults. *Diabetes research and clinical practice* **75**, 72-80, doi:10.1016/j.diabres.2006.04.013 (2007).

Lee, Y. H., Bang, H., Kim, H. C., Kim, H. M., Park, S. W. & Kim, D. J. A simple screening score for diabetes for the Korean population: development, validation, and comparison with other scores. *Diabetes care* **35**, 1723-1730, doi:10.2337/dc11-2347 (2012).

Lee, Y. H., Kang, E. S., Kim, S. H., Han, S. J., Kim, C. H., Kim, H. J., Ahn, C. W., Cha, B. S., Nam, M., Nam, C. M. & Lee, H. C. Association between polymorphisms in SLC30A8, HHEX, CDKN2A/B, IGF2BP2, FTO, WFS1, CDKAL1, KCNQ1 and type 2 diabetes in the Korean population. *Journal of*

human genetics **53**, 991-998, doi:10.1007/s10038-008-0341-8 (2008).

Liu, Z. P., Wang, Y., Zhang, X. S. & Chen, L. Network-based analysis of complex diseases. *IET systems biology* **6**, 22-33, doi:10.1049/iet-syb.2010.0052 (2012).

Mortaz, S., Wessman, C., Duncan, R., Gray, R. & Badawi, A. Impact of screening and early detection of impaired fasting glucose tolerance and type 2 diabetes in Canada: a Markov model simulation. *ClinicoEconomics and outcomes research : CEOR* **4**, 91-97, doi:10.2147/CEOR.S30547 (2012).

Narayan, K. M., Boyle, J. P., Thompson, T. J., Gregg, E. W. & Williamson, D. F. Effect of BMI on lifetime risk for diabetes in the U.S. *Diabetes care* **30**, 1562-1566, doi:10.2337/dc06-2544 (2007).

Narayan, K. M., Boyle, J. P., Thompson, T. J., Sorensen, S. W. & Williamson, D. F. Lifetime risk for diabetes mellitus in the United States. *JAMA : the journal of the American Medical Association* **290**, 1884-1890, doi:10.1001/jama.290.14.1884 (2003).

Neumann, A., Schwarz, P. & Lindholm, L. Estimating the cost-effectiveness of lifestyle intervention programmes to prevent diabetes based on an example from Germany: Markov modelling. *Cost effectiveness and resource allocation : C/E* **9**, 17, doi:10.1186/1478-7547-9-17 (2011).

Noble, D., Mathur, R., Dent, T., Meads, C. & Greenhalgh, T. Risk models and scores for type 2 diabetes: systematic review. *Bmj* **343**, d7163, doi:10.1136/bmj.d7163 (2011).

Park, J. H., Kwon, H., Oh, S. W., Lee, C. M. & Cho, B. Update and validation of a national health risk appraisal tool in Korea. *Journal of public health* **35**, 107-114, doi:10.1093/pubmed/fds061 (2013).

RACGP. *Guidelines for preventive activities in general practice*. 8th edn, (The Royal Australian College of General Practitioners, 2012).

Shankaracharya, Odedra, D., Samanta, S. & Vidyarthi, A. S. Computational intelligence in early diabetes diagnosis: a review. *The review of diabetic studies : RDS* **7**, 252-262, doi:10.1900/RDS.2010.7.252 (2010).

Silventoinen, K., Pankow, J., Lindstrom, J., Jousilahti, P., Hu, G. & Tuomilehto, J. The validity of the Finnish Diabetes Risk Score for the prediction of the incidence of coronary heart disease and stroke, and total mortality. *European journal of cardiovascular prevention and rehabilitation : official journal of the European Society of Cardiology, Working Groups on Epidemiology & Prevention and Cardiac Rehabilitation and Exercise*

Physiology **12**, 451-458 (2005).

Smerecnik, C., Grispen, J. E. & Quaak, M. Effectiveness of testing for genetic susceptibility to smoking-related diseases on smoking cessation outcomes: a systematic review and meta-analysis. *Tobacco control* **21**, 347-354, doi:10.1136/tc.2011.042739 (2012).

Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S. F., Payne, F., Lowe, C. E., Szeszko, J. S., Hafler, J. P., Zeitels, L., Yang, J. H., Vella, A., Nutland, S., Stevens, H. E., Schuilenburg, H., Coleman, G., Maisuria, M., Meadows, W., Smink, L. J., Healy, B., Burren, O. S., Lam, A. A., Ovington, N. R., Allen, J., Adlem, E., Leung, H. T., Wallace, C., Howson, J. M., Guja, C., Ionescu-Tirgoviste, C., Genetics of Type 1 Diabetes in, F., Simmonds, M. J., Heward, J. M., Gough, S. C., Wellcome Trust Case Control, C., Dunger, D. B., Wicker, L. S. & Clayton, D. G. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature genetics* **39**, 857-864, doi:10.1038/ng2068 (2007).

USPSTF. *Screening for Type 2 Diabetes Mellitus in Adults*, <http://www.uspreventiveservicestaskforce.org/uspstf/uspsdiab.htm> (2012).

Wagner, E. H., Beery, W. L., Schoenbach, V. J. & Graham, R. M. An assessment of health hazard/health risk appraisal. *American journal of public health* **72**, 347-352 (1982).

Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:10.1038/nature05911 (2007).

Whiting, D. R., Guariguata, L., Weil, C. & Shaw, J. IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes research and clinical practice* **94**, 311-321, doi:10.1016/j.diabres.2011.10.029 (2011).

WHO. The global burden of disease: 2004 update. *World Health Organization* (2008).

Wikepedia. *Hill climbing*, <http://en.wikipedia.org/wiki/Hill_climbing> (2013).

Wikipedia. *Markov chain*, <http://en.wikipedia.org/wiki/Markov_chain> (2013).

Wilkins, J. T., Ning, H., Berry, J., Zhao, L., Dyer, A. R. & Lloyd-Jones, D. M. Lifetime risk and years lived free of total cardiovascular disease. *JAMA : the journal of the American Medical Association* **308**, 1795-1801,

doi:10.1001/jama.2012.14312 (2012).

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H. & Kannel, W. B. Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837-1847 (1998).

Woodward, M., Zhang, X., Barzi, F., Pan, W., Ueshima, H., Rodgers, A., MacMahon, S. & Asia Pacific Cohort Studies, C. The effects of diabetes on the risks of major cardiovascular diseases and death in the Asia-Pacific region. *Diabetes care* **26**, 360-366 (2003).

국가통계포털. *생명표*, <http://kosis.kr/>

대한당뇨병학회. 당뇨병 진료지침 2011. (2011).

보건복지부 & 질병관리본부. 2011 국민건강통계. (2012).

보건복지부 & 질병관리본부. *국민건강영양조사 5기*, <http://knhanes.cdc.go.kr/knhanes/index.do> (2012).

윤재문 & 손현석. 당뇨병 연구의 전산 기법적 접근. *보건학 논집* **49**, 88-98 (2012).
Jae-Moon Yoon and Hyeon S. Son Computational Approach for Understanding Diabetes Mellitus. The Korean Journal of Public Health 49(2), 88-98 (2012

조비룡. 건강나이 알아보기(HRA) 업데이트 및 개선. *국민건강보험공단* (2012).

통계청. 2011 death and cause of death in Korea. *Korea National Statistical Office* (2012).

한학용. *패턴인식 개론*. 개정판, 429-479 (한빛미디어(주), 2009).

부 록: 본 연구에 사용된 코드

```
# R code of functions for estimation
# make_tmat: transitional matrix를 만드는 함수
# fp: dataframe, death(전체사망률), pr(당뇨병유병률), sn(당뇨병인지율)
# hr: 당뇨병이 있는 대상자의 사망 위험비
make_tmat <- function(fp, hr){
        max=nrow(fp)
        m_tmat<-array(data=c(1,0,0,0,1,0,0,0,1),dim=c(3,3,max))
        for(i in 1:(max-1)){
                m_tmat[1,1,i]<- 1-(fp$pr[i+1]-fp$pr[i]+fp$death[i]*
                        (hr*fp$pr[i]/(1-fp$pr[i]+fp$pr[i]*hr)-fp$pr[i+1]))/(1-
fp$pr[i])-
                        fp$death[i]/(1-fp$pr[i]+fp$pr[i]*hr)
                m_tmat[2,1,i]<- (fp$pr[i+1]-fp$pr[i]+fp$death[i]*
                        (hr*fp$pr[i]/(1-fp$pr[i]+fp$pr[i]*hr)-fp$pr[i+1]))/(1-
fp$pr[i])
                m_tmat[3,1,i]<- fp$death[i]/(1-fp$pr[i]+fp$pr[i]*hr)
                m_tmat[2,2,i]<- 1-fp$death[i]*hr/(1-fp$pr[i]+fp$pr[i]*hr)
                m_tmat[3,2,i]<- fp$death[i]*hr/(1-fp$pr[i]+fp$pr[i]*hr)
        }
        m_tmat[1,1,max]<- 0
        m_tmat[2,1,max]<- 0
        m_tmat[3,1,max]<- 1
        m_tmat[2,2,max]<- 0
        m_tmat[3,2,max]<- 1
        m_tmat
}

# make_smat: observational matrix를 만드는 함수
# fp: dataframe, death(전체사망률), pr(당뇨병유병률), sn(당뇨병인지율)
make_smat <- function(fp){
        max=nrow(fp)
```

```
        m_smat<-array(data=c(1,0,0,0,1,0,0,0,1),dim=c(3,3,max))
        for(i in 1:max){
                m_smat[1,2,i]<- 1-fp$sn[i]
                m_smat[2,2,i]<- fp$sn[i]
        }
        m_smat
}

# mat_cmult: 행렬의 누적곱 함수
# arr: 행렬들의 집합으로 이루어진 3차원 배열
# start: 곱을 시작할 index
# end: 곱을 종료할 index
mat_cmult <- function(arr,start,end){
        if (start>end) return(diag(3)) else
                return(mat_cmult(arr, start+1, end)%*%arr[,,start])
}

# hmm_est: HMM에 의해 당뇨병 위험도 추정함수
# arr: trainsitional matrix들의 3차원 배열
# arr2: observational matrix들의 3차원 배열
# age: 현재 (추정시작) 연령
hmm_est <- function(arr,arr2,age){
        normal=1
        diabetes=0
        know_DM=0
        d_cDM=0

        for(i in (age-19):64){
                temp<-mat_cmult(arr,(age-19),i)
                normal=normal+temp[1,1]
                diabetes=diabetes+temp[2,1]
                know_DM=know_DM+(arr2[,,i]%*%temp)[2,1]
                d_cDM=d_cDM+temp[2,1]*arr[3,2,i+1] #death with DM
        }
```

```
        return(c(normal, diabetes-know_DM, know_DM, d_cDM, normal+diabetes,
diabetes))
}

# risk_tmat: 위험도 교정 함수
# arr: 기본 trainsitional matrix들의 3차원 배열
# or: 표준 인구집단과 비교하였을 때 당뇨병 발생 OR
risk_tmat <- function(arr, or){
        for(i in 1:64){
                temp<-arr[,,i]
                temp[1,1]<-(arr[1,1,i]+arr[2,1,i])/(1+or*arr[2,1,i]/arr[1,1,i])
                temp[2,1]<-(or*arr[2,1,i]/arr[1,1,i])*temp[1,1]
                arr[,,i]<-temp
        }
        arr
}

# mix_or: 복합위험도를 계산하는 함수
# rf: 개인의 risk factor를 저장하고 있는 dataframe
# rft: risk factor에 따른 표준화 log(OR)들을 저장하고 있는 dataframe
mix_or <- function(rf, rft){
        i <- if(rf$age<80) as.integer(rf$age/10)-1 else 6
        sex <- if(rf$sex==1) "male" else "female"
        ht <- if(is.null(rf$ht) || is.na(rf$ht)) 0 else rft[[sex]]$ht[i,rf$ht+1]
        wc <- if(is.null(rf$wc) || is.na(rf$wc)) 0 else rft[[sex]]$wc[i,rf$wc+1]
        fhx <- if(is.null(rf$fhx) || is.na(rf$fhx)) 0 else rft[[sex]]$fhx[i,rf$fhx+1]
        smk <- if(is.null(rf$smk) || is.na(rf$smk)) 0 else rft[[sex]]$smk[i,rf$smk+1]
        bmi <- if(is.null(rf$bmi) || is.na(rf$bmi)) 0 else rft[[sex]]$bmi[i,rf$bmi+1]
        exp(ht+wc+fhx+smk+bmi)
}

# load.rf.table: 표준화 log(OR)들을 저장하는 dataframe을 생성함
# risk_coef.dta: 연령, 성별 log(OR) 정보를 갖고 있는 STATA file
```

```
# risk_prop.dta: risk factor의 연령, 성별 distribution이 있는 STATA file
load.rf.table<-function(){
        library(foreign)
        r_coef<-data.frame(read.dta("risk_coef.dta"))
        r_prop<-data.frame(read.dta("risk_prop.dta"))
        tempvec<-r_coef$male[1:2]
        temp_pr<-r_prop[c("ht_m0","ht_m1")]
        m_ht_coef<-matrix(tempvec, byrow=T, nrow=6, ncol=2)-apply(temp_pr, 1,
function(x) crossprod(x, tempvec))
        tempvec<-r_coef$female[1:2]
        temp_pr<-r_prop[c("ht_f0","ht_f1")]
        f_ht_coef<-matrix(tempvec, byrow=T, nrow=6, ncol=2)-apply(temp_pr, 1,
function(x) crossprod(x, tempvec))
        tempvec<-r_coef$male[3:4]
        temp_pr<-r_prop[c("wc_m0","wc_m1")]
        m_wc_coef<-matrix(tempvec, byrow=T, nrow=6, ncol=2)-apply(temp_pr, 1,
function(x) crossprod(x, tempvec))
        tempvec<-r_coef$female[3:4]
        temp_pr<-r_prop[c("wc_f0","wc_f1")]
        f_wc_coef<-matrix(tempvec, byrow=T, nrow=6, ncol=2)-apply(temp_pr, 1,
function(x) crossprod(x, tempvec))
        tempvec<-r_coef$male[5:6]
        temp_pr<-r_prop[c("smk_m0","smk_m1")]
        m_smk_coef<-matrix(tempvec, byrow=T, nrow=6, ncol=2)-apply(temp_pr,
1, function(x) crossprod(x, tempvec))
        tempvec<-r_coef$female[5:6]
        temp_pr<-r_prop[c("smk_f0","smk_f1")]
        f_smk_coef<-matrix(tempvec, byrow=T, nrow=6, ncol=2)-apply(temp_pr, 1,
function(x) crossprod(x, tempvec))
        tempvec<-r_coef$male[7:8]
        temp_pr<-r_prop[c("fhx_m0","fhx_m1")]
        m_fhx_coef<-matrix(tempvec, byrow=T, nrow=6, ncol=2)-apply(temp_pr,
1, function(x) crossprod(x, tempvec))
        tempvec<-r_coef$female[7:8]
```

```
        temp_pr<-r_prop[c("fhx_f0","fhx_f1")]
        f_fhx_coef<-matrix(tempvec, byrow=T, nrow=6, ncol=2)-apply(temp_pr, 1,
function(x) crossprod(x, tempvec))
        tempvec<-r_coef$male[9:11]
        temp_pr<-r_prop[c("bmi_m0","bmi_m1","bmi_m2")]
        m_bmi_coef<-matrix(tempvec, byrow=T, nrow=6, ncol=3)-apply(temp_pr,
1, function(x) crossprod(x, tempvec))
        tempvec<-r_coef$female[9:11]
        temp_pr<-r_prop[c("bmi_f0","bmi_f1","bmi_f2")]
        f_bmi_coef<-matrix(tempvec, byrow=T, nrow=6, ncol=3)-apply(temp_pr, 1,
function(x) crossprod(x, tempvec))
        list(male=list(bmi=m_bmi_coef, wc=m_wc_coef, ht=m_ht_coef,
smk=m_smk_coef, fhx=m_fhx_coef),
        female=list(bmi=f_bmi_coef, wc=f_wc_coef, ht=f_ht_coef,
smk=f_smk_coef, fhx=f_fhx_coef))
}
```

```
# R code for test and validation step
# initialize step
hr<-1.62                                # mortality ratio between non-DM and DM
fp<-read.table("male_data.txt",header=T, sep="\t", dec=".")
m_tmat<-make_tmat(fp, hr)   # transitional matrix of men with standard risk
m_smat<-make_smat(fp)                   # observational matrix of men

fp<-read.table("female_data.txt",header=T, sep="\t", dec=".")
f_tmat<-make_tmat(fp, hr)    # transitional matrix of women with standard risk
f_smat<-make_smat(fp)                   # observational matrix of women

rft <-load.rf.table()               # load OR table by risk factors

# test, standard-risk men without DM, by age
result<-matrix(0, nrow=0, ncol=7)
for(i in seq(20,80,by=5)){
        result<-rbind(result,c(i,(hmm_est(noDMm_tmat, m_smat, i))))
}
colnames(result)<-c( "age","nl", "DM1", "DM2", "LR", "life", "DM")
write.table(result,file="result_m_sDM.txt",row.names=F,sep="\t")

# test, standard-risk women without DM, by age
result<-matrix(0, nrow=0, ncol=7)
for(i in seq(20,80,by=5)){
        result<-rbind(result,c(i,(hmm_est(noDMf_tmat, f_smat, i))))
}
colnames(result)<-c( "age","nl", "DM1", "DM2", "LR", "life", "DM")
write.table(result,file="result_f_sDM.txt",row.names=F,sep="\t")

# test, with individual risk
m_or <- mix_or(rf,rft)                  # calculate mixed OR according to individual
risk factors
mr_tmat <- risk_tmat(m_tmat, m_or)   # transitional matrix according to individual
risk factors
mr_tmat <- risk_tmat(m_tmat, 15)
```

```
for(i in 20:60){
        print(hmm_est(mr_tmat, m_smat, i))
}

# test with sample data
ind_inf <- read.dta("sampledata.dta")
result<-matrix(nrow=0, ncol=8)
for(i in 1:nrow(ind_inf)){
        temp_inf<-ind_inf[i,]
        temp_or<-mix_or(temp_inf, rft)
        if(temp_inf$sex==1){
                temp_tmat<-risk_tmat(m_tmat, temp_or)
                result<-rbind(result, c(temp_inf$id, temp_inf$age,
hmm_est(temp_tmat, m_smat, temp_inf$age)))
        } else {
                temp_tmat<-risk_tmat(f_tmat, temp_or)
                result<-rbind(result, c(temp_inf$id, temp_inf$age,
hmm_est(temp_tmat, m_smat, temp_inf$age)))
        }
}
colnames(result)<-c( "id","age","nl", "DM1", "DM2", "LR", "life", "DM")
write.table(result,file="result_ind.txt",row.names=F,sep="\t")

# correction of risk factors in P5
temp_p5<-ind_inf[5,]
temp_p5<-rbind(temp_p5, temp_p5, temp_p5, temp_p5)
temp_p5[2,]$bmi<-0
temp_p5[3,]$wc<-0
temp_p5[4,]$wc<-0
temp_p5[4,]$bmi<-0
result<-matrix(nrow=0, ncol=8)
for(i in 1:nrow(temp_p5)){
        temp_inf<-temp_p5[i,]
        temp_or<-mix_or(temp_inf, rft)
        if(temp_inf$sex==1){
```

```
                temp_tmat<-risk_tmat(m_tmat, temp_or)
                result<-rbind(result, c(temp_inf$id, temp_inf$age,
hmm_est(temp_tmat, m_smat, temp_inf$age)))
        } else {
                temp_tmat<-risk_tmat(f_tmat, temp_or)
                result<-rbind(result, c(temp_inf$id, temp_inf$age,
hmm_est(temp_tmat, m_smat, temp_inf$age)))
        }
}
colnames(result)<-c( "id","age","nl", "DM1", "DM2", "LR", "life", "DM")
write.table(result,file="result_P5.txt",row.names=F,sep="\t")

# validation with validation set
val_set<-read.dta("val_set.dta")
val_set$nl_p<-NA
val_set$dm_p<-NA
for(i in 1:nrow(val_set)){
        temp_inf<-val_set[i,]
        temp_or<-mix_or(temp_inf, rft)
        if(temp_inf$sex==1){
                temp_tmat<-risk_tmat(m_tmat, temp_or)
        } else {
                temp_tmat<-risk_tmat(f_tmat, temp_or)
        }
        temp_cmat<-mat_cmult(temp_tmat, 1, temp_inf$age-20)
        val_set[i,]$nl_p<-temp_cmat[1,1]
        val_set[i,]$dm_p<-temp_cmat[2,1]
}
val_set$pr<- val_set$dm_p/(val_set$nl_p+val_set$dm_p)
library(pROC)
roc1<-roc(val_set$dm~val_set$pr)
roc2<-roc(val_set$dm~val_set$k_score)
plot.roc(roc1,lty="solid", print.auc=T, print.auc.x=0.9, print.auc.y=1.0)
plot.roc(roc2,lty="dashed", print.auc=T, print.auc.x=0.5, print.auc.y=0.8, add=T)
write.dta(val_set, file="est_val_set.dta")
```

```
// 기초 자료 생성을 위한 STATA do file
// 사전 data cleaning & data step
use hn10_all.dta , clear
append using hn11_all.dta
keep if inrange(age, 20, 84)
gen dm=.
replace dm=0 if de1_31<. & de1_32<. & !inrange(he_glu, 126, 1000) & !inrange(he_hba1c, 6.5, 100)
replace dm=1 if (de1_31==1) | (de1_32==1) | inrange(he_hba1c, 6.5, 100) | inrange(he_glu, 126, 1000)
replace dm=. if he_glu==. & he_hba1c==. & !( de1_31==1 | de1_32==1)
drop if dm==.
gen recog=cond(de1_pr==1, 1, 0) if dm==1
gen fhx=.
replace fhx=1 if (he_dmfh1==1) | (he_dmfh2==1) | (he_dmfh3==1)
replace fhx=0 if (he_dmfh1==0) & (he_dmfh2==0) & (he_dmfh3==0)
gen sbp=cond(year==2010, he_sbp_tr, he_sbp)
gen dbp=cond(year==2010, he_dbp_tr, he_dbp)
gen ht= inrange(di1_2,1,4) | ((sbp>=140) & (sbp<.)) | ((dbp>=90)&(dbp<.))
replace ht=. if di1_2==. & sbp==. & dbp==.
gen wc=1
replace wc=0 if sex==1 & he_wc<90
replace wc=0 if sex==2 & he_wc<85
replace wc=. if he_wc==.
gen smk=sm_pr
recode he_bmi (min/25=1)(25/30=2)(30/max=3), gen(bmi)
recode age (20/29=1)(30/39=2)(40/49=3)(50/59=4)(60/69=5)(70/max=6), gen(age10)

set seed 12345678
gen group=cond(runiform()<0.5, 0, 1)

save hmm_base.dta, replace    // training set + validation set

/* training set */
```

```
keep if group==0
/* restricted spline curve와 logistic regression을 통해 유병률 구하기 */
preserve
keep if sex==1
mkspline age = age, cubic knots(25 40 50 60 75)
logistic dm age1-age4
predict pr
keep age pr
duplicates drop age, force
sort age
save male_pr.dta, replace
restore

preserve
keep if sex==2
mkspline age = age, cubic knots(25 40 50 60 75)
logistic dm age1-age4
predict pr
keep age pr
duplicates drop age, force
sort age
save female_pr.dta, replace
restore

/* restricted spline curve와 logistic regression을 통해 인식률 구하기 */
preserve
keep if sex==1
mkspline age = age, cubic knots(25 40 50 60 75)
logistic recog age1-age4 if dm==1
predict sn
keep age sn
duplicates drop age, force
sort age
save male_sn.dta, replace
```

```
restore



preserve
keep if sex==2
mkspline age = age, cubic knots(25 40 50 60 75)
logistic recog age1-age4 if dm==1
predict sn
keep age sn
duplicates drop age, force
sort age
save female_sn.dta, replace
restore

//연령대별 위험도 계산
logistic dm b0.ht b0.wc b0.smk b0.fhx b1.bmi age if sex==1
matrix male=e(b)'
logistic dm b0.ht b0.wc b0.smk b0.fhx b1.bmi age if sex==2
matrix female=e(b)'
foreach x of varlist ht wc fhx smk bmi {
proportion `x' if dm==0 & sex==1, over(age10)
matrix `x'_m=e(b)'
proportion `x' if dm==0 & sex==2, over(age10)
matrix `x'_f=e(b)'
}

preserve
clear
matrix gr=(0\0\0\0\0\0\1\1\1\1\1\1\2\2\2\2\2\2)
matrix agegr=(0\1\2\3\4\5\0\1\2\3\4\5\0\1\2\3\4\5)
svmat gr
svmat agegr
foreach x in ht wc fhx smk bmi {
```

```
svmat `x'_m
svmat `x'_f
}
rename *1 *
reshape wide ht_m-bmi_f, i(agegr) j(gr)
order *, alphabetic
save "risk_prop.dta", replace
outsheet using "risk_prop.txt", replace

clear
svmat male
svmat female
rename *1 *
drop in 12/13
gen factor=""
replace factor="ht0" in 1
replace factor="ht1" in 2
replace factor="wc0" in 3
replace factor="wc1" in 4
replace factor="smk0" in 5
replace factor="smk1" in 6
replace factor="fhx0" in 7
replace factor="fhx1" in 8
replace factor="bmi0" in 9
replace factor="bmi1" in 10
replace factor="bmi2" in 11
save "risk_coef.dta", replace
outsheet using "risk_coef.txt", replace
restore

// 사망률 자료 만들기
clear
set obs 65
gen age=_n+19
```

```
merge 1:1 age using "mortality_base.dta"
drop _merge
regress ln_m age
predict ln_m_p, xb
regress ln_f age
predict ln_f_p, xb
gen death_m=exp(ln_m_p)
gen death_f=exp(ln_f_p)

preserve
keep age death_m
keep if inrange(age, 20, 84)
sort age
rename death_m death
save "male_death.dta", replace
restore

preserve
keep age death_f
keep if inrange(age, 20, 84)
sort age
rename death_f death
save "female_death.dta", replace
restore

// 자료 합치기
use "male_death.dta", clear
merge 1:1 age using "male_pr.dta"
drop _merge
merge 1:1 age using "male_sn.dta"
drop _merge
save "male_data.dta", replace
outsheet using "male_data.txt", replace
```

```
use "female_death.dta", clear
merge 1:1 age using "female_pr.dta"
drop _merge
merge 1:1 age using "female_sn.dta"
drop _merge
save "female_data.dta", replace
outsheet using "female_data.txt", replace

/* make validation set and data step */
use "hmm_base.dta", clear
keep if group==1
keep if inrange(age, 21, 83)

gen alcf=.
replace alcf=0 if inlist(bd1_11, 1, 8)
replace alcf=0.5 if bd1_11==2
replace alcf=1 if bd1_11==3
replace alcf=3 if bd1_11==4
replace alcf=10 if bd1_11==5
replace alcf=20 if bd1_11==6

gen alcd=.
replace alcd=0 if bd2_1==8
replace alcd=1.5 if bd2_1==1
replace alcd=3.5 if bd2_1==2
replace alcd=5.5 if bd2_1==3
replace alcd=8 if bd2_1==4
replace alcd=12 if bd2_1==5

gen alctotal=alcf*alcd/30
recode alctotal (5/max=2)(1/5=1)(min/1=0),gen(k_alc)
recode age (min/34=0)(35/44=2)(45/max=3), gen(k_age)
gen k_wc=3
replace k_wc=2 if sex==1 & he_wc<90
replace k_wc=2 if sex==2 & he_wc<84
```

```
replace k_wc=0 if sex==1 & he_wc<84
replace k_wc=0 if sex==2 & he_wc<77
replace k_wc=. if he_wc==.
gen k_score=k_age+fhx+ht+k_wc+smk+k_alc
drop if k_score==.

preserve
keep age sex dm- k_score
replace bmi=bmi-1
save val_set.dta, replace
restore
```

Abstract

Lifetime Risk Estimation of Diabetes Mellitus using Hidden Markov Model

Name : Jae Moon Yun
Department and Major: Bioinformatics,
Graduate School of Public Health,
Seoul National University

Due to high prevalence and mortality, public concerns of DM (diabetes mellitus) have been increasing. Addressing for the needs, prediction model and lifetime risk of DM may be useful for both public and clinicians. However, there has been no study about lifetime risk of DM based on Korean data and previous similar studies did not apply individual risk factors. This study was conducted to estimate the lifetime risk of DM according to individual risk factors for Korean adults and to validate the prediction model. A prediction model was designed using HMM (Hidden Markov Model). Biologically hyperglycemic status was regarded as hidden status, and recognition of DM as observational status. From death report of Statistics Korea, mortalities by age group were extracted. The fifth KNHANES (Korea National Health and Nutrition Examination Survey) was randomly separated into training and validation sets in equal sizes. The prevalence of DM and other data of risk factors were calculated from the training set. The basal parameters of HMM were calculated with all-cause mortality, prevalence and recognition rate of DM. The parameters were adjusted with odd ratios according to individual risk factors. Estimation and validation were performed with validation set. Firstly, estimations with

standard risk were done by gender and age. Life expectancies were lower than previous study, however the differences were consistent regardless of initial age. Secondly, estimations of individual risk factors were performed. The predicted prevalence showed a good results comparing to Korean diabetes prediction score. The expected durations of unrecognized and recognized DM were also estimated, however these were not validated because of lack of data currently available. In spite of several limitations, the prediction model of this study showed a new type of results as well as plausible performance and is proved to be easily modifiable for new risk factors in case of future improvements. As a result, this model is expected to help motivation for risk reduction and to provide useful information for medical decisions.

Keywords: diabetes mellitus, lifetime risk, hidden Markov model, individual risk, validation

Student Number: 2011-22112