The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.1.013
30(1), 13–24

A visual query database system for the Sample Research DB of the National Health Insurance Service

Sang-Hoon Cho[a] · HeeChan Kim[b] · Gunseog Kang[a,1]

[a]Department of Statistics and Actuarial Science, Soongsil University;
[b]Department of Software Convergence, Graduate School, Soongsil University

(Received October 27, 2016; Revised December 12, 2016; Accepted December 12, 2016)

Abstract

The Sample Cohort DB supplied by the National Health Insurance Service is a valuable resource for statistical studies as well as for health and medical studies. It takes significant time and effort to extract data from this Cohort DB having a large size. As such, we introduce a database system, conveniently called the **N**ational Health **I**nsurance Service **C**ohort DB **E**xtract **Tool** (NICE Tool), which supports several useful operations for effectively and efficiently managing the Cohort DB. For example, researchers can extract variables and cases related with study by simply clicking a computer mouse without any prior knowledge regarding SAS DATA step or SQL. We expect that NICE Tool will facilitate the faster extraction of data and eventually lead to the active use of the Cohort DB for research purposes.

Keywords: visual query, database system, sample survey DB, NHISS

1. 서론

국민건강보험공단(약칭: 건보공단)에서 보유하고 있는 국민건강정보DB는 전 국민의 자격 및 보험료, 건강검진결과, 진료내역, 노인장기요양보험 자료, 요양기관 현황, 암 및 희귀난치성질환자 등록정보 등 2조 1천억 건에 달하는 방대한 빅데이터를 포함하고 있다(www.nhis.or.kr). 그리고 이와 같은 국민건강정보 자료들이 정책이나 학술연구에 원활하게 활용될 수 있게 하기 위한 방안으로 건보공단에서는 '국민건강보험자료공유서비스'(National Health Insurance Sharing Service; NHISS)를 운영하고 있다(nhiss.nhis.or.kr). 이 사이트에서는 자료 이용자들에게 국민건강정보 자료의 신청에서부터 연구 성과를 공유하는데 이르기까지 다양한 정보를 제공하고 있다. 또한 건보공단은 2015년부터 매분기마다 표본연구DB 활용 교육을 실시하고 있으며, 이를 통해 연구자들이 자료를 효율적으로 사용하여 통계분석을 원활히 수행할 수 있도록 도움을 주고 있다.

이와 같은 건보공단의 노력으로 표본연구DB를 활용한 분석이 활성화되고 있으며, 특히 보건의료 분야에서 중요한 연구 결과들이 학술지들을 통해 출간되고 있다. 2016년 9월까지 NHISS에서 조사한 논문 편수는 약 33편으로 집계되어 있지만, 조사에 포함되지 않은 논문 수도 상당할 것으로 짐작된

This study was supported by a faculty research grant of Yonsei University College of Medicine (6-2016-0058).
[1]Corresponding author: Department of Statistics and Actuarial Science, Soongsil University, 369, Sangdo-ro, Dongjak-gu, Seoul 06978, Korea. E-mail: gskang@ssu.ac.kr

다(nhiss.nhis.or.kr/bbs/boards/faq.do). 초창기에는, 예를 들어, 표본코호트DB에서 사례를 추출하는 과정을 설명하는 기술적 측면의 논문도 있었지만 (Yu 등, 2015), 최근에는 의학적인 측면에서 상당히 중요한 연구결과를 도출한 논문들도 상당수에 이른다 (Kim 등, 2016; Ko 등, 2015; Park 등, 2015; Rim 등, 2015, 2016).

건보공단에서 실시하는 표본연구DB 활용 교육에는 매분기마다 50명–100명 정도의 연구자들이 참여하고 있으며, 교육 내용은 database(DB)에 대한 설명, 자료처리, 모의실습 및 시연, 자료를 활용한 연구사례 및 분석과정 소개와 함께 자료분석에 가장 많이 사용되는 SAS 프로그램 활용 등으로 구성되어 있다. 전체적인 자료분석 과정에서 연구자들이 가장 큰 어려움을 겪는 부분은 건보공단의 빅데이터에서 특정 질병을 진단받은 사람들이나 특정 약을 처방받은 사람들을 추출하는 과정이다. 이 과정은 일반적으로 SAS에서 DATA문이나 structured query language(SQL)문을 사용하여 해당되는 사람들의 자료를 DB를 구성하고 있는 여러 테이블에서 추출하는 일련의 연속작업이다. SAS나 SQL문에 익숙한 통계전문가들은 어느 정도 추가적인 교육을 받은 후에 어렵지 않게 자료를 추출과정의 수행할 수 있지만 의료 또는 보건계통의 일반 연구자들에게는 많은 시간과 노력이 소모되는 작업이라 할 수 있다.

본 논문에서는 이러한 연구에 필요한 사례 추출과정에 도움을 주는 데이터베이스 시스템인 **N**ational Health **I**nsurance Service **C**ohort DB **E**xtract **Tool**(NICE Tool)을 소개하고자 한다. SAS의 DATA 명령문이나 SQL문에 익숙하지 않은 연구자들도 마우스 클릭만으로 각 테이블에서 필요한 변수들과 조건에 맞는 사례들을 추출할 수 있는 기능을 제공하며, 또한 추출 결과를 실시간으로 확인할 수도 있다. 그리고 추출할 내용에 해당하는 쿼리(query)를 자동으로 생성해주므로 작업오류를 줄일 수 있으며, NICE Tool을 사용한 반복적인 쿼리 작업을 통하여 SQL문을 배울 수 있어 전문가로서의 기량을 향상시킬 수도 있을 것이다. 의학 또는 보건계통의 연구자들도 NICE Tool을 활용하면 연구에 사용할 자료를 추출하는데 필요한 시간을 크게 단축할 수 있으며, 상대적으로 자료 분석에 더 많은 시간을 할애 할 수 있으므로 더 의미 있는 분석 결과를 도출하는데 도움이 되리라 생각한다. 또한 저자들의 경험상 표본연구DB의 자료의 크기가 대용량이라는 문제점과 SAS를 이용하여 분석할 때 생성되는 임시파일의 용량이 매우 커지는 문제점 때문에 일반 PC에서 통계패키지를 사용하여 자료를 추출하는 것은 바람직하지 않으며, 데이터베이스 시스템을 활용하여 자료를 추출하는 것이 더욱 효과적이다.

2. 건강보험공단의 표본연구DB

국민건강보험공단에서는 현재 '표본연구DB', '맞춤형DB', '건강질병지표' 등을 제공하고 있다. '건강질병지표'는 건보공단의 빅데이터를 활용하여 만성질환의 위험요인뿐만 아니라 발생 및 합병증 과정을 포괄하여 체계적으로 산출한 만성질환 관리 지표를 말하며, '맞춤형DB'는 건보공단이 수집, 보유, 관리하는 건강정보자료를 정책 및 학술 연구 목적으로 이용할 수 있도록 수요맞춤형 자료로 가공하여 제공하는 데이터이다. 그리고 '표본연구DB'는 국민건강정보DB의 방대한 규모와 개인정보보호 문제 등으로 연구자의 접근과 활용이 제한적이었던 점을 획기적으로 개선하고자 표본을 추출하여 외부 반출이 가능한 형태로 규격화한 데이터를 말하며, 동일 대상자에 대해 사회·경제적 변수(거주지, 사망년월, 사망사유, 소득수준 등)가 포함된 자격자료, 진료내역 및 건강검진자료 등을 연결한 코호트 자료로 장기간의 관찰이 가능하여 시간적 선후관계나 인과적 관계 분석이 가능한 자료이다(nhiss.nhis.or.kr). 현재는 '표본코호트DB', '건강검진코호트DB', '노인코호트DB'가 제공되고 있으며, 앞으로 '직장여성코호트DB'와 '영유아검진코호트DB'가 추가로 제공될 예정이다.

일반적으로 많은 연구자들이 관심을 갖는 DB는 '표본연구DB'이며, 본 논문에서 소개하는 데이터베이스 시스템도 기본적으로 이 DB를 위한 것이다. 따라서 시스템을 설명하기 전에 먼저 '표본연구DB'에 대한 정보를 간단하게 정리하고자 한다.

Table 2.1. Status of the Sample Cohort DB (2002–2013) (NHIS, 2016a)

연번	세부DB			파일수 (개)	건수 (백만건)	파일크기 (GB)
1	자격DB			12	12.0	0.62
2	진료DB	의과_보건기관(T1)	명세서(20t)	12	119.4	14.49
3			진료내역(30t)	12	577.0	46.02
4			상병내역(40t)	12	299.4	15.74
5			처방전교부상세내역(60t)	12	399.3	33.44
6		치과_한방(T2)	명세서(20t)	12	27.1	3.03
7			진료내역(30t)	12	128.7	10.26
8			상병내역(40t)	12	32.4	1.71
9			처방전교부상세내역(60t)	12	7.5	0.61
10		약국(T3)	명세서(20t)	12	99.2	11.43
11			진료내역(30t)	12	913.4	73.07
12	건강검진DB			12	2.0	0.20
13	요양기관DB			12	1.2	0.04
전체	합계			156	2,618.6	210.67

먼저 '표본코호트DB'는 5천만 전 국민의 2%에 해당되는 약 100만명을 2002년 자격 대상자 중 성별, 연령, 소득수준을 층화변수로 추출하여 2002년부터 2013년까지 12년 동안의 사회/경제적 자격 변수(장애 및 사망 포함), 의료이용(진료 및 건강검진) 현황, 요양기관 현황 자료를 포함하고 있다.

'표본코호트DB'는 세부적으로 '자격DB', '진료DB', '건강검진DB', '요양기관DB'로 구성되어 있다. '자격DB'는 건강보험가입자 및 의료급여수급권자(외국인 제외)를 대상으로 성, 연령대, 지역, 가입자 구분, 소득분위 등 대상자의 사회경제적 변수 및 장애, 사망관련 총 14개 변수로 구성되어 있다. '진료DB'는 대상자가 요양기관에 방문하여 진료 등을 받은 내역에 대해 요양기관으로부터 요양급여가 청구된 자료를 포함하고 있으며, 다시 의과/보건기관(T1), 치과/한방(T2), 약국(T3)자료에 대한 명세서(20t), 진료내역(30t), 상병내역(40t), 처방전교부상세내역(60t) 등의 10개 세부DB로 구성되어 있다. '건강검진DB'는 건강검진 주요 결과 및 문진에 의한 생활습관 및 행태관련 자료로 2009년도의 검진제도 개편으로 인하여 2002년–2008년, 2009년–2013년으로 나누어 별도로 구성되어 있다. 그리고 '요양기관DB'는 요양기관의 종별, 설립구분별, 지역(시도)별 현황 및 시설, 장비, 인력관련 자료 변수 등 총 10개의 변수로 구성되어 있다.

'표본코호트DB'의 자세한 구성은 Table 2.1에 주어져 있다. 그리고 '표본코호트DB'의 작성과정에 대한 설명은 Lee와 Kim (2012)과 Lee 등 (2016)을 참조하면 되며, DB 내용과 사용방법에 대한 자세한 설명은 국민건강보험공단의 빅데이터운영실에서 작성한 '표본 코호트DB 사용자 매뉴얼' (NHIS, 2016a)을 참조하면 된다.

'표본코호트DB' 내의 '건강검진DB'도 건강검진의 주요 결과에 대한 정보를 제공하여 주지만 더욱 정확한 분석을 위하여 2016년부터 추가적으로 '건강검진코호트DB'를 제공하고 있다. 이 DB는 2002년 자격유지자 중 2002년–2003년 사이에 일반건강검진을 받은 40세–79세의 수검자(약 51만명)들에 대한 2002년–2013년(12개년) 동안의 정보를 포함하고 있으며, 그 기본적인 구성은 '표본코호트DB'와 동일하다. 그리고 '노인코호트DB'는 2002년 자격유지자 중 만 60세 이상 대상자(약 55만명)에 대한 12개년 동안의 정보를 포함하고 있다.

3. 시스템 설명

이 장에서는 국민건강보험공단에서 제공하는 표본코호트DB에서 필요한 자료를 쉽게 추출 및 관리할 수

(a) (b)

Figure 3.1. Start-up window of NICE Tool(a) and a window with the Sample Cohort DB(b).

있는 도구인 NICE Tool에 대해 소개한다. NICE Tool을 사용하면 DB와 SQL에 대한 이해가 없는 연구자라도 마우스 클릭(click)만으로 건보공단 자료에서 연구에 사용할 자료를 직관적으로 손쉽게 추출할 수 있다.

3.1. NICE Tool 구현 도구, 사용 환경 및 기능

NICE Tool은 기본적으로 비주얼 쿼리(visual query)를 제공하는 데이터베이스 시스템이다. NICE Tool은 내부 DB로 오픈 소스(open source)인 MySQL(www.mysql.com; Welling과 Thomson, 2016)을 사용하고 있는데, 이 시스템은 MySQL을 위한 웹-인터페이스(web-interface)인 phpMyAdmin(www.phpmyadmin.net)으로 구현되었다. 명칭에서 알 수 있듯이 phpMyAdmin은 PHP 컴퓨터 언어(Naramore, 2015; Welling과 Thomson, 2016)를 사용하여 작성되었는데, 오픈 소스이기 때문에 사용자들이 직접 소스 코드를 수정할 수 있다.

NICE Tool 시스템 운영을 위한 서버 컴퓨터에는 데이터베이스 처리를 위한 MySQL, 웹서비스 제공을 위한 Apache, 그리고 PHP 언어 구동 환경만 구축되어 있으면, 운영체계(MS Window, Linux, MacOS)에 상관없이 설치가 가능하다. 그리고 NICE Tool 사용자는 컴퓨터 운영 체계에 제한을 받지 않고 웹브라우저(web browser)가 설치된 모든 PC에서 NICE Tool이 설치된 서버에 접속하여 사용할 수 있으며, 테블릿 PC 및 휴대폰 등의 모바일 환경에서도 사용이 가능하다.

현재 NICE Tool은 국민건강보험공단의 표본연구DB 사용을 위하여 개발되었지만 DB의 종류에 관계없이 사용할 수 있다. NICE Tool은 자료추출에 필요한 거의 모든 주요 기능을 갖추고 있는데, 앞으로 사용경험이 축적됨에 따라 필요한 새로운 기능이나 편리한 기능 등을 언제든지 추가하는 것이 가능하다.

Figure 3.1은 NICE Tool의 초기화면을 보여주며, Table 3.1에는 NICE Tool의 대표적인 기능들이 요약되어 있다. 표본코호트DB에 포함되어 있는 테이블들의 내부 관계가 테이블 간에 자료 검색 및 통합을 위해 지정되어 있으며, 테이블들은 '개인일련번호', '청구일련번호', '요양기관식별대체번호'를 참조

Table 3.1. Important features of NICE Tool

기능	설명
탑재된 DB들과 소속 테이블 탐색	Tree형태와 테이블형태로 탐색 가능 테이블형태에서는 자료 수도 함께 표시
각 테이블의 변수명과 원자료값 탐색	테이블형태로 제공 스크롤 기능, 검색, 행 필터링, 정렬 기능을 제공
테이블 간의 내부 관계 표시	DB를 탑재할 때 설정된 테이블 간의 내부 관계를 자동으로 표시
테이블 또는 변수 선택 기능	각 테이블 또는 변수 옆에 체크박스가 있어 이를 이용하여 선택 또는 해제가 가능
변수의 조건 설정	변수 각각에 대한 추출 조건을 관계연산자 등을 이용하여 지정할 수 있음 설정된 조건을 확인/수정/삭제가 가능
자료 추출을 위한 쿼리의 생성	'질의 마법사'기능을 이용하여 쿼리를 자동 생성
추출된 자료 보여주기	작성된 쿼리를 통하여 추출된 자료를 기존의 테이블과 동일한 양식으로 보여줌
임시 인덱스 작성	추출과정에서 임시 인덱스를 작성하여 인덱스에 해당하는 자료들만 추출 또는 제거 가능
추출된 테이블과 기존 테이블과의 병합	새로 작성된 테이블과 기존의 테이블을 이용하여 새로운 쿼리 작업 가능
테이블 내보기	기존 테이블 또는 새로 작성된 테이블을 CSV(default), Excel, PDF, XML, MS-WORD 등의 다양한 형태로 내보내기가 가능
관계 연산자	=, <, >, <=, >=, NOT, IN, NOT IN, LIKE
합계 연산자	DISTINCT, SUM, MIN, MAX, AVG, COUNT

키(foreign key)로 하여 서로 연결되어 있다. 따라서 여러 테이블을 통합하여 자료를 추출하는 경우에는 이들을 연결하고 있는 참조키에 대한 정보가 필요한데, NICE Tool의 경우에는 이를 시각화하여 표시해 주고 있으므로 테이블 간의 내부 관계 및 참조키 값을 쉽게 파악할 수 있다.

Table 3.1의 '관계 연산자' 중에서 'IN'은, 예를 들어, 여러 개의 상병 목록을 주고 그 중에 해당되는 상병이 있는 사례를 모두 추출하는 연산자이고, 'LIKE'는, 예를 들어, LIKE 'I9%'이라 설정하면 I9로 시작되는 3자리–5자리 코드로 표시되는 심근경색(stroke)과 관련된 상병들을 모두 추출하는 연산자이다. 그리고 '합계 연산자'의 'DISTINCT'는 자료들 중에 동일한 ID를 갖는 행들이 여러 개 있을 때 그 중에서 중복없이 한 개의 행만 추출하는 기능이다(SAS DATA문에서의 NODUPKEY 기능에 해당). 이 외에도 임시 인덱스 테이블을 작성한 후, 이 인덱스 테이블에 속하는 사례들만을 추출하든지, 또는 인덱스 테이블에 속하는 사례들을 기존의 테이블에서 모두 제거하는 등의 유용한 기능도 포함하고 있다(4장의 [예제 2] 참조).

3.2. NICE Tool의 일반적인 사용방법

대부분의 사례 추출과정은 비슷한 절차를 따르게 된다. 먼저 해당되는 DB에서 필요한 테이블을 선택하고 내용을 탐색한다. 이때 검색과 정렬 등의 기능이 유용할 수 있다. 그리고 추출과정을 위해서는 대부분 '질의 마법사'를 사용하게 되는데, 이 기능을 선택하면, 표본코호트DB의 경우에는 우선 '검진'과 '자격' 테이블은 기본적으로 선택되어 보이도록 설정되어 있으며, 두 테이블이 개인일련번호를 참조키로 사용하여 내부적으로 연관되어 있음을 시각적으로 확인할 수 있다(Figure 4.1 참조).

각 테이블에 나열된 변수들의 왼쪽에 위치한 사각형 모양의 체크박스를 이용하여 해당 변수의 선택 여부를 표시할 수 있고, 오른쪽의 '옵션' 아이콘을 선택하면 각 변수에 대해 특정값을 지정하는 조건들

을 설정할 수 있다. 테이블 선택, 변수 선택, 변수 조건 설정 등이 마무리 되면 해당 작업에 대한 쿼리(query)를 '쿼리 생성' 버튼을 클릭하여 자동으로 생성하여 실행할 수 있다. 쿼리 결과 보기가 가능하며, '내보내기' 기능을 사용하여 추출된 자료를 별도의 파일로 저장하여 통계분석에 사용할 수 있다. 또는 쿼리 결과로 작성된 테이블과 기존의 테이블을 함께 사용하여 다른 쿼리 작업을 수행할 수 있다.

NICE Tool은 현재 http://statistics.ssu.ac.kr/~statistics/YM에 탑재되어 있다. 이 주소로 서버에 접속한 후, 로그인 화면에서 사용자명으로 tester, 암호로 tester를 입력하면 시스템을 사용할 수 있다.

4. 사용 예제

NICE Tool은 이미 약 210GB 크기의 '표본코호트DB'와 약 160GB 크기의 '건강검진코호트DB'를 탑재하여 필요한 자료를 추출하는데 사용해 보았으며, 실행한 작업들을 모두 성공적으로 수행해냈다. 그러나 이들 DB를 사용하기 위해서는 건보공단의 사전승인을 받아야하고, 또한 사용범위가 제한되기 때문에 NICE Tool 시스템을 소개하기 위한 자료로는 적당하지 않다. 그래서 본 논문에서는 건보공단의 표본연구DB 활용 교육에서 사용하는 표본코호트DB의 샘플 데이터를 탑재하여 활용한 예제를 소개하고자 한다. 이 샘플 데이터는 NHISS에서 내려 받을 수 있으며, 2010년 자격자료에서 무작위 추출한 1천명에 대한 2010년과 2011년의 자료로 전체 자료구조는 표본코호트DB와 동일하다 (NHIS, 2016a).

4.1. [예제 1] 연구주제: 2010년도 30, 40대 직장 건강보험 가입자들의 비만도와 고혈압, 고혈당 및 고콜레스테롤혈증의 관련성 연구

위의 연구를 수행하기 위해서는 30, 40대 직장인의 허리둘레, 혈압, 혈당, 총콜레스테롤값들을 추출하여야 한다. 위의 변수들 중 나이와 직장인 여부는 '자격' 테이블에 포함되어 있으며, 허리둘레, 혈압, 혈당과 총콜레스테롤값은 '건강검진' 테이블에 포함되어 있다. 그리고 두 테이블에는 '개인일련번호'가 모두 포함되어 있으므로 이를 참조키로 하여 두 테이블을 조인(join)하여 해당되는 대상자와 관련 변수들의 값들을 추출할 수 있다. 이 작업을 위한 SAS 코드는 아래에 주어져 있다. 기본적으로 SAS에서는 DATA 명령문들을 사용하거나, PROC SQL문을 사용할 수 있다. 여기에는 건보공단 교육에서 사용된 PROC SQL문을 그대로 인용하였다 (NHIS, 2016b).

[예제 1]을 위한 SAS 명령문

```
/* 2010년도 30, 40대 직장인 중 2010년도 건강검진 수검자 추출 및 필요 변수 가져오기 */
PROC SQL;
CREATE TABLE TEST1 AS
SELECT A.STND_Y, A.PERSON_ID, A.AGE_GROUP, A.IPSN_TYPE_CD,
         B.HCHK_YEAR, B.YKIHO_GUBUN_CD, B.HEIGHT, B.WEIGHT, B.WAIST,
         B.BP_HIGH, B.BP_LWST, B.BLDS, B.TOT_CHOLE
FROM TEST.NHID_JK_2010_TEST AS A INNER JOIN TEST.NHID_GJ_2010_TEST AS B
ON A.PERSON_ID=B.PERSON_ID
WHERE A.AGE_GROUP IN ('7','8','9','10') AND A.IPSN_TYPE_CD='5';
QUIT;
```

위의 작업을 NICE Tool을 이용하여 수행하는 과정은 다음과 같다.

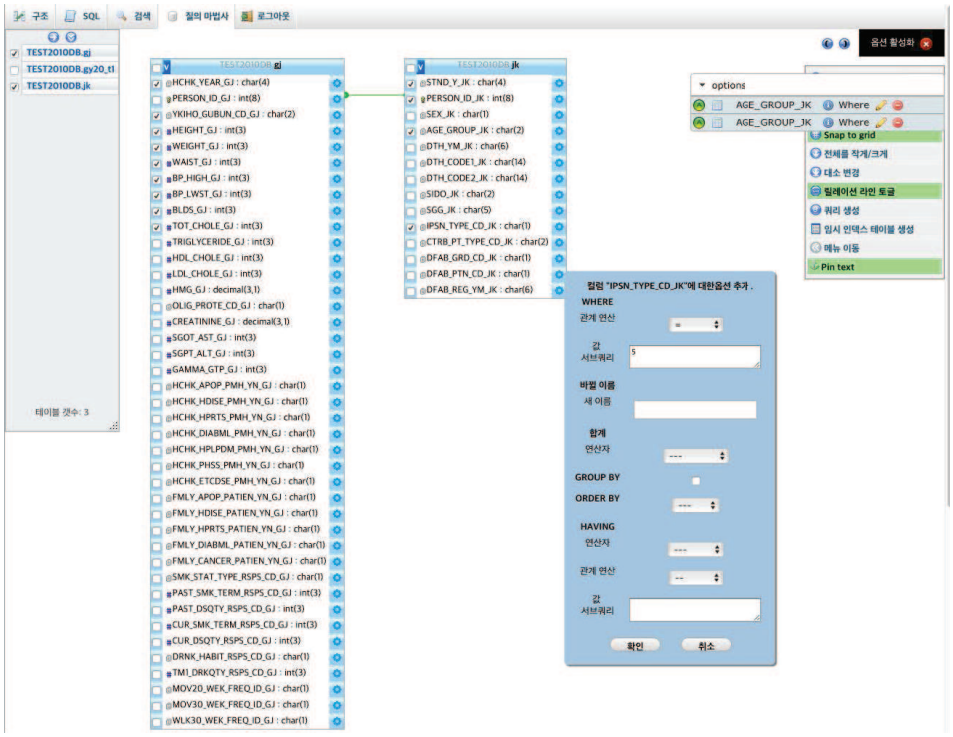

Figure 4.1. Variable option setting window for [Example 1].

(1) 초기화면에서 좌측 패널의 DB tree에서 'TEST2010DB'를 선택한 후, 화면 상단의 '질의 마법사'를 클릭하면 우측 패널에서 '자격DB'와 '건강검진DB' 테이블을 확인할 수 있다.

(2) '자격DB'에서 변수 AGE_GROUP과 IPSN_TYPE_CD를 선택하고(각 변수 왼쪽의 체크박스를 클릭), '건강검진DB'에서 해당 변수들(HEIGHT부터 TOT_CHOLE까지)을 모두 선택한다.

(3) 각 변수에 대한 추출 조건을 입력한다. 예를 들어, 분석대상이 '직장인'이므로 이를 가입자구분을 나타내는 변수인 IPSN_TYOPE_CD의 우측에 있는 '옵션' 버튼을 클릭하여 관계 연산 항목에서 '='을 선택하고, '값 서브쿼리'에 5를 입력한다. AGE_GROUP의 경우에는 7, 8, 9, 10에 해당되는 값을 '>= 7'과 '<= 10'의 두 개의 옵션을 설정하여 지정한다. 이들 옵션들은 '옵션 활성화' 메뉴를 이용하여 확인, 수정, 및 삭제할 수 있다.

(4) 화면 우측에 있는 '질의 마법사 실행 패널'의 '쿼리 생성' 메뉴를 클릭한 후 생성되는 SQL문을 실행시킨다.

Figure 4.1은 위의 단계 (1), (2), (3)을 수행하는 화면을 보여주며, Figure 4.2는 생성된 쿼리와 쿼리를 실행하여 만들어진 테이블을 보여주고 있다.

4.2. [예제 2] 연구주제: 2010년–2011년 30대 이상 제2형 당뇨환자 관련 연구

위의 연구를 수행하기 위해서 2011년 자료에서 2형 당뇨병(E11)을 주상병 또는 부상병으로 처음 진단받은 30대 이상 신규환자들을 추출하여야 한다. 이를 위해서는 '자격' 테이블과 '진료DB' 내의 '명세서'

(a) (b)

Figure 4.2. Created query window in [Example 1](a) and extracted cases(b).

테이블을 사용한다. 두 테이블에는 모두 '개인일련번호'가 포함되어 있으므로 이를 참조키로 하여 두 테이블을 조인(join)하여 해당되는 대상자와 관련 변수들의 값들을 추출할 수 있다. 이 예제에서 주의할 점은 2011년 이전에(즉, 2010년 말까지) 동일한 병으로 청구된 적이 있는 환자는 제외시켜야한다는 것이다. 이 작업을 위한 SAS 코드는 다음과 같다 (NHIS, 2016b).

[예제 2]를 위한 SAS 명령문

```
/* 2011년 2형 당뇨병(E11)을 주/부상병으로 처음 진단 받은 30대 이상 신규 환자 발췌 */

/* 2011년 30대 이상 중 2형 당뇨병(E11) 진단 받은 환자 */
PROC SQL;
 CREATE TABLE TEST2_1_2011 AS
 SELECT DISTINCT
  A.PERSON_ID, A.SEX, A.AGE_GROUP
 FROM TEST.NHID_JK_2011_TEST AS A INNER JOIN TEST.NHID_GY20_T1_2011_TEST AS B
 ON A.PERSON_ID=B.PERSON_ID
 WHERE A.AGE_GROUP IN ('7','8','9','10','11','12','13','14','15','16','17',
 '18') AND (SUBSTR(B.MAIN_SICK,1,3)='E11'  OR SUBSTR(B.SUB_SICK,1,3)='E11');
QUIT;

/* 2010년 30대 이상 중 2형 당뇨병(E11) 진단 받은 환자 */
PROC SQL;
 CREATE TABLE TEST2_1_2010 AS
 SELECT DISTINCT
  A.PERSON_ID, A.SEX, A.AGE_GROUP
 FROM TEST.NHID_JK_2010_TEST AS A INNER JOIN TEST.NHID_GY20_T1_2010_TEST AS B
 ON A.PERSON_ID=B.PERSON_ID
 WHERE A.AGE_GROUP IN ('7','8','9','10','11','12','13','14','15','16','17',
  '18') AND (SUBSTR(B.MAIN_SICK,1,3)='E11'  OR SUBSTR(B.SUB_SICK,1,3)='E11');
QUIT;
```

```
/* 2011년 30대 이상 중 2형 당뇨병(E11)을 처음으로 진단 받은 환자 */
PROC SQL;
 CREATE TABLE TEST2_FIRST AS
 SELECT A.PERSON_ID, A.SEX, A.AGE_GROUP, B.PERSON_ID AS PERSON_ID_2010
 FROM TEST2_1_2011 AS A LEFT JOIN TEST2_1_2010 AS B
 ON A.PERSON_ID=B.PERSON_ID
 WHERE PERSON_ID_2010=.;  /* 2010년도 대상자를 제거하기 위한 조건 */
QUIT;
```

위의 작업을 NICE Tool을 이용하여 수행하는 과정은 두 단계로 나누어진다. 첫 번째는 2010년도 DB에서 당뇨병을 진단 받은 환자를 추출한 후(2010년 명단), 두 번째는 2011년의 환자 명단을 작성하여 2010년 명단의 환자들을 제거하는 과정이다.

(1) [예제 1]과 같은 방법으로 'TEST2010DB'를 선택한 후 질의 마법사를 사용하여 자격(jk) 테이블과 명세서(gy20_t1) 테이블을 선택한 후 자격 테이블에서 PERSON_ID 변수를 선택하여 '합계 연산자' 항목에서 'DISTINCT'를 선택하여 옵션을 만든다(동일한 사람이 여럿 있는 경우 하나만 남기고 제거한다).

(2) AGE_GROUP 변수에 대해 '관계 연산'에서 '>=7' 조건을 설정하고, 주상병 또는 부상병에서 당뇨병(E11x)을 추출하기 위해 MAIN_SICK(주상병)과 SUB_SICK(부상병)변수에 대해 '관계 연산'에서 'LIKE'를 선택하고 "값 서브쿼리"에 'E11%'를 입력하여 조건을 설정한다. 이 경우 당뇨병 코드가 두 변수 중에 하나라도 나타나면 추출되므로 쿼리에서 AND/OR를 명확하게 하기 위하여 '옵션 활성화 패널' 왼쪽의 괄호기호('(', ')')를 사용해야한다. 모든 조건은 AND로 주어지므로 SUB_SICK변수 조건 앞의 'A' 아이콘(AND 의미)을 클릭하여 'O' 아이콘(OR 의미)으로 변환시킨다.

(3) '질의마법사 실행 패널'에서 '임시 인덱스 테이블 생성'을 클릭하여 임시 테이블을 생성한다(DB tree 패널에서 'temp'를 클릭하면 고유번호의 테이블이 생성되었음을 확인할 수 있다. 예: 't1e51' DB).

(4) 같은 방법으로 2011년도 자료에 대해 당뇨병 환자를 추출하는 쿼리를 작성한다.

(5) 2011년도 자료에 대한 쿼리를 생성한 후, 'SELECT 쿼리 창'에서 '임시 인덱스 조건 NOT IN' 버튼을 클릭하면 'temp'DB의 't1e51' 테이블에 포함된 환자가 제외되는 SQL이 자동으로 추가된다. 최종 쿼리를 실행하여 자료를 생성한다.

Figure 4.3은 위의 단계 (4), (5)를 수행한 화면을 보여준다(단계 (1), (2)에 대한 쿼리는 'temp' DB만 제외하면 동일하다).

5. 결론

'표본코호트DB'는 전체 크기가 약 210GB로 일반 데스크탑 PC나 노트북 PC에서 다루기가 부담스럽다. 물론 실제 분석에는 DB내의 일부 변수만을 가지고 사례를 추출하기 때문에 통계분석에 사용되는 최종 자료파일은 대부분 크지는 않다. 그렇지만 자료가 연도별 12개의 DB로 나누어져 있으며, 각 DB는 여러 개의 테이블들로 구성되어 있다. 예를 들어, '개인일련번호'와 '청구일련번호' 등을 매칭

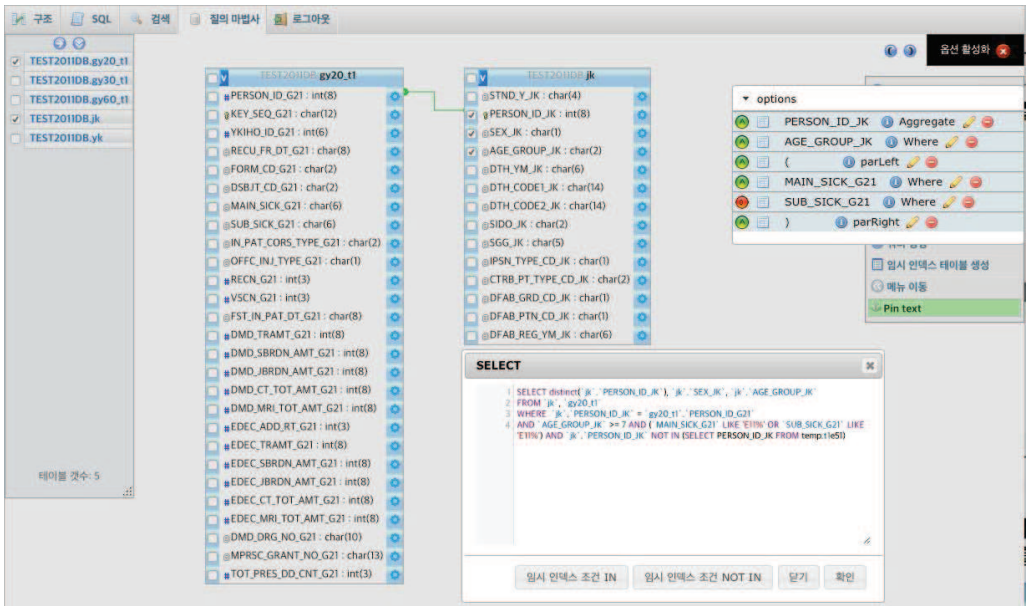

Figure 4.3. Variable option settings for [Example 2] and created query window.

시키면서 사례들을 추출하기 위해서는 서로 다른 DB의 테이블들을 join할 필요가 있다. 이 과정에서 SAS 등에서는 임시파일을 내부적으로 만드는데 이들 임시파일들의 크기도 무시할 수 없을 정도로 크다. 실제로 저자들은 DB를 외장하드에 저장하여 사용함에도 불구하고 특정질병 1–2개의 사례를 추출하는 경우에도 임시파일의 크기로 인하여 작업수행에 실패한 경우를 많이 경험하였고, 이것이 본 논문에서 소개하고자 하는 시스템을 만든 계기가 되었다고 할 수 있다.

본 연구에서 개발된 국민건강보험공단의 표본연구DB를 위한 비주얼 쿼리 데이터베이스 시스템인 NICE Tool은 기본적으로 SAS의 DATA 또는 SQL에 익숙하지 않은 의학 및 보건계통의 연구자들이 표본연구DB에서 필요한 변수나 분석조건에 맞는 사례들을 쉽게 추출할 수 있도록 만들어졌다. 표본연구DB의 자료를 활용하면 유용한 의학적 사실이나 통계적 모형을 구축할 수 있음에도 불구하고 자료추출의 어려움 때문에 연구가 활발히 진행되지 못하는 실정이다. 이 시스템을 활용하면 SAS가 없는 경우에도 필요한 자료를 추출하여 R 등의 통계패키지를 이용하여 필요한 통계분석을 수행할 수 있으며, SAS가 있는 경우에도 사례 추출작업 단계를 단순화하여 전체적인 분석 기간을 단축시킬 수 있다.

NICE Tool은 우선적으로 국민건강보험공단의 표본연구DB 사용을 위하여 작성되었지만 DB의 종류에 관계없이 사용될 수 있다. 예를 들어, 각 병원 또는 연구자들이 보유하고 있는 일반 자료 또는 코호트 자료들을 이 시스템에 탑재하여 편리하게 사용할 수 있다.

현재 NICE Tool은 초기 개발 단계로 앞으로 많은 면에서 기능의 확장이 필요하다. 현재로서도 대부분의 사례추출작업을 할 수 있지만 사용자의 편의를 위한 여러 가지 기능들을 앞으로 추가할 예정이다. 예를 들어, 표본연구DB에는 모든 변수명이 영어 약자로 되어 있어 의학용어에 익숙하지 않은 연구자들에게는 많은 어려움이 있다. 따라서 각 테이블에서 영어 변수명 옆에 한글 변수명을 추가하여 사용의 편리성을 높일 예정이다. 또한, 예를 들어, 상병 등의 변수들은 모두 코드로 입력되어 있다. 따라서 각 코드의 의미를 알기 위해서는 따로 필요한 표들을 참조하여야 한다. 표본연구DB의 사용에 필요한 이러한 표들을 시스템에 탑재해놓아 실시간으로 참조할 수 있도록 기능을 추가할 예정이다. 그리고 NICE

Tool은 기본적으로 데이터베이스 시스템으로 통계분석에 필요한 자료를 관리하고 추출하는 것이 주목적이다. 자료가 추출된 후 필요한 통계분석은 통계 소프트웨어를 이용하여 수행하는 것이 맞을 것이다. 그러나 추출된 자료의 간단한 분포를 파악하는 기능으로, 기초통계량을 계산하고 도수분포표나 간단한 막대그래프 등을 작성할 수 있는 기능을 추가하고자 한다.

References

Kim, T., Lee, H., Ahn, S., Kwon, O.-K., Bang, J. S., Hwang, G., Kim, J. E., Kang, H.-S., Son, Y.-J., Cho, W.-S., and Oh, C. W. (2016). Incidence and risk factors of intracranial aneurysm: a national cohort study in Korea, *International Journal of Stroke*, **11**, 917–927.

Ko, M. J., Park, C. M., Kim, Y. J., Kang, S. H., and Park, D. W. (2015). Clinical application and potential effects of 2014 hypertension guidelines on incident cardiovascular events, *American Heart Journal*, **170**, 1042–1049.

Lee, J. and Kim, K. (2012). *Construction of an Appropriate Sampling Design and a Sample Database Using the National Health Information Database*, Research Report, National Health Insurance Service.

Lee, J., Lee, J. S., Park, S. H., Shin, S. A., and Kim, K. (2016). Cohort profile: the national health insurance service-national sample cohort (NHIS-NSC), South Korea, *International Journal of Epidemiology*, Available from: 10.1093/ije/dyv319

Naramore, E. (2015). *Beginning PHP5, Apache, and MySQL Web Development*, Wiley Publishing, Indianapolis.

NHIS (2016a). *User's Manual for the Sample Research DB Ver4.0*, Big Data Steering Department, National Health Insurance Service.

NHIS (2016b). *Training Book for Utilizing the Sample Research DB*, National Health Insurance Service.

Park, J., Suh, B., Shin, D. W., Hong, J. H., and Ahn, H. (2015). Changing patterns of primary treatment in Korean men with prostate cancer over 10 years: a nationwide population based study, *Cancer Research and Treatment*, **48**, 899–906.

Rim, T. H., Kim, D. W., Han, J. S., and Chung, E. J. (2015). Retinal vein occlusion and the risk of stroke development: a 9-year nationwide population-based study, *Ophthalmology*, **122**, 1187–1194.

Rim, T. H., Oh, J., Kang, S. M., and Kim, S. S. (2016). Association between retinal vein occlusion and risk of heart failure: a 12-year nationwide cohort study, *International Journal of Cardiology*, **217**, 122–127.

Welling, L. and Thomson, L. (2016). *PHP and MySQL Web Development* (5th ed), Addison & Wesley, Indianapolis.

Yu, S., Wee, J., Kim, J. W., and Yoon S. (2015). Methodology for big data analysis using data from national health insurance service: preliminary methodologic study and review about the relationship between sinus surgery and asthma, *Journal of Rhinology*, **22**, 28–33.

국민건강보험공단의 표본연구DB를 위한 비주얼 쿼리 데이터베이스 시스템 개발 연구

조상훈[a] · 김희찬[b] · 강근석[a,1]

[a]숭실대학교 정보통계보험수리학과, [b]숭실대학교 대학원 융합 소프트웨어학과

(2016년 10월 27일 접수, 2016년 12월 12일 수정, 2016년 12월 12일 채택)

요 약

국민건강보험공단에서 제공하는 표본코호트DB는 보건의료계뿐만 아니라 통계학 연구를 위한 중요한 자원이다. 일반적으로 이들 자료에서 연구에 필요한 정보를 얻기 위하여 관련 사례들을 추출하는 과정에는 많은 시간과 노력이 들게 된다. 본 논문에서는 표본코호트DB를 이용하고자 할 때 사례 추출과정에 도움을 주는 데이터베이스 시스템인 **N**ational Health **I**nsurance Service **C**ohort DB **E**xtract **Tool**(NICE Tool)을 소개한다. SAS의 DATA 명령문이나 SQL문에 익숙하지 않은 연구자들도 쉽게 마우스 클릭만으로 DB에서 필요한 변수들과 조건에 맞는 사례들을 추출할 수 있는 기능을 제공한다. 이 시스템을 활용하면 빠른 사례추출이 가능하여 표본코호트DB를 사용한 연구들이 더욱 활성화되리라 판단된다.

주요용어: 비주얼 쿼리, 데이터베이스 시스템, 표본코호트, 국민건강보험공단

이 연구는 연세대학교 의과대학 2016년도 정책과제연구비(6-2016-0058) 지원에 의하여 이루어졌음.

[1]교신저자: (06978) 서울시 동작구 상도로 369, 숭실대학교 정보통계보험수리학과. E-mail: gskang@ssu.ac.kr