AI for Healthcare & Privacy Protection for Healthcare (sooyong.shin@amc.seoul.kr)

- Healthcare Data
    - 의료기관 (Clinical data) - EMR (Electronic Medical Records)
    - 공공기관 (Claim data) - 건강보험공단, 건강보험심사평가원
    - 공공기관 (Environmental data) - 기상청, 환경부
    - 개인 (Patient-Generated Healthcare Data) [PGHD] **
        - mHealth, wearable, bio-sensor, home-monitoring devices, ...
    - 개인 (Life Log) *
        - wearable, SNS, purchase data, ...

  ** MIT Technologoy Review Vol.117 No.5

    - Electronic Medical Records
    - Family Health History
    - Public Health Data
    - Mobile Health Data
    - Environmental Data
    - Genomic Data
    - Insurance Claims Data +
    - Patients
    - Doctors
    - Reserachers

- Healthcare Data의 특성
    - Text, Image, Video, Sound, ...
    - 대부분이 비정형화

- Size of Individual Healthcare Data
    - Clinical Data: 10% -- 0.4 TB / per Life
    - Genomics Data: 30 % -- 6 TB / per Life
    - Exogenous data (behavior, Environmental, ..) : 60 % -- 1100 TB / per Life

- Facts & Limitations on EMR
    - 국내 EMR 보급률이 92% 정로라고 하지만 대부분은 무의미
    - EMR 데이터가 대부분 Word Processor 수준 (대부분 Text, 많은 약어, 의사마다 표기가 다름, 통일성 Zero)
    - OCR, NLP 기술로 국내에서 의미있는 추출 불가능 : 통일성 없는 축약(약어), 전문적 의학 용어, 문장이 아닌 단어 나열, 콩글리시, 심볼 난발, ...
    - 일부 검사 보고서(병리보고서)는 semi-structure 형태로 패턴만 알면 99% 정확도로 데이터 추출 가능
    - NER(Named Entity Recognition)에 Regular Expression으로 대부분 가능

    - 하지만, 의료진의 높은 기대치에는 못 미침. (정확도 95%, 리콜 90% -- 사용불가)
    - Clinical Data의 부정확성
        - 오타가 많음 (키 15cm, 몸무게 1000 kg)
        - 진단명에 아주 일부만 나타내어 기술함.
        - Post-annotation이 반드시 필요한 상황
- Facts & Limitations on Claim Data
    - 실제 환자 정보와 불일치 : 오직 돈을 벌기 위한 진단명으로 변경 처방 (청구용)
    - 아주 일부 데이터만 남아있음. (돈을 청구하기 위한 최소 데이터만 남아있음.)
- Practical Direction of AI for Healthcare
    - Images/Video에 집중
        - 의사들 중 영상의학과 교수들이 가장 컴퓨터에 익숙함
        - Blackbox algorithm에 대한 거부감이 적음 (어차피, 통계기법은 의미 없음)
        - Deep Learning이 가장 잘 할수 있는 분야!!!
    - 해석이 가능한 모델부터 시작
        - Decision tree와 같이 의료진이 쉽게 이해할 수 있는 모델 부터 접근
        - 의사들은 통계분석을 선호
        - 본인들이 이해하지 못하는 AI기법에 대해 거부감이 큼
    - PGHD가 새로운 대안 **
        - Clinical data는 구하는 것이 어려움. Cliam Data는 큰 쓸모가 없음. (비협조적)
        - 정확도는 낮을 수 있으나 continuous monitoring이 가능.
    - 의료기기 log data를 분석 **
        - 중환자실의 환자감시장치의 경우 생체신호를 real-time으로 생성 --> 하지만, alarm 용도로 만 쓰고 저장 및 분석하지 않음.
        - 신생아실에서의 인큐베이터 같은 경우 --> 저장 / 분석하지 않음.
    - SNS 분석
        - Influenza outbreak 예측, 자살 예측, ...
- 뛰어넘어야 할 허들
    - 사용하기 전에 임상실험이 필요함
    - CDSS(Clinical Decision Support System)의 경우 식품의약안전처 승인이 필요함
    - Common Data model : 국내의 경우 data integration를 위한 표준 준수가 미비
    - Clinical Data 활용을 위해서는 개인동의 필요 --> 익명화가 대안!!
- Health IT 엔지니어
    - IT + BT + HT
    - 의학용어, Health IT 표준 Review

    - 의료진과의 Communication !!!
    - NLP를 이용한 Post-processing보다 SDE(Structured Data Entry)를 통한 입력시점부터 자료 구조화 필요
    - Cleansing, Cleansing, Cleansing !!!
    - Annotating
    - PMI-cohort 프로그램
- 의료와 관련된 개인정보보호 법률 체계
    - 개인정보보호법 (공공/민간분야에 모두 적용) --> 신체적정보 (신체정보, 의료/건강정보) !!!
    - 생명윤리 및 안전에 대한 관한 법률 (15.12.29) --> 인간연구대상 및 인체유래물연구에 대한 윤리심사 의무화, 익명화 처리후 이용
    - 의료법 (환자 진료내역, 병력 등에 대한 정보보호) --> 비밀누설 금지, 기록열람 제한
    - 건강검진기본법
    - 국민건강보험법
    - 보건의료기본법 :
- 개인정보보호법 •2011년 9월 30일 시행 •2014년 8월 7일 개정안 시행 •개인정보 이용 제한 •제18조 2항 “통계작성 및 학술연구 등의 목적을 위하여 필요한 경우로서 특정 개인을 알아볼 수 없는 형태로 개인정보를 제공하는 경우” 예외로 인정 •제24조 고유식별정보* 사용 원칙적 금지 • 제24조 2항 고유식별정보 중 주민등록번호는 다음의 경우를 제외하고는 처리 금지 •법령에서 구체적으로 주민등록번호의 처리를 요구 또는 허용한 경우 (예: 의료법) •정보주체 또는 제3자의 급박한 생명, 신체, 재산의 이익을 위한 경우 •안전행정부 장관이 고시하는 경우 •보유하고 있는 주민등록번호는 법령에 구체적인 근거가 없는 경우 법 시행 후 2년 이내 파기 (’16년 8월까지 파기)
- 생명윤리 및 안전에 대한 법률 •2015년 12월 29일 일부개정안 시행 •개인정보 이용 제한 •제18조, 제38조, 제43조: 연구를 위해서 반드시 개별 동의를 받거나 또는 익명화 처리를 하여 기관위원회의 심의를 받아야 함 •인간대상연구 •사람을 대상으로 물리적으로 개입하는 연구 •설문조사, 행동관찰 등으로 자료를 얻는 연구 •개인을 식별할 수 있는 정보를 이용하는 연구 •인체유래물연구
- 의료기관의 개인정보
    - 환자로 부터 수집한 정보 : 성명, 성별, 주소, 직업, 주민등록번호, 전화번호, 휴대폰번호, 이메일주소, 보호자 성명 및 주소, 병력 및 가족력, 주된 증상 등
    - 의료인으로부터 생성되는 정보 : 의료행위에 따른 결과 및 의견정보인 진단결과, 검사결과, 진료경과 및 예견, 의료행위의 내용, 의료인의 성명, 종별면허번호, 서명날인, 진료일시 등
- 의료정보 활용
    - 일차적 이용: 동의 획득 필요 없음. (진료 예약, 진단, 검사, 치료 등), 병원 경영
    - 이차적 이용: 개별 동의 필요, 또는 익명화!! (연구, 홍보, 환자 교육 등) --> 연구용 의료 정보 활용이 이에 해당됨. --> 모든 연구 (인간연구대상 & 인체유래물연구)는 IRB의 심의를 받아야 함. ** IRB(Institutional Review Board: 의학연구윤리심의위원회) 임상연구에 참여하는 연구대상자의 권리・안전・복지를 위하여 인간을 대상으로 하는 모든 생명의과학연구의 윤리적, 과학적 측면을 심의하여, 연구계획을 승인할 수 있는 독립된 합의제 의결기구이다
- 익명화 (De-identification, Anonymization, Pseduonymization)

    - ISO/TS 252237 : Health Informatics - Pseduonymization
    - IHE IT : De-identification
        - http://ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_Handbook_De-Identification_Rev1.1_2014-06-06.pdf
    - NIST 8053 : De-identification Personal Information
        - http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

- 비식별화 혹은 익명화를 하더라도 상업적 목적은 불가 ??

    - 개인정보 활용방법: 비식별화 부분 (제 7조 민감정보 생성의 금지)
    - 빅데이터 활용을 위한 개인정보 비식별화 기술 활용 안내서 (v 1.0)
        - https://kbig.kr/?q=지식자료실/15596
    - 개인정보 비식별화에 대한 적정성 자율평가 안내서
        - http://www.privacy.go.kr/inf/rfr/selectBoardArticle.do?nttId=5951&bbsId=BBSMSTR_000000000044

Personalized vs. Precision Medical

Precision Medicine (정밀의료)

- 용어: 개인 유전자 정보에 진료 정보, 생활습관 정보를 하나로 모아 개인의 건강을 예측하는 서비스

CDSS

http://www.yoonsupchoi.com/tag/%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5/

헬스케어 이노베이션 / Yoon Sup Choi

IT와 헬스케어의 컨버젼스를 통해 사회적 가치를 창출하는 것을 화두로 삼고 있는 융합 생명과학자, 미래의료학자, 블로거, 작가입니다. 포항공대에서 컴퓨터공학과 생명과학을 복수 전공하였고, 동대학원에서 전산생물학으로 이학박사를 취득했습니다. Stanford University 방문연구원, 서울대학교 의과대학 암연구소 연구조교수, KT종합기술원 컨버전스연구소 팀장, 서울대병원 의생명연구소 연구조교수 등을 역임하였습니다. 현재 최윤섭 디지털 헬스케어 연구소의 소장이며, 국내 유일의 헬스케어 전문 스타트업 엑셀러레이터 디지털 헬스케어 파트너스(DHP)의 대표 파트너를 맡고 있습니다. VUNO, Zikto, Promisope, Souling, HB 인베스트먼트, 녹십자 홀딩스의 자문이며, 매일경제신문의 필진입니다

디지털 헬스케어 관련 클라우드 플랫폼 구축 본격화

인텔(Intel), 미국 오레곤 의과대학 연구소와 의료용 클라우드 플랫폼 공동 개발 착수 ▶ 글로벌 칩셋 제조사 인텔(Intel)이 인텔 개발자 포럼(Intel Developer Forum)에서 암환자의 유전자 및임상 데이터 공유가 가능한 클라우드 플랫폼을 발표하여 주목('15.8.19.) • 해당 플랫폼의 명칭은 'CCC(Collaborative Cancer Cloud)'로 인텔과 미국의 오레곤 의과대학 산하 나이트 암 연구소(Knight Cancer Institute)가 공동 개발 중 • 인텔에 따르면 'CCC'는 병원 및 연구기관에서 환자의 유전자 정보, 의학용 시각 자료, 임상 데이터 등을 안전하게 공유할 수 있도록 지원하는 클라우드 시스템 • 인텔은 2016년 1분기, 해당 플랫폼에서 사용 중인 핵심 기술 중 하나를 개발자 커뮤니티에 오픈 소스로 공개하고 이후 나머지 기술 역시 전부 공개하면서 관련 개발자 양성에 주력할 것이라고 언급

▶ 인텔은 유전자 배열 분석 에서 발생하는 대규모의 데이터 자원을 'CCC'를 통해 수집·공유하여 맞춤형 치료를

구현할 수 있을 것으로 예상 • 나이트 암 연구소 소장 브라이언 드러커는 미래 의학의 핵심 과제는 모든 질병의 유전자 배열을 분석하는 것에 있으며 이를 실현하기 위해서는 'CCC'와 같은 원활한 데이터 공유 플랫폼이 요구됨을 강조 • 드러커 소장은 CCC가 완전 상용화될 경우 현재 2%의 환자들만이 누리는 유전자 맞춤형 치료를 2020년까지 보편화할 수 있을 것으로 기대 • 이를 실현하기 위해 나이트 암 연구소와 인텔은 2016년 1분기 내에 2개 이상의 새로운 연구기관과의 협약을 체결할 예정

디지털 헬스케어 산업 내 경쟁 초점, 건강 데이터의 공유 및 분석 가능한 클라우드 솔루션으로 이동 ▶ 인텔에 앞서 IBM 역시 의료 영상 분석 플랫폼 머지 헬스케어(Merge Healthcare)를 7억 달러(8,200억 원)에 인수하며 클라우드 기반의 대규모 헬스케어 데이터 분석 시스템 구축에 착수('15.8.12.) • 머지 헬스케어는 엑스레이(X-ray), MRI 스캔, 컴퓨터 단층촬영(CT) 사진 등 대규모의 의료용 시각 데이터 보유 업체 IBM은 머지 헬스케어의 데이터를 자사의 클라우드 인프라 및 슈퍼컴퓨터 '왓슨(Watson)1'과 결합하여 각종 질병 분석을 실시할 예정 ▶ 인텔, IBM 등 디지털 헬스케어 사업에 관심을 표명하고 있는 주요 사업자들이 클라우드 시스템 구축에 주력하는 이유는, 맞춤형 의료 실현의 기초 인프라로서 건강 데이터의 원활한 공유·분석이 필수이기 때문 • 웨어러블 단말의 보급, 사물 인터넷의 대중화 등으로 건강 데이터를 수집할 수 있는 채널은 빠르게 증가하고 있는 추세 • 그러나 현재 수집된 데이터들은 여전히 개별적으로 측정 및 보관되고 있으며, 이는 질병 등 건강 관련 정보의 단편적인 측면만을 보여준다는 한계를 보유 • 반면, 인텔, IBM 등의 헬스케어 클라우드 시스템이 현실화 될 경우, 수집한 데이터를 실시간으로 통합하여 건강 상태에 대한 전반적인 정보를 파악하고 이에 따라 맞춤형 진료를 실시하는 것이 가능 한편, 관련 사업자의 수요에 힘입어 헬스케어 클라우드 컴퓨팅 시장은 폭발적인 성장을 거듭할 것으로 보이며, 기술 진화 속도 역시 가속화 될 전망 • 시장조사기관 마켓앤마켓(MarketsAndMarkets)의 조사에 따르면, 2015년 현재 37억 달러(4조3,771억원) 수준인 헬스케어 클라우드 컴퓨팅 시장 규모는 연평균 20.5%의 성장률을 기록하며 2020년 95억 달러(11조 2,385억 원)까지 성장할 전망('15.6.) • 또한 월 스트리트 저널(The Wall Street Journal)은 헬스케어 클라우드 시스템에 딥러닝(Deep Learning) 기술이 접목되면서 맞춤형 의료 서비스에 획기적인 진화가 발생할 것이라고 언급('15.8.11.) • 질병 패턴 분석 등 딥러닝 기술의 주요 결과물이 성과를 나타내기 위해서는 대량의 데이터 풀(pool)이 필수라는 점에서 클라우드 시스템과의 시너지 효과가 상당할 것이라는 의견

-- 아산병원, 신수용 교수

https://www.youtube.com/watch?v=pFUKQ2eNJGA

[AI for Healthcare]

◦ Healthcare의 데이터들의 Source는 의료기관(Clinical data), 공공기관(Claim 데이터), 공공기관(기상청, 환경부 등), 환자가 자가 생성한 개인데이터(Patient-Generated Health Data와 구매목록이나 SNS Data와 같은 Lifelog)

◦ Healthcare의 데이터 형태는 Text, Image, Video, Code, Sound 등으로 구성

◦ 2011년 기준, 전 세계의 의료 데이터는 500 petabytes이며 2020년 기준 25000 petabytes가 될 것으로 추정됨

◦ IBM 자료에 의하면 Clinical data가 10%에 불과하고 유전체 데이터는 30%, 나머지 60%가 일생동안 생성하는 외부(라이프)데이터임

◦ 국내 EMR 보급률은 92% 이상이나 실제적으로는 부분적 도입(전체의 34.1%)과 영상 EMR(종이로 쓰고 스캔하여 저장)을 사용한 부분 등을 포함된 수치임

◦ 현재 EMR은 Word Processor수준으로 대부분 Text, 수많이 많은 약어, 복잡하고 전문적 의학용어와 약어의 경우, 의사마다 진료과마다 다름(통일성 zero)

◦ NLP(자연어처리)는 미국/외국의 경우는 가능하나, 한국은 통일되지 않은 약어, 전문의학용어, 문장이 아닌 phrase 형태 등의 이슈로 현재 수준으로는 잠정적으로 어려움

◦ Regular expression은 잘됨, 일부 검사보고서(병리보고서)는 semi-structure로 패턴만 알면 99%정도의 정확도로 data추출이 가능하나 의료진의 높은 기대치가 제약사항

◦ Clinical data의 정확성이 보장되지 않음- 타과 진료 시 진단명이 정확히 기재된 환자가 일부분으로 Post-annotation이 반드시 필요

◦ Claim data의 경우 전국민 데이터라는 상징성은 존재하나 돈을 받기 위한 청구용 데이터로서 실제 환자 진료 데이터와 불일치되거나 일부 데이터만 존재

◦ 컴퓨터에 익숙한 영상의학과 교수들을 대상으로 딥러닝이 가장 잘 할수 있는 Image/Video에 집중하는 것이 효율적

◦ 데이터 확보는 Clinical data가 얻기 어려우므로 환자가 생성한 PGHD(Patient-Generated Health Data) 역시 새로운 대안이 될 수 있음

- POCT(Point-of-care testing) 장비들의 발전으로 손쉽게 데이터 획득 가능하며 지속적 모니터링이 가능하다는 장점이 존재

◦ 극복해야 할 장벽으로는 진료현장에서 사용하기 위한 임상시험/CDSS의 경우, 식품의약안전처의 승인이 필요/data integration을 위한 표준 준수 미비/Clinical data 활용을 위한 개인동의 필요 등이 있음

◦ 의사는 면허직이므로 인공지능 의사에게 라이선스를 줄 것인지? 의료 사고시 책임은 누가 질 것인지에 대한 고민이 필요

[Privacy Protection for Healthcare]

◦ 의료와 관련된 개인정보보호 법률체계는 개인정보보호법, 의료법, 건강검진기보넙, 국민건강보험법, 보건의료기본법, 응급의료에 관한법률, 생명윤리 및 안전에 관한 법률 등 다수 존재

◦ 개인정보 vs. 민감정보 vs. 고유식별정보

- 집에서 잰 키/몸무게가 민감정보인가? 명확히 규정할 수 없는 것이 현실

◦ 개인의 정보를 활용하여 연구를 진행하기 위해서는 개인정보보호법과 생명윤리 및 안전에 관한 법률을 따라야 함

- 개인정보보호법에 의해서 환자 정보를 사용하기 위해서는 동의를 받아야 함, 대규모 사용을 위해서는 익명화가 답
- 생명윤리 및 안전에 관한 법률에 따르면 사람을 대상으로 한 물리적 개입(약 처방 등), 설문조사/행동관찰 등으로 자료를 얻는 연구 등도 심의를 받아야 함

◦ “개인의료정보”가 무엇인지에 대한 명확한 정의가 없음

◦ 연구용으로 활용하기 위해서는 비식별화가 필요하나 “무엇을”,“어떻게” 해야 할지는 명확하지 않음 (즉, 무엇이 개인을 식별할 수 있는 정보인지 명확하지 않음)

◦ 개인식별 정보를 구분하자면 직접적 개인식별정보/조합을 통해 개인식별이 가능한 정보/간접적 개인식별 정보로 구분 가능

◦ 의료용 개인정보 보호은 미국의 HIPAA가 대표적이며 의료 정보의 18가지 정의와 가이드라인 존재

◦ “개인정보 활용 방법:비식별화”에 의해 의료 데이터의 삭제/총계처리/범주화 등을 진행하면 개인의 개별 데이

터의 중요성이 사라짐

◦ 비식별화 혹은 익명화를 하더라도 상업적 목적의 활용은 불가함

◦ 결국 Healthcare data중에 Personal Healthcare data를 구분하여 대상 정보만 비식별화 처리를 하는 것이 중요함

AI기반의 바이오 헬스

헬스케어 스타트업(신생 벤처기업) 뷰노코리아는 인공지능 기술을 의료에 적용해 '골 연령 판독 보조 소프트웨어'를 개발했다. 의사가 환자의 뼈 나이를 알려면 골밀도측정기로 환자의 뼛 속 밀도를 확인한다. 그 다음 환자의 증상, 자세, 관절 위치에 따라 종합적으로 뼈 나이를 판단한다.

이 회사는 골밀도 측정결과와 환자 상태에 따른 뼈 나이 정보를 대량으로 컴퓨터에 학습시켰다. 의사가 골밀도 측정기로 환자의 뼈를 촬영하면, 이 솔루션이 일차적으로 환자의 예상 뼈 나이를 알려준다.

원문보기:
http://biz.chosun.com/site/data/html_dir/2016/03/30/2016033002398.html#csidx1e312388c74adfeb774743093baa0f6

연구중심병원 활성화

바이오헬스 산업이 성장하려면 병원이 진료 중심에서 '연구 중심'으로 바뀌어야 한다. 의사가 연구개발(R&D)에 많이 참여할수록 의약품, 의료기기 등 바이오헬스 산업이 활성화될 수 있다.

서울대병원, 서울아산병원, 세브란스병원 등 10개 병원은 2013년부터 3년간 '1차 연구중심병원'으로 활동했다. 복지부 조사결과에 따르면, 10개 연구중심병원에서 연구에 전담하는 의사수는 2013년 76명에서 올해 3월 현재 174명으로 123% 늘었다. 특허 등록 등 지적재산권을 획득한 연구도 2013년 745건에서 올해 1926건으로 159% 급증했다. 병원이 기업에 기술이전으로 벌어들인 수입액은 2013년 24억원에서 올해 93억원으로 282% 증가했다.

원문보기:
http://biz.chosun.com/site/data/html_dir/2016/03/30/2016033002398.html#csidxd49b1e728a51fa2b7ac0576843e3423