National Taiwan University

Integrity, Deep Learning
Love our country & people worldwide

Next Generation Healthcare Information System to Support Precision Medicine

賴飛羆 Feipei Lai 9/19/2019

National Taiwan University

flai@ntu.edu.tw

IFQSH 2019, Taipei

Essentials of AI

- 1. Big Clean Data (90%)
- 2. Supercomputers with GPUs (7~9%)
- 3. Deep Learning Algorithms (3~1%)

AI Applications in Medical Domain

- 1. Image (Now)
- 2. EMR (To Be)
- 3. Gene, Life Style and Environment (Will Be)
- 4. Protein–Protein interactions (May Be)

Outline

- Integrated Medical Database, *i*MD-Taiwan
- What is Precision Medicine?
- Preliminary results of Gene and EMR
- Preliminary results of Life Styles & Environment Open Data
- PM to-do-list

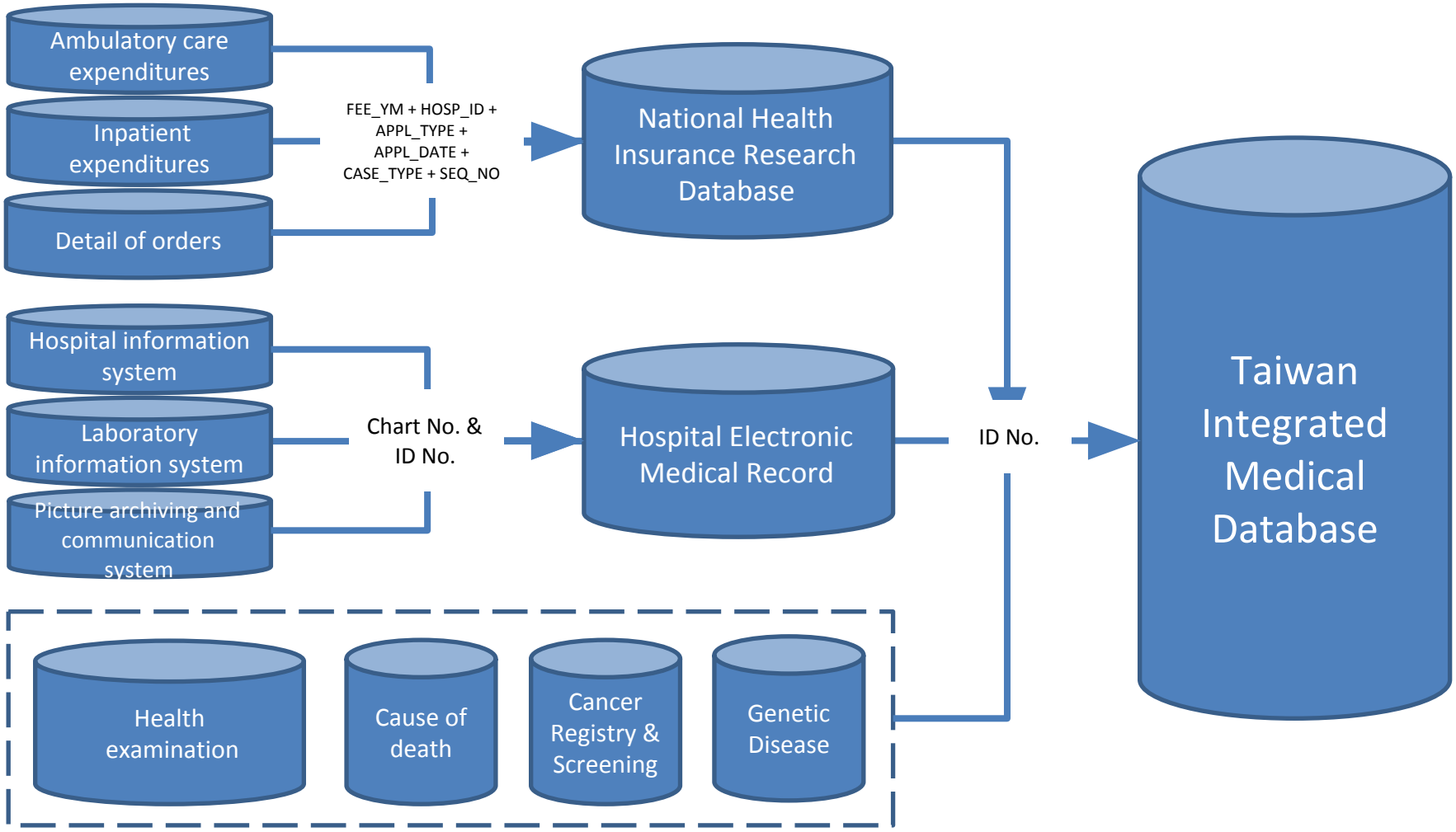
Taiwan National Health Databases

Database	Main information	Data period/ Data point
National Health Insurance Research Database	outpatient, inpatient, emergency, dental, and traditional Chinese medicine: diagnosis and treatment procedure (ICD code), medication use, utilization, age, sex, occupation, hospitalizations, clinical visits, cost	1998 – 2017 48,414,163,600
Registry for Catastrophic Illness	catastrophic illness (ICD code), disease type, approval/valid date, death mark/date	2001 – 2017 4,795,498
Birth Certificate	pregnant nationality/week/time, delivery way, complication, newborn status/weight/defect, baby number	2001 – 2016 3,400,487
Cause of Death	date of death, cause of death, death place, marriage status	1971 – 2017 6,981,999
Cancer Registry	tumor size/TNM stage/histology/behavior/grade, diagnosis date/pattern, therapeutic type, invasion of lymph node	1979 – 2015 3,987,369
Cancer Screening (included in MyHealthBank)	colorectal: fecal occult blood test, test result breast: mammography, family/MC history oral: oral mucosa, betel nut chewing/smoke cervix: pap smear test, pathology	2004 – 2014 63,225,491

Comparison between Taiwan National Health Database and Hospital EMR

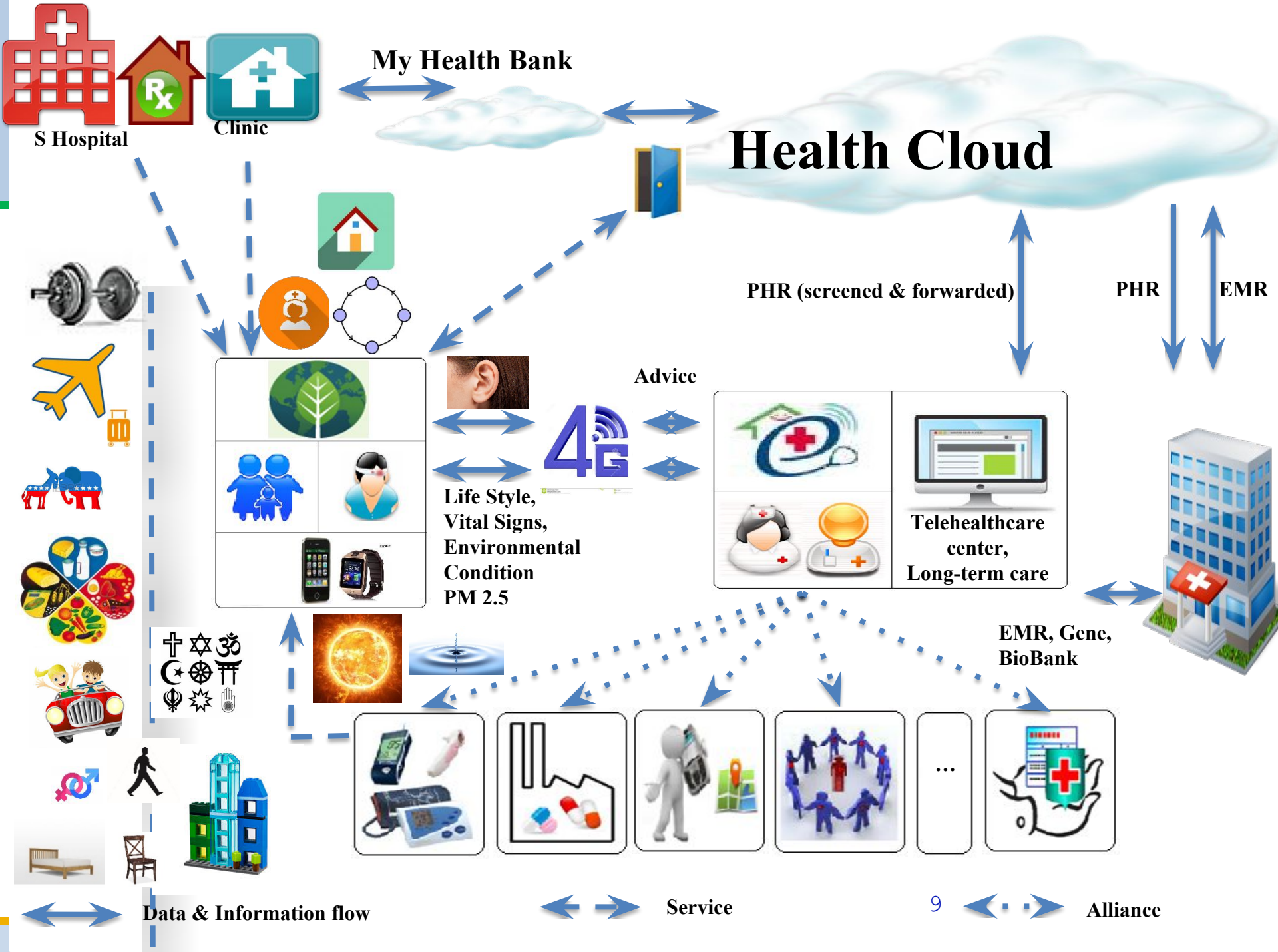
	National Health Insurance Research DB	Hospital EMR
Strengths	1. Nationwide population 2. Constructing longitudinal histories for cohort study design	1. Include potentially important factors in clinical studies, such as laboratory data, image findings, smoking, alcohol use, exercise, diet, and family history 2. Include information of over-the-counter (OTC) procedure 3. Real time 4. Detailed information of disease severity 5. Detailed timing information of medications and procedures
Weaknesses	1. Without potentially important factors in clinical studies, such as laboratory data, image findings, smoking, alcohol use, exercise, diet, and family history 2. Without information of over-the-counter (OTC) procedure 3. The time lag for database released to the public could be as long as 12 to 24 months 4. Difficult to identify disease severity 5. Without timing information among events during hospitalization	1. Without follow-up information and status of patients, such as readmission, complication, and long-term prognosis 2. No medical visit records by any other medical institutions in Taiwan 3. Do not link with other public health databases

Databases integration



Outline

- Integrated Medical Database, *i*MD-Taiwan
- What is Precision Medicine?
- Preliminary results of Gene and EMR
- Preliminary results of Life Styles & Environment Open Data
- PM to-do-list


My Health Bank
S Hospital
Clinic
Health Cloud
PHR (screened & forwarded)
PHR
EMR
Advice
4G
Life Style,
Vital Signs,
Environmental
Condition
PM 2.5
Telehealthcare
center,
Long-term care
EMR, Gene,
BioBank
...
Data & Information flow
Service
Alliance

Outline

- Integrated Medical Database, *i*MD-Taiwan
- What is Precision Medicine?
- Preliminary results of EMR and Gene
- Preliminary results of Life Styles & Environment Open Data
- PM to-do-list

Background

- When expert coding, a patient takes about 20 to 40 minutes
- Coding is both laborious and time consuming

ICD-10 (extreme multi labeling problem)

I	A00–B99	Certain infectious and parasitic diseases
II	C00–D48	Neoplasms
III	D50–D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00–E90	Endocrine, nutritional and metabolic diseases
V	F00–F99	Mental and behavioural disorders
VI	G00–G99	Diseases of the nervous system
VII	H00–H59	Diseases of the eye and adnexa
VIII	H60–H95	Diseases of the ear and mastoid process
IX	I00–I99	Diseases of the circulatory system
X	J00–J99	Diseases of the respiratory system
XI	K00–K93	Diseases of the digestive system
XII	L00–L99	Diseases of the skin and subcutaneous tissue
XIII	M00–M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00–N99	Diseases of the genitourinary system
XV	O00–O99	Pregnancy, childbirth and the puerperium
XVI	P00–P96	Certain conditions originating in the perinatal period
XVII	Q00–Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00–R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00–T98	Injury, poisoning and certain other consequences of external causes
XX	V01–Y98	External causes of morbidity and mortality
XXI	Z00–Z99	Factors influencing health status and contact with health services
XXII	U00–U99	Codes for special purposes

Differences Between ICD-9-CM and ICD-10 Code Sets		
	ICD-9-CM	**ICD-10 code sets**
Procedure	3,824 codes	71,924 codes
Diagnosis	14,025 codes	69, 823 codes

ICD-10 Code Structure Changes (selected details)		
	Old	**New**
Diagnosis Structure	ICD-9-CM • 3 -5 characters • First character is numeric or alpha • Characters 2-5 are numeric	ICD-10-CM • 3 -7 characters • Character 1 is alpha • Character 2 is numeric • Characters 3 – 7 can be alpha or numeric
Procedure Structure	ICD-9-CM • 3-4 characters • All characters are numeric • All codes have at least 3 characters	ICD-10-PCS • ICD-10-PCS has 7 characters • Each can be either alpha or numeric • Numbers 0-9; letters A-H, J-N, P-Z

Objectives

– Build an automated ICD-10 coding system by machine learning methods

Right inguinal hernia
1.Undescended testes, bilateral; 2.E
1. Benign prostatic hyperplasia stat
1. Left inguinal hernia2. Allergic rh
1. Right inguinal hernia, status post
C18.9 Colon cancerE66.9 ObesityM
1. Umbilical hernia
1. Right inguinal hernia2. Bilateral
1.Left inguinal hernia
Right inguinal hernia
1. Bilateral undescended testis2.Bil
1. Recurrent right inguinal hernia

K40.90		
K40.90		
K40.20	B18.1	H16.003
K40.20	N43.3	
K40.20	K65.9	N40.1
K40.90		
K40.90		
K40.90	Q21.0	
K40.20	I48.91	
K40.90	Z87.74	
K40.91	G20	E03.9
K40.20		

Missing value problems, different length, unbalanced labels

A	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
ACCOUNTID	入院診斷	出院診斷	轉出加護病	主訴	病史	身體檢查	手術	住院治療經	併發症	檢驗紀錄	檢查紀錄	影像報告	轉出出院情	病理報告	轉出出院指
aDvIs	1. right inguir	Right inguina	Nil	Informant:The patient, patient's family and previous records Right] inguinal bulging mass noted for several weeks	on: Past Surgical History: Travel History: This is a 84-years-old male patient who has history of: 1. Colon cancer, mucinous adenocarcinoma, pT3N1M0, stage IIIB, status post anterior resection on 2006-01-14, status post FOLFOX (VI) on 2006-02-21,	體檢查 (Physical Examination at admission) BH:158.9 cm BW:62.4 kg 1. General appearance: fair 2. Consciousness: clear, Activity: well 3. HEENT: gross anomaly(-); Conjunctiva: not pale; Sclera: anicteric; Pupils: isocoric 3mm/3mm,	Date: 2015/12/30 手術醫師: 梁金銅 Pre-operative Diagnosis: Inguinal hernia, right side Post-operative Diagnosis: Inguinal hernia, right side Operative Method: Tension-free hernia repair with mesh-plug method Operative Findings: 1. Right side indirect type inguinal hernia, no	Tension-free hernia repair with mesh-plug method was done, tolerable pain was noted. Due to stable condition, he can be discharged and followed up at OPD. 住院用藥摘要: Cefazolin Sodium, Amiodarone HCl, Dutasteride, Propranolol HCl, Tamsulosin HCl,	nil	檢驗室:HE******** 群組:CBC+PLT WBC RBC HB HCT MCV time/item (k/μL) (M/μL) (g/dL) (%) (fL) 1041229 [1208] 7.23 4.56 13.7 42.5 93.2	nil	nil	治療出院 治	nil	制：無特殊限制，可依個人日常生活進行 飲食注意事項：無特別飲食限制，請依個人日常飲食進行 其他指示：請依預約日期返院門診 Magnesium Oxide【MgO 250 mg/tab 限住院藥局用】1 tab PO QID 出院藥天數7天 Acetaminophen【PARAMOL 500 mg/tab】1 tab PO QID 出院藥天數

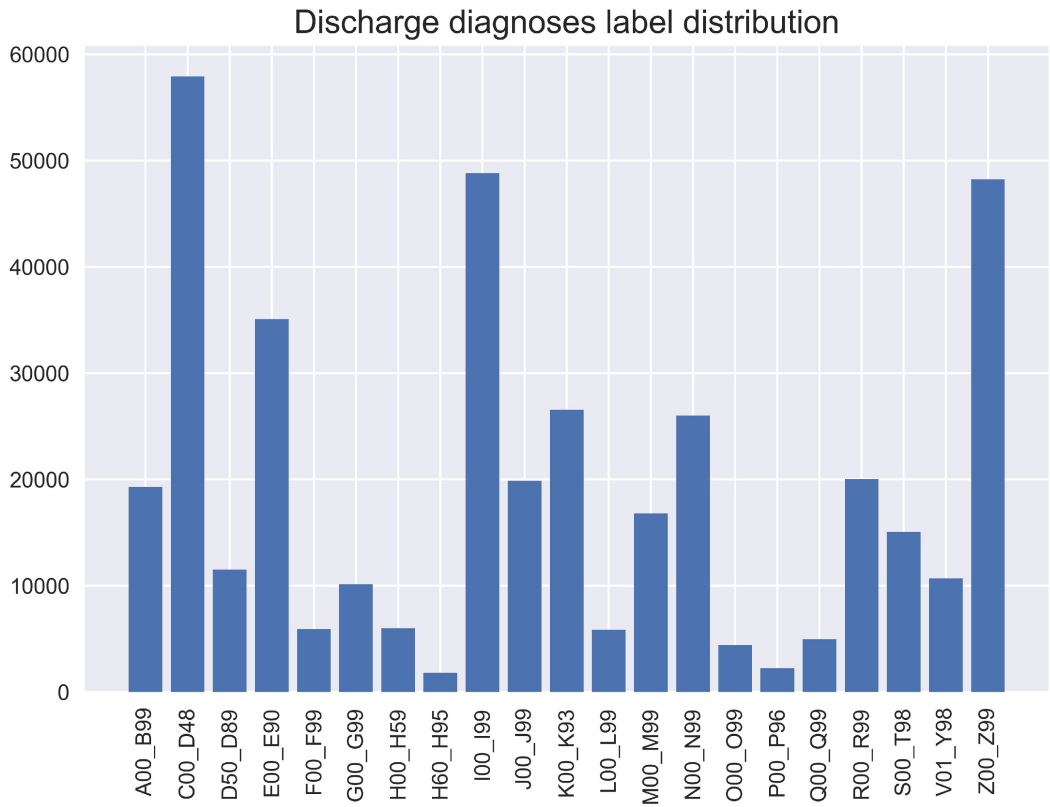
Data description

- ICD-10 CM code: 12,291
- Data volume: **144,120**
- Training data type:
  - Chief complaint
  - Progress
  - History
  - Pathology report
  - Physical examination
  - Discharge diagnosis
  - Transfer out of ICU diagnosis

Data distribution

- discharge diagnosis 139,565

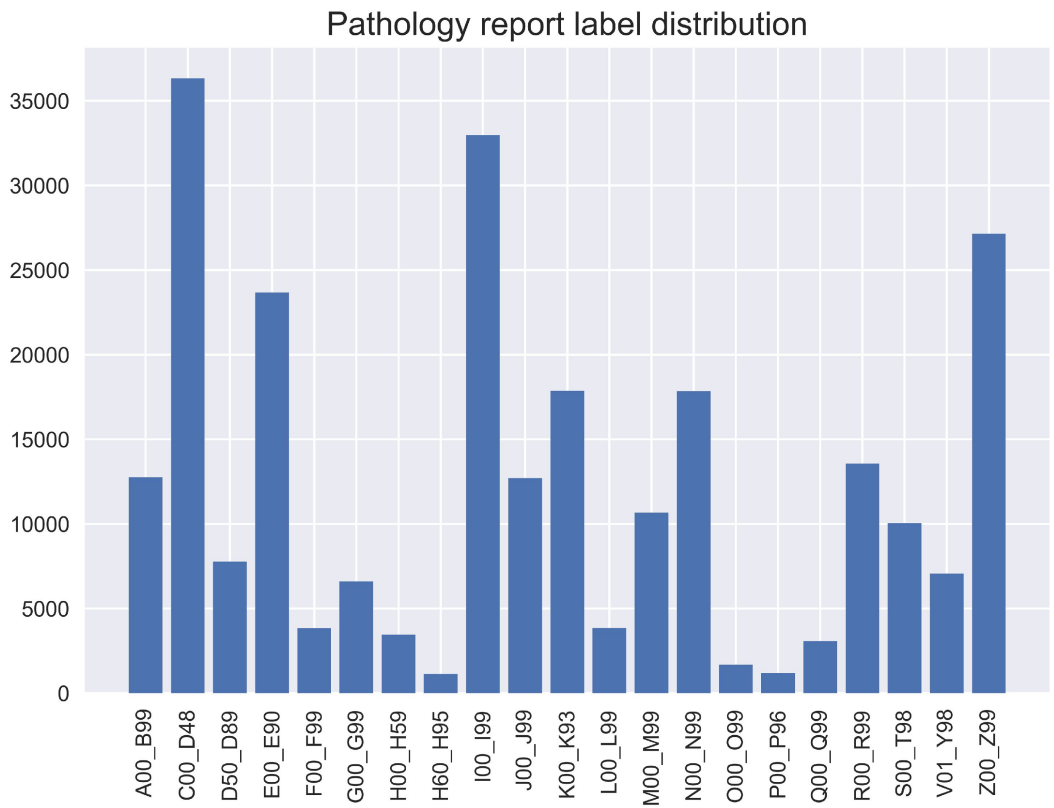
- Max length: 306
- unique words: 14,858



Data distribution

- pathology report 83,384

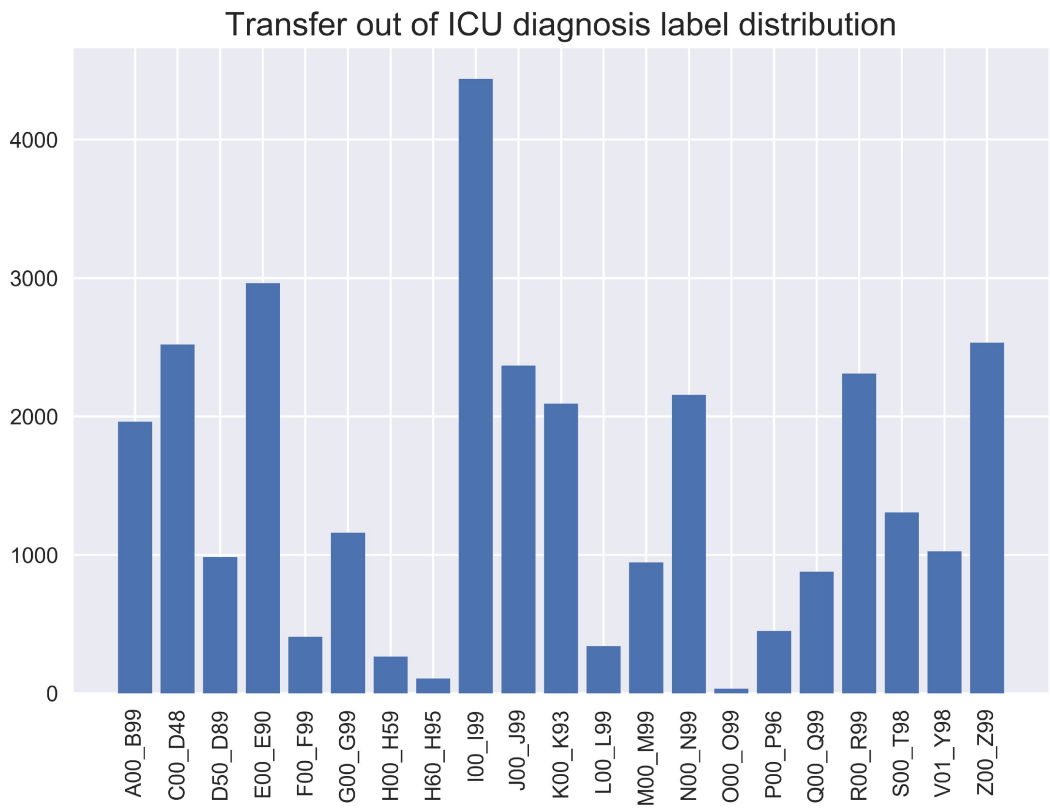
- Max length: 354
- unique words: 3,023 (too little)



Data distribution

- transfer out of ICU diagnosis 7,973 (too little)

- Max length: 270
- unique words: 4,048



Data preprocessing

'1. Ovarian cancer, serous adenocarcinoma, high-grade, stage IIIC, status post optimal debulking surgery (left salpingo-oophorectomy, cytoreduction, infracolic omentectomy and appendectomy) on 2008/05/22 and (Paclitaxel and Carboplatin)*6(2008/06/22-09/16); recurrence with liver metastasis, status post radiofrequency ablation on 2014/02/21, with peritoneal metastasis induced ileus, status post adhesionalysis and jejunoileal bypass on 2014/4/23, status post (weekly Paclitaxel and Carboplatin)*3(2014/05/24-07/31), status post optimal debulking surgery (small bowel segmental resection, cytoreduction, adhesionalysis and bilateral DBJ insertion) on 2014/08/22; complicated with small bowel perforation, status post Repair of small bowel perforation and small bowel side-to-side anastomosis on 2014/08/26; recurrence with liver metastasis, status post radiofrequency ablation on 2015/04/16, 05/21, and 06/25, status post (weekly Paclitaxel and Carboplatin)C5D15, with small bowel perforation and enterocutaneous fistula\n2. Peritonitis with intraabdominal abscess, with small bowel perforation and enterocutaneous fistula, ovarian cancer peritoneal metastasis related\n3. Bilateral knee joint synovitis.'

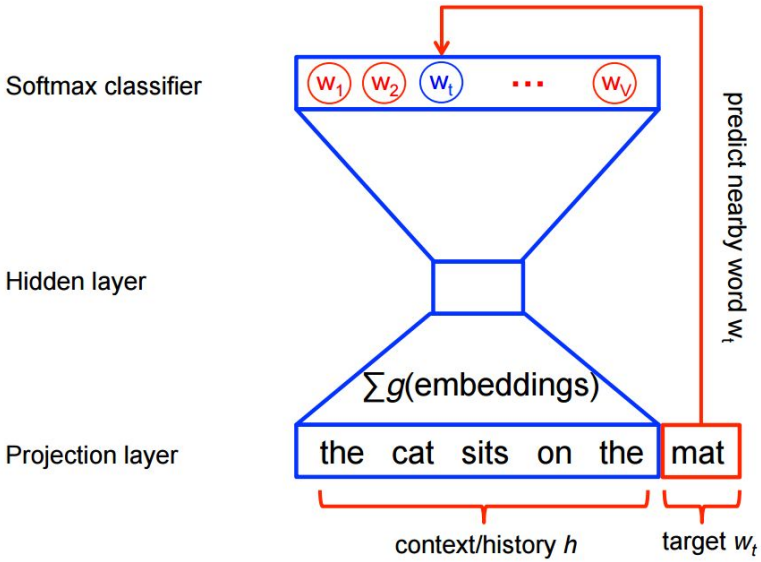
1. converting into lower case
2. removing stop words (*the*, *is*, *at*, *which*, *on*)
3. eliminated the words, not medical related, less appearing

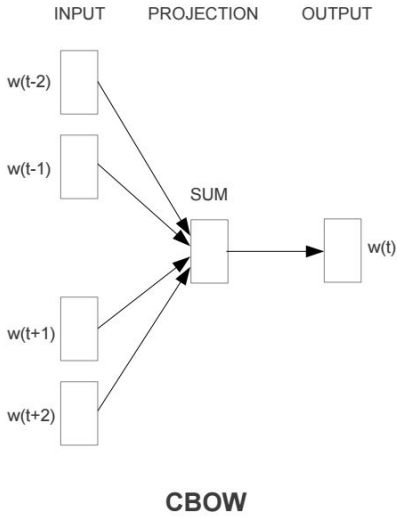
'ovarian cancer serous adenocarcinoma high grade stage iiic optimal debulking surgery salpingo oophorectomy cytoreduction infracolic omentectomy appendectomy paclitaxel carboplatin recurrence liver metastasis radiofrequency ablation peritoneal metastasis induced ileus adhesionalysis jejunoileal bypass weekly paclitaxel carboplatin optimal debulking surgery small bowel segmental resection cytoreduction adhesionalysis bilateral dbj insertion complicated small bowel perforation repair small bowel perforation small bowel side side anastomosis recurrence liver metastasis radiofrequency ablation weekly paclitaxel carboplatin small bowel perforation enterocutaneous fistula peritonitis intraabdominal abscess small bowel perforation enterocutaneous fistula ovarian cancer peritoneal metastasis related bilateral knee joint synovitis'

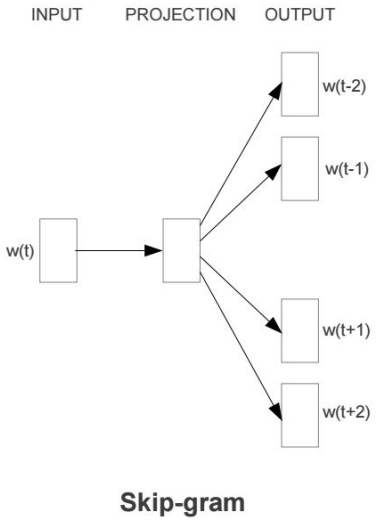
Feature extraction

- Word2vec

- Created by a team of researchers led by Tomas Mikolov at Google
- A group of related models that are used to produce word embedding







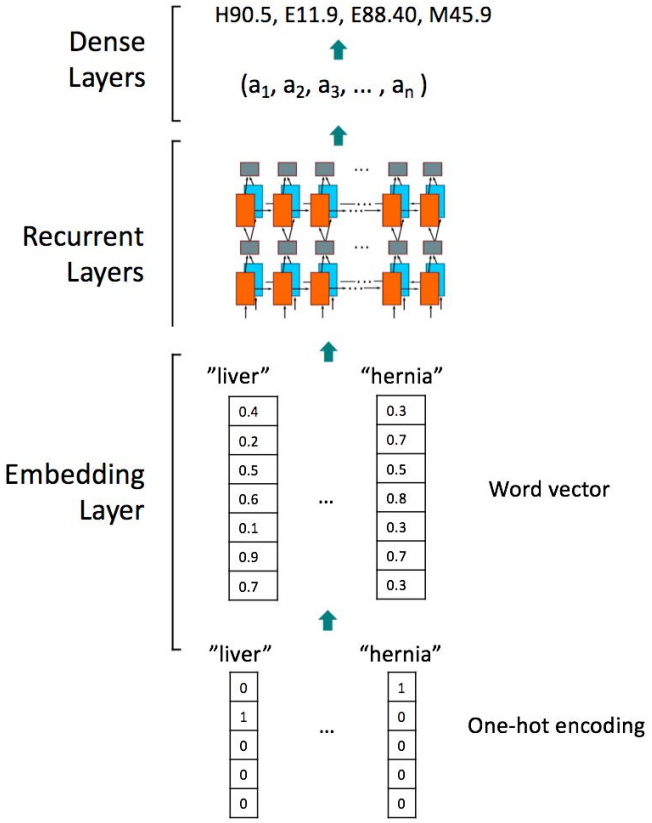
Feature extraction

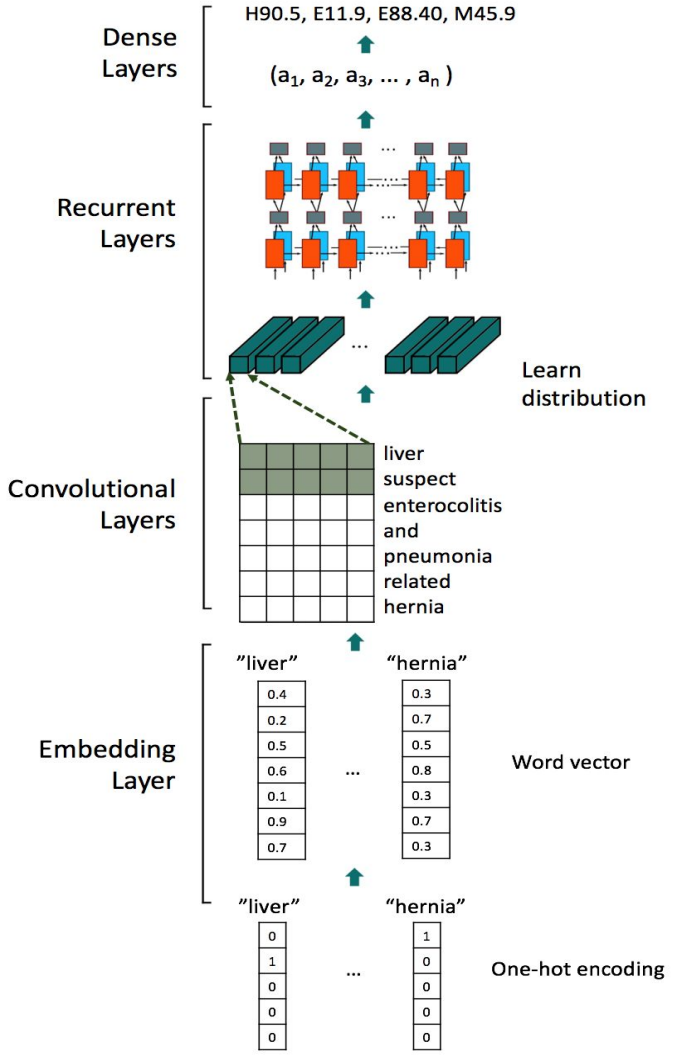
- Word2vec

```
'liver': array([ 1.21615875,  1.68940961, -2.49957013,  1.98355567,  1.59935582,
       -0.83841294,  0.65176499,  2.56678414,  4.14427185, -4.13278484,
       -2.12083054, -1.8522439 ,  1.09076416,  1.04200613, -0.68361241,
        2.09606862, -2.64920306,  0.46346927,  1.13365781, -1.0932641 ,
       -3.07526779, -1.47863853,  2.42217994,  2.38274717,  0.04512066,
        1.08634627, -0.20155224,  1.97593939,  0.8243534 ,  0.15649199,
       -1.24295926, -1.08944464,  0.0867281 , -2.50360346, -0.77610105,
       -1.03488839, -1.00410104, -2.89629769,  1.04792035, -3.28821707,
        0.83427483, -0.30315682, -2.53073049,  3.50550079,  1.11031449,
        4.8127408 , -2.38508463, -3.43210483, -2.59828782, -1.02613699,
        4.6447382 ,  5.2100606 , -1.5950321 , -5.76604414,  0.96173126,
       -0.88581026, -0.4082042 ,  3.1841085 ,  0.76301599, -0.58662415,
       -3.99700499, -1.18123984, -0.99098474, -2.46303773, -0.51321268,
       -1.85507703, -2.40457559,  0.6285612 ,  0.89840859,  0.59829599,
       -1.51530409, -1.69485164, -0.40196013, -0.39235681, -3.36257768,
       -0.56782615,  1.5169872 ,  2.24017382, -3.58359075, -0.05919989,
       -0.8911503 ,  0.91509557,  2.01442504, -0.84691536, -1.11354053,
        1.87302494,  0.24436112,  1.66490316, -1.04922938,  4.15188026,
       -3.79901242,  1.7169323 ,  1.5759362 , -0.18215163, -0.36818993,
       -5.24973822, -0.94957596,  0.42252877, -2.39204788, -1.776389  ], dtype=fl
```

Classification algorithm

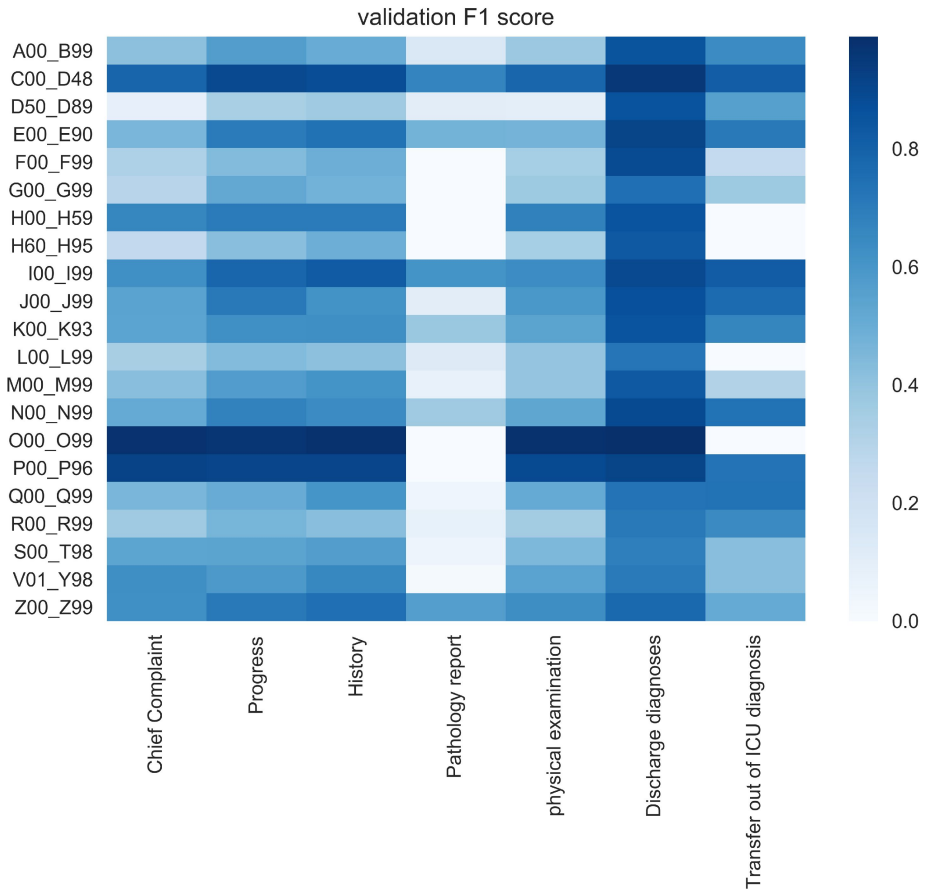
- Recurrent: 1, Dense: 4, Convolutional: not help here





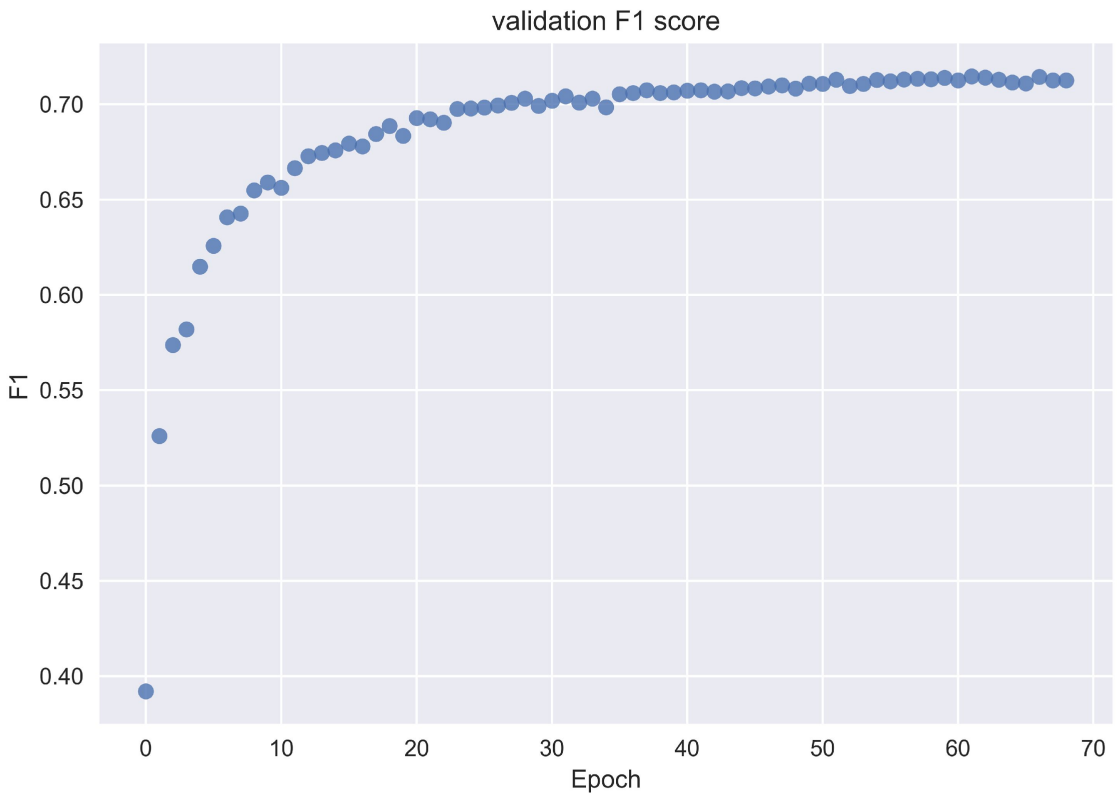
21-label F1 score

- Chief complaint: 0.572
- Progress: 0.695
- History: 0.692
- Pathology report: 0.443
- Physical examination: 0.583
- Discharge diagnosis: 0.876
- ICU diagnosis: 0.658



The first three digits prediction W59

- Max length: 306
- Unique words: 14,858
- Min count: 5
- Embedding dim: 300
- Total code: 1,598
- F1 score: 0.715



All label prediction W59.22, I25.119X

- Max length: 306
- Unique words: 14,858
- Min count: 5
- Embedding dim: 300
- Total code: 12,291
- F1 score: 0.66

- ICD 9, 0.41 achieved by Google
- (n: 85,522), Nature, digital medicine
- 08 May 2018

Web ICD10 predictor

Discharge diagnosis:

1. Amyotrophic lateral sclerosis, with hypercapnic respiratory failure
2. Pneumonia, with right lower lung atelectasis
3. Diastolic heart failure,
4. Anemia, suspect gastrointestinal bleeding related,
5. Hypertension,
6. Subclinical hyperthyroidism

Input free text diagnosis

submit

CM | PCS

Predict 20th highest ICD10 codes

Code	Title
G12.21	Amyotrophic lateral sclerosis
K92.2	Gastrointestinal hemorrhage, unspecified
D50.0	Iron deficiency anemia secondary to blood loss (chronic)
J96.90	Respiratory failure, unspecified, unspecified whether with hypoxia or hypercapnia
J18.9	Pneumonia, unspecified organism
J96.92	Respiratory failure, unspecified with hypercapnia
E05.90	Thyrotoxicosis, unspecified without thyrotoxic crisis or storm
J98.11	Atelectasis
I10	Essential (primary) hypertension

Contact info.: hb2506.t619@gmail.com

Problem

- Difficult to learn, more than 4,000 ICD 10 codes appear just once!
- Spelling error, mixed with Chinese
- Golden standard not built
- Abbreviation
- Combinational codes
- Standard usage

Data Aggregation (心肌梗塞)

Concept ID: A SNOMED CT Identifier that uniquely identifies a Concept (meaning).

Term	Description ID	Concept ID
myocardial infarction	37436014	22298006
cardiac infarction	37442013	22298006
heart attack	37443015	22298006
myocardial infarct	1784873012	22298006
MI - Myocardial infarction	1784872019	22298006
infarction of heart	37441018	22298006

General Form of Clinical LOINC Names

LOINC codes are created systematically using a six axis model

<component> : <property> :
<timing> : <body system> :
<scale> : <method>

8331-1 Body Temperature: TEMP: PT: MOUTH: QN

The first 5 parts are mandatory, but *method* is optional.

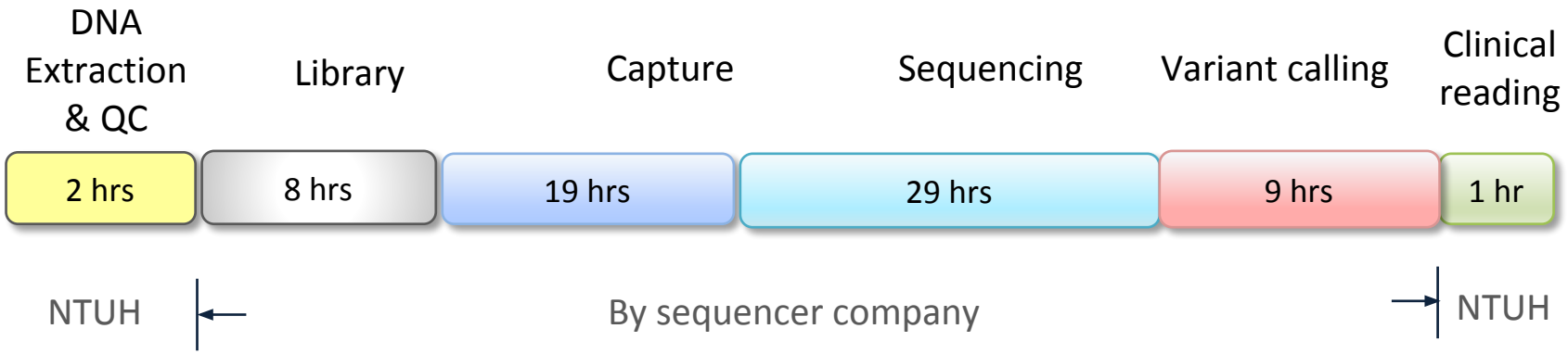
Fast Healthcare Interoperability Resources (FHIR)

- **id**: example
- **status**: final
- **category**: Vital Signs (Details: {http://hl7.org/fhir/observation-category code 'vital-signs' = 'Vital Signs', given as 'Vital Signs'})
- **code**: Body Weight (Details: {LOINC code '29463-7' = 'Body weight', given as 'Body Weight'}; {LOINC code '3141-9' = 'Body weight Measured', given as 'Body weight Measured'}; {SNOMED CT code '27113001' = 'Body weight', given as 'Body weight'}; {http://acme.org/devices/clinical-codes code 'body-weight' = 'body-weight', given as 'Body Weight'})
- **subject**: Patient/example
- **context**: Encounter/example
- **effective**: 28/03/2016
- **value**: 185 lbs (Details: UCUM code [lb_av] = 'lb_av')

Could rapid NGS benefit Pediatric emergent/intensive care?

http://greysanatomy.wikia.com/wiki/If_Only_You_Were_Lonely?file=8x16-17.jpg

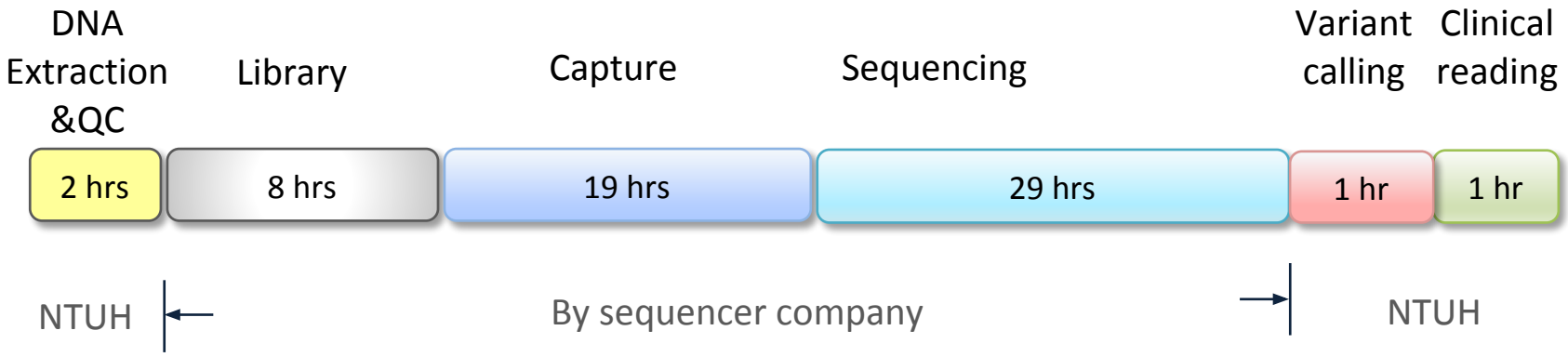
Rapid Whole Exome Sequencing for Pediatric Intensive Care Unit



Turn-around time: 7 days

Modified from Meng LY, et al. ACMG poster 2018

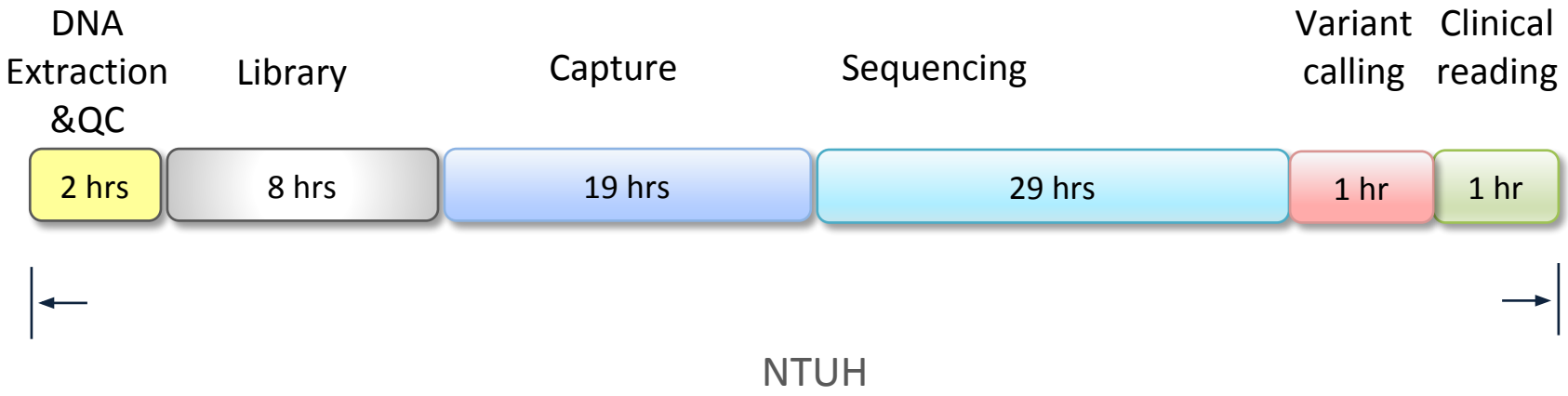
Rapid Whole Exome Sequencing
2nd Phase



Turn-around time: 6 days

Modified from Meng LY, et al. ACMG poster 2018

Rapid Whole Exome Sequencing
3rd Phase



Turn-around time: 3 days

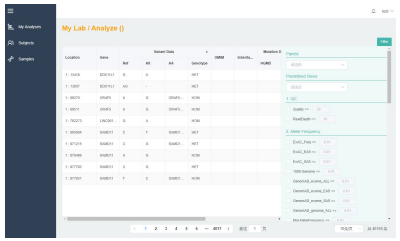
Modified from Meng LY, et al. ACMG poster 2018

Web-based MViewer + Variants Prioritizer

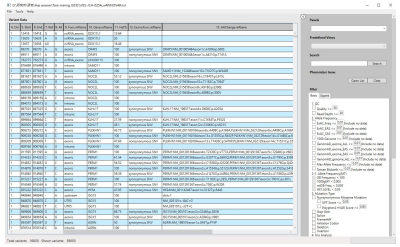
Whole Exome Sequencing

Alignment and Variant Calling

MViewer



Web Application



Single User Program

SNVs and small Indels Annotated from wANNOVAR

1. Query other Databases
   a. *Human Gene Mutation Database*
   b. *Online Mendelian Inheritance in Man*
   c. *Taiwan Biobank*
   d. *ClinVar*
   e. *Nirvana*
   f. *Variant Effect Predictor*

2. Trio Analysis

3. Filter

4. Interpretation Tool

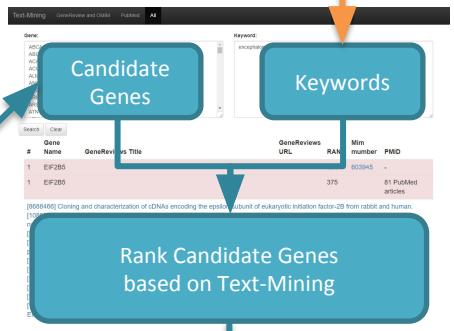
Clinical Report

Database

EMRs

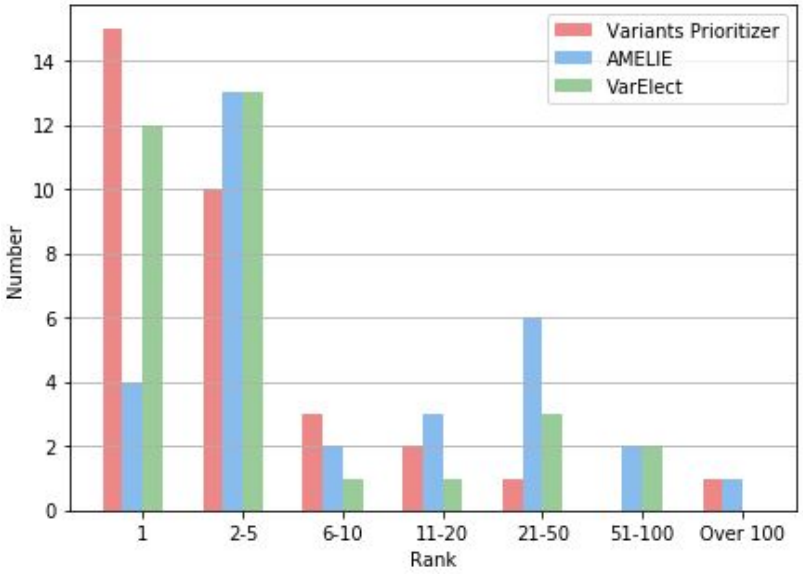
Variants Prioritizer

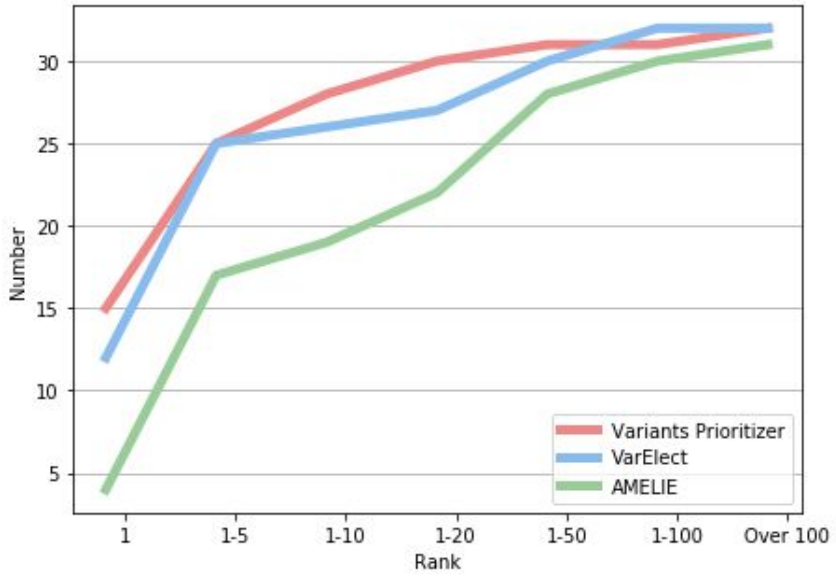
Web Application



Gene Ranking List (Result)

Result and Discussion – AMELIE (Stanford) and VarElect (Genecards)





Rank	1	2-5	6-10	11-20	21-50	51-100	Over 100
Variants Prioritizer	15	10	3	2	1	0	1
	46.88%	*31.25%*	*9.38%*	*6.25%*	*3.13%*	*0.00%*	*3.13%*
AMELIE	4	13	2	3	6	2	1
	12.5%	*40.63%*	*6.25%*	*9.38%*	*18.75%*	*6.25%*	*3.13%*
VarElect	12	13	1	1	3	2	0
	37.5%	*40.63%*	*3.13%*	*3.13%*	*9.38%*	*6.25%*	*0.00%*

Rank	1	1-5	1-10	1-20	1-50	1-100	Over 100
Variants Prioritizer	15	25	28	30	31	31	32
	46.88%	*78.13%*	*87.5%*	*93.75%*	*96.88%*	*96.88%*	*100.00%*
AMELIE	4	17	19	22	28	30	31
	12.5%	*53.13%*	*59.38%*	*68.75%*	*87.5%*	*93.75%*	*96.88%*
VarElect	12	25	26	27	30	32	32
	37.5%	*78.13%*	*81.25%*	*84.38%*	*93.75%*	*100.00%*	*100.00%*

Outline

- Integrated Medical Database, *i*MD-Taiwan
- What is Precision Medicine?
- Preliminary results of Gene and EMR
- Preliminary results of Life Styles & Environment Open Data
- PM to-do-list

The Architecture of Life style and Environment Information System

AQI, CO, O3, NO2, PM2.5, PM10, SO2, UVI, Humidity, Temperature

Environmental Data

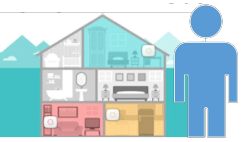
氣象資料開放平臺 OPEN DATA

OpenData.epa 行政院環境保護署・環境資源資料開放平臺

Open API

Input → Valuable Data

Personal Health Advice APP



Environment and Health Information

Smart Phone Sensor Data

Upload Environmental Data

EDIMAX

EDIMAX API

- **Health Related Datasets**
- **Environment Related Datasets**

Doctor observe platform

GARMIN

Upload Physiological Data

GARMIN HEALTH

Garmin Health API

Hospital Data Analytics

Upload Physiological Data

fitbit

Fitbit API

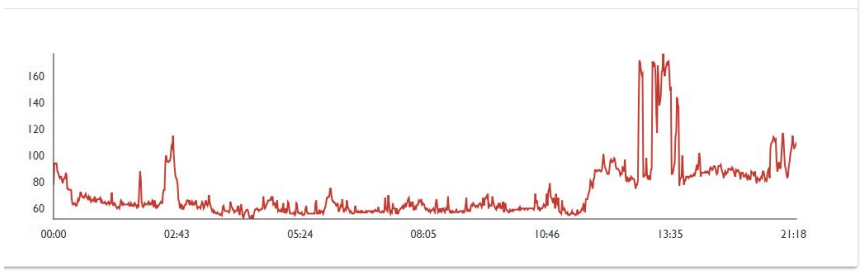
Fitbit Versa Smart Watch

Personal Health Data

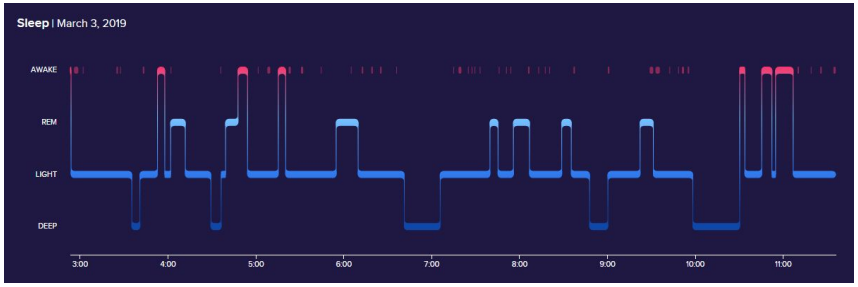
Valuable Health Data

	A	B	C	D	E	F	G	H
1	date	serial_numl	name	phone	steps	floors	distance	calories
2	2019/3/2	test002	[redacted]	[redacted]	5471	7	5.9	2437.2
3	2019/3/3	test002			3171	8	2.3	2539.3
4	2019/3/4	test002			16161	11	11.9	3600.6
5	2019/3/5	test002			7462	5	5.5	2595.3
6	2019/3/6	test002			10488	2	8.3	2969.1
7	2019/3/7	test002			6591	4	4.9	2769.7
8	2019/3/8	test002			6782	1	5	1894.9

steps, floors, distance, calories



Heart rate



Sleep Hypnogram

Outdoor Environmental Monitoring

Open Data Application

- **Source**: EPA, CWB
- **Government Open Data API**: Real-time and Historical
- **Data Integration**:
  - Data preprocessing (Different open data schema)
  - Import environmental standards

交通部中央氣象局 Central Weather Bureau

OpenData.epa 行政院環境保護署・環境資源資料開放平臺

Outline

- Integrated Medical Database, *i*MD-Taiwan
- What is Precision Medicine?
- Preliminary results of Gene and EMR
- Preliminary results of Life Styles & Environment Open Data
- NTUH PM to-do-list & future work

To-Do-List

- Integrate NHI MyHealthBank - to get complete EMR
- Integrate post acute, long term, hospice care - to get complete EHR
- Use environmental sensors to get environment & climate open data
- Promote mobile healthcare to get complete life style data, SmartPhone + SmartWatch

To-Do-List

- Lab. Data transformed to LOINC forms (RELMA)
- EMR transformed to SNOMED form
- Diseases management platform and service