공공 및 의료분석

공공데이터 개방과 관련한 건강보험심사평가원의 빅데이터 사업과 SAS 활용사례

- 김록영(건강보험심사평가원 부연구위원)

보건의료정보 현황

건강보험심사평가원은 진료 심사 및 평가를 하는 전문기관입니다. 건강보험 청구자료는 다음과 같은 프로세스로 처리됩니다. ▶요양기관에서 서면 및 전산매체, EDI 등을 통해 건강보험심사평가원에 급여를 청구하면 ▶전문가 점검, 처방전 인덱스 등 다양한 기준으로 심사를 하고 ▶심사결과를 국민건강보험공단에 통보합니다. 동시에 건강보험심사평가원의 데이터 웨어하우스 시스템에 데이터를 저장합니다.

한편, 진료비 청구 명세서에는 120개 항목의 데이터가 있습니다. 이 명세서에서 건강보험심사평가원으로 들어오는 데이터의 양은 1년에 약 50테라바이트에 이르며, 건수로는 약 13억 건입니다. 이 외에도 의약품안심서비스(DUR), 요양기관현황, 병원평가정보, 안전상비의약품 편의점 정보, 의약품생산 수입 공급, 의사별 처방, 병원 영상정보 교류 등 총 60여 종, 200억 건(약 52테라바이트)의 정보를 보유하고 있습니다. 특히 건강보험 청구자료는 매일 새롭게 생성되므로 그 용량이 크고, 방대하여 빅데이터에서 굉장히 중요한 부분을 차지합니다.

건강보험심사평가원이 보유한 데이터 특히, '명세서 진료정보'에는 몇 가지 특징이 있습니다. 첫째, 개인정보입니다. 명세서 진료정보는 의료데이터로서 굉장히 민감한 정보입니다. 진료기록 자체가 개인정보를 기반으로 구성되어 있으며, 특정 개인의 실시간 의료이용 현황을 알 수 있습니다. 둘째, 임상데이터가 없습니다. 급여비용을 청구하기 위한 데이터이므로 병원에서는 환자의 건강상태 등을 기입할 의무가 없습니다. 따라서 EMR 데이터의 검사기록, 비급여 내역의 데이터, 건강검진기록 등이 존재하지 않습니다. 셋째, 빅 사이즈입니다. 연간 14억 건에 이르는 명세서를 기반으로 상세진료내역, 처방조제내역 등 데이터 사이즈가 너무 큽니다. 넷째, 복잡성입니다. 데이터의 구조가 복잡하고 전문적이어서 분석과 해석 시 주의가 필요합니다.

현재 건강보험심사평가원은 공공데이터 제공 및 이용 활성화에 관한 법률, 공공기관의 정보공개에 관한 법률 등 법적 근거를 기준으로 정보공개를 하고 있으며, 앞으로도 법이 허용되는 범위에서 추가적으로 정보를 공개할 예정입니다.

보건의료 빅데이터시스템

2013년 분석시스템 구축을 완료하고, 2014년 포털시스템 구축을 완료할 예정입니다. 총 예산은 150억 규모에 이릅니다. 보건의료와 관련된 통계분석시스템, 개방용 DB시스템, 원격접속시스템, 정보 포털을 구축하여 보건의료 정보제공 허브역할을 위한 '보건의료 빅데이터 플랫폼'이 구축됩니다.

이 같은 플랫폼을 통해 건강보험심사평가원의 정보는 ▶보유 정보를 직접 제공하는 '정보 개방' ▶타 기관에서 정보를 결합하여 분석을 원하는 경우 원내시스템에서 분석을 하는 '정보연계 분석' ▶개방된 타 기관 정보를 수집, 융합하여 정보의 가치를 높이는 '정보 융합' ▶타 기관에서 건강보험심사평가원 정보와 결합을 원하는 경우 이를 결합하여 제공하는 '정보연계 개방', 그리고 다시 '정보 개방'으로 연결되는 선순환 구조를 갖게 됩니다. 이를 통해 보건의료 분야의 산학연 연구를 지원할 것입니다.

의료정보지원센터 운영계획

이러한 데이터와 시스템 등의 인프라를 기반으로 건강보험심사평가원은 4월 17일, 의료정보지원센터를 개소했습니다. 지원센터에는 ▶외부 연구자를 위한 '시설 및 지원 인프라' ▶정보개방을 위한 빅데이터분석시스템, 원격분석시스템, 의료빅데이터DB 등 'IT인프라' ▶산학연 공동 커뮤니티와 전문가 그룹 등 '협력 인프라' ▶건강보험청구자료를 활용할 수 있는 '인력 양성 인프라' 등을 구축했습니다.

의료정보지원센터의 궁극적인 목적은 '정보제공을 통한 보건의료산업 생태계 조성'입니다. 이를 위해 원시 데이터를 중심으로 '데이터 서비스'를

제공하고, 원격분석시스템을 통해 지방에서도 활용할 수 있도록 지원합니다. 아울러 의료정보지원센터 청구데이터에서 2차 데이터인 '가공 정보 서비스'를 제공합니다. 한편 정보 제공에서 그치지 않고, 인력양성(데이터를 활용할 수 있는 교육프로그램, R&D 컨설팅, 전문인력양성) 등도 지원합니다.

이를 위해 별도의 전담팀(RAD; Research Assistance Desk)를 구성했습니다. 이 전담팀은 건강보험 청구자료 활용을 위한 접속환경 설정, 데이터 속성, 정보분석 등의 교육, 프로젝트 진행을 위한 일대일 맞춤형 컨설팅, 정보활용 결과를 수집/공유하여 피드백 할 수 있는 피드백 체계 확립, 보건의료 빅데이터 전문인력 양성 등의 업무를 담당합니다. 아울러 정보활용 및 활성화를 위한 운영위원회를 구성하고, 연구중심병원 워킹그룹(Working Group) 운영 및 지원을 추진하고 있습니다.

특히 보건의료 분야에서 데이터 사이언티스트 양성이 무엇보다 중요합니다. 이를 위해 SAS코리아와 전략적 제휴를 체결하고, 제11회 'SAS 데이터마이닝 챔피언십 2013'을 공동 개최했으며, 올해에도 추진 예정입니다. 이와 함께 보건의료 데이터 사이언티스트 인증 프로그램(연 2회, 5일 과정)을 공동으로 개발하여 지금까지 교육을 2회 실시했습니다.

한편, 의료정보지원센터에서 제공하는 정보는 ▶1년 단위 표본자료(전체환자표본, 입원환자표본, 노인환자표본, 소아청소년환자표본 등) ▶오픈 API 소스 개발(의료경영지원 DB, 의약품 유통정보 DB, 진료비 예측정보 DB 등) ▶원격분석시스템(보유자료 연계서비스, 5%/10%/20% 샘플링 자료 제공, 연구자 맞춤형 정보 제공) ▶의료정보지원센터 등입니다.

SAS를 활용한 의료정보 연구•개발

환자표본 자료 구축_ 건강보험으로 청구되는 환자 수가 1년에 4,600만 명에 이릅니다. 이 많은 데이터에서 표본데이터 셋을 추출하려니 데이터가 너무 커서, 2009년 SAS 프로그램을 기반으로 '환자표본(HIRA-NPS)'을 산출했습니다. 총 4종의 표본 데이터셋을 제공하며, 샘플링 환자수는 약 150만명, 데이터 량은 약 30기가 정도입니다. 건수로는 2,500만 건입니다.

이 표본 자료를 활용하여 만성신부전 환자 추정 및 뇌졸중 발병 분석을 수행했습니다. 만성신부전 환자의 경우 실제 모집단을 갖고 추정해보니 모집단 환자 빈도가 약 11만 명이 나왔습니다. 반면 환자표본 자료의 신뢰구간을 구해보니 오차 구간 내에 모집단 환자가 모두 들어갔습니다.

진료경향 모니터링 시스템_ 진료비 청구 정보를 체계적으로 구조화한 시스템으로, 건강보험 내/외부 자료를 통합하여 분석할 수 있는 시스템입니다. 데이터마이닝 기법으로 통계분석을 고도화하여 객관적인 분석 및 예측, 이상징후 감지 등을 할 수 있습니다. 이를 통해 진료비 지출 변동에 대한 과학적인 평가 및 예측으로 지출관리를 효율적으로 할 수 있게 됩니다.

이 시스템에는 크게 두 가지 기능이 있습니다. 첫째, 진료 경향 모니터링. 총 진료비, 내원 일수, 청구 건수 등의 지표에 대한 분류 유형별 통계현황을 분석하고, 기여도 및 외부영향요인 통합분석으로 진료비 변동의 실질적인 원인을 파악합니다. 아울러 다변량회귀분석으로 단기(12개월), 중기(3년) 진료비 예측정보를 제공합니다. 이 예측에 SAS 프로그램을 사용했습니다. 둘째, 이상징후감지. 상병, 진료행위, 약품, 치료재료에 대하여 개별 항목별 이상변동징후를 감지하고, 이상징후 항목에 대해서는 해당부서(심사, 급여, 약제 등) 조치체계를 마련했습니다. 이 시스템을 활용하면 데이터 마이닝을 통해 객관성 있는 진료비 지출평가 및 예측 정보가 가능하고, 4대 중질환 보장성 강화 등 정책 쟁점 항목에 대한 모니터링도 가능합니다.