한국인 유전체역학 정보 DB 설계 및 구축

양은주°, 박용철, 고인송, 오범석, 김규찬

국립보건원 유전체연구소 유전체역학정보실

{ejyang°, ycpark, insong}@ngri.re.kr, {ohbs, k2kimm}@nih.go.kr

Design and Development of the Database for the Korean Genome and Health Study

Eunjoo Yang°, Yongcheol Park, InSong Koh, Bermseok Oh, Kuchan Kimm

Division of Epidemiology and Bioinformatics, National Genome Research Institute, KNIH

요 약

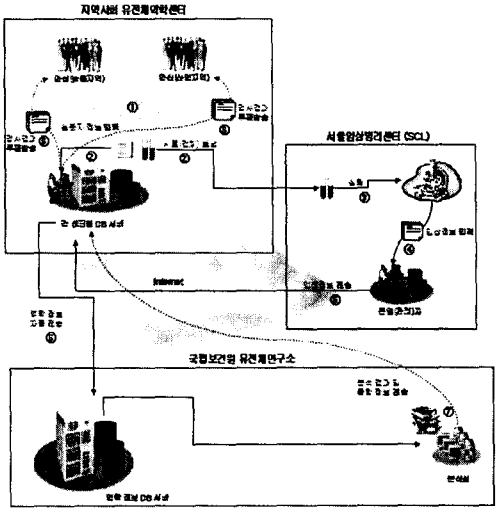
국립보건원 유전체연구소는 안산-안성 지역에 거주하는 45세 이상 69세 이하의 성인을 대상으로 고혈압, 당뇨, 골다공증, 천식, 비만 등 한국인의 총 국민의료비용에서 큰 부분을 차지하는 중요 만성질환에 초점을 맞추어 코호트 연구를 수행하고 있다. 이에 검진대상자의 설문 및 임상검사를 통하여 수집되는 개인식별 데이터, 생활습관 데이터 등의 설문정보와 다양한 임상검사정보에 대한 체계적 저장· 관리와 향후 수행될 대규모 정보 분석을 위해 유전체역학 정보 DB를 설계·구축하였다.

1. 서론

미국을 중심으로 한 인간유전체계획(Human Genome Project; HGP)이 예상보다 약 2년여 앞당겨 완결됨에 따라 그 결과를 활용하는 새로운 의과학 기술 개발 경쟁이 가열되고 있다. 개인 간 유전체 서열의 차이(유전염기서열 30억 염기 중 약 0.1 %)가 신체적 특성, 질병에 대한 감수성 등 개인 간 차이의 근본이므로 개인별 유전체형의 다양성 규명이 '유전체이후시대'(Postgenomic Era)의 핵심화두이며, 이는 개인간 뿐만 아니라 민족, 종족 간 특유한 유전형 차이가 존재함을 적시하고 있다. 이러한 배경으로 현재 국립보건원 유전체연구소 안산-안성 지역사회유전체역학센터에서는 한국인의 중요 만성질환 즉, 고혈압, 당뇨, 골다공증, 천식, 비만 등을 초점으로 역학정보(epidemiology information)를 수집하고 있다. 그외 국내의 경우, 서울대-서울 남성 코호트의 암연구, 연세대-강화 프로젝트의 고혈압 연구가 수행 중에 있다. 선진국의 경우, 영국 MRC-UK BioBank 협력, 미국 NIH-Coriell Repositories/DNA Sciences 협력, 아이슬란드의 deCODE 계획, 영국-Finland 국가간 협력연구 등 대규모로 종족별, 또는 질병별로 환자와 정상인으로부터 대규모의 유전체 및 임상정보를 확보하기 위한 노력이 진행되고 있으며, 이들의 역학·시료정보 연계를 통해 미국, 영국, 일본, 캐나다 및 중국이 참여하고 민족간 유전체형 다양성 분석을 목적으로 하는 국제협력연구(International Haplotype Map Project)가 진행 중에 있다. 미국 심혈관 질환의 risk factor를 규명하여 심혈관 질환 이환률을 감소시킨 것으로 유명한 Framingham Heart Study에서도 50여년 전부터 주민의 역학정보를 수집하고 있다. 그외에 영국(health survey), 일본(국립 순환기연구소의 Suita 연구), 캐나다(hearth health survey)에서는 질환의 발생률, 진행 정도, 사망률을 조사하기 위해 수십 년 전부터 대규모의 국가적 역학사업이 진행 중이다[4]. 그러나, 시공간적인(spatio-temporal) 속성을 지닌 이들 역학정보에 대한 체계적인 데이터 모델링 및 DB 설계에 대한 국외 논문을 찾기란 쉽지 않다. 전세계 computer science 관련 논문 검색 엔진인 DBLP(DataBase systems and Logic Programming)를 통해서도 역학정보 DB 설계 및 구축에 대한 논문은 1편도 없었다[1]. 단지, cancer cluster 분석을 위해 고안된 독일의 CARLOS(Cancer Registry Lower Saxony) 프로젝트에서 시행한 역학정보, 통계정보 통합 분석 시스템 구축 시 고려한 시간, 공간, 통계적 특성에 기반한 multidimensional data 분석 모델링 기법을 소개한 논문만 찾아볼 수 있었다[8, 9]. 또한, 모든 생물정보에 대한 논문 검색 시스템인 PubMed에서 조차 역학정보 DB 설계에 관한 체계적이고 명료한 논문을 찾아볼 수 없었다[7]. 이러한 이유는 몇 십 년 전에 이미 대규모 역학정보 DB 구축을 시도했던 선진국의 경우, 당시 데이터 중심(data-oriented)이 아닌, 업무 처리(process-oriented) 중심으로 DB를 구축 후 현재는 이를 통한 정보 분석 측면으로 초점이 맞춰져있기 때문이 아닌가 한다. 또한, 국내의 경우, 역학정보를 취급하는 연구자들이 대부분 의료정보 통계학자들로 이루어져 있기 때문에, 체계적인 정보 DB 설계 및 구축이 아닌, 정보 분석을 위한 데이터 파일 역할만 기대시하고 있기 때문이 아닌가 추정된다. 따라서, 향후 이 분야의 논문이 국내에서 계속 발표될 것으로 예상된다.

2. 업무 흐름도 및 DB 스키마 설계

2.1 업무 흐름도



[그림 1] 역학정보 흐름도

코호트 연구(cohort study)로부터 생산되는 검진대상자의 개인식별 데이터, 생활습관 데이터 등과 같은 설문정보와 임상검사정보를 총괄적으로 역학정보라 한다. 그림 1은 안산-안성 코호트 연구로부터 생산·수집되어 유전체연구소로 통합되기까지의 검진대상자에 대한 역학정보 및 업무 흐름도를 표현한 것이다.

2.2 DB 스키마 설계

국립보건원 유전체연구소로 통합되는 역학정보의 총 entity type 개수가 20개 이상이며, attribute의 개수가 2,000개 이상이기 때문에 본 논문에서는 E-R 다이어그램 대신, 이를 관계형 스키마로 매핑한 각각의 테이블에 대한 명세를 설명한다.

[표 1] 역학 정보 테이블 명세

테이블명	설 명	분류명
BREATH	호흡/순환기질환정보: 폐질환 진단여부/유해물질 노출정도/과거력/폐기능에 관한 설문정보	생활습관정보
DIET1	영양정보1: 식이습관정보	생활습관정보
DIET2	영양정보2: 식품섭취빈도조사정보	생활습관정보
DRSM	음주 및 흡연정보	생활습관정보
EMO	정서정보: 스트레스 및 긴장도 측정정보	생활습관정보
FAMD	가족력정보: 부모형제에 대한 과거병력정보	의료정보
FEMD	여성력정보: 여성병력정보	의료정보
GEN	일반정보: 종교/직업/결혼여부/좋아하는 색 등의 정보	일반정보

<표 계속>

테이블명	설 명	분류명
JOINT	관절염정보: 퇴행성 및 류마티스 관절염에 대한 증상정보	의료정보
MEDIC	의료기본정보: 사고및입원경험/틀니 및 치아상태 정보	의료정보
PASTD	과거병력정보: 약물복용상태/치료중인 질환정보	의료정보
PATIENT	검진자 기본정보: 검진센터/생년월일/성별/나이/검사일정보	일반정보
RLT1	신체측정정보: 혈압/키/몸무게/골밀도측정 결과정보	임상/검사정보
RLT2	신체측정정보: 심전도(EKG) 기록결과정보	임상/검사정보
RLT3_1 RLT3_2 RLT3_3	신체측정정보: 흉부 X-ray 결과정보	임상/검사정보
RLT3_4	신체측정정보: 손/무릎 X-ray 결과정보	임상/검사정보
RLT4	신체측정정보: 폐기능/Inbody 측정결과정보	임상/검사정보
SCL_RESULT	생화학적 검사정보: 혈액 및 뇨 검사결과정보	임상/검사정보
SLEEP	수면활동정보: 수면습관 및 장애요인 설문정보	생활습관정보
SPE_RECEIPT	NGRI내 시료 접수정보	시료정보
TYPEA	정서정보: TYPE-A 진단정보	생활습관정보
T_RES	영양정보: 식품섭취빈도조사의 영양소 계산정보	생활습관정보

각 테이블들은 안산-안성 유전체역학센터에서 검진한 검진자의 ID, name, 그리고 국립보건원 유전체연구소에서 정의한 EPI_ID로 서로 연계되어 있으며, 각 컬럼들은 PK(Primary Key), NN(Not Null) 제약 조건(constraint)으로 정의되어 있다.

3. DB 및 데이터 전송 스케줄러 구축

3.1 DB 구축

역학정보 DB 구축시 검진자의 생년월일, 이름, 주민등록번호와 같은 개인식별 데이터에 대한 보안을 위해 현재, Oracle DBMS 8i부터 지원하고 있는 DBMS_OBFUSCATION_TOOLKIT을 이용해 암호화된 형태로 데이터를 저장하였다. 이는 데이터를 암호화된 형태로 저장하기 위한 기존의 3rd party tool이나, application logic으로 구현하던 암호화 정책을 DB 차원에서 구현할 수 있도록 해준다.

[표 2] DB 구축 환경

구 분			설 명
안산-안성 유전체역학센터	server	모델명	HP workstation x1100 Base SPU
		OS	Linux Red Hat r. 7.1
	DBMS	모델명	Oracle DBMS Standard Edition 9i
		DB유형	Relational DB
	사용 목적		각 local DB 저장 및 관리
국립보건원 유전체 연구소	server	모델명	HP workstation xw4000
		OS	Linux Red Hat r. 7.2
	DBMS	모델명	Oracle DBMS Standard Edition 9i
		DB유형	Relational DB
	사용 목적		각 유전체역학센터로부터 전송되어온 역학정보 통합 저장 및 관리

3.2 데이터 전송 스케줄러

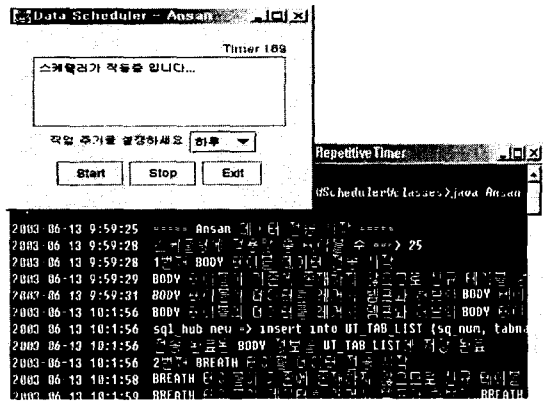
1) 기능

스케줄링 하고자 하는 안산-안성의 각 local DBMS에 접속 후 테이블 및 컬럼 정보를 추출해 오는 역할을 하는데, 처음 유전체연구소로 데이터가 전송되어 올 경우, 안산-안성의 local DB에 별도의 temp 테이블스페이스 및 데이터 화일(dbf)를 새로 생성해 복사본을 만든다. 그 후, local DBMS에 저장되어 있는 기존 테이블에 insert, delete와 같은 DML 명령이 수행되면, temp 데이터 화일에 있는 데이터와 변경된 데이터를 서로 비교한 후, 변경된 데이터만을 국립보건원 유전체연구소로 전송한다. 아래는 각 세부 기능들에 대해 요약한 것이다. 아래 그림 2는 안산 지역의 local DBMS내 데이터를 '하루' 주기로 스케줄링 해오는 작업을 표현한 것이다.

o selection 부분
- 작업 주기 설정: 하루, 일주일, 한달 등으로 설정
- start 버튼: 대기상태(설정된 작업 주기로 대기 상태, 버튼을 누르면 비활성화된 후 정지버튼이 활성화됨
- stop 버튼: 정지상태(동작 상태 해제, 버튼을 누르면 비활성화된 후 동작 버튼이 활성화 됨
- exit 버튼 : 스케줄러가 종료됨

o display 부분
- Idle 상태인 경우 caption 은 "No Operation"으로 뜸
- 현재 작업 처리중인 경우 작업 설정 주기 등에 대한 caption 이 나타남

o log file 생성
- 스케줄링 작업 처리 내용에 대한 log가 일자별 txt file 형태로 저장됨
- 안산-안성의 각 local DBMS내 DML 명령을 수행한 사용자 계정 및 일시, 해당 테이블, 해당 컬럼명도 같이 수록되어 전송되어옴



[그림 2] 역학정보 스케줄링 작업 화면

4. 결론 및 고찰

본 논문에서 정의한 DB 설계 및 구축은 안산-안성의 역학 1기(2001. 5. - 2003. 2.) 연구에 적용한 것이다. 국내외 대부분의 역학 연구는 일정 term을 기준으로 검진자의 반복적 역학정보를 수집한다. 국립보건원 유전체연구소 유전체역학센터의 안산-안성 코호트 1기 연구와 현재 진행 중에 있는 2기 연구의 경우, 검진 항목이 1기 연구와 다른 것이 많다. 또한, 검진대상자의 이력정보 및 가계도정보가 발생하고 있다. 따라서, 이력정보 및 가계도정보에 대한 여러 속성들을 반영토록 계속 보완해 나갈 예정이다.

참고문헌

[1] DBLP(DataBase systems and Logic Programming), http://dblp.uni-trier.de/, 2003.
[2] Genome Frontier Project of Japan http://www.genome.ad.jp/documents/gf/frontier.html, 2003.
[3] Introduction to the UK Biobank; http://www.ukbiobank.ac.uk/, 2003.
[4] Korea National Institute of Health, Central Genome Research Center plan (Health and Medical Technology Development Project), 2000
[5] Medical Research Council, Public news; http://www.mrc.ac.uk/index/public-interest/public-news/public-news_archive/publicnews_archive_1_2002/public-biobank_uk.htm, 2003.
[6] National Human Genome Research Institute, Haplotype map development report, http://www.genome.gov/, 2002.
[7] PubMed, http://www.ncbi.nih.gov/entrez/query.fcgi, 2003.
[8] V. Kamp, Frank Wietek: Database System Support for Multidimensional Data Analysis in Environmental Epidemiology, IDEAS 1997: 180-190.
[9] V. Kamp and F. Wietek, Intelligent Support for Multidimensional Data Analysis in Environmental Epidemiology, In Advances in Intelligent Data Analysis, Second International Symposium, IDA'97, LNCS 1280, Springer, 1997.