Khoa học dữ liệu trong chăm sóc sức khoẻ và nghiên cứu y học

Ho Tu Bao

Vietnam Institute for Advanced Study in Mathematics
John von Neumann Institute, Vietnam National Universiry at HCMC
Japan Advanced Institute of Science and Technology
bao@viasm.edu.vn, bao.ho@jvn.edu.vn, bao@jaist.ac.jp

JAIST
JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY
1990

JVN Institute
Vietnam National University HCM

VIASM
Vietnam Institute for
Advanced Study in Mathematics

Japan Advanced Institute of Science and Technology

1. World level research
2. Excellent faculties
3. Motivated students

4. Systematic education
5. Advanced laboratory facilities
6. Innovative administration

- Three graduate schools on **information science** (1992) **materials science** (1993) and **knowledge sciences** (1998).
- Totally 309 faculty and adminitrative staffs (154 professors, 149 staffs), 1138 students (807 master students, 331 PhD students).

John Von Neumann Institute
Excellence in Synergy

Research
Innovation
Education

Project members and collaborators

Pr. K. Takabayashi Pr. Takahiro Suzuki Pr. Tatsuo Kanda Pr. Dam Hieu Chi Pr. Dong T.B.Thuy Dr. Nguyen V. Binh Pr. Nguyen D. Cong

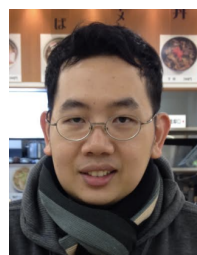Pr. Cao Hoang Tru Dr. Vo T.N. Chau Dr. Nguyen T.M. Huyen Pr. Ly Le Mr. Nguyen N. Hop Mr. Huynh T. Anh Pr. Ho Tu Bao

PhD Studdents
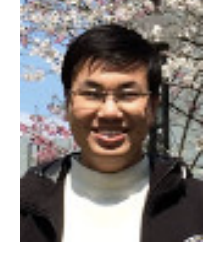Hoang K. Hung G. Moharasan S. Nuttapong P. Ouankhamchan M. Matsuo Phan S. Thanh S. Taewijit Dang T. Thai

Outline

- **Brief of data science**

- The data-driven approach and electronic medical records (EMRs)

- Our project on EMRs data analytics

How does people collect data?

- Observing, measuring, or collecting the **values of features** (features, attributes, properties, variables) of the **objects** under consideration.
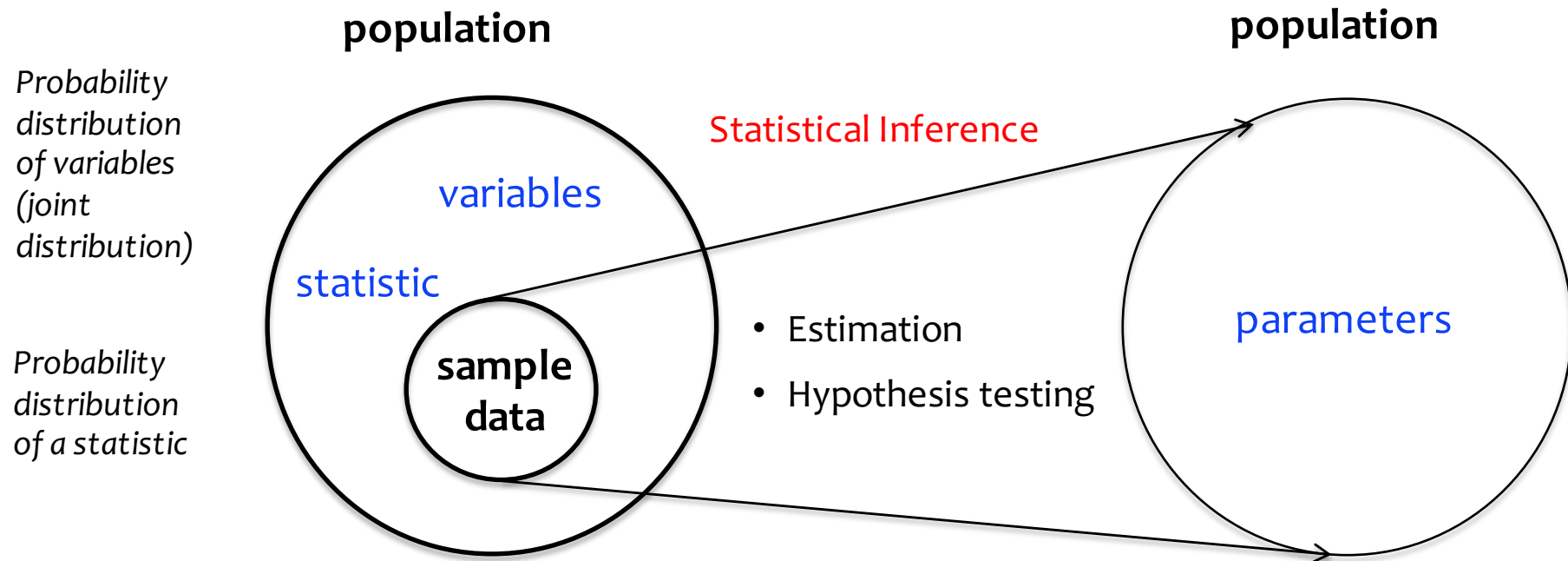- Two ways of collecting data

Randomly sampling	Collecting all available data
Conventional statistics, methods were created when small or medium-sized data sets were common.	Many innovative multivariate techniques being developed to solve large-scale data problems.

Essence of statistics

population   **population**

Probability distribution of variables (joint distribution)

Statistical Inference

variables

statistic

Probability distribution of a statistic

sample data

parameters

- Estimation
- Hypothesis testing

Statistical inference is the ways of drawing conclusions about population parameters from an analysis of the sample data.

- A **parameter** is a *numerical feature* of the population, such as mean, proportion, standard deviation.

- A **statistic** is a single measure of some feature of a sample. It is defined as a *numerical-valued function* of the sample data. It is used to infer the corresponding population parameter.

Multivariate analysis

- Simultaneously analyze the relationship of multiple random variables

- Testing hypothesis by data in **Confirmatory data analysis** (CDA) vs. producing hypotheses from data in **Exploratory data analysis** (EDA)

  - Factor analysis, PCA, Linear discriminant analysis
  - Regression analysis
  - Cluster analysis

- What we can see from conventional methods?

  - Poor results on large and complex data
  - Traditional methods are suitable for analyzing small datasets.
  - Price of storage and data processing are quickly decreasing.

Machine learning

- Mục đích của học máy là xây dựng các hệ máy tính có khả năng học tập như con người.

- *Given*
  - $\{(x_i, y_i)\}, x_i$ is description of an object in some space, $y_i \in \{C_1, C_2, \ldots, C_K\}$ or $y_i \in \mathbb{R}$ is viewed as label of $x_i, i = 1, \ldots, n$.
  - Examples: Set of electronic medical records.

- *Find*
  - Function $p(y|x)$ for labeled data and $p(x)$ for unlabeled data.
  - Diagnosis or treatment regimen for a patient.
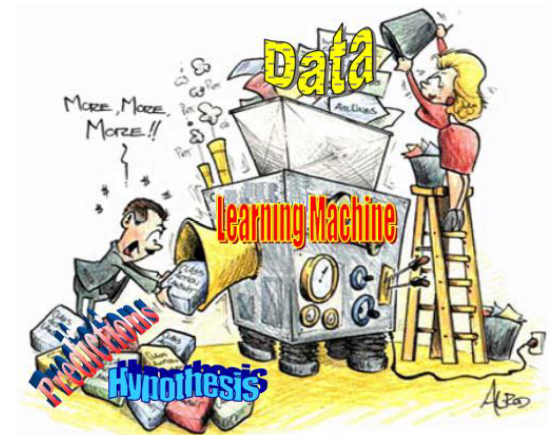
(Source: Eric Xing lecture)

Khai phá dữ liệu – Data Mining

Tự động khám phá, phát hiện các tri thức tiềm ẩn từ các tập dữ liệu lớn và đa dạng.

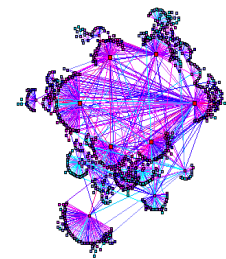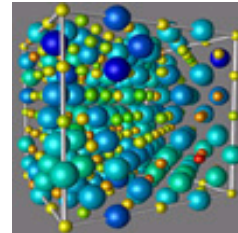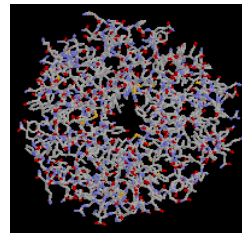**Statistics**

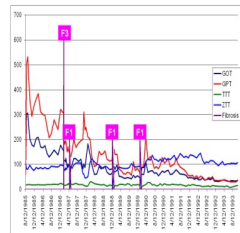Data mining metaphor:
Extracting ore from rock

Large and unstructured real-life data

KDD

Databases   **Machine Learning**

Tìm ra quy luật từ dữ liệu kinh doanh

Dữ liệu kinh doanh

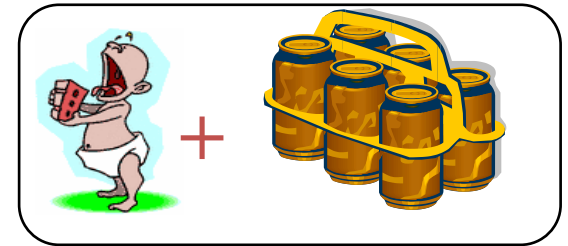"Nhiều đàn ông trẻ tìm mua bia và bỉm trong siêu thị"

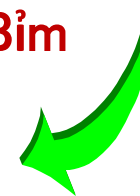
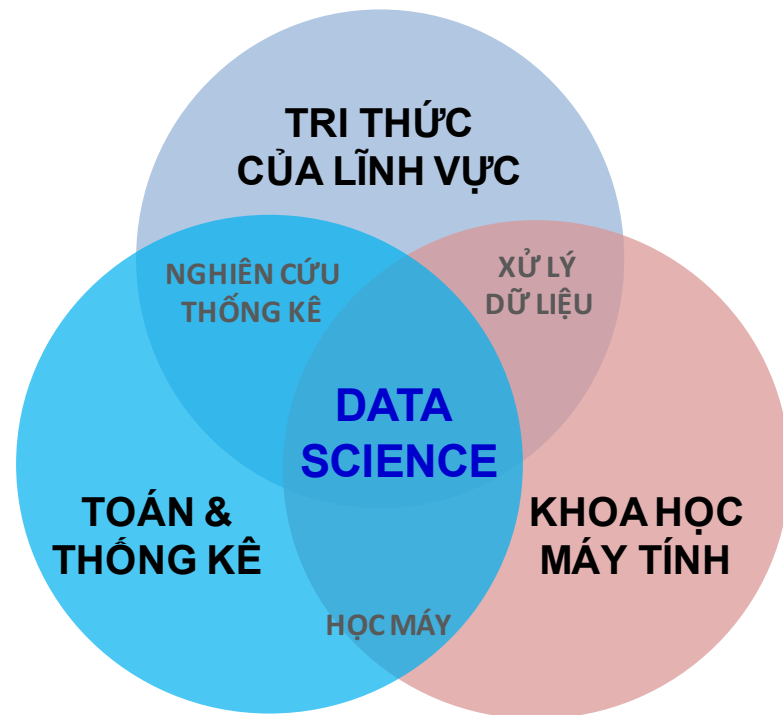Dữ liệu kinh doanh   Phân tích   Đàn ông 20–30 tuổi   **Bỉm**   +   **Bia**

Đang mùa World Cup, những ông bố trẻ mua bia và bỉm để cuối tuần vừa trông con vừa xem bóng đá.

➡️ **sắp xếp bia và bỉm gần nhau**

Khoa học dữ liệu – Data science

Khoa học về phân tích dữ liệu

"In God we trust. All others bring data".
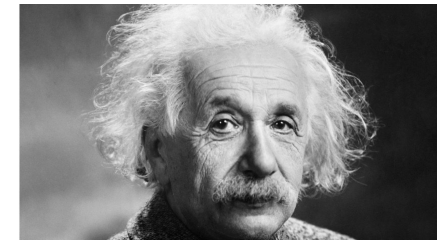"Ta tin Thượng đế. Ngoài ra, là dữ liệu".
W.E. Deming

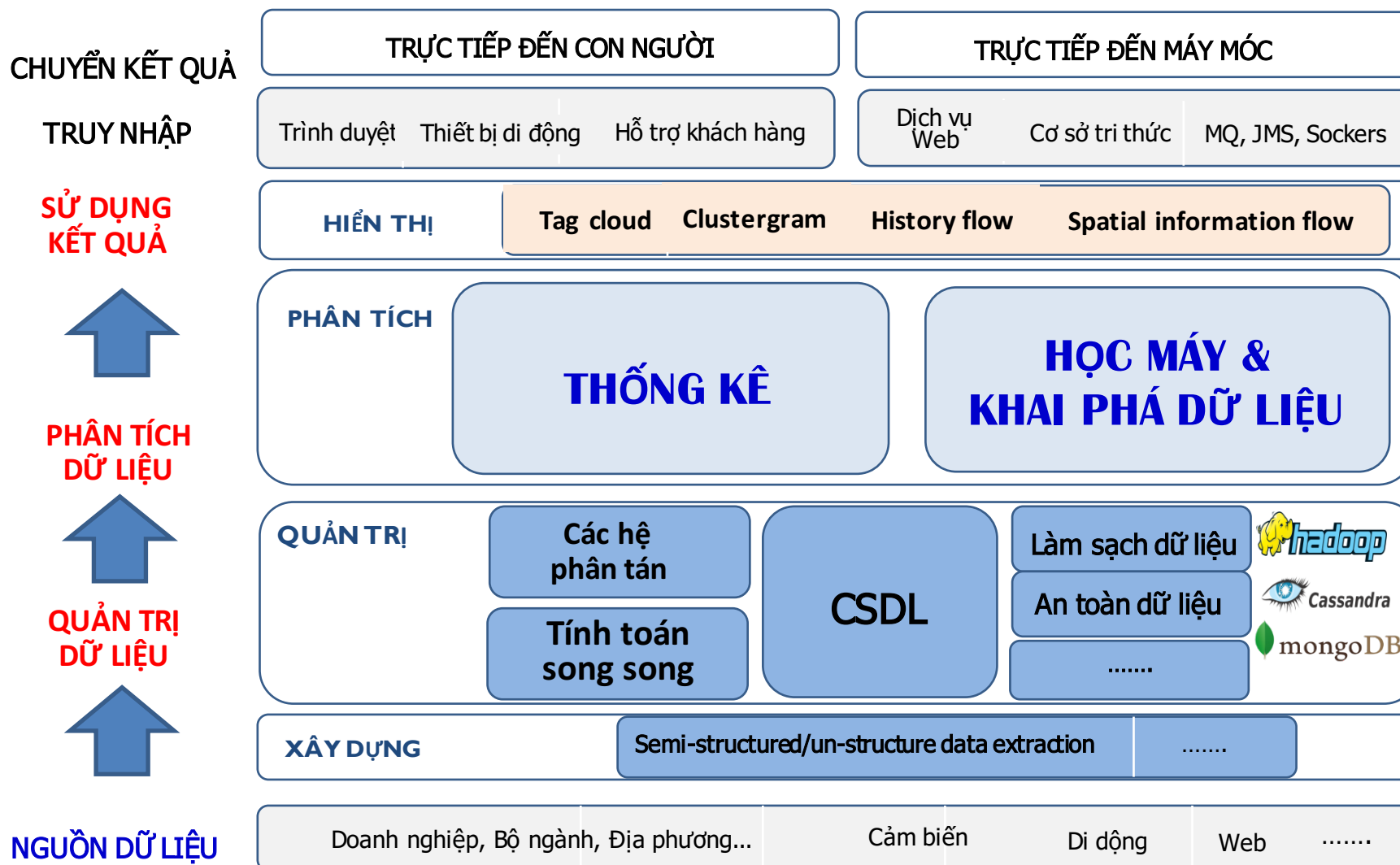"We cannot solve problems by using the same kind of thinking we used when we created them"
Ta không thể giải quyết các vấn đề với chính cách nghĩ ta đã dùng khi đặt vấn đề
Albert Einstein

Kết hợp của Toán học và Tin học là cốt lõi của khoa học dữ liệu

Một lược đồ của khoa học dữ liệu

CHUYỂN KẾT QUẢ

TRỰC TIẾP ĐẾN CON NGƯỜI	TRỰC TIẾP ĐẾN MÁY MÓC

TRUY NHẬP

Trình duyệt	Thiết bị di động	Hỗ trợ khách hàng	Dịch vụ Web	Cơ sở tri thức	MQ, JMS, Sockers

SỬ DỤNG KẾT QUẢ

HIỂN THỊ — Tag cloud · Clustergram · History flow · Spatial information flow

PHÂN TÍCH DỮ LIỆU

PHÂN TÍCH

THỐNG KÊ — **HỌC MÁY & KHAI PHÁ DỮ LIỆU**

QUẢN TRỊ DỮ LIỆU

QUẢN TRỊ
- Các hệ phân tán
- Tính toán song song
- CSDL
- Làm sạch dữ liệu
- An toàn dữ liệu
- ....

hadoop · Cassandra · mongoDB

XÂY DỰNG — Semi-structured/un-structure data extraction · ....

NGUỒN DỮ LIỆU

Doanh nghiệp, Bộ ngành, Địa phương...	Cảm biến	Di động	Web	....

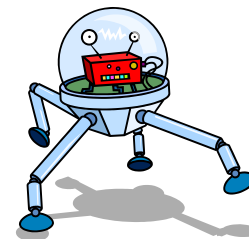Source: WAMDM, Web group

14

Statistics and Machine learning

Statistics

- Formal statistical inference

- Models for small size problems and mostly with numerical data.

- Changing culture and adapt to computational power.
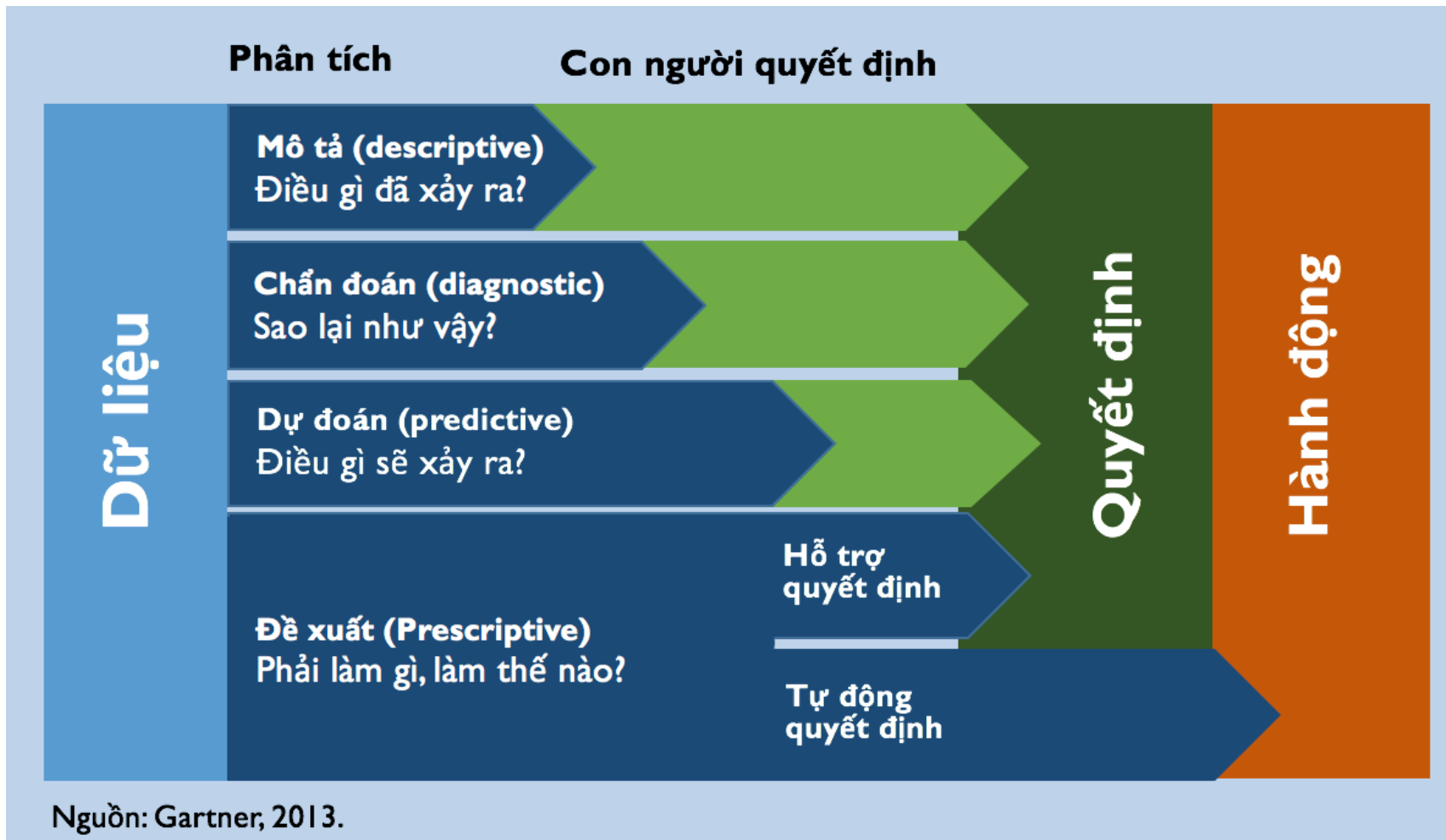
- Trend to move to machine learning.

Machine learning

- Prediction problems in high dimensionality and with symbolic data.

- In early days with heuristics algorithms.

- Tend to statistical models underlying the algorithms.

Khoa học dữ liệu: Quyết định & Hành động

Phân tích   Con người quyết định

Dữ liệu

Mô tả (descriptive)
Điều gì đã xảy ra?

Chẩn đoán (diagnostic)
Sao lại như vậy?

Dự đoán (predictive)
Điều gì sẽ xảy ra?

Đề xuất (Prescriptive)
Phải làm gì, làm thế nào?

Hỗ trợ quyết định

Tự động quyết định

Quyết định

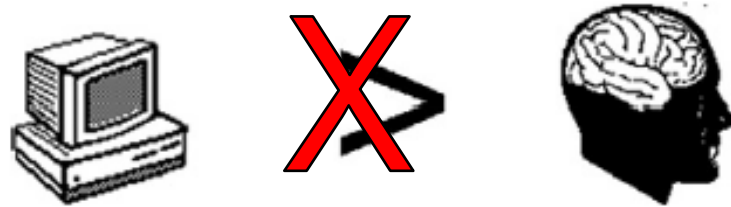**Hành động**

Nguồn: Gartner, 2013.

Data science is **the essential tool for using data**

Outline

- Brief of data science

- **The data-driven approach and electronic medical records (EMRs)**

- Our project on EMRs data analytics

"Fundamental theorem" on using computers in medicine

Charles P. Friedman. J Am Med Inform Assoc. 2009;16:169–170.

Expert systems in medicine

the deduction approach

An expert system is a computer program that behaves like an expert in some narrow area of expertise, using expert knowledge.

Knowledge acquisition

Expert Knowledge

MYCIN (Shortliffe, Feigenbaum, 1979): Infection Diagnosis.
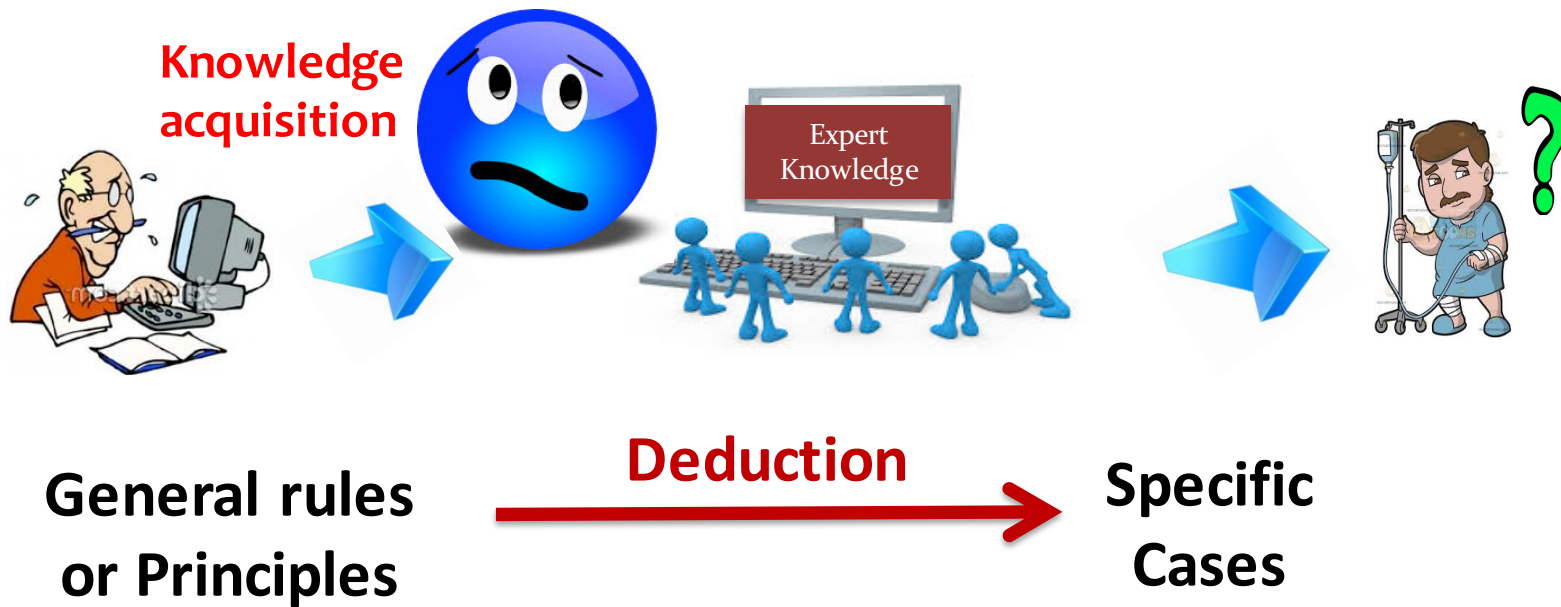
IF   1. the infection is primary bacteremia, and

        2. the site of the culture is one of the sterile sites, and

        3. the suspected portal of entry of the organism is the gastro intestinal tract

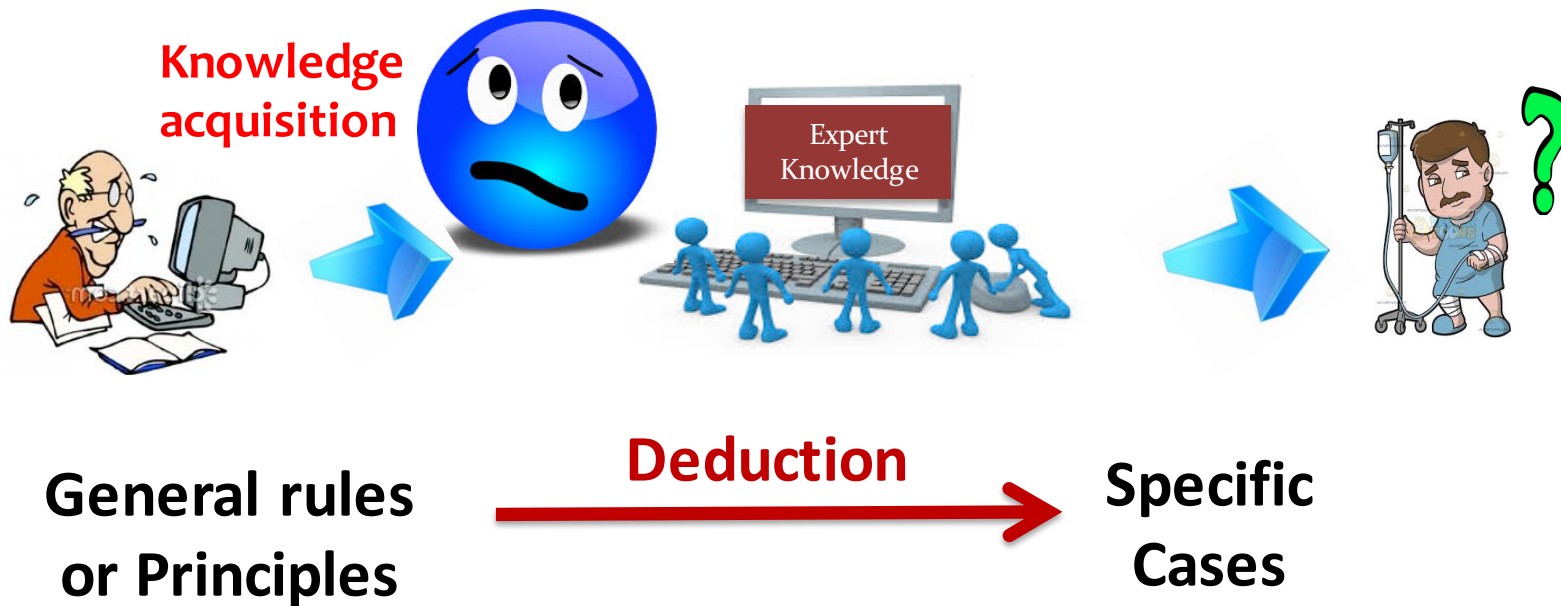THEN there is suggestive evidence (0.7) that the identity of the organism is bacteroides.
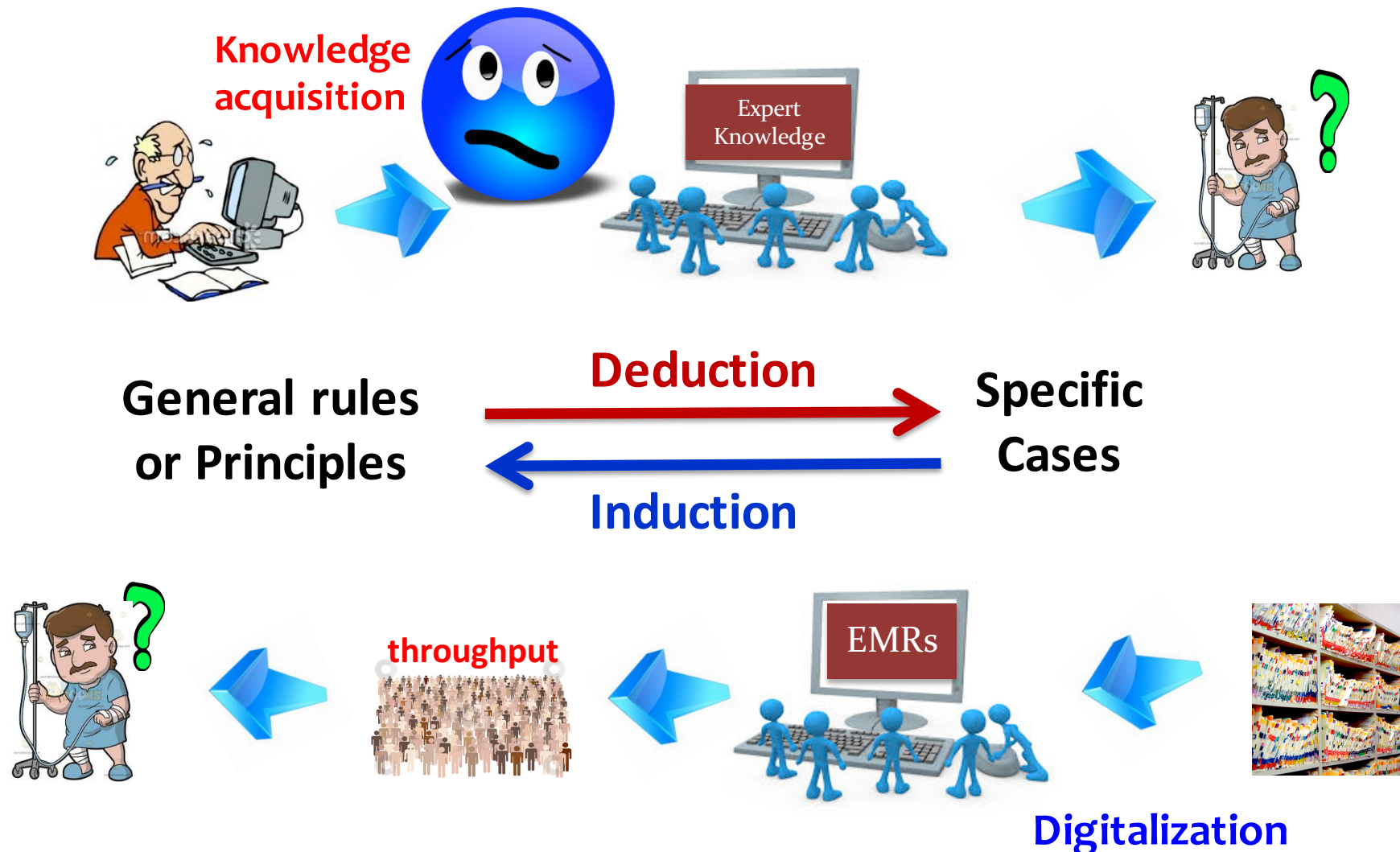
EMRs in medicine
the induction approach

Knowledge acquisition

Expert Knowledge

General rules or Principles

Deduction

Specific Cases

EMRs in medicine
the induction approach

Knowledge acquisition

Expert Knowledge

General rules or Principles → **Deduction** → **Specific Cases**

Jaundice is yellowing of the skin and eyes and can indicate a serious problem with liver, gallbladder, or pancreas function

EMRs in medicine
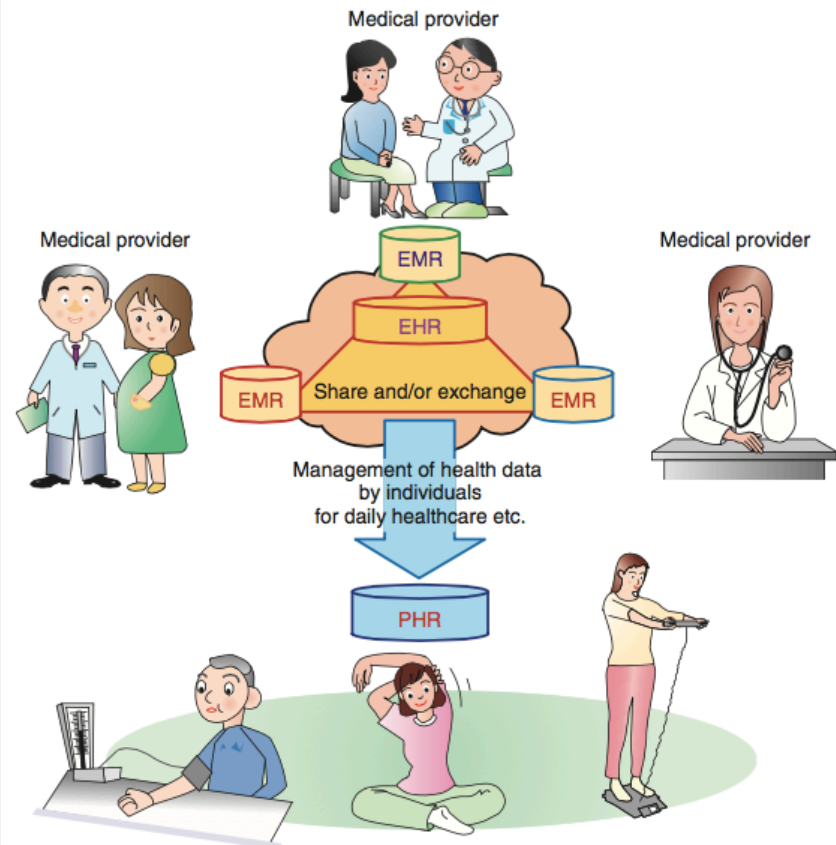## the induction approach (data-driven approach)

Knowledge acquisition

Expert Knowledge

General rules or Principles

Deduction →

Specific Cases

← Induction

throughput

EMRs

Digitalization

EMR, EHR, and PHR?

- **EMR - electronic medical records** – created, gathered, managed, and consulted by authorized clinicians and staff *within one health care organization.*

- **EHR - electronic health records** - conforms to nationally recognized interoperability standards and that can be created, managed, and consulted by authorized clinicians and staff *across more than one health care organization.*

- **PHR - personal health records** - conforms to nationally recognized interoperability standards and that can be drawn from multiple sources while being *managed, shared, and controlled by the individual.*

Yasuo Ishigure, Trends, Standardization, and Interoperability of Healthcare Information, NTT Technical Review 2017
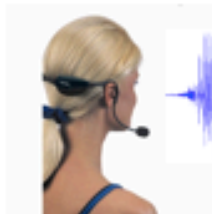
EMR: implementation & exploitation

Language normalization and mobile devices

EMR implementation support

Using speech technology to support EMR data entry

Hospital information system (HIS)

Return the result to HIS

Creating EMR secondary databases

Analysis of EMR secondary databases

1
2
3
4

24

EMR and paradigm shift in medicine

- **Main sources from EMRs**
  - Doctor daily notes
  - Nurse narratives
  - Discharge summary

- **Essence**
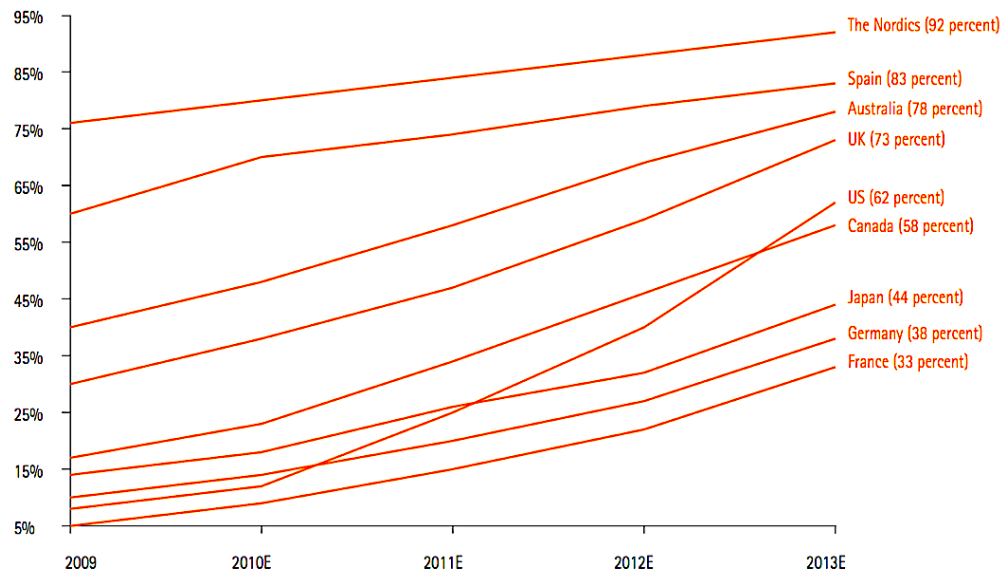 - All medical knowledge were found by the observation and analysis of the patient health care.
  - EMRs are a huge source of medical tacit knowledge accumulated on patients' diagnosis and treatment.

- **Paradigm shift**

  EMRs can play an evolutional role in medical care and research.

Obama's big idea: Digital health records

President-elect Barack Obama, as part of his effort to revive the economy, is proposing a massive effort to modernize health care by making all health records standardized and electronic. The government estimates about 212,000 jobs could be created by this program, CNNMoney reports. full story
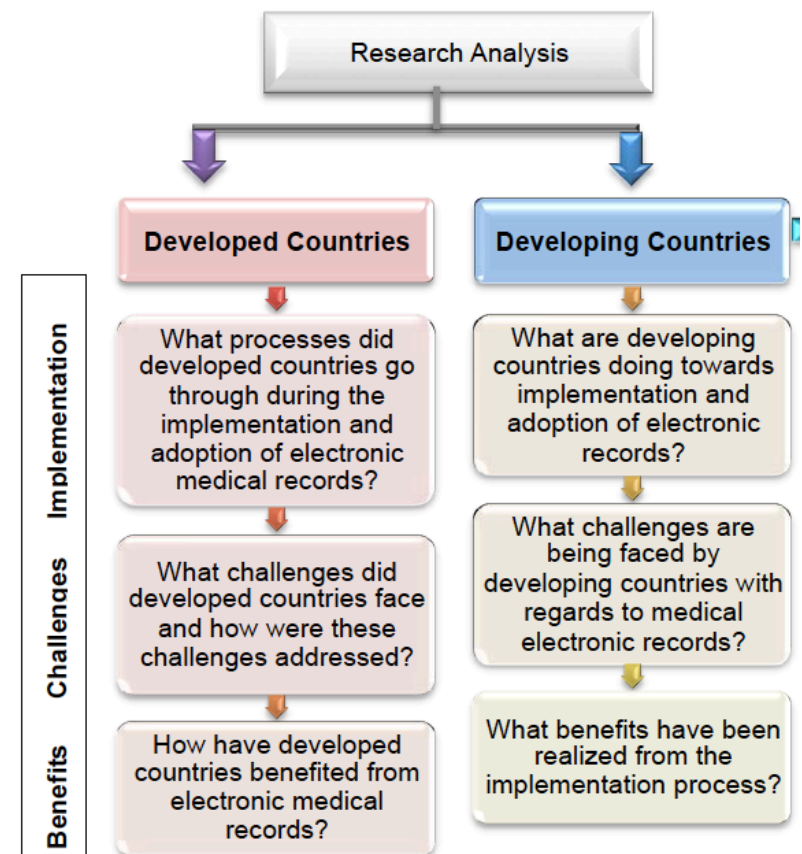

EMR Documentation Style
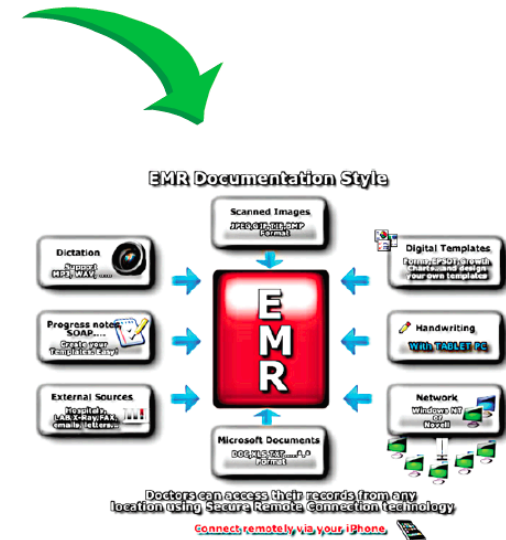
"...within five years, all of America's medical records are computerized..." (Jan. 2009)

EMRs around the world

Estimated Hospital-based EMR Adoption Rate Projections by Country

The Nordics (92 percent)
Spain (83 percent)
Australia (78 percent)
UK (73 percent)
US (62 percent)
Canada (58 percent)
Japan (44 percent)
Germany (38 percent)
France (33 percent)

EMR implementation has been well done in developed countries, and now the beginning of EMRs exploitation. The EMRs implementation has started in some developing countries.

Research Analysis

Developed Countries

Developing Countries

Implementation

What processes did developed countries go through during the implementation and adoption of electronic medical records?

What are developing countries doing towards implementation and adoption of electronic records?

Challenges

What challenges did developed countries face and how were these challenges addressed?

What challenges are being faced by developing countries with regards to medical electronic records?

Benefits

How have developed countries benefited from electronic medical records?

What benefits have been realized from the implementation process?

Percentage of primary care physicians in selected countries using EMR in 2015

Country	Percentage
New Zealand	100%
Sweden	99%
Norway	99%
Netherlands	98%
United Kingdom	98%
Australia	92%
United States	84%
Germany	84%
France	75%
Canada	73%
Switzerland	54%

Percentage of primary care physicians

27

Tradition medical records vs. Electronic medical records

Tradition medical records vs. Electronic medical records

Mostly individually used for single patient.

Tradition medical records vs. Electronic medical records

Can simultaneoulsly analyze EMRs of many patients.

Tradition medical records vs. Electronic medical records

Two kinds of data in EMRs

984,20123,1216,0,3354-02-05   05:40:00   EST,3354-02-05   06:01:00
EST,4270,"N",54,"Nursing/Other", "MICU nursing admission note 7AM","
MICU nursing admission note 7AM
Pt is 68 yo male adm [**Hospital1 2**] EW [**2-4**] s/p fall 2 weeks ago while
in [**State 552**] where he landed on his left side, having left sided pain. CXR x
2 at hospital, no fx, pt sent home. Took motrin for pain steadily last 2 weeks. ^
SOB, anorexia last 2-3 days. Flew to [**Location (un) 175**] for medical care. In
EW, + EKG changes, + troponin/MB. ARF, cr 3.4, K 5.4. Given Kayexelate, D50,IV
insulin, CaGluc. Heparin gtt started for EKG changes, ?PE. No CT d/t ARF. VQ
scan showed low prob PE.   Also FSBS 300s, covered by SQ insulin.
Vanco/levoquin for ? UTI. Desatted on RA, 100% NRB with SAts 100%. CXR no
rib Fx. Bicarb gtt for acidosis, gap 26. Hemodynamically stable, BP decreased
80s x 2 while sleeping, increased when awake. A&O x 3. Tx MICU for further
management. ARF probably d/t motrin use.

Neuro - A&O x 3. C/O left sided pain when turning, otherwise comfortable.
MAE.

Resp - Weaned O2 NC 6L, SAts 94%. Lungs clear, diminished at bases. No SOB.

CV - BP 103-118/54-59. NSR 70s-80s, no ectopy. Heparin gtt 1450U/hr. PTT
>150, shut off @ 4:30. Restarted 6:00 @ 1200U/hr. K 5.7->6.5. EKG unchanged.
2amps CaGluc, insulin 10U IV, 30gm Kayexelate given. Pt has had no stool from
any kayexelate given. 4:30 lytes will not reflect

X-ray, CT scan, MRI, ... in PACS

Lab examination (blood, cardiogram...)

CLINICAL DATA (clinical text)   **PARA-CLINICAL DATA (numbers)**

Two kinds of data in EMRs

984,20123,1216,0,3354-02-05   05:40:00   EST,3354-02-05   06:01:00
EST,4270,"N",54,"Nursing/Other", "MICU nursing admission note 7AM","
MICU nursing admission note 7AM
Pt is 68 yo male adm [**Hospital1 2**] EW [**2-4**] s/p fall 2 weeks ago while
in [**State 552**] where he landed on his left side, having left sided pain. CXR x
2 at hospital, no fx, pt sent home. Took motrin for pain steadily last 2 weeks. ^
SOB, anorexia last 2-3 days. Flew to [**Location (un) 175**] for medical care. In
EW, + EKG changes, + troponin/MB. ARF, cr 3.4, K 5.4. Given Kayexelate, D50,IV
insulin, CaGluc. Heparin gtt started for EKG changes, ?PE. No CT d/t ARF. VQ
scan showed low prob PE.   Also FSBS 300s, covered by SQ insulin.
Vanco/levoquin for ? UTI. Desatted on RA, 100% NRB with SAts 100%. CXR no
rib Fx. Bicarb gtt for acidosis, gap 26. Hemodynamically stable, BP decreased
80

Ne
MA

Resp - Weaned O2 NC 6L, SAts 94%. Lungs clear, diminished at bases. No SOB.

CV - BP 103-118/54-59. NSR 70s-80s, no ectopy. Heparin gtt 1450U/hr. PTT
>150, shut off @ 4:30. Restarted 6:00 @ 1200U/hr. K 5.7->6.5. EKG unchanged.
2amps CaGluc, insulin 10U IV, 30gm Kayexelate given. Pt has had no stool from
any kayexelate given. 4:30 lytes will not reflect

Heterogeneous and longitudinal

X-ray, CT scan, MRI, ... in PACS

RBC(Hồng cầu)	4.67	/mm^3	4.0 - 5.9	4.0 - 5.9	06/10/2016 14:5
RDW	40.1	%	8.0 - 12.0	8.0 - 12.0	06/10/2016 14:5
THR(Tiểu cầu)	238	/mm^3	150 - 450	150 - 450	06/10/2016 14:5
WBC(Bạch cầu)	6.9	/mm^3	4.0 - 10.0	4.0 - 10.0	06/10/2016 14:5
Tổng phân tích nước tiểu (B					
pH	7.0		4.8-7.4	4.8-7.4	06/10/2016 14:5
BIL (Bilirubin)	Âm tính	umol/L	<3.4	<3.4	06/10/2016 14:5
BLO (Hồng cầu)	VẾT	/µ	<5	<5	06/10/2016 14:5
GLU (Glucose nước tiểu)	Âm tính	mmol/L	3.7 - 6.2	3.7 - 6.2	06/10/2016 14:5
KET (Ketone)	Âm tính	mmol/L	<5	<5	06/10/2016 14:5
LEU (Bạch cầu)	-	/µ	<10	<10	06/10/2016 14:5

Lab examination (blood, cardiogram...)

CLINICAL DATA (clinical text)   **PARA-CLINICAL DATA (numbers)**

Public EMR database MIMIC II & III

- Shared EMR database for research (more than 50 thousand de-identified EMRs)
- MIMIC II (2001-2008), MIMIC III (2009-2012)

https://mimic.physionet.org/

Tables in MIMIC
- ADMISSIONS
- CALLOUT
- CAREGIVERS
- CHARTEVENTS
- CPTEVENTS
- D_CPT
- D_ICD_DIAGNOSES
- D_ICD_PROCEDURES
- D_ITEMS
- D_LABITEMS
- DATETIMEEVENTS
- DIAGNOSES_ICD
- DRGCODES
- ICUSTAYS
- INPUTEVENTS_CV
- INPUTEVENTS_MV
- LABEVENTS
- MICROBIOLOGYEVEN
- NOTEEVENTS
- OUTPUTEVENTS
- PATIENTS
- PRESCRIPTIONS
- PROCEDUREEVENTS_
- PROCEDURES_ICD
- SERVICES
- TRANSFERS

Some pieces from an EMR in MIMIC

ICUSTAY

```
26,2538-10-29,4320,"N",1,1,"Y","Y",2538-10-26 03:18:00 EST,2538-10-29
16:25:00EST,58.95198,"adult",5107,"N","CCU","CCU","CCU","CCU",185.42,100.4,100.4,
100.4,16,5,16,5,1,5,
```

ICD_DISGNOSIS AGE

```
SUBJECT_ID,HADM_ID,SEQUENCE,CODE,DESCRIPTION
25,5726,1,"410.71","SUBENDOCARDIAL INFARCTION INITIAL EPISODE OF CARE"
25,5726,2,"250.11","DIABETES MELLITUS WITH KETOACIDOSIS TYPE I NOT STA"
25,5726,3,"414.01","CORONARY ATHEROSCLEROSIS OF NATIVE CORONARY ARTERY"
25,5726,4,"401.9","UNSPECIFIED ESSENTIAL HYPERTENSION"
```

DEMOGRAPHIC EVENTS DATA

```
SUBJECT_ID,HADM_ID,MARITAL_STATUS_ITEMID,MARITAL_STATUS_DESCR,ETHNICITY_ITEMID,ET
HNICITY_DESCR,OVERALL_PAYOR_GROUP_ITEMID,OVERALL_PAYOR_GROUP_DESCR,RELIGION_ITEMI
D,RELIGION_DESCR,ADMISSION_TYPE_ITEMID,ADMISSION_TYPE_DESCR,ADMISSION_SOURCE_ITEM
ID,ADMISSION_SOURCE_DESCR
25,5726,200050,"MARRIED",200083,"WHITE",200067,"PRIVATE",200081,"UNOBTAINABLE",20
0028,"EMERGENCY",200029,"EMERGENCY ROOM ADMIT"
```

MEDEVENTS DATA

```
SUBJECT_ID,ICUSTAY_ID,ITEMID,CHARTTIME,ELEMID,REALTIME,CGID,CUID,VOLUME,DOSE,DOSEUOM,SOLUTIONID,SOLVOLUME,
LUNITS,ROUTE,STOPPED
25,28,45,2538-10-26 04:30:00 EST,1,2538-10-26 04:57:00 EST,2691,1,0,8,"Uhr",18,100,"ml","IV Drip",
25,28,142,2538-10-26 04:30:00 EST,1,2538-10-26 05:00:00 EST,2691,1,0,2,"mcgkgmin",13,100,"ml","IV Drip",
25,28,45,2538-10-26 04:45:00 EST,1,2538-10-26 04:57:00 EST,2691,1,0,10,"Uhr",18,100,"ml","IV Drip",
25,28,142,2538-10-26 04:45:00 EST,1,2538-10-26 05:00:00 EST,2691,1,0,2,"mcgkgmin",13,100,"ml","IV Drip",
25,28,45,2538-10-26 05:00:00 EST,1,2538-10-26 05:23:00 EST,2049,1,0,10,"Uhr",18,100,"ml","IV Drip",
25,28,142,2538-10-26 05:00:00 EST,1,2538-10-26 05:23:00 EST,2049,1,0,2,"mcgkgmin",13,100,"ml","IV Drip",
25,28,45,2538-10-26 05:15:00 EST,1,2538-10-26 06:07:00 EST,2691,1,0,10,"Uhr",18,100,"ml","IV Drip",
25,28,142,2538-10-26 05:15:00 EST,1,2538-10-26 06:07:00 EST,2691,1,0,2,"mcgkgmin",13,100,"ml","IV Drip",
25,28,45,2538-10-26 05:30:00 EST,1,2538-10-26 06:07:00 EST,2691,1,0,10,"Uhr",18,100,"ml","IV Drip",
25,28,142,2538-10-26 05:30:00 EST,1,2538-10-26 06:07:00 EST,2691,1,0,2,"mcgkgmin",13,100,"ml","IV Drip",
```

MEDURATIONS DATA

```
SUBJECT_ID,ICUSTAY_ID,ITEMID,ELEMID,STARTTIME,STARTREALTIME,ENDTIME,CUID,DURATION
25,28,45,1,2538-10-26 04:30:00 EST,2538-10-26 04:57:00 EST,2538-10-29 16:25:00 EST,1,5035
25,28,142,1,2538-10-26 04:30:00 EST,2538-10-26 05:00:00 EST,2538-10-29 16:25:00 EST,1,5035
25,28,45,1,2538-10-26 04:45:00 EST,2538-10-26 04:57:00 EST,2538-10-29 16:25:00 EST,1,5020
25,28,142,1,2538-10-26 04:45:00 EST,2538-10-26 05:00:00 EST,2538-10-29 16:25:00 EST,1,5020
25,28,45,1,2538-10-26 05:00:00 EST,2538-10-26 05:23:00 EST,2538-10-29 16:25:00 EST,1,5005
25,28,142,1,2538-10-26 05:00:00 EST,2538-10-26 05:23:00 EST,2538-10-29 16:25:00 EST,1,5005
```

POE-MED DATA

```
POE_ID,DRUG_TYPE,DRUG_NAME,DRUG_NAME_GENERIC,PROD_STRENGTH,FORM_RX,DOSE_VAL_RX,DOSE_UNIT_RX,FORM_VAL_
DISP,FORM_UNIT_DISP,DOSE_VAL_DISP,DOSE_UNIT_DISP,DOSE_RANGE_OVERRIDE
1930588,"BASE","D5W",,"250mL Bag",,"250","ml","250","ml",,,
1930589,"BASE","NS",,"500mL Bag",,"500","ml","500","ml",,,
1936709,"BASE","SW",,"100mL Bottle",,"100","ml","100","ml",,,
1929791,"MAIN","Aspirin","Aspirin","325mg Tab",,"325","mg","1","TAB",,,
1929796,"MAIN","Potassium Chloride","Potassium Chloride","20mEq Packet",,"20","mEq","1","PKT",,,
1929797,"MAIN","Atorvastatin","Atorvastatin","40mg Tab",,"80","mg","2","TAB",,,
1929819,"MAIN","Potassium Chloride","Potassium Chloride","20mEq Packet",,"40","mEq","2","PKT",,,
1930558,"MAIN","Potassium Chloride","Potassium Chloride","20mEq Packet",,"40","mEq","2","PKT",,,
1930691,"MAIN","Pantoprazole","Pantoprazole","40mg Tab",,"40","mg","1","TAB",,,
1931503,"MAIN","Calcium Gluconate","Calcium Gluconate","1g/10mL Vial",,"2","gm","2","VIAL",,,
1931745,"MAIN","Zolpidem Tartrate","Zolpidem Tartrate","5mg Tab",,"5-10","mg","1-2","TAB",,,
1931746,"MAIN","Acetaminophen","Acetaminophen","325mg Tab",,"325-650","mg","1-2","TAB",,,
...
```

POR-ORDER DATA

```
POE_ID,SUBJECT_ID,HADM_ID,ICUSTAY_ID,START_DT,STOP_DT,ENTER_DT,MEDICATION,PROCEDURE_TYPE,STATUS,ROUTE
,FREQUENCY,DISPENSE_SCHED,IV_FLUID,IV_RATE,INFUSION_TYPE,SLIDING_SCALE,DOSES_PER_24HRS,DURATION,DU
RATION_INTVL,EXPIRATION_VAL,EXPIRATION_UNIT,EXPIRATION_DT,LABEL_INSTR,ADDITIONAL_INSTR,MD_ADD_INST
R,RNURSE_ADD_INSTR
1929790,25,5726,28,2538-10-26 05:00:00 EST,2538-10-27 03:00:00 EST,2538-10-26 04:00:00
EST,"Insulin","IV Piggyback","Inactive (Due to a change order)","IV
DRIP","INFUSION",,,,,,,"Ongoing",,"Enter on Label",,,"Fingersticks every hour  IV Drip Rate: 8
UNIT/HR","Specify blood glucose goal",
1929795,25,5726,28,2538-10-26 05:00:00 EST,2538-10-26 04:00:00 EST,2538-10-26 04:00:00 EST,"Potassium
Chloride","IV Piggyback","Discontinued","IV","ONCE","5",,,,,1,1,"Doses",,"Enter on
Label",,,,"CARDIAC MONITORING AND CENTRAL LINE ARE REQUIRED WHEN SELECTING CONCENTRATED PRODUCT
(20 mEq/50 mL).  20 mEq/50 ml preparations are given via central line only. Fluid restricted
patients may recieve 40 mEq in 500 ml NS or D5W. No more than 60 mEq placed in one liter of fluid
per BIDMC policy. ","Cardiac monitoring and central lines are required for rates > 10 mEq/hr."
```

NOTEEVENTS DATA

```
"",,25,5726,28,0,2538-10-26 07:51:00 EST,2538-10-26 08:33:00 EST,1807,"N",1,"Nursing/Other","NURSING PROGRESS NOTE",
```

NURSING PROGRESS NOTE
58 Y/O MALE ADMITTED FROM [**Hospital1 2**] ER (TRANSFERED FROM [**Hospital6 110**]). HE INITIALLY PRESENTED TO [**Hospital6 110**] WITH C/O N/V, DIZZINESS. HE IS S/P INSULIN PUMP INSERTION IN [**2538-5-6**]. HIS PUMP FAILED ON SATURDAY AND BEGAN FEELING POORLY. HE WAS ADMITTED WITH A BLOOD GLUCOSE > 575. HE ALSO HAD ST CHANGES ON EKG. HE WAS TREATED WITH IV LOPRESSOR, INTEGRILLIN, IV NS, INSULIN. HE REFUSED ASA STATING IT MAKES HIS STOMACH UPSET. ADMITTED TO CCU FOR R/O MI PROTOCOL.

This is a 58 yr old male Pt who presented to [**Hospital6 **] with C/O N/V & dizziness- He had an insulin pump inserted in 6/04 & on Saturday [**10-25**] it failed- blood sugar was > 500- Also, his EKG showed new ST depressions (no C/O CP & cardiac enzymes negative)- Pt was transferred to [**Hospital1 2**] EW on integrilin & insulin gtts for further care- Pt was admitted to CCU- R radial A line was placed- Pt developed a sinus arrythemia HR 40-70's with hypotension (SBP 60-70's)- atropine given for ? bradycardia induced hypotension, IV fluids wide open & dopamine gtt started- EKG SA HR 50-70's with return of ST-T waves changes in lateral leads- PA line inserted into R IJ- RA 8- PAP 42/22-PCWP 15-16- decision was made to cath Pt due to persistent hypotension Cardiac cath revealed moderately severe single vessel CAD (LCx large vessel proximal 60-70%) normal LV systolic function- no intervention done- ? elective stent LCx when stable- CO high with low SVR- ? sepsis

CV-R/I MI. HR 70-80NSR, BP by R radial Aline 110-140/60-70. ASA, plavix (loaded w/ 300mg this am) cont., lopressor 12.5mg bid added. No c.o CP, weakness, dizz. PA line- CVP 8-10, PA 28-38/16-18, CO [**7-15**], SVR 500. Has received ~10liters of IVF over 48hr, u/o 3000 over same time. R femoral Aline d/c @ 12n w/o complication by Card fellow, site C&D w/ transparent dsg, no hematoma, no oozing. Pulses dpl-1+, baseline. Endo/Fluids- IDDM on insulin gtt @2-3u/hr w/ small and improving po intake.FS 92-152. IVF D5.45NS @ 100cc/h (dec'd from 150/hr this am). U/O 80-120/hr clear urine. + 2500 for day.

CCU Nursing Progress Note-7a-7p 58 y/o male admitted [**10-25**] w/ N/V/dizz, IDDM w/ failed pump, FS 576 to [**Hospital6 **], EKG changes. Placed on insulin gtt, IVF and tx to [**Hospital1 2**]. Over w/e,hypotensive- Dopa and Levo; PA line placed w/ High CO, low SVR; cathed, RCA 70% stenosis, RI MI; DKA. Much improved overnight and today. Anion gap now closed. Heparin, R fem Aline d/c. Cont INS gtt, IVF, antibx. Plan for Stent of RCA [**10-28**]. NPO p MN.
Neuro- A&O x3, MAE, much less irritable w/ cardiac explanation/education by MD/RN CCU team. Able to assist w/ position change. To be OOB this evening when PA line D/C.

CHARTDURATION DATA

```
SUBJECT_ID,ICUSTAY_ID,ITEMID,ELEMID,STARTTIME,STARTREALTIME,ENDTIME,CUID,DURATION
25,28,781,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1536,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1535,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1534,0,2538-10-26 03:59:00 EST,2538-10-26 09:29:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1532,0,2538-10-26 03:59:00 EST,2538-10-26 09:29:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1529,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1525,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1523,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1522,0,2538-10-26 03:59:00 EST,2538-10-26 09:29:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1162,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
```

EMR is the core of the e-health paradigm shift

Improving the quality and efficiency of hospital jobs

Improving safety by preventing incidents

Improving security in relation to the personal information protection

Providing real data for pragmatic clinical trials in clinical research.

Quality & efficiency

Improving safety

Improving security

Medical research

EMR
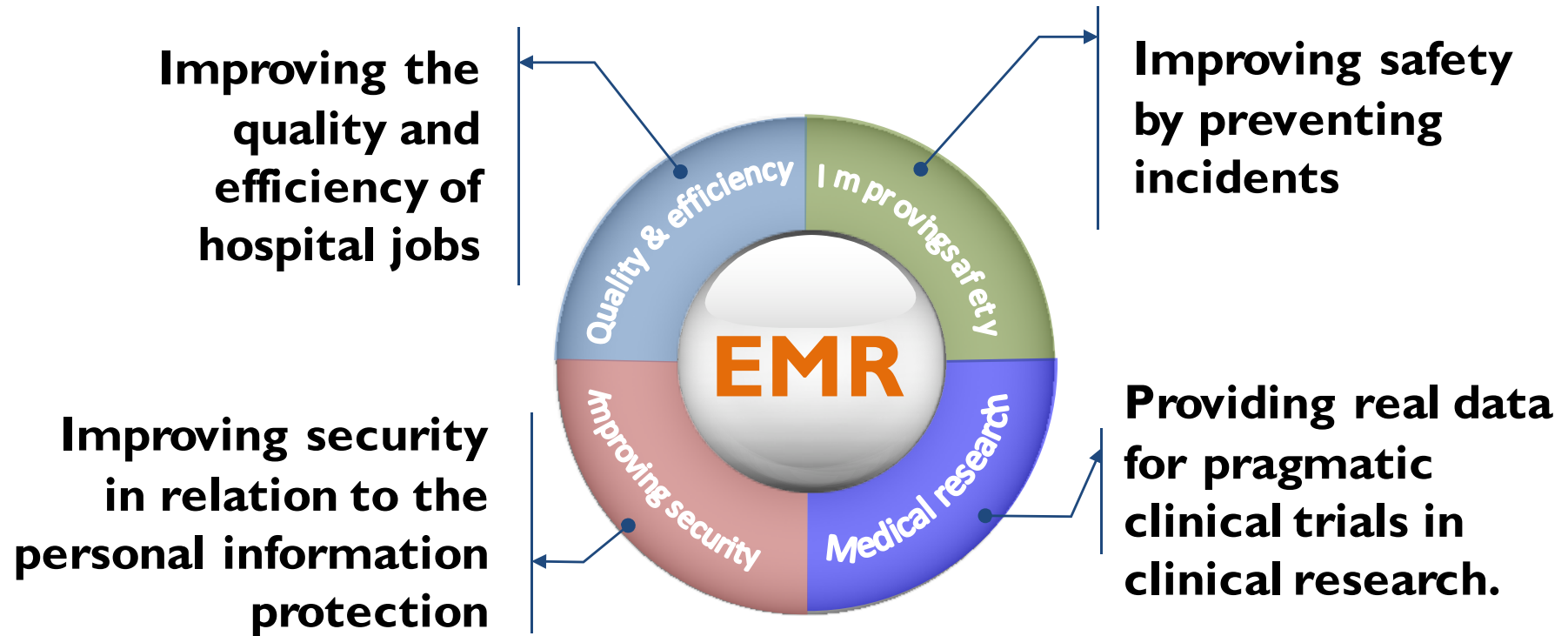
EMR Data Analytics

Example: IBM Watson health

IBM Watson Health   IBM

Systems of Insight have the potential to redefine clinical decision support in the context of both Knowledge-driven and Data-driven Analytics.

Knowledge-Driven   **Data-Driven**

- Published Knowledge
  - Journals
  - Books
  - Guidelines

- Observational Data
  - Longitudinal records
  - Claims, Rx, Labs
  - Genomics
  - Exogenous
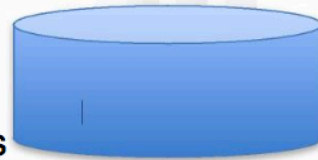
- Discovery Advisor
- Oncology Advisor
- Engagement Advisor
- Clinical Trials Matching
- EMR Assist
- Genomics Advisor

- Patient similarity
- Predictive modeling
- Real World Evidence
- Genomics
- Visual analytics

© 2015 IBM Corporation   13

IBM Watson's Next Step: Health, 2017

Research for EMR exploitation

- **I2B2 Challenges (English)**

  (Informatics for Integrating Biology and the Bedside)

  - 2006: De-identification and Smoking

  - 2008: Obesity

  - ...

  - 2011: Co-reference Challenge

  - 2012: Temporal Relations Challenge

  - 2014: De-identification & risk factors for heart disease

  - 2015: Temporal Relations Challenge

  - 2016: Patient severity for a disease

- **NTCIR Challenges (日本語)**

  - 2013: De-identification

  - 2014: Recognize disease complaints and diagnosis,
    and links to ICD-10.

  - 2015: Disease recognition from ICD codes

  - 2016: Temporal Relations Challenge

https://www.i2b2.org
http://research.nii.ac.jp/ntcir/publication1-en.html

MIMIC
EMR open access DB

Pittsburgh Univ. DB
of clinical reports
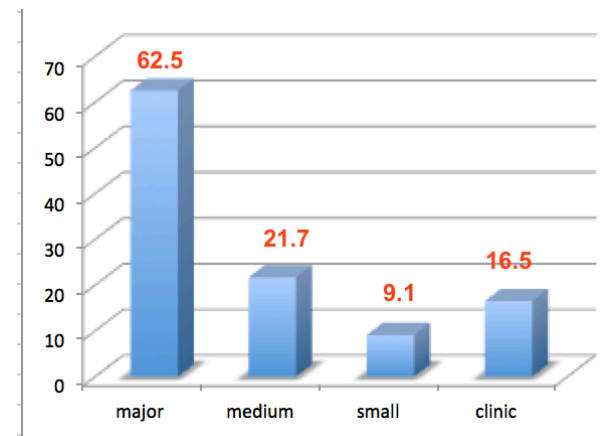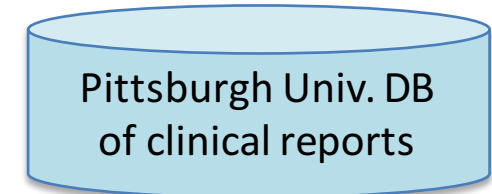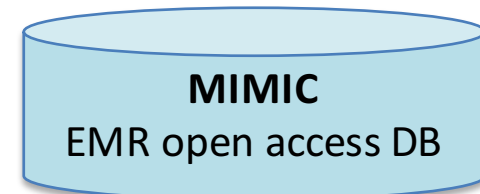
Rates of Japanese hospitals with EMRs (2011)

Research for EMR exploitation
Journal of Machine Learning Research

JMLR Special Topic on Learning from Electronic Health Data

Interleaved Text/Image Deep Mining on a Large-Scale Radiology Database for Automated Image Interpretation
Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, Ronald M. Summers; 17(107):1−31, 2016.
[abs][pdf][bib]

Patient Risk Stratification with Time-Varying Parameters: A Multitask Learning Approach
Jenna Wiens, John Guttag, Eric Horvitz; 17(209):1−23, 2016.
[abs][pdf][bib]

Extracting PICO Sentences from Clinical Trial Reports using Supervised Distant Supervision
Byron C. Wallace, Joël Kuiper, Aakash Sharma, Mingxi (Brian) Zhu, Iain J. Marshall; 17(132):1−25, 2016.
[abs][pdf][bib]

The Factorized Self-Controlled Case Series Method: An Approach for Estimating the Effects of Many Drugs on Many Outcomes
Ramin Moghaddass, Cynthia Rudin, David Madigan; 17(185):1−24, 2016.
[abs][pdf][bib]

Decrypting ``Cryptogenic'' Epilepsy: Semi-supervised Hierarchical Conditional Random Fields For Detecting Cortical Lesions In MRI-Negative Patients
Bilal Ahmed, Thomas Thesen, Karen E. Blackmon, Ruben Kuzniekcy, Orrin Devinsky, Carla E. Brodley; 17(112):1−30, 2016.
[abs][pdf][bib]

Electronic Health Record Analysis via Deep Poisson Factor Models
Ricardo Henao, James T. Lu, Joseph E. Lucas, Jeffrey Ferranti, Lawrence Carin; 17(186):1−32, 2016.
[abs][pdf][bib]

Cross-Corpora Unsupervised Learning of Trajectories in Autism Spectrum Disorders
Huseyin Melih Elibol, Vincent Nguyen, Scott Linderman, Matthew Johnson, Amna Hashmi, Finale Doshi-Velez; 17(133):1−38, 2016.
[abs][pdf][bib]

Integrative Analysis using Coupled Latent Variable Models for Individualizing Prognoses
Peter Schulam, Suchi Saria; 17(234):1−35, 2016.
[abs][pdf][bib]

Structure-Leveraged Methods in Breast Cancer Risk Prediction
Jun Fan, Yirong Wu, Ming Yuan, David Page, Jie Liu, Irene M. Ong, Peggy Peissig, Elizabeth Burnside; 17(235):1−15, 2016.
[abs][pdf][bib]

Multi-Objective Markov Decision Processes for Data-Driven Decision Support
Daniel J. Lizotte, Eric B. Laber; 17(211):1−28, 2016.
[abs][pdf][bib]

Outline

- Brief of data science

- The data-driven approach and electronic medical records (EMRs)

- **Our project on EMRs data analytics**

Project: Core technologies for exploiting EMRs

- Developed countries have established EMR implementation and going to EMR exploitation. Developing countries are mostly in the EMR enfancy (implemention phase).

- Build the core technologies for EMRs in Vietnamese.

- Do some pilot research.

Electronic Medical Records

Phase 1
EMR Transformation

EMR Cleansing   EMR de-identification   EMR Computable Forms

Phase 2
EMR Utilization

Various ways of using EMR computable forms

EMR-based diagnosis support

EMR-based study on the drug-disease relationship

Project tasks

(1) Implementation tools

(2) Vietnamese Universal Medical Language System (V-UMLS)

(3) Secondary EMR DB

(4) Insurance and hospital cost

(5) Diagnosis and treatment support

(6) Drug safety and effectiveness

EMR implementation

Resource and tools

EMR exploitation

EMR implementation: speech technology

- More than 2/3 among ~600000 physicians in the US are using speech recognition for EMR data entering (speech up 3 times).
- We are successfully used voice technologies for EMRs implementation in Vietnamese.

Secondary databases of EMRs

Hospital Information Systems

Extract and transform data into the secondary EMR DB (1)

Analyze and exploit EMR in the secondary DB (3) (2)

– Identify data to be extracted from the original EMRs.

– Build the data scheme for the secondary DB.

– Create tools to transform EMRs to secondary DB.

– Algorithms of de-identification

V-UMLS vs. UMLS

(Unified Medical Language System)

Metathesaurus

1 million+ biomedical **concepts** from over 100 sources

Semantic Network

135 broad **categories** and 54 **relationships** between them

SPECIALIST Lexicon +Tools

lexical information and programs for **language processing**

UMLS (1984-)
V-UMLS (2014-)
used for analyzing clinical text in Vietnamese
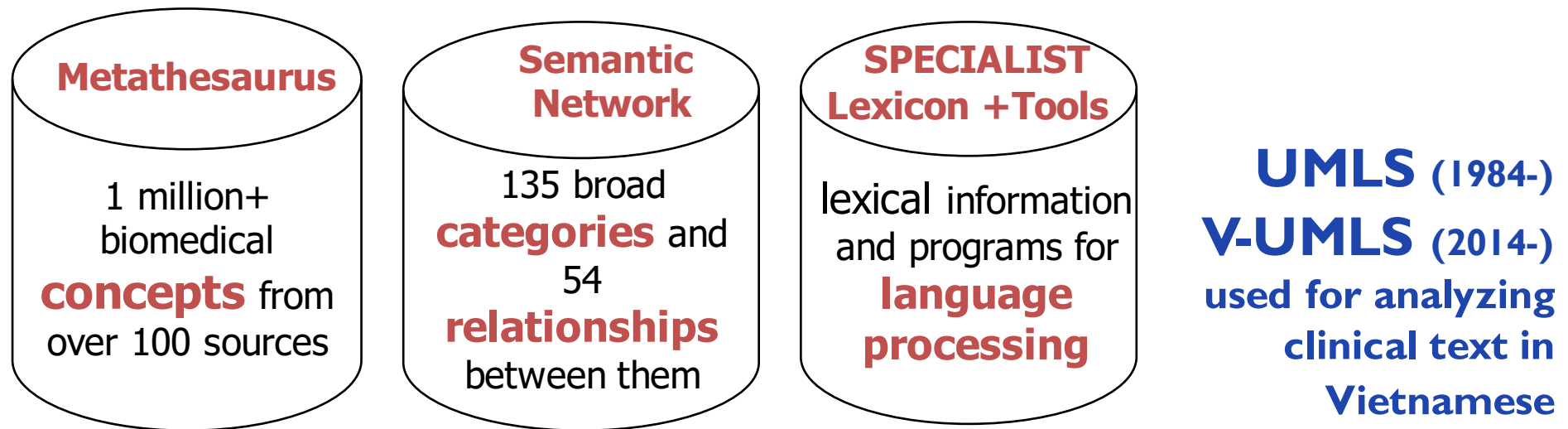
- V-Thesaurus: The concept in V-Thesaurus is mapped into semantic classes and the relationship between them.

- V-MetaMap: Medical text analysis tools and mapping of phrases into conceptual classes in V-Thesaurus.

45

UMLS: 3 components

- SPECIALIST Lexicon
  - 200,000 lexical items
  - Part of speech and variant information

    Lexical resources

- Metathesaurus
  - 5M names from over 100 terminologies
  - 1M concepts
  - 16M relations

    Terminological resources

- Semantic Network
  - 135 high-level categories
  - 7000 relations among them

    Ontological resources

UMLS Characteristics

- Number of
  - Concepts: 1.5M (2008AA)
  - Terms: ~6M
- Major organizing principles (Metathesaurus):
  - Concept orientation
  - Source transparency
  - Multi-lingual through translation
- Formalism: Proprietary format (RRF)

UMLS Integrating subdomains
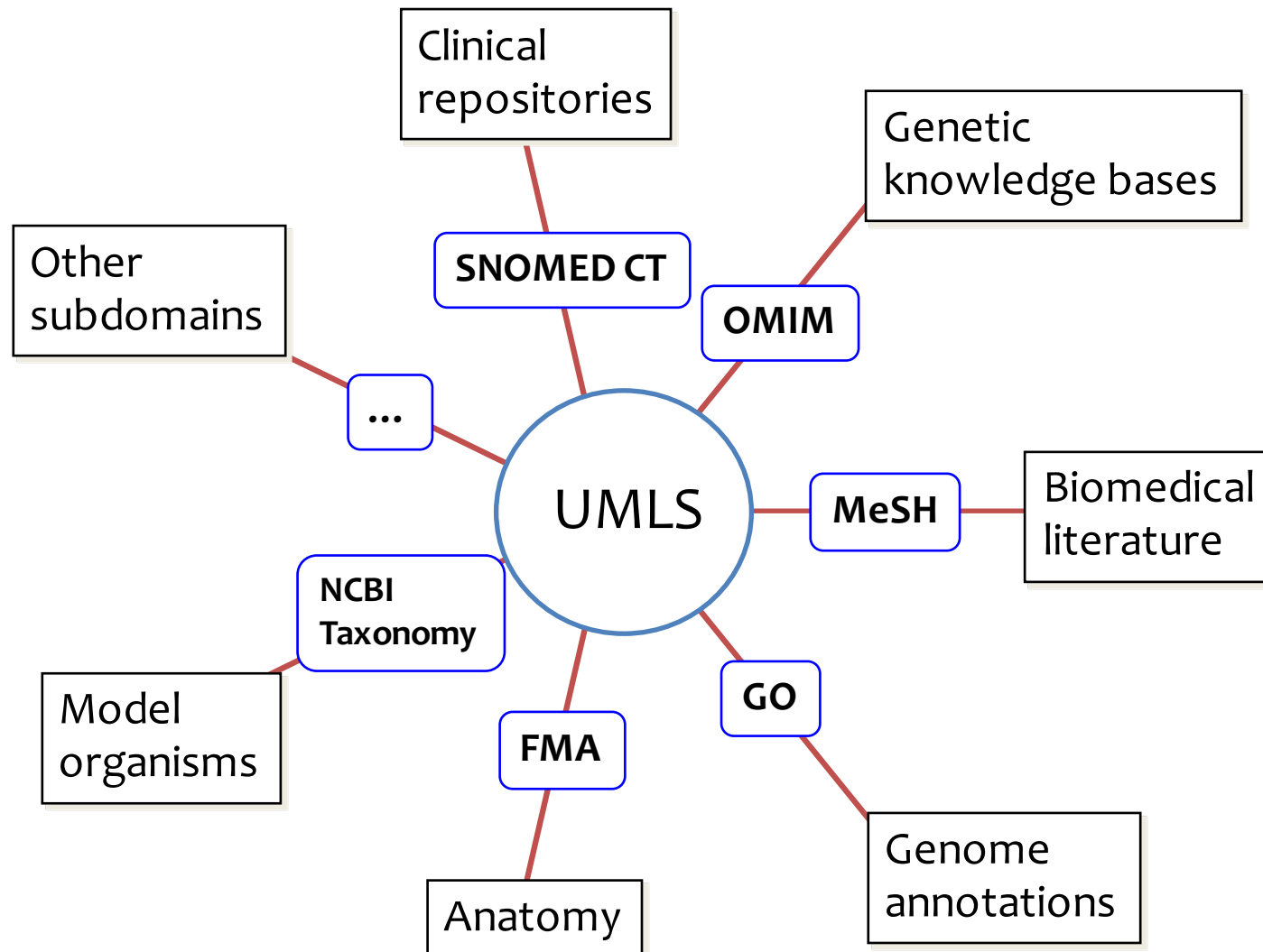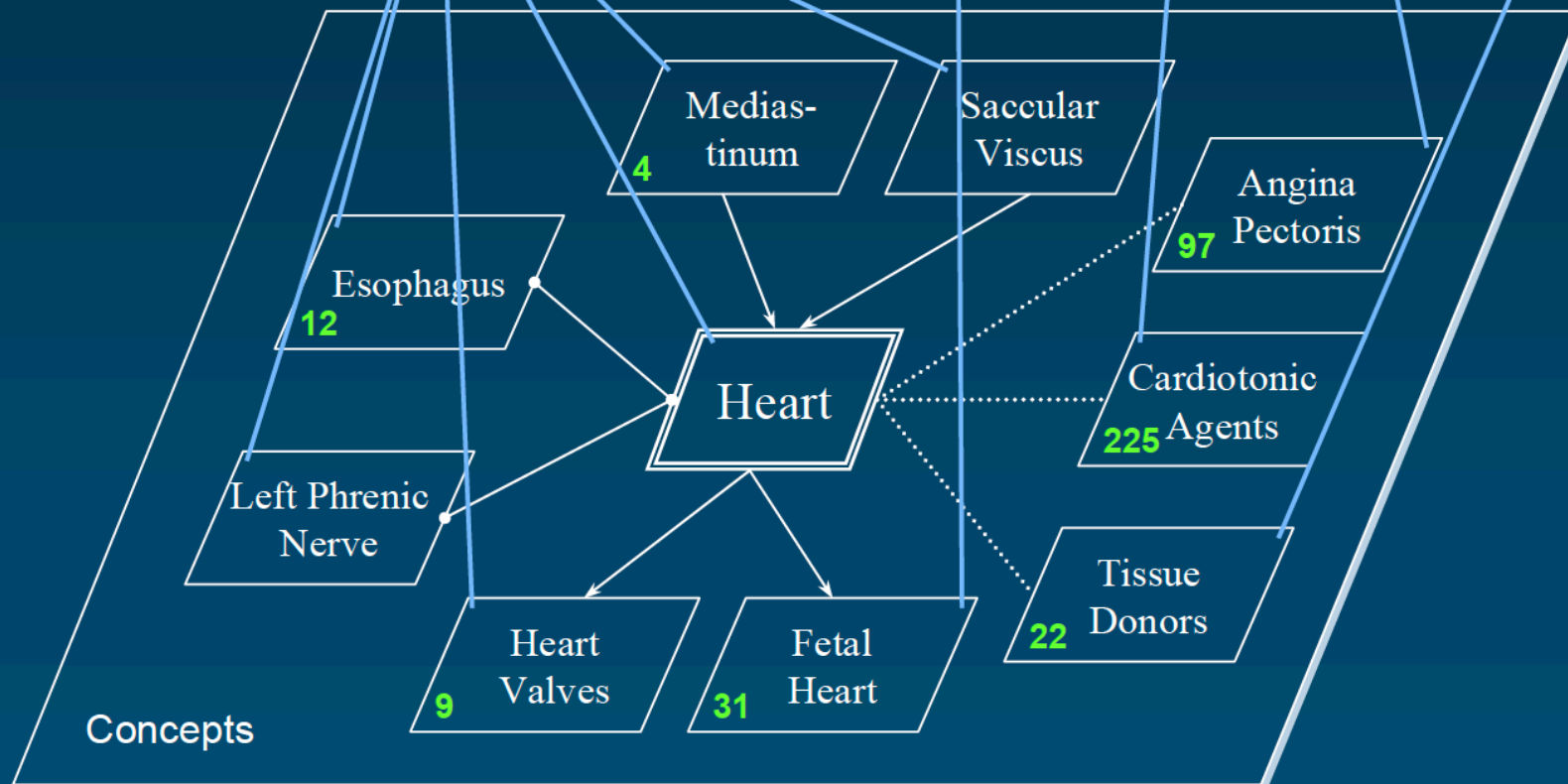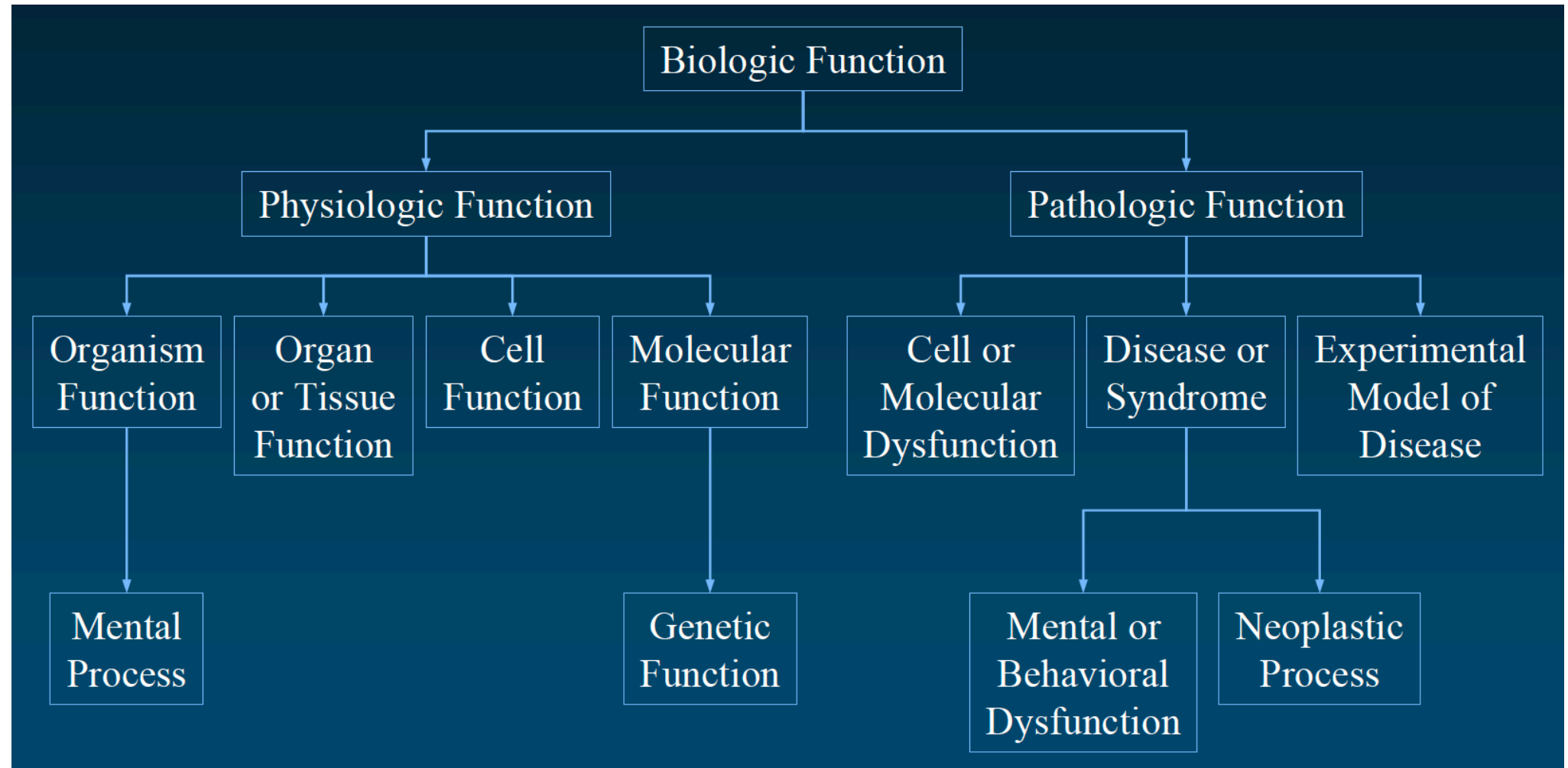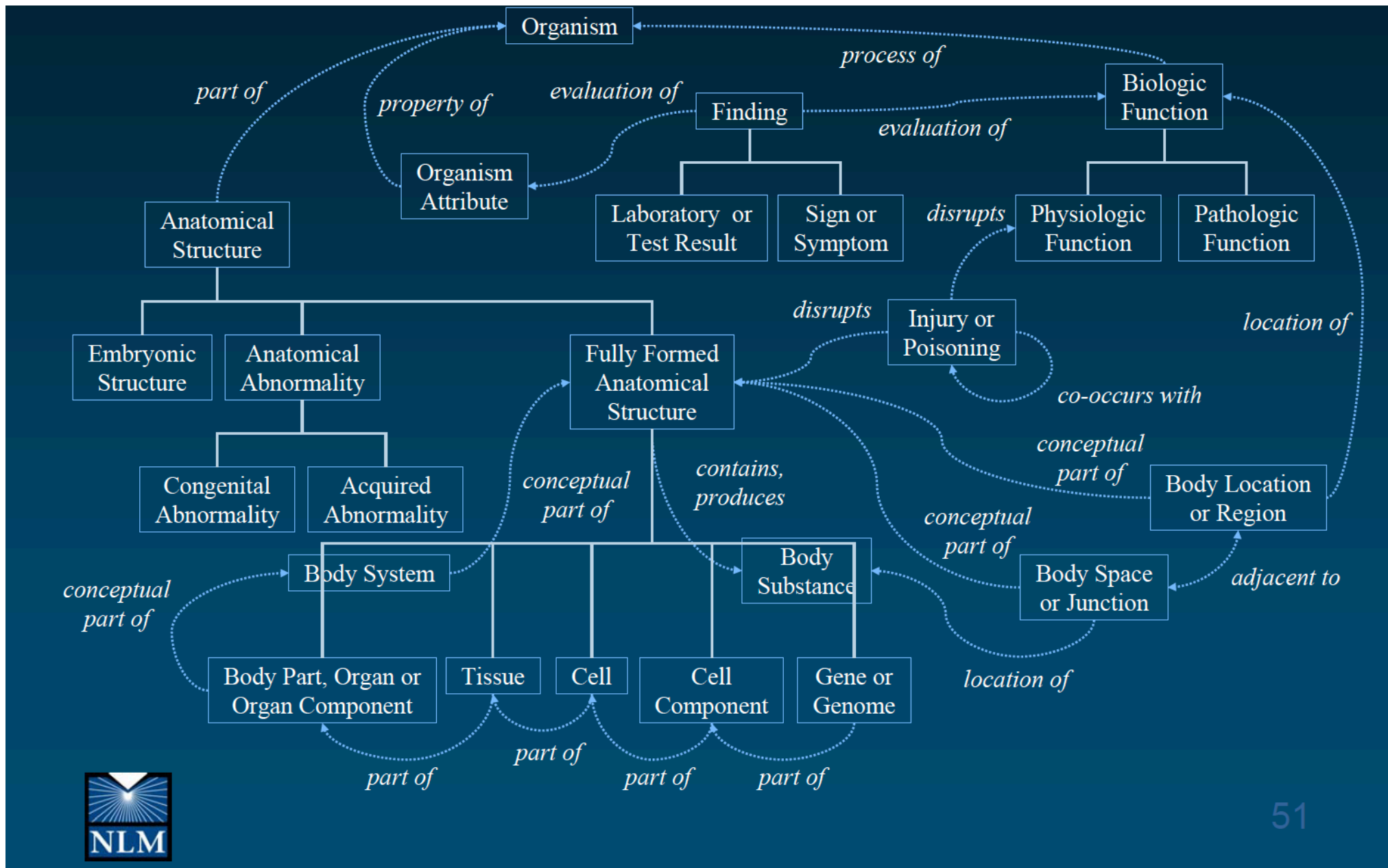
EMR clinical text

A 58-year-old women presented to her primary care physician after several days of dizziness, anorexia, dry mouth, increased thirst, and frequent urination. She reported no pain in her abdomen, back, or flank and no cough, shortness of breath, diarrhea, or dysuria. Her history was notable for cutaneous lupus, hyperlipidemia, osteoporosis, frequent urinary tract infections, three uncomplicated cesarean sections, a left oophorectomy for a benign cyst, and primary hypothyroidism, which has been diagnosed a year earlier. Her medications were levothyroxine, hydroxychloroquine, pravastatin, and alendronate. She had a 20-pack-year of smoking but had quit 3 weeks before presentation.

EMR clinical text: named entity recognition

Modifiers **Symptoms** **Diseases** **Medications** **Temporal**

A 58-year-old women presented to her primary care physician after several days of dizziness, anorexia, dry mouth, increased thirst, and frequent urination. She reported no pain in her abdomen, back, or flank and no cough, shortness of breath, diarrhea, or dysuria. Her history was notable for cutaneous lupus, hyperlipidemia, osteoporosis, frequent urinary tract infections, three uncomplicated cesarean sections, a left oophorectomy for a benign cyst, and primary hypothyroidism, which has been diagnosed a year earlier. Her medications were levothyroxine, hydroxychloroquine, pravastatin, and alendronate. She had a 20-pack-year of smoking but had quit 3 weeks before presentation.

Unified Medical Language System® (UMLS®)

TENTH REVISION
ICD-10
INTERNATIONAL CLASSIFICATION OF DISEASES
ICD-10 is a new code set for reporting medical diagnoses & inpatient procedures.

53

Sentiment analysis for clinical text

- **Purpose:** To evaluate treatment outcomes or judge the impact of a medical condition on patients.

- **Methodology:** Determining the information of patient's health status noted in clinical text to be positive or negative

- **Challenges**

   1. Less sentiment words

   2. Implicit sentiments: implicit description of health status, critical symptoms

   3. Negation diversity

   4. Short text
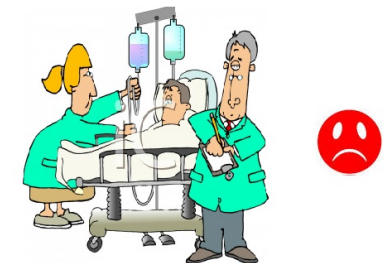
Denecke, Kerstin, and Yihan Deng. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine* 64.1, 17-27, 2015.

Sentiment analysis for clinical text

Problems

- **Polarity classification**

Sentence in clinical note	Sentiment label
There has significant improvement in pleural effusion.	Positive
There is moderate cardiomegaly.	Negative

Challenges: Negation diversity, shortness of text

- **Aspect-based classification**

Aspect: Lung

Sentence in clinical note	Sentiment label
There has significant improvement in pleural effusion.	Positive
There is moderate cardiomegaly.	Negative

Aspect: Heart

Dang, Tran-Thai, and Tu-Bao Ho. "Mixture of Language Models Utilization in Score-Based Sentiment Classification on Clinical Narratives." *IEA-AIE* 2016.

Two problems under investigation

- Recommendation of treatment regimens for patients based on the past treatment data from EMRs.

Case study

	Treatment 1	Treatment 2
Effective level 1	n11	n12
Effective level 2	n21	n22
Effective level 3	n31	n32

- **Construction of contingency tables from EMRs for evaluating the effectiveness of drug utilization.**

- **Detection and prediction of adverse drug reaction when using multiple drugs.**

Drug utilization research

Factors influencing drug utilization	Prescribing, dispensing and consumption of drugs	Outcomes of drug therapy
Patient and provider characteristics, disease patterns, marketing, regulations and reimbursement, etc.		Benefits and risks, e.g. mortality, morbidity, hospitalizations, quality of life

DRUG UTILIZATION RESEARCH

The principal aim of drug utilization research is to assess whether drug therapy is rational or not.

PHARMACO-EPIDEMIOLOGY   DRUG UTILIZATION RESEARCH   HEALTH SERVICES RESEARCH
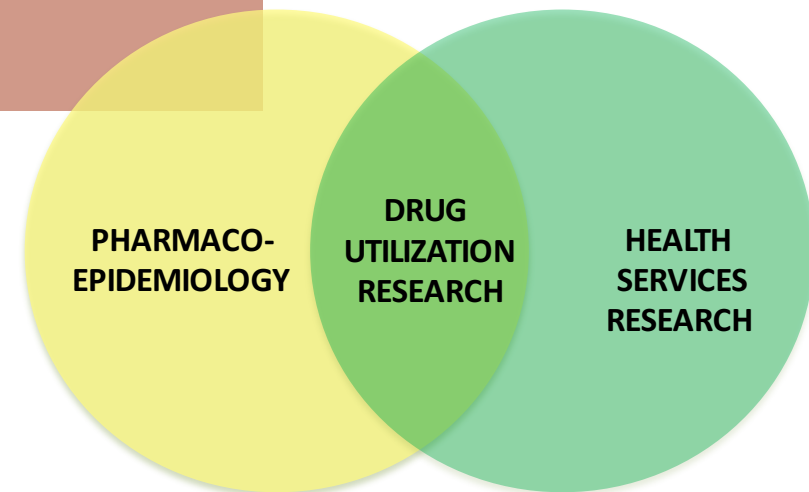
WHO. Introduction to Drug Utilization Research. 2003.

Post-market surveillance of drug

SAFETY   DOSE, DOSE FORM   EFFICACY, ADR   EFFECTIVENESS

100 strong men   100-500 patients   1000-5000 patients

0 1 2 3 4 5 6 7 8 Year

Investigational New Drug Application
Clinical trial Phase 1
Clinical trial Phase 2
Clinical trial Phase 3
Regulatory Review
Full Availability
Post-approval Studies
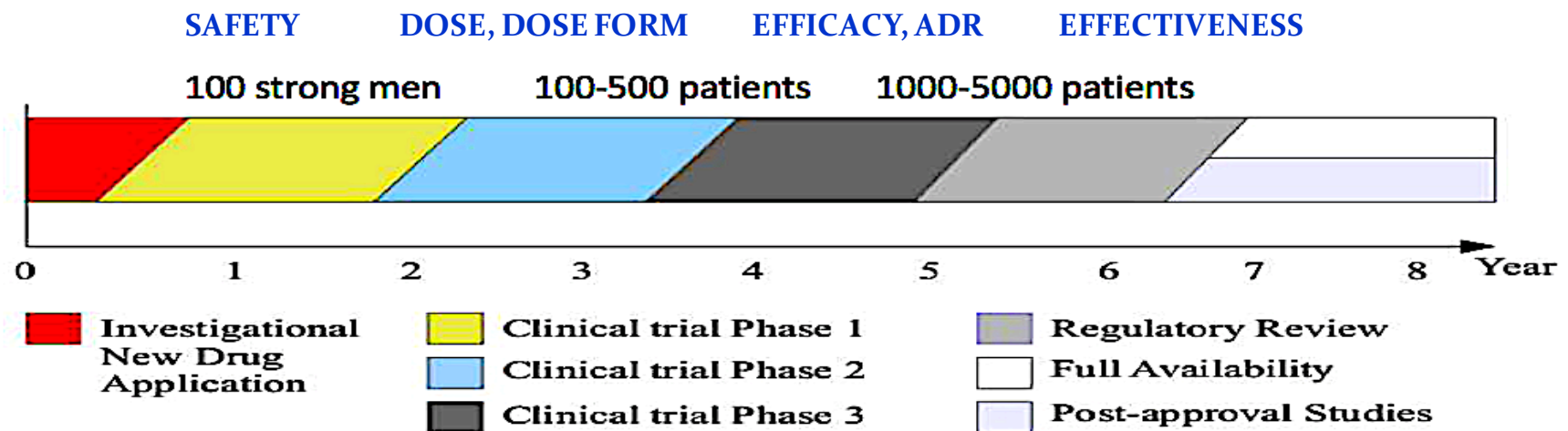
Traditional approval process

| Clinical study | Phased clinical trials (confirmation of efficacy and safety) | Marketing authorization | Marketing |
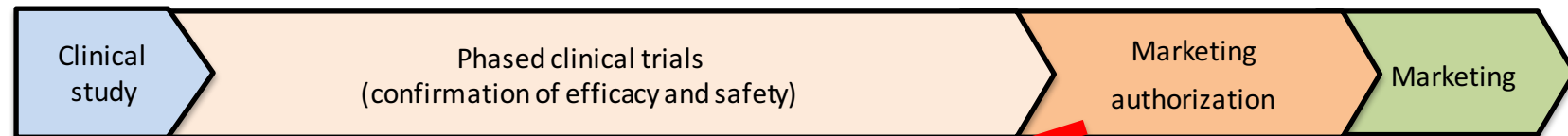
New scheme for regenerative product

| Clinical study | Clinical trials (likely to predict efficacy, confirming safety) | Conditional /term-limited authorization | Marketing (Further confirmation of efficacy & safety) | Marketing Authorization or Revocation | Marketing continuous |

Post-marketing safety measures must be taken, including prior informed consent of risk to patients

58
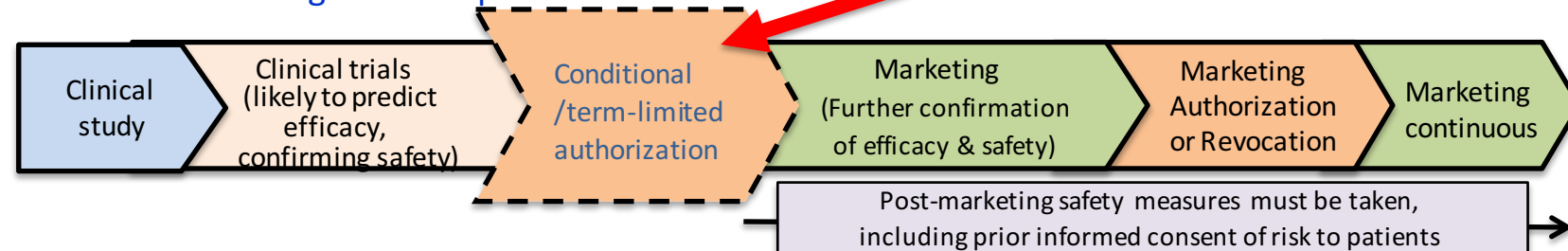
Drug safety and effectiveness

- Support for **presciption**: over or under dose, precision, risk, targeted drugs, replacement drugs...

- Detection and prediction of **adverse drug reaction**

- Pragmatic evaluation of **drug utilization effectiveness**, relation between treatment factors and patient groups: drug groups, dose, time, utilization methods...

- EMRs are golden data to understand drug utilization in the **post-marketing**.

Ho, T.B., Le, L., Dang, T.T., Siriwon, T. Data-driven Approach to Detect and Predict Adverse Drug Reactions, *Current Pharmaceutical Design Journal*, Vol. 22, No. 23, 3498-3526, 2016.

Methods for drug utilization research

Descriptive (qualitative) methods

Analytical (quantitative) methods

M. Elseviers et al. (Eds.). Drug Utilization Research: Methods and Applications, 2016.
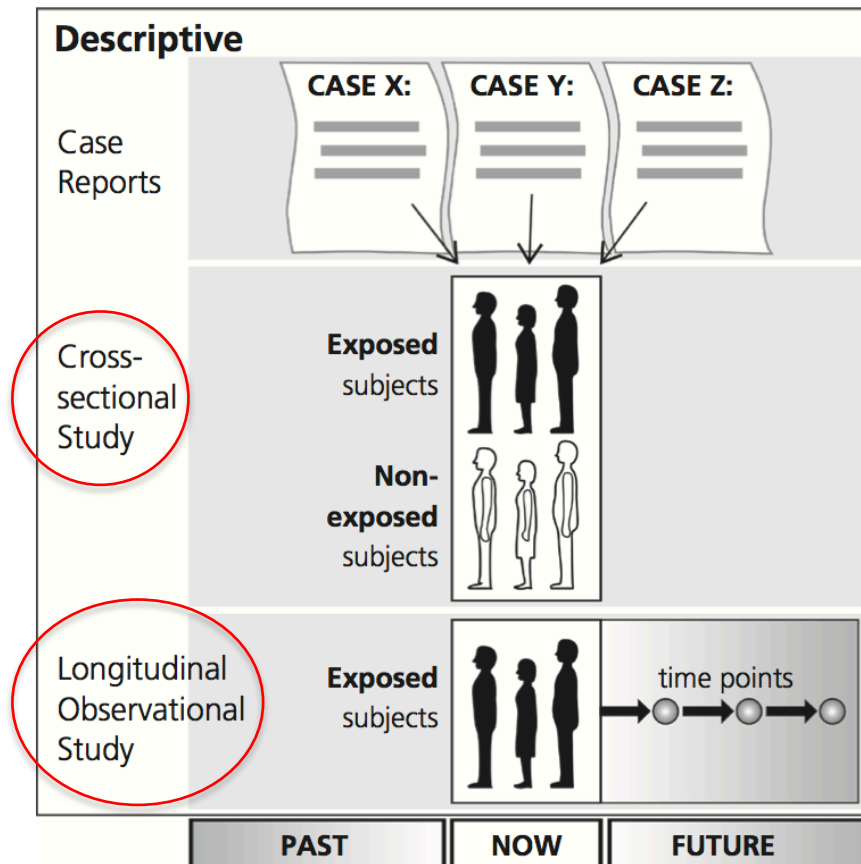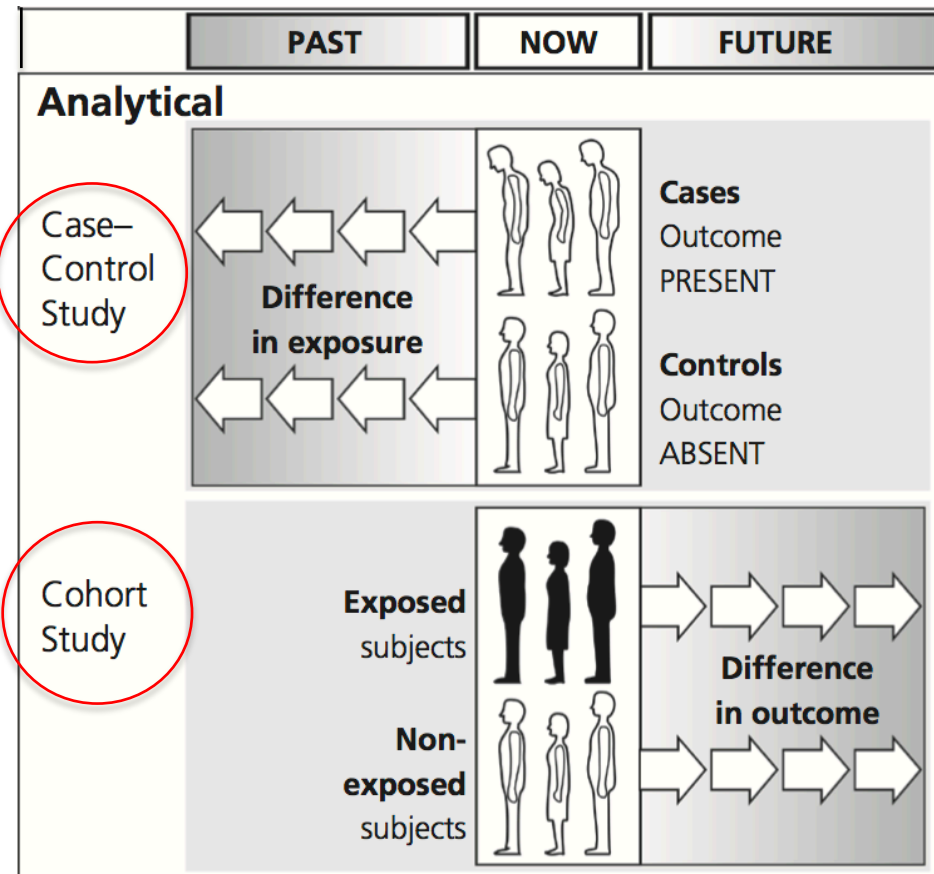
Methods for drug utilization research

Descriptive (qualitative) methods
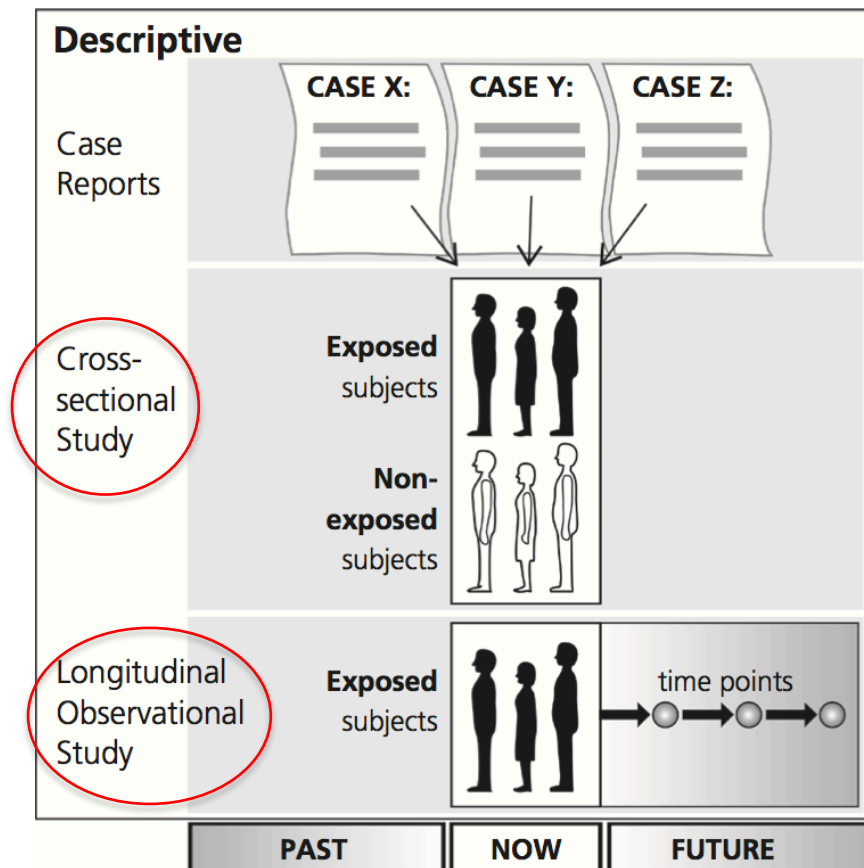
Identifying patterns or trends in drug utilization

- **Case Reports**
  present drug consumption in a single patient or the prescribing pattern at an individual clinic.

- **Cross-sectional Study**
  describe the utilization of drugs in a given population at a given point in time.

- **Longitudinal Observational Study**
  involve repeated observations of the same variables over time.

M. Elseviers et al. (Eds.). Drug Utilization Research: Methods and Applications, 2016.

Methods for drug utilization research

Deeper understanding of the explanatory factors behind utilization patterns or the effectiveness or safety of medication use.

- **Case-Control Study**
  On having the outcome of interest or not. Previous exposure of interest be compared retrospectively between those with and without the study outcome.

- **Cohort Study**
  Exposed and non-exposed subjects are followed over a period of time to investigate the exposure effects.

Analytical (quantitative) methods

M. Elseviers et al. (Eds.). Drug Utilization Research: Methods and Applications, 2016.
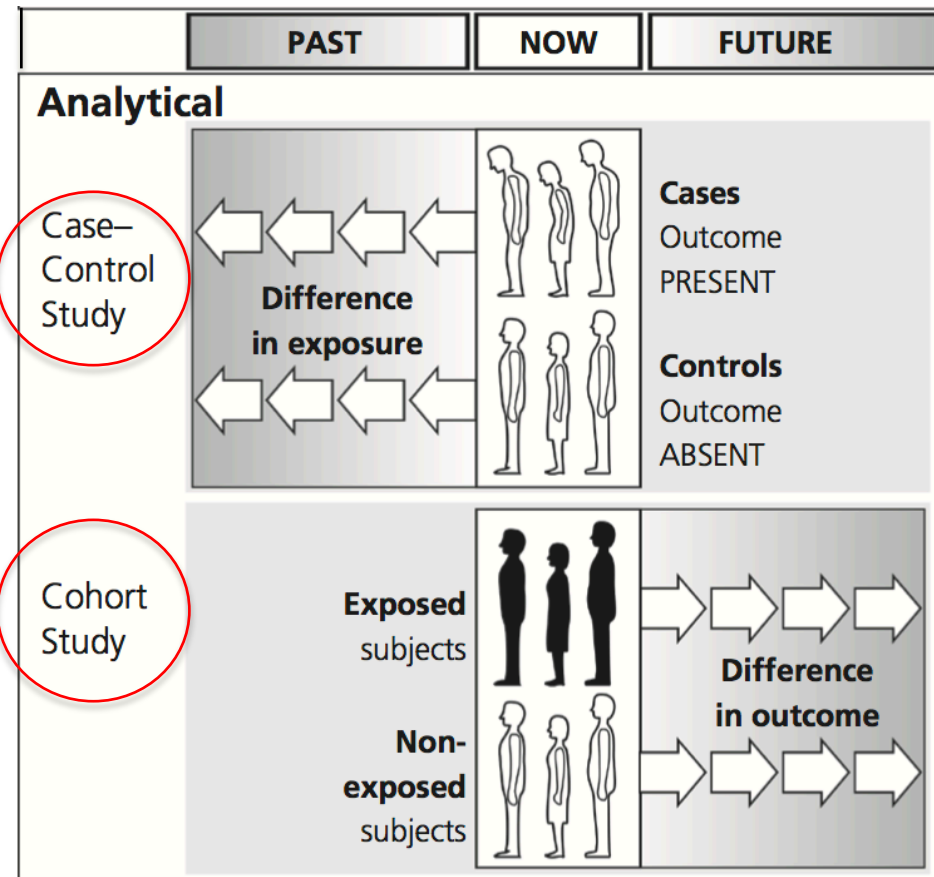
Efficacy vs. effectiveness

- **Efficacy** is the capacity to produce an effect. In medicine, it is the ability of an intervention or drug to produce a desired effect. It is tested by explanatory clinical trials under ideal and controlled circumstances.

- **Effectiveness** is the capability of producing a desired result. In medicine effectiveness relates to how well a treatment works in practice. It is tested by pragmatic clinical trials.

Ian Ford, Pragmatic trials, *The New England Journal of Medicine*, 454-463, 2016.

PICOT comparison of ECT and PCT

	Explanatory clinical trials (ECT)	Pragmatic clinical trials (PCT)
Population	Homogeneous patients	Real-life patients
Intervention	Tightly defined intervention	Flexible intervention with changes
Comparison	Clearly defined control group and often placebo	Active comparator instead of placebo
Outcome	Objective/surrogate outcomes	Clinically important outcomes
Time	Short-term follow-up time, e.g., 6 weeks	Long-term follow-up time, e.g., 6 months

Pragmatic clinical research is much more significant and needed than explanatory clinical trial but it is very difficult to carry out as requiring **real world conditions**.

J.P. A. Ioannidis. Why most clinical research is not useful. *PLOS Medicine,* 2016

Designs for drug utilization research

- **Data on drug use** has to be collected and analyzed by **statistical methods**.

- Typically, categorical data is often collected, represented in **contingency tables** (cross tables) and analyzed by two methods of Fisher's exact test (for the 2x2 contingency tables) and RxC chi-square test

(a) Typical table in drug safety evaluation

	Treatment Group	Comparison Group	Total
Test positive	a (TP)	b (FP)	a + b
Test negative	c (FN)	d (TN)	c + d
Total	a + c	b + d	

(b) Typical table in cross-sectional studies in descriptive research

	Exposes subjects	No-exposed subjects	Total
Patient group 1	a	b	a + b
Patient group 2	c	d	c + d
Total	a + c	b + d	

(c) Typical table in cohort study in analytical research

	Exposes subjects	No-exposed subjects	Total
Outcome 1	a	b	a + b
Outcome 2	c	d	c + d
Total	a + c	b + d	

(d) Large table in drug use effectiveness evaluation

	Male	Female	Total
Placebo	a	b	a + b
Dose 1	c	d	c + d
Dose 2	e	f	e + f
Dose 3	g	h	g + h
Total	a + c + e + g	b + d + f + h	

M. Elseviers et al. (Eds.). Drug Utilization Research: Methods and Applications, 2016.

PICOT comparison of ECT and PCT

	Explanatory clinical trials (ECT)	Pragmatic clinical trials (PCT)
Population	Homogeneous patients	Real-life patients
Intervention	Tightly defined intervention	Flexible intervention with changes
Comparison	Clearly defined control group and often placebo	Active comparator instead of placebo
Outcome	Objective/surrogate outcomes	Clinically important outcomes
Time	Short-term follow-up time, e.g., 6 weeks	Long-term follow-up time, e.g., 6 months

In the force of **pragmatism** in clinical research, an emerging approach is **electronic medical records pragmatic clinical research** (**EMRPCT**).

J.P. A. Ioannidis. Why most clinical research is not useful. *PLOS Medicine,* 2016

The framework

For each comment of a patient

- Compute the ratio $\frac{|p^+|}{|p^-|}$

- Do regression on the chain $\frac{|p^+|}{|p^-|}$

- Find the trend coefficient to classify drug effectiveness into outcomes

Original EMRs
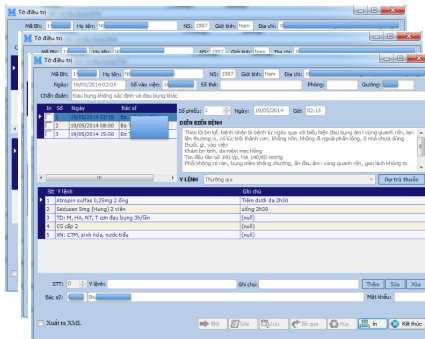
Cleaning & De-identification
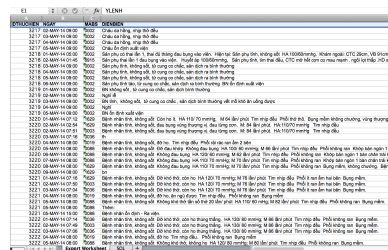
Secondary EMR DB

Patient groups

Drug associations

Evaluating the comments (opinion mining)

	Coversyl 5mg	Amidile G (Amlodipin 5mg)
Less effective (0, 0.01)	2	14
Effective (0.01, 0.55)	9	50
Strong Effective (0.55, 1)	2	1

Contingency tables

Clinical note evaluation

- **Bases on sentiment classification algorithm**

  - Phrases containing adjectives or adverbs which are good indicators of subjectivity and opinions are extracted if they conform some prior patterns.

  - The semantic orientation of extracted phrases is estimated using pointwise mutual information, which is defined as

  $$PMI(term_1, term_2) = \frac{log_2(P(term_1 \wedge term_2)}{P(term_1)P(term_2)}$$

  - The semantic orientation (SO) of a phrase:

  $$SO(phrase) = PMI(\ phrase, positive\ terms) - PMI\ (phrase, negative\ terms)$$

  - Given a comment, the algorithm computes the average SO of all phrases in that comment and classify it as positive or negative meaning.

- **Adapt the sentiment classification algorithm for Vietnamese**
- **Build prior patterns by looking through a part of EMRs**

EMRPCT for drug effectiveness study

Data: EMRs from Van Don hospital from 3.2013-21.2016. Extracted 10 most popular diseases

Objectives: To automatically extract contingency tables from EMRs on the effect of drug groups on patient groups

Disease and number of patients

ICD	Disease	Frequency
I10	Primary hypertension	690
K30	Degestive disorders	589
I99.1	Cerebral circulation disorder	435
J20	Acute bronchitis	354
H81	Vestibular dysfucntion	317
J18.0	Bronchopneumonia, unspecified	235
K29.1	Acute gastritis	233
B34	Virus infection in unknown location	164
N39.0	Urinary tract infections	152
T78.4	Allergy undetermined	148

FREQUENCY OF 10 POPULAR DISEASES AT VÂN ĐỒN HOSPITAL

Ho T.B., Hoang K.H., Dang T.T., Drug utilization research with pragmatic clinical trials using electronic medical records (to appear)
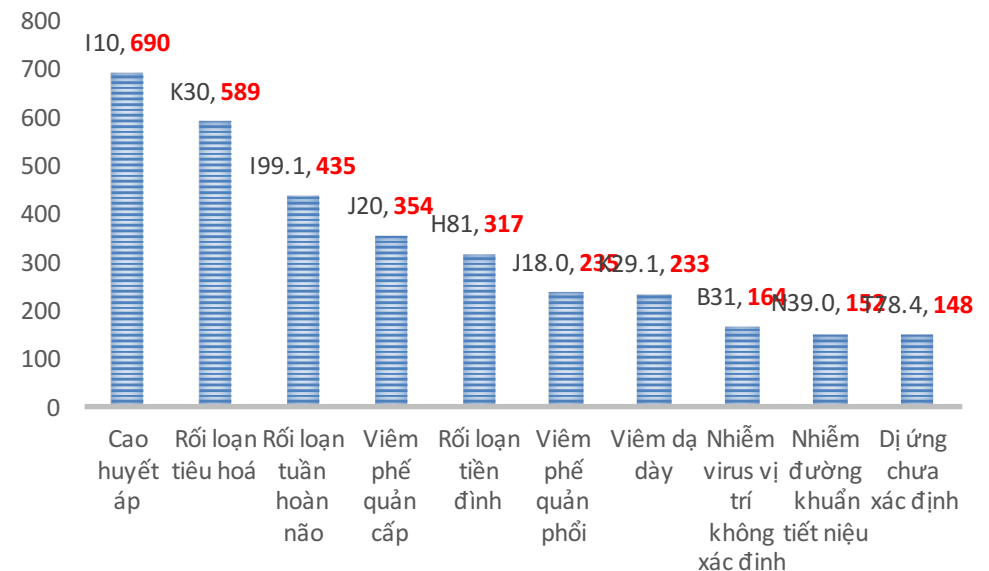
Drug used for hypertention patients

Drug	Code
Actelsar 40mg	TEL002
Adalat Cap 10	ADA001
Adalat	ADA001
Amidile G(Amlodipin 5mg)	AMI002
Amlodipin (Amlor 5mg)	AML002
Amlodipin 5mg (Ampori)	AML001
Amlodipin 5mg (Primodil)	AMLODIPI
Captopril 25mg (Khánh Hoà)	CAP001
Captopril stada 25mg	CAP003
CarvesyL-12.5mg	CAVESS
Coversyl 5mg	COV001
Dorodipin 10mg(Amlodipin)	DOR001
Enalapril (Renitec 5mg) Anh	ENA001
Furocemid	FUR001
Furosemide	FUR002
Idatril 5mg	IDA001
Lodimax 5mg	AML003
Lopilcar - A	LOP003
Nifedipin Stada 10 mg	NIF001
Nifedipine (uống)	NIF001
S-Lopilcar 2	
Vinzix	FUR001
coversyl	COV001

Drug	Frequency
Actelsar 40mg	10
Adalat Cap 10	13
Idatril 5mg	15
Dorodipin 10mg(Amlodipin)	16
Furocemid	17
Amlodipin (Amlor 5mg)	18
Lodimax 5mg	36
Coversyl 5mg	41
Furosemide	43
Amlodipin 5mg (Ampori)	45
Amlodipin 5mg (Primodil)	62
Amidile G(Amlodipin 5mg)	90
Furosemide, Coversyl 5mg	7
Furosemide, Amlodipin 5mg	7
Coversyl 5mg, Amlodipin	7
Captopril 25mg	9
Coversyl 5mg, Amidile G(Amlodipin 5mg)	9
Amlodipin 5mg (Primodil), Furosemide	9

Hypertension drug effectiveness

	Coversyl 5mg	Amidile G (Amlodipin 5mg)
Less effective (0, 0.01)	2	14
Effective (0.01, 0.55)	9	50
Strong Effective (0.55, 1)	2	1

- Compare Coversyl 5mg vs Amidile G(Amlodipin 5mg) using EMRs of hypertension patients who took only those drugs for the study.

$$H_0: p_{Coversyl\_lessEff} = p_{AmildileG\_lessEff};$$
$$p_{Coversyl\_Eff} = p_{AmildileG\_Eff}$$
$$p_{Coversyl\_strongEff} = p_{AmildileG\_strongEff}$$

	Less effective	Effective	Strong effective	Sample size
Coversyl 5mg	2 (2.67)	9 (9.83)	2 (0.5)	13
Amidile G(Amlodipin 5mg)	14 (13.33)	50(49.17)	1(2.5)	65
Total	16	59	3	78

- With d.f. = 2, the tabulet upper 5% point of χ^2 is 5.99, the null hypothesis is not rejected.

The values of $(O - E)^2/E$

	Less effective	Effective	Strong effective
Coversyl 5mg	0.168	0.07	4.5
Amidile G(Amlodipin 5mg)	0.034	0.014	0.9
Total	$\chi^2 = 5.686$		
	$d.f$ of $\chi^2 = (3-1)(2-1) = 2$		

- No difference in the patients treated by Coversyl 5mg and Amidile G (Amlodipin 5mg).

Ho T.B., Hoang K.H., Dang T.T., Drug utilization research with pragmatic clinical trials using electronic medical records, PAKDD 2018

Hypertension drug effectiveness

	Lodimax 5mg	Amidile G (Amlodipin 5mg)
Less effective (0, 0.01)	2	14
Effective (0.01, 0.55)	20	50
Strong Effective (0.55, 1)	2	1

	Less effective	Effective	Strongly effective	Sample size
Lodimax 5mg	2 (4.315)	20 (18.876)	2 (0.809)	24
Amidile G(Amlodipin 5mg)	14 (11.685)	50(51.124)	1(2.191)	65
Total	16	70	3	89

The values of $(O - E)^2/E$

	Less effective	Effective	Strong effective
Coversyl 5mg	1.242	0.067	1.753
Amidile G(Amlodipin 5mg)	0.459	0.025	0.647
Total	$\chi^2 = 4.193$		
	$d.f \text{ of } \chi^2 = (3 - 1)(2 - 1) = 2$		

- Compare Lodimax 5mg vs Amidile G (Amlodipin 5mg) using EMRs of hypertension patients who took only those drugs for the study.

H_0: $p_{Lodimax\_lessEff} = p_{AmildileG\_lessEff}$;
$p_{Lodimax\_Eff} = p_{AmildileG\_Eff}$
$p_{Lodimax\_strongEff} = p_{AmildileG\_strongEff}$

- With d.f. = 2, the tabulet upper 5% point of χ^2 is 5.99, the null hypothesis is not rejected.

- No difference in the proportion of patients treated by Lodimax 5mg and Amidile G (Amlodipin 5mg)

Acute bronchitis drug effectiveness

	Acetyl cystein - (ESOMEZ 200mg)	Acetyl cystein (Andomuc 200mg)
Less effective (0, 0.01)	7	7
Effective (0.01, 0.55)	20	26
Strong Effective (0.55, 1)	1	2

	Less effective	Effective	Strongly effective	Sample size
ESOMEZ 200mg	7 (6.222)	20 (20.444)	1 (1.333)	28
Andomuc 200mg	7 (7.778)	26(25.556)	2(1.667)	35
Total	14	46	3	63

The values of $(O - E)^2/E$

	Less effective	Effective	Strong effective
ESOMEZ 200mg	0.097	0.01	0.083
Andomuc 200mg	0.078	0.008	0.0665
Total	$\chi^2 = 0.3425$ $d.f \ of \ \chi^2 = (3-1)(2-1) = 2$		

- Compare ESOMEZ 200mg and Andomuc 200mg using EMRs of hypertension patients who took only those drugs for the study.

$$H_0: p_{Esomez\_lessEff} = p_{Andomuc\_lessEff};$$
$$p_{Esomez\_Eff} = p_{Andomuc\_Eff}$$
$$p_{Esomez\_strongEff} = p_{Andomuc\_strongEff}$$

- With d.f. = 2, the tabulet upper 5% point of χ^2 is 5.99, the null hypothesis is not rejected.

- No difference in the patients treated by ESOMEZ 200mg and Andomuc 200mg.

Drug used for acute bronchitis

Drug	Code
Acetyl cystein (Andomuc 200mg)	AND001
Acetyl cystein -(ESOMEZ 200mg)	ACE003
Cotrimoxazole	COT001
Novahexin 5 ml	NOV001
Penicilin V Kali	PEN004
Pms-Opxil 500mg	CEP011
Terpincodein (Amucopect)	TER004
CefacLor 125mg Domesco	CEF021
Cefuroxine 750mg	CEF020
Cephalexin 0.5g (medofalexi LD/MDP)	CEP008
Diaphylin 4.8% Hung	DIA0012
Diaphylin 4.8%	DIA
Drenoxol	AMB002
Pulmicort	PUL001
Gentamicin	GEN002
Gentamicin	GEN001
Medoclor 250mg	CEF003
Medotase 10mg(an)	MED012
Salbutamol 4mg	SAL
Salbutamol	SAL
Sultasin 0.75g	SUL002
Trichopol	MET006
Ventolin 2.5mg	VEN002
Ventolin Neb Sol 2.5mg/2.5ml 6x5's	VEN001

Drug	Frequency
Gentamicin	3
Diaphylin 4.8%	3
Cephalexin 0.5g (medofalexi LD/MDP)	3
Sultasin 0.75g	3
Diaphylin 4.8%, Acetyl cystein (Andomuc 200mg)	3
Terpincodein (Amucopect),Acetyl cystein (Andomuc 200mg)	3
Acetyl cystein -(ESOMEZ 200mg), Sultasin 0.75g	3
Acetyl cystein (Andomuc 200mg), Salbutamol 4mg	3
Terpincodein (Amucopect), Acetyl cystein -(ESOMEZ 200mg)	4
Salbutamol 4mg	5
Drenoxol, Ventolin Neb Sol 2.5mg/2.5ml 6x5's	5
Ventolin Neb Sol 2.5mg/2.5ml 6x5's, Pulmicort	7
Pulmicort	9
Ventolin Neb Sol 2.5mg/2.5ml 6x5's	12
Gentamicin	12
Drenoxol	13
Terpincodein (Amucopect)	18
Acetyl cystein -(ESOMEZ 200mg)	36
Acetyl cystein (Andomuc 200mg)	45

Hypertension drug effectiveness

	Lodimax 5mg	Amidile G (Amlodipin 5mg)
Less effective (0, 0.01)	2	14
Effective (0.01, 0.55)	20	50
Strong Effective (0.55, 1)	2	1

	Less effective	Effective	Strongly effective	Sample size
Lodimax 5mg	2 (4.315)	20 (18.876)	2 (0.809)	24
Amidile G(Amlodipin 5mg)	14 (11.685)	50(51.124)	1(2.191)	65
Total	16	70	3	89

The values of $(O - E)^2/E$

	Less effective	Effective	Strong effective
Coversyl 5mg	1.242	0.067	1.753
Amidile G(Amlodipin 5mg)	0.459	0.025	0.647
Total	$\chi^2 = 4.193$ $d.f \ of \ \chi^2 = (3-1)(2-1) = 2$		

- Compare Lodimax 5mg vs Amidile G(Amlodipin 5mg) using EMRs of hypertension patients who took only those drugs for the study.

$H_0: p_{Lodimax\_lessEff} = p_{AmildileG\_lessEff};$
$p_{Lodimax\_Eff} = p_{AmildileG\_Eff}$
$p_{Lodimax\_strongEff} = p_{AmildileG\_strongEff}$

- With d.f. = 2, the tabulet upper 5% point of χ^2 is 5.99, the null hypothesis is not rejected.

- No difference in the proportion of patients treated by Clodimax 5mg and Amidile G (Amlodipin 5mg)

NỘI DUNG 5: Xây dựng công cụ phân tích quan hệ bệnh-thuốc

STT	Hoạt động	Thời gian	Người phụ trách thực hiện
5.1	Tạo bảng tần suất từ BAĐT về hiệu quả dùng các loại thuốc (*)	4/2016 – 9/2016	GS Hồ Tú Bảo, PGS Lê Thị Lý
5.2	Phân tích ý kiến trong BAĐT về đánh giá hiệu quả dùng thuốc (*)	4/2016 – 3/2017	GS Hồ Tú Bảo, PGS Lê Thị Lý
5.3	Phát hiện hiệu ứng phụ của thuốc (*)	4/2016 – 3/2017	PGS Lê Thị Lý, GS Hồ Tú Bảo
5.4	Tìm khả năng mới của thuốc đã có (*)	4/2016 – 3/2017	PGS Lê Thị Lý, GS Hồ Tú Bảo

Chỉ tiêu đánh giá

Các phương pháp và bước đầu nghiên cứu quan hệ giữa bệnh và thuốc từ BAĐT:
- Tạo bảng tần suất từ BAĐT về hiệu quả dùng các loại thuốc.
- Phân tích ý kiến trong BAĐT về đánh giá hiệu qủa dùng thuốc.
- Phát hiện hiệu ứng phụ của thuốc.
- Tìm khả năng sử dụng mới của thuốc đã có.

NỘI DUNG 5: Xây dựng công cụ phân tích quan hệ bệnh-thuốc

5.1 Tạo bảng tần xuất về hiệu quả dùng thuốc từ BAĐT

	Coversyl 5mg	Amidile G (Amlodipin 5mg)
Less effective (0, 0.01)	2	14
Effective (0.01, 0.55)	9	50
Strong Effective (0.55, 1)	2	1

	Less effective	Effective	Strong effective	Sample size
Coversyl 5mg	2 (2.67)	9 (9.83)	2 (0.5)	13
Amidile G(Amlodipin 5mg)	14 (13.33)	50(49.17)	1(2.5)	65
Total	16	59	3	78

Efficacy → **Effectiveness**

Mở ra một hướng nghiên cứu mới về nghiên cứu lâm sàng thực chứng: electronic medical records pragmatic clinical research (EMRPCT)

Ho T.B., Hoang K.H., Dang T.T., Drug utilization research with pragmatic clinical trials using electronic medical records (under preparation)

NỘI DUNG 5: Xây dựng công cụ phân tích quan hệ bệnh-thuốc

5.2 Phân tích ý kiến trong đánh giá hiệu quả dùng thuốc từ BAĐT

Phần lớn các văn bản lâm sàng là ý kiến của bác sĩ và điều dưỡng về bệnh nhân

| Sentence in clinical note	Sentiment label
There has significant improvement in pleural effusion.	Positive
There is moderate cardiomegaly.	Negative

Aspect: Lung

Aspect: Heart

Method	P (%)	R(%)	F(%)	R_{new} (%)
Reverb	53.19	5.12	9.35	2.38
PMI without constraints	49.45	73.16	59.01	74.6
PMI + Constraint 1 ()	54.27	46.93	50.33	45.24
PMI + Constraint 2 ()	51.05	64.95	57.17	67.85
PMI + Constraint 3 (K=40)	52.26	56.97	54.51	59.92

- Dang, Tran-Thai, and Tu-Bao Ho. "Mixture of Language Models Utilization in Score-Based Sentiment Classification on Clinical Narratives." IEA-AIE 2016.
- Đặng Trần Thái, Mixture of language models utilization in score-based sentiment classification on clinical narratives. Master Thesis
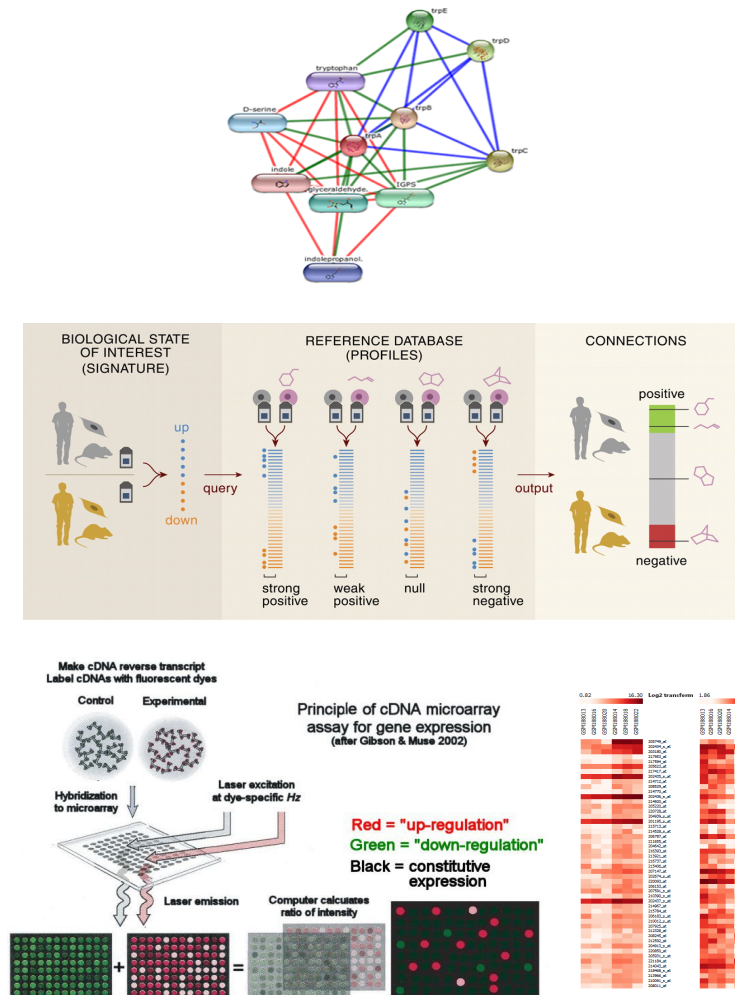
78

NỘI DUNG 5: Xây dựng công cụ phân tích quan hệ bệnh-thuốc

5.3 Phát hiện hiệu ứng phụ của thuốc
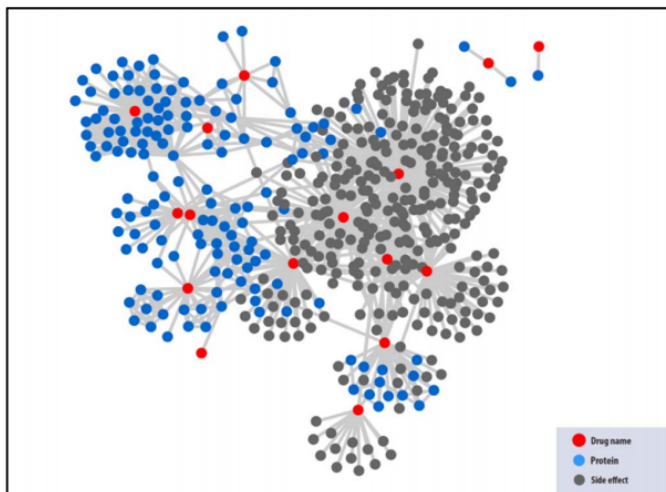
Hệ thống mạng kết nối

Pham D., Le B.K., L. Ly, Ho T.B., System pharmacology: application of network theory in predicting potential adverse drug reaction based on gene expression data, IEEE Inter. Conf. RIVF 2016, November 7-9, 2016.

NỘI DUNG 5: Xây dựng công cụ phân tích quan hệ bệnh-thuốc

5.3 Phát hiện hiệu ứng phụ của thuốc từ BAĐT

$$P(d_t^{a_i}|o_{t-1}^{a_i}) = \frac{c(o_{t-1}^{a_i}, d_t^{a_i})}{c(o_{t-1}^{a_i})}$$

$$P(o_k^{a_i}|o_{k-1}^{a_i}) = \frac{c(o_{k-1}^{a_i}, o_k^{a_i})}{c(o_{k-1}^{a_i})}$$

$$P(d_t^{a_i}|o_{t-1}^{a_i} o_{t-2}^{a_i}) = \frac{c(o_{t-2}^{a_i}, o_{t-1}^{a_i}, d_t^{a_i})}{c(o_{t-2}^{a_i}, o_{t-1}^{a_i})}$$

$$P(o_k^{a_i}|o_{k-1}^{a_i} o_{k-2}^{a_i}) = \frac{c(o_{k-2}^{a_i}, o_{k-1}^{a_i}, o_k^{a_i})}{c(o_{k-2}^{a_i}, o_{k-1}^{a_i})}$$


| Method	$Prec_5$	$Prec_{10}$	$Prec_{15}$	$Prec_{20}$	$Prec_{25}$	$Prec_{30}$
RR	0.331	0.33	0.33	0.337	0.333	0.339
$conf$	0.403	0.375	0.386	0.387	0.389	0.39
lev	0.373	0.337	0.343	0.343	0.339	0.335
χ^2 test	0.373	0.346	0.356	0.367	0.369	0.363
Sequence-based measure	**0.437**	**0.447**	**0.439**	**0.439**	**0.433**	**0.427**
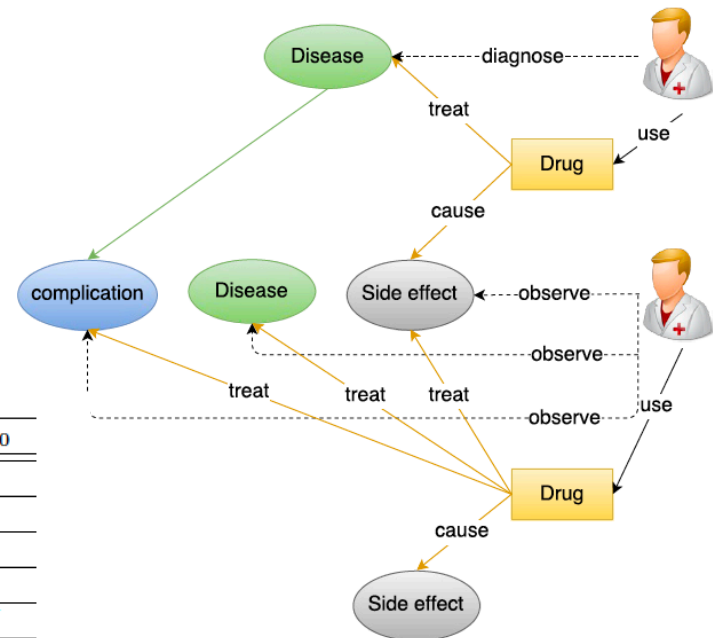
- Ho T.B., L. Ly, Dang T.T., S. Taewijit. Data-driven Approach to Detect and Predict Adverse Drug Reactions, Current Pharmaceutical Design Journal, Vol. 22, No. 23 (May 2016), 3498-3526 (SCI).

- Dang T.T., Ho T.B., Sequence-Based Measure for Assessing Drug-Side Effect Causal Relation from Electronic Medical Records, Inter. Symp. on Knowledge and Systems Sciences (KSS), Nov 2017.

NỘI DUNG 5: Xây dựng công cụ phân tích quan hệ bệnh-thuốc

5.4 Tìm khả năng sử dụng mới của thuốc đã có
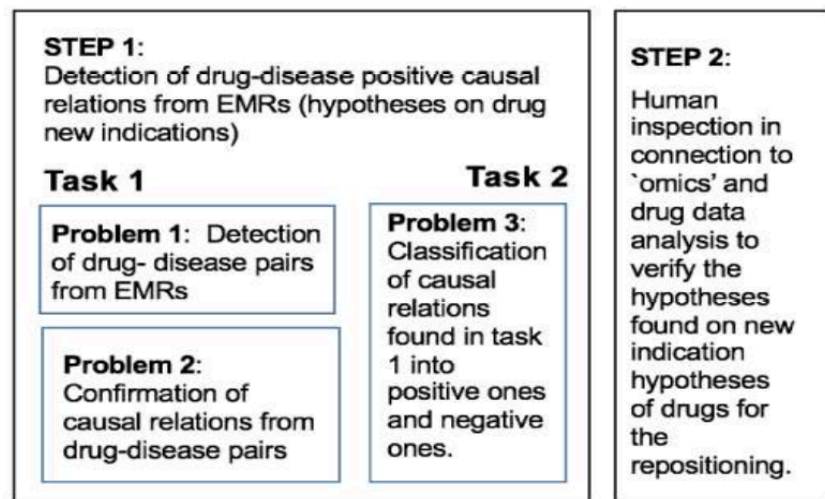
ĐỀ XUẤT QUÁ TRÌNH TÌM HIỆU ỨNG MỚI CỦA THUỐC
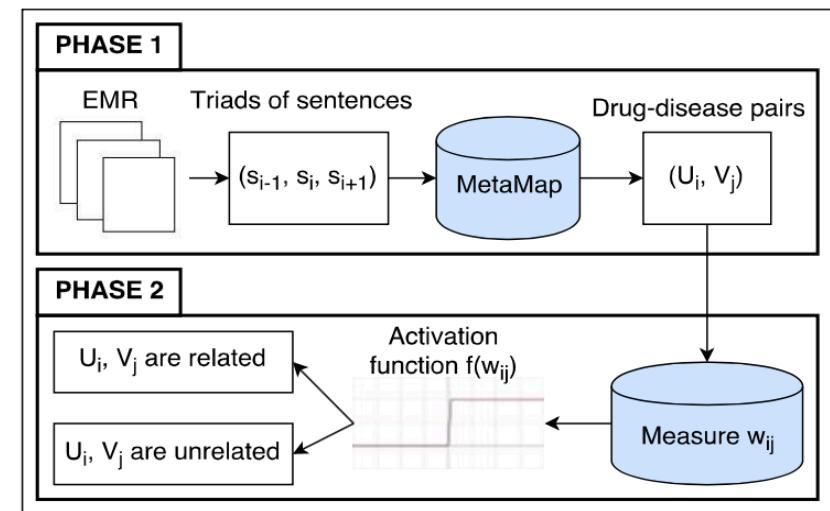
ĐỀ XUẤT LƯỢC ĐỒ GIẢI HAI BÀI TOÁN CỦA NHIỆM VỤ 1

- Dang, T.T., Ouankhamchan, P., Ho, T.B., Detection of New Drug Indications from Electronic Medical Records, *IEEE Inter. Conf.* RIVF 2016, November 7-9, 2016.
- Luận văn thạc sĩ của P. Ouankhamchan.

Two problems under investigation

- **Recommendation of treatment regimens for patients based on the past treatment data from EMRs**

- Construction of contingency tables from EMRs for evaluating the effectiveness of drug utilization.

- Detection and prediction of adverse drug reaction when using multiple drugs

Support for diagnosis and treatment
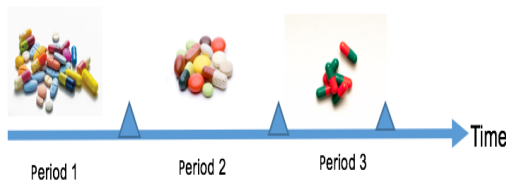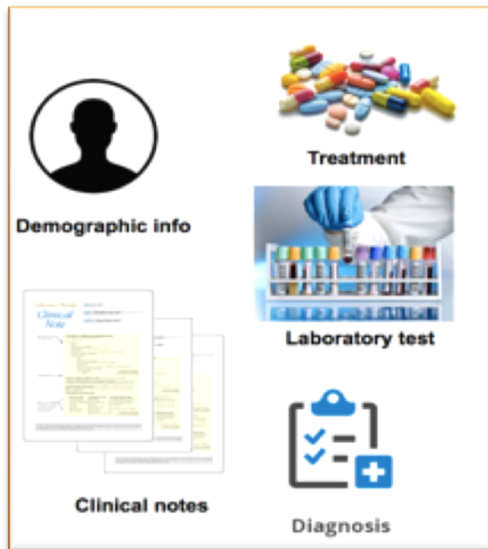
- Compare and evaluate treatment regiments
- Finding appropriate treatment regiments
- Finding similar patients for rare and difficult cases
- etc.

- Support for interpreting lab examination data (anomaly, disorders...
- Compare and visualize the index relations in treatement
- Suveillance of the treatment process
- etc.

LIS
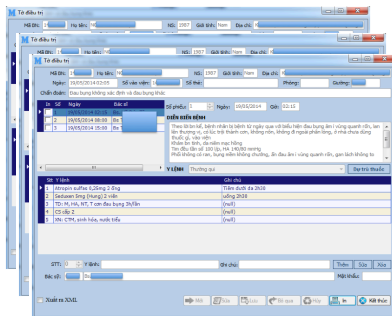
RIS

Learning treatment regimens
Method

EMR with many data types

Encoding each data type

Fit them into Restricted Boltzmann Machine

hidden variables

h

W

Bipartite Structure

Image   **v**   visible variables

Output EMRs as hidden binary vectors

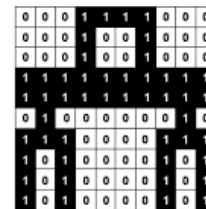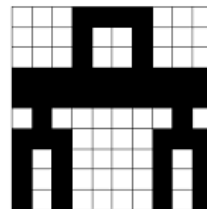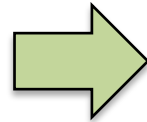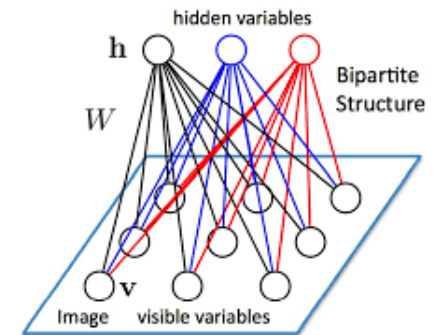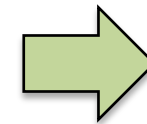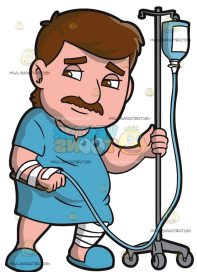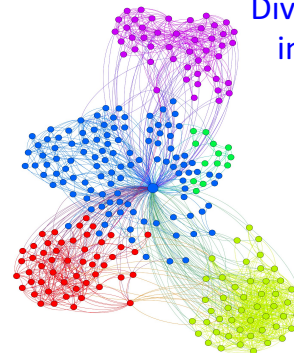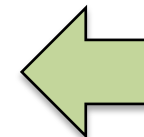Use group regimen to recommend the patient
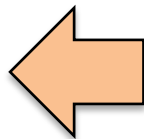
Learn treatment regimen for each group
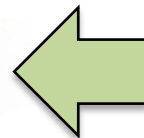
Divide patients into similar groups

Learning treatment regimens
Heterogeneous data

- Disease: coronary artery with no shared commoribity score

- Number of patients : 707

Prescirption table

subject_id	hadm_id	startdate	enddate	drug_type	drug	formulary_drug_cd	dose_val_rx	dose_unit_rx	route
13	143045	1/8/67 0:00	1/9/67 0:00	MAIN	Metoprolol	METO50	50	mg	PO
13	143045	1/8/67 0:00	1/9/67 0:00	MAIN	Atorvastatin	ATOR20	20	mg	PO
13	143045	1/8/67 0:00	1/9/67 0:00	MAIN	Captopril	CAPT25	25	mg	PO
13	143045	1/8/67 0:00	1/9/67 0:00	MAIN	Pantoprazole	PANT40	40	mg	PO
13	143045	1/8/67 0:00	1/9/67 0:00	MAIN	Acetaminophen	ACET325	325-650	mg	PO
13	143045	1/8/67 0:00	1/12/67 0:00	MAIN	Zolpidem Tartrate	AMBI5	5	mg	PO
13	143045	1/8/67 0:00	1/9/67 0:00	MAIN	Docusate Sodium	DOCU100	100	mg	PO
13	143045	1/8/67 0:00	1/9/67 0:00	MAIN	Insulin	INSULIN	0	UNIT	SC
13	143045	1/9/67 0:00	1/9/67 0:00	MAIN	Potassium Chloride	MICROK10	40	mEq	PO

Numerical data

itemid	charttime	cgid	value	valuenum	valueuom
52	2167/1/10 8:30	20670		84	84 mmHg
55	2167/1/10 8:30	20670		102	102 mmHg
59	2167/1/10 8:30	20670		102	102 mmHg
62	2167/1/10 8:30	20670		51	51 mmHg
113	2167/1/10 8:30	20670		8	8 mmHg
128	2167/1/10 8:30	20670	Full Code		
153	2167/1/10 8:30	20670		12	12 %
161	2167/1/10 8:30	20670	None		
52	2167/1/10 9:00	20670		70	70 mmHg
55	2167/1/10 9:00	20670		109	109 mmHg
59	2167/1/10 9:00	20670		106	106 mmHg

Text data

History of Present Illness:
This is a 71 year old male with known CAD. He underwent PTCA to
LAD and diagonal in [**3467**]. Prior to hernia repair operation, an
ETT in [**3476-6-16**] was notable for EKG changes. An ECHO in [**Month (only) 202**]
[**3475**] was notable for mild MR [**First Name (Titles) **] [**Last Name (Titles) 1
estimated at 1.1 cm2 with peak/mean gradients of 34 and 22 mmHg.
The was mild concentric LVH with an LVEF of 60%. He was
subsequently referred for cardiac catheterization. This was
performed at the [**Hospital1 18**] on [**3476-7-6**]. Angiography showed a right
dominant system with 80% ostial LAD lesion; first diagonal had a
60% stenosis; the circumflex had a 60% lesion while the RCA had

Learning treatment regimens
Encoding vector & mixed-variate RBM

1652 mixed type features: 1477 binary &155 numerical features

Gender	Admission_Type0	Admission_Type1	Admission_Type2	Marital_Status-1	Marital_Status0	Marital_Status1	Marital_Status2	Marital_Status3	Marital_Status4	Marital_Status5	itemid_50	itemid_51	itemid_52	itemid_58	itemid_69
1	1	0	0	0	0	1	0	0	0	0	1.183957232	0.717319613	0.715479568	2.736176602	0.665098375
1	0	1	0	0	0	1	0	0	0	0	1.183957232	0.50409911	0.481893293	-0.285465976	0.635011189
1	0	1	0	0	0	1	0	0	0	0	1.183957232	0.450793984	0.533801354	-0.349303495	0.595149427
1	0	1	0	0	0	0	0	1	0	0	-0.842817013	0.966076867	0.923111812	0.289071697	0.850600724
1	0	1	0	0	0	1	0	0	0	0	1.183957232	0.006584603	0.222352988	-0.349303495	0.76572525
1	0	1	0	0	0	1	0	0	0	0	-0.842817013	1.108223869	1.052881965	-0.455699361	0.70729166
1	1	0	0	0	0	1	0	0	0	0	1.183957232	0.433025609	0.611663446	-0.455699361	0.689076437
1	1	0	0	0	0	1	0	0	0	0	-0.842817013	-1.414885417	-1.412750934	-0.455699361	-1.352222445
1	0	1	0	0	0	1	0	0	0	0	-0.842817013	-1.414885417	-1.412750934	-0.455699361	-1.352222445
0	1	0	0	0	0	1	0	0	0	0	1.183957232	1.072687118	1.026927934	-0.24290763	0.434582315
0	0	1	0	0	0	0	0	0	1	0	-0.842817013	-1.414885417	-1.412750934	-0.455699361	-1.352222445
1	0	1	0	0	0	1	0	0	0	0	1.183957232	1.55243325	1.753640788	-0.200349284	1.051823643
1	0	1	0	0	0	1	0	0	0	0	1.183957232	0.752856364	0.81929569	-0.349303495	0.876024744
1	0	1	0	0	0	1	0	0	0	0	-0.842817013	-1.414885417	-1.412750934	-0.455699361	-1.352222445
1	1	0	0	0	0	1	0	0	0	0	1.183957232	0.415257234	0.663571507	-0.24290763	0.933612825
1	1	0	0	0	0	1	0	0	0	0	1.183957232	0.912771741	0.845249721	-0.136511764	0.884078322
1	1	0	0	0	0	0	0	1	0	0	1.183957232	0.699551238	0.507847324	-0.455699361	0.720752078
0	0	1	0	0	0	0	0	0	0	1	-0.842817013	-1.414885417	-1.412750934	-0.455699361	-1.352222445
0	1	0	0	0	0	0	0	0	0	1	1.183957232	0.646246112	0.689525537	-0.349303495	0.306759929
1	0	1	0	0	1	0	0	0	0	0	-0.842817013	0.575172611	0.845249721	-0.455699361	0.953355451
1	1	0	0	0	0	1	0	0	0	0	-0.842817013	-1.414885417	-1.412750934	-0.455699361	-1.352222445
1	1	0	0	0	0	0	0	0	1	0	1.183957232	0.184268356	0.248307019	-0.349303495	0.493758972
1	0	1	0	0	0	1	0	0	0	0	-0.842817013	0.664014487	0.79334166	-0.24290763	0.715538955
1	0	1	0	0	0	1	0	0	0	0	-0.842817013	1.747885378	0.378077171	-0.455699361	1.01952704

Fit preprocessed data to mixed-variate RBM: hidden nodes: 200, # iterations: 10000

|																		
1	0	0	1	0	1	0	0	1	0	1	0	1	1	1	1	0	0	1
1	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	0	0	0
1	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	1	0	1	0	0	1	1	0	0	1	1	0	1	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	1	0	0	1
1	1	0	0	0	0	1	1	1	1	0	0	0	1	1	0	1	0	1
1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
1	0	0	0	1	0	1	0	0	1	0	1	0	0	1	0	0	1	1
0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	1	0	0	0	1	0	1	1	1	0	1	0	0	0
1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

Học phác đồ điều trị
Clustering & Learning treatment regimes

Heterogenous objects fitted into a mixed-variate Restricted Boltzmann Machines to output a homogenous latent representation, upon which clustering is performed to learn action protocol for recommendation

Hoang H., Ho T.B., PAKDD 2018.
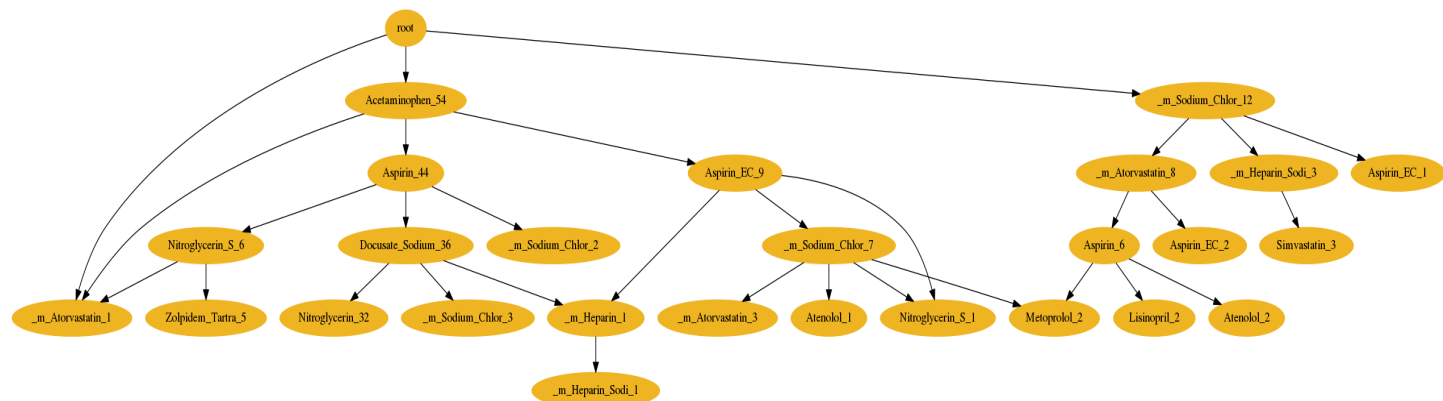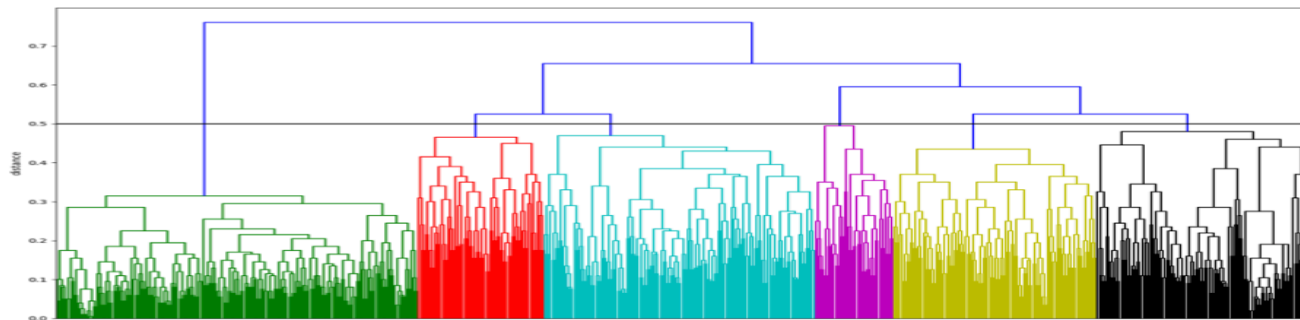
Học phác đồ điều trị
Recommend treatment regimen
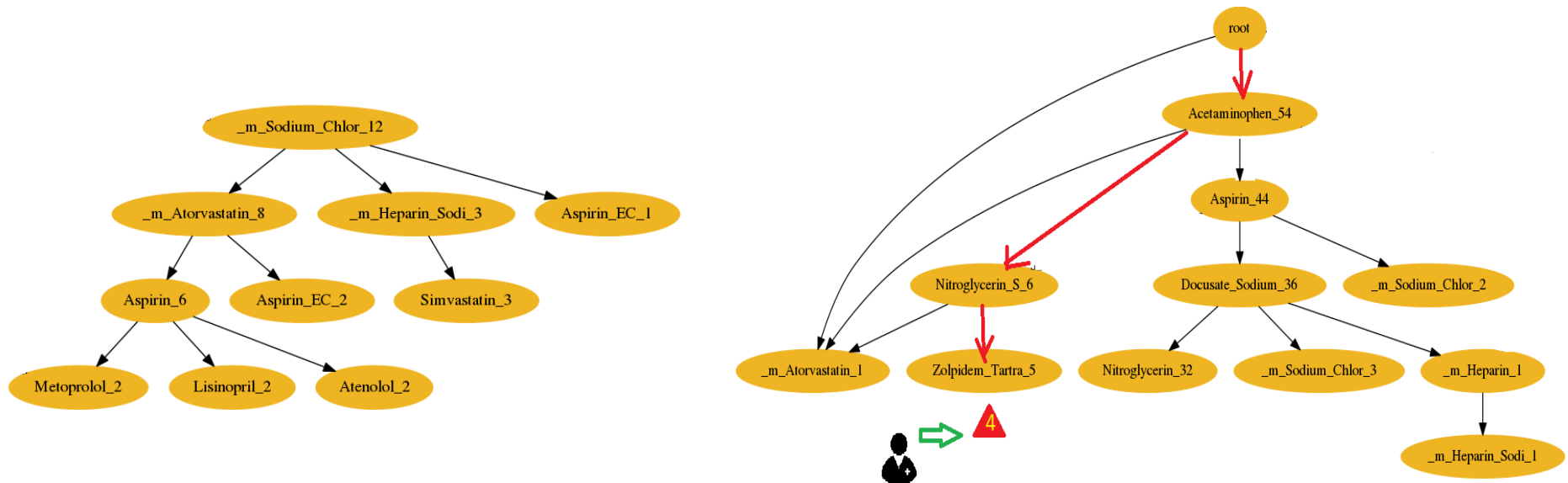
Depth Framework	2	3	4	5	6
Baseline	0.375	0.377	0.374	0.382	0.369
Symptom + Nearest	0.772	0.735	0.702	0.678	0.6597
Symptom + Ensemble	0.818	0.796	0.758	0.728	0.703
Treatment (K) + Nearest	0.745	0.714	0.684	0.655	0.638
Treatment (K) + Ensemble	0.783	0.781	0.745	0.715	0.69
Treatment (A) + Nearest	0.725	0.696	0.672	0.651	0.638
Treatment (A) + Ensemble	0.784	0.775	0.737	0.707	0.682
Dual Ensemble (K)	0.815	0.795	0.758	0.729	0.708
Dual Ensemble (A)	0.814	0.797	0.756	0.729	0.705

The ensemble recommendation framework achieves better performance in comparison

Hoang Khanh Hung and Tu Bao Ho, submitted

88

Take home message

- Electronic medical records is playing a paradigm shift in healthcare and medical research: e-health with data-driven care and research.

- Open new chances to learn treatment regimens.

- Besides genomic data, EMRs is a golden resource for post-market drug study.

- Processing both clinical data and para clinical data from EMRs will lead to big change in healthcare and medical research.

Thank you