



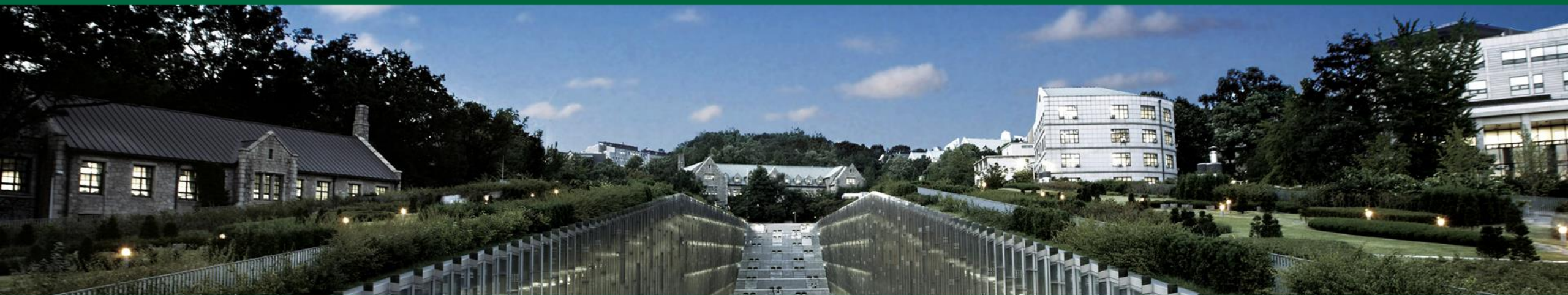
의료빅데이터의 개념, 현황, 활용전략

이화여대 의과대학 환경의학교실 김이준

강연자 소개







- ✓ 이화여대 의대 졸업
- ✓ 방사선종양학과 전문의
- ✓ 전 서울대학교병원 정밀의료센터 연구교수 (Big data, genomics)
- ✓ 전 이대목동병원 융합의학연구원 임상조교수 (Machine learning)
- ✓ 전 하버드의대 브리검여성병원 방문연구원 (Single cell sequencing)
- ✓ 현 이대의대 환경의학교실 조교수 (Wet lab & Dry lab)

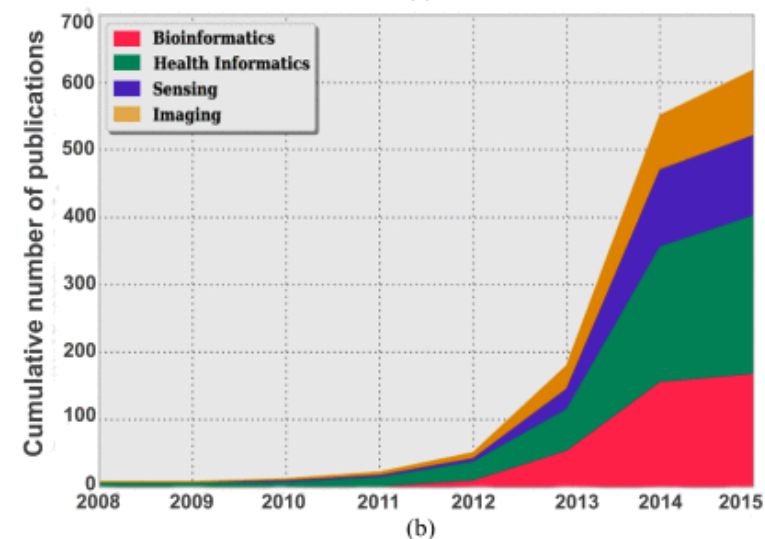
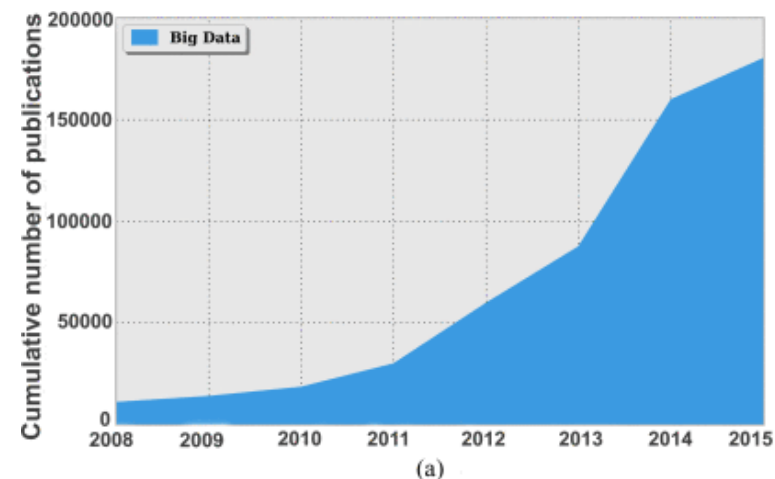
- # 지금까지 다루어 본 의료빅데이터
- 병원 임상 데이터
 - 미국암등록사업데이터 (SEER)
 - 건강보험공단데이터
 - 심사평가원데이터
 - CDW 데이터
 - CDM 데이터
 - TCGA 데이터 (RNA)
 - GEO 데이터
 - Bulk seq 데이터 (From Bcl files~)
 - Single cell seq 데이터
 - 등...



의료 빅데이터의 개념

What is big data?

Value		Clinically relevant data Longitudinal studies
Volume		High-throughput technologies Continuous monitoring of vital signs
Velocity		High-speed processing for fast clinical decision support Increasing data generation rate by the health infrastructure
Variety		Heterogeneous and unstructured data sources Differences in frequencies and taxonomies
Veracity		Data quality is unreliable Data coming from uncontrolled environments
Variability		Seasonal health effects and disease evolution Non-deterministic models of illness and health



Andreu-Perez et al. 2015 doi: 10.1109/JBHI.2015.2450362.

의료 빅데이터의 종류

임상 데이터

- ✓정형
- ✓비정형
 - 줄글
 - 영상
 - 음성

유전체 데이터

- ✓DNA 돌연변이
- ✓RNA 유전체 발현량
- ✓단백질 발현량
- ✓Pharmacogenomics

Life-log 데이터

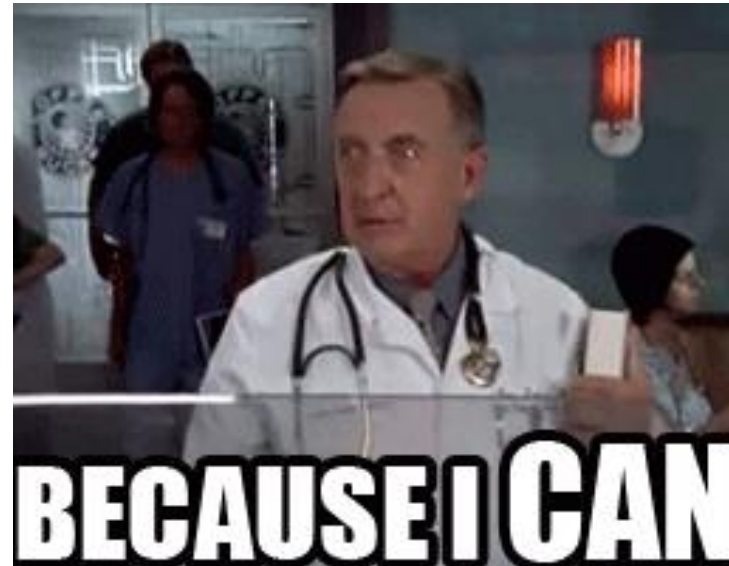
- ✓시계 심박동 체크
- ✓실시간 혈당 체크

사회적 데이터

- ✓미세먼지
- ✓소셜 네트워크



왜 빅데이터의 시대가 도래하였는가?



컴퓨터 성능의 발전

- 메모리
- 프로세서
- 클라우드 시스템

정형 데이터의 전자화

- EMR
- 정부 데이터
- 병원간 데이터 (OMOP-CDM)

비정형 데이터의 정형화

- Natural Language Process (MLP)
- 영상 데이터 처리
- 음성 데이터 처리

데이터의 양은 왜 중요한가?

- ✓ 단순한 산술적인 다다익선 (多多益善)
 - 예전에는 모집단에 대한 표본 연구를 했었음
 - 이제는 모든 환자를 통계처리할 수 있음
 - 예. 미국 암환자등록사업(SEER)은 암환자는 의무적으로 등록됨. 이 기록에서 특정 rare disease 환자 기록을 모두 사용한다면 미국내 모든 환자 기록을 활용하여 분석하는 것임.

- 작은 sample size 의 한계가 없어.
 - 예. "연구 결과가 어떠한 경향성을 보이거나 limitation of sample size 때문에 통계적 유의성에 도달하지 못했다..."



Vs.



Real-world evidence (RWE), Real-world data (RWD)

Real-World Evidence



미국 FDA의 정의

Real-world data (RWD) and real-world evidence (RWE) are playing an increasing role in health care decisions.

- FDA uses RWD and RWE to **monitor postmarket safety** and adverse events and to make regulatory decisions.
- The health care community is using these data to support coverage decisions and to **develop guidelines** and **decision support tools** for use in clinical practice.
- Medical product developers are using RWD and RWE to **support clinical trial designs** (e.g., large simple trials, pragmatic clinical trials) and **observational studies** to generate innovative, new treatment approaches.

The 21st Century Cures Act, passed in 2016, places additional focus on the use of these types of data to **support regulatory decision making**, including approval of new indications for approved drugs. Congress defined RWE as data regarding the usage, or the potential benefits or risks, of a drug derived from sources other than traditional clinical trials. FDA has expanded on this definition as discussed below.

Why is this happening now?

The use of **computers, mobile devices, wearables and other biosensors** to gather and store huge amounts of health-related data has been rapidly accelerating. This data holds potential to allow us to better design and conduct clinical trials and studies in the health care setting to answer questions previously though infeasible. In addition, with the development of sophisticated, new analytical capabilities, we are better able to analyze these data and apply the results of our analyses to medical product development and approval.

What are RWD and where do they come from?

Real-world **data** are the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. **RWD** can come from a number of sources, for example:

- **Electronic health records (EHRs)**
- **Claims and billing activities**
- Product and disease registries
- Patient-generated data including in home-use settings
- Data gathered from other sources that can inform on health status, such as **mobile devices**

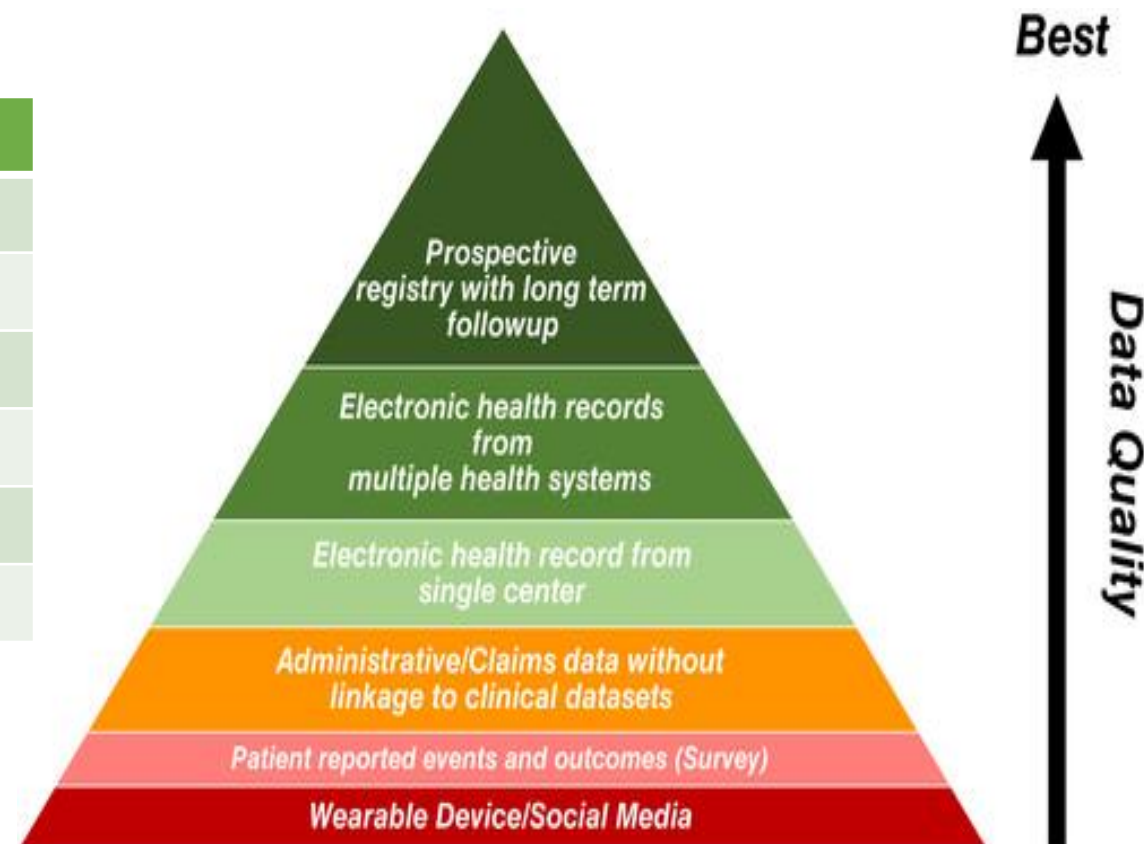
What is RWE?

Real-world **evidence** is **the clinical evidence** regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD. RWE can be generated by different study designs or analyses, including but not limited to, randomized trials, including large simple trials, pragmatic trials, and observational studies (prospective and/or retrospective).

<https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>

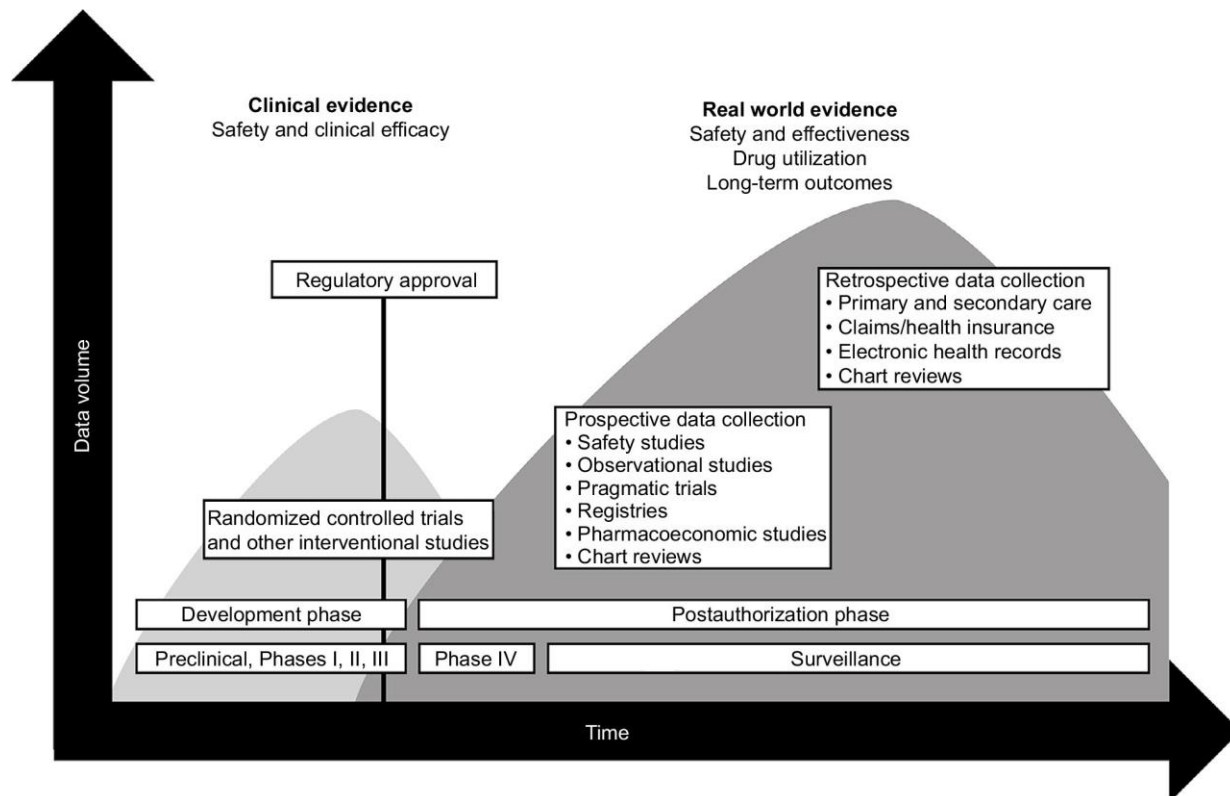
Relationship between sources of real world data and the ability to control for confounding variables

Rank	데이터 종류
1	장기 추적 관찰한 전향적 등록 데이터
2	다기관의 EMR 데이터
3	단일 기관의 EMR 데이터
4	보험청구데이터 (임상데이터셋과 연결 안된)
5	설문조사
6	웨어러블 디바이스/SNS



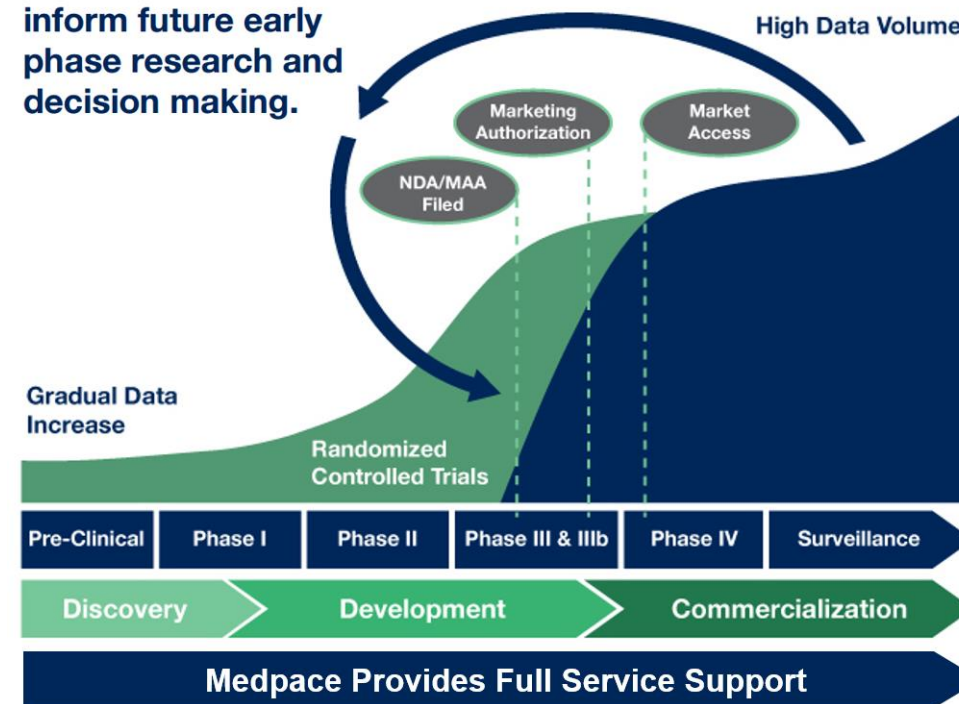
O'Leary et al. Diabetes, Obesity and Metabolism. 2020 Apr;22:3-12.

RCT 와 RWD/RWE



J Multidiscip Healthc. 2018;11:295-304

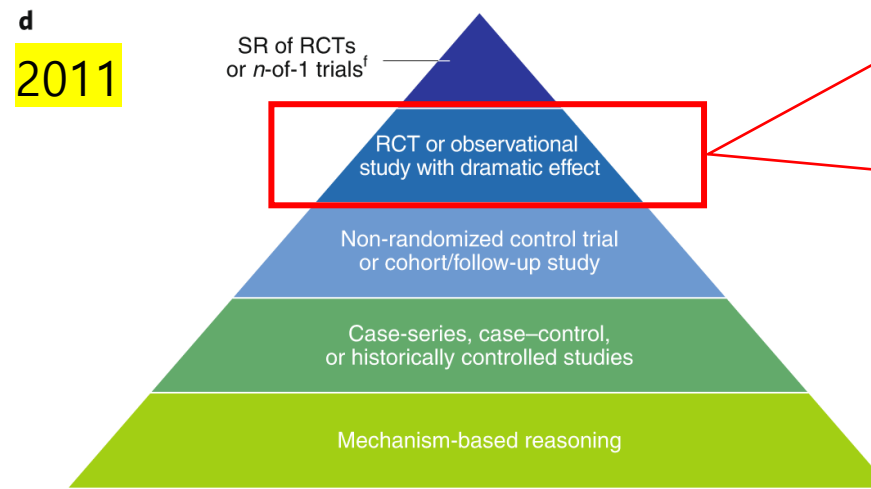
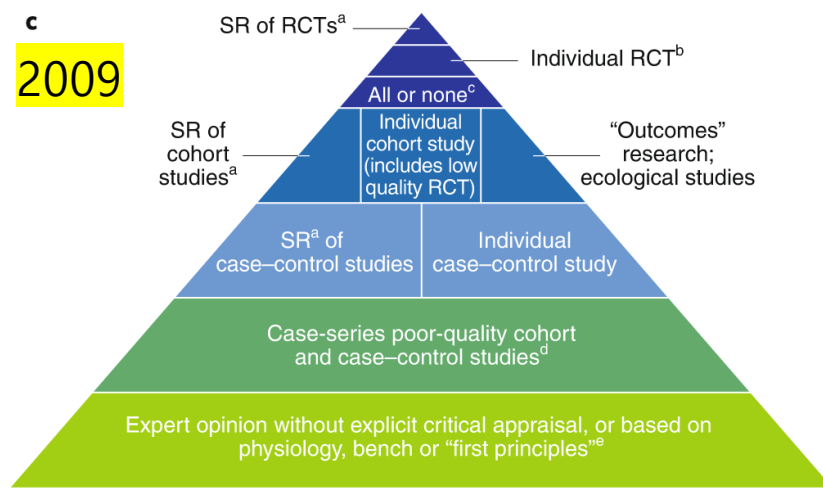
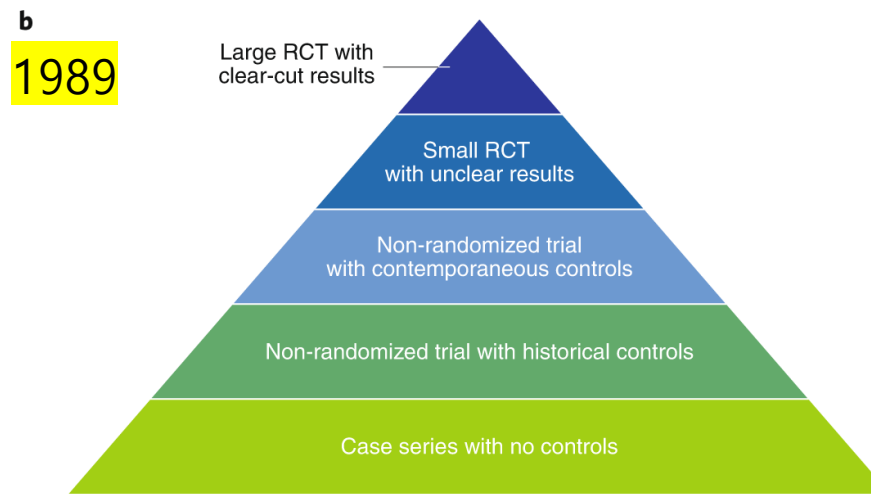
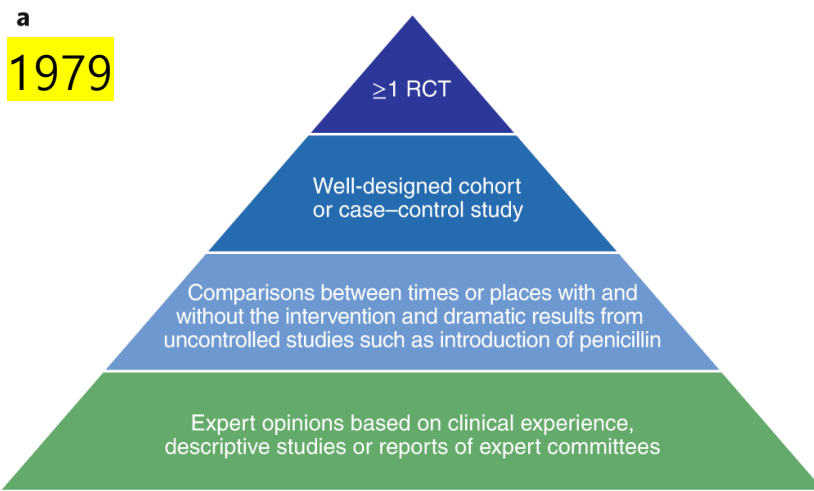
Later phase RWE research can inform future early phase research and decision making.



<https://www.medpace.com/solutions/rwe-late-phase-clinical-research/>

Balancing clinical evidence in the context of a pandemic

Nature Biotechnology volume 39, pages270–274 (2021)



"Indeed, in these modernized guidelines, some types of observational studies with striking effects now occupy the first or second tier of the levels-of-evidence pyramid."

a, Canadian Task Force on the Periodic Health Examination's Levels of Evidence (1979)

b, Levels of evidence from Sackett (1989)

c, The Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence Working Group (March 2009)

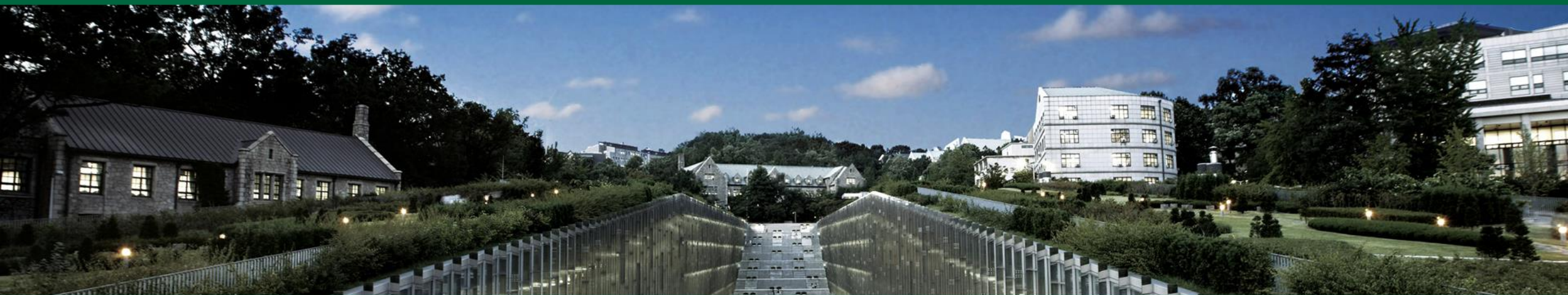
d, OCEBM Levels of Evidence Working Group (2011)

NEJM에 출판된 COVID-19 치료제 임상시험 결과들

Table 1 | Clinical trials on COVID-19 treatments published in the *New England Journal of Medicine* from 12 March through 17 July 2020

Trial	Patients (n)	Design	Results	Reference
Lopinavir-ritonavir in adults hospitalized with severe COVID-19	199	Open-label RCT: Patients were randomly assigned in a 1:1 ratio to receive either lopinavir and ritonavir (400 mg and 100 mg, respectively; $n = 99$ patients) twice a day for 14 days in addition to standard care, or standard care alone ($n = 100$). The primary end point was the time to clinical improvement.	In hospitalized adult patients with severe COVID-19, no benefit was observed with lopinavir-ritonavir treatment beyond standard care.	7
Compassionate use of remdesivir for patients with severe COVID-19	61	Compassionate use: Remdesivir was provided to patients hospitalized with COVID-19 who had an oxygen saturation of $\leq 94\%$ while breathing room air or who were receiving oxygen support. Patients received a 10-day course of remdesivir (200 mg administered intravenously on day 1, followed by 100 mg daily for the remaining 9 days of treatment).	Clinical improvement was observed in 36 of 53 evaluable patients (68%).	26
HCQ	1,376	Observational study: HCQ (600 mg twice on day 1, then 400 mg daily for a median of 5 days).	In the main analysis, there was no significant association between HCQ use and intubation or death (hazard ratio, 1.04, 95% confidence interval (CI), 0.82 to 1.32).	27
HCQ	821	Double-blind RCT: Participants were randomly assigned in a 1:1 ratio to receive either HCQ (800 mg once, followed by 600 mg in 6 to 8 h, then 600 mg daily for 4 more days) ($N = 414$) or placebo ($n = 407$). The primary endpoint was incidence of either laboratory-confirmed COVID-19 or illness compatible with COVID-19 within 14 days.	The incidence of new illness compatible with COVID-19 was not significantly different between experimental arm (49 of 414, 11.8%) and control arm (58 of 407, 14.3%); the absolute difference was -2.4 percentage points (95% CI, -7.0 to 2.2; $P = 0.35$).	8
Dexamethasone	6,425	Open-label RCT: Patients were randomly assigned to receive oral or intravenous dexamethasone (6 mg once daily; $n = 2,104$) for up to 10 days or to receive usual care alone ($n = 4,321$). The primary endpoint was 28-day mortality.	482 patients (22.9%) in the dexamethasone group and 1,110 patients (25.7%) in the usual care group died within 28 days after randomization (age-adjusted rate ratio (RR), 0.83; 95% CI, 0.75 to 0.93; $P < 0.001$). Among patients receiving invasive mechanical ventilation, the patients who received dexamethasone had a lower incidence of death compared to the usual care group (29.3% versus 41.4%; RR, 0.64; 95% CI, 0.51 to 0.81). The dexamethasone group compared with the usual care group had a lower incidence of death in those receiving oxygen without invasive mechanical ventilation (23.3% versus 6.2%; RR, 0.82; 95% CI, 0.72 to 0.94), but not among those who were receiving no respiratory support at randomization (17.8% vs. 14.0%; RR, 1.19; 95% CI, 0.91 to 1.55).	10

Nature Biotechnology volume 39, pages270–274 (2021)



의료 빅데이터의 현황

최근 논문들...

건강보험공단 데이터 (NHIS, Claims) 등 빅데이터를 활용한 연구 (NHIS), 국내연구진

Title	Methods	비고	Date	Journal	IF
Short-term exposure to PM10 and cardiovascular hospitalization in persons with and without disabilities: Invisible population in air pollution epidemiology	<ul style="list-style-type: none"> We conducted a time-stratified case-crossover analysis using conditional logistic regression to investigate the association between short-term exposure to PM₁₀ and cardiovascular hospital admissions. A case-crossover design is a variant of the matched case-control study, in which each case serves as his/her own control (Maclure, 1991). 	<ul style="list-style-type: none"> 환경역학 Case-cross over design (like a matched case-control) Conditional logistic regression PM10 → cardiovascular hospitalization (Disability condition) 	November 2022	Sci. Total Environ.	10.75
Long-term opioid use and mortality in patients with chronic non-cancer pain: Ten-year follow-up study in South Korea from 2010 through 2019	<ul style="list-style-type: none"> Owing to the large sample size, data of 2.5% of adult patients (≥20 years of age) were newly extracted using a stratified random sampling technique. Age and sex were used as an exclusive stratum for sampling. Next, we carried out survival analyses using multivariable Cox regression modelling for all-cause mortality factors spanning a 10-year period. 	<ul style="list-style-type: none"> Random sampling technique Cox regression modelling Opioid use → mortality 	September 2022	EClinical Medicine	10.04

Title	Methods	비고	Date	Journal	IF
Cardiovascular Implications of the 2021 KDIGO Blood Pressure Guideline for Adults With Chronic Kidney Disease	From the cross-sectional Korea National Health and Nutrition Examination Survey (KNHANES) and longitudinal National Health Insurance Service (NHIS) data, adults with nondialysis CKD were identified and categorized into 4 groups based on concordance/discordance between guidelines: 1) above both targets; 2) above 2021 KDIGO only; 3) above 2012 KDIGO or 2017 ACC/AHA only; and 4) controlled within both targets. We determined the nationally representative proportion and CVD risk of each group.	<ul style="list-style-type: none"> • KNHANES (국민건강영양조사) – cross sectional • NHIS (건강보험공단) - longitudinal • CKD 환자군을 가이드라인 따라 4개군으로 분류 • CKD → CVD • (blood pressure) 	May 2022	J. Am. Coll. Cardiol.	24.09
Outcomes of living liver donors are worse than those of matched healthy controls	This cohort study included 12,372 LLDs who donated a liver graft between 2002 and 2018, and were registered in the Korean Network for Organ Sharing . They were compared to 3 matched control groups selected from the Korean NHIS and comprising a total of 123,710 individuals: healthy population (Group I); general population without comorbidities (Group II); and general population with comorbidities (Group III).	<ul style="list-style-type: none"> • Cohort study • 한국장기기증네트워크 • 3 matched control groups • Liver donor → outcome 	Nov 2021	J. Hepatol.	30.08
Long-term Survival of 10,116 Korean Live Liver Donors	Data of 10,116 live liver donors were drawn from a mandated national registry of Korean live liver donors between 2000 and 2015. Matched controls were selected from the Korean National Health Insurance System-National Sample Cohort (NHIS-NSC). Median (range) follow-up of liver donors was 5.7 (0–15.9) years. Donors were 1:3 individually matched to controls by sex and 5-year age group ; potential controls were from the whole NHIS-NSC (Control 1) or from NHIS-NSC after excluding people with contraindications to be organ donors (Control 2) (donor, n = 7538; Control 1, n = 28,248; Control 2, n = 28,248).	<ul style="list-style-type: none"> • 장기기증등록데이터 • Matched control 을 건강보험공단 표본코호트에서 생성 • Liver donor → survival 	Aug 2021	Ann. Surg.	13.79

Title	Methods	비고	Date	Journal	IF
Incidence of cancer after asthma development: two independent population-based cohort studies	Two independent, population-based, longitudinal cohorts were examined, and estimated hazard ratios were determined using Cox regression . One group consisted of an unmatched cohort of 475,197 participants and a propensity score-matched cohort of 75,307 participants from the National Health Insurance Service-National Sample Cohort (NHIS-NSC; claims-based data from 2003 to 2015). The other group consisted of 5,440 participants from the Ansan-Ansung cohort (interview-based data from 2001 to 2014).	# 2개의 longitudinal cohorts <ul style="list-style-type: none"> • 건강보험공단 표본코호트 (matched-control) • 안산-안성 코호트 (인터뷰 기반) <ul style="list-style-type: none"> • Asthma → cancer 	May 13, 2020	J. Allergy Clin. Immunol.	14.29
Effect of hypertension duration and blood pressure level on ischaemic stroke risk in atrial fibrillation: nationwide data covering the entire Korean population	A total of 246 459 oral anticoagulant-naïve non-valvular AF patients were enrolled from Korea National Health Insurance Service (NHIS) database (2005–2015). The risk of ischaemic stroke according to the duration of hypertension and systolic BP (SBP) levels were assessed.	<ul style="list-style-type: none"> • 건강보험공단의 AF patients • cohort study • Ischaemic stroke risk <ul style="list-style-type: none"> • AF → ischemic stroke • (hypertension duration, blood pressure) 	January 2019	Eur. Heart J.	35.86
Metabolic syndrome and risk of Parkinson disease: A nationwide cohort study	Health checkup data of 17,163,560 individuals aged ≥40 years provided by the National Health Insurance Service (NHIS) of South Korea between January 1, 2009, and December 31, 2012, were included, and participants were followed up until December 31, 2015. The mean follow-up duration was 5.3 years. The hazard ratio (HR) and 95% confidence interval (CI) of PD were estimated using a Cox proportional hazards model adjusted for potential confounders.	<ul style="list-style-type: none"> • 건강보험공단의 건강검진 데이터 • Cox proportional hazard ratio <ul style="list-style-type: none"> • Metabolic syndrome → Parkinson 	August 21, 2018	PLoS Med	11.61
Cumulative Dose Threshold for the Chemopreventive Effect of Aspirin Against Gastric Cancer	<ul style="list-style-type: none"> • A total of 461,489 individuals in a population-based longitudinal cohort provided by the National Health Insurance Services (NHIS) in the Republic of Korea were observed from 2007 to 2012 to identify gastric cancer incident cases. The pharmacy claims data of these individuals from 2002 to 2006 were reviewed to assess cumulative medication exposure using the defined daily dose (DDD) system. Hazard ratios (HRs) of aspirin use for gastric cancer were estimated using multivariate Cox Proportional Hazard regression. Sensitivity analyses, including propensity-score matching and a nested case-control design, were performed to evaluate the variability caused by study design. 	<ul style="list-style-type: none"> • Cohort study • Defined daily dose (DDD) system • Multivariate Cox proportional hazard regression # Sensitivity analysis (study design) <ul style="list-style-type: none"> • Propensity score matching • Nested case-control design <ul style="list-style-type: none"> • Aspirin → Gastric cancer (preventive) 	June 2018	Am. J. Gastroenterol	10.38

Title	Methods	비고	Date	Journal	IF
Gamma-glutamyl transferase predicts future stroke: A Korean nationwide study	<ul style="list-style-type: none"> In Korea, the National Health Insurance Service (NHIS) provides full-coverage health insurance service for all citizens. Using data from the NHIS, the NHIS–National Sample Cohort was designed by randomly selecting 2% of Koreans, carefully considering demographic characteristics. We analyzed eligible individuals from this standardized cohort. The Cox proportional hazards model was used for the study investigating the relationship between GGT and stroke. Sex, age, and measurements of height, weight, systolic blood pressure, fasting blood glucose, total cholesterol, hemoglobin, aspartate transaminase (AST), alanine transaminase (ALT), and GGT were routinely obtained for all participants at the time of their first general health examination. 	<ul style="list-style-type: none"> 건강보험공단 표본코HORT 건강검진자료 Cox proportional hazard ratio Gamma-glutamyl transferase → stroke (predictive) 	Feb 2018	Ann. Neurol.	10.42
Weight gain after smoking cessation does not modify its protective effect on myocardial infarction and stroke: evidence from a cohort study of men	A prospective cohort study using the National Health Insurance Service (NHIS) data set collected from 2002 to 2013 was implemented. Based on the first (2002–03) and second (2004–05) NHIS health check-up periods, 108 242 men aged over 40 years without previous diagnoses of MI or stroke were grouped into sustained smokers, quitters with BMI gain, quitters without BMI change, quitters with BMI loss, and non-smokers.	<ul style="list-style-type: none"> Prospective cohort study 건강검진 데이터 Grouping Smoking cessation → MI (protective) (Weight gain) 	Jan 2018	Eur. Heart J.	35.86
Clinical implication of an impaired fasting glucose and prehypertension related to new onset atrial fibrillation in a healthy Asian population without underlying disease: a nationwide cohort study in Korea	We included 366 507 subjects (age ≥20 years) not diagnosed with non-valvular AF from the Korean National Health Insurance Service-National Sample Cohort (NHIS-NSC) from 2003 to 2008. In total, 139 306 subjects diagnosed with AF-related comorbidities were excluded, and a 227 102 healthy population was followed up until 2013 . The body mass index (BMI), blood pressure (BP), and fasting blood glucose (BG) level were acquired during National health check-ups .	<ul style="list-style-type: none"> 건강보험 표본코HORT 건강검진 Cohort study Impaired fasting glucose, prehypertension – AF 	Sep 2017	Eur. Heart J.	35.86

Title	Methods	비고	Date	Journal	IF
Association of prediabetes with death and diabetic complications in older adults: the pros and cons of active screening for prediabetes	<ul style="list-style-type: none"> a total of 36,946 adults aged ≥ 65 years who underwent national health examinations from 2006 to 2008. follow-up was until 31 December 2015. Cox's proportional hazards models estimated hazard ratios (HRs) and 95% confidence intervals (CIs) for death and diabetic complications. 	<ul style="list-style-type: none"> 건강검진자료 Cox proportional HR Prediabetes \rightarrow death, diabetic complications 	June 2022	Age Ageing	10.67
Association of Chronic Hepatitis B Infection and Antiviral Treatment With the Development of the Extrahepatic Malignancies: A Nationwide Cohort Study	<p>We conducted an 18-month landmark analysis using nationwide claims data from the National Health Insurance Service of South Korea. Patients newly diagnosed with CHB in 2012-2014 (n = 90,944) and matched-controls (n = 685,436) were included. Patients with CHB were further classified as the NA-treated (CHB+/NA+, n = 6,539) or the NA-untreated (CHB+/NA-, n = 84,405) group. Inverse probability of treatment weighting analysis was applied to balance the treatment groups. Time-varying Cox analysis was performed to evaluate time-varying effect of NA treatment. The primary outcome was the development of any primary extrahepatic malignancy. Development of intrahepatic malignancy and death were considered as competing events.</p>	<ul style="list-style-type: none"> Matched-control Inverse probability of treatment weighting analysis Time-varying Cox analysis Chronic hepatitis B infection \rightarrow Extrahepatic malignancy (Antiviral treatment) 	May 2022	J. Clin. Oncol.	44.54
Leukotriene-receptor antagonist and risk of neuropsychiatric events in children, adolescents, and young adults: a self-controlled case series	<p>A self-controlled case series study was conducted using the Korean National Health Insurance Service claims database from two three-year observation periods (observation period 1 [Obs1]: 2005 to 2007, observation period 2 [Obs2]: 2016 to 2018). Asthma or AR patients aged 3–30 years who were prescribed LTRAs and diagnosed with NPEs were included. The incidence rate ratios (IRRs) for exposed period and risk periods (1–3, 4–7, 8–14, 15–30, 31–90, >90 days from initiation of LTRA) compared to unexposed periods were calculated using conditional Poisson regression. Subgroup analysis according to age group, type of NPEs and indication of LTRA was performed.</p>	<ul style="list-style-type: none"> Self-controlled case series Incidence rate ratio – conditional Poisson regression Subgroup analysis Leukotriene-receptor antagonist \rightarrow neuropsychiatric events (children, adolescents, young adults) 	May 2022	Eur. Resp. J.	33.80

Title	Methods	비고	Date	Journal	IF
Risk of COVID-19 Infection and of Severe Complications Among People With Epilepsy A Nationwide Cohort Study	We included participants who underwent at least 1 severe acute respiratory syndrome coronavirus 2 real-time reverse-transcription PCR test between January 1 and June 4, 2020, from the Korean nationwide COVID-19 dataset . Epilepsy was defined according to the presence of diagnostic code in health claims data before the COVID-19 diagnosis. To investigate the association between epilepsy and the susceptibility for or severe complications of COVID-19, a 1:6 ratio propensity score matching (PSM) and logistic regression analysis were performed. Severe complications with COVID-19 infection were defined as a composite of the incidence of mechanical ventilation, intensive care unit admission, and death within 2 months after COVID-19 diagnosis.	<ul style="list-style-type: none"> • 한국 COVID 데이터셋 • 건강보험공단자료 연계 • Propensity score matching • COVID-19 → Complications • (with Epilepsy) 	March 2022	Neurology	11.80
Risk of Incident Dementia According to Glycemic Status and Comorbidities of Hyperglycemia: A Nationwide Population-Based Cohort Study	Using a health insurance claims database and the results of biennial health examinations in South Korea, we selected 8,400,950 subjects aged ≥40 years who underwent health examinations in 2009–2010. We followed them until 2016. Subjects' baseline characteristics were categorized by presence of diabetes (yes/no) and glycemic status as normoglycemia, impaired fasting glucose (IFG), new-onset diabetes, or known diabetes (duration <5 years or ≥5 years). We estimated adjusted hazard ratios (aHRs) for dementia occurrence in each category.	<ul style="list-style-type: none"> • 건강보험공단자료 • 건강검진데이터 • Adjusted hazard ratio • Diabetes → Dementia 	October 2021	Diabetes Care	19.11
High Risk of Fractures Within 7 Years of Diagnosis in Asian Patients With Inflammatory Bowel Diseases	Using data from the Korean National Health Insurance claims database gathered between 2007 and 2016, we calculated the incidence rate ratios (IRRs) of vertebral and hip fractures in patients with newly diagnosed IBD (n = 18,228; 64.1% male, 65.9% ulcerative colitis) compared with an age- and sex-matched control population (matching ratio, 1:10; n = 186,871).	<ul style="list-style-type: none"> • Incidence rate ratio (IRR) • Matched control • IBD → Fracture 	June 2021	Clin. Gastroenterol. Hepatol.	11.38
Cardiovascular risk associated with allopurinol vs. benzbromarone in patients with gout	Using the Korean National Health Insurance claims data (2002–17), we conducted a cohort study of 124 434 gout patients who initiated either allopurinol (n = 103 695) or benzbromarone (n = 20 739), matched on propensity score at a 5:1 ratio . The primary outcome was a composite CV endpoint of myocardial infarction, stroke/transient ischaemic attack, or coronary revascularization. To account for competing risk of death, we used cause-specific hazard models to estimate hazard ratios (HRs) and 95% confidence intervals (CIs) for the outcomes comparing allopurinol initiators with benzbromarone.	<ul style="list-style-type: none"> • Matched control (PSM) • Cause-specific hazard model • Allopurinol vs. Benzbromarone → Cardiovascular risk 	Sep 2021	Eur. Heart J.	35.86
Impact of smoking on the development of idiopathic pulmonary fibrosis: results from a nationwide population-based cohort study	Using the Korean National Health Information Database, we enrolled individuals who had participated in the health check-up service between 2009 and 2012. Participants having a prior diagnosis of IPF were excluded. The history of smoking status and quantity was collected by a questionnaire. We identified all cases of incident IPF through 2016 on the basis of ICD-10 codes for IPF and medical claims. Cox proportional hazards models were used to calculate the adjusted HR (aHR) of the development of IPF.	<ul style="list-style-type: none"> • 건강보험공단 건강검진 • Cox proportional hazard ratio • Smoking → idiopathic pulmonary fibrosis 	Sep 2021	Thorax	10.31

Title	Methods	비고	Date	Journal	IF
Risk of tuberculosis in patients with cancer treated with immune checkpoint inhibitors: a nationwide observational study	<ul style="list-style-type: none"> While some recent studies have reported the development of tuberculosis (TB) in patients exposed to immune checkpoint inhibitors (ICIs), there is limited evidence to date. Therefore, we evaluated the risk of TB in patients with cancer exposed to ICIs using the National Health Insurance claims data in South Korea. Patients with diagnostic codes for non-small cell lung cancer, urothelial carcinoma or melanoma between August 2017 and June 2019 were identified. The incidence rate and standardized incidence ratio (SIR) of TB were calculated for both the ICI exposure and non-exposure groups. The risk of TB according to ICI exposure was assessed using a multivariable Cox regression model. 	<ul style="list-style-type: none"> Incident rate, standandardized incidence ratio (SIR) ICI exposure group vs non-exposure group Multivariate Cox regression Immune check point inhibitor → TB 	Sep 2021	J. Immuno ther. Cancer	12.47
Awareness of the use of hyponatraemia-inducing medications in older adults with hyponatraemia: a study of their prevalent use and association with recurrent symptomatic or severe hyponatraemia	<ul style="list-style-type: none"> To evaluate the use of hyponatraemia-inducing medications (HIMs) after treatment for symptomatic or severe hyponatraemia and to investigate the impact of HIMs on the recurrence of symptomatic or severe hyponatraemia in older patients. A cross-sectional and nested case-control study using data obtained from national insurance claims databases. The rate of prescribing HIMs during the 3 months before and after the established index date was analysed in a cross-sectional analysis. Multivariable logistic regression was performed to investigate the association between HIM use and recurrence of symptomatic or severe hyponatraemia after adjusting for covariates in a case-control study. 	<ul style="list-style-type: none"> Cross-sectional : prescription rate before and after hyponatraemia diagnosis Nested case-control study : hyponatraemia recurrence rate, Multivariate logistic regression Hyponatraemia-inducing medication with hyponatraemia → recurrence of hyponatraemia 	July 2021	Age Ageing	10.67
Lower risk of stroke after alcohol abstinence in patients with incident atrial fibrillation: a nationwide population-based cohort study	Using the Korean nationwide claims and health examination database, we included subjects who were newly diagnosed with AF between 2010 and 2016. Patients were categorized into three groups according to the status of alcohol consumption before and after AF diagnosis: non-drinkers; abstainers from alcohol after AF diagnosis; and current drinkers. The primary outcome was incident ischaemic stroke during follow-up. Non-drinkers, abstainers, and current drinkers were compared using incidence rate differences after the inverse probability of treatment weighting (IPTW).	<ul style="list-style-type: none"> 건강보험공단 데이터 건강검진 데이터 3 groups according to alcohol consumption Incidence rate differences IPTW AF → ischaemic stroke (alcohol consumption, abstainer) 	June 2021	Eur. Heart J.	35.86

Title	Methods	비고	Date	Journal	IF
Risk of Hematologic Malignant Neoplasms From Abdominopelvic Computed Tomographic Radiation in Patients Who Underwent Appendectomy	This nationwide population-based cohort study used the National Health Insurance Service claims database in South Korea to assess 825 820 patients who underwent appendectomy for appendicitis from January 1, 2005, to December 31, 2015, and had no underlying risk factors for cancer. Patients were divided into CT-exposed (n = 306 727) or CT-unexposed (n = 519 093) groups. The study was terminated on December 31, 2017, and data were analyzed from October 30, 2018, to September 27, 2020.	<ul style="list-style-type: none"> • CT-exposed vs. CT-unexposed • Abd CT (appendectomy) → Hematologic malignant neoplasm 	January 20, 2021	JAMA Surg.	16.68
Effect of Asthma and Asthma Medication on the Prognosis of Patients with COVID-19	The study included 7590 de-identified patients, who were confirmed to have COVID-19 using the severe acute respiratory syndrome coronavirus 2 RNA-PCR tests conducted up to May 15, 2020; we used the linked-medical claims data provided by the Health Insurance Review and Assessment Service. Asthma and asthma severity (steps suggested by the Global Initiative for Asthma) were defined using the diagnostic code and history of asthma medication usage.	<ul style="list-style-type: none"> • COVID-19 data • 연계 건강보험데이터 • Asthma, asthma medication → Prognosis of COVID-19 	September 25, 2020	Eur. Resp. J.	16.67
Altered Risk for Cardiovascular Events With Changes in the Metabolic Syndrome Status: A Nationwide Population-Based Study of Approximately 10 Million Persons	A total of 27 161 051 persons who received national health screenings from 2009 to 2014 were screened. Those with a history of major adverse cardiovascular events (MACE) were excluded. We determined the MetS status of 9 553 042 persons using the following harmonizing criteria: MetS-chronic (n = 1 486 485), MetS-developed (n = 587 088), MetS-recovery (n = 538 806), and MetS-free (n = 6 940 663).	<ul style="list-style-type: none"> • 건강검진 데이터 • Metabolic syndrome 에 따른 group 분류 • Metabolic syndrome → Cardiovascular event 	November 2019	Ann. Intern. Med.	51.6
Nasal Polyps and Future Risk of Head and Neck Cancer: A Nationwide Population-based Cohort Study	The 2005-2017 National Health Insurance claims and National Health Screening program databases were used to construct a cohort of patients with nasal polyps and matched comparators in Korea. The relative risk of NCPS and nasopharyngeal cancer in patients with nasal polyps was examined.	<ul style="list-style-type: none"> • Matched coparators • Nasal polyps → Head and neck cancer 	July 2019	J. Allergy Clin. Immunol.	14.29

Title	Methods	비고	Date	Journal	IF
Markedly Reduced Risk of Internal Malignancies in Patients With Vitiligo: A Nationwide Population-Based Cohort Study	We conducted a population-based retrospective cohort study using data from the Korean National Health Insurance claims database obtained from January 2007 to December 2016. All patients age 20 years or older with vitiligo who had at least two contacts with a physician from 2009 to 2016, during which a principal diagnosis was made, were identified (vitiligo group). Controls were randomly selected (two per patient with vitiligo) after frequency matching with the vitiligo group for age and sex during the same period (control group).	<ul style="list-style-type: none"> • Cohort study • Vitiligo → internal malignancies • Matched control 	February 2019	J. Clin. Oncol.	44.54
Antithyroid Drugs and Congenital Malformations: A Nationwide Korean Cohort Study	A cohort of 2 886 970 completed pregnancies linked to live-born infants in 2 210 253 women between 2008 and 2014 in NHIS (Nationwide cohort study). The risk for overall and organ-specific congenital malformations in offspring, with logistic regression models used to control for potential confounders.	<ul style="list-style-type: none"> • 산모/신생아 코호트 • Logistic regression • Antithyroid drug → Congenital malformation 	January 2018	Ann. Intern. Med.	51.6
High liver fibrosis index FIB-4 highly predictive of hepatocellular carcinoma in chronic hepatitis B carriers	Our retrospective cohort study involved 986 Korean HBsAg carriers 40 years of age or older who visited Seoul National University Hospital for a health checkup. National medical service claims data were used to determine HCC incidence. Median follow-up time was 5.4 years (interquartile range: 4.4 years).	<ul style="list-style-type: none"> • 병원 건강검진 자료와 NHIS 의 HCC 진단 자료를 연계함. • FIB-4 index → HCC 	DEC 2014	Hepatology	17.42
Job Loss and Re-Employment of Cancer Patients in Korean Employees: A Nationwide Retrospective Cohort Study	All employees except for the self-employed in Korea who were diagnosed with cancer during the 2001 calendar year (n=5,396) were identified as the first baseline patients and were followed every 3 months over 6 years to estimate the time taken to job loss. Patients who lost their job within the first year after a diagnosis of cancer (n=1,398) were identified as the second baseline patients and were followed up over 5 years to estimate the time taken to re-employment using the National Health Insurance claims data . Patient demographic, socioeconomic, and clinical variables were investigated as factors that affected job loss and re-employment.	<ul style="list-style-type: none"> • Cancer 진단 후 job loss • Job loss 후 re-employment • 사회과학적 분야 연구 	March 2008	J. Clin. Oncol.	44.54

Title	Methods	비고	Date	Journal	IF
Sarcopenia is associated with significant liver fibrosis independently of obesity and insulin resistance in nonalcoholic fatty liver disease: Nationwide surveys (KNHANES 2008-2011)	This study investigated whether sarcopenia is associated with significant liver fibrosis in subjects with NAFLD. Data from the Korean National Health and Nutrition Examination Surveys 2008-2011 database were analyzed.	<ul style="list-style-type: none"> • Sarcopenia → liver fibrosis • (Nonalcoholic fatty liver disease) 	December 2015	Hepatology	17.42
Low vitamin D levels are associated with atopic dermatitis, but not allergic rhinitis, asthma, or IgE sensitization, in the adult Korean population	A cross-sectional study was performed by using data collected from 15,212 individuals 19 years or older who participated in the Korean National Health and Nutrition Examination Survey from 2008 to 2010.	<ul style="list-style-type: none"> • Cross sectional study • Vitamin D → atopic dermatitis 	January 2014	J. Allergy Clin. Immunol.	14.29



의료 빅데이터의 활용전략

보건의료빅데이터개방시스템

의료이용지도
로그인 | 회원가입

공공데이터
의료빅데이터
의료통계정보
고객지원
시스템소개

공개는 늘리! 제공은 **빨리!** 이용은 **편리!**

건강보험심사평가원에서 보유하고 있는 다양한 의료데이터를 국민에게 개방합니다.

데이터 서비스 현황

- 공공데이터: 112 중 다빈도이용 Top10
- Open API: 24 중
- 의료통계정보: 126 SHEET, 40 CHART, 2 MAP
- 원격분석시스템: 총 470 계정

국민관심질병

국민관심 진료행위

다빈도질병

질병(소분류)

진료행위

데이터결합

빅데이터 시각화

의료빅데이터

과제목록 ▶ 더보기

- 바이러스질환과 소아신경질환 연관성 분석
- 혈액투석 환자에서 의료 형평성과 예후의 영향 분석
- COVID-19 전 후 천식 환자 비교
- 비소세포폐암과 소세포폐암의 연도별 치료 성적 개선...

전문적인 빅데이터 분석

- 이용안내 ▶
- 이용신청 ▶
- MY 분석과제 ▶

원격분석시스템

- 원격분석시스템 ▶
- 빅데이터분석연습 ▶

공공데이터 신청 안내

공공데이터
바로가기

Open API
바로가기

Q&A

FAQ

용어사전

데이터 레아웃

원격지원요청 | 이용약관 | 개인정보처리방침 | 사이트맵 | 원격분석시스템

-심평원 관련사이트- v 이동

보건복지부

DATA 공공데이터포털
GO.KR

KOSIS 국가통계포털
KOSIS.Stat.go.kr

- 나라지표

NHSS National Health Insurance Sharing Service

로그인 회원가입 사이트맵 ENGLISH

서비스이용안내
데이터신청
성과공유
통계
의료이용지표
공공데이터
고객센터

표본코호트DB 안내

☞ 데이터신청 > 표본코호트DB 안내

데이터 제공안내
표본코호트DB 안내
연구DB 신청
환경성질량DB 신청
교육용DB 신청
데이터 다운로드

표본코호트DB

건강검진코호트DB

노인코호트DB

직장여성코호트DB

영유아검진코호트DB

- [기준] 2006년 1년간 건강보험가입자 및 의료급여수급권자 자격을 유지한 전국민
- [대상자] 100만명
- [연도] 2002 ~ 2015년(14년)
- [표본추출] 전국민 모집단의 2%, 성·연령·가입자구분·보험료 분위·지역별 층화추출
- [내용] 사회·경제적 현황(자격 및 보험료, 장애 및 사망), 의료이용 현황(진료 및 건강검진), 영양기관 현황
- [참고] 표본코호트DB 참고자료 다운로드

자격 및 보험료 테이블

	대상	건강보험가입자 및 의료급여수급권자(외국인 제외)
	내용	성, 연령대, 거주지역, 가입자 구분, 소득분위 등 대상자의 사회경제적 변수 및 장애, 검진대상자 여부 등
	변수	10개 변수로 구성

출생 및 사망 테이블

	대상	대상자의 출생정보 및 통계청의 사망원인DB와 연계하여 사망정보가 확인된 대상자
	내용	표본 대상자의 출생년도, 사망연월 및 사망원인
	변수	5개 변수로 구성

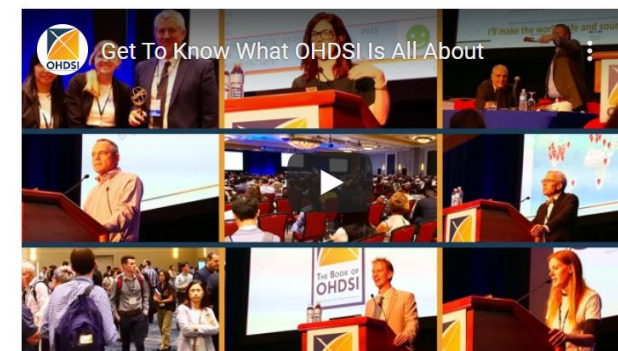
진료 테이블

	내용	대상자가 요양기관에 방문하여 진료 등을 받은 내역에 대해 요양기관으로부터 요양급여가 청구된 자료																									
	구성	의과_보건기관(M), 치과(D), 한방(K), 약국(P) 자료에 대한 명세서 일반내역(T20), 진료내역(T30), 상병내역(T40), 처방전교부상세내역(T60)의 10개 세부DB로 구성																									
	변수	<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th>구분</th> <th>의과_보건기관</th> <th>치과</th> <th>한방</th> <th>약국</th> </tr> </thead> <tbody> <tr> <td>일반내역 (T20)</td> <td>○</td> <td>○</td> <td>○</td> <td>○</td> </tr> <tr> <td>진료내역 (T30)</td> <td>○</td> <td>○</td> <td>○</td> <td>○</td> </tr> <tr> <td>상병내역 (T40)</td> <td>○</td> <td>○</td> <td>○</td> <td>-</td> </tr> <tr> <td>처방전 교부상세내역 (T60)</td> <td>○</td> <td>○</td> <td>-</td> <td>-</td> </tr> </tbody> </table>	구분	의과_보건기관	치과	한방	약국	일반내역 (T20)	○	○	○	○	진료내역 (T30)	○	○	○	○	상병내역 (T40)	○	○	○	-	처방전 교부상세내역 (T60)	○	○	-	-
구분	의과_보건기관	치과	한방	약국																							
일반내역 (T20)	○	○	○	○																							
진료내역 (T30)	○	○	○	○																							
상병내역 (T40)	○	○	○	-																							
처방전 교부상세내역 (T60)	○	○	-	-																							

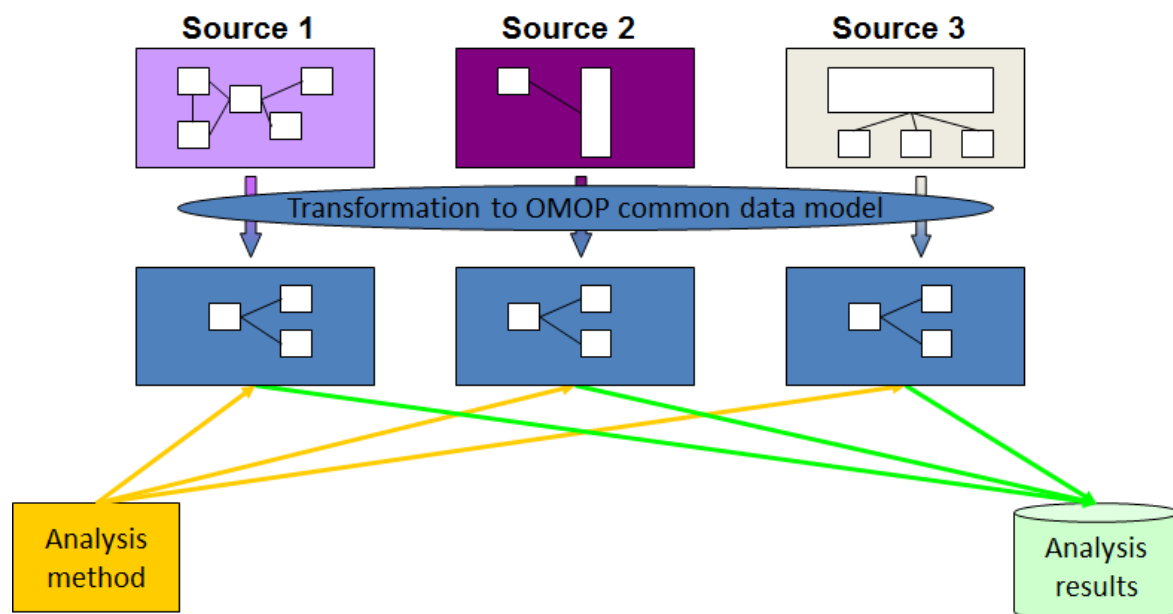
WOMANS UNIVERSITY

E/W/H/A,
 THE FUTURE
 WE CREATE

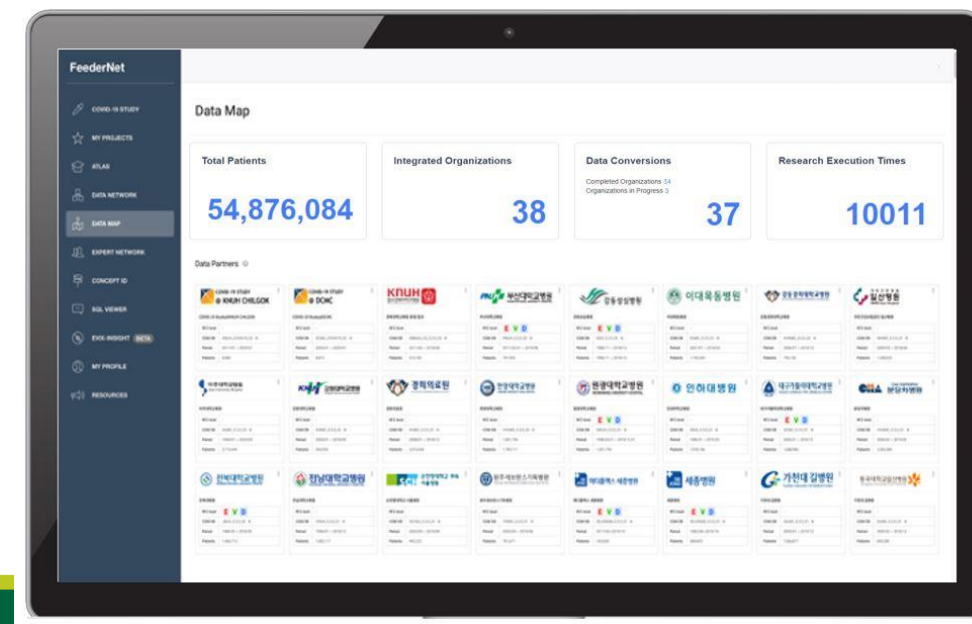
글로벌 CDM network



OMOP-CDM (Common data model)



한국의 CDM network



전략 - Big data가 어디에 있는지 알아야

- ✓ 직접 구축한 병원 cohort
- ✓ 공개 빅데이터들 (국내)
 - 건강보험데이터
 - 심사평가원데이터
 - 국민건강영양조사
- ✓ 공개 빅데이터들 (국제)
 - 미국/영국 등 공개 데이터
 - 우리 과의 international association 에서 구축한 공개 코호트

전략 – 데이터 선택 시 고려할 점

- ✓ 내 연구를 하기에 가장 효과적인 데이터는 무엇인가?
- ✓ 내가 관심있는 value가 포함되어 있는가?
- ✓ Longitudinal 인가, 아니면 cross-sectional 인가?
- ✓ 공신력 있는 데이터인가?
- ✓ 여러 cohort를 해볼 것인가? (2개?)

전략 – 어떤 주제의 연구를 하고 싶은지?

- ✓ 결국 중요한 것은 domain knowledge
- ✓ 진료를 보면서 떠올랐던 궁금증, 연구 주제들
- ✓ 문헌 검색
- ✓ $A \rightarrow \text{new } B$
- ✓ $\text{New } A \rightarrow B$ (other disease, drug, biomarker...)
- ✓ $A \rightarrow B$ (new sub-condition?)

전략 – 그 연구를 하기 위한 연구 디자인?

- ✓ Evidence pyramid
- ✓ Cohort study
- ✓ Matched case-control (propensity score matching, IPTW)
- ✓ Cross sectional

전략 – 효과적인 통계 기법은 무엇인지?

- ✓ 문헌 검색
- ✓ 통계 자문 구하기

전략 – target journal 은 무엇인지?

- ✓ Observational study 에서 STROBE 기준을 고려하면서 연구 setting
- ✓ Target journal 들이 내가 생각하는 연구 주제에 관심이 있는지? (scope 고려)
- ✓ 이미 출판된 최근 논문들의 기법과 수준을 확인

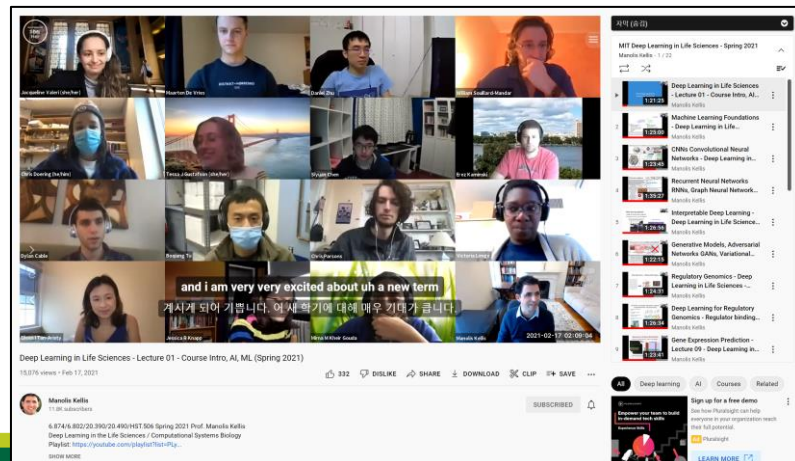
누가 분석을 할 것인지?

- ✓ 직접 한다.
 - R programming
 - SAS
 - Python
- ✓ 공동 연구자(분석자)와 함께 한다.
 - 어떻게 구하는가?
 - 공동 연구자 찾기
 - 병원 분석 지원
 - 연구원 채용
 - 학생과 함께
 - 이 때 본인(PI)의 역할은 무엇인가?
 - 연구 주제 선정
 - 연구 디자인
 - 연구를 위한 빅데이터 마련 (IRB, 데이터 신청, 데이터 조건 설정, 조작적 정의 지정, 데이터 신청 비용을 위한 연구비 수주 등), 결과 분석, 논문 작성, 논문 투고 등...
 - 사실 할 일이 매우 많음...
 - 프로그래밍 돌리는 것 빼고 모두 다 본인이 할 일임

학습 빅데이터가 넘쳐나는 세상

- ✓ 국내 무료/유료 강좌
- ✓ 대학 공개강좌 (MIT, Stanford)
- ✓ Stack Overflow
- ✓ (Almost) 모든 것이 공개되어 있어
- ✓ 사람들 간의 **초격차**가 가능한 세상

- Q. 왜 정보를 공유할까?
- 데이터를 공개함으로써 오히려 영향력 증가
 - Citation number is the power.
 - "Like" is the power. : 구독과 좋아요!





Thank you!

yijunkim@ewha.ac.kr