Big Data Analysis in Respiratory Research

연세대학교
의과대학
내과학교실

김 영 삼

BIG DATA의

정의와 특징

관심을 가져야 하는 이유

호흡기 질환에서의 연구

나아가야 할 길

BIG DATA의 정의

기존 데이터베이스 관리도구로 데이터를 수집·저장·관리·분석의 역량을 넘어서는 대량의 정형 또는 비정형 데이터 세트 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술

BIG DATA의 정의

수십에서 수천 데라바이트 정도의 거대한 크기를 갖고 여러 가지 다양한 비정형 데이터를 포함하고 있으며, 생성, 유통, 소비가 몇 초에서 몇 시간 단위로 일어나 기존의 관리 및 분석 체계로는 감당할 수 없을 정도의 거대한 데이터의 집합으로 대규모 데이터와 관계된 인력, 조직, 기술 및 도구(수집, 저장, 검색, 공유, 분석, 시각화 등)까지 모두 포함하는 개념

BIG DATA의 특징

규모(Volume)의 증가

: 기술적인 발전과 IT의 일상화가 진행되면서 해마다 디지털 정보량이 기하급수적으로 폭증하여 제타바이트(ZB)시대로 진입

다양성(Variety)의 증가

: 로그기록, 소셜, 위치, 소비, 현실데이터 등 데이터의 종류의 증가와 멀티미디어 등 비정형화된 데이터 유형의 다양화

BIG DATA의 특징

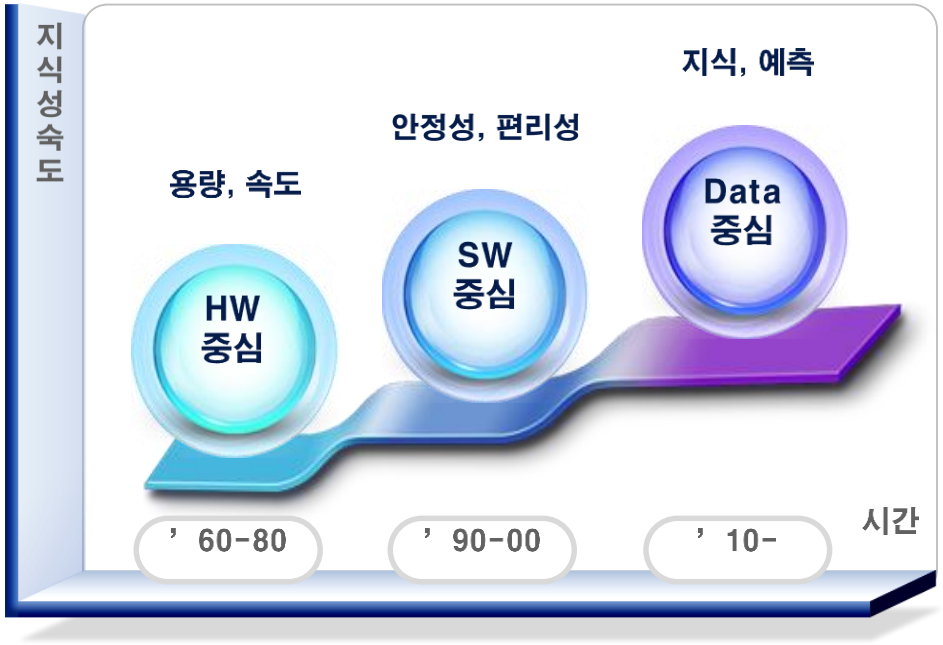
복잡성(Complexity)의 증가

: 구조화되지 않은 데이터, 저장방식의 차이, 중복성 문제, 데이터의 종류 확대, 데이터 관리 및 처리의 복잡성이 심화

속도(Velocity)의 증가

: 사물정보(센서, 모니터링), 스트리밍 정보 등 실시간 정보의 증가로 데이터의 생성과 이동(유통) 속도가 증가, 대규모 데이터 처리와 정보의 활용을 위한 데이터 처리 및 분석 속도가 중요

BIG DATA의 특성


지식성숙도
지식, 예측
안정성, 편리성
용량, 속도
HW 중심
SW 중심
Data 중심
'60-80
'90-00
'10-
시간

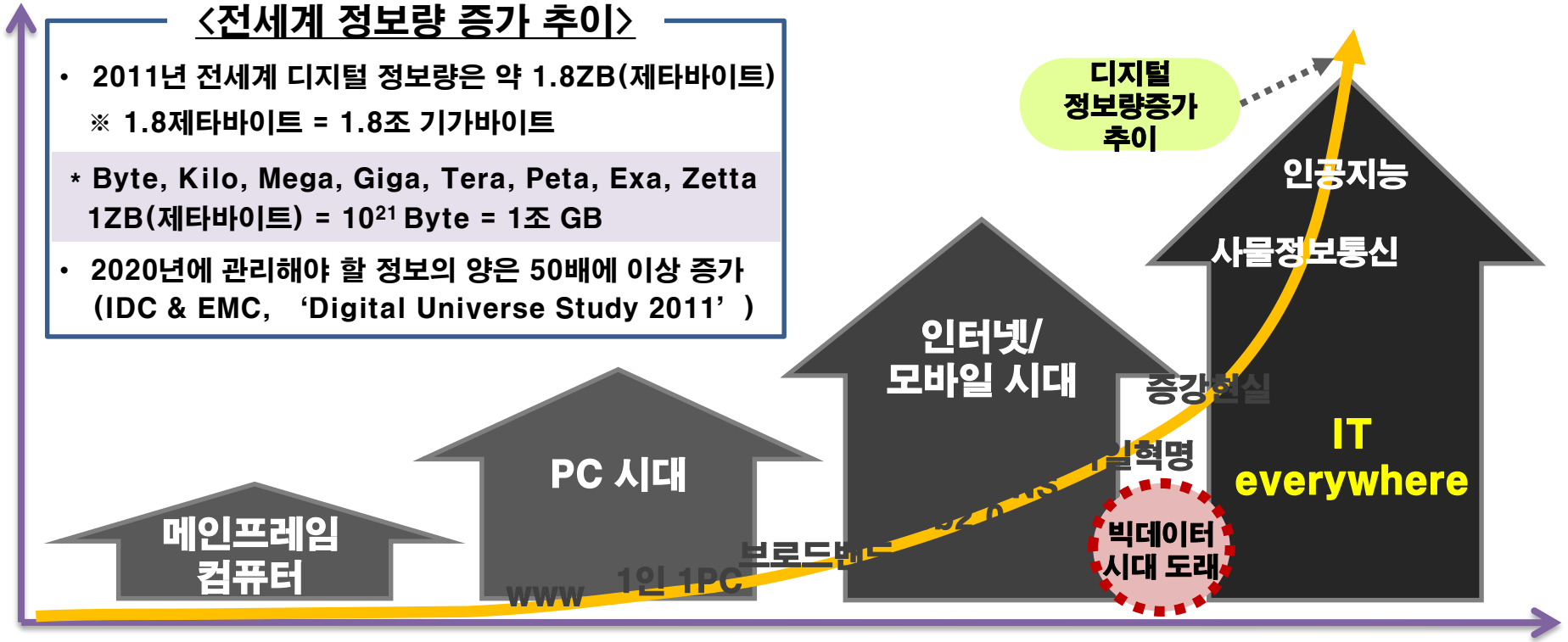

전달
참여 소통
분석 예측
폐쇄형 정보전달 홈페이지, 게시판
개방형 집단지성 블로그, WIKI, P2P
지능화 정보분석/예측 지식제공 웹
Web 1.0
Web 2.0
Web 3.0

<전세계 정보량 증가 추이>

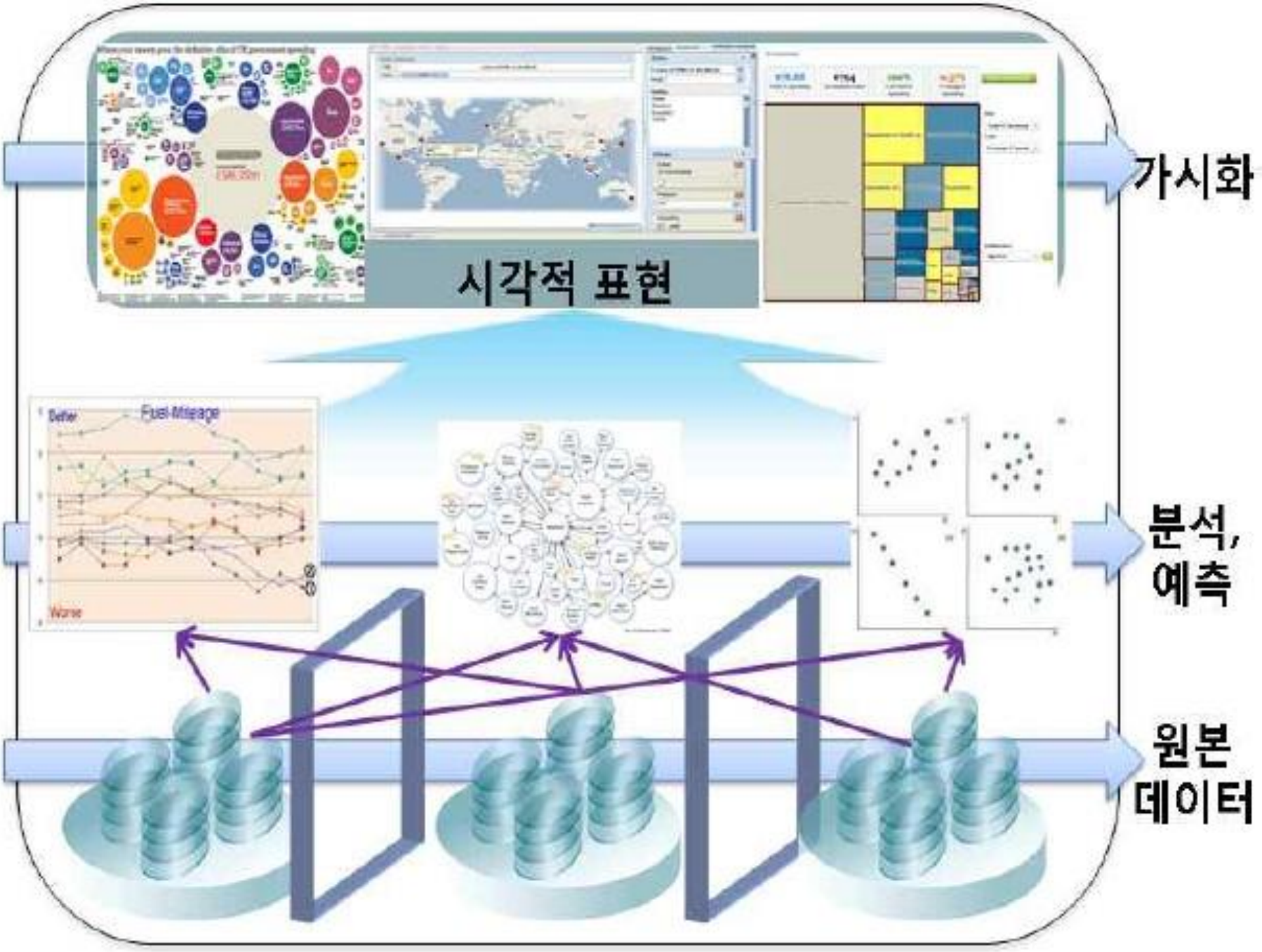
- 2011년 전세계 디지털 정보량은 약 1.8ZB(제타바이트)
  ※ 1.8제타바이트 = 1.8조 기가바이트

* Byte, Kilo, Mega, Giga, Tera, Peta, Exa, Zetta
  1ZB(제타바이트) = 10^{21} Byte = 1조 GB

- 2020년에 관리해야 할 정보의 양은 50배에 이상 증가
  (IDC & EMC, 'Digital Universe Study 2011')



데이터 규모	EB(Exa Byte) (90년대 말=100EB)	ZB(Zetta Byte) 진입 (2011년=1.8ZB)	ZB 본격화 시대 ('20년= '11년대비 50배 증가)
데이터 유형	정형 데이터 (데이터베이스, 사무정보)	비정형 데이터 (이메일, 멀티미디어, SNS)	사물정보, 인지정보 (RFID, Sensor, 사물통신)
데이터 특성	구조화	다양성, 복합성, 소셜	현실성, 실시간성


시각적 표현
가시화
분석,
예측
원본
데이터

BIG DATA의 사례

미국 국립보건원은 다양한 질병을 연구하기 위해 유전자 데이터를 공유 분석할 수 있는 유전자 데이터 공유를 통한 질병치료체계를 마련하여 주요 관리 대상에 해당하는 질병에 대한 관리 및 예측을 실시하고 있다. 현재 1,700명의 유전자 정보를 아마존 클라우드에 저장하여 누구나 데이터를 이용 가능하게 구축하였다(www.1000genomes.org/).

BIG DATA의 사례

구글 독감예보 서비스(구글 플루 트렌드; www.google.org/flutrends/)는 다양한 사용자의 검색어 분석을 통하여 사용자에게 다시 유의미한 데이터로 가공하여 정확한 정보를 실시간으로 제공하고 있다.

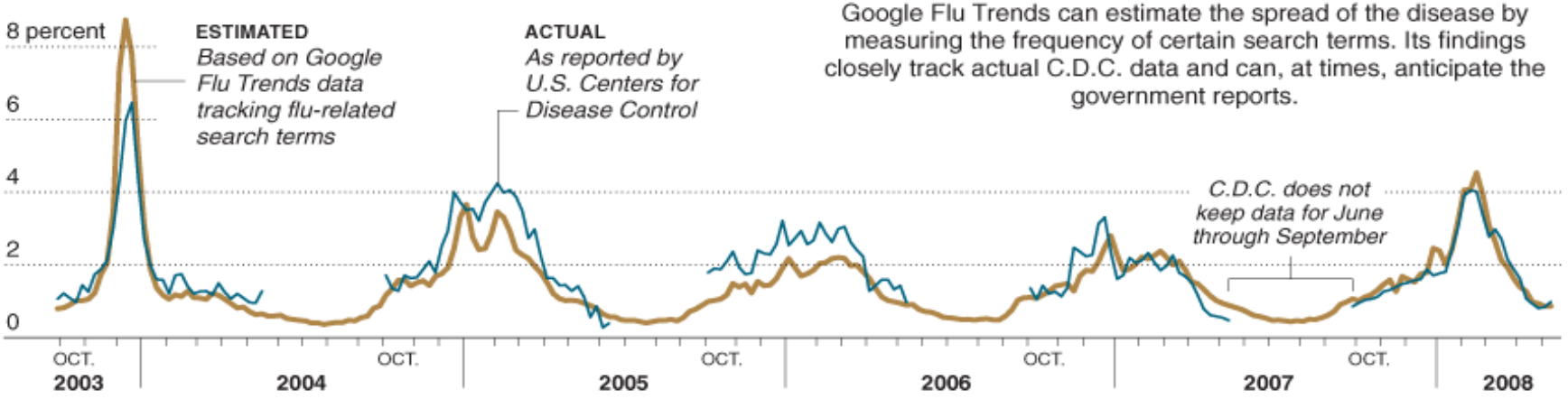
The New York Times

November 12, 2008

PERCENT OF HEALTH VISITS FOR FLU-LIKE SYMPTOMS *Mid-Atlantic region*

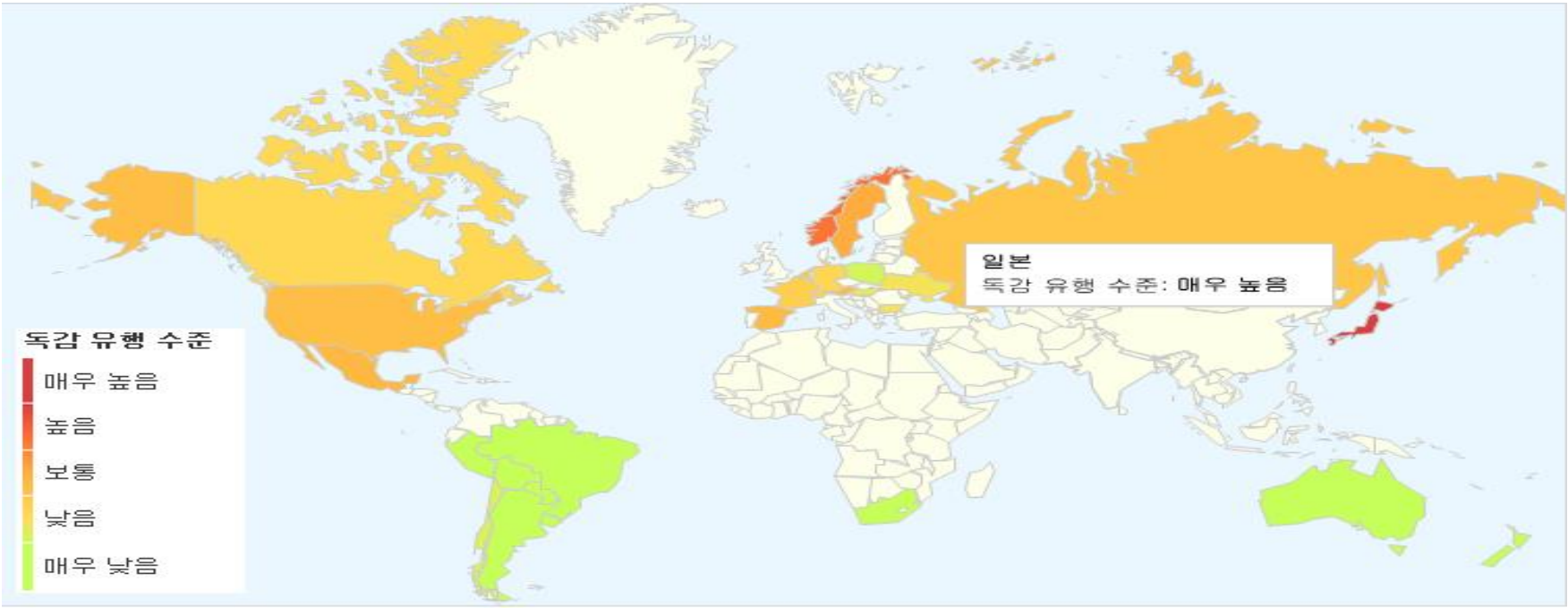
Using Google to Monitor the Flu

Google Flu Trends can estimate the spread of the disease by measuring the frequency of certain search terms. Its findings closely track actual C.D.C. data and can, at times, anticipate the government reports.



Sources: Google; Centers for Disease Control

THE NEW YORK TIMES



BIG DATA의 사례

보건 분야 국내 활용 사례로 질병관리본부에서 운영하는 한국인체자원은행네트워크(kbn.cdc.go.kr/)는 16개 병원을 통해 36만 명의 인체자원 확보하여 질병지표 발굴 및 질병조기 진단을 위해 활용하고 있다.

- 현재 국내 DNA 데이터의 보존•활용은 선진국의 약 1/100 수준에 불과함

BIG DATA에 관심을 가져야 하는 이유

정 보 통 신 기 술 (Information Communication Technology: ICT)이 다른 산업들과 융복합되면서 방대한 양의 데이터들이 생산되고 있는 가운데 사회변화에 따른 삶의 질에 대한 욕구 및 현안해결을 위해 빅데이터를 활용하고 있다.

BIG DATA에 관심을 가져야 하는 이유

빅데이터 활용 시 미국 의료분야에서 연 3,000억 달러, 유럽 공공분야에서 연 2,500억 달러의 경제적 효과가 있을 것으로 예측함. 우리나라는 약 10.7조의 정부지출을 감소시킬 것으로 예측함. 일본에서는 빅데이터의 활용이 촉진되면 부가가치의 창출이나 사회적 비용을 절감하여 총 16조 원이상의 경제적인 효과를 얻을 것으로 예상함

BIG DATA에 관심을 가져야 하는 이유

연구 및

보건정책의 수립에 이용

호흡기 질환 연구에서 BIG DATA 이용 사례

데이터 연계를 통한
새로운 코호트 구축

2001년도 국민건강영양조사와
건강보험청구자료를 연계함

INT J TUBERC LUNG DIS 15(6):824–829

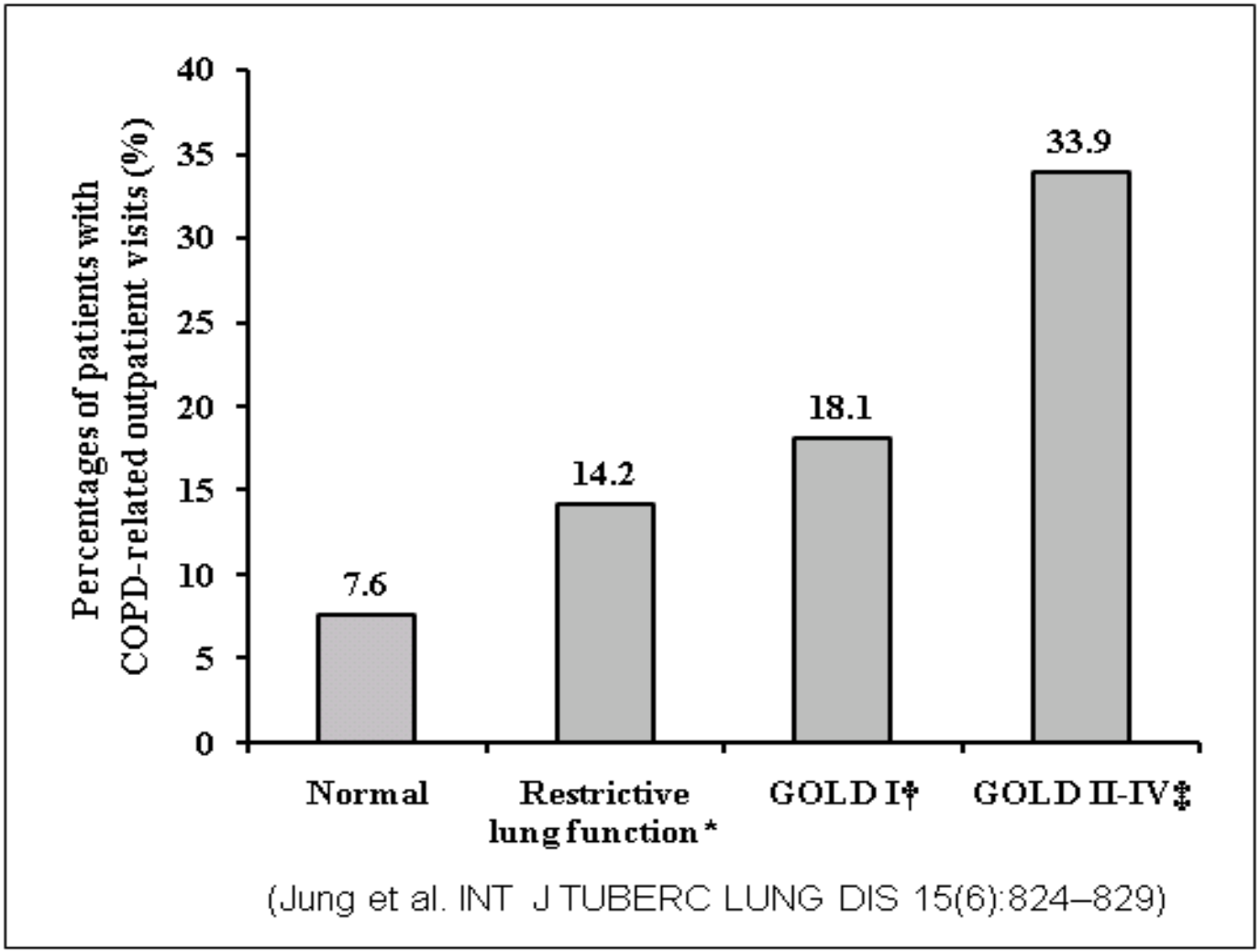
doi:10.5588/ijtld.10.0432

Chronic obstructive lung disease–related health care utilisation in Korean adults with obstructive lung disease

J. Y. Jung,* Y. A. Kang,* M. S. Park,* Y. M. Oh,† E. C. Park,‡ H. R. Kim,§ S. D. Lee,† S. K. Kim,* J. Chang,* Y. S. Kim*

*Department of Internal Medicine and Institute of Chest Diseases, Severance Hospital, Yonsei University College of Medicine, Seoul, †Department of Pulmonary and Critical Care Medicine and Clinical Research Center for Chronic Obstructive Airway Diseases, Asan Medical Center, University of Ulsan College of Medicine, Seoul, ‡Department of Preventive Medicine and Institute of Health Services Research, Yonsei University College of Medicine, Seoul, §Korea Institute for Health and Social Affairs, Seoul, Republic of Korea

INT J TUBERC LUNG DIS 15(6):824–829



2001년도에 COPD로 진단 받은 환자들의

5년간 의료이용에 대한 첫 번째 연구 결과

데이터 연계를 통한

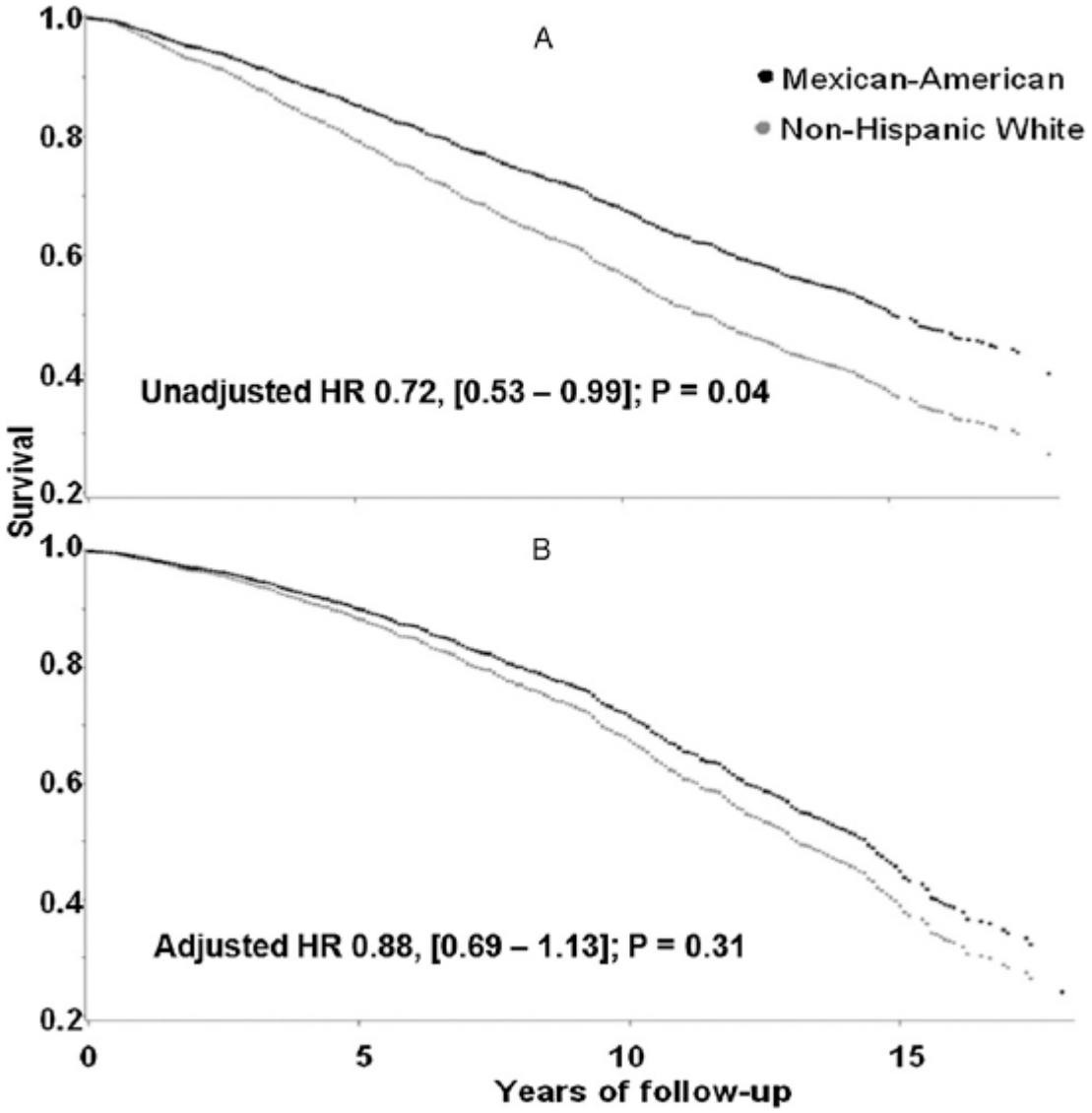
새로운 코호트 구축

Obstructive Lung Disease in Mexican Americans and Non-Hispanic Whites

An Analysis of Diagnosis and Survival in the National Health and Nutritional Examination Survey III Follow-up Study

Alejandro A. Diaz, MD, MPH; Carolyn E. Come, MD, MPH; David M. Mannino, MD, FCCP; Victor Pinto-Plata, MD, FCCP; Miguel J. Divo, MD; Carol Bigelow, PhD; Bartolome Celli, MD, FCCP; and George R. Washko, MD

CHEST 2014; 145(2):282–289



미국의 NHANES III에서 COPD로 진단 받은 성인의 사망률 추적 결과

건강보험청구자료를 이용한 의료이용에 대한 코호트 구축

1. The association between inhaled long-acting bronchodilators and less in-hospital care in newly-diagnosed COPD patients.

Respiratory Medicine (2014) 108, 153-161

2. The health care burden of high grade chronic obstructive pulmonary disease in Korea: analysis of the Korean Health Insurance Review and Assessment Service data.

International Journal of COPD 2013:8, 561-568

건강보험청구자료를 이용한 의료이용에 대한 코호트 구축

3. Medical Utilization and Cost in Patients with Overlap Syndrome of Chronic Obstructive Pulmonary Disease and Asthma.

COPD, 00:1–8, 2013

4. Association between chronic obstructive pulmonary disease and gastroesophageal reflux disease: a national cross-sectional cohort study.

BMC Pulmonary Medicine 2013, 13:51

BIG DATA를 이용한

새로운 코호트 구축

Pneumonia and pneumonia related mortality in patients with COPD treated with fixed combinations of inhaled corticosteroid and long acting β_2 agonist: observational matched cohort study (PATHOS)

Conclusions: There is an intra-class difference between fixed combinations of inhaled corticosteroid/long acting β2 agonist with regard to the risk of pneumonia and pneumonia related events in the treatment of patients with COPD.

BMJ 2013;346:f3306

BIG DATA를 이용한 새로운 코호트 구축

Study design, protocol, and data sources

: We carried out an observational retrospective cohort study, matched for propensity score, linking primary care medical records to data from national mandatory Swedish registries.

The Swedish National Board of Health and Welfare performed the data linkage.

BMJ 2013;346:f3306

BIG DATA를 이용한

새로운 코호트 구축

ORIGINAL ARTICLE

Use of inhaled corticosteroids and the risk of tuberculosis

Conclusions: ICS use increases the risk of TB in an intermediate-TB-burden country. Clinicians should be aware of the possibility of TB development among patients who are long-term high-dose ICS users.

Thorax 2013;68:1105–1113.

BIG DATA를 이용한 새로운 코호트 구축

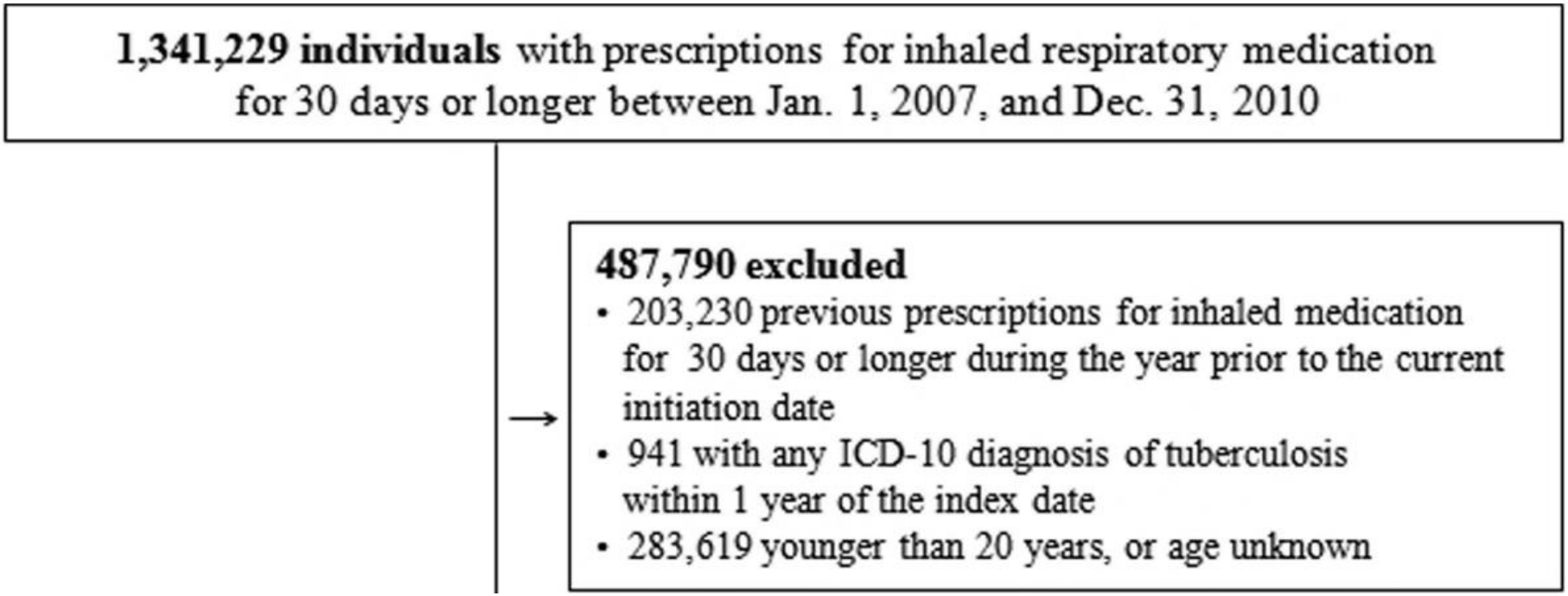
Study design

: A nested case-control study based on the HIRA database was conducted. The source population consisted of all individuals who were dispensed at least one of the following inhaled respiratory medications between 1 January 2007 and 31 December 2010:

Thorax 2013;68:1105–1113.

BIG DATA를 이용한

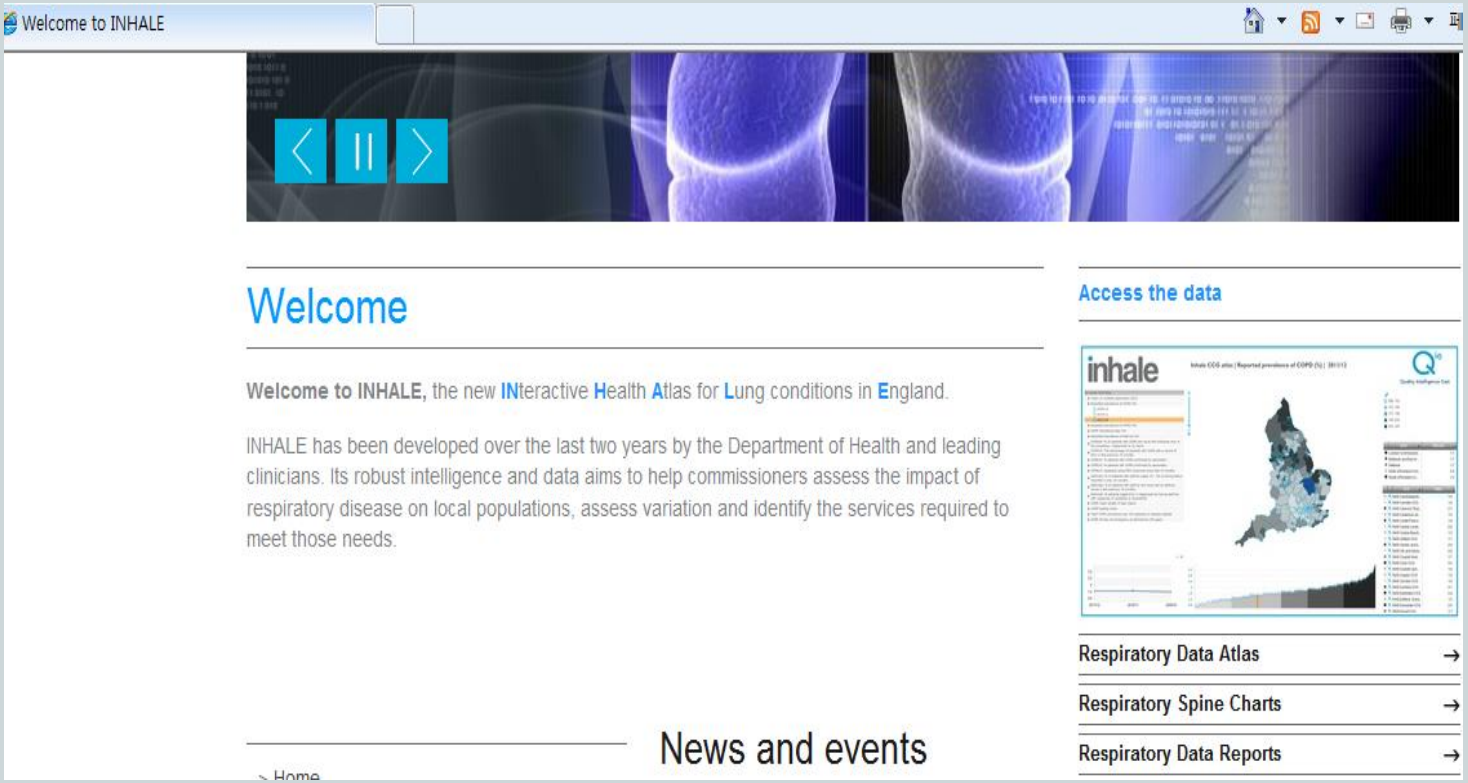
새로운 코호트 구축



Thorax 2013;68:1105–1113.

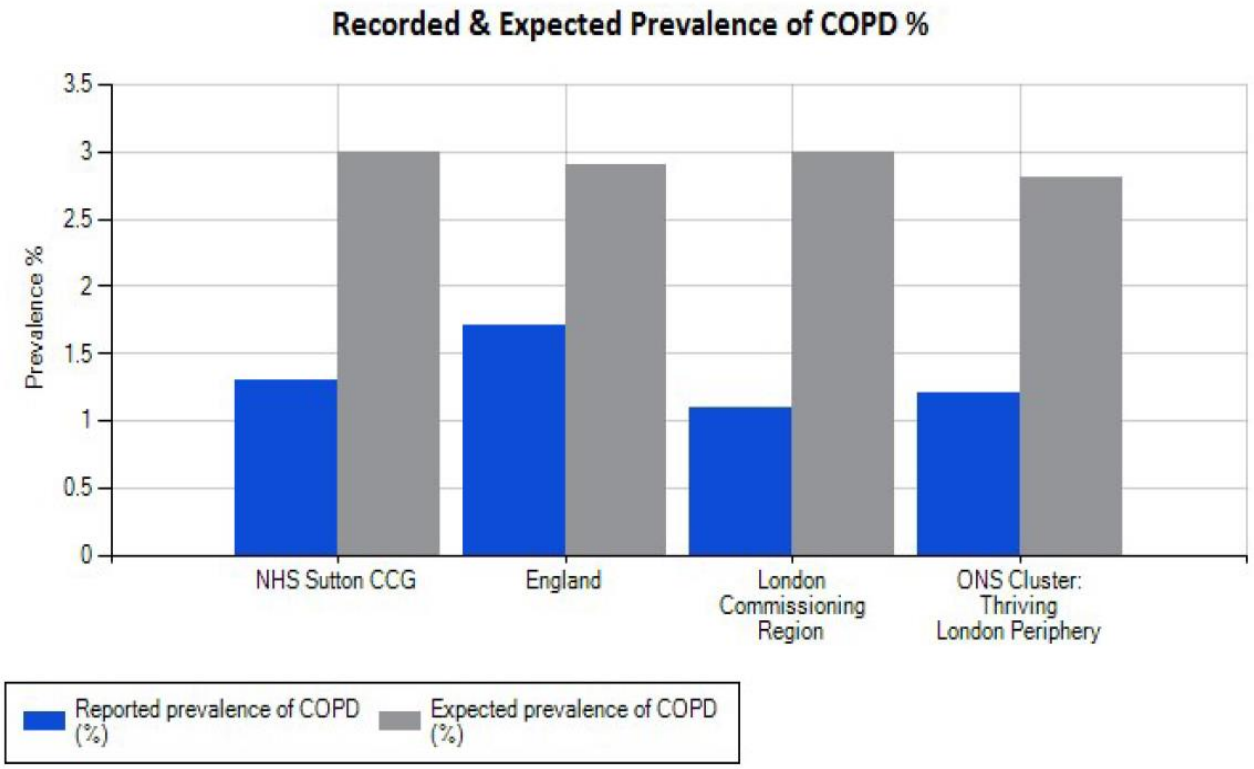
만성질환의 정도 관리지표 개발

영국의 INHALE (the new Interactive Health Atlas for Lung conditions in England)의 홈페이지 화면



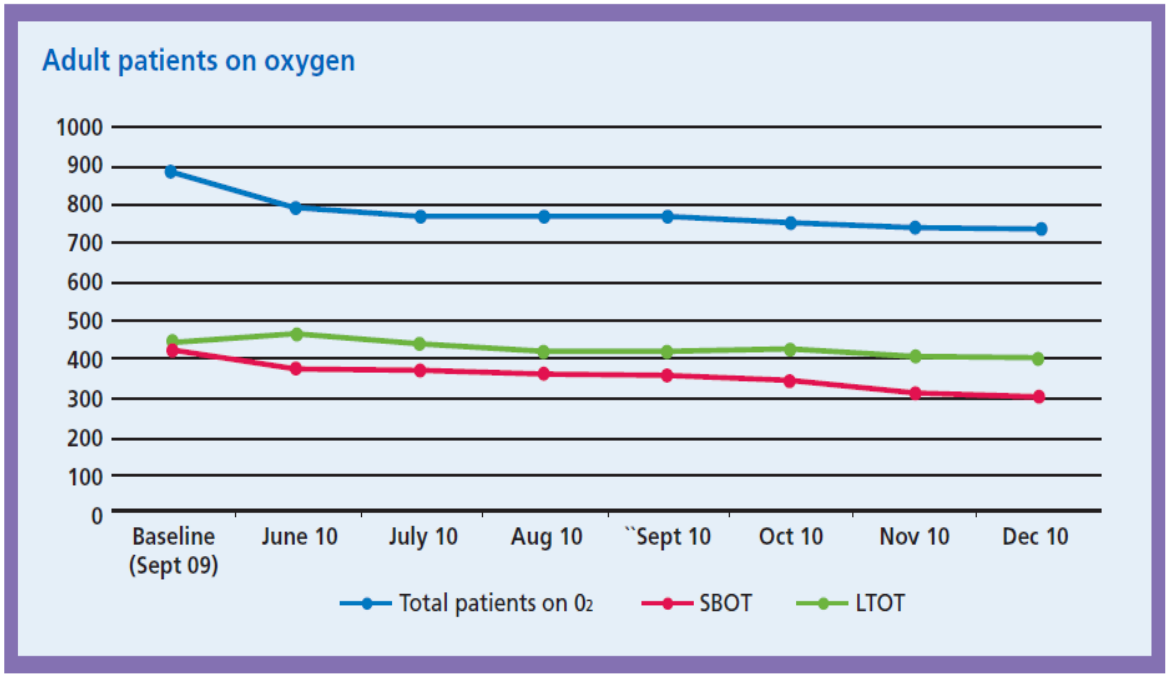
COPD 관리지표의 예: (1)
(DEGREE OF UNDER-DIAGNOSIS)

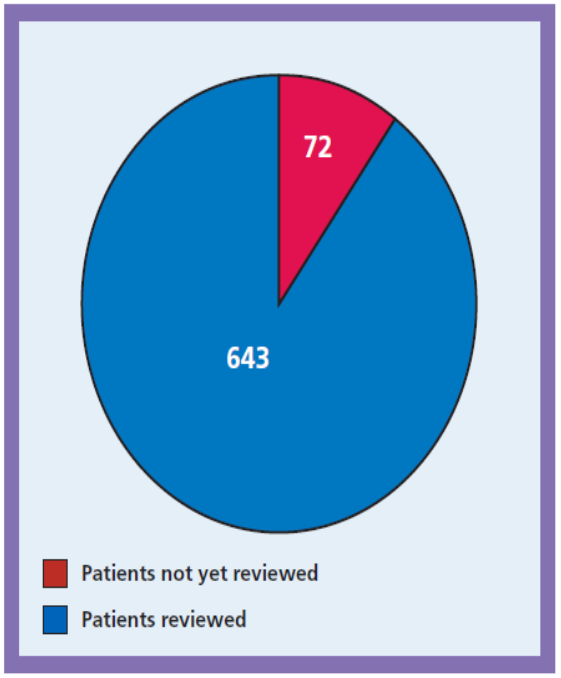
The gap between recorded and expected prevalence is a measure of the degree of under-diagnosis of COPD.



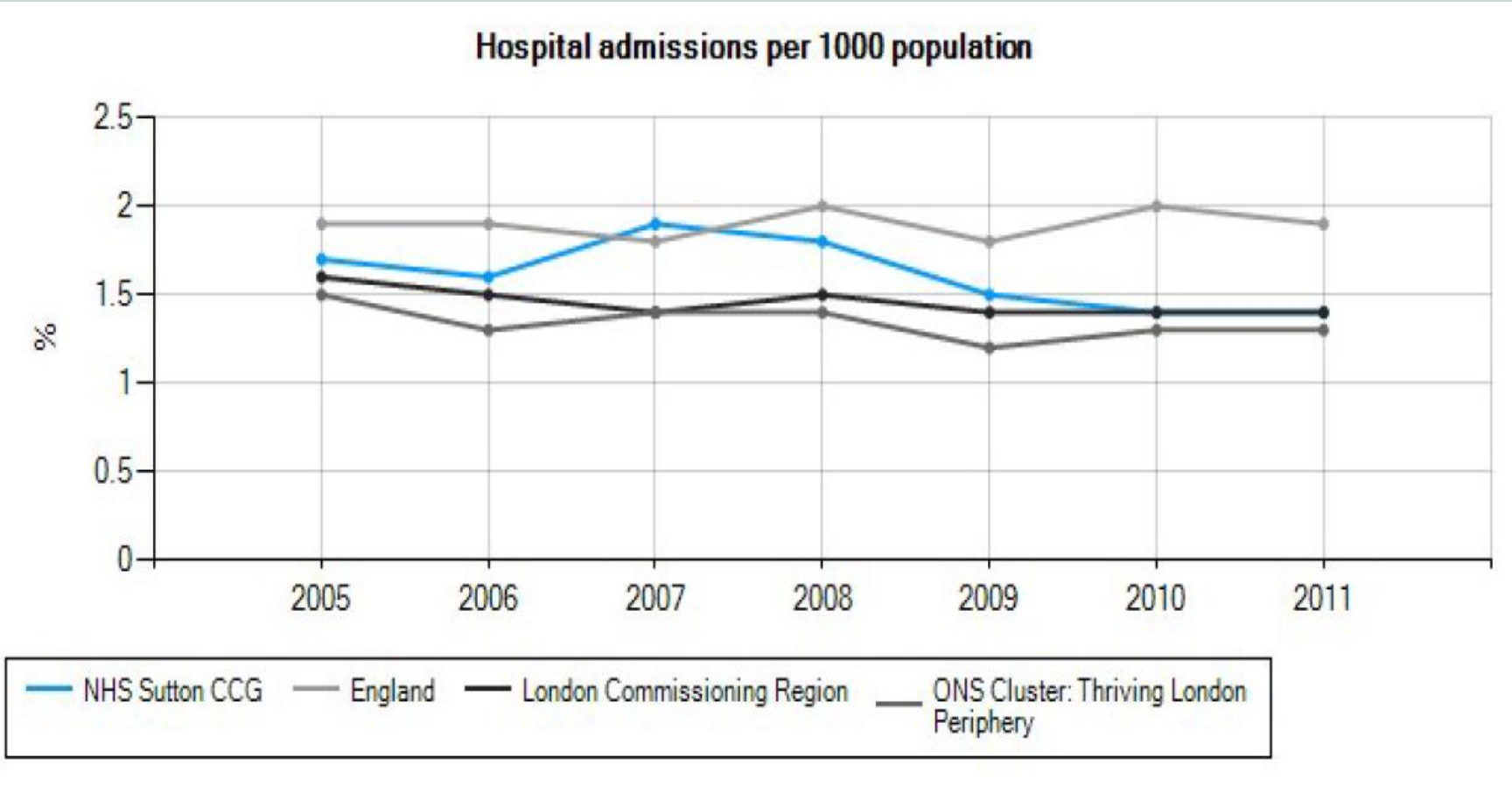
Source: Reported Prevalence: Information Centre. QOF. Estimated Prevalence: Eastern Region Public Health Observatory (ERPHO)

COPD 관리지표의 예 (2)
(ADULT PATIENT ON OXYGEN)

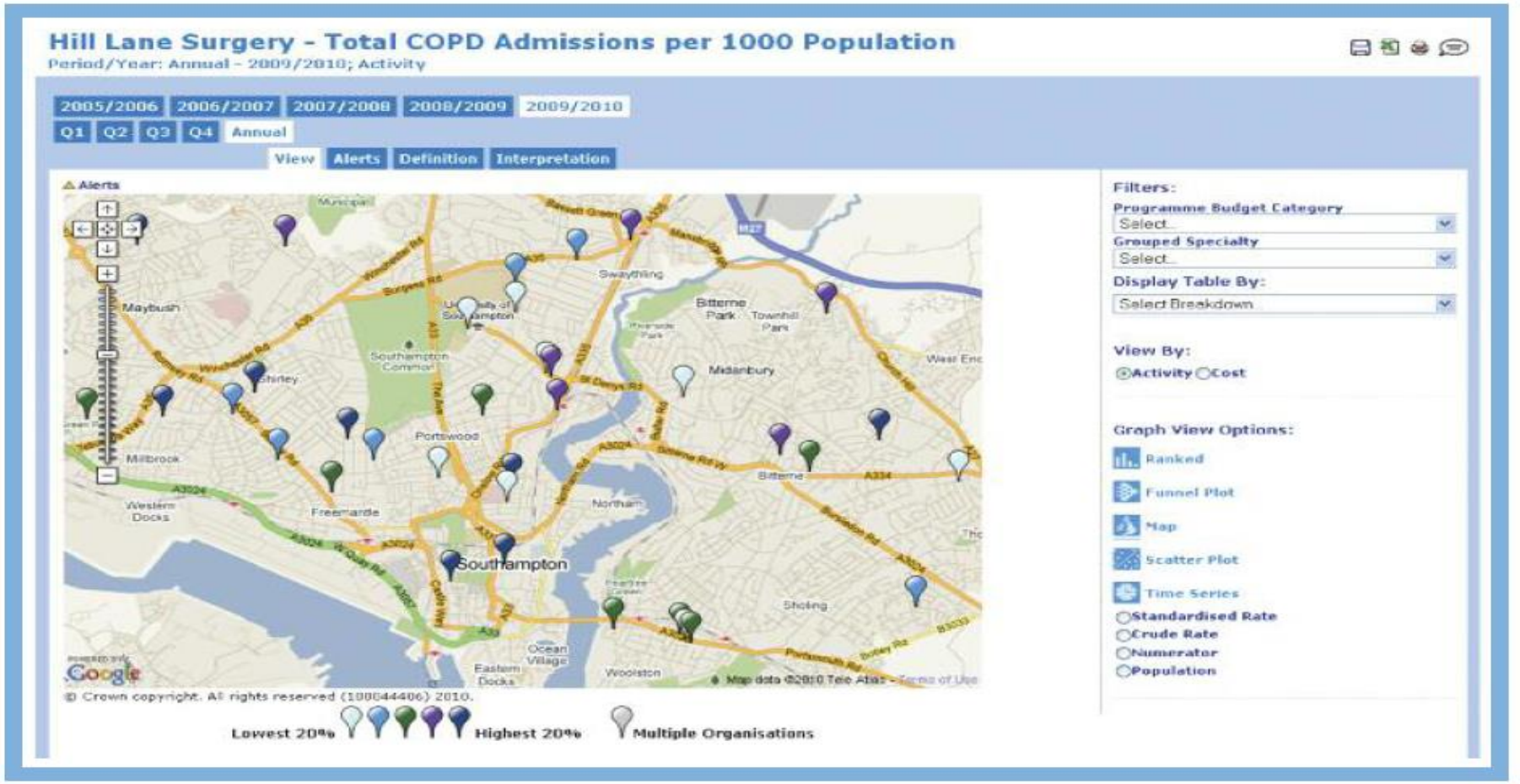




COPD 관리지표의 예 (3) (HOSPITAL ADMISSION)



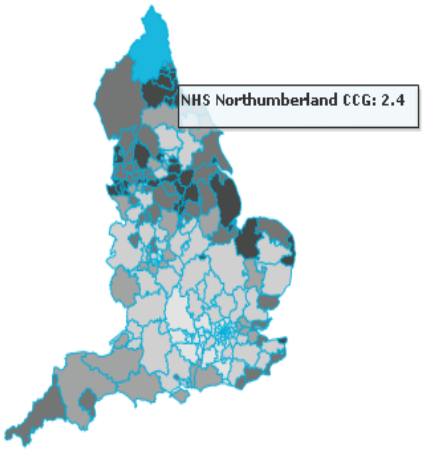
COPD 관리지표의 예 (3) (HOSPITAL ADMISSION)

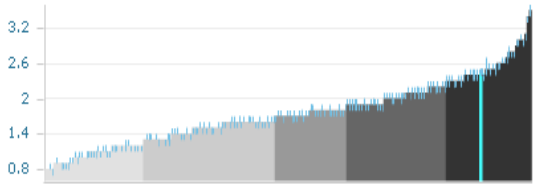


NHS Comparators mapping example

Reported prevalence of COPD (%) | 2011

Select Indicator | Clear | Print | Help



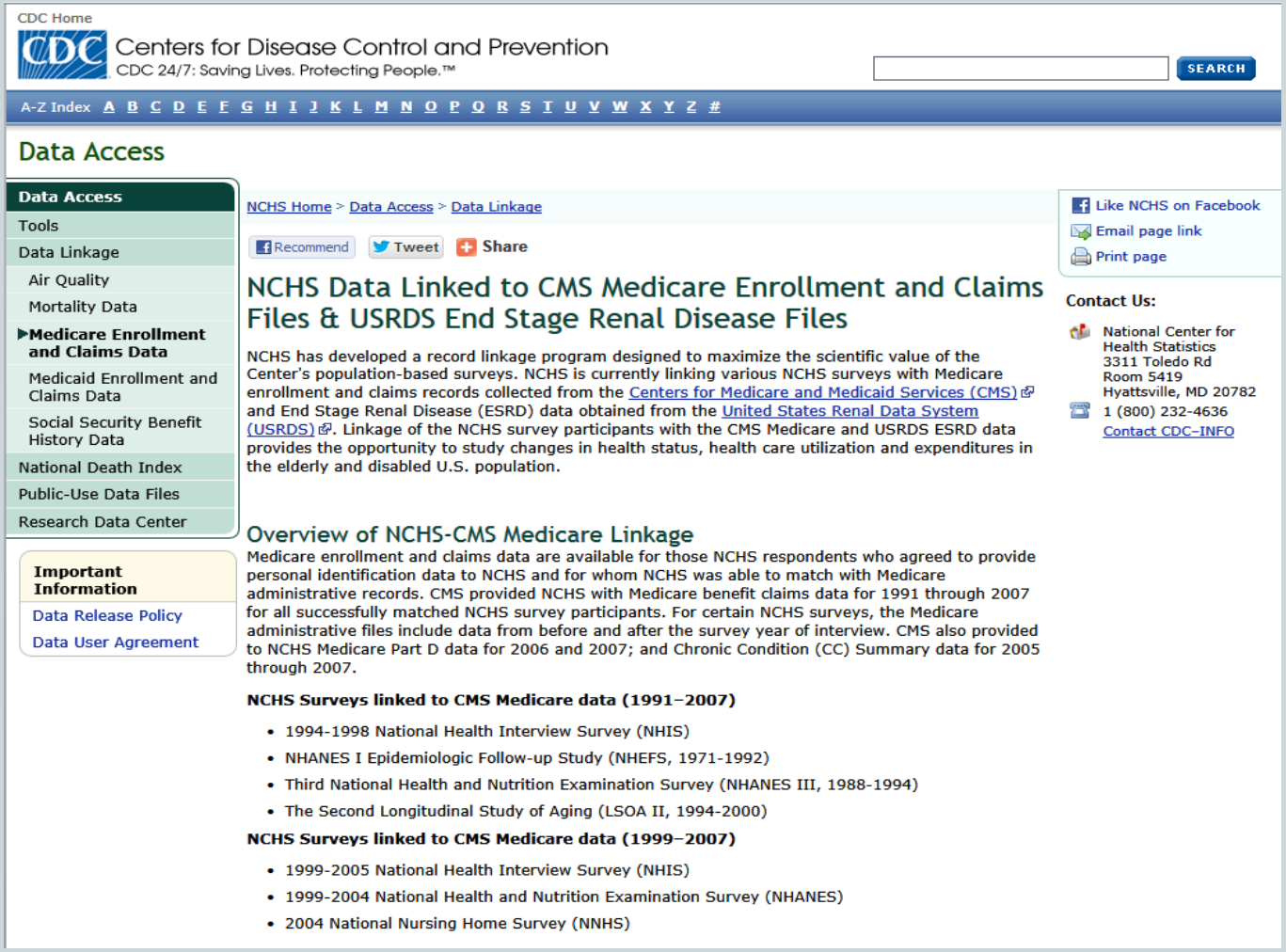


CCGs by DH 20120813_r
- 0.8 - 1.2
- 1.3 - 1.6
- 1.7 - 1.8
- 1.9 - 2.2

Indicator	Area	Num...	CCG ...	Engl...	England Worst	Current Performance	England Best
▼ Prevalence							
Reported prevalence of COPD (%)	NHS Northumberland CCG	7,747	2.4	1.7	3.5		0.8
Late Diagnosis - Observed/Expecte...	NHS Northumberland CCG	7,747	0.86	0.6	1.1		0.2
Expected prevalence of COPD %	NHS Northumberland CCG	8,982	2.8	2.9	5		1.8
▼ General Practice							
COPD10: The percentage of patient...	NHS Northumberland CCG	6,332	90.3	88.8	74.6		93.4
COPD12: % patients with COPD con...	NHS Northumberland CCG	1,461	90.4	89.7	81.4		94.5
COPD15: % patients with COPD con...	NHS Northumberland CCG	598	94.2	93.0	82.1		97.3
Smoking status recorded in last 15 ...	NHS Northumberland CCG	81,285	96.0	95.4	93.5		97.5
Smoking cessation advice/referral o...	NHS Northumberland CCG	12,676	93.3	92.9	88.6		96.4
COPD08: % of patients with COPD w...	NHS Northumberland CCG	6,351	92.8	93.1	88		96.8
Exception rate for COPD indicators	NHS Northumberland CCG	2,438	10.2	11.8	6.2		17.5
COPD13: assessed using MRC dysp...	NHS Northumberland CCG	6,538	92.2	91.8	85.6		94.5
▼ Secondary Care							
COPD 30 Day emergency re-admiss...	NHS Northumberland CCG		24.2	21.2	13.1		29.6
COPD admissions per 100 patients ...	NHS Northumberland CCG	995	13.3	12.6	21.7		7.8
Emergency COPD Admissions per 1...	NHS Northumberland CCG	976	13.1	12.0	19		7.1
COPD admissions per 1,000 popula...	NHS Northumberland CCG	1,070	3.3	2.0	5.1		0.8
COPD emergency admissions per 1,...	NHS Northumberland CCG	1,034	3.2	1.9	4.6		0.7
Mean Length of Stay in Hospital for ...	NHS Northumberland CCG	7,445	7.3	6.0			
Mean Length of Stay in Hospital for ...	NHS Northumberland CCG	7,623	7.4	6.1	8.7		3.1
COPD 30 Day emergency re-admiss...	NHS Northumberland CCG	275	24.2	21.2	13.1		29.6
▼ Spend							
Average cost per emergency hospit...	NHS Northumberland CCG	46,681	2,301.9	2,288.1	2,850.9		1,616.3

Significance compared with England average: worse ● better ● no difference ● could not be calculated ★
England Average | Commissioning Board average ◆ ONS Cluster average ★
Quartile 0 to Quartile 1 Quartile 1 to Quartile 3 Quartile 3 to Quartile 4
− +

DATA 연계를 통한 BIG DATA의 구축 (미국)



CDC Home

Centers for Disease Control and Prevention
CDC 24/7: Saving Lives. Protecting People.™

SEARCH

A-Z Index A B C D E F G H I J K L M N O P Q R S T U V W X Y Z #

Data Access

- Data Access
- Tools
- Data Linkage
  - Air Quality
  - Mortality Data
  - ▶Medicare Enrollment and Claims Data
  - Medicaid Enrollment and Claims Data
  - Social Security Benefit History Data
- National Death Index
- Public-Use Data Files
- Research Data Center

Important Information

- Data Release Policy
- Data User Agreement

NCHS Home > Data Access > Data Linkage

Recommend Tweet Share

NCHS Data Linked to CMS Medicare Enrollment and Claims Files & USRDS End Stage Renal Disease Files

NCHS has developed a record linkage program designed to maximize the scientific value of the Center's population-based surveys. NCHS is currently linking various NCHS surveys with Medicare enrollment and claims records collected from the Centers for Medicare and Medicaid Services (CMS) and End Stage Renal Disease (ESRD) data obtained from the United States Renal Data System (USRDS). Linkage of the NCHS survey participants with the CMS Medicare and USRDS ESRD data provides the opportunity to study changes in health status, health care utilization and expenditures in the elderly and disabled U.S. population.

Overview of NCHS-CMS Medicare Linkage

Medicare enrollment and claims data are available for those NCHS respondents who agreed to provide personal identification data to NCHS and for whom NCHS was able to match with Medicare administrative records. CMS provided NCHS with Medicare benefit claims data for 1991 through 2007 for all successfully matched NCHS survey participants. For certain NCHS surveys, the Medicare administrative files include data from before and after the survey year of interview. CMS also provided to NCHS Medicare Part D data for 2006 and 2007; and Chronic Condition (CC) Summary data for 2005 through 2007.

NCHS Surveys linked to CMS Medicare data (1991–2007)

- 1994-1998 National Health Interview Survey (NHIS)
- NHANES I Epidemiologic Follow-up Study (NHEFS, 1971-1992)
- Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994)
- The Second Longitudinal Study of Aging (LSOA II, 1994-2000)

NCHS Surveys linked to CMS Medicare data (1999–2007)

- 1999-2005 National Health Interview Survey (NHIS)
- 1999-2004 National Health and Nutrition Examination Survey (NHANES)
- 2004 National Nursing Home Survey (NNHS)

Like NCHS on Facebook
Email page link
Print page

Contact Us:

National Center for Health Statistics
3311 Toledo Rd
Room 5419
Hyattsville, MD 20782
1 (800) 232-4636
Contact CDC–INFO

미국의 DATA 연계를 통한 BIG DATA의 구축

	Total Person Sample (Column 1)	Sample Eligible for the 1999-2007 Medicare Linkage[2] (Column 2)	Respondents with Information on any Medicare Denominator File[3] (Column 3)	Linkage Rate for Total Sample (Column 3/Column 1)	Linkage Rate for Eligible Sample (Column 3/Column 2)
NHIS 1994	116,179	87,079	23,819	20.5%	27.4%
<65	89,116	65,794	2,938	3.3%	4.5%
>=65	27,063	21,285	20,881	77.2%	98.1%
NHIS 1995	102,467	73,809	18,930	18.5%	25.6%
<65	80,861	57,113	2,535	3.1%	4.4%
>=65	21,606	16,696	16,395	75.9%	98.2%

이에 비해 한국의 연계율은 높다!!

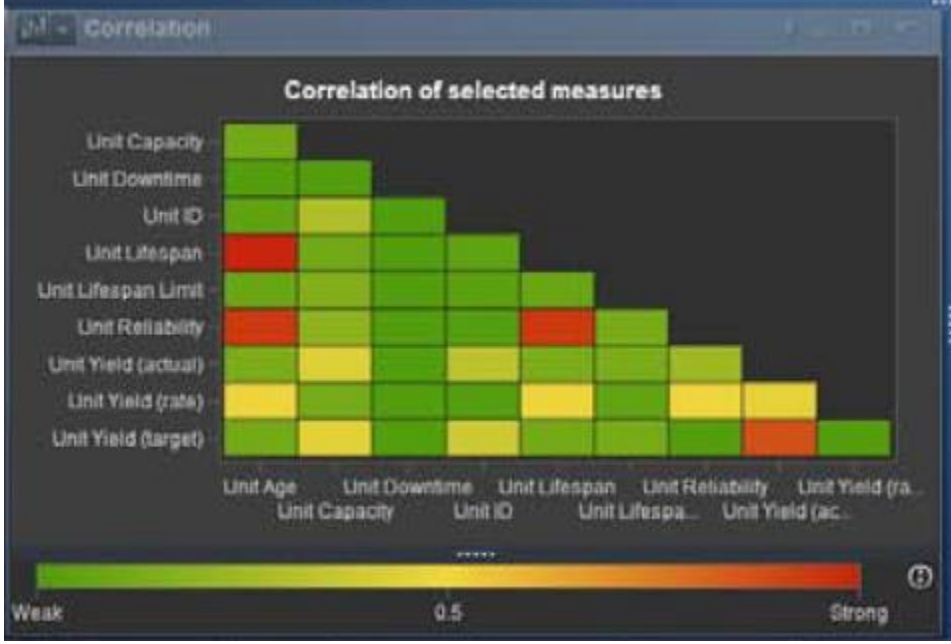
우리가

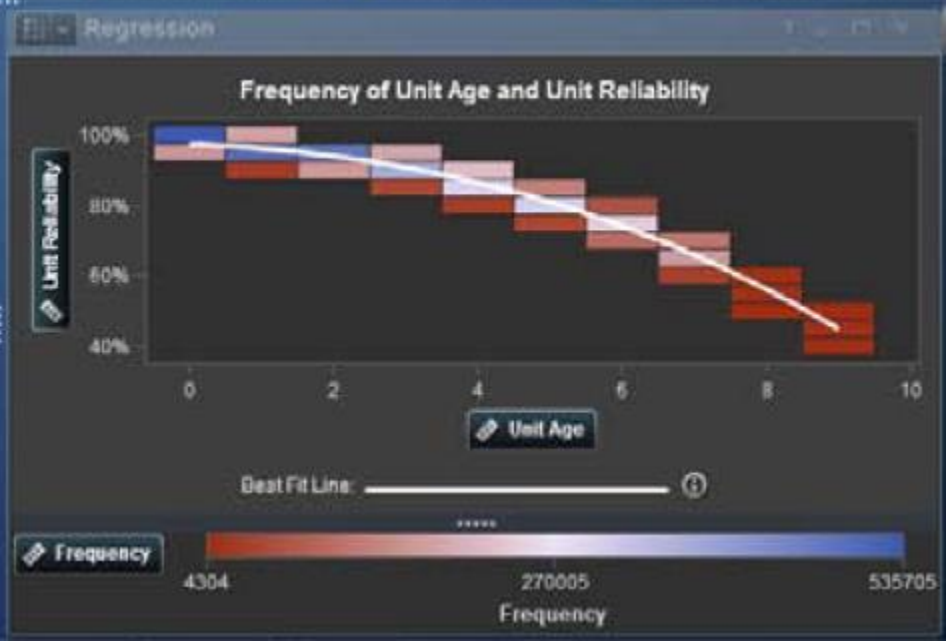
나아가야 할 길


Expenses
Crosstab
Line w/ Forecast
BoxPlot
Correlation
Regression
Geomap
Analytics
Drop controls here to create a section prompt
Novelty
Toy
Expenses Report
sas
THE POWER TO KNOW.
East
North
South
West
$9.60
$5.78
$1.96
($1.85)
$1.75
$3.11
$6.00
$2.79
Expenses (millions)
$6
$4
$2
$0
Action Figure
Game
Stuffed Animal
Product Line
Expenses (capital) / Expenses (material) / Expenses (operational) / Expenses (staffing)
$600,000
$400,000
$300,000
$200,000
$100,000
$0
Jan2009
Jul2009
Jan2010
Jul2010
Jan2011
Jul2011
Date by Month
Expenses (capital)
Expenses (material)
Expenses (operational)
Expenses (staffing)
Facility Region: All > South
Santa Fe
Amarillo
Tulsa
OKLAHOMA
ARKANSAS
TENNESSEE
Asheville
NORTH CAROLINA
SOUTH CAROLINA
Atlanta
Lubbock
Fort Worth
Shreveport
MISSISSIPPI
GEORGIA
Odessa
Ciudad Juárez
Biloxi
Destin
Gainesville
Del Rio
Seguin
Corpus Christi
FLORIDA
Expenses
$5,997,097
OpenStreetMap
Facility State
AL
AR
LA
TX
Date by Year
2009
2010
2011
Total
Product
Expenses
$1,695,655.70
$1,657,346.44
$1,970,378.03
Athlete
$166,848.67
$191,490.31
$195,295.38
Bear
$2,559.41
$4,461.40
Big Cats
$2,487.42
$3,318.04
Board
$581,578.04
$657,453.59
$674,829.36
Card
$341,339.82
$401,967.85
$384,145.30
Cat
$1,641.63
$4,376.06
Dog
$1,500.43
$4,746.87
Elephant
$1,362.00
$5,736.88
Firefighter
$180,584.04
$167,467.05
$181,329.19
Horse
$2,167.59
$3,799.98
Movie Star
$172,771.39
$178,765.18
$178,017.75
Musician
$185,642.07
$189,999.57
$185,782.48
Police
$172,516.40
$169,709.32
$184,955.00
Primate

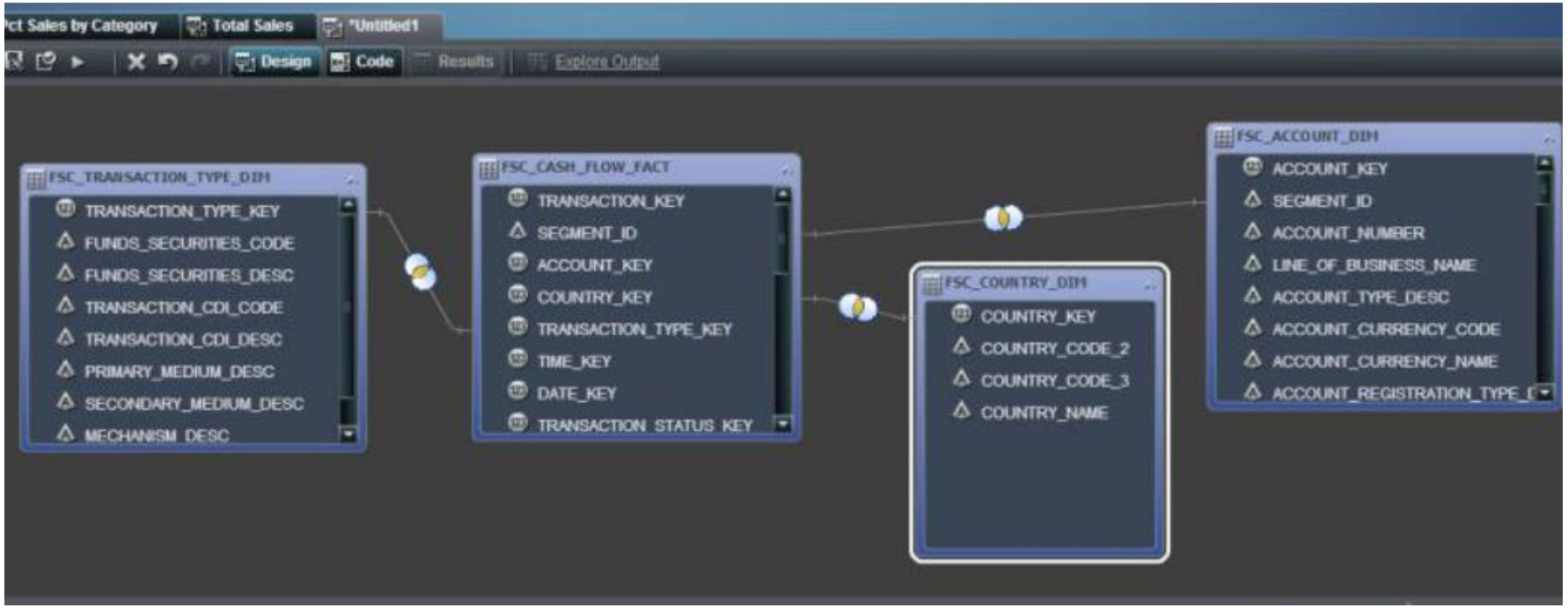

BoxPlot
All Geo
Unit Yield (rate), Unit Reliability by Geo
100%
80%
60%
40%
20%
Unit Yield (rate)
Unit Reliability
East
North
South
West
Geo
Unit Reliability
Unit Yield (rate)


Line w/ Forecast
Revenue, Profit, Expenses by Date by Year
Revenue (S...
Profit (Sum)
Expenses (...
(millions)
$20
$10
$0
($10)
1980
1985
1990
1995
2000
2005
2010
2015
Date by Year
Expenses
Profit
Revenue
Revenue (95% confidence)
Profit (95% confidence)
Expenses (95% confidence)


Correlation
Correlation of selected measures
Unit Capacity
Unit Downtime
Unit ID
Unit Lifespan
Unit Lifespan Limit
Unit Reliability
Unit Yield (actual)
Unit Yield (rate)
Unit Yield (target)
Unit Age
Unit Capacity
Unit Downtime
Unit ID
Unit Lifespan
Unit Lifespa...
Unit Reliability
Unit Yield (ra...
Unit Yield (ac...
Weak
0.5
Strong

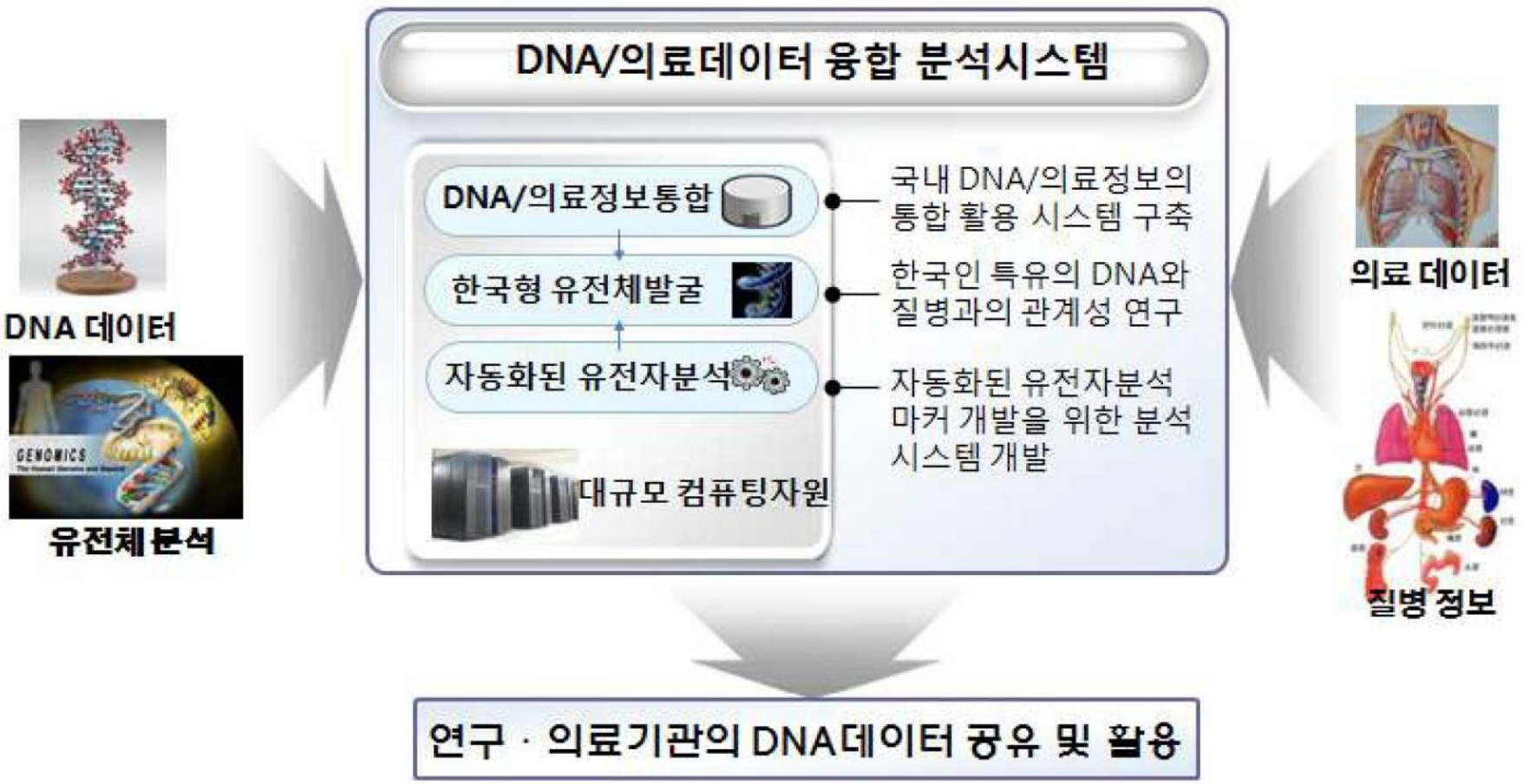

Regression
Frequency of Unit Age and Unit Reliability
Unit Reliability
100%
80%
60%
40%
0
2
4
6
8
10
Unit Age
Best Fit Line:
Frequency
4304
270005
535705
Frequency

BIG DATA의 연계와 분석이 힘든 일이 아님



분석하는 단계에서 실시간 예측하는 시대로 진입하고 있음

BIG DATA의 연계를 위한 시도와 부처간의 협력이 증가하고 있다.



현재 우리의 문제점은?

가장 좋은 연구 환경을 가지고 있었지만

- 관심의 부재
- 전문가의 부재
- 협력관계의 부재

최근에 관심과 지원이

증가하고 있음 !!

준비된 자, 준비된 집단만이

기회가 주어 졌을 때

성공할 수 있고 앞서 나갈 수 있다.

감사합니다.