Big Data Era in Medicine
brought by Genome Omics Information

Hiroshi Tanaka
Tohoku Medical Megabank Organization
Tohoku University
and
Dept. Biomedical Informatics
Tokyo Medical and Dental University

the 71th General Meeting of
the Japanese Society of Gastroenterological Surgery

COI Disclosure

Hiroshi Tanaka

The author have no financial conflicts of interest to disclose concerning the presentation.

Big Data?

Difficult to treat by conventional information processing method because it is too large, too many kinds and too frequently changing

So what is

Medical Big Data?

Arrival of **Big Data Era in Medicine**

Rapid and Huge Accumulation of

(1) Comprehensive Molecular Medical Data brought by the advance of Genome Omics Medicine due to next generation sequencer

(2) Physiological and Behavioral Data brought by Mobile Health (mHealth) Monitoring by Wearable Sensor

(3) Genomic and Exposomic Data by World-wide Spread of Biobank and Genome Cohort

Enormously **Cost Reduced**, nevertheless **High Quality** Massive Data

Genome data : 13 yr, 3,500 M$ (2003) → 1day 1000$ (2016)

Personalized (Precision) Medicine

Tremendous Improvement of **Preciseness** of Medical Care

New type of Big Data emerges

Medical Big Data Revolution

- Clinical "Big Data"
  - Clinical Lab Tests, Prescriptions, Images
  - Ex. NDB (claim DB) . J. Sentinel Project DB
- Socio-Medical "Big Data"
  - Epidemiological survey data
  - life style, health exams, questionnaire

    Due to recent spread of "Digitalization"

Conventional Medical Big data

- Comprehensive Molecular Big Data
  - Genome Omics Medicine
  - Due to Rapid Advance of **Clinical Sequencing**
- Life - log Big Data
  - Epidemiological Survey data
  - Life style, health exams, questionnaire

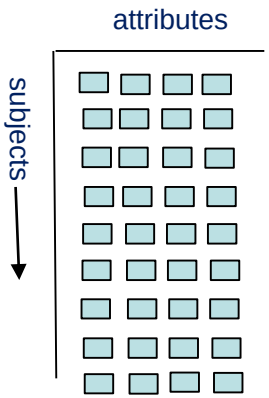
    Due to Rapid Advance of **Wearable Sensor**

New type of (Genuine) Medical Big Data

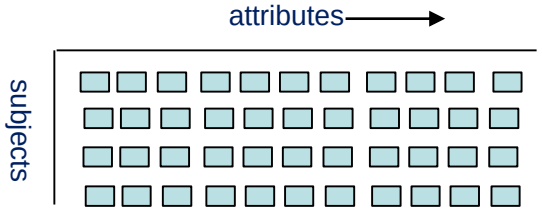
TMDU

New type of Medical Big Data

Data Structure



- Conventional Medical "Big Data"
  - **Big "Small Data"**
    - For one subject (patient)
      Num. of attributes is "Small" (n>>p)
    - But Num. of subjects (patients) is "Big"
    - Conventional statistical method works well



- **Molecular Big Data** (genome, omics)
  - **Small "Big Data"**
    - Num. of attributes for one subject is "Big"
      - Whole genome sequence (x30 cover), 100Gbp for one patient
    - But Num. of subject (patients) is comparatively "Small"
    - Conventional statistical method does not work well

Necessity of
New Data Science of Medicine

TMDU

New type of Medical Big Data

Purpose to Collect Big Data

- Conventional Medical "Big Data"
  - **Population Medicine**
  - To reveal the **"collective law"** ("law of large numbers") by collecting large number of samples
  - Can not be found by seeing each individual subject
- **Molecular Big Data (genome, omics)**
  - **Personalized Medicine**
  - To comprehensively enumerate all the individualized (stratified) patterns exist under the same nameb of disease
  - For exhaustive search, Big number of samples is necessary

Direction to Collect Big Data is Quite Opposite

TMDU

Paradigm Changes

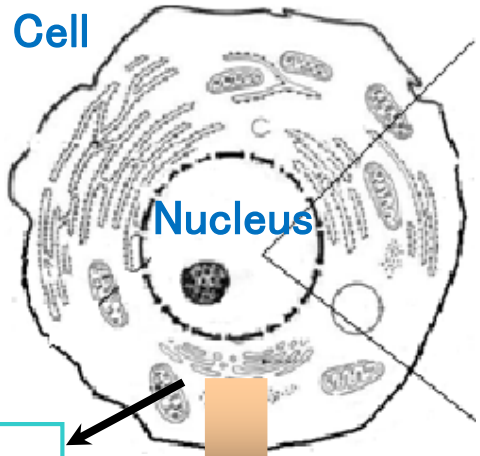
Medical Big Data Revolution Causes

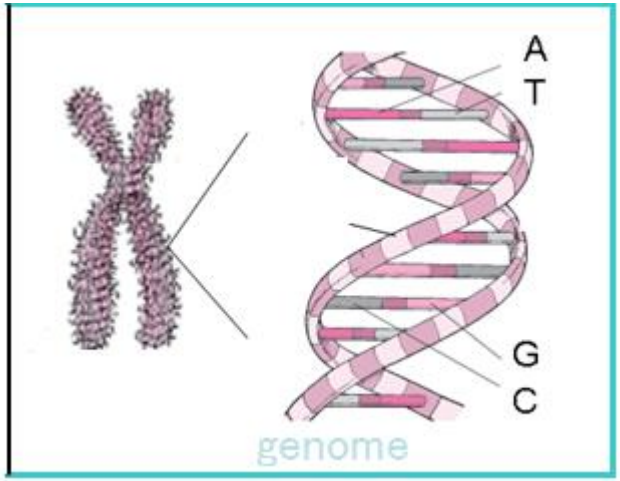
- "**Population medicine**" paradigm changes
  - **"One size fit for all" medicine** no more valid
  - **Towards "Personalized (Precision) Medicine"**
    - Comprehensively Survey is necessary
    - How many "Personalized (Stratification) Patterns" of Disease (intrinsic subtype) exit
    - How fine granularity of stratification should be?
    - Big Data is needed for realization of Personalized Medicine
- "**Evidence-based Medicine**" paradigm changes
  - Liberation from the "gold standard" of RCT and EBM
  - RCT: Controlled (Artificial) Clinical Trials with Small-ish populations outside the Real World
  - **Towards Learning from "Real World Data"**
  - (Disease registry, EHR big data) for clinical evaluation of drug, devise, procedure

Genome and Omics

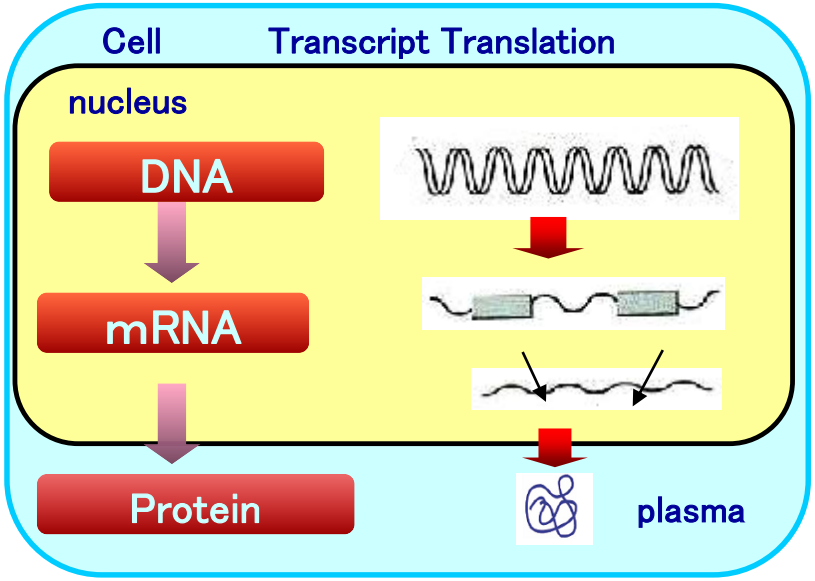
1990 Human Genome Project
2003 HGP finished

Cell





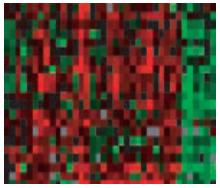
Omics
＝ −Ome (Whole)＋−ics (study)

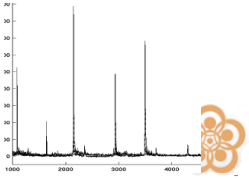


genome
genomics
transcriptome
transcriptomics
proteome
proteomics
metabolome
metabolomics


sequencer

DNA microarray




Mass Spect

TMDU

Impact of Next Generation Sequencer

Enormously rapid advance of High-throughput Molecular Device

Outstanding spead-up and cost reduction of Next Generation Sequencer

2005～ NGS 454 (LS,Roche)
2007/8～454, Solexa (Ilumina),
SOLiD (LT,TF)
Sequence Revolution
Faster than Moor's law

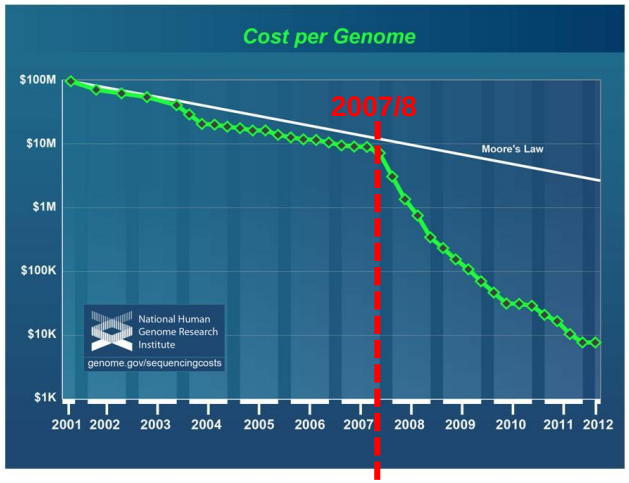
Ilumina 2500

Ion Torrent

Ilumina 2500

WGS(Whole genome sequencing)
3GB (1 person) X 30 = about 100Gbps
1 person WGS 27 hours
WES(Whole exome sequencing)
60Mb (1 person) X 100 = 6Gbps
15 persons WES for 27 hours

Hiseq X system 10 set (cost 1/5）



千ドル
ゲノム

DNA Sequencing Cost: the National Human Genome Research Institute

Sequence Revolution 2007/8

TMDU

Genome omics medicine and Big Data

Practice of Genome medicine

NGS, high-throughput technology

Clinical Implementation of genome sequencing, omics.

Accumulation of Genome, omics data

Integration of Molecular & Clinical Data

Clinical phenotyping (EMERGE project)

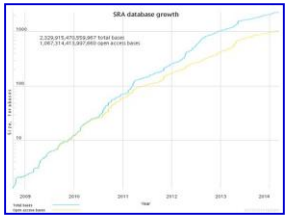
Medical Big Data

Knowledge Discovery

Genome-omics knowledge

Medical Big Data

2Pbps
NCBI:SRA



Mayo Clinic
100K Genome

TMDU

Major Areas of Genome Omics Medicine

Start of Clinical Implementation (2010～)

1． Identification of unknown disease causative gene
at the point of clinical routine practice
Wisconsin Univ. (2010 First Clinical Implementation)
Baylor Medical College (2011)

2． Identification of cancer driver mutation
Dana Faber CC, MD Anderson CC (2012~)

3． Identification of polymorphism of drug metabolizing enzyme (preemptive PGx, EMR implementation)
Vanderbilt Univ., Mayo Clinic (2010~)

Medical College of Wisconsin

First Clinical Sequencing
3 yo boy, unknown intestinal disease, exome seq. identifies the causative mutation, BM transplantation, Complete Remission

Baylor Medical College

Whole genome laboratory In-house, Sequencing

Alert of Mismatch in EMR

Vanderbilt University Hospital (PREDICT)

TMDU

Three Major Streams of Genome Omics Medicine in US

2005～ NGS (454,Solexa,SOLID) 2007/8～
Sequence Revolution

Identification of Causative Gene for Undiagnosed Disease Exome Sq.
MCW Children H

Polymorphism genome Hapmap project 2002～ GWAS 2003～

Polymorphism of Drug Metabolizer, Preemptive PGx
Integrated with EHR
Vanderbilt UH

TCGA (2006), ICCG (2008) ,Results were Published from 2011～

Identification of Cancer Driver Gene, Molecular Targeted Drug
Dana Faber CC

2008
2009
2010
2011
2012
2013

Genome Omics Medicine
Clinical Implementation

TMDU

Clinical Implementation in United States
Genome/Omics Medicine

More than 20 hospitals have implemented Genome/Omics medicine

Institution	Major Projects
MC Wisconsin	Using whole genome sequencing to establish diagnosis in patients with currently undiagnosed genetic disorders
Mount Sinai	• CYP2C19 testing for antiplatelet rx post percutaneous coronary intervention • Personalized decision support for CVD risk management incorporating genetic risk info
Northwestern	Using pharmacogenomics evidence (from GWA genotyping) to guide prescriptions in primary care and assess risk for other conditions such as HFE/hemochromatosis
Cleveland Clinic	Tumor-based screening for Lynch syndrome, endometrial cancer
UCSD	• Screening for actionable mutations in malignant gliomas and glioblastomas for biomarker based RCTs • Targeted rx (such as RET inhibitor) of metastatic solid tumors based on tumor mutation status
Morehouse	• Exome sequencing of 1200 early onset severe African American hypertension cases and 1200 controls
Duke	• Computer-based family hx collection and CDS tool with 1-yr follow-up for perceptions, attitudes, behaviors related to thrombosis and breast, ovarian, and colon cancer • SLCO1B1*5 genotyping and statin adherence • Effect of genetic risk info on anxiety and adherence in T2DM

Institution	Major Projects
Alabama	Planning stages for projects in risk assessment, pharmacogenetic analysis, identification of families for further research
Baylor	Whole exome and whole genome sequencing in Mendelian disorders to improve diagnosis
Geisinger	• Selection for gastric bypass surgery vs other wt loss means based on genetic variants predictive of long-term benefit from surgery • IL28B variants and response to hepatitis C treatment • KRAS and BRAF mutational analysis in thyroid cancer patients
Ohio State	• Personalized genomic med study of CHF and HTN pts randomized to genetic counseling vs usual care • CYP2C19 testing in interventional cardiovascular procedures for clopidogrel
Harvard	Whole genome sequencing with integration in EMR and CDS; pilot of 3 patients to start
U Penn	Genotyping for assessment of MI risk in Preventive Cardiology program
St. Jude's	Pre-emptive PGx genotyping in children
Vanderbilt	Pre-emptive PGx genotyping for clopidogrel, warfarin, or high-dose simvastatin
U Maryland	Develop and apply evidence-based gene/drug guidelines that allow clinicians to translate genetic test results into actionable medication prescribing decisions
Mayo	• PGx driven selection/dosing of antidepressants • CYP2C19 genotyping for antiplatelet rx post PCI
Inter-Mountain	Tumor-based screening for Lynch syndrome

TMDU

Progress of Genome Omics Medicine and Big Data

2005～ NGS (454 Life sci)
2007～**sequence revolution**

2010

1st gen.

Start of Clin. Implementation
first clinical WES（**MCW**）
first preemptive PGx （**VU**）

- **MCW Nic (3yo). Undiagnosed Enteropathy, WES, XIAP mutation**
- **Vanderbilt preemptive PG (PREDICT project) Start**

MC Wisconsin
Clin. Implement.
Big Impact

Baylor **Med College**
Mayo Clinic

Early adopter period

about 2013

2nd gen.

Nation-wide project
NIH "BD2K "initiative
Genome consortiums

Big Data Concept

NIH "Big Data to Knowledge" **(2012/13)**
ACGM incidental finding list 56 genes (2013)
NACHGR report "Future is here" (2013)
CPIC guideline, EGAPP guideline 2013.14

Nation-wide Consortium Period

2015

President Obama
Precision Medicine initiative

Clinical Implementation of Genome Medicine
Several Tens Hospital in US

NIH "BD2K "COE in Data Science, DDI (2014)
ASCO "CancerLinQ", Cancer Common
"Precision Medicine (Obama) " 1 M genomic cohort

Genome/Omics medicine in Japan

- National Cancer Center: Hospital East
  - Research Center for Innovative Oncology (2014 ~)
  - Targeted sequencing to find driver mutation of cancer
  - Allocate a patient to the clinical trial for anticancer molecular target drug, **SCRUM JAPAN**
  - Supported by pharmaceutical companies

- Shizuoka Cancer Center
  - **HOPE project** (High-tech Omics-based Patient Evaluation)
  - Multi-omics based evaluation technology for driver mutation of cancer
  - Supported by research fund

- The University of Kyoto:
  - Identify driver mutation
  - Search for appropriate molecular-targeted drug trial
  - Patient's own expense
  - **OncoPrime**

- **AMED**
  - **iRUD** (initiative on rare and undiagnosed disease)
  - Clinical sequencing of unknown causative mutation

The First Year of Genome Medicine In Japan

TMDU

NIH: *eMERGE* network

electronic MEdical Record + GEnome

phase I (2007-2011)

- **Phenotyping from EMR**
  - Develop, disseminate, and apply approaches to research that **combine biorepositories with electronic medical record (EMR) systems for genomic discovery and genomic medicine implementation** research. In addition, the consortium includes a focus on social and ethical issues such as privacy, confidentiality, and interactions with the broader community
- **EMR-based GWAS**
  - Developing methods and best practices for the utilization of EMR as a tool for genomic research.
  - Each with its own biorepository (DNA etc) linked to phenotypic data contained within EMRs
- **eMERGE-I**: 5 Institutes, **PheKB**
  - Mayo Clinic, Vanderbilt University, Northwestern University, University of Washington, Marshfield Clinic

phase II (2011-2015)

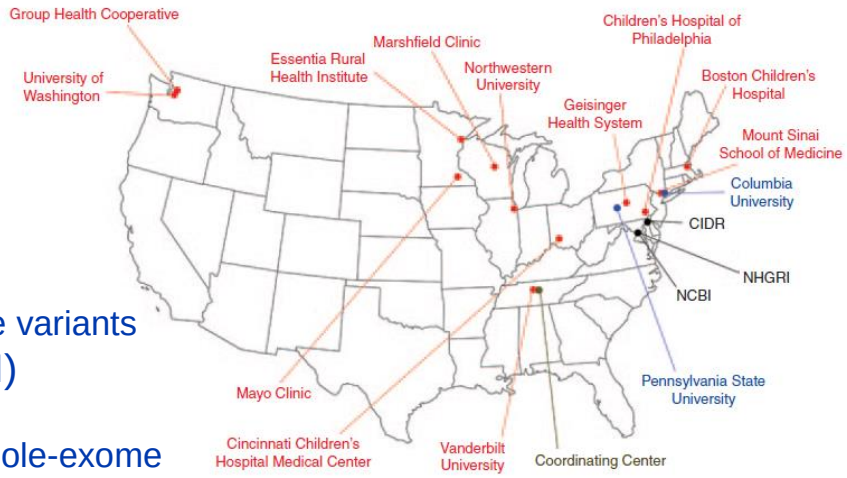
- **Integration of Genomic Information into EMR** (Clinical Implementation)
- PGx implementation in EMR
- Return of (Genomic) Result (RoR)
- 4 new institutes joined in eMERGE-II
  - Children's Hospitals, Mount Sinai/Geisinger

Phase III (2015～2019)

- Specially added: phenotypic implication of rare variants

Collaboration with **CSER consortium** " (NHGRI)

- "Clinical Sequencing Exploratory Research
- explore the potential of whole-genome and whole-exome sequencing to generate new knowledge and improve patient outcomes



TMDU

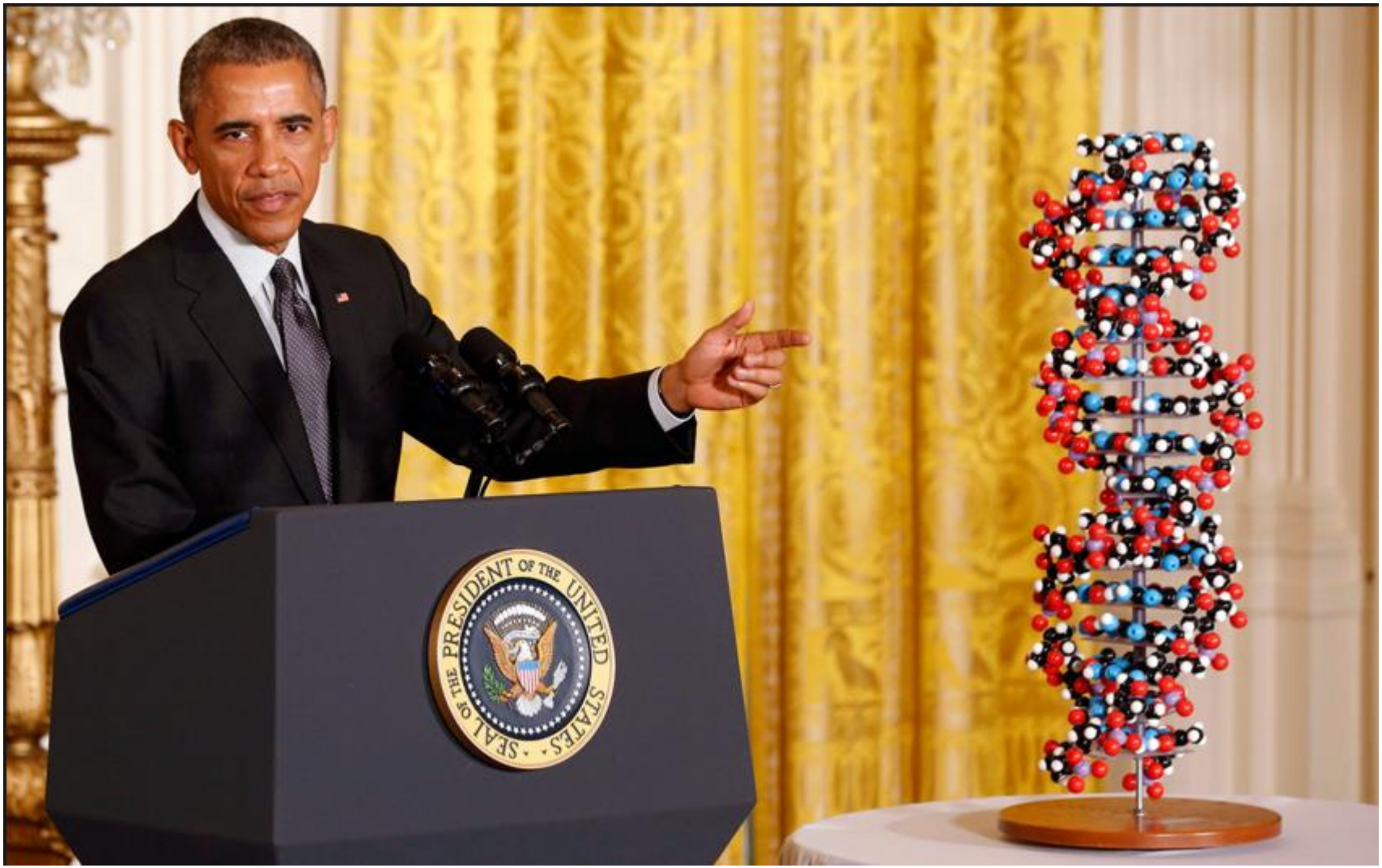
NIH

"***Big Data to Knowledge***" *(BD2K)initiative*

- BD2K: Big Data to Knowledge Initiative 2013 start
  - WG on Data and Informatics for Advisory Committee to the Director of NIH
    - Centers of Excellence (COE),
    - Data discovery index (DDI),
    - Training programs of data scientist,
    - **Associate Director of Data Sciences- Bourne, PhD**.
  - Francis Collins said,
    - The era of '**Big Data**' has arrived,
    - NIH-wide priority initiative to take better advantage of the exponential growth of biomedical research datasets.
    - NIH play a major role in coordinating access to and analysis of many different data types that make up this revolution in biological information."
  - http://bd2k.nih.gov

TMDU

President Obama　Precision Medicine Initiative



2015.1　State of the Union Address

TMDU

New Features of "Precision Medicine"

- New Concept of "Precision Medicine"
  - Essential Same as Concept of "Personalized Medicine"
- Difference from Personalized Medicine
  - More emphasis on "Stratification of disease"
  - Include the effects by environment factors on disease occurrence (GxE interaction)
  - Estimation of the importance of Life-log data mHealth by Wearable Sensor
  - Recognize the importance of Biobank/Genomic Cohort

    As the **information Source/Basis** of **Genomic Medicine**

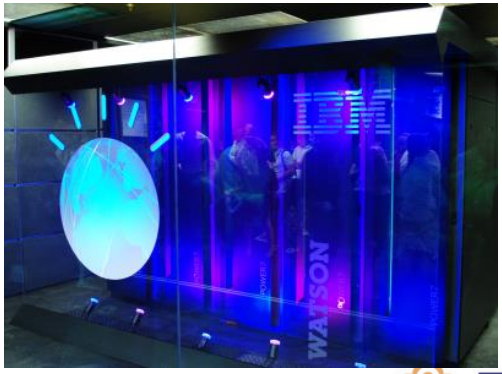
    PMI 1 million cohort project

ASHG2015 Oct

F. Collins

Big Data and Learning system

- **Artificial Intelligence for Learning system**
  - Neural network: Deep Learning to extract characteristic features
  - Data Mining: Sparse Modeling to reduce dimension
- The ASCO (American Society of Clinical Oncology) **CancerLinQ** initiative
  - Building a “learning health system”
  - Collect and analyze cancer care data from millions of patient visits and expert guidelines
  - Pilot prototype (2013～)
    - a 170,000-record prototype by 2015
    - For any given tumor type, DB of 10,000 to 20,000 patients, and with 50 to 100 common tumor types, records of at least one million patients
    - Uses statistical functions and an artificial neural network to learn, structure, and map data fields
- IBM Watson for Cancer centers
  - Memorial Sloan-Kettering Cancer Center
    - The Oncology Expert Adviser software (OEA)
  - New York Genome Center
    - Glioblastoma as a target

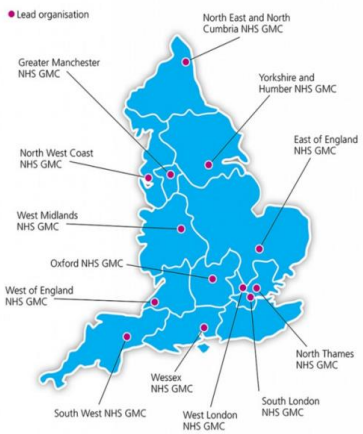


IBM Watson Learning Big Data

TMDU

Biobank/Genome Cohort

- **Biobank**
  - An organized collection of human biological material and associated information stored for research purposes
  - **Past**: tissue sample for regenerative medicine, resource preservation for clinical study
  - **Present**: information basis for realizing clinical genome medicine
  - World-wide trends to promote Biobank project
- **Types of Biobank/Genome Cohort**
  - **Population-type Biobank**:
    - Prospective able-bodied subjects, Estimation of incidence rate of disease by long-termed follow up
    - Genomic and exposomic information
  - **Disease-type Biobank**:
    - Subjects contracting specific disease. Course of disease, genomic and clinico-pathological information
- **Major Biobanks**
  - **UK biobank**;
    - 500,000 persons (2006-2016), population type
  - **Genomics England**;
    - 100,000 persons (2013-2017), WGS, cancer/rare disease
  - **BBMRI** (Biobanking and BioMolecule Resourse Research Infra)
    - More than 300 biobanks in Europe recruited to join BBMRI.
    - Harmonization and Standardization to pool biobank data
  - **Tohoku Medical Megabank**;
    - 150,000 persons (2012～)
    - Community-Based / Residents Cohort 80,000 residents
    - Birth and Three Generation Cohort 70,000 people



NHS Genome Medical Cente
(Genomic England)

These Two Trends would merge and support the genome/omics medicine

within hospital

Clinical genome medicine

Integrated genome-phenome DB
EHR

Nation-wide basis

New knowledge, New information

Large scale Medical Big Data
(both genomic phenomic information)

Disease Genome Cohort

Population Genome Cohort

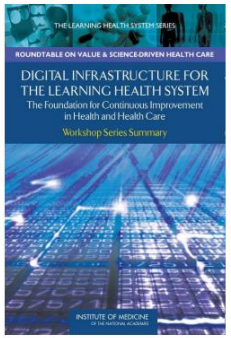
Learning Health System

From Discovery of Biological knowledge to Clinical Implementation: 17 yr
While practicing healthcare, acquire the new knowledge

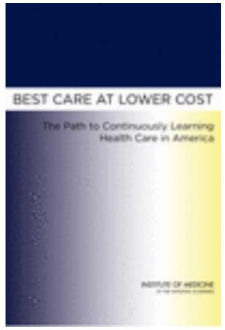
- IOM: “Clinical Data as a Basic Staple of Health Learning”
- “Data obtained from routine medical practice is the Key to support LHS”
  Sharing and learning data improves Health care system
- RCT: Gold standard, but conducting outside the ordinary healthcare systems.
- Is RCT representing the patient group, healthcare is actually directed to
- RCT takes a time and cost
- Effective knowledge accelerate data accumlation

IOM(Institute of Medicine) report 2007, proposed as the paradigm replacing EBM/RCT

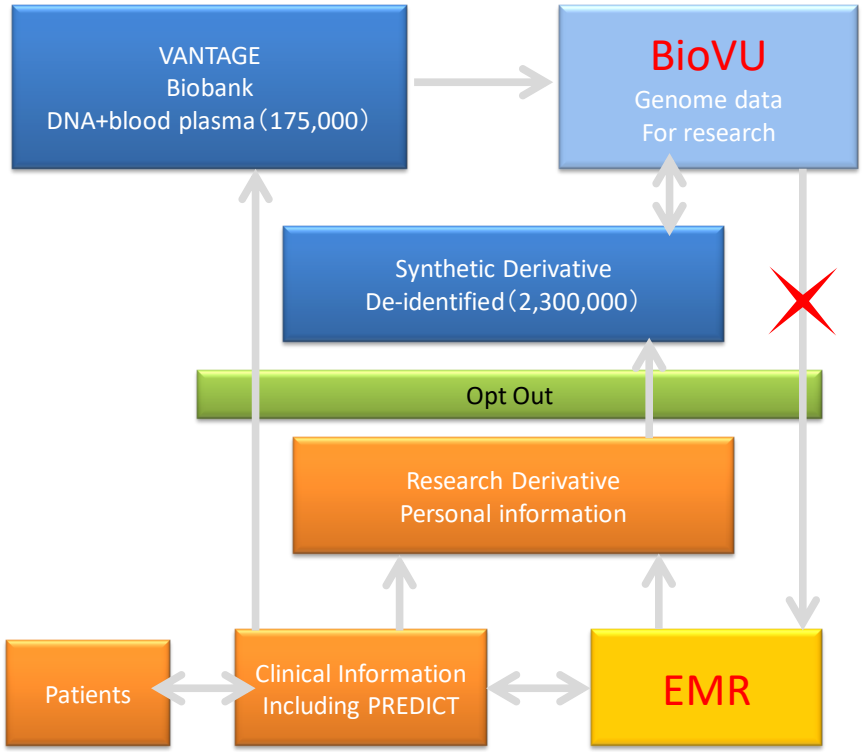
Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care



Best Care at Lower Cost: The Path to Continuously Learning Health Care in America



Typical example of LHS
Integration of Genomic and Clinical information
BioVU Vanderbilt UH



EMR

Synthetic Derivative :
De-identified EMR information
Opt out (2,300,000 records)

Biobank and Genome Analysis

BioVU :
Genome (DNA)
InformationIntegration with
Synthetic Derivative

VANTAGE Core :
175,000 specimen,
DNA extracted from blood,
Genomic analysis

TMDU

EBM changes to **BDM** (Big Data based Medicine)

Paradigm Shift of Clinical Research

- **Disparity** between **RCT Study Population** and **Real World Data**
  - **Impossible in reality** to make study population including **all the stratified (personalized) patterns** of disease
  - Current clinical research study population is in **"artificial environment" outside real world data**
- **Directly use Big Real World Data**
  - No need for **unbiased sampling** from population
  - Because Big Real World Data is very close to population data
  - But still exist the **bias and confounder** (causality) problem

TMDU

Possible Solution

Registry-based Clinical Randomized Trial

- Advantage to use "Real World Data" and the rigorous "Randomization" is fused
  - Thrombus Aspiration in ST-Elevation Myocardial Infarction in Scandinavia (**TASTE**)
    - first trial of RRCT with cost 50 $ per participant
  - Large scaled trial build on already-existing high quality registry
- RRCT process
  - Select the **study population** from the **disease registry** where already exist much of clinical information (7244 STEMI patients)
  - **Randomized allocation of study and control drug** to selected population among registry
  - **End point of trial** is observed by registry.

TMDU

Thank you for kind attention