Characteristics of Clinical Data: AI Perspective

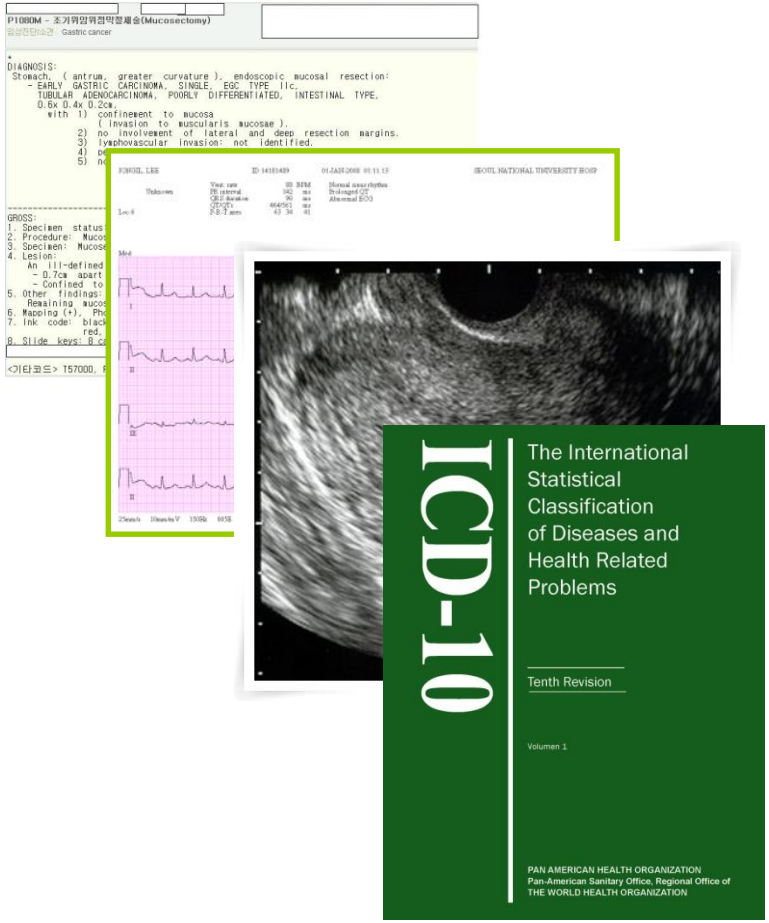
Soo-Yong Shin, Ph.D.
Department of Biomedical Informatics
Asan Medical Center

2016.6.29
@KCC 2016

ASAN FOUNDATION 아산재단
ASAN Medical Center

Types of Healthcare Data

- Text
- Image
- Video
- Code
- Sound
- …



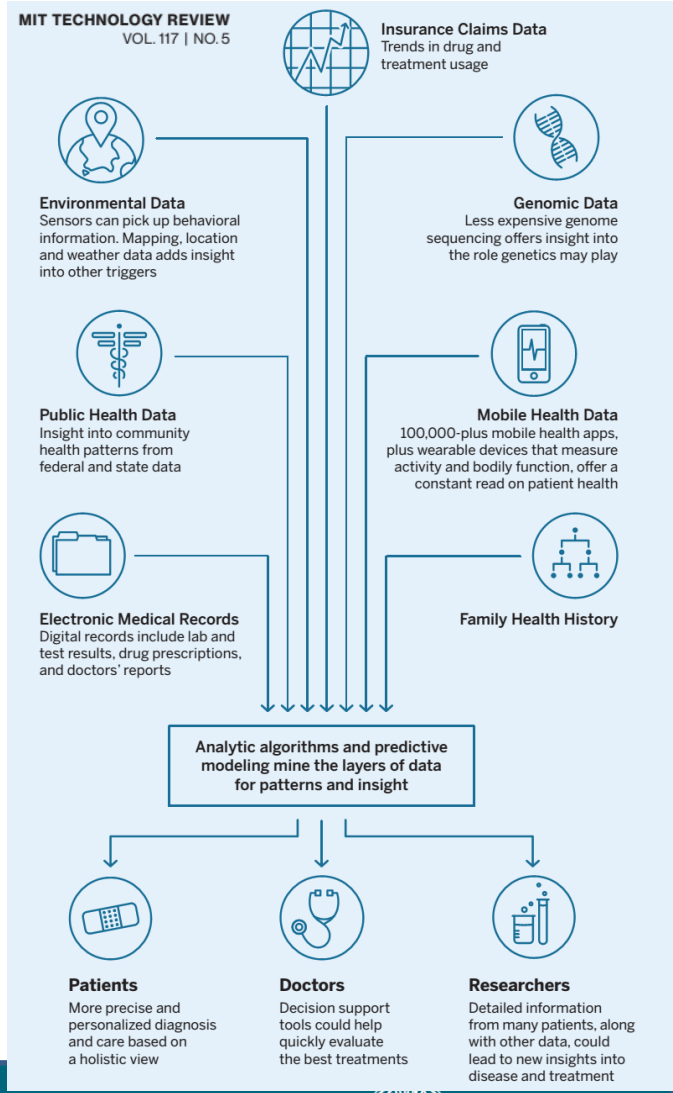
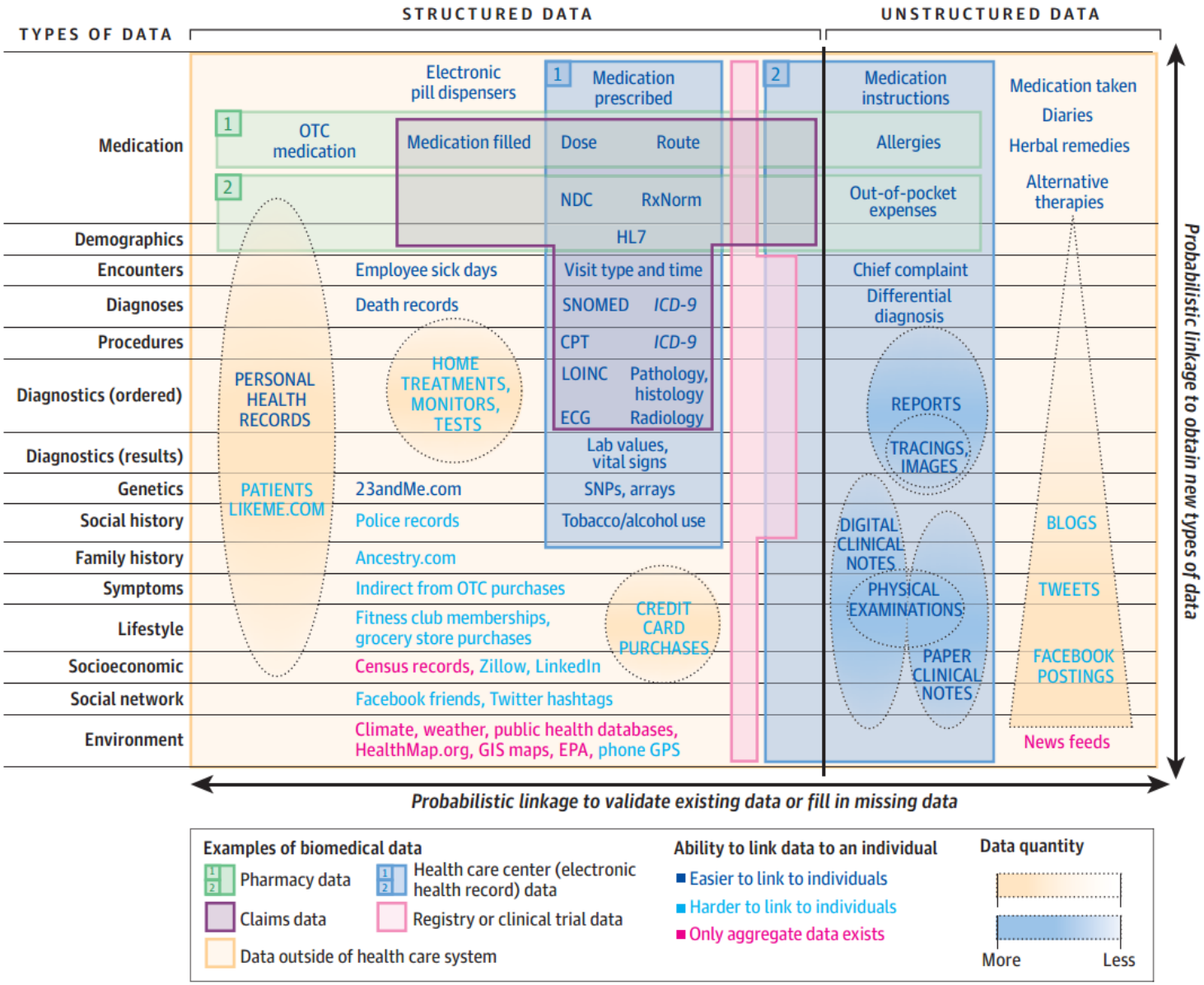


Figure. The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health Care



CPT indicates current procedural terminology; ECG, electrocardiography; EPA, US Environmental Protection Agency; GIS, geographic information systems; GPS, global positioning system; HL7, Health Level 7 coding standard; *ICD-9*, *International Classification of Diseases, Ninth Revision*; LOINC, Logical Observation Identifiers Names and Codes; NDC, National Drug Code; OTC, over-the-counter; SNOMED, Systematized Nomenclature of Medicine; SNP, single-nucleotide polymorphism.

http://www.ncbi.nlm.nih.gov/pubmed/24854141

Size of Individual Healthcare Data

Sources of Healthcare Data

의료기관
(clinical data)

기상청 Korea Meteorological Administration

환경부

공공기관
(environmental data)

국민건강보험공단
National Health Insurance Corporation

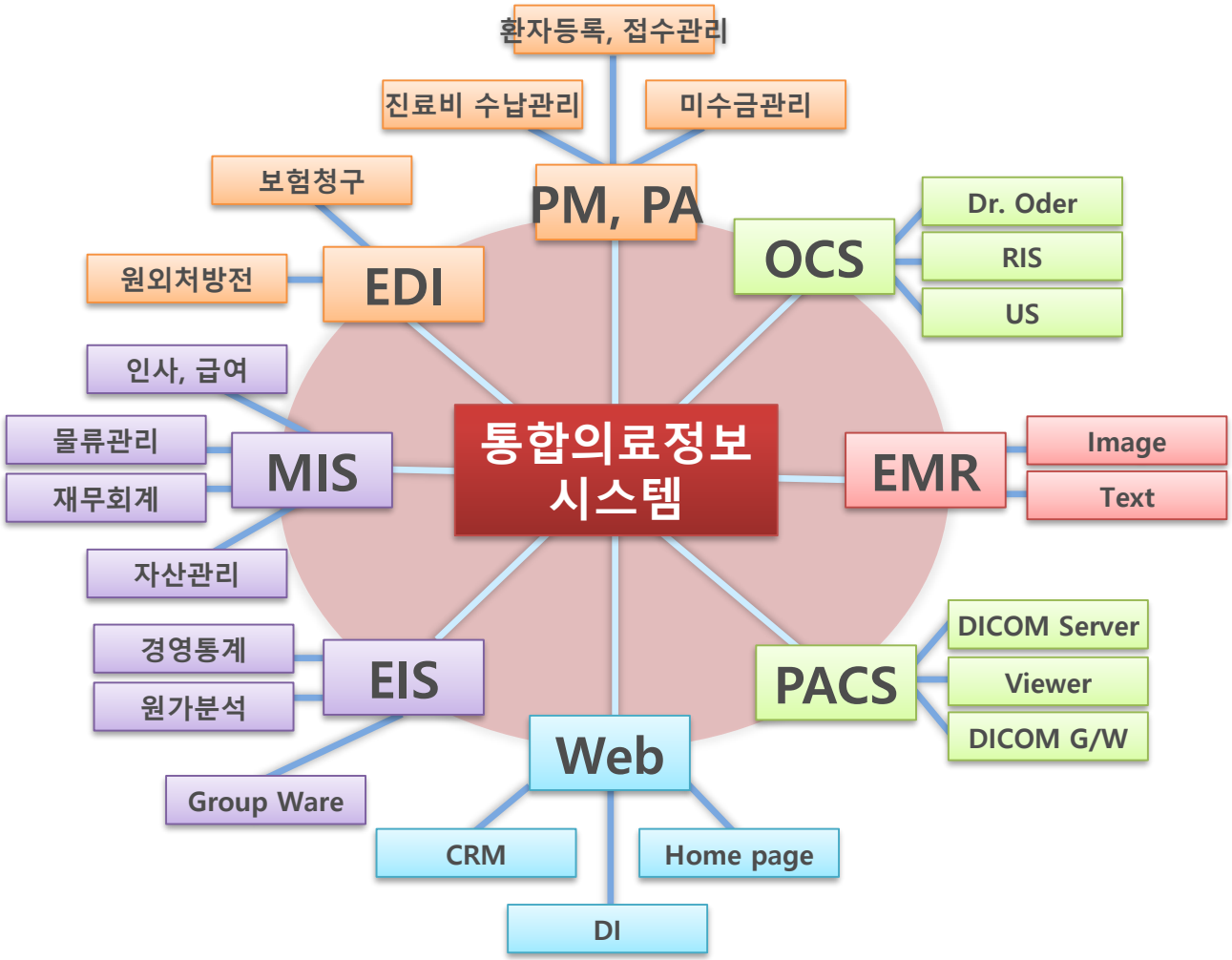
H I R A
건강보험심사평가원
Health Insurance Review & Assessment Service

공공기관
(claim data)

제약회사
연구기관
(R&D data)

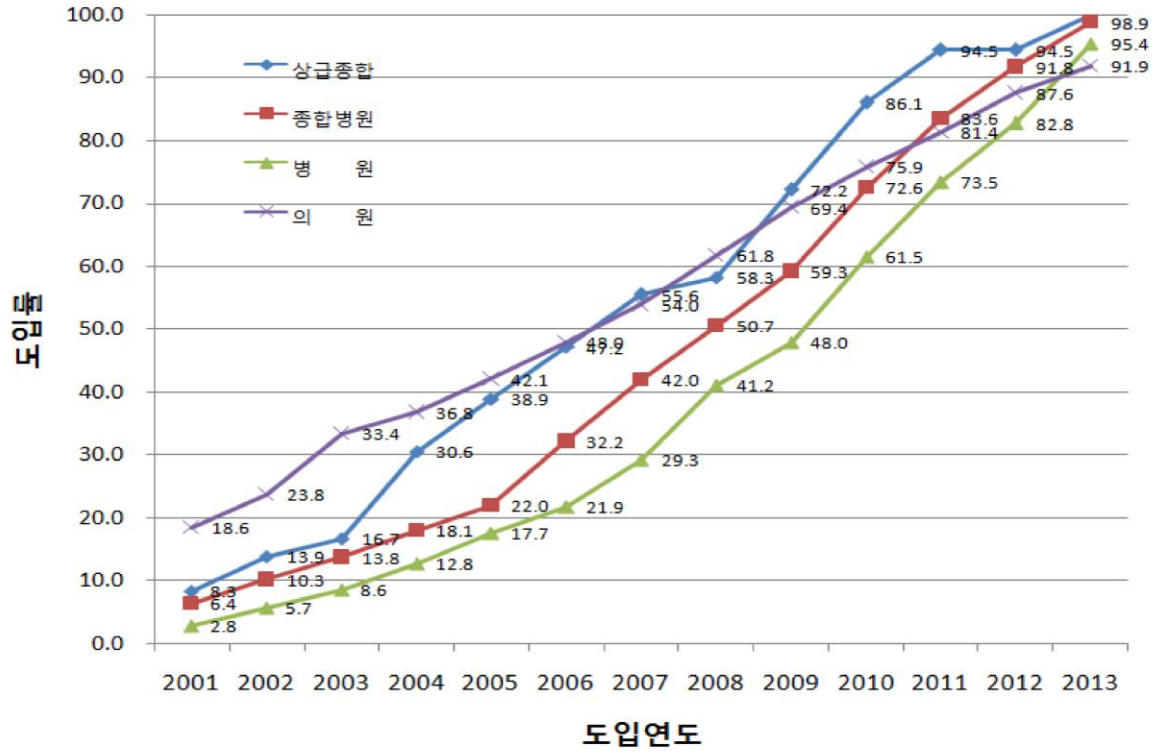
개인
(Wearable,
home-monitoring devices,
SNS, purchase data)

Hospital Information System



Facts: 전산화 정도

- 국내 EMR 보급률은?



Facts: 전산화 정도

- 과연? (2014년 설문조사결과)

(단위: 개, %)

구 분	상급종합		종합병원		병원		의원		치과병원		치과의원	
	기관수	%	기관수	%	기관수	%	기관수	%	기관수	%	기관수	%
도 입	32	84.2	193	72.6	183	63.5	614	67.8	28	56.0	175	40.0
부분도입	6	15.8	70	26.3	93	32.3	253	27.9	15	30.0	198	45.3
도입안함	0	0.0	3	1.1	12	4.2	39	4.3	7	14.0	64	14.6
전 체	38	100.0	266	100	288	100	906	100	50	100	437	100.0

  - Comprehensive EMR 비율은 더 낮음!

Facts: 전산화 정도

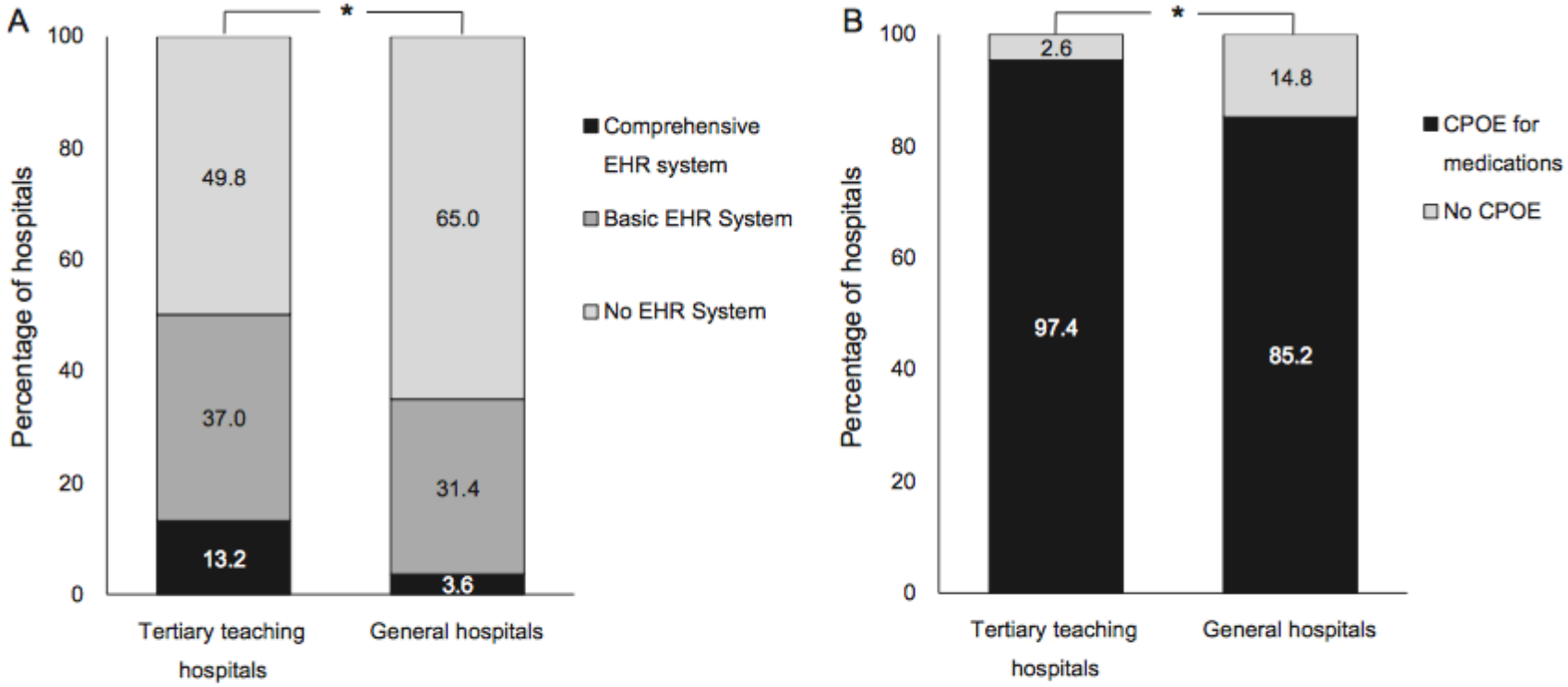


Fig. 1 – A: Levels at which EHRs were adopted by tertiary teaching and general hospitals. B: Levels at which CPOE for medications was adopted by tertiary teaching hospitals and general hospitals. Tertiary teaching hospitals: $n = 44$, general hospitals: $n = 269$. $^{*}p < 0.05$.

Int J Med Inform. 2012 Mar;81(3):196-203.

Facts: 의료정보화

MAIN GOAL: 4LESS HOSPITAL!

- 기존 의무기록 작성
  - 수기로 작성
- IT와의 결합
  - 컴퓨터로 입력 ⇒ 의료정보화?!

Limitations

- 현재 EMR은 **Word Processor**
  - 대부분 Text
  - 수없이 많은 약어
    - 심지어, 의사마다 진료과마다 다름 (통일성 zero)
  - 복잡하고 전문적인 의학 용어

Limitations: NLP

NLP..?

- In US or other foreign countries
  - YES!
- In Korea
  - Probably **NO**!!
  - 통일되지 않은 약어
  - 전문 의학 용어
  - phrase not sentence
  - Bilingual (Korean + English) + symbols..

Limitation: Examples from Clinical notes

경기 용인시;

◈Subject # Hypothyroidism (1981 synthroid 복용중) # Rt breast cancer s/p MRM (2003/4/18) pT1(1.3cm)N0M0, ER/PR -/-, c-erbB2 (3/3) s/p adjuvant AC #4 doing well intermittent anal bleeding

◈Object CBC & chem B: OK P/Ex: no palpable mass Bone scan & CXR & MMG: OK

◈Assessment NED 3 year 8 month

◈Plan 1. RTC 6 month later with CBC, chem B, breast US

Limitations: NLP

- **Personal experiences**
  - 차라리 Regular expression이 잘 됨
  - 일부 검사보고서 (병리보고서)는 semi-structure
    - 패턴만 알면 99% 정도의 정확도로 data 추출 가능
  - NER (Named Entity Recognition)도 regular expression으로 대부분 가능
  - BUT, 의료진의 높은 기대치가 제약사항.
    - Precision > 95%, Recall > 90% (is it possible?)

Limitations: incomplete data of EMR

- Clinical data의 정확성?
  - **NO**!
  - 성인 키가 15cm, 몸무게가 1,000kg… (오타!)
  - 병원의 진단명은 환자의 Phenotype 중 아주 일부분만 나타냄
    - Ex) 당뇨병 환자 중 정말 당뇨병 진단명이 붙은 환자는 ?
  - Post-annotation이 반드시 필요
  - Phenotyping!

Limitations: Claim data

- 실제 환자 진료 데이터랑 불일치
  - 왜냐하면 claim data는 돈을 받기 위한 청구용 data
  - 돈을 벌기 위해서 진단명이 변경됨
- 아주 일부 데이터만 있음
  - 진단명, 투약력, 검사 시행 여부, 극히 일부 검사 결과 (돈을 받기 위한) 등

Practical Direction of AI for Healthcare

- **Images/Video**에 집중!
  - 의사들 중 가장 컴퓨터에 익숙한 그룹이 영상을 판독하는 영상의학과 교수들
  - Blackbox algorithm에 대한 거부감이 적음. (어차피 통계기법으로 안 되기 때문에)
  - Deep learning이 가장 잘 할 수 있는 분야

Practical Direction of AI for Healthcare

- **해석이 가능한 모델부터 시작!**
  - Decision tree와 같이 의료진이 이해할 수 있는 model부터 접근
  - 의사들은 통계분석을 선호함
  - 본인들이 이해할 수 없는 AI기법 *(통계도 이해하는지 의문이지만..)*에 대한 거부감이 큼
  - BUT, AlphaGo..?!

Practical Direction of AI for Healthcare

- **PGHD**도 새로운 대안
  - Clinical data는 구하는 것이 어려움
  - IRB, 의료진의 비협조, ….
  - POCT 장비들의 발전으로 손쉽게 Data 획득 가능
  - 단, 정확도는 낮아서 의료용으로는 쓰기 어려움
  - 하지만, continuous monitoring이 가능

Practical Direction of AI for Healthcare

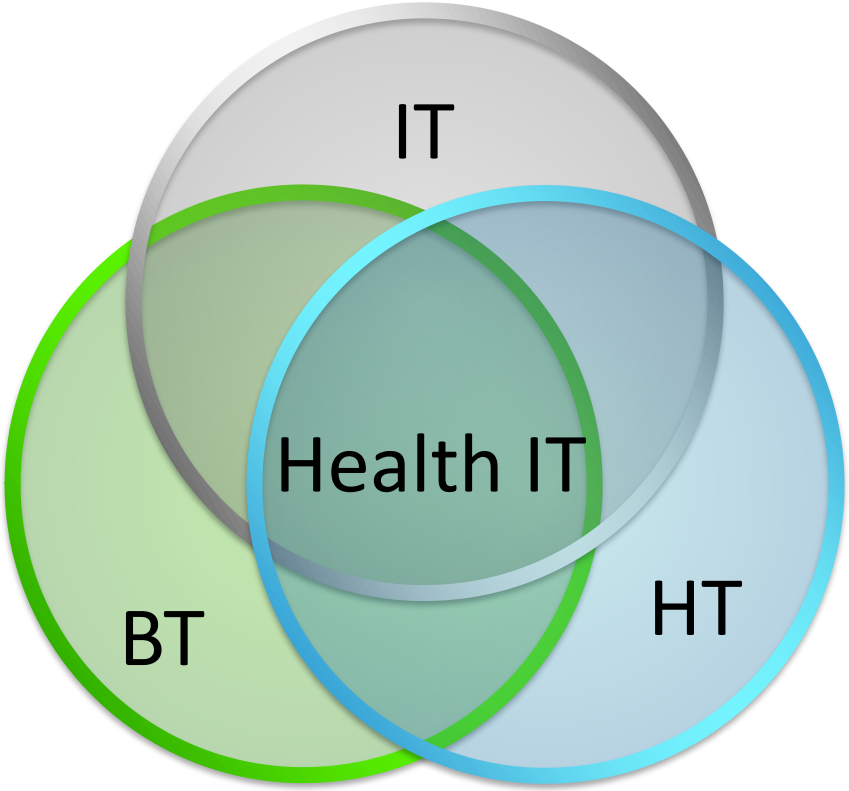
- **의료기기 log data**를 분석!
  - 중환자실의 환자감시장치의 경우 생체신호를 real-time으로 생성
  - 하지만, 의료 기관에서는 무시되고 있음. 오로지 alarm용으로만 활용
  - 질환 사전 예측이 가능
- **SNS** 분석
  - Influenza outbreak 예측, 자살 예측…

Remaining Hurdles

- 진료현장에 쓰기 위해서는 **임상시험이 필요**
- CDSS의 경우 **식품의약안전처 승인이 필요**함!
- Common data model
  - 국내의 경우 **data integration을 위한 표준 준수 미비**
- Clinical data를 활용하기 위해서는 **개인 동의 필요**
  - 익명화가 대안

More Practical Hurdles

- Health IT?



More Practical Hurdles

- 의학용어, Health IT 표준 review

More Practical Hurdles

- 의료진과의 COMMUNICATION!
  - 화성에서 온 남자 금성에서 온 여자

sooyong.shin@amc.seoul.kr

@likesky3

ASAN FOUNDATION 아산재단

ASAN Medical Center