퇴원손상심층조사 자료의 효율적 활용을 위한 데이터베이스 구축에 대한 연구

김성수

청주대학교 보건행정 · 헬스케어학부 교수

A Study on Database Construction for Efficient Utilization of Korean National Hospital Discharge In-depth Injury Survey Data

Sung-Soo Kim

Professor, Department of Health Administration & Healthcare, Cheongju University, Cheongju, Korea

Objectives: The present study is focused on extracting various data such as diseases, surgical operations and treatments of the Korean National Hospital Discharge In-depth Injury Survey Data to construct an effective database that can be widely used in development of policies and research. **Methods:** The database construction methodology was applied to construct an effective database using the Korean National Hospital Discharge In-depth Injury Survey Data from 2006 to 2016. And for the data integration, the items and properties of data for each year were checked. As for the software to construct the database, the Relational Database Management System (RDBMS) MySQL was used. And through normalization, the Entity Relationship Diagram was designed to record the DDL and DML to conduct physical modeling. **Results:** As a result of checking the collected items per year, there were no problems in constructing the database. There were total of 68 items and 2,253,001 cases of sample patients. Through the normalization process, they were divided into 3 different entities and each entity was issued a primary key consisting of the year and patient number. Finally, the integrated database construction was completed by compiling the SQL for physical modeling. **Conclusions:** The integrated database model of the Korean National Hospital Discharge In-depth Injury Survey Data was produced. In the future, its usage will increase as it enables effective data extraction for various researches related to diseases, surgical operations, and treatments. Furthermore, it could also be used in constructing the web-based supplying system for the Korean National Hospital Discharge In-depth Injury Survey Data.

Key words: Database construction, DBMS (Database Management System), KNHDS (Korean National Hospital Discharge In-depth Injury Survey), MySQL

서 론

정보기술이 발전하면서 각 산업분야의 빅데이터 이용이 활성화되고 있다. 정부차원에서도 공공데이터포털 사이트를 통해서 국가 중점 데이터를 공개하고 활용할 수 있도록 하거나 각 공공기관에 사용 신청서를 제출하면 사용을 허가하고 있는 실정이다[1]. 이러한 빅데이터는

정보보호차원에서 우려와 부정적 시각도 존재하지만 다각적이고 효율적인 국가 정책개발의 필요성이 강조되고 있는 4차 산업혁명 시대를 맞아 새로운 가치창출이라는 긍정적 평가가 높아지면서 활용 수준이 점차 높아지고 있다[2,3].

보건으로 분야에서도 빅데이터 제공이 활발하다. 건강보험심사평가원의 보건의료빅데이터 개방시스템은 주요 의료통계, 국가승인통계, 질

Corresponding author: Sung-Soo Kim

298 Daeseong-ro, Cheongwon-gu, Cheongju 28503, Korea.
Tel: +82-43-229-7998, E-mail: mra7033@naver.com

Received: January 12, 2020 Revised: February 13, 2020 Accepted: February 21, 2020

*This work was supported by the research grant of Cheongju University (2018.03.01.-2020.02.29.)

No potential conflict of interest relevant to this article was reported.

How to cite this article:

Kim SS. A study on database construction for efficient utilization of Korean National Hospital Discharge In-depth Injury Survey Data. J Health Info Stat 2020;45(1):44-51. Doi: <https://doi.org/10.21032/jhis.2020.45.1.44>

© It is identical to the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permit unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2020 Journal of Health Informatics and Statistics

병·행위별의료통계, 의약품 통계, 의료자원 통계 등과 같이 정형화된 자료뿐 아니라 사용자 맞춤형태의 진료정보 및 의료자원과 같은 데이터를 제공하고 있다. 이 같은 자료는 건강보험심사평가원의 업무과정에서 발생하는 방대한 데이터이기 때문에 서울, 원주 등 전국 10곳의 빅데이터센터에 방문하여 자료를 분석하거나, 원하는 장소에서 원격 접속하여 분석할 수 있도록 지원하고 있다[4].

보건의료빅데이터개방시스템처럼 자료가 수시로 발생하고 수집되는 범위가 넓은 대용량의 빅데이터는 원활한 자료제공과 이용을 위해 오픈 API (Open Application Programming Interface) 기반으로 운영하는 반면, 국민건강영양조사, 퇴원손상심층조사, 한국의료패널조사와 같은 대표적인 보건의료 공공데이터는 조사연도별로 자료를 제공하고 있는 실정이다[5,6]. 이들 자료를 이용하는 연구자들은 점차 자료가 축적되면서 특정 건강지표나 질병, 치료 등과 같은 주제들에 대하여 장기적인 추세를 분석할 필요성을 느낄 수 있을 것이다. 다년도 자료를 이용하여 시계열이나 횡단면적 연구를 하기 위해서는 연구자가 직접 연도별로 제공된 자료를 병합하여 하나의 통합된 파일로 관리해야 하는 어려움이 발생한다. 또한 연구주제의 예비타당성을 확인하기 위한 기초조사를 플랫폼이나 웹기반에서 실시간 분석이 불가능하다. 관계형 데이터베이스 개념 구조상 정규화의 필요성이 없는 공공데이터의 경우 연구에 필요한 해당 변수들을 순서대로 병합하여 분석하는 정도는 큰 어려움이 없으나 퇴원손상심층조사 자료처럼 질병과 수술처치 자료를 정규화(normalization)하여 여러 개체(entity)로 나누어야 할 경우에는 별도의 데이터베이스시스템을 구축해야 하는 과정을 거쳐야 한다.

퇴원손상심층조사는 2005년부터 매년 20만 건 이상의 자료가 수집되고 있는데, 전국의 100병상 이상 일반 의료기관을 대상으로 표본환자의 질병 및 수술, 처치 데이터 등을 수집한 자료이다. 건강보험환자만을 대상으로 하는 보건의료빅데이터개방시스템, 국민건강보험자료공유서비스와는 달리 산재환자나 자동차보험, 건강보험을 적용 받지 않는 일반환자들도 포함하고 있으며 의료진이 작성한 의무기록을 기반으로 하기 때문에 조사대상의 주관이 개입되지 않는 장점이 특징이다. 이러한 장점에도 불구하고 앞서 기술한 데이터베이스시스템을 구축해야만 동반질환이나 수술, 처치 등 다년도 연구가 가능하기 때문에 이용 활성화를 위해 적절한 데이터베이스 모델을 제안하는 것이 필요하다.

본 연구에서는 보건의료분야에서 다년간 축적된 퇴원손상심층조사 자료의 질병, 수술, 처치 등 다양한 데이터를 손쉽게 추출하고 정책 개발과 연구에 폭넓게 활용할 수 있는 방안을 모색하기 위해 실시하였다. 관계형 데이터베이스를 근간으로 데이터의 구조를 분석하고 여러 개체로 분리하는 정규화 과정을 거쳐 조사연도별로 제공된 자료를 하나의 통합 데이터베이스로 구축하는 방안을 제안하고자 한다.

연구 방법

연구자료

본 연구는 2005년부터 매년 조사사업이 진행되고 있는 퇴원손상심층조사 자료를 이용하였다. 표본환자의 퇴원일을 기준으로 100병상 이상의 의료기관을 대상으로 1차 추출 단위를 결정하고, 추출된 표본 의료기관의 퇴원환자를 2차 추출하는 층화집락 2단추출 표본설계에 따라 자료가 수집된다. 1차 추출 단위인 의료기관의 병상규모에 따라서 Neyman배분을 실시하고 층별 표본크기는 비례배분법을 적용하여 전국의 시도를 대표하고 지역별 균형을 이루도록 하고 있다. 2005년도 첫째 조사에서는 한국표준질병사인분류(KCD-4)를 기준으로 부진단코드 7개까지만 수집하였고, 부수술 및 처치코드는 국제의료행위분류(ICD-9-CM)를 사용하여 5개가 수집되었다. 이후 2차 조사에서는 부진단코드와 부수술 및 처치코드 모두 20개까지 수집하고 있으며, 치료결과, 퇴원 후 항방, 손상발생장소, 손상 시 활동, 손상기전, 자살위험요인 등 1차 조사에 포함되지 않았거나 부분 수정이 있던 조사항목이 2차부터 안정화되기 시작하였다[7]. 따라서 본 연구에서는 조사항목의 표준화가 적용된 2006년 2차 조사자료부터 2016년까지 11년간 2,253,001명의 표본환자 자료를 대상으로 한다. 질병관리본부의 퇴원손상심층조사 자료 담당자에게 연구자료 사용신청서를 작성하여 제출하고 승인 후 자료를 획득하였다.

데이터베이스 설계과정

퇴원손상심층조사 자료의 효율적 활용을 위한 데이터베이스 구축 설계과정은 다음과 같다(Figure 1). 데이터베이스를 구축하기 위해서는 해당 자료의 구조를 먼저 살펴보는 것이 중요하다. 조사항목에 대해서 조사연도별 기준을 확인하여 하나의 데이터베이스로 통합이 가능

Figure 1. Data modeling process. RDBMS, relational database management system; ERD, entity relationship diagram.

한지를 확인하고 별도의 개체나 테이블로 분리해야 할 것이 있는지 검토하여 정규화를 진행할 수 있다. 다음으로 데이터베이스 구축을 위한 논리적 모델링을 실시하여야 하는데, 본 연구에서는 Database Management System (DBMS)으로 MySQL을 사용하고자 한다. 퇴원심층조사자료가 매년 수집되고 현재 200만 건 이상의 자료가 쌓이는 등 양적 증가가 계속되고 있다. 10년 미만의 누적자료는 그동안 개인용 DBMS인 마이크로소프트사의 ACCESS를 활용할 수 있었으나 자료의 크기가 2GB(gigabyte)를 넘으면 데이터를 추가 저장하거나 질의어를 실행하는데 제한이 따른다[8]. 이같은 이유로 Oracle, Infomix, Sybase 등 여러 서버급 DBMS가 있으나 무료로 사용이 가능한 공개 DBMS인 MySQL이 적합하다고 판단하였다[9]. MySQL은 Central Process Unit (CPU)를 효율적으로 사용할 수 있도록 Multithread를 지원하여 속도가 빠르고 웹기반 등 다양한 플랫폼을 지원한다[10]. 다음으로 정규화 및 논리적 모델링의 결과로 개체관계도인 Entity Relationship Diagram (ERD)를 작성하여 물리적 모델링을 위한 준비과정을 수행할 것이다. ERD는 관계형데이터베이스시스템에서 데이터가 물리적 저장에 어떻게 이루어지는지 도식적인 설계도 역할을 한다. 관련된 데이터들의 속성(attribute)들과 레코드들의 집합을 의미하는 것이 개체이며, 여러 개체들 사이의 관계(1:1, 1:N)를 정의하여 데이터 추출을 위한 효과적인 Structured Query Language (SQL)을 수행할 수 있도록 할 것이다. 마지막으로 ERD를 기반으로 Data Definition Language (DDL)을 작성하여 퇴원손상심층조사 DBMS를 구축하고자 한다.

연구 결과

연도별 퇴원손상심층조사 자료분석

데이터베이스를 구축하기 위해서 연도별 퇴원손상심층조사 자료의

Table 1. Korean National Hospital Discharge In-depth Injury Survey sample size by year

Survey year	Sample size	Diagnostic classification system	Procedure classification system
2006	161,997	KCD-5th	ICD-9-CM vol. III
2007	177,011		
2008	177,619		
2009	190,074	KCD-5th	
2010	198,586		
2011	220,838		
2012	225,105	KCD-6th	
2013	235,526		
2014	214,569		
2015	224,061	KCD-7th	
2016	227,615		
Total	2,253,001		

표본크기와 진단코드, 수술 및 처치코드의 표준을 확인하였다. 2006년 표본환자수는 161,997명이며 2016년에는 227,615명까지 증가하여 전체 11년간 총 2,253,001명의 자료가 수집되었다. 한국표준질병사인분류 업데이트에 따라서 2011년 까지는 5차 개정판이 적용되었고 2015년까지는 6차판, 2016년부터는 7차 개정판이 기준이다. 수술 및 처치코드는 모든 조사년도에 대해 ICD-9-CM vol. III를 기반으로 조사되었다 (Table 1).

퇴원손상심층조사 자료의 구조

퇴원손상심층조사 자료의 구조를 살펴본 결과 2006년도 조사자료부터는 조사항목에 차이가 없는 것으로 나타났다. 따라서 연도별 자료

Table 2. Data structure of Korean National Hospital Discharge In-depth Injury Survey

Group	Physical name	Logical name
Identifier	Num	Patient identifier
Hospital information	Hident	Hospital identifier
	Area	Hospital area
	Bed	Bed size
Socio-demographic information of inpatient	Sex	Patient sex
	Age	Patient age in years
	Bdate	Date of birth of discharged patient
	Zip	ZIP code for patient's residence
	Paysou	Medical bill payment source
Hospital visit information	Admdate	Date of admission
	Disdate	Date of discharge
	Admrout	Admission route
Disease and treatment information	Pridia	Principal diagnosis*
	Adddia1-adddia20	Additional diagnosis code 1-20*
	Injdia1	External cause code of injury 1
	Injdia2	External cause code of injury 2
	Pripro	Principal procedure*
	Prodate	Date of procedure
	Addpro1-addpro20	Additional procedure code 1-20*
	Txres	Discharge status
	Dispo	Disposition
	Deathca	Underlying cause of death
Information of injury patient	Injint	Intent of injury
	Injpla	Place of injury occurrence
	Injact	Activity at injury
	Injmec	Mechanism of injury
	Injdate	Date of injury
	Tamode	Mode of transport
	Suifac	Risk factors of suicide or suicide attempt
	Poissub	Poisoning substance
weight	Weight	Weight

*Need normalization.

를 통합하는데 문제가 되지 않아 별도의 조정과정을 거칠 필요가 없었다. 조사항목은 총 7개의 그룹으로 나눌 수 있다(Table 2). 표본환자의 표식자 1개 항목, 의료기관정보에 해당하는 표식자, 지역, 병상규모 3개 항목, 표본환자의 사회인구학적 정보로는 성별, 나이, 생년월일, 우편번호, 보험자격 5개 항목, 의료기관 방문정보로 입원일자, 퇴원일자, 입원경로 3개 항목, 질병과 치료과련 정보는 주진단, 부진단(최대 20개), 상해외인코드 1, 2, 주수술, 부수술 및 처치(최대 20개), 수술일자, 퇴원상태, 퇴원 후 향방, 선행사인 48개 항목, 손상환자정보는 손상의 도성, 손상발생장소, 손상시 활동, 손상기전, 손상발생일, 운수사고유형, 자살위험요인, 중독물질 8개 항목, 마지막으로 전국규모 추정을 위해 가중치 1개 항목이다. 총 조사항목은 68개 이다. 주진단과 20개의 부진단, 주수술, 부수술 20개가 각각의 속성으로 표현되어 열(column)로 배열되어 있어 진단이나 수술 및 처치에 대해 검색 시 효과적이지 못하다. 따라서 이들의 정규화가 필요하다. 즉, 진단코드와 수술 및 처

치코드를 각각 별도의 개체로 분리하는 2차 정규화 과정이 필요한 대상이다.

정규화

퇴원손상심층조사 자료의 정규화를 실시한 결과 아래와 같다(Figure 2). 정규화 결과 총 3개의 테이블로 구성되었다. Table 2에서 주진단 pridia, 부진단 adddia1-20을 diag_tb 테이블로 분리하고, 주수술 pripro, 부수술 및 처치 addpro1-20을 proc_tb 테이블로 분리하였다.

정규화는 데이터의 중복을 최소화하고 빠른 데이터 프로세싱이 가능하도록 데이터베이스의 설계과정에서 여러 개의 테이블로 나누는 작업이다. 1차 정규화는 각 테이블의 필드(field)가 더 이상 분해할 수 없을 때까지 원자값(atomic value)을 갖도록 분리하는 것을 말한다. 검토한 결과 더 이상 분해할 수 있는 필드가 존재하지 않기 때문에 1차 정규화를 실행할 부분은 없었다. 환자의 거주지 우편번호에 해당하는

Figure 2. Entity relationship diagram for logical datamodeling.


```

1 • create table nhds_tb
2 (
3     YEAR      int not null,
4     NUM       int not null,
5             primary key(YEAR, NUM),
6     HIDENT    varchar(10),
7     AREA      int,
8     BED       int,
9     SEX       int,
10    AGE       int,
11    BDATE     varchar(8),
12    ZIP       varchar(6),
13    PAYSOU    int,
14    ADMDATE   varchar(8),
15    DISDATE   varchar(8),
16    ADMROU    int,
17    INJDIA1   varchar(9),
18    INJDIA2   varchar(9),
19    TXRES     int,
20    DISPO     int,
21    DEATHCA   varchar(10),
22    INJINT    int,
23    INJPLA    int,
24    INJACT    int,
25    INJMEC    int,
26    INJDATE   varchar(8),
27    TAMODE    int,
28    SUIFAC    int,
29    POISUB    int,
30    WEIGHT    decimal(20,15)
31 );

```

(A)

```

1 • create table diag_tb
2 (
3     year      int(4),
4     num       int(8),
5             primary key(YEAR, NUM),
6     flag      int(1),
7     serl      int(2),
8     code      varchar(10)
9 );

```

(B)

```

1 • create table proc_tb
2 (
3     year      int(4),
4     num       int(8),
5             primary key(YEAR, NUM),
6     prodate   varchar(8),
7     flag      int(1),
8     serl      int(2),
9     code      varchar(10)
10 );

```

(C)

Figure 3. Data definition language for table creation. (A) Nhds_thtable, (B) Diag_thtable, (C) Proc_thtable.

zip 필드가 과거 3자리씩 나눌 수 있었으나 최근 도로명 주소표기로 변경되면서 5자리로 통합되었기 때문에 1차 정규화를 시행할 필요가 없다고 판단하였다. 기본키(primary key)는 각 테이블에서 유일한 하나의 레코드를 검색할 수 있는 표식자 역할을 한다. 표본환자의 표식자 역할을 하는 num 필드는 조사년도마다 serial number로 생성되기 때문에 연도별 자료를 통합할 경우 표식자 역할을 하는 기본키 역할을 할 수가 없다. 따라서 year라는 필드를 모든 테이블에 생성한 후 year, num 2개 필드를 각각의 테이블에 기본키로 지정하여야만 세 개의 테이블에 대한 릴레이션(relation)이 가능해진다. 주진단과 부진단을 하나의 테이블로 통합하여 분리하였기 때문에 diag_tb 테이블에서 특정 표본환자의 질병코드인 code 필드가 주진단인지 부진단인지 구분할 필드가 필요하다. 따라서 flag 필드를 추가적으로 생성하였다. 주수술과 부수술 및 처치도 같은 방식으로 flag 필드를 생성하였다. seri 필드는 diag_tb, proc_tb 테이블에서 code의 입력된 순서를 표기하기 위해 추가 생성한 필드이다. seri는 표본환자에 따라서 최대 20까지 serial number가 발생할 수 있다.

데이터베이스 생성

앞서 정규화 과정을 통해서 논리적 모델의 설계가 완성됨에 따라 이를 바탕으로 데이터베이스 언어인 DDL을 이용하여 데이터의 물리적

```

1 # totl_tb의 주진단, 부진단을 diag_tb에 옮기기
2 • insert into diag_tb (year, num, flag, serl, code)
3     select year,
4           num,
5           1 flag,
6           1 serl,
7           code
8     from totl_tb
9     where pridia is not null;
10
11 # totl_tb의 주수술, 부수술을 proc_tb에 옮기기
12 • insert into proc_tb (year, num, prodate, flag, serl, code)
13     select year,
14           num,
15           prodate,
16           1 flag,
17           1 serl,
18           code
19     from totl_tb
20     where pripro is not null;

```

Figure 4. Data manipulation language for data transportation.

모델링 과정을 거쳐야 한다. 본 연구에서는 퇴원손상심층조사 자료를 활용할 여러 연구자를 위하여 SQL기반으로 작성된 DDL 소스코드를 공유하고자 한다.

먼저 <https://dev.mysql.com/downloads> 접속하여 MySQL 프로그램(MySQL Installer for Windows)과 MySQL for Excel을 다운로드하여 각각 설치한다[9]. 질병관리본부의 퇴원손상심층조사 자료를 요청하여 연도별로 Excel 파일로 변환한 다음 Excel의 데이터 메뉴에서 우측

Result Grid

Filter Rows:

Export:

Wrap Cell Content:

Fetch rows:

	year	num	hidnt	sex	age	bdate	zip	paysou	admdate	disdate	admrou
	2006	1	A1	2	30	19750512	133821	1	20060228	20060308	2
	2006	2	A1	2	83	19220625	412702	1	20060320	20060327	2
	2006	3	A1	2	61	19440801	110811	1	20060612	20060627	2
	2006	4	A1	1	24	19820218	143822	2	20061128	20061205	1
	2006	5	A1	2	80	19250915	100450	1	20060306	20060316	1
	2006	6	A1	1	76	19291222	595892	1	20061030	20061107	1
	2006	7	A1	2	63	19421120	110542	2	20060224	20060314	1
	2006	8	A1	1	58	19480511	142762	3	20060613	20060623	1
	2006	9	A1	1	65	19401113	152050	4	20060630	20061031	2
	2006	10	A1	1	51	19550428	136041	2	20060615	20060626	2

Result Grid

Filter Rows:

	year	num	flag	serl	code
	2005	1	1	1	I638
	2005	1	0	2	I10
	2005	1	0	3	E119
	2005	1	0	4	I083
	2005	1	0	5	I493
	2005	1	0	6	E786
	2005	1	0	7	L82
	2005	2	1	1	Z470
	2005	2	0	2	K760

Result Grid

Filter Rows:

	year	num	prodate	flag	serl	code
	2006	3	NULL	0	2	5187
	2006	3	NULL	0	3	2309
	2006	4	NULL	0	2	3995
	2006	10	NULL	0	2	3995
	2006	11	20061124	1	1	194
	2006	11	20061124	0	2	2049
	2006	12	20060816	1	1	3228
	2006	21	20060801	1	1	2188
	2006	21	20060801	0	2	2169

Figure 5. Final view of database construction in MySQL. (A) Nhds_tb table, (B) Diag_tb table, (C) Proc_tb table.

MySQL for Excel을 이용하여 MySQL 프로그램에 데이터를 전송할 수 있는데, 이를 시행하기 전에 Figure 3과 같은 데이터 정의를 활용하여 테이블을 각각 생성한다.

Excel의 MySQL for Excel을 이용하여 MySQL로 연도별 자료를 반복하여 total_tb 테이블로 전송하였다고 하면 Figure 4와 같은 데이터 조작어를 이용하여 진단 테이블과 수술 및 처치 테이블에 데이터를 저장할 수 있다. 진단정보가 축적되는 diag_tb 테이블에는 퇴원손상심층조사 자료 항목에는 없는 flag 필드를 추가하였다. flag는 진단코드가 1이면 주진단을 뜻하고 0이면 부가진단을 뜻한다. 같은 방법으로 수술정보인 proc_tb 테이블에도 flag 필드를 추가하였다. 부진단과 부수술이 20개까지 있는 표본환자들이 있기 때문에 Figure 4의 SQL을 반복하되 serl의 앞 숫자를 1씩 늘려주고, 주진단과 주수술을 제외하고는 flag 앞 숫자를 0으로 변경하여 실행하면 된다. 최종 구축된 퇴원손상심층조사 자료의 데이터베이스는 Figure 5와 같다.

고찰 및 결론

본 연구에서 구축한 데이터베이스는 퇴원손상심층조사 자료를 보다 효과적으로 활용하는데 목표를 두고 개발하였다. 공공데이터의 개방정책 개선에 관한 Yoon and Hyun [11]의 연구에 의하면 2019년 2월을 기준으로 Excel, 한글, PDF 형태 등의 파일데이터로 자료를 제공하

는 비율이 88.9%로 나타났다. 파일데이터는 같은 주제이지만 기간이 다른 자료를 합치거나 가공하는 것이 불편하기 때문에 이용자 편의성이 높은 자료제공 방식을 고려할 필요가 있다고 하였다. 퇴원손상심층조사도 현재까지 Statistical Package for the Social Sciences (SPSS)나 Statistical Analysis System (SAS)와 같은 통계분석 프로그램의 파일형태로 제공되고 있다. 10년 이상의 대용량 자료가 누적되면서 동반질환을 분석하기 위한 연구나, 시계열 분석, 종단적 연구를 시행하기에는 자료의 통합과 분석에 어려움이 있었다. 본 연구는 선행연구의 이러한 개선점을 충족할 수 있도록 데이터베이스 구축 방법론에 따라 퇴원손상심층조사 자료의 데이터베이스 모델을 제시하였다.

연구결과 표본환자의 기본정보를 포함하는 nhds_tb, 진단 정보를 나타내는 diag_tb, 수술 및 처치정보를 나타내는 proc_tb entity로 분리하여 논리적 모델을 설계하였고, 물리적 저장을 위한 DDM 및 DML SQL을 작성하였다. 의료기관에 입원한 환자들은 주진단 외에도 다양한 동반질환을 진단받고 수술이나 처치가 다발적으로 이루어진다. 따라서 파일 기반의 여러 속성으로 진단과 수술 및 처치정보를 하나의 테이블로 제공하면 동반질환이나 동반수술의 검색에 제한을 받게 된다. 본 연구를 통해 이러한 문제를 해결할 수 있는 데이터베이스를 제시함으로써 퇴원손상심층조사 자료의 이용 활성화에 기여할 수 있을 것이다.

최근 대용량 자료의 고성능 처리를 위한 Not only SQL (NoSQL)

DBMS가 각광받고 있다. NoSQL은 Facebook, Instagram, Twitter 등 Social Network Service (SNS)로 처리할 수 있는 분산 저장 시스템이다. Open source형태로 대부분 제공되며 확장성, 가용성이 높은 성능을 자랑한다. 윈도우를 포함한 유닉스, 리눅스 등 다양한 운영체제를 지원하고 설치와 사용이 용이하며 보안 성능이 우수하다[12,13]. 데이터베이스를 구축하고자 한다면 자료의 특성에 맞는 DBMS를 적절히 선택하여야 효율적인 자료처리가 가능하다. 퇴원손상심층조사 자료는 SNS서비스에 비해 대용량 자료가 아니며, 데이터 형태도 수집항목이 정형화 되어 있는 특성을 고려하여 1970년대부터 널리 사용되어 일관성과 신뢰성을 갖춘 RDBMS를 선택하였다. 본 연구에서 선택한 DBMS는 RDBMS의 일종인 MySQL이다. MySQL은 Oracle에서 개발한 서버급 Open source DBMS로 표준화된 SQL을 사용하여 자료검색과 처리가 용이하고 다양한 프로그램 언어를 지원하기 위한 API를 제공하고 있다[14]. 본 연구는 앞서 기술한 기존 공공데이터의 자료제공 형태를 극복하고 데이터 특성에 맞는 관계형 데이터베이스를 근간으로 MySQL을 활용하여 구축한 바, 향후 이를 바탕으로 연구자가 개별적으로 데이터베이스를 효과적으로 구축할 수 있고 나아가 퇴원손상심층조사 자료를 이용한 다양한 연구가 활성화 되는데 기여할 수 있을 것이다.

경제협력개발기구(Organization for Economic Cooperation and Development, OECD)가 발간하는 정부백서에 따르면 데이터개방과 활용을 위한 가용성, 접근성, 정부지원의 3개 분야를 측정한 공공데이터 개방지수에서 우리나라는 2015년과 2017년 1위를 차지하였다. 데이터 제공 건수는 2013년 5,272건에서 2018년 25,184건으로 5배 가까이 증가하는 등 양적 성장을 이루었으나 국내 빅데이터 이용률은 7.5%로 낮은 수준을 보이고 있다[11]. 미국 오픈 데이터 기업을 대상으로 네트워크 분석을 실시한 결과 보건의료에 대한 높은 관심을 갖는 것으로 나타나 향후 이용이 활성화될 것으로 전망된다[15]. 이러한 문제를 해결하기 위해 빅데이터의 자료수집과 분석 활용을 늘리기 위한 방법으로 오픈 API 방식의 서비스가 각광받고 있다. Lee and Kim [16]은 의료정보서비스의 접근성 제고와 건강정보의 수집, 관리 및 활용에 대한 서비스를 활성화 하는데 오픈 API 서비스가 가능할 것이라고 제시하였다. 본 연구는 진단, 수술 및 처치 데이터를 별도의 entity로 분리한 퇴원손상심층조사 자료의 데이터베이스 구축방안을 제시하였으므로 이를 바탕으로 질병관리본부 주도로 오픈 API 방식의 자료수집 및 추출 플랫폼을 개발하는데 활용될 수 있을 것이다. 향후 연구에서는 표본환자의 거주지역별 분석이 용이하도록 우편번호 entity를 추가하여 데이터베이스를 구축하는 것이 필요하고 한국표준질병사인분류의 중분류, 대분류 체계를 포함하여 확장이 가능하다. 수술 및 처치코드 또한 해부학적 분류기준에 따른 entity를 추가할 수 있을 것이다. 향후 퇴원손상심층조사 자료뿐만 아니라 다양한 공공데이터의 활발한 이용을 위해

연구자의 편의성을 증진시킬 수 있는 데이터베이스 구축 관련 연구가 활성화 되기를 기대한다.

ORCID

Sung-Soo Kim <https://orcid.org/0000-0002-7205-0758>

REFERENCES

1. Lim JW, Choi GH. The influence of open data policies on public innovation. J Korean Inst Ind Eng 2017;43(1):19-29 (Korean). DOI: 10.7232/JKIE.2017.43.1.019
2. Kim KY, Kwon HY. Tasks for improvements of the personal information protection legislation for using the public data. J Law Econ Regulat 2014;7(2):22-37 (Korean).
3. Song SH, Lee SY, Shin Y, Lee JY. A study on the effectiveness of Korea's open government data policy: ecosystem perspective. J Korean Assoc Reg Inf Soc 2017;20(4):1-34 (Korean).
4. Lee EH, Kim HS, Lee BH, Choi HJ, Kim JA. Utilization of open government data for health policy. Suwon: Gyeonggi Research Institute; 2016 (Korean).
5. Cho SY, Park K. Trends in metabolic risk factors among patients with diabetes mellitus according to income levels: the Korea National Health and Nutrition Examination Surveys 1998~2014. J Nutr Health 2019; 52(2):206-216 (Korean). DOI: 10.4163/jnh.2019.52.2.206
6. Hwang JS, Shin SJ, Kim JH, Oh SH, Kang HK, Park SH, et al. Domestic secondary data resources utilization in healthcare research. Seoul: National Evidence-based Healthcare Collaborating Agency; 2013 (Korean).
7. Korea Centers for Disease Control and Prevention. Korean National Hospital Discharge In-depth Injury Survey Guidebook. Cheongju: Korea Centers for Disease Control and Prevention; 2010 (Korean).
8. Access specifications. Available at <https://support.office.com/en-us/article/access-specifications-0cf3c66f-9cf2-4e32-9568-98c1025bb47c> [accessed on December 23, 2019].
9. MySQL Community Downloads. Available at <https://dev.mysql.com/downloads/installer/> [accessed on December 26, 2019].
10. Sung HI, Kim SC, Yom IS, Kim BG, Ahn YS, Nam HW, et al. Construction of astronomical database based on mysql and jsp. Pub Korean Astron Soc 2004;19(1):109-119 (Korean). DOI: 10.5303/PKAS.2004.

- 19.1.109
11. Yoon SO, Hyun JW. A study on the actual condition analysis and improvement of public data opening policy: focusing on the case of national data opening of public data portal. *Korean Public Manag Rev* 2019;33(1):219-247 (Korean). DOI: 10.24210/kapm.2018.33.1.010
12. Kim HS, Lee BH. Interoperability between NoSQL and RDBMS via Auto-mapping scheme in distributed parallel processing environment. *J Korea Inst Inf Commun Eng* 2017;21(11):2067-2075 (Korean). DOI: 10.6109/jkiice.2017.21.11.2067
13. Mahajan D, Blakeney C, Zong Z. Improving the energy efficiency of relational and NoSQL databases via query optimizations. *Sustainable Computing: Informatics and Systems* 2019;22:120-133. DOI: 10.1016/j.suscom.2019.01.017
14. Jeon DC, Lee BM, Hwang HJ. Performance analysis of RDBMS and MongoDB through YCSB in medical data processing system based HL7 FHIR. *J Korea Multimedia Soc* 2018;21(8):934-941 (Korean). DOI: 10.9717/kmms.2018.21.8.934
15. Shin HK, Park HW. Ego network analysis of open data in United States healthcare companies. *J Korean Data Anal Soc* 2015;17(6):3107-3118 (Korean).
16. Lee HJ, Kim YH. Open platform for improvement of e-Health accessibility. *J Dig Contents Soc* 2017;18(7):1341-1346 (Korean). DOI: 10.9728/dcs.2017.18.7.1341