빅데이터를 이용한 폐암 임상연구 방안

original article

Annals of Oncology
doi:10.1093/annonc/mdr606

Proteins involved in DNA damage response pathways and survival of stage I non-small-cell lung cancer patients

	ID	성명	Sex	Age	Diagnosis date (by Bx)	Smoking hx	smoking (pack-years)	FEV1(L)	FEV1(% predicted)	FVC(L)	FVC(% predicted)	FEV1/FVC (%)	DLco (%)	OP type	OP method	OP date	병리번호	Histologic type	Histologic type 분류	pathologic grade	Tumor location
1																					
29			1	59	2006-10-26	1	60	2.26	82	3.31	88	93	97	3	2	2006-11-06	06S 061925	squamous	2	2	Lt
30			1	55	2006-11-07	1	60	2.47	85	3.06	77	111		1	2	2006-11-23	06S 065968	squamous	2	2	RLL
31			1	72	2006-10-12	1	30	1.88	68	2.63	63	106	99	1	2	2006-11-27	06S 066493	squamous	2	2	LUL
32			1	60	2006-11-29	2	20	2.03	68	3.21	77	88	107	3	2	2006-12-14	06S 070698	squamous	2	3	LUL
33			1	76	2006-12-01	1	50	2.47	118	3.84	121	93	69	1	2	2006-12-27	06S 072979	squamous	2	2	LLL
34			2	60	2007-01-03	0		2.44	111	2.72	92	122	112	1	2	2007-01-11	07S 002443	adenoca	1	1	LUL
35			2	50	2007-01-17	0		2.69	115	3.42	111	105	116	1	2	2007-02-02	07S 007514	adenoca	1	1	RUL
36			1	66	2007-01-23	1	40	1.42	55	2.64	71	76	94	1	2	2007-02-08	07S 008827	squamous	2	2	RUL
37			2	49	2007-02-23	0		2.27	111	2.88	106	105	82	2	2	2007-03-08	07S 014415	squamous	2	2	RML
38			2	75	2007-03-09	0		1.51	95	2.07	88	103		1	1	2007-03-15	07S 016049	adenoca	1	2	LLL
39			1	64	2007-03-22	1	30-60	2.88	107	4.06	105	100	83	1	2	2007-04-05	07S 020876	squamous	2	2	RUL
40			1	51	2007-04-16	1	3	3.57	103	4.35	93	112	97	1	2	2007-04-19	07S 024195	adenoca	1	2	RML

characteristics of the patients

Table 1 shows the clinical characteristics of the 889 patients (401 squamous cell carcinoma and 488 adenocarcinoma cases)

통계용어 완전정복!!

전수조사 vs 표본조사

전수조사란 집단을 이루는 모든 개체들을 조사하여 모집단의 특성을 측정하는 방법

표본조사란 전체 모집단 중 일부를 선택하고 이로부터 전체 집단의 특성을 추정하는 방법

ELSEVIER

Lung Cancer 30 (2000) 15–22

LUNG CANCER

IASLC

www.elsevier.nl/locate/lungcan

Characteristics of lung cancer in Korea, 1997

Choon-Taek Lee [a,*], Kyung Ho Kang [a], Younsuck Koh [a], Joon Chang [a], Hee Soon Chung [a], Sue Kyung Park [b], Keun-Young Yoo [b], Jeong Sup Song [a]

[a] *Scientific Committee,[1] Korean Academy of Tuberculosis and Respiratory Diseases, 14 Woomyun-Dong, Sochou-Gu, Seoul 137-140, South Korea*

[b] *Department of Preventive Medicine, Seoul National University College of Medicine, 28 Yongon-Dong, Chongno-Gu, Seoul 110-744, South Korea*

Received 9 November 1999; received in revised form 4 February 2000; accepted 17 February 2000

In 1997 - 3794 subjects were selected – Retrospective nationwide survey

Contents lists available at ScienceDirect

Lung Cancer

journal homepage: www.elsevier.com/locate/lungcan

Lung cancer patients who are asymptomatic at diagnosis show favorable prognosis: A Korean Lung Cancer Registry Study

Kwang-Ho In[a,b], Yong-Soo Kwon[c], In-Jae Oh[c], Kyu-Sik Kim[c], Maan-Hong Jung[a,d], Kwan-Ho Lee[a,e], Sun-Young Kim[a,f], Jeong-Seon Ryu[a,g], Sung-Yong Lee[a,b], Eun-Taik Jeong[a,h], Sang-Yeub Lee[a,b], Ho-Kee Yum[a,i], Chang-Geol Lee[a,j], Woo-Sung Kim[k], Jae-Ill Zo[l], Hojoong Kim[m], Young-Whan Kim[n], Se-Kyu Kim[j], Jae-Cheol Lee[o], Young-Chul Kim[a,c,*]

[a] *The Survey Committee of Korean Association for the Study of Lung Cancer, South Korea*

%	All patients
% (Number)	100 (8788)
Female	24.2
No smoking history	28.9
Adenocarcinoma	36.1
I–II/III/IV	25.0/34.5/40.6
Asymptomatic patients	6.5
Performance status (0–1)	75.0
Any treatment	73.4
Surgery	22.1
Radiation	7.8
Chemo-radiation	5.4
Chemotherapy	38.0
Median survival (95% CI)	28.0 m (26.5–29.5)

In 2005 8788 were selected – Retrospective nationwide survey

국방의료정보체계(DEMIS) : Defence Medical Information System

Pulmonary tuberculosis in young Korean soldiers: incidence, drug resistance and treatment outcomes

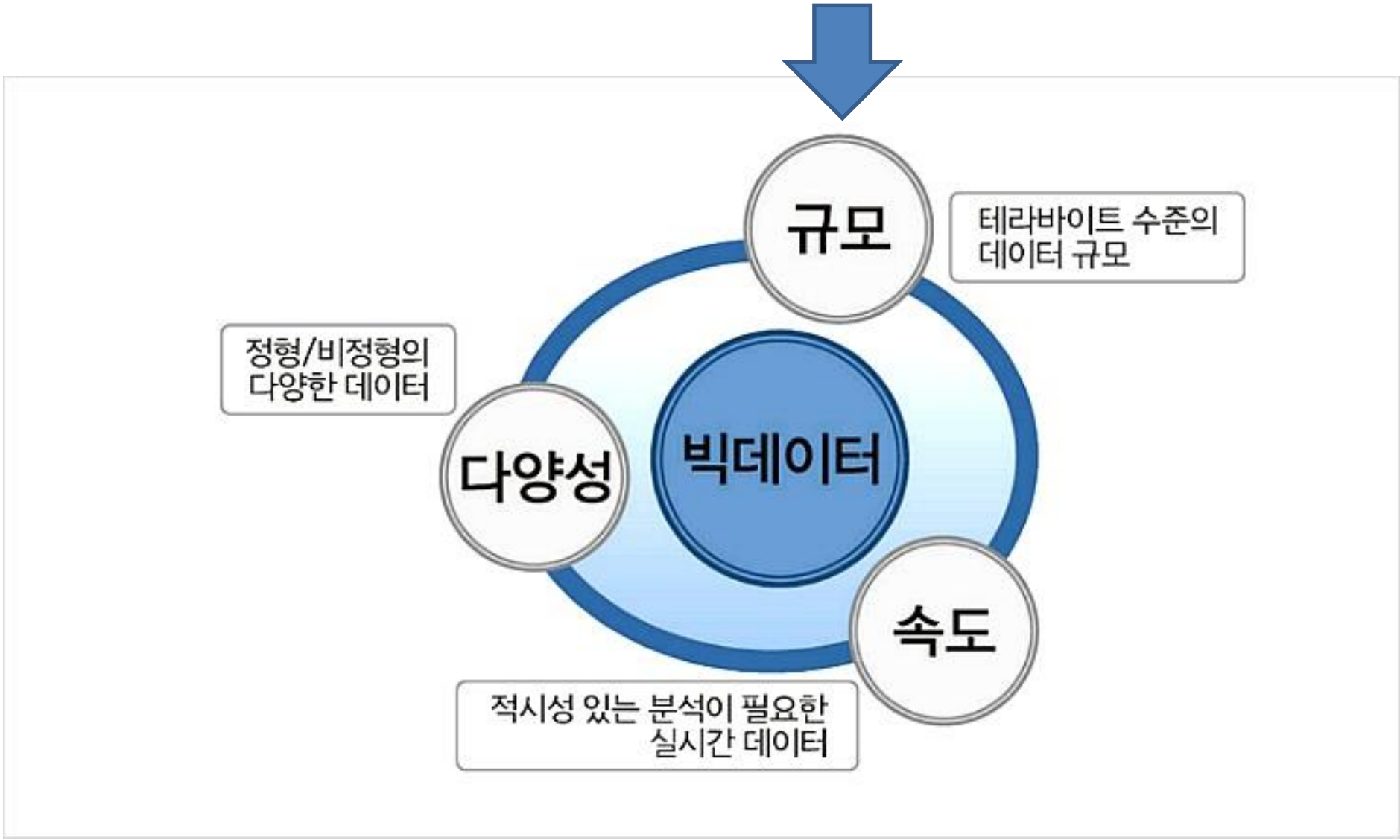
젊고 건강한 한국 군인들에서 대상포진의 발생률

Incidence and Seroprevalence of Hepatitis A Virus Infections among Young Korean Soldiers

Clinical and laboratory predictors of oliguric renal failure in haemorrhagic fever with renal syndrome caused by Hantaan virus

Tuberculosis among Dislocated North Koreans Entering Republic of Korea since 1999

Incidence of herpes zoster and seroprevalence of varicella-zoster virus in young adults of South Korea


규모
테라바이트 수준의
데이터 규모
정형/비정형의
다양한 데이터
다양성
빅데이터
속도
적시성 있는 분석이 필요한
실시간 데이터

건보공단, 10년 축적한 1조 3,034억 건의 빅데이터로 『건강보험 빅데이터 운영센터』 본격 가동

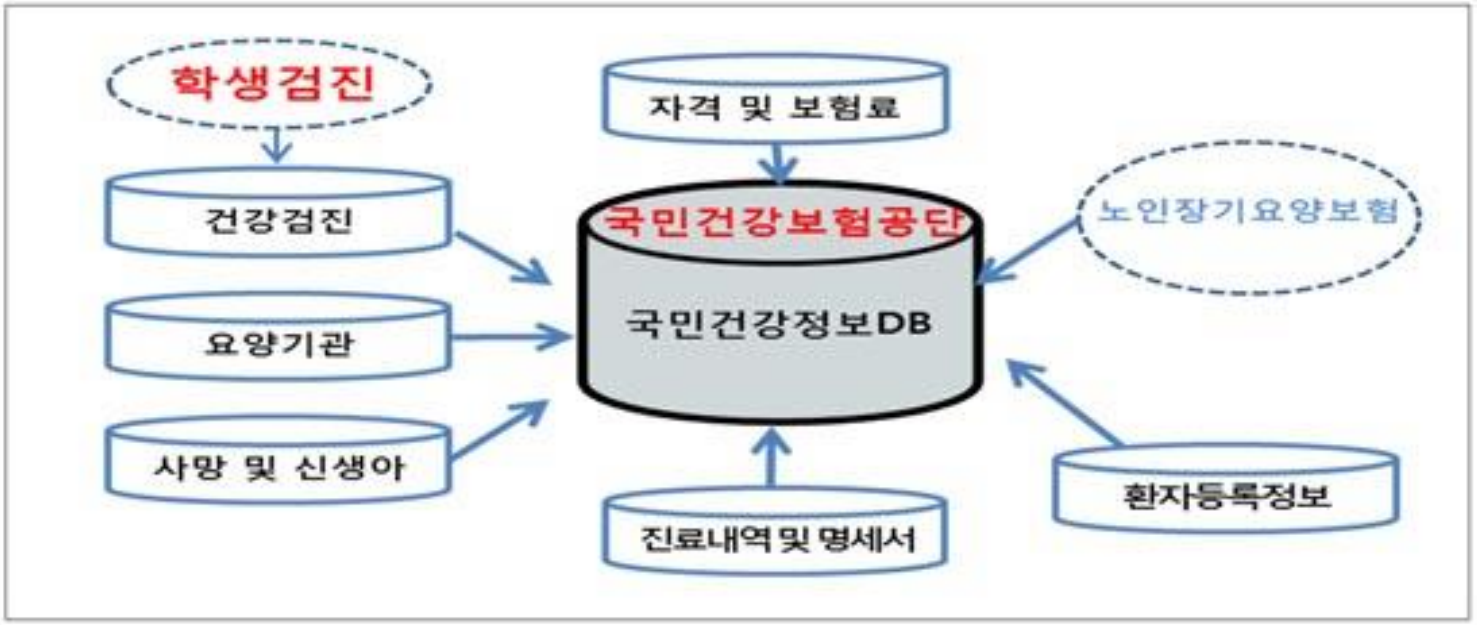
– '개인별 맞춤형 건강정보 서비스 제공', '정보의 공개·개방'으로 보건의료분야의 새로운 부가창출에 기여 –

☐ 국민건강보험공단(이사장 김종대)은 전국민 건강정보와 다양한 비정형 데이터를 융합한 빅데이터(거대자료)를 바탕으로 **개인별 평생 맞춤형 건강서비스를 제공**하고 관련정보를 공개·개방함으로써 **보건의료분야의 새로운 부가가치 창출**을 위하여 **'건강보험 빅데이터 운영센터'**를 **본격적으로 가동**한다고 밝혔다.

- **'빅데이터 운영센터'**는 '서비스개발팀', '데이터분석팀', 'ICT지원팀'으로 구성하여 빅데이터의 체계적인 구축과 그 활용을 극대화하고, 향후 데이터를 지속적으로 축적·유지·관리할 예정이다.

○ 공단은 전국민 5천만명의 출생에서 사망까지 자격 및 보험료 자료, 병의원 이용내역과 건강검진결과, 가입자의 희귀난치성 및 암 등록정보 등 **10년 동안 축적된 1조 3,034억 건의 빅데이터를 보유**하고 있으며,

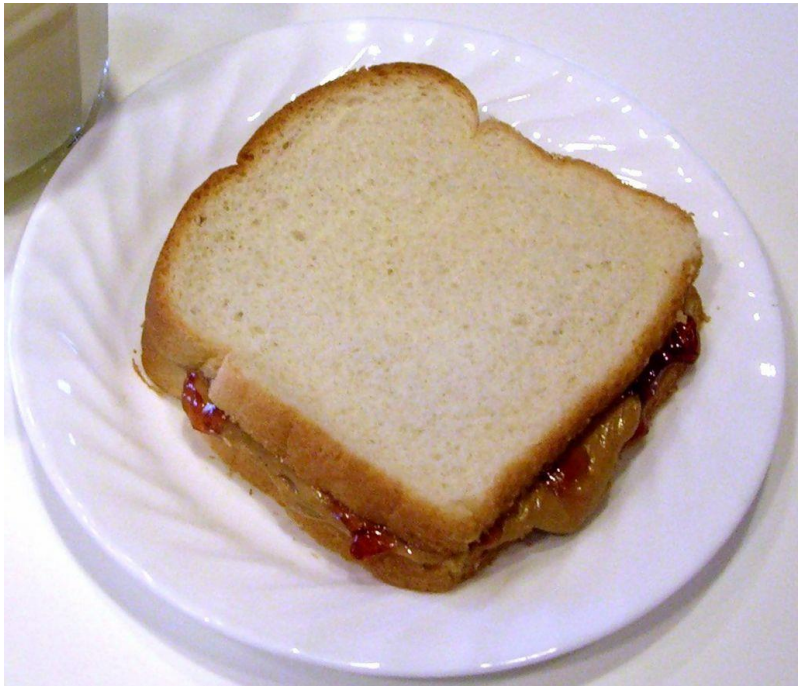
- 작년 6월 과거 10년간의 가입자 자격 및 보험료, 진료내역, 건강검진 내역 등이 포함된 **747억건의 '국민건강정보DB' 구축을 완료**한데 이어, 금년 1월에는 국민건강정보DB를 대표하는 3종의 연구용 **'표본DB' 구축**을 완료하고 그 **완성도 및 질을 높이는 작업을 계속**하고 있다.


학생검진
건강검진
요양기관
사망 및 신생아
자격 및 보험료
국민건강보험공단
국민건강정보DB
진료내역 및 명세서
노인장기요양보험
환자등록정보

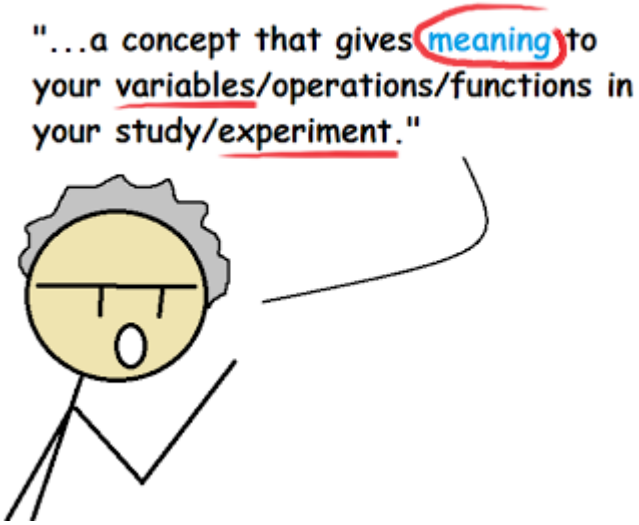

구슬이 서 말이라도...
헉헉
M
25

Operational Definition

Peanut Butter Sandwich



‘the result of putting peanut butter on a slice of bread with a butter knife and laying a second equally sized slice of bread on top’



조작적 정의

- 사물 또는 현상을 객관적인 경험적으로 기술하기 위한 정의 – 수량화할 수 있는 내용
- 조작 (Operational) – 대상을 경험적으로 다룰 수 있도록 서술하는 것
- 산소는 그것이 들어있는 용기에 꺼져가는 성냥개비를 넣으면 (실천적 행동) 불꽃이 다시 일어나게 하는 (관찰할 내용) 기체이다.

폐암 1기는 대부분 수술만 한다.
전산상태가 좋고 다른 질환이 없을 경우 폐엽절제술이 표준치료이다.

건강보험공단 청구코드를 통해 폐암 진단 후 수술 (폐엽절제술)만을 받은 환자는 폐암 1기와 같은 예후를 가질 것이다.

이 환자들의 생존률은?

공공데이터 이렇게 활용한다

서비스	발생산업	일자리
•기상예보	•기상산업	•기상사업자 •기상통보관

서비스	신규사업	신규 일자리
•재해보험컨설팅 •맞춤형 날씨정보 •재해예보 서비스 •에너지 소비량 예측	•기상컨설팅업 •기상 장비업 •기상 감정업	•기상컨설턴트 •기상 감정기사 •기상 예보사

자료·기상청

폐렴 발생위험 예측모형개발 연구

- 폐렴 발생과 환경 및 기상인자와의 관계 분석-

연구 방법

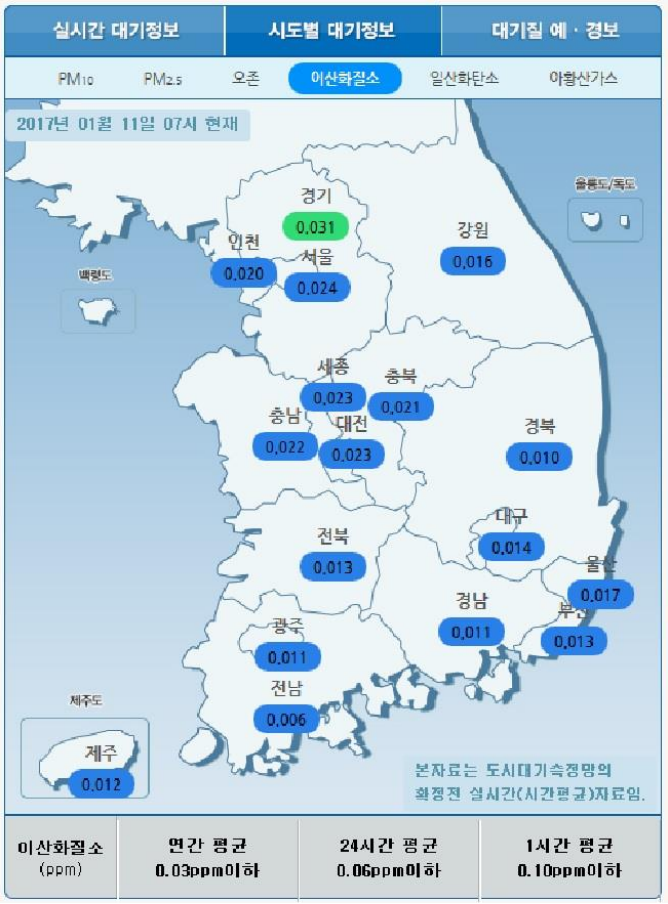
- **연구 목적**
  - ✓ 목적: 2011-2013년 표본 코호트 DB를 대상으로 폐렴의 지역별 평균 발생률에 지역별 대기환경인자 및 기상인자가 미치는 영향을 확인하고자 함

- **연구 대상 질환의 조작적 정의**
  - ✓ 폐렴 : 진단상병코드 J10-J18를 진단받은 환자
  - ✓ 폐렴 신규발생의 정의 : 2010년까지 폐렴 진단을 받은적이 없은 환자 중 연구기간 (2011-2013년)중 새롭게 폐렴을 진단받은 환자를 신규 발생으로 정의

- **연구 결과변수**
  - ✓ 1차 결과변수 : 폐렴 발생에 영향을 주는 환경인자(대기오염, 기상인자)의 파악
  - ✓ 2차 결과변수 : 폐렴 발생에 영향을 주는 환경인자의 time lag을 파악

연구 방법

- **사용변수**
  - ✓ 표본코호트 지역별 평균 일별 및 주별 폐렴발생률(16개 시도)
  - ✓ 대기환경 데이터 : CO,NO2,O3,PM10,SO2
    - 국립환경원 자료 : Air Korea (http://wwww.airkorea.or.kr/) 에서 자료 추출
    - 대표값 산정 방식 : NO2, SO2, PM10는 시간
      CO, O3는 일중 시간별 최
  - ✓ 기상 데이터 : 상대습도, 일조시간, 일교차,평균풍
    - 기상청 자료

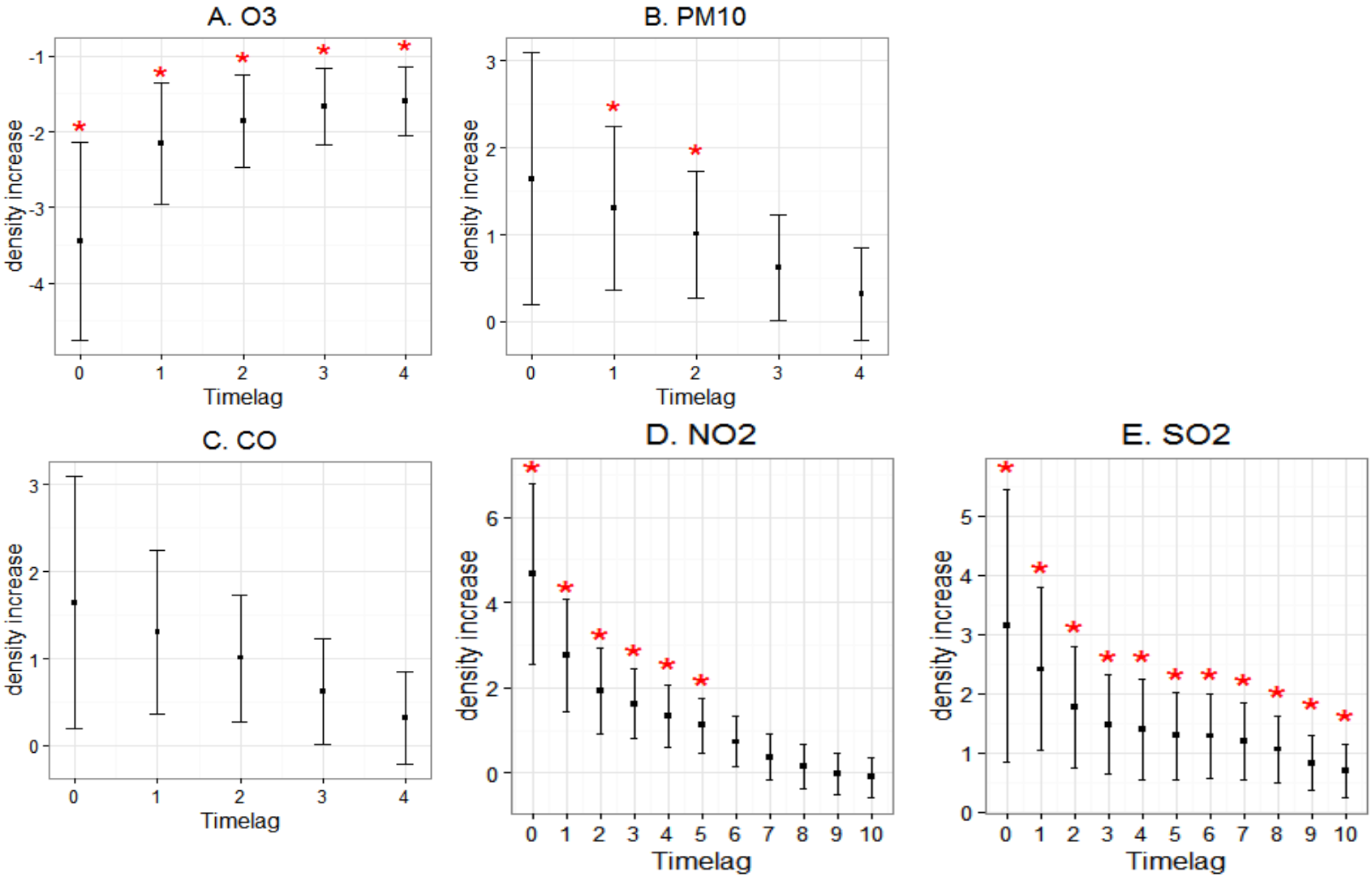


- **분석방식**
  - ✓ 1단계 : 16개 시,도별 포아송 회귀분석 (Generaliz
    regression analysis)
  - ✓ 2단계 : 16개 시,도별 결과값의 메타분석(Meta an

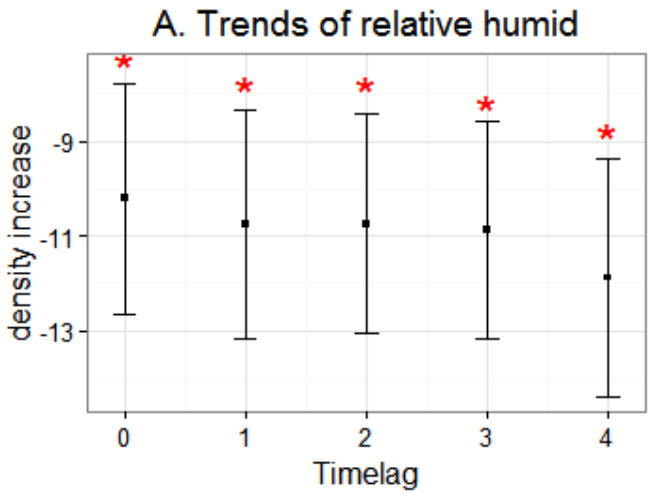
연구결과 – 3년간 폐렴 발생자수

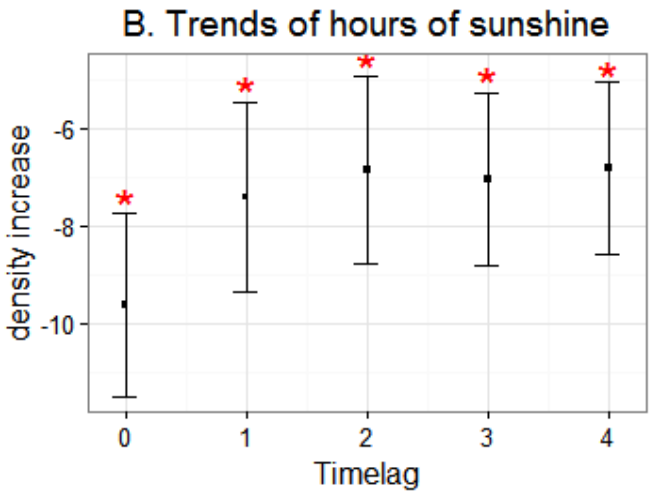
시도코드	표본코호트 DB 수			폐렴 발생자 수(표본코호트 DB추출)		
	2011	2012	2013	2011	2012	2013
	소계	소계	소계	소계	소계	소계
	N	N	N	N	N	N
전체	**1,006,481**	**1,011,123**	**1,014,730**	**24,597**	**25,600**	**26,631**
서울특별시	202,595	201,644	200,395	4,246	4,384	4,540
부산광역시	71,722	71,433	71,190	1,436	1,469	1,499
대구광역시	50,332	50,253	50,164	1,208	1,190	1,202
인천광역시	55,973	56,763	57,459	1,337	1,402	1,439
광주광역시	29,641	29,815	29,864	825	1,028	1,082
대전광역시	30,228	30,444	30,615	826	725	804
울산광역시	23,446	23,674	23,788	676	576	718
세종특별자치시	.	2,199	2,378		67	71
경기도	235,956	239,361	242,246	5,660	5,698	5,734
강원도	28,363	28,443	28,485	707	666	668
충청북도	30,974	31,096	31,202	876	975	1,062
충청남도	41,552	40,199	40,637	1,077	987	1,014
전라북도	37,074	37,075	37,008	1,233	1,292	1,467
전라남도	38,138	38,028	37,963	959	1,152	1,202
경상북도	53,847	53,794	53,889	1,410	1,412	1,398
경상남도	65,640	65,788	66,099	1,844	2,238	2,394
제주특별자치도	11,000	11,114	11,348	277	339	337

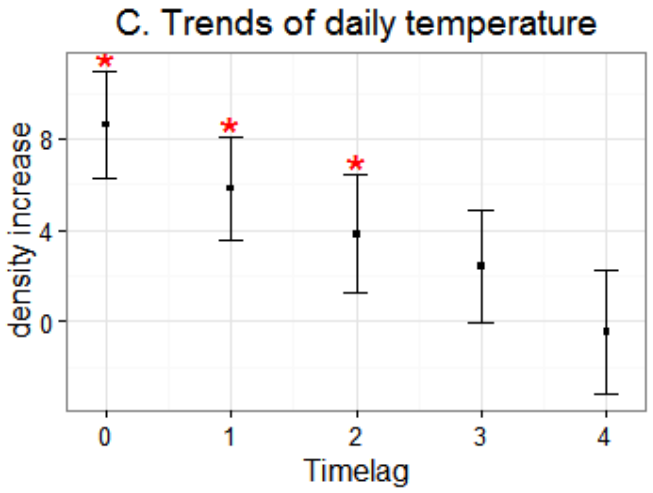
대기환경인자에 따른 포아송 회귀분석 및 메타분석 결과

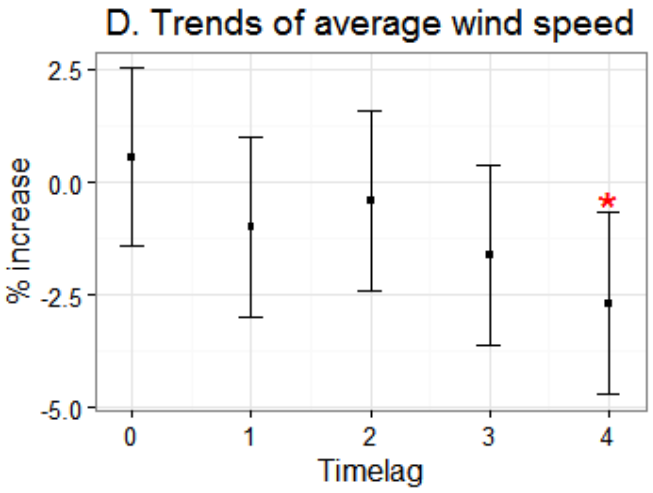


기상인자에 따른 포아송 회귀분석 및 메타분석 결과



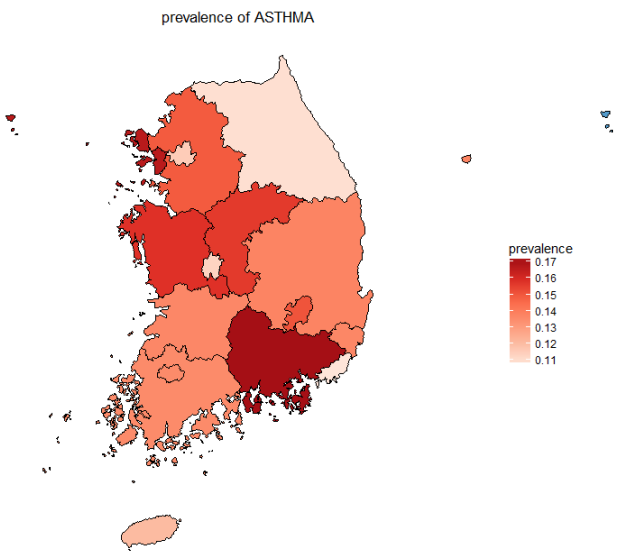




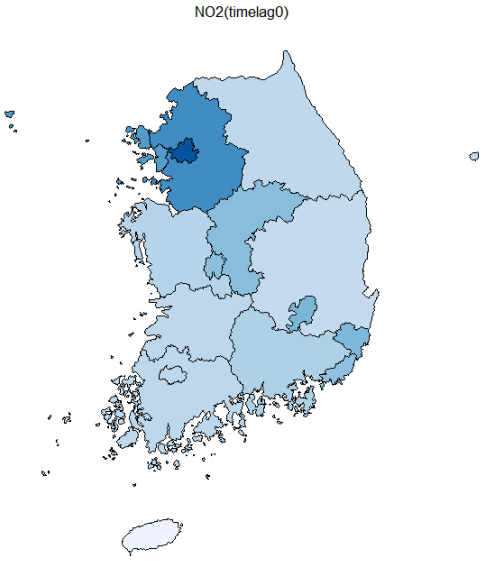


Asthma 발생과 NO2 노출과의 연관성 분석

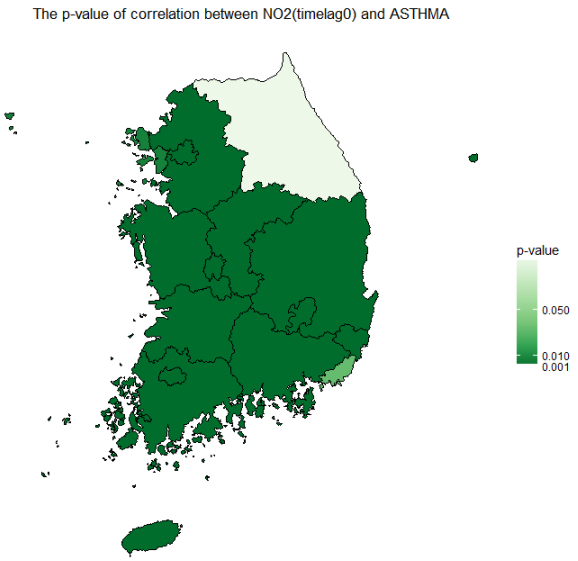
ASTHMA와 NO2(timelag0)의 영향



Prevalence of ASTHMA



NO2(timelag0)

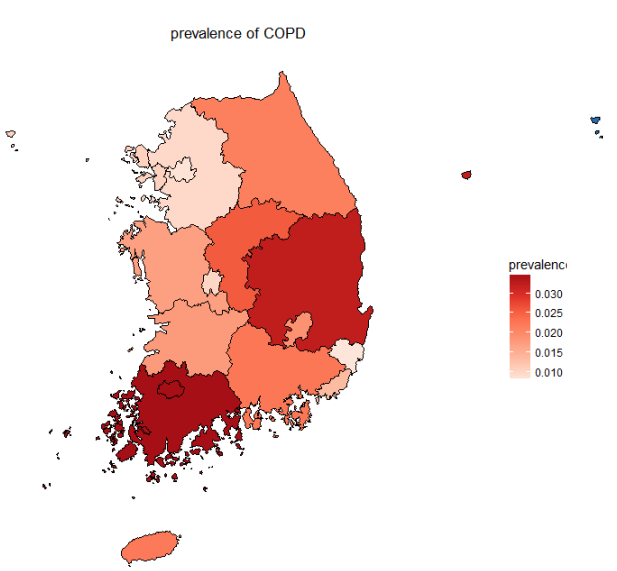


P value

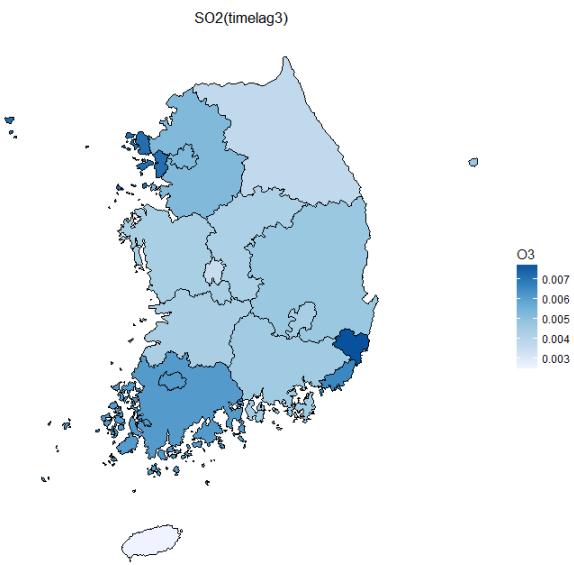
Disease	Time Lag (week)	NO2(이산화질소)		
		estimate	Bonferroni Corrected p	
ASTHMA	0	4.4019	1.73E-17	***

COPD 발생과 SO2(lag3) 노출과의 연관성 분석

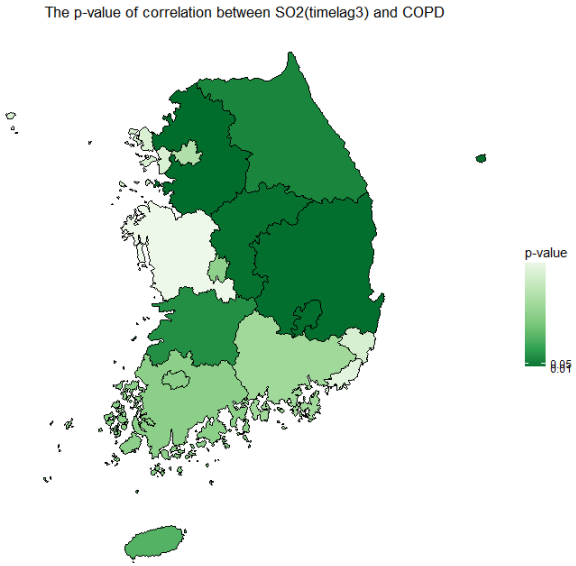
COPD와 SO2(lag3)의 영향



Prevalence of COPD



SO2(timelag3)

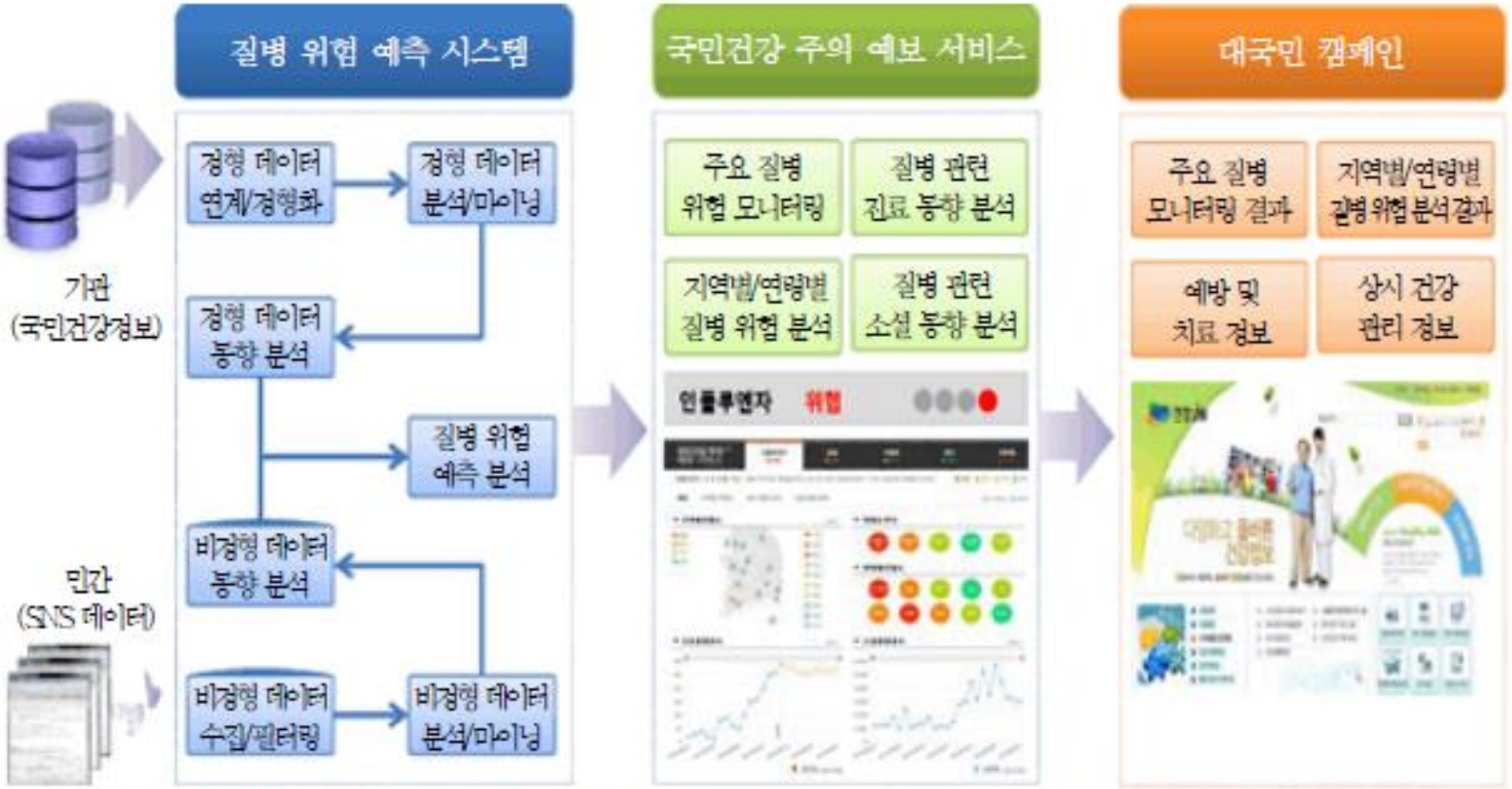


P value

Disease	Time Lag (week)	SO2(아황산가스)		
		estimate	Bonferroni Corrected p	
COPD	3	0.5601	0.022385	*

건강 보험 공단 DB 의 한계

- 환자의 임상정보가 없다.
- 폐암의 조직형태와 유전자 정보가 없다.
  - 6개월 이상 EGFR-TKI를 사용했다면 양성
- 비보험 치료와 임상연구 자료는 없다.



<자료>: "국민건강 주의 알람서비스", 개시, 건보공단, 국민건강보험 보도자료, 2014. 5. 16.

국민건강 주의 예보서비스 개요

HIS-based Big Data Platform

REDCap
Research Electronic Data Capture

I/F (de-identification)

I/F

I/F

I/F

HIS
(EMR/OCS/PACS..)

Disease Registries
/eCRFs

BioBank

Genome DB

Other HIS

Disease
Cohorts

External BioBanks

Public
Genome DB

EMBL-EBI

NCBI

PubMed

h·well
국민건강보험

건강보험심사평가원
Health Insurance Review & Assessment Service

환경부

기상청
Korea Meteorological Administration

AMIS 2.0

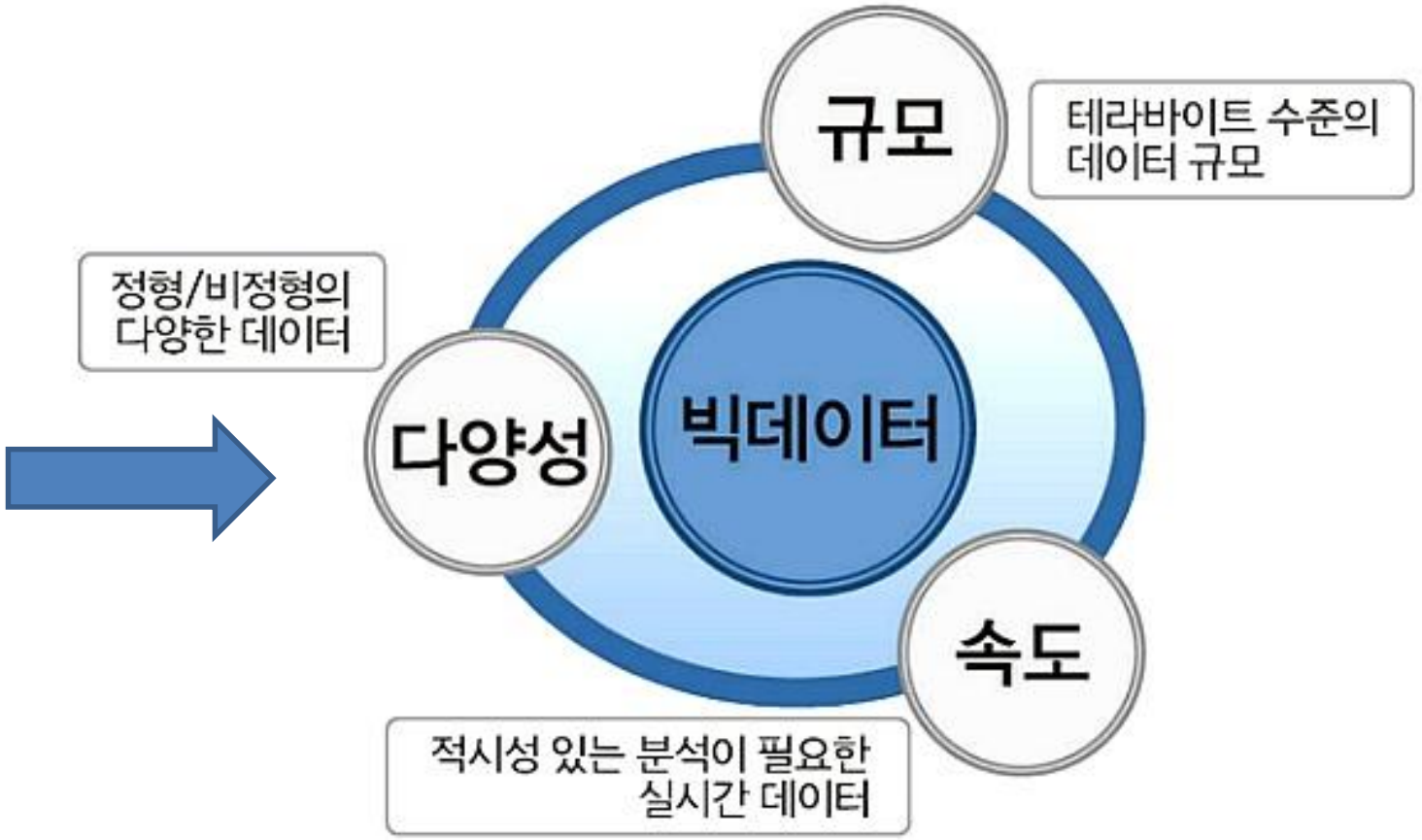

AsanERP
AMIS2
New PetaVisi...
EMR
입원처방
외래처방
호흡기검사 결과입력


Asan BiomedicaL Research Environment
ABLE
Welcome to the ABLE!
Please login enteryour id and password.
아이디
사용자 ID 입력
비밀번호
비밀번호 입력
YES
NO
서울아산병원
Asan Medical Center
아이디 저장
Copyright © 2013 ABLE All Rights Reserved.


Asan Medical Center Information System
AMIS2
자주쓰는 시스템
ABLE
CHD센터
CIS Registry
WorkUp
corelab
감염관리
기타CIS
실행중인 업무시스템
사용중인 메세지시스템
AMIS 종료
재 설 치
개발자 Login
메세지보기
최소화

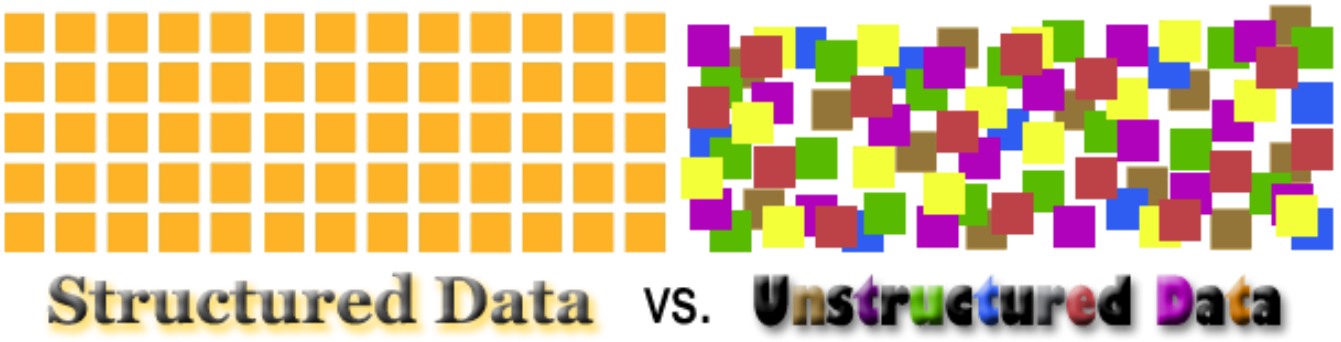
ABLE Statistics (as of Feb. 2014)

4M	Total registered patients
600M+	Total orders
191M+	Total medications
715M+	Total lab results
257M+	Total clinical notes
4M+	Total DICOM images

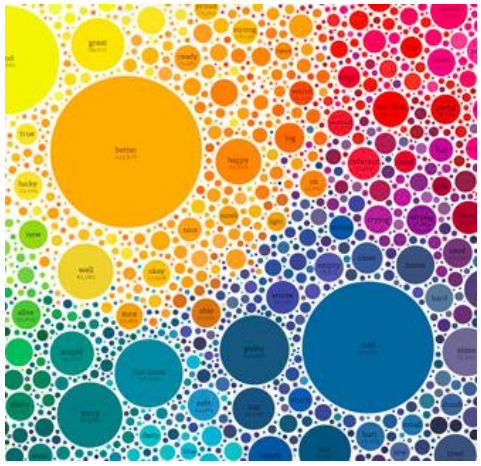

규모
테라바이트 수준의
데이터 규모
정형/비정형의
다양한 데이터
다양성
빅데이터
속도
적시성 있는 분석이 필요한
실시간 데이터

Structure Standarad

"80% of business-relevant information originates in unstructured form, primarily text."



DW (Data Warehouse) **ABLE**




90 kcal
감 1개 160g
130 kcal
사과 1개 250g
80 kcal
귤 1개 100g
90 kcal
오렌지 1개 315g
150 kcal
배 1개 360g
150 kcal
참외 1개 200g
30 kcal
키위 1개 70g
40 kcal
토마토 1개 300g
240 kcal
포도 1송이 340g
30 kcal
방울토마토 5개 100g
50 kcal
수박 2개 250g
100 kcal
바나나 1개 135g
270 kcal
메론 1개 250g
30 kcal
딸기 6개 100g
70 kcal
건대추 11개 25g

Data Warehouse

ABLE

임상 빅데이터 익명화와 연구활용

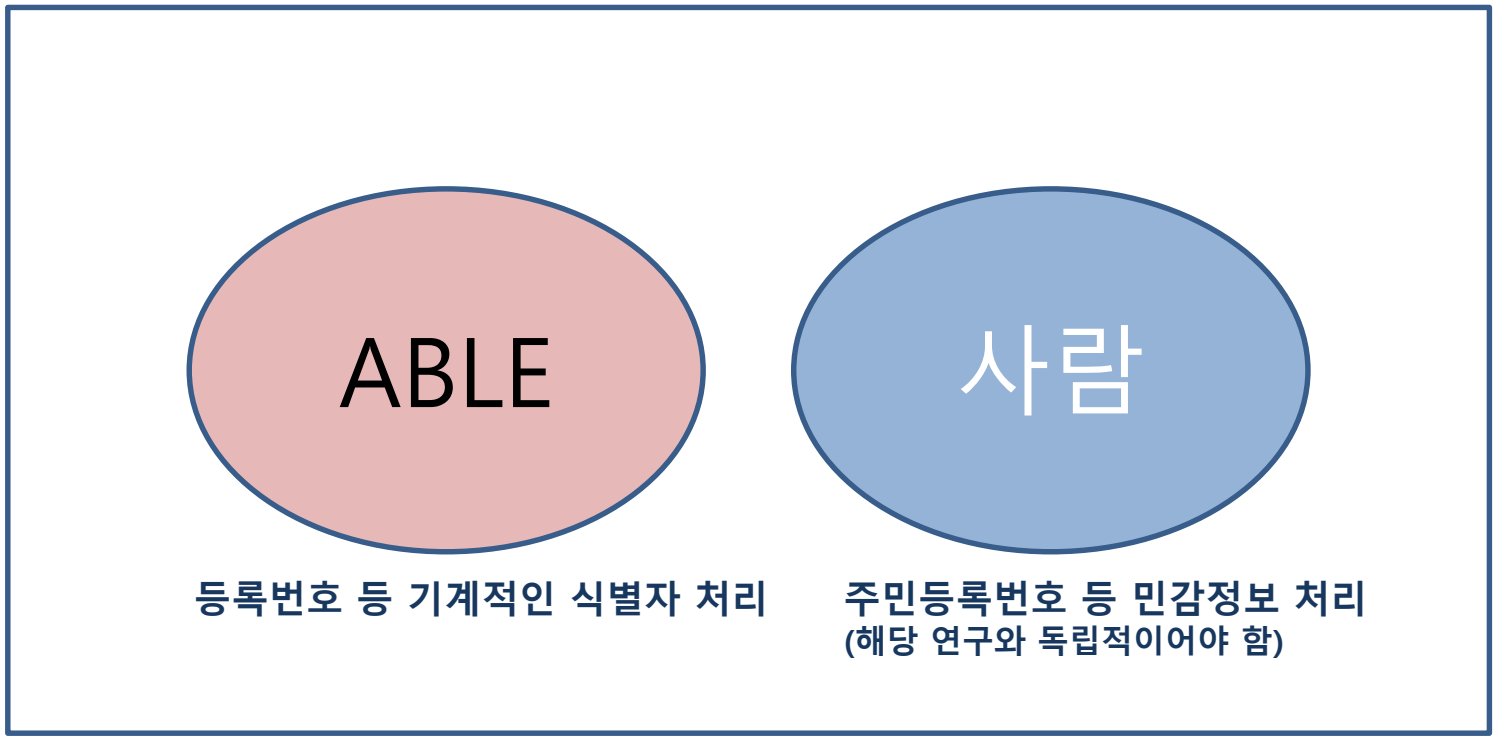
Now, We've got ABLE!

개인건강식별정보 (AMC)

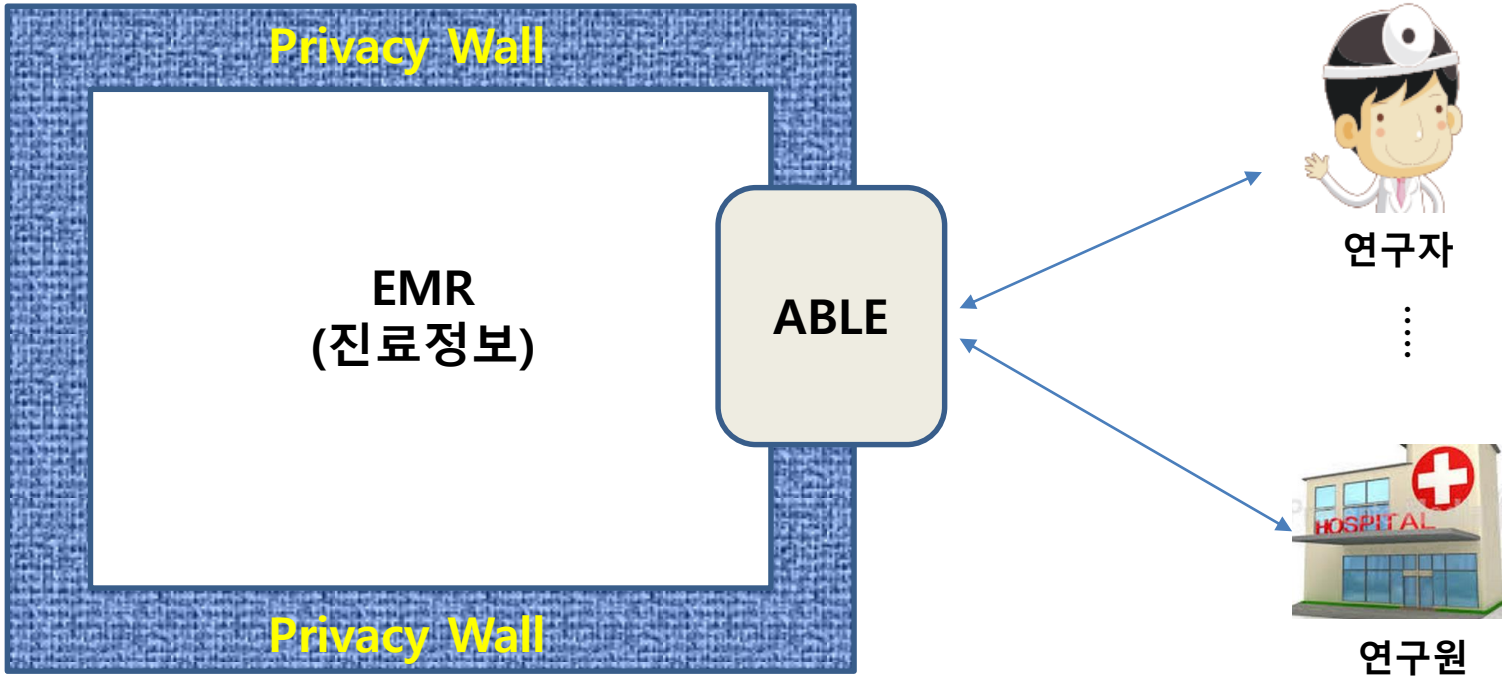
No	개인식별정보
1	성명 (의료진 이름 제외)
2	시/군/구 보다 작은 단위의 지역정보(읍/면/동 이하 상세 주소)
3	전화번호(집전화번호, 직장전화번호, 이동전화번호, Fax번호)
4	이메일주소
5	주민등록번호
6	외국인등록번호
7	여권번호
8	**등록번호**
9	건강보험증번호
10	은행계좌번호
11	신용카드번호
12	자격/면허번호
13	차량번호
14	바이오정보: 지문, 얼굴, 홍체, 정맥, 음성, 필적 등
15	유전자정보
16	홈페이지 회원ID
17	사번
18	비밀번호
19	IP 주소
20	URLs
21	**생년월일 (생년월까지 허용)**

Honest Broker

Honest Broker의 역할 분담



Honest Broker



국민건강보험공단과의 연계문제

4. 이 임상시험에 참여함으로써 문제가 발생할 경우에는 누구에게 연락을 해야 하는지를 알고 계십니까? ☐

본인은 다음 기관에 아래와 같은 목적으로 개인정보를 제공하는 것에 관한 설명을 이해하고 이에 동의합니다. ● 개인정보 제공기관: 국민건강보험공단, 건강보험심사평가원 ● 개인정보제공항목: 청구명세서번호 ● 개인정보제공목적: 타병원에서 받은 진료기록 조회	예 ☐ 아니오 ☐

Big Data for Biomedical Fields

쉽게 생각하면

대규모 Multi-center Trial or Cohort Study!

확장하면

지금까지 따로 따로 사용하고 있던 data를

통합하여 분석하는 것

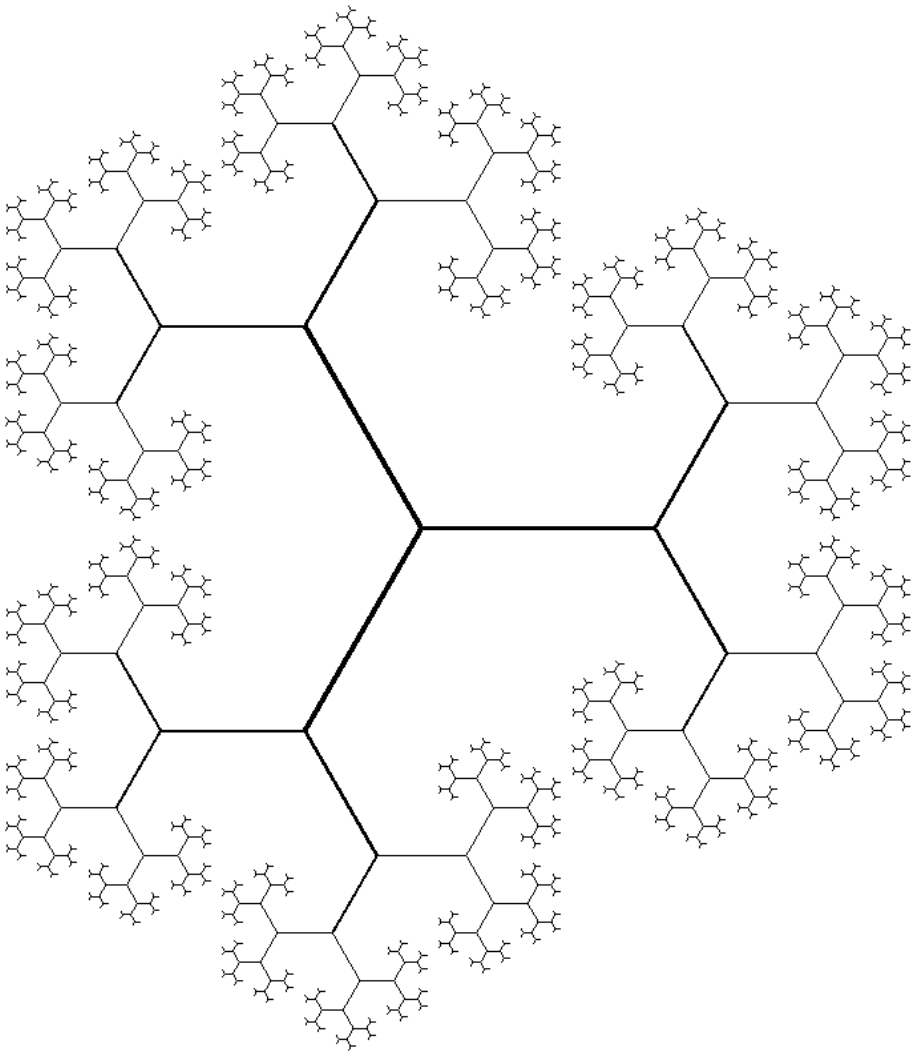
더 확장하면

지금까지 사용하고 있던 데이터에

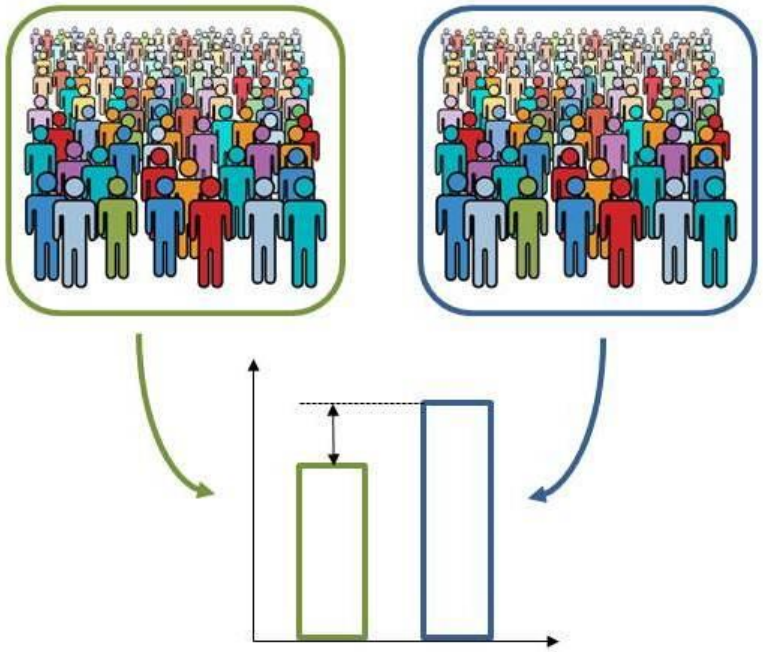
사용하지 않고 있었지만 추가하면 좋을 것 같은 data를

(ex. SNS, 날씨, 환자 생활 data, 자가 측정 data, 의료 기기 log 등)

통합하여 분석하는 것



Randomization
Sampling
Cohort



국가암등록사업 추진체계

암등록병원

시작년도; 1980
병원중심의 암등록사업
암발생의 90%를 등록받음

지역암등록본부

11개 지역암등록본부
부산, 대구/경북,
인천, 광주/전남,
대전/충남, 울산, 제주
강원, 충북,
전북, 경남

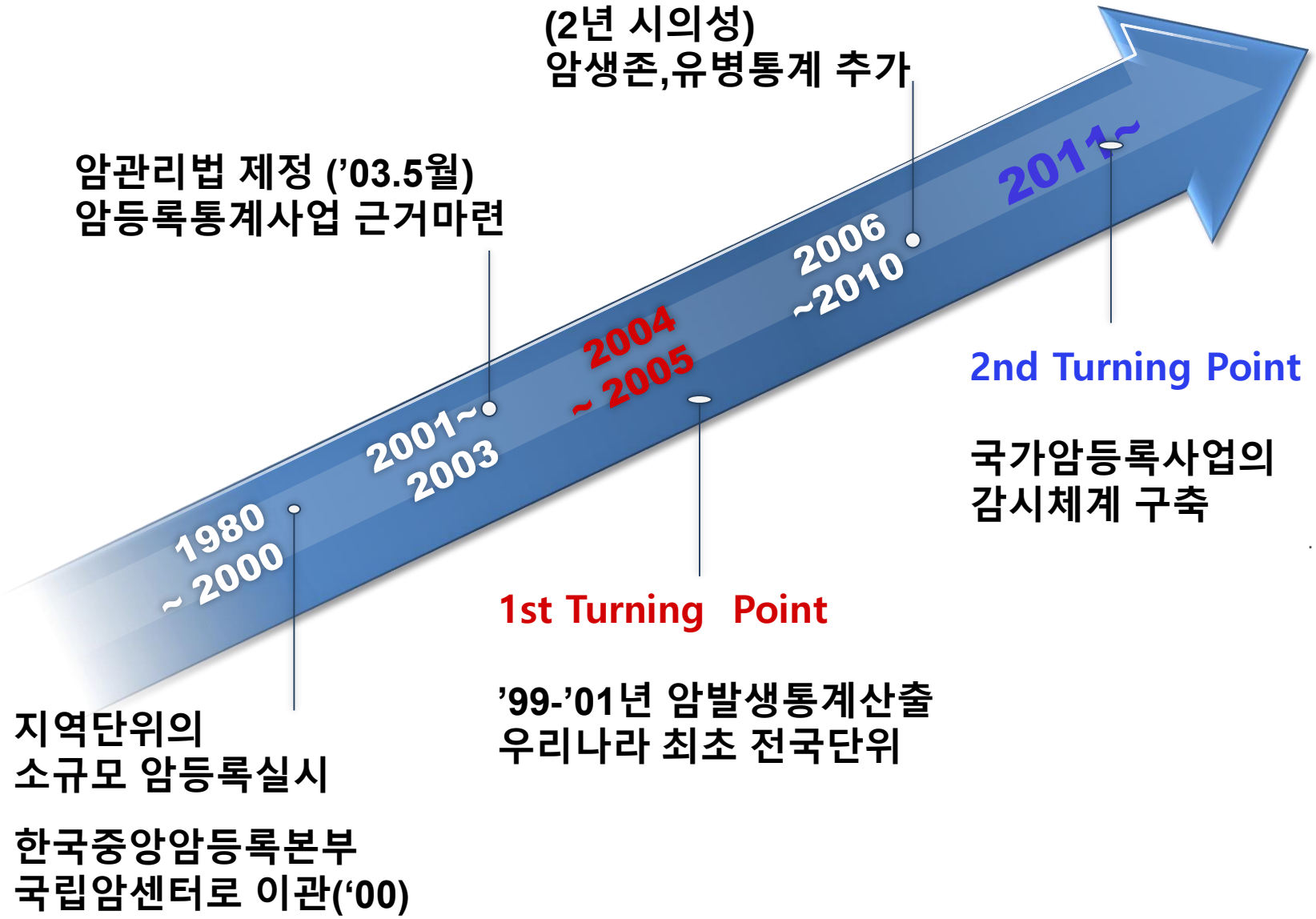
중앙암등록본부

600여개 소형병원자료 조사취합
(중앙 & 11 지역암등록본부)
우리나라 암 발생환자의 96%이상 확보

암종별 암등록

암등록자료 수집정보

환자정보	• 병원번호, 환자번호, 이름, 성별, 나이 • 주소, 직업
암정보	• 초진일, 원발부위, 조직학적 진단명, 최종 진단방법 • 치료(4개월내), 요약병기, 진단경로, 편측성, 분화도, 전이부위
관리정보	• 입력일, 입력자 • 입원일, 퇴원일
타기관 정보	• 사망일, 사망원인 (통계청) • 주민등록주소지 (행정자치부)



지역단위의
소규모 암등록실시

한국중앙암등록본부
국립암센터로 이관('00)

1st Turning Point

'99-'01년 암발생통계산출
우리나라 최초 전국단위

사업의 정착기

'02-'07년 국가암발생률산출
(2년 시의성)
암생존,유병통계 추가

암관리법 제정 ('03.5월)
암등록통계사업 근거마련

2nd Turning Point

국가암등록사업의
감시체계 구축

폐암등록사업

추진 배경

- 전 세계에서 가장 많이 진단, 암사망 원인 중 1위
- 2013년 국내 전체 암 발생의 10.3% (23,177명) 차지
- 우리나라 전체 암 사망자의 22.8% (17,177명) 차지
- 폐암 통계 요구도 증가
- 국가암등록자료에 임상정보, 병기(病期)정보, 환자의 사회경제적 상황 등에 대한 변수가 부재

진행 절차

	내 용	일 정
1	1차 자료 수합 및 강연회	2015. 7. 11
2	2,3,4 차 자료 수합 – 변수 구조 완성	2016. 1. 18
3	중앙암등록 본부 폐암 DB 개발	2016. 2
4	폐암 등록 pilot study 위원회 소속된 위원의 의료기관(11개) 5case, 직접조사 수행	2016. 4
5	pilot 조사 feedback 에 따른 DB 구조 보완	2016. 5
6	폐암등록자료의 활용 규약 제정	2016. 5~6
7	2016 대한 폐암학회 춘계학술대회에서 사업계획 발표	2016. 6. 24
8	자료입력 지침서 제작, 전산프로그램 개발	2016. 6~9
9	폐암등록사업 교육 프로그램	2016. 9
10	폐암 등록 시범 조사 개시	2016. 10 ~

TNM stage 영상진단소견	T인자	T인자	hc_t	영상기록에서의 T stage 표시	
		종양최대직경	hc_maxsize	영상검사에서 기록된 구체적인 크기 입력	종양크기: 5×4×3cm → 5cm 입력
		종양의 위치	hc_location_1 ~ hc_location_6	종양의 위치 * 중복선택가능	
		Main bronchus	hc_tsite4	0. 침윤없음 1. 침윤있음 9. 모름	
		폐쇄성 폐렴	hc_tsite8	0. 없음 1. 있음 9. 모름	
		종양 침범 정도	hc_tsite6_1 ~ hc_tsite6_16	종양이 침범한 주변 기관 및 장기 * 중복선택가능	
	N인자	N인자	hc_n	임상적 국소림프절 침범 여부	
		N1_site	hc_nsite1_1 ~ hc_nsite1_5	N1인 경우, 침범된 동측(Ipsilateral) 림프절 * 중복선택가능	
		N2_site	hc_nsite2_1 ~ hc_nsite2_8	N2인 경우, 침범된 동측(Ipsilateral) 림프절 * 중복선택가능	
		N3_site	hc_nsite3_1 ~ hc_nsite3_6	N3인 경우, 침범된 반대측(Contra-lateral) 림프절 * 중복선택가능	
	M인자	M인자	hc_m	임상적 원격전이 여부 (CT와 PET-CT, Brain MRI 등 영상자료 검토)	
		M1a_site	hc_msite1_1 ~ hc_msite1_4	M1a 인 경우, 전이 부위 * 중복선택가능	
		Pleura nodule	hc_msite1_3	☐ Ipsilateral ☐ Contralateral ☐ 모름	
		M1b_site	hc_msite2_1 ~ hc_msite2_6	M1b 인 경우, 전이 부위 * 중복선택가능	
		M1b_기타	hc_msite2_t	전이부위가 '7. 기타' 인 경우 기타에 해당하는 전이부위 입력	Kidney
		전이개수	hc_msite3	M1b인 경우; 전이된 개수의 총합 1. 1개 2. 2개 이상 9. 모름	2. 2개
	Clinical Stage		hc_stage_1	Clinical Stage 산출 (자동계산)	cT1aN2M0 → Stage IIIA

공공자료 활용

- 기관의 고유목적에 따라 공공자료원 수집 및 관리
- 구축된 공공자료원을 연구에 활용할 수 있음

➔ **비용효과적**으로 **양질**의 연구결과 도출 가능

표 1. 공공자료원별 보유정보

변수	건강보험청구자료	검진자료	자격자료	사망원인자료	중앙암등록 자료
	공단/심평원	공단		통계청	국립암센터
인구사회학적 특성	-	○	○	-	-
생활습관 및 행태	-	○	-	-	-
질병이환	○	○	-	-	○
약물정보	○	-	-	-	-
검진 및 신체계측자료	-	○	-	-	-
사망	-	-	사망만 표기	○	-

자료 : 박기수 외. 근거개발을 위한 보건의료 자료연계 전략계획 연구, 한국보건의료연구원, 2010. 수정

공공자료 활용

- 정부 3.0 및 세계적 이슈
- 공공자료 개방 및 포털 구축

건보공단, 표본코호트DB 학술연구용 제공

건보자료 개방공유 통한 사회경제적 가치 증대 기대

배준열 기자 |jun@medifonews.com

등록일: 2014-06-23 오후 3:45:06

국민건강보험공단(이사장 김종대 이하 공단)은 자사가 보유한 빅데이터를 학술연구용으로 제공해 사회경제적 가치 증대를 도모하기로 했다.

공단은 지난 2012년에 구축한 표본코호트DB를 2013년 시범연구를 통해 자료의 완성도를 높여 오는 25일부터 공고기간을 거친 후 일반 연구자에게 학술연구용으로 제공한다고 밝혔다.

표본코호트DB는 2002년을 기준으로 전 국민의 2%인 약 100만 명을 표본 추출해 2010년까지 동일 대상자에 대해 사회·경제적 변수(거주지, 사망년월, 사망사유, 소득수준 등)가 포함된 자격자료, 진료내역 및 건강검진자료를 9년간 연결한 코호트 자료로 장기간의 관찰이 가능하여 시간적 선후관계나 인과적 관계 분석이 가능한 자료이다.

동 자료는 익명화된 자료이지만 국민의 민감한 건강정보임을 감안하여 우선 정책 및 학술 연구과제에 한하여 공단 내부의 심의기구인 "연구지원 심의위원회"의 심의를 거쳐 최소한의 수수료를 받고 제공할 예정이다.

빅데이터 개방과 현실

NECA

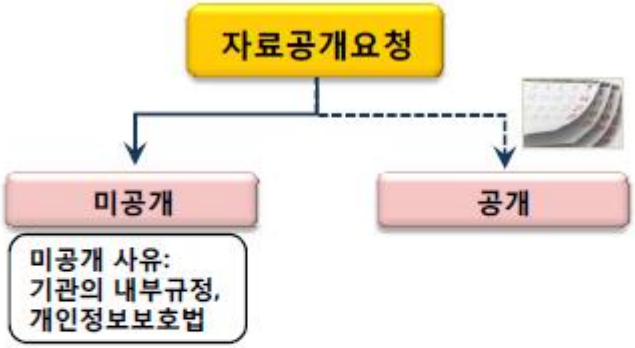
개방

- 정부 3.0 및 세계적 이슈
- 공공자료 개방 및 포털 구축

현실

- 보건의료 관련
  공공기관 자료 공개 요청

➔ **자료 공개까지 긴 시간 소요**



- **개인정보 활용 제한**으로 자료연계를 활용한 공익연구 어려움

공공데이터 활용 및 향후 추진 방향

AMC – 여성 폐암 건강보험공단 사례

2014.11

맞춤형 자료제공이 가능하나
지원인력, 인프라 한계 (직접 방문)
단일 병원 사업으로는 부적절

추진 방향

호흡기학회와 건강보험공단 - 폐암 DB 구축 및 활용에 대한 MOU 체결
대한폐암학회 폐암병기조사 사업 참여
분자폐암연구회 - 임상정보가 포함된 다기관 코호트 구축

다양한 정보가 유기적으로 연결된 폐암 DB 구축

HIS-based Big Data Platform

REDCap
Research Electronic Data Capture

I/F (de-identification)

I/F

I/F

I/F

HIS
(EMR/OCS/PACS..)

Disease Registries
/eCRFs

BioBank

Genome DB

I/F

Other HIS

Disease
Cohorts

External BioBanks

Public
Genome DB

EMBL-EBI

NCBI

PubMed

h·well
국민건강보험

건강보험심사평가원
Health Insurance Review & Assessment Service

환경부

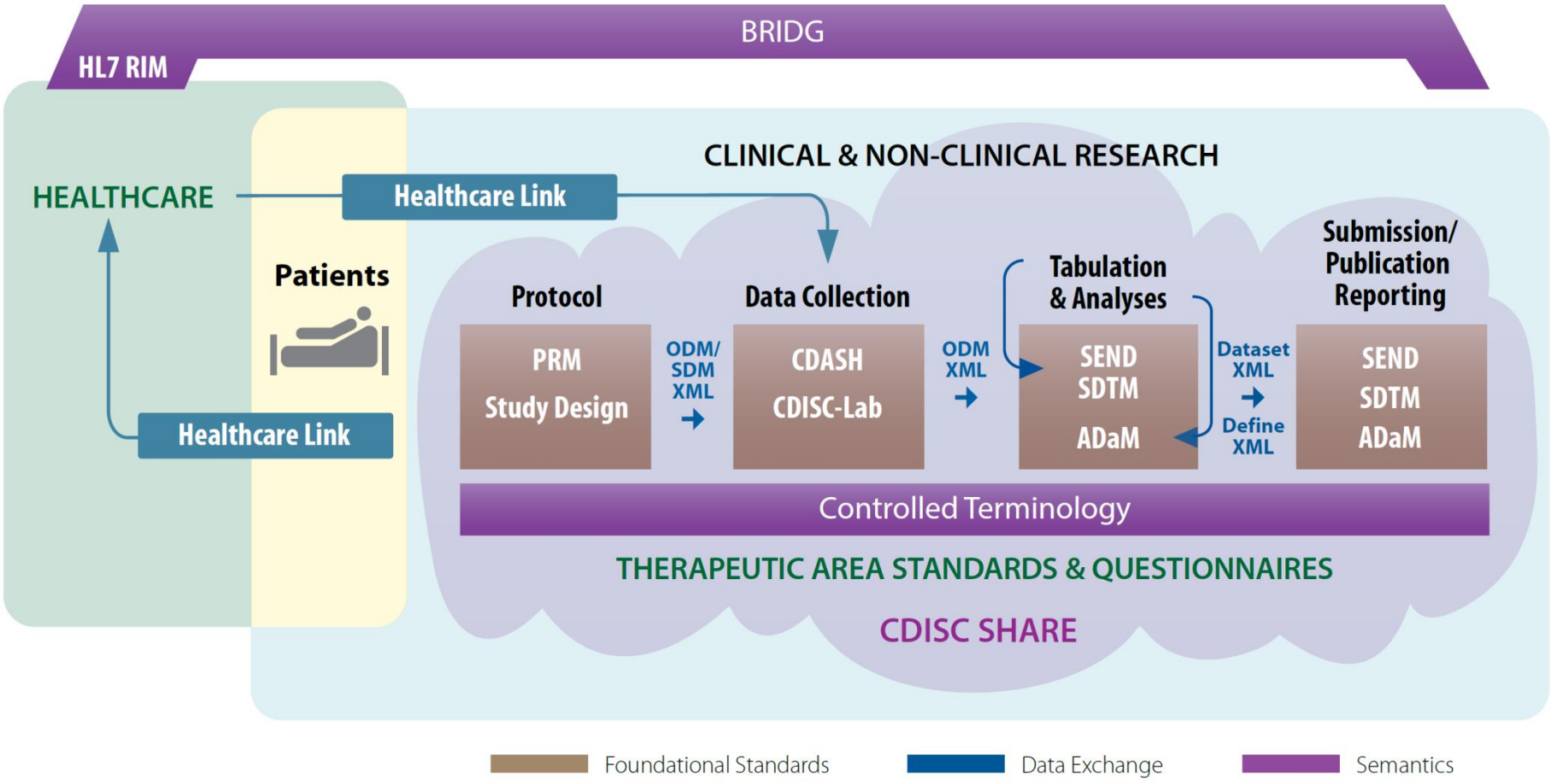
기상청
Korea Meteorological Administration

Clinical Data Interchange Standards Consortium

The CDISC Vision is to inform patient care and safety through higher quality medical research.

The CDISC Mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare.

Achieving Interoperability



BRIDG

Biomedical Research Integrated Domain Group (BRIDG) UML model of the semantics of protocol-driven clinical research.

CDASH

Clinical Data Acquisitions Standards Harmonization is a specification describing basic data collection domains and variables for CRF data with standard question text, implementation guidelines, and best practices.

Glossary

Glossary with definitions of acronyms and terms commonly used in clinical research.

Standard	Description
SDTM, SEND	SDTM : Study Data Tabulation Model SEND : Standard for Exchange of Nonclinical Data
ODM	Operational Data Model
Define.xml	Case Report Tabulation Data Definition Specification
LAB	Laboratory Standards-Content standard
ADaM	Analysis Data Model
Protocol Representation	Collaborative effort to develop macine-readable standard protocol with data layer
Terminology Codelists	Developing standard terminology to support all CDISC standards
CDASH	Data acquistion(CRF) standards

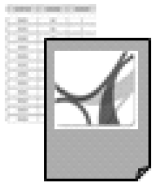
*Specification referenced in FDA Final Guidance

Clinical Trial Flow ; CDISC WAY

Protocol	Form Setup & Config	Data Capture	Analysis	Submission	Review








Protocol

CDASH

ODM

LAB

SDTM & ADaM

CDISC Standards

Table 2. CDSC Metadata

Attributes	Description
Variable Name	실험값 기술에 필요한 변수를 정의, 8자로 제한
Variable Label	변수명에 대한 상세 설명, 40자로 제한
Data Type	실험값에 대한 타입으로 Character나 numeric등이 있음
Controlled terminology	값이나 변수의 포맷을 표현할 때 사용하는 용어집
Origin of each variable	변수(항목)이 어디에서 유래된 것인지를 나타냄 CRF 에서 수집되거나 파생된 것
Role of the variable	변수명이 데이터 셋에서 어떻게 사용되는지 정해 놓은 것
Comments or other relevant information	코멘트나 Data와 관련된 정보를 나타냄

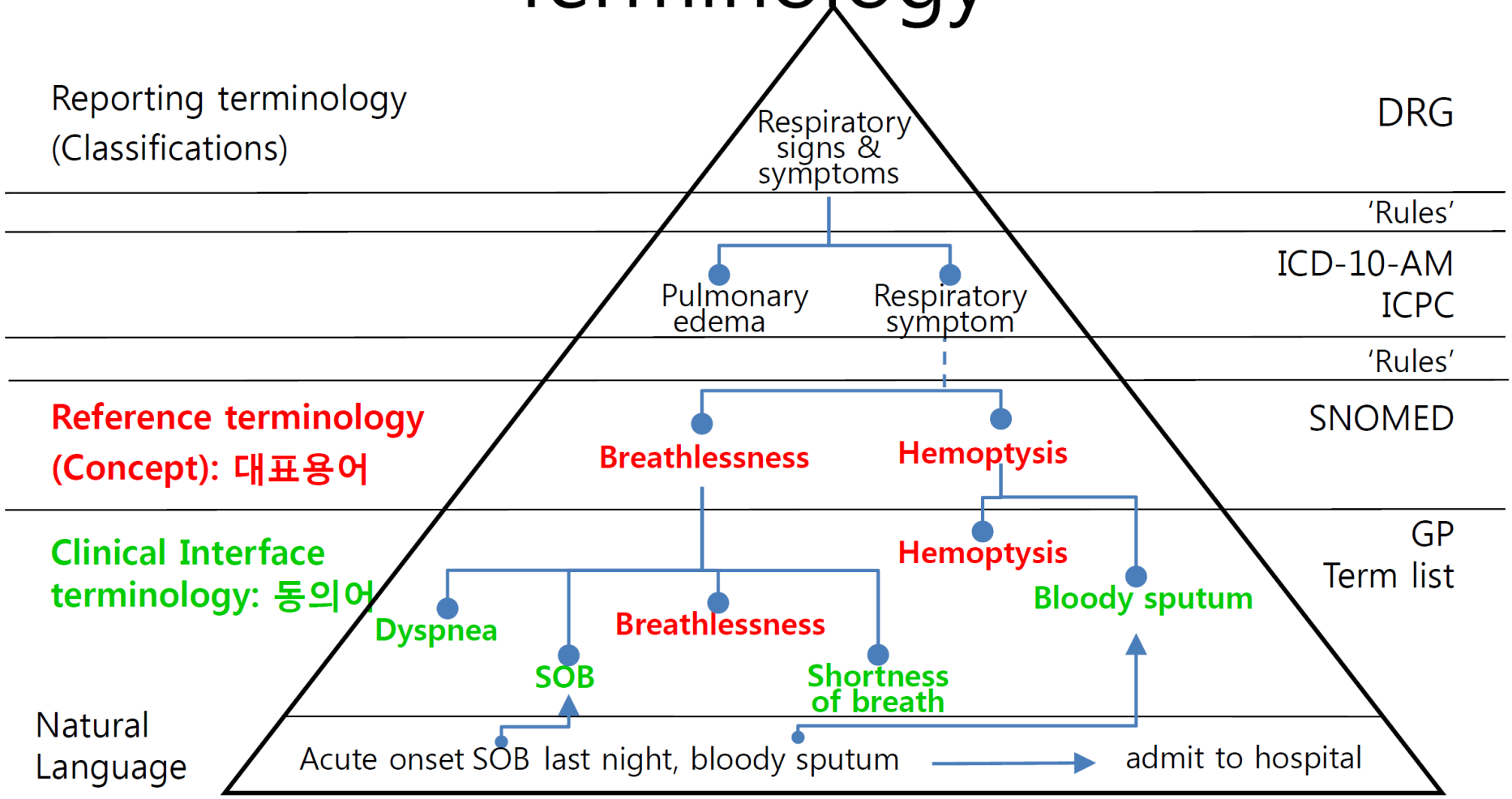
Row	STUDYID	DOMAIN	USUBJID	CMSEQ	CMTRT	CMDOSE	CMDOSU	CMDOSFRQ	CMSTDTC	CMENDTC
1	ABC123	CM	ABC-0001	1	ASPIRIN	100	MG	ONCE	2004-01-01	2004-01-01
2	ABC123	CM	ABC-0001	2	ASPIRIN	100	MG	ONCE	2004-01-02	2004-01-02
3	ABC123	CM	ABC-0001	3	ASPIRIN	100	MG	ONCE	2004-01-03	2004-01-03
4	ABC123	CM	ABC-0001	4	ASPIRIN	100	MG	ONCE	2004-01-07	2004-01-07
5	ABC123	CM	ABC-0001	5	ASPIRIN	100	MG	ONCE	2004-01-07	2004-01-07

electronic data capture (EDC). The process of collecting clinical trial data into a permanent electronic form. NOTE: Permanent in the context of these definitions implies that any changes made to the electronic data are recorded with an audit trail. EDC usually denotes manual entry of CRF data by transcription from source documents. The transcription is typically done by personnel at investigative sites. *See also data entry, direct data entry, data acquisition.*

monitoring. The act of overseeing the progress of a clinical trial and of ensuring that it is conducted, recorded, and reported in accordance with the protocol, standard operating procedures (SOPs), good clinical practice (GCP), and the applicable regulatory requirement(s). [ICH E6 Glossary]

interim analysis(es). Analysis comparing intervention groups at any time before the formal completion of the trial, usually before recruitment is complete. [CONSORT statement]

Terminology



Source : Conceptual Framework; The Language of Health Concept Representation, Australia Standards 2004

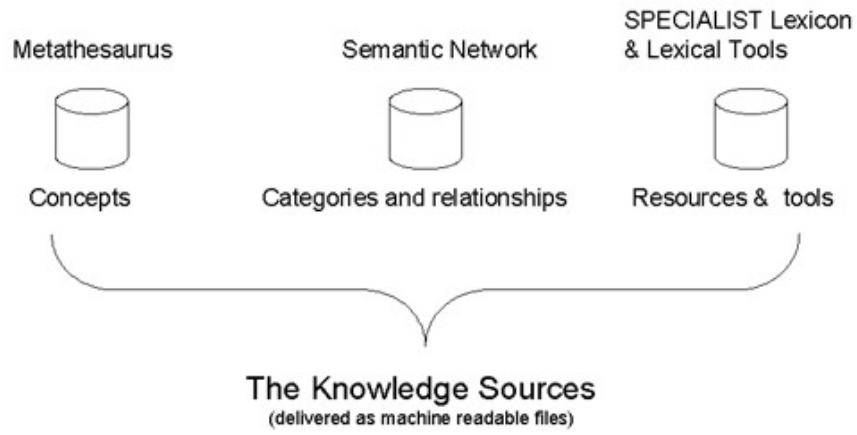
The Unified Medical Language System (UMLS)

The purpose of the National Library of Medicine® Unified Medical Language System (UMLS) is to facilitate the development of computer systems that bel they "understand" the meaning of the language of biomedicine and health. The UMLS provides data for system developers as well as search and report fu less technical users.

There are three UMLS Knowledge Sources:

- The Metathesaurus®, which contains over one million biomedical concepts from over 100 source vocabularies
- The Semantic Network, which defines 135 broad categories and fifty-four relationships between categories for labeling the biomedical domain
- The SPECIALIST Lexicon & Lexical Tools, which provide lexical information and programs for language processing

They are distributed with flexible lexical tools and MetamorphoSys, the UMLS install and customization program.



Collecting more (BIG) data..

OHDSI

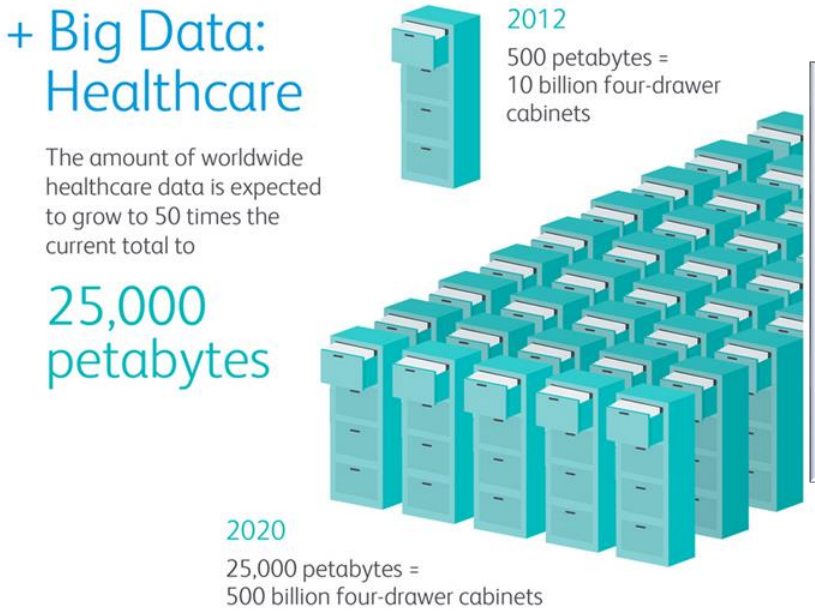
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

http://youtu.be/wGdqGOQNkuM

Big Data, Big Data하는데 대체 그게 뭐야?

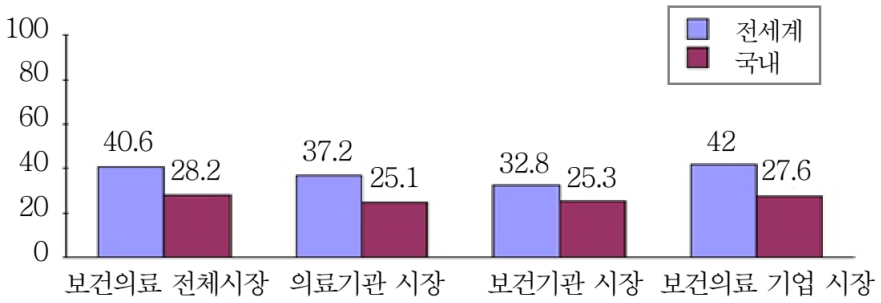
큰 데이터....?

빅데이터 시대



- 의료 관련 정보의 급증
- 다양한 데이터 포맷
- 생성 및 이용속도 급증
- 데이터 분석 및 정보 추출에 대한 요구(Needs) 폭발
- 임상현장에 적용할 수 있는 기술에 대한 관심 증가

연평균 25% 이상의 고성장 시장



출처: 정보통신산업진흥원, 국내외 보건의료 빅데이터 현황 및 과제, 2014 (재인용)

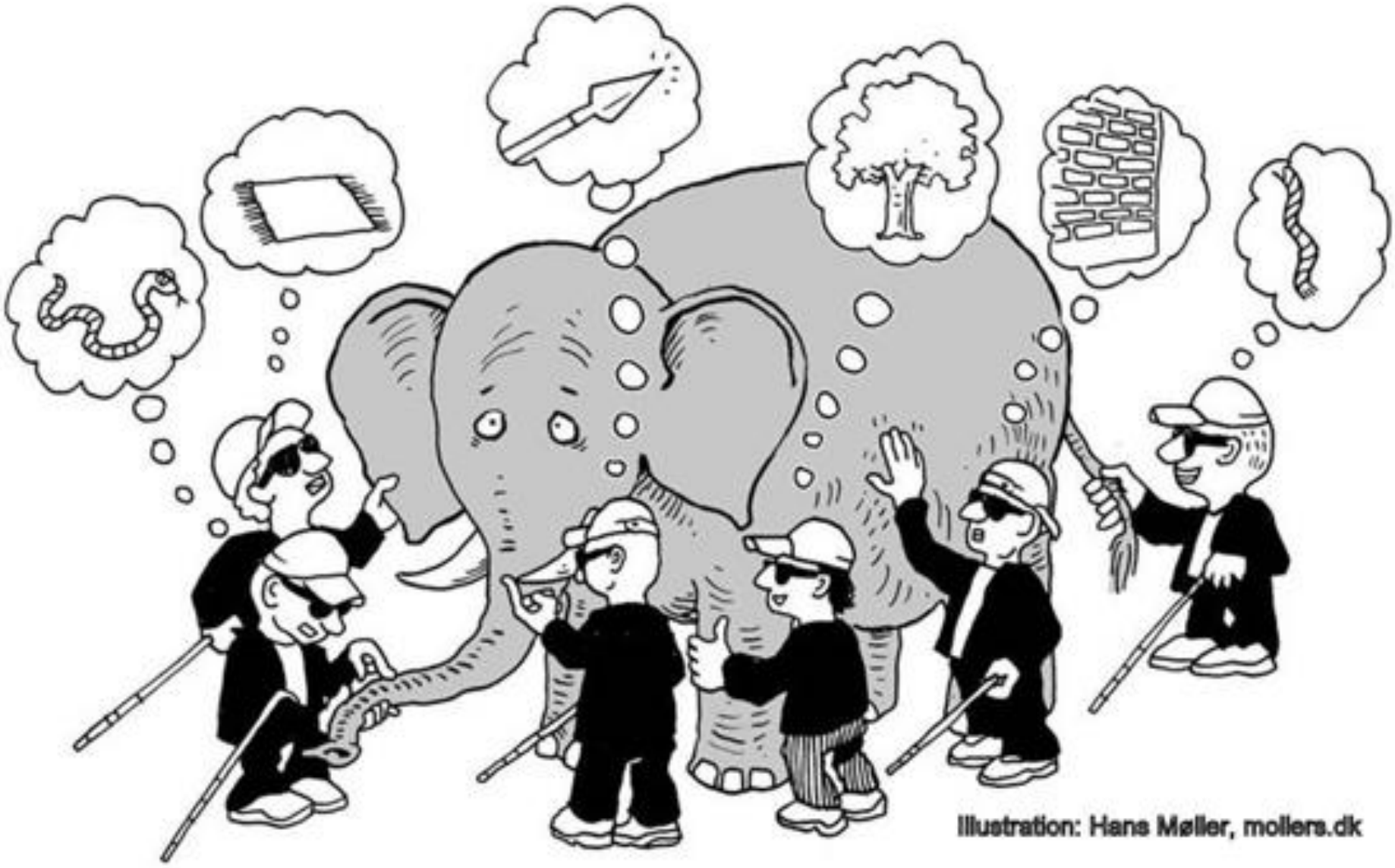
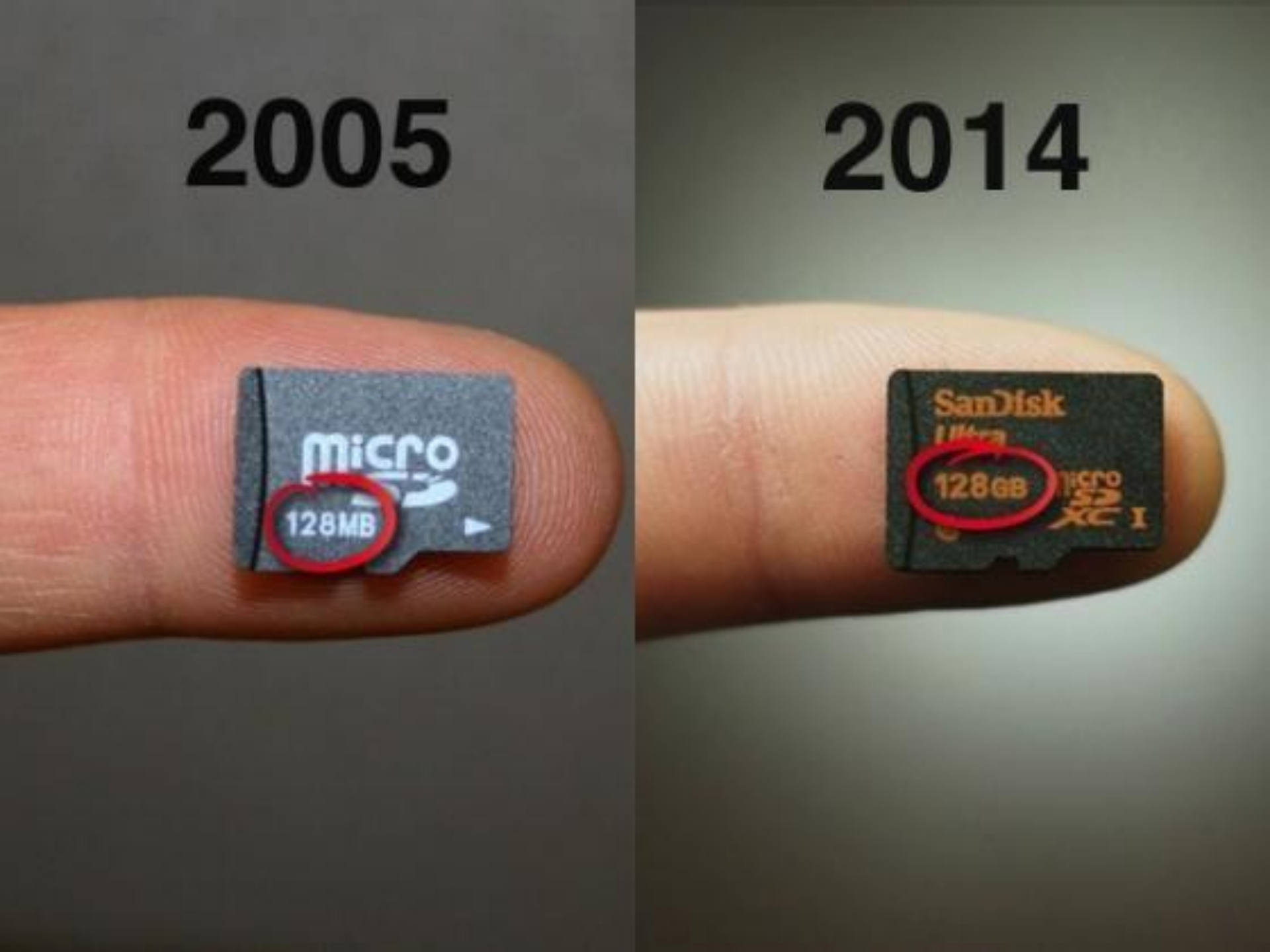
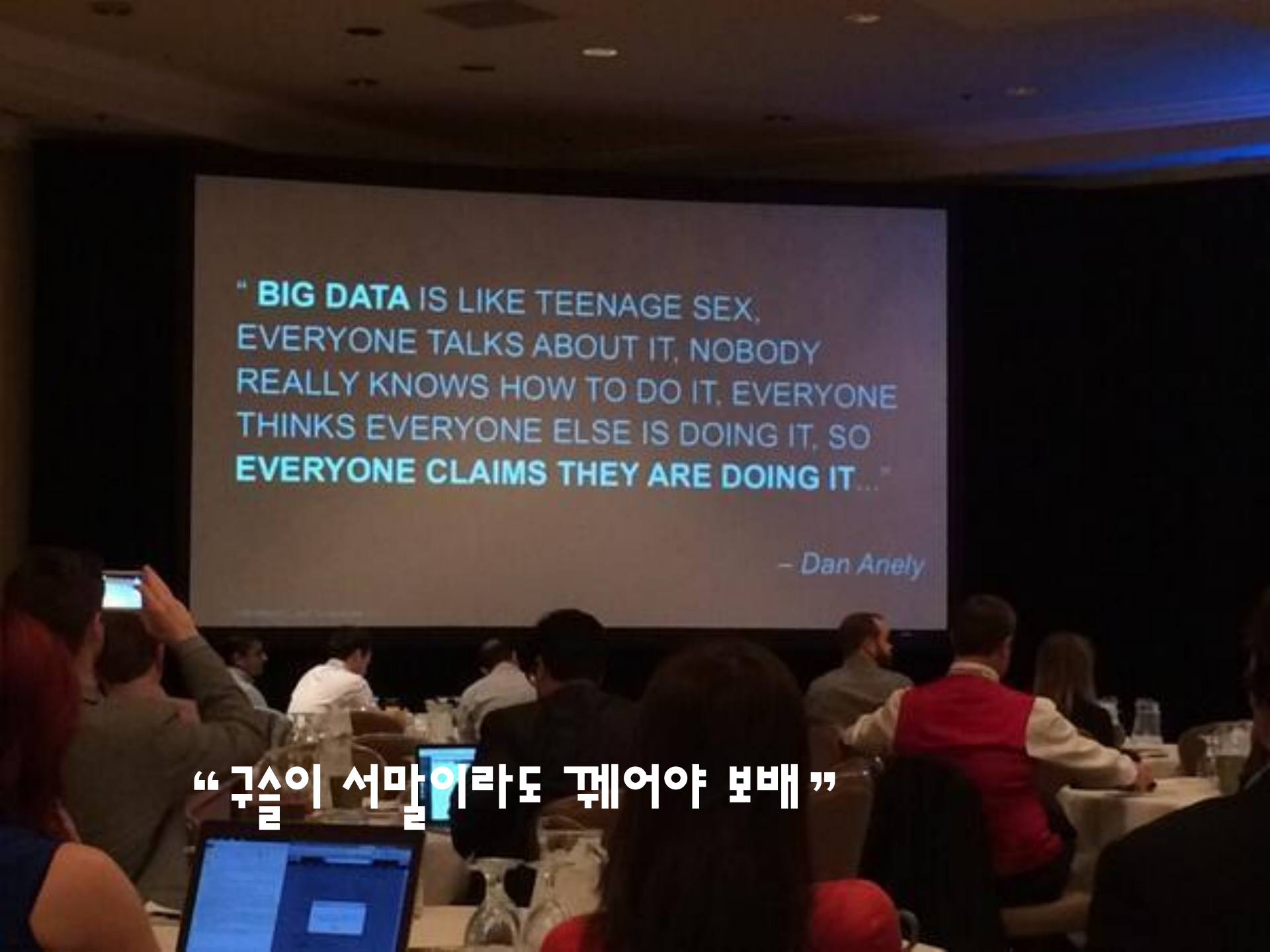


Illustration: Hans Møller, mollers.dk

40+ though leaders opinion on Big Data

http://datascience.berkeley.edu/what-is-big-data/


2005
micro
128MB
2014
SanDisk
128GB



"구슬이 서말이라도 꿰어야 보배"