J Korean Med Assoc 2012 August; 55(8): 711-719
http://dx.doi.org/10.5124/jkma.2012.55.8.711
pISSN: 1975-8456 eISSN: 2093-5951
http://jkma.org



# 포스트 EHR시대를 대비한 임상연구 전략: 전자건강기록을 이용한 관찰연구

박 래 웅* | 아주대학교 의과대학 의료정보학과

# A clinical research strategy using longitudinal observational data in the post-electronic health records era

**Rae Woong Park, MD***

Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea

***Corresponding author:** Rae Woong Park, E-mail: veritas@ajou.ac.kr

Received June 29, 2012 · Accepted July 12, 2012

Adoption of electronic health records (EHRs) is increasing worldwide. The worldwide EHR adoption rate is estimated to be around 9% to 12%. Thus, the accumulation of medical records in electronic form is also sharply increasing and is expected to be a precious asset for clinical research. Longitudinal observational studies based on EHRs are also increasing. Observational studies covering more than a million people are not rare at present. However, much of the current EHR data are equivalent in form to those of paper records, but are just stored in electronic stor-age devices, rather than as electronic data that can be transferred and shared without loss of clinical semantics. Current EHR systems must be improved in many ways to be used for anal-yses to yield important clinical knowledge. These improvements, which are addressed in this review, include the adoption of clinical data warehouses, use of controlled vocabulary, avoidance of personal/departmental research databases, a standardized interface of many diagnostic devices with the EHR system, control of time-stamp granularity, preparedness for whole-genome sequencing of every patient, confederation or consolidation of multi-institutional EHR data, protection of privacy and confidentiality, and an education system for clinical informaticians.

**Keywords:** Electronic health records; Data warehouse; Medical informatics; Longitudinal observation study

## 서 론

임상연구는 크게 실험연구와 관찰연구로 분류할 수 있다. 실험연구는 비뚤림과 교란요인을 최소화시킬 수 있는 연구방법이지만, 윤리적인 문제, 연구의 수행가능성, 또는 비용 문제로 인하여 수행하기 어려운 경우가 있다. 관찰연구, 특히 이미 수집된 이차자료를 이용하는 경우는 자료수집에 필요한 시간과 비용을 크게 줄일 수 있으며, 실험연구에서 수행하기 어려운 다양한 상황에 대한 분석이 가능하다. 하지만 관찰연구의 경우 의무기록 검토를 수반하기 때문에, 많은 연구대상자를 대상으로 연구하기에 어려움이 있다.

전자건강기록(electronic health record, EHR)이 보급되면서 임상관찰연구의 방식이 바뀌고 있다. 종이에 기록된 환자의 의무기록을 일일이 찾아서 쌓아놓고 검토하는 방식

에서 벗어나, 컴퓨터 앞에서 바로 연구 대상 환자의 의무기록을 검색하여 검토할 수 있게 되었다. 더 나아가 환자의 진료과정이 기록된 EHR시스템에 연구자가 직접 접근하여, 연구대상자를 선정하고, 가설에 맞추어 변수들을 추출하여 빠른 시간 내에 가설을 검증할 수 있게 되었다. 최근 들어 EHR을 이용하여 수백만 명의 환자들을 대상으로 하는 관찰연구 결과가 드물지 않게 보고되고 있다[1].

우리나라의 EHR 보급률은 전세계에서 가장 높다고 할 수 있다[2,3]. Yoon 등[4]의 조사 결과, 국내 종합병원급 이상 의료기관의 EHR 도입률은 37.2%로서, 미국의 주요 교육병원의 도입률 21.1%보다 높으며, 미국의 전체 EHR 도입률 11.9%(2009년)[5], 오스트리아 11.9%(2007년), 독일 7.0%(2007년)[6], 일본 10.0%(2007년)[7]보다도 월등히 높은 보급률을 보였다. 이러한 점을 고려할 때, 한국은 EHR 자료를 이용한 대규모의 관찰연구에 매우 유리한 여건을 가지고 있다고 할 수 있다.

정보통신 및 컴퓨터산업의 발전으로, 수십 년간 누적된 수억에서 수십억 건의 자료에 대한 복잡한 분석을 비교적 짧은 시간 내에 저렴한 비용으로 수행할 수 있게 됨에 따라[1,4,8,9], 과거에 상상할 수 없었던 방식의 연구가 가능하게 되었다. 하지만, 현재 EHR에 저장된 원천자료는 자료의 질이 떨어지기 때문에 이를 이용하여 신뢰할만한 결과를 도출하기 위해서는 정교한 역학적 설계와 분석은 물론이고, 자료의 질을 향상시키기 위한 많은 전처리 단계가 필요하다[10]. 근본적으로는 쌓이는 데이터가 대규모의 분석이 가능한 형태가 되도록 EHR에 자료를 입력하는 단계에서 많은 노력을 기울여야 한다.

본 논문의 목적은 EHR자료를 이용한 관찰연구에 필요한 도구와 개념들에 대하여 살펴보고, EHR내 임상자료의 활용에 걸림돌이 될 문제점들과 향후 발전방향에 대하여 논의하는 것이다.

## 임상데이터웨어하우스

EHR의 주된 목적은 임상데이터의 저장소로서 환자의 진료 및 진료와 관련된 행정업무를 지원하는데 있다. 따라서 EHR은 많은 자료의 정밀한 검색에 최적화 되어 있지 않다. 따라서, 연구 목적으로 EHR을 이용하는 과정에서 낮은 접근성, 성능저하, 자료분석기능의 부족과 같은 문제들에 직면하게 된다. EHR 자료를 연구에 활용하기 위해서는 EHR 데이터의 복사본을 정리한 별도의 데이터베이스 및 분석시스템이 필요하며, 이처럼 의사결정을 지원하기 위하여 시간의 선후관계를 가지고, 주제별로 정리 및 통합한 비휘발성의 데이터 저장소를 데이터웨어하우스라고 한다[11]. 특히 병원에서 환자진료 중에 발생한 임상자료와 분석도구를 의료진에게 제공하는 시스템을 임상데이터웨어하우스(clinical data warehouse, CDW)라고 한다[10,12].

연구를 위한 분석계 데이터베이스를 갖춤으로써 운영계 시스템에 영향을 주지 않고 과거 자료를 검색할 수 있다는 면에서 CDW는 임상연구에 매우 중요한 시스템이지만, 실사용자들이 연구에 필요한 수준의 데이터를 CDW에서 추출하는 것은 여전히 어려운 문제이다. 전통적인 의무기록을 이용한 연구에서는 개별 증례 별로 연구자의 판단에 따라 적절한 자료를 수집할 수 있지만, CDW를 이용한 관찰연구에서는 일률적 기준에 의하여 자료를 추출하기 때문에 개별 증례에 많은 예외 사항을 고려하기 힘들다. 또한 EHR 혹은 CDW의 자료는 낮은 수준의 이질적인 자료의 통합체로서, 충분한 정제작업을 거치지 않으면 임상자료로서 신뢰받기 어렵다[10]. 낮은 수준의 원천자료를 분석할 수 있는 형태의 자료로 만드는 과정이 전체 분석에 소모되는 시간과 노력의 70%이상을 차지한다[11]. EHR원천자료를 연구자료로 만들기 위해서 관계형 데이터베이스에서 자료의 검색과 관리를 위해 개발된 structured query language (SQL)와 같은 컴퓨터 프로그래밍 언어로 많은 양의 프로그래밍이 필요하다. 일반적으로 CDW시스템도 자료검색을 위한 기본적인 검색기능을 지원하지만, CDW시스템에서 제공하는 정도의 기능으로는 연구에 필요한 수준의 잘 정제된 자료를 추출하기 어렵다. SQL 언어를 일부 지원하는 CDW도 있지만, 조건분기, 루프(loop), 커서(cursor) 등 자료처리에 필수적인 SQL의 고급기능을 지원하지는 못한다. 전문 프로그래머가 임상연구자의 요청에 따라 SQL등을 이용하여 미리 검색문을 만들어 놓고, 사용자의 요청에 따라 이용하는 방법이 일

```
(TEXT like '% Tb %'OR TEXT like '% Tbc %' OR TEXT like '% tubercul%') AND (TEXT like '% cicatri%' OR TEXT like '% destr%' OR TEXT like '% bronchiect%' OR TEXT like '% deformit%' OR TEXT like '% scar%' OR TEXT like '% cicatr%' OR TEXT like '% distort%' OR TEXT like '% destro%' OR TEXT like '% atelecta%' OR TEXT like '% fibro%' OR TEXT like '% sequela%' OR TEXT like '% volume loss %' OR TEXT like '% volume decreas%' OR TEXT like '% decreased volume%' OR TEXT like '% 파괴%' OR TEXT like '% 변형%') AND (TEXT NOT like '%radiat%') AND (TEXT NOT like '%metasta%')
```

**Figure 1.** An example query to find cases with 'tuberculous destructed lung' from the computed tomography readings.

반적이다. 임상연구자와 SQL프로그래머는 데이터에 대한 개념과 관점 및 언어체계가 서로 다르다는 점을 인식하여, 상호간에 의사소통 및 의견의 일치를 위해서 많은 노력을 기울여야 한다.

## 전자건강기록 문서검색

임상의가 임상소견을 EHR에 기록할 때, 통상 통제어휘(controlled vocabulary)를 이용하거나 자유진술문(free text)을 이용하게 된다. 임상소견의 기록을 어떤 방식으로 했는가에 따라 해당 문서를 검색하는 방식도 달라지게 된다.

통제어휘란 미리 정의된 용어체계로서 특정 분야에서 사용되는 개념을 기술하는 용어체계를 말한다. 한국표준질병사인분류(Korean Classification of Diseases 5th revision, KCD-5), 국제질병 및 사인분류(International Statistical Classification of Diseases and Related Health Problems 10th revision, ICD-10), Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) 등이 통제어휘의 대표적인 예이다. 통제어휘를 사용함으로써 지식체계를 조직화하고, 검색 시에 일관된 결과를 도출할 수 있다. 통제어휘를 사용하면 짧은 시간 안에, 입력자의 의도에 충실한 자료를 검색할 수 있다. 많은 EHR시스템에서 진단명 조차 자유진술문 입력을 허용하는 경우를 종종 접하곤 하는데, 이런 경우 향후 EHR의 활용측면에서 매우 불리하다는 점을 인식하여야 한다. 따라서 자료의 활용성을 높이려면 최소한 진단명 입력에 대해서는 통제어휘 사용을 엄수해야 할 필요가 있다. 기존에 사용하는 KCD-5나 보험청구코드가 임상상황을 적절히 반영하지 못하기 때문에 어쩔 수 없이 자유진술문을 쓰게 되는 경우가 많은데, SNOMED CT와 같은 온톨로지 수준의 용어체계는 대부분의 임상상황을 잘 반영하고 있으며, ICD-10과도 연계되어 있기 때문에, 임상의사들은 SNOMED CT를 이용해 입력하고, 행정 또는 원무 측면에서는 SNOMED CT 코드와 연결된 ICD-10을 이용하는 방법을 적극적으로 도입할 필요가 있다. 다만 SNOMED CT는 포함하고 있는 개념이 30만 개가 넘기 때문에 따로 각 임상과에서 자주 사용하는 용어세트를 만들어야 하는데, 이러한 작업에 많은 시간과 노력이 소모된다.

자유진술문은 특별한 규정 없이 일상적 언어를 이용하여 표현한 문장을 말한다. 자유진술문으로 기록된 문건에서 의미 있는 데이터를 뽑아내는 대표적 방법으로 자연어처리(natural language processing) 기법을 들 수 있다. 자연어처리를 이용하여 자유진술문 형태로 기록한 임상문서에서 약처방, 문제목록, 종합적인 임상정보를 추출하는 시도가 진행되고 있다[13-16]. 의료분야에서는 Medical Language Extraction and Encoding System이 대표적인 자연어처리 엔진이다[17]. 하지만 우리나라의 경우 의무기록에 한글과 영어가 혼용되어 사용되므로 자연어처리에 어려움이 있다. 의학이외의 분야에서는 한영혼용문에 대한 자연어처리 연구가 많이 진행되고 있지만, 의학분야에서는 자연어처리 연구가 거의 없는 실정이다. 자유진술문에 대한 질의어검색은 우리가 인터넷검색이나 PubMed 등에서 원하는 문서를 찾는데 가장 흔히 사용하는 검색방법이다. 한영혼용 의학문서에 대한 자연어처리가 아직까지 실용화 단계까지는 거리가 멀다는 점을 고려하면, 질의어를 통한 검색이 당분간 우리가 이용할 수 있는 유일한 대안이라고 볼 수 있다. 하지만 이는 시스템에 많은 부하를 주기 때문에 EHR은 물론, CDW에서도 사용하지 못하는 경우가 많다. 자유진술문 검색을 위해서는 정교한 검색기법이 필요한데, CDW에서 제공하는 일반적인 검색기능으로 이를 구현하기는 쉽지 않다. Figure 1은 한 삼차의료기관에 누적된 18년간의 컴퓨터단층촬영술 판독소견으로부터 'tuberculous destructed lung' 소견을 찾기 위한 SQL 질의어의 예이다.

질의어를 이용한 검색을 할 때 결과 테이블을 주제나 검사 별로 분리한 후 따로 검색하는 것이 검색시간을 줄이는데 유리하다. Figure 1을 예로 들면, 전체 방사선 판독결과(n=7,470,072)에 대하여 검색을 수행한 결과 약 440초의 시간이 소모되었으나, 컴퓨터단층촬영술 판독결과(n=71,878)만을 분리하여 별도의 테이블을 구성한 후, 동일한 검색을 수행한 결과 단지 18초의 시간이 소요되었다. 테이블을 어떤 주제로 분리할 것인지는 실 사용자인 임상의사들이 결정하여야 한다.

## 교실/과에서 관리하는 연구용 데이터베이스

교실이나 과에서 관리하는 연구용 데이터베이스는 연구에 필요한 증례를 유연한 형태로 상세하게 기록하고, 필요한 자료를 빠르게 찾을 수 있다는 장점이 있다. 하지만 이러한 유연성으로 인해 자료의 구조가 계속 바뀌며, 입력변수의 의미가 시간이 지나면서 달라지는 문제가 있다. 또한 자료입력을 여러 사람이 하게 됨에 따라 처음 설계자가 의도한 변수의 정의와 달리 잘못 입력되는 경우가 흔하다. 이로 인해 수십 년간 모은 자료를 막상 분석하고자 할 때, 그 변수가 무엇을 의미하는지 알기 어렵고 분석결과도 신뢰하기 힘든 경우가 많다. 따라서 연구용 데이터베이스는 EHR내부에 구축하는 것이 바람직하다. 또한 과거에 구축한 자료도 EHR로 이관할 필요가 있다. 그리고 반드시 연구용 데이터베이스구축 및 개정에 대한 명문화된 문서를 만들어 보관하고 주기적으로 개정하여야 한다.

## 기능검사 결과와 전자건강기록과의 연동

검사장비로부터 나오는 자료는 표준 프로토콜이나 혹은 제조업체가 제공하는 프로토콜을 이용하여 EHR과 연동될 수 있다. 심전도 관리시스템의 경우 심전도기기에서 측정한 값을 중앙의 서버로 보내면, 서버는 심전도 결과와 파형을 자체 데이터베이스에 보관하고, 동시에 측정한 결과 수치들은 연계서버를 통하여 EHR에도 기록할 수 있다. 호흡기능검사 결과도 이처럼 자동으로 EHR에 수치데이터로 저장하는 것이 가능하다. 하지만 장비가 구형이거나 또는 신형장비와 구형장비가 혼재되어 사용되는 경우 연동을 포기하고 종이로 출력한 후 스캔하여 보관하는 경우도 많다. 만일 검사결과지를 스캔하거나 또는 PDF파일 형태로 보관하고 있다면, 광학문자인식이나 패턴인식을 통하여 해당 검사의 결과수치를 일괄해서 추출하는 것이 가능하다. 실제로 Park 등[18]은 과거 18년간 소속 의료기관에 누적된 모든 심전도에 대하여, 웹파싱기법과 광학문자인식기술을 적용하여, 심전도 관리시스템 내부에 저장된 심전도 자료는 물론, 종이로 출력된 후 스캔되어 보관 중이던 이미지 심전도 자료에 대하여 주요변수를 수치값으로 추출하여, 총 74만 건의 결과값을 데이터베이스화하여 공개하였을 뿐만 아니라, 추출에 필요한 프로그램과 프로그램 소스도 공개하였다(http://www.ecgview.org).

## 자료발생 시간의 정확도

데이터베이스기반의 관찰연구에서는 데이터가 기록된 시간의 선후관계를 제대로 고려하지 않으면 자료 추출과정에서 오류가 발생할 수 있다. 이때 기록된 시간 값이 얼마나 정확한 지에 대해서 확인과정을 거쳐야 하며, 발생시점의 정확도는 자료에 따라 다르다는 점을 고려하여야 한다. 예를 들어 입원자료 중에서 약물투약시간은 통상 하루 중에 일정 시간에 순차적으로 투여하고 일괄해서 입력하는 방식이므로, 시간단위로 기록이 된다(분 이하 시간값은 00으로 기록). 이에 비하여 검사장비에서 출력되는 결과값은 초 단위까지 기록이 된다. 하지만 장비에서 출력되는 시간값이 얼마나 정확한 것인지 확인해 보아야 한다. 각 장비의 시간값을 주기적으로 확인하여 수정하는 절차가 있는지, 혹은 장비에서 참조하는 기준시간이 무엇인지 확인하여야 한다. 시간의 동기화 방법으로 지피에스나 타임서버를 이용한 동기화가 가장 신뢰할 만 하지만, 이런 방식으로 시간값을 동기화는 경우는 거의 없다. 따라서, 초/분 단위로 자료간의 전후 상관성 유무를 확인하는 것은 많은 오류를 포함할 수 있음을 인식하여야 하며, 자료 별로 기록되는 시간의 특성을 이해하고 있어야

한다. 약처방의 경우 간호수행 시간을, 검사결과는 검체채취 시간이나 검사수행 시간을 기준으로 삼아야 한다.

## 다기관 통합 전자건강기록시스템

여러 기관의 임상의료정보를 실시간으로 익명화하고 통합한 후, 이를 기관 내외부 연구자들에게 제공하는 다기관 통합 EHR시스템에 대한 관심이 증가하고 있다. 한 기관에 국한된 자료를 이용한 연구는 결과의 일반화에 한계가 존재한다. 만약 다른 기관의 자료에서도 결과가 동일하게 반복된다면 결과의 재현성과 외적타당도를 확보할 수 있겠지만, 비교할 수 있는 외부자료는 구하기 쉽지 않다. 개인건강정보 보호에 대한 의무와 두려움, 또는 자료공유에 대한 거부감 때문에, 자료간의 통합이나 혹은 외부연구자에게 자료를 제공하는 것은 극히 제한적이다. 국외에서는 여러 기관의 의료정보를 통합하여 활용하고자 하는 다양한 노력이 진행되고 있다. 미국의 i2b2 프로젝트는 임상연구자들이 기존의 임상데이터와 유전체 정보를 결합하여 공유하려는 프로젝트이다[19]. 미국의 mini-Sentinel System이나[20], 유럽의 EU-ADR 프로젝트는[21] 보험청구자료나 EHR 자료를 통합하여 약물 유해반응을 조기에 감지하기 위한 프로젝트이다. 국내에서는 생물자원관리본부, 인체자원중앙은행, 건강보험심사평가원의 건강보험 청구 자료, 국민건강보험공단의 건강검진 자료 등이 여러 의료기관의 임상자료를 수집하여 통합하고 있다. 하지만, 유료이거나 비표준 데이터베이스를 사용하거나, 자료가 특수한 목적으로 제한되어 있거나(mini-Sentinel System, EU-ADR), 비개방성(건강보험심사평가원, 국민건강보험공단), 또는 임상데이터 부재(생물자원관리본부, 인체자원중앙은행)로 인해 임상연구에 활용하기에 많은 제약이 있다. 최근 민간의료기관을 중심으로 EHR자료를 익명화하여 서로 통합하고, 통합된 자료에 대해서 참여기관간에 자유롭게 이용하고자 하는 프로젝트가 시작되었다. 이를 위해서는 병원 간 상이한 자료구조를 공통의 자료구조로 바꾸어야 하는데, 이때 필요한 것이 공통데이터모델(common data model, CDM)이다. 현재 컨소시엄에 참여한 3개 의료기관이 EHR을 분석하여 16개의 주요 테이블을 통합 대상으로 선정하고, 이에 대한 공통데이터모델을 개발중에 있다. 컨소시엄에 참여하지 않은 의료기관이라 할지라도, 공통데이터모델에 맞추어 자료를 출력하면 자료공유가 가능하다. 실시간으로 다기관 의료정보를 통합하고 공유할 수 있게 된다면, 전국적 규모의 여러 인구집단을 반영하는 연구자료로서 충분한 표본 수 확보가 가능하게 되며, 전염성 질환 등 현재 발생하는 보건학적 문제에 대한 즉각적인 자료분석이 가능하게 될 것이다.

## 개인정보 보호, 익명화 및 개인식별정보 제거

EHR에 생성저장된 의무기록을 학술적 연구목적으로 공유하기 위해서는, 환자의 개인정보가 보호되고 환자 개인을 식별할 수 없도록 충분한 조치가 취해져야 한다. 익명화는 개인을 식별할 수 있는 모든 정보를 비가역적으로 제거하는 과정이다[22]. 하지만, 이렇게 완전히 익명화된 자료는 연구목적으로 적절하지 못한 경우가 많다. 개인식별정보 제거는 데이터에서 개인을 식별할 수 있는 식별자를 제거하는 것으로, 개별적으로는 개인식별자가 아니지만, 조합하여 개인을 알아 볼 수있는 자료도 개인식별자에 해당한다. 우리나라 개인정보보호법에서는 "개인정보란 살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보(해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함한다)"로 정의하고 있다. 미국의 Health Insurance Portability and Accountability Act에서는 이름, 사회보장번호, 병록번호 등 18개 항목을 민감건강정보(protected health information)로 규정하고, 임상자료 연구활용의 기본조건으로 민감건강정보의 익명화를 규정하고 있다[23]. 개인식별정보 제거를 위한 구체적인 방법으로 개인식별정보 삭제, 90세 이상의 나이를 90세로 변경, 양 극단에 있는 결과값을 95% 또는 99%선의 값으로 대체(top-coding)[24], 사회적 차별의 가능성이 있는 민감한 진단명 삭제, 약물 상품명의 성분명 변환, 일정한 범위 내에서 환자의 모든 날짜정보 동시 이동(random shift of date) 등의 방

법을 적용할 수 있다[18]. 하지만 현실적으로 자료의 유용성을 유지하면서, 동시에 개인정보를 100% 완벽하게 보호할 수 있는 방법은 없다고 보아야 한다. 개인정보 보호 조치의 강도에 따라 자료의 유용성과 개인정보 누출 간에 상반관계가 형성된다. 어느 수준의 개인정보 보호조치를 강구할 것인지는 기술적 문제가 아니라 사회적 합의의 문제임을 인식하고, 이해 당사자간의 활발한 논의를 통하여 적절한 합의를 이끌어 내야 한다.

## 유전체검사 대중화

Science는 2006년에 "The race for the $1,000 genome"이라는 기사를 통해 2016년경에는 1,000달러의 비용으로 유전체 분석이 가능할 것이라고 전망하였다[25]. 2007년 솔렉사에 의해서 차세대 염기서열 분석(next geration sequencing, NGS)기술이 상용화되면서 이 가능성은 현실이 되었고, 실제로 2011년 약 400만 원 선이던 분석비용이 2012년 200만 원 선까지 하락했다. 조만간 100달러의 비용으로 전유전체분석이 가능한 시기가 도래할 것으로 보인다. 이처럼 whole genome sequencing (WGS) 및 whole genome analysis가 저렴한 비용으로 가능하게 된다면, 개인 맞춤형 예방, 진단, 치료를 위하여 환자들에 대한 유전체 서열분석을 일상적으로 수행할 시기가 오게 될 것이다. 하지만 아직까지는 생산되는 유전체 자료의 질이 임상에 직접 적용할 수준의 질에 이르지 못하고 있으며, 또한 해당 환자의 임상자료와 연계되어 있지 않아 질병과의 상관성 분석에 사용되기에는 부적합하다. 유전적 변이와 질병과의 상관성 분석을 위해서는 개인의 유전체정보와 충분히 누적된 임상자료가 같이 연계되어 인구집단 수준으로 존재하여야 한다. NGS 가격이 지금보다 훨씬 더 저렴해지고, 환자의 임상정보가 충분히 쌓인 시점에서의 유전체자료 및 임상정보의 통합분석을 염두에 두고, 지금부터 적절한 절차를 거쳐서 환자의 검체를 수집보관할 필요가 있다. 예를 들어 특정 검사를 위해 환자로부터 채혈한 검체를 본래 목적한 검사를 실시한 후에 관련 법에 정한 절차에 따라 환자의 동의를 구하고, 검체를 조직은행에 보관하는 방식이 가능할 것이다. 수 년이 지나서 보관된 혈액을 이용하여 WGS를 수행한다면, 유전자의 변이와 질병과의 상관성 분석을 지금과 비교할 수 없는 저렴한 비용으로 수행할 수 있게 될 것이다. 다만 내가 그 실과를 거두지 않을 사과나무를 누가 심느냐의 문제가 남아있다. 또한 수 Tbytes에서 수백 Gbytes에 달하는 NGS 데이터를 어느 수준에서 보관할 것인지도 고민하여야 한다. 잘 정돈된 염기서열 전체는 6 Gbyte정도의 용량(30억 bp)에 불과하며, 참고데이터와의 차이점만을 저장하거나, 유전자의 존재유무나 변이 유무 정보만을 저장한다면 몇 Mbyte 수준으로 줄일 수도 있을 것이다. 이것은 자기공명영상을 촬영할 때 생산되는 영상자료의 양에 비교하면 EHR 또는 영상저장 및 전송체계 내에 저장할 수 있을만한 양의 자료이다. 다만, 데이터를 축약할수록 그 활용 가능성은 줄어들게 된다. 유전체검사 대중화와 관련하여 사전동의의 범위, 알지 않을 권리, 유전정보의 보관기간, 태아/소아에 대한 검사, 사회적 차별, 개인정보 보호 등의 많은 윤리적 문제가 존재한다. 이에 대한 활발한 논의와 사회적 합의에 대한 노력이 병행되어야 할 것이다[26].

## 임상데이터 과학자 및 정보의학 인력 양성

앞으로 전통적인 임상자료뿐만 아니라, 영상자료, 생체신호자료, 각종 분자생물학 자료 및 소셜네트워크 자료 등의 이질적인 자료가 기관별 EHR, 또는 개인건강기록(personal health record)을 중심으로 수집되어 누적될 것으로 전망되고 있다. 또한 한 기관의 정보만이 아니라 적절한 익명화 단계를 거쳐서 여러 기관의 정보가 통합되거나 연계되어 갈 것이며, 이를 이용하여 환자 개인에 맞는 진단 및 치료법을 개발하기 위한 노력이 점차 증가할 것으로 보인다. 하지만 이러한 이질적이고 다차원적인 대량의 데이터는 기존에 사용하던 전통적인 자료처리 방식으로 다루기에는 많은 어려움이 있다. 대량 데이터의 정제 및 통합능력, 고성능의 컴퓨터를 이용한 고속 계산 및 해석능력이 필요하다. 이를 위해서는 의학, 통계, 역학 등 전통적인 의생명자료 분석기술뿐만 아니라, 정보통신, 컴퓨터공학 분야의 기술을 많이 필요로 한다. 일선의 임상의학자들이 이러한 기술을 익힌다는 것은

대단히 어려운 일이므로 이러한 기술을 익힌 의료정보/정보의학 전문가들이 필요하다. 현재 정규 의료정보/정보의학 과정을 수료한 인력은 전국적으로 극소수에 불과한 실정이다. 최근 대한의료정보학회는 임상의사들을 대상으로 1년 - 1년 반 정도의 기간 동안 축약된 정보의학 교과과정을 운영할 예정이며, 이를 바탕으로 정보의학 인증의 또는 이와 유사한 제도의 도입을 고려하고 있다. 복잡한 의생명정보가 급증하는 상황을 고려하면, 정보의학 전공자, 정보의학 인증의 등 다양한 수준의 임상 데이터 과학자 양성이 시급한 것으로 보인다.

## 결 론

지금까지 EHR자료가 임상연구, 특히 관찰연구에서 유용한 자료로 활용되기 위해서 현재의 EHR시스템을 개선하거나, 또는 우리가 극복하여야 할 많은 문제점들에 대하여 기술하였다. 이러한 문제를 해결하기 위해서는 무엇보다도 기관차원에서의 CDW구축이 선행되어야 한다. 진단명과 같이 의무기록의 중심을 이루는 핵심개념에 대해서는 자유진술문의 사용을 배제하고 통제어휘를 사용하여야 한다. 개인 또는 과에서 유지하고 있는 개별 연구용 데이터베이스는 EHR 시스템 내로 이관하여 관리하는 것이 장기적인 유지관리 및 자료의 활용성 측면에서 훨씬 유리할 것이다. EHR 자료에 기록되는 각종 기록들의 시간의 정밀도가 다르다는 점을 고려하고, 각종 의료장비들의 시간값을 동기화하기 위한 절차와 방법을 모색하여야 한다. 대규모 자료에 대한 자동화된 분석을 시도하기 위해서는 자료수집과 분석에서 발생할 수 있는 여러 가지 예외사항을 모두 고려하여야 한다. 생성되는 의무기록을 한 기관 내에서만 활용하는 차원을 넘어서서 여러 기관의 자료를 통합 및 관리하여 이용할 필요가 있다. 이 때, 개인정보보호를 위한 적절한 익명화/개인식별정보 삭제 등의 기술적 조치 및 자료활용에 대한 절차확립은 물론, 그에 대한 사회적 합의를 이루기 위한 노력이 병행되어야 한다. EHR자료의 낮은 질, 다양한 이질적 데이터, 자료양의 급증 등을 고려할 때, 임상연구자가 EHR 원천자료를 직접 이용하기는 대단히 어려운 일이다. 따라서, 정보의학 전공자 혹은 정보의학 인증의 제도 등을 통하여 숙련된 임상 데이터 과학자들을 양성하여 활용할 필요가 있다.

## Acknowledgement

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (2012-0000995), and by a grant of the Korea Health technology R&D Project, Ministry of Health & Welfare, Republic of Korea (A112022).

**핵심용어:** 전자건강기록; 데이터웨어하우스; 의료정보학; 관찰연구

## REFERENCES

1. Ray WA, Murray KT, Hall K, Arbogast PG, Stein CM. Azithromycin and the risk of cardiovascular death. N Engl J Med 2012;366:1881-1890.
2. Yoon D, Chang BC, Kang SW, Bae H, Park RW. Adoption of electronic health records in Korean tertiary teaching and general hospitals. Int J Med Inform 2012;81:196-203.
3. Park RW, Shin SS, Choi YI, Ahn JO, Hwang SC. Computerized physician order entry and electronic medical record systems in Korean teaching and general hospitals: results of a 2004 survey. J Am Med Inform Assoc 2005;12:642-647.
4. Yoon D, Park MY, Choi NK, Park BJ, Kim JH, Park RW. Detection of adverse drug reaction signals using an electronic health records database: Comparison of the Laboratory Extreme Abnormality Ratio (CLEAR) algorithm. Clin Pharmacol Ther 2012;91:467-474.
5. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, Shields A, Rosenbaum S, Blumenthal D. Use of electronic health records in U.S. hospitals. N Engl J Med 2009;360:1628-1638.
6. Hubner U, Ammenwerth E, Flemming D, Schaubmayr C, Sellemann B. IT adoption of clinical information systems in Austrian and German hospitals: results of a comparative survey with a focus on nursing. BMC Med Inform Decis Mak 2010;10:8.
7. Yasunaga H, Imamura T, Yamaki S, Endo H. Computerizing medical records in Japan. Int J Med Inform 2008;77:708-713.
8. Park MY, Yoon D, Lee K, Kang SY, Park I, Lee SH, Kim W, Kam

HJ, Lee YH, Kim JH, Park RW. A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database. Pharmacoepidemiol Drug Saf 2011;20:598-607.

9. Sheen SS, Choi JE, Park RW, Kim EY, Lee YH, Kang UG. Overdose rate of drugs requiring renal dose adjustment: data analysis of 4 years prescriptions at a tertiary teaching hospital. J Gen Intern Med 2008;23:423-428.
10. Lee SM, Park RW. Basic concepts and principles of data mining in clinical practice. J Korean Soc Med Inform 2009;15:175-189.
11. Han J, Kamber M. Data warehouse and OLAP technology for data mining. In: Han J, Kamber M, editors. Data mining: concepts and techniques. San Francisco: Morgan Kaufmann Publishers; 2001. p. 39-43.
12. Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. Medical data mining: knowledge discovery in a clinical data warehouse. Proc AMIA Annu Fall Symp 1997:101-105.
13. Friedman C, Hripcsak G, Shagina L, Liu H. Representing information in patient reports using natural language processing and the extensible markup language. J Am Med Inform Assoc 1999;6:76-87.
14. Meystre SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. Int J Med Inform 2008;77:602-612.
15. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010;17:19-24.
16. Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. Clin Pharmacol Ther 2012;92:228-234.
17. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995;122:681-688.
18. Park MY, Yoon D, Choi NK, Lee J, Lee K, Lim HS, Park BJ, Kim JH, Park RW. Construction of an open-access QT database for detecting the proarrhythmia potential of marketed drugs: ECG-ViEW. Clin Pharmacol Ther 2012 Jul 25 [Epub]. DOI: 10.1038/clpt.2012.93.
19. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc 2006:1040.
20. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, Nelson JC, Racoosin JA, Robb M, Schneeweiss S, Toh S, Weiner MG. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiol Drug Saf 2012;21 Suppl 1:1-8.
21. Trifiro G, Fourrier-Reglat A, Sturkenboom MC, Diaz Acedo C, Van Der Lei J; EU-ADR Group. The EU-ADR project: preliminary results and perspective. Stud Health Technol Inform 2009;148:43-49.
22. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002;26:1-24.
23. U.S. Department of Health and Human Services; National Institutes of Health. Research repositories, databases, and the HIPAA privacy rule [Internet]. Bethesda: National Institutes of Health; 2004 [cited 2012 Jun 27]. Available from: http://privacyruleandresearch.nih.gov/research_repositories.asp.
24. El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Rose S, Howard J, Gluck J. De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. J Med Internet Res 2012;14:e33.
25. Service RF. Gene sequencing. The race for the $1000 genome. Science 2006;311:1544-1546.
26. Dondorp WJ, de Wert GM. The 'thousand-dollar genome': an ethical exploration. The Hague: Centre for Ethics and Health; 2010.

## Peer Reviewers' Commentary

본 논문은 EHR 이후에 중요하게 부각될 임상연구 데이터웨어하우스의 중요성에 대해서 언급하고 있다. EHR 자료를 이용하여 대규모 관찰적 임상연구를 수행함으로써 임상적 의사결정에 필요한 과학적 근거를 저비용 고효율적으로 생산할 수 있게 된 것이다. 본 논문은 EHR 자료를 임상연구에 활용하기 위해 필요한 선결조건과 해결방안을 소개하고 있어 시의 적절한 논문으로 판단된다. 이러한 자료의 활용성을 높이려면 기관차원의 CDW 구축, 진단명과 같은 핵심개념의 자유진술문 대신 통제어휘 사용, 개별 연구용 자료의 EHR시스템 내 이관, 각종 정보의 시간정밀도를 높이는 방법 및 자동분석방법의 개발 등이 선행되어야 한다. 다기관 통합 EHR시스템을 구축하는 것이 연구결과의 대표성을 높이기 때문에 관심이 높아지고 있으나 개인정보보호와 자료공유에 대한 거부감을 극복하여야 한다. 아울러 이러한 여건의 변화를 주도적으로 이끌어갈 수 있는 전문인력의 양성이 시급하다.

**[정리: 편집위원회]**