Healthcare Transformation from Data and System Perspectives

Beng Chin OOI

www.comp.nus.edu.sg/~ooibc
Contents

• Healthcare Problems
• Challenges
• Our Healthcare Data Analytics Stack
  • GEMINI
    • Cleaning, De-biasing, Regularizing
  • ForkBase
    • Storage Engine for Collaborative Analytics and Forkable Applications
  • Foodlg / Foodhealth
    • Pre-diabetes app
  • MediLOT
    • A blockchain solution
• Conclusions
An Obamacare success: financial penalties reduce hospital readmission rates

By JASON H. WASFY, FRANCESCA DOMINICI, and ROBERT W. YEH / DECEMBER 27, 2016

The Ministry Of Health (MOH) Office for Healthcare Transformation (MOHT) (formed in 2018) aims to shape the future of healthcare in Singapore. This is done by identifying, developing and experimenting with game-changing systems-level concepts and innovations in the key areas of health promotion, illness prevention and the delivery of care.

AI in Health Grand Challenge (Ongoing large grant call by AI.SG – 3 x 5 mil in the first phase and 1 x 20 mil in the second phase)

“How can Artificial Intelligence (AI) help primary care teams stop or slow disease progression and complication development in 3H – Hyperglycemia (diabetes), Hypertension (high blood pressure) and Hyperlipidemia (high cholesterol) patients by 20% in 5 years?”
3H Problems: Where/what Can We Contribute?

- **Hyperlipidemia**
- **Hyperglycemia**
- **Hypertension**

**Life Style**

**Drug Compliance + Pharmacogenomics**

**Eye (DME, retinopathy, glaucoma, ...)**
**Kidney (AKI, ESRF ...)**
**Cardiac (AMI)**
**Stroke (AF, fall...)**
**Limb Salvage/amputation**

**Personal Health Coach**

**Sensors + Cameras**

**Telemedicine**

**Hospital System**

**Healthcare Analytics**

**Primary Care**

**Secondary Care ++**
Healthcare System/AI’s Objective

A unified end-to-end engine to integrate all available data sources and provide a holistic view of medical data, from where we support all sorts of medical applications.

- Increase the accuracy of diagnoses
- Improve preventive medicine
- Optimize insurance product costs
- Better understand the needs for medications
- Cut costs on healthcare facility management etc

This is beyond typical database query processing
The Reality of Exploiting AI

• The actual implementation of the ML algorithm is usually less than 5% lines of code in a real, non-trivial application

• The main effort (i.e. those 95% LOC) is spent on:
  • Data cleaning & annotation
  • Data extraction, transformation, loading
  • Data integration & pruning
  • Parameter tuning
  • Model training & deployment
  • … …

• This blurs the line between DB and “non-DB” processing, and calls for better integration

These are what we have been doing!
The BIG Data Analytics Pipeline*

*Alexandros Labrinidis, H. V. Jagadish:
Challenges and Opportunities with Big Data. PVLDB 5(12): 2032-2033 (2012)
Challenges
Identifying Common Challenges

- Readmission
- Disease Progression Modelling (DPM)
- Radiology
- App
- Prediabetes Prev.

Support

GEMINI Platform

Research

Clinical Needs
- Readmission
- Disease Progression Modelling (DPM)
- ...

National University Health System

NUH
National University Hospital

Jurong Health
Ng Teng Fong General Hospital

SingHealth
Defining Tomorrow’s Medicine

Tan Tock Seng Hospital

SingHealth

KK Women’s and Children’s Hospital
China Healthcare Providers/Hospitals

...... more
Challenges

Time-consuming data extraction
• Different storage formats
• Unstructured data

Difficult data cleaning
• Missing data
• Duplications
• Different coding standards

Doctors-in-the-loop data annotation (medical expertise)
• Missing code filling
• Standardized diagnoses

Bias in observation data
• Observation data is biased from the actual conditions of the patients

Complexity of medical features
• Numerous concepts
• Heterogeneous data
• Complex relations

Demanding data storage requirements
• Multi-source and heterogeneous data formats
• Reuse of datasets
• Provenance
Challenge 1: Data Preprocessing

- **time-consuming data extraction**
  - different storage formats, un-structured data

- **difficult and expensive data cleaning**
  - missing data, duplications, different coding standards

- **medical expertise required for data annotation**
  - standardizing diagnoses, missing code filling

---

Diagnoses

Lab Tests

Medications

Procedures

Image Data

Unstructured Text Data
Challenge 2: Bias in EMR Data
NUH surgery dataset:
22987 medical features
12319 diagnosis codes
2335 lab test codes
6932 medication names
1401 procedure codes
8 demographic features (BirthYear, Gender etc)

Numerous Concepts
UMLS consists of over 2.97 million concepts and 10+ million terms.

Multi-source and Heterogeneous Data
Medical data consists of diagnoses, lab tests, procedures, etc.

Complex Relations
Complex relations among different sources of medical data
Challenge 4: Dataset Management in Healthcare

- Dataset Cleansing
  - Track evolution history to ensure correctness

- Dataset Transformation
  - Save different formats for future reuse

- Dataset Sharing/redundancy
  - Avoid data redundancy to reduce storage overhead

- Dataset Security
  - Impose access control to healthcare data
Challenge 5: Data Prior

- Existing ML algorithms work well for image classification and sequence prediction, but not healthcare problems

- Images are not random pixels
  - Neighbor pixels are most correlated --> CNN
  - Color channel prior --> haze removal/super-resolution

- Sequences are not random numbers/words
  - Latent state at each time point --> RNN LSTM

- Prior for healthcare?
  - How to find and formulate?
  - How to create algo/model to utilize them?
Matching Data and Model/Algorithm

• No Free Lunch Theorem [1997]

• Checklist for useful AI:
  • Lots of data
  • Flexible models
  • Efficient system and algorithm design
  • Powerful priors that can defeat the curse of dimensionality

• Opportunities come from utilizing data distribution information
  • Can we learn prior from data? (Domain-specific AutoML)
Development Pipeline

• Parameterize existing data processing solutions to meet the characteristics of healthcare data

**Data Acquisition:**
- Hospital Data
- Genomic Data
- Medical KB
- CT/MRI Images

**Integration & Augmentation:**
- AE/D Data Cleaning
- Collaborate Analytics
- KB Data Enrichment
- Image Augmentation

**Understanding & Interpretation:**
- EMR Bias Resolving
- EMR Imputation
- EMR Embedding
- EMR Pattern Mining

**Application Deployment:**
- Standard Model Pool
- Adaptive Regularizer
- KB Hashing Model
- Bagging & Evaluation

---

Extensive Raw Data → Cleaned Data with Rich Semantics → Extracted Effective Feature Sets → Medical Insights
## Enabling Global Optimization

- **SINGA – RAFIKI (MLaaS) -- PANDA mainly for healthcare**

<table>
<thead>
<tr>
<th>PANDA Healthcare</th>
<th>Current AI systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aim</strong></td>
<td>Defining new AI problems</td>
</tr>
<tr>
<td><strong>Iteration</strong></td>
<td>Optimizing for existing AI problems</td>
</tr>
<tr>
<td>Doctors take part in the</td>
<td>Data scientists as the agent</td>
</tr>
<tr>
<td>development circle</td>
<td></td>
</tr>
<tr>
<td><strong>Key Techs</strong></td>
<td>Efficient declarative interaction</td>
</tr>
<tr>
<td><strong>Domain Knowledge</strong></td>
<td>ML model and platform</td>
</tr>
<tr>
<td>Instilled by doctors</td>
<td>Understood by data scientists</td>
</tr>
<tr>
<td><strong>Delivery</strong></td>
<td>Explored together with doctors</td>
</tr>
<tr>
<td></td>
<td>Plain model outputs</td>
</tr>
</tbody>
</table>

Healthcare Data Analytics

Stack

GEMINI (GEneralisable Medical Information aNalysis and Integration platform)

AI Implementation at NUH

Demographic information
ED notes
Dispensed medication
Visits and encounters
Lab test results
Radiology reports
Procedures
Discharge summaries
Vital signs
Inpatient medications
Inpatient notes
Outpatient notes

CDOC
CCDR

Pre-processing filter matrix

GEMINI
Production AI Modules

Diagnosis module
Readmissions module
Complications module
Disease progression module
VDO module
Future Extensions

H-Cloud

Deep machine learning

Reinforced learning

Predicted clinical WARNING

Deep machine learning

Reinforced learning

National University Hospital

Deep machine learning

Reinforced learning
Example: Readmission Prediction
WARNING

88.6% Chance of readmission

Ranked Factors:
1. Uncontrolled diabetes H/C 16
2. > 6 medications
3. 72.3% chance of post-op wound infection
4. Past readmissions due to social factors

Acknowledge
GEMINI Platform (2011 - )

Application

Healthcare

Data Analysis
Pipeline

Crowdsourcing

EMR Transformation

Machine/Deep Learning

Visualization

Raw Data

CDAS

Data Integration

Big Data Processing

Cohort Analysis

iDat

CDAS

EMR-T

SINGA

GAM

ForkBase

Malleable, Semantic Storage

CPU-GPU Cluster

Application

Data Integration

Big Data Processing

Cohort Analysis

Visualization

Infrastructure

ForkBase

Malleable, Semantic Storage

CPU-GPU Cluster

Application

Data Integration

Big Data Processing

Cohort Analysis

Visualization

Infrastructure

ForkBase

Malleable, Semantic Storage

CPU-GPU Cluster
Making Healthcare Data Usable


Healthcare Data Usability

If a doctor wants to analyze the medical records related to “chronic kidney disease” ...
Healthcare Data Usability

- Two reasons cause the healthcare data usability.
  - Different writing styles.

**Real-world healthcare data**

- 2 recent cva
- posterior circulation transient ischaemic infarct
- multi infarct cva with dementia
- massive ischemic stroke with hemorrhagic conversion
- acute stroke infarct
- 2 rt sided cva with gd recovery 1994 5
- r groin hematoma
- cerebellar stroke
- acute left pontine cva
- acute cva left ic laci
- acute cva left sided weakness
- basal ganglion infarct

**refer to**

<table>
<thead>
<tr>
<th>concept code</th>
<th>Canonical description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I63.50</td>
<td>Cerebral infarction due to unspecified occlusion or stenosis of unspecified cerebral artery</td>
</tr>
</tbody>
</table>
Healthcare Data Usability

• Two reasons cause the healthcare data usability.
  • Different writing styles.
  • Different medical standards.

<table>
<thead>
<tr>
<th>Standard</th>
<th>Concept code</th>
<th>Canonical description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-10-CM</td>
<td>K64.2</td>
<td>Third degree hemorrhoids</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>455.0</td>
<td>Internal hemorrhoids without mention of complication</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>455.1</td>
<td>Internal thrombosed hemorrhoids</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>455.2</td>
<td>Internal hemorrhoids with other complication</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>455.5</td>
<td>External hemorrhoids with other complication</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>455.6</td>
<td>Unspecified hemorrhoids without mention of complication</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>455.7</td>
<td>Unspecified thrombosed hemorrhoids</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>455.8</td>
<td>Unspecified hemorrhoids with other complication</td>
</tr>
</tbody>
</table>

Real-world healthcare data

- internal haemorrhoid prolapsed
- haemorrhoid bleeding ligated
- 3 degree pile
- prolapsed haemorrhoid
- 3rd degree prolapsed piles, not thrombosed
- thrombosed internal haemorrhoid
- 3rd degree pile x 1
- haemorrhoid
- 3rd degree external hemorrhoids
- hemorrhoids prolapsing piles
- haemorrhoids no complication
- prolapsed and thrombosed haemorrhoid at 4 clock
Healthcare Data Usability

• Two reasons cause the healthcare data usability.
  • Different writing styles.
  • Different medical standards.

• To improve the healthcare data usability, we need a linker that is able to automatically link a medical record to a unified concept ontology.
Neural Concept Linking

• We have developed a neural concept linking framework to accomplish the healthcare concept linking.
Neural Concept Linking

Concept representations

Word representations

\[ p(s | c) = 0.016 \]

ckd secondary to dm nephropathy
Example Results

We cleaned 13 years of NUHS data – 90% done by machine, 10% done by human
Resolving “bias”


Knowledge Driven Regularization. K. Yang, Z. Luo, J. Gao, J. Zhao, B.C. Ooi, B. Xie. 2019
Similar Pattern and yet Different Results

- Patient1 always visits hospital due to respiratory infection
  - Can we conclude that Patient1 has respiratory infection every day?

- Patient2 always visits hospital due to chronic kidney disease
  - Can we conclude that Patient2 has chronic kidney disease every day?

- What is the difference?
Bias in EMR Data

- If a doctor or analyst want to analyze the EMR data with missing values, they may employ traditional imputation methods directly

→ Misinterpretation

### Acute kidney failure (AKF)

- $t_1$: N17.9
- $t_2$: ?
- $t_3$: ?
- $t_4$: N17.9
- $t_5$: ?
- $t_6$: ?

? Last observation carried forward

### Glomerular filtration rate (GFR)

- $t_1$: 20
- $t_2$: ?
- $t_3$: ?
- $t_4$: ?
- $t_5$: 40
- $t_6$: ?

? Mean imputation

Mean imputation = $\frac{20 + ? + ? + ? + 40 + ?}{6}$
Bias in EMR Data

- Bias – recorded EMR series is different from patients’ actual hidden conditions
  - Patients tend to visit hospital more often when they feel sick
  - Doctors tend to prescribe the lab examinations that show abnormality

- To Solve Bias Challenge – EMR Regularization
  - Transform the biased EMR series into unbiased EMR series
Resolving Bias in EMR Data

- **Condition Change Rate (CCR)**
  - measures how a medical feature is likely to change from its condition in the previous observation

- **Observation Rate (OR)**
  - measures the probability that a medical feature is exposed at a time point based on its actual condition at that time point
Resolving Bias in EMR Data

- Imputation accuracy evaluation

- Benefits for analytic tasks
  - In-hospital mortality prediction, Diagnosis by category prediction
  - Disease progression modelling
Disease Progression Modeling

Comparably Stable Progression Trajectory

Deteriorating Progression Trajectory

Medical Features

Longitudinal Patient Matrix

Severity Labeled

Prediction Time Point

Time

Severity

Time

2012-01-01 2012-03-01 2012-04-30 2012-06-29 2012-08-28 2012-10-27 2012-12-26

0 10 20 30 40 50 60 70

GFR Value

0 10 20 30 40 50 60 70

Time

2012-01-01 2012-03-01 2012-04-30 2012-06-29 2012-08-28 2012-10-27 2012-12-26

Diabetes Kidney Disease

HbA1C Blood Pressure

Cholesterol Insulin

Amputation

\[ f_m : f_2 \quad f_k \]

\[ s_1 \quad s_2 \quad s_3 \quad \ldots \quad s_k \]

Prediction Time Point

Age Race Gender Education

0 10 20 30 40 50 60 70

2012-01-01 2012-03-01 2012-04-30 2012-06-29 2012-08-28 2012-10-27 2012-12-26

GFR Value
Advice to Doctors on Intervention

- Our model would suggest to guarantee the monitoring for Patient 1 → may need dialysis or kidney transplant
- Our model would suggest healthcare workers to provide more aggressive interventions to Patient 2 in advance
- Our model would suggest to guarantee the monitoring for Patient 3
Facilitating Data Sharing and Provenance

ForkBase Designs

Versioning & Tamper Evidence

Merkle DAG

Indexing & Deduplication

SIRI indexes

Collaboration Workflows

Fork Semantics

ForkBase

git

(versioning)

database

(query)

blockchain

(integrity)
ForkBase Storage Stack

Documents Hosting

Access Control
branch-based

Blockchain

Data Security
integrity

Git

Collaborative
Dataset Mgmt

Consistency
merge semantics

Applications

Semantic Views
(application-oriented)

Data Access APIs
(data types, fork semantics)

Branch Representation
(versioning, tamper evidence)

Chunk Storage
(deduplication, immutability)

| put(object) → version |
| get(version) → {objects} |
| merge({objects}) → object |

Node A

Node B
SIRI Indexes & POS-tree

• An Index Class: Structurally-Invariant Reusable Indexes
  • Structurally Invariant, Recursively Identical, Universally Reusable ...
• An Implementation: Pattern-Oriented-Split Tree

Content-determined Structure  
(-> Deduplication)

Native Merkle Tree  
(-> Tamper Evidence)

Probabilistically Balanced Tree  
(-> Query Efficiency)
Blockchain Data Model in ForkBase

• KV Store
  • Customized structures
    • Linked block
    • State Merkle tree
    • State delta
    • ...
  • Hard to implement

• ForkBase
  • Achieve with built-in types
    • UBlob
    • UMap
    • ...
  • Easy to maintain
    • 10+ lines for each structure
Analytic-Ready Blockchain Backend

- Analytic on blockchain is expensive
  - Need to scan whole block history to extract information
- Built-in data types in ForkBase to support fast analytics

State Scan Query

Block Scan Query
Prevention is Better Than Cure

Lifestyle Intervention Programme (LIVEN)

The effect of a behaviour-based lifestyle change program using combined face and remote sessions on weight, diet intake and physical activity level in people at-risk of diabetes: a Randomised Controlled Trial
Effecting Behavioral Change

Snap

- Quick and Easy way to record dietary intake
- A deep learning image-based food recognition for a faster, closest food match and handy recording

Track

- Self-monitoring with pre-set goals and intuitive nutrition information
- Peer-to peer monitoring of dietary and physical activity goals
- Daily and weekly reports of progress

Feedback

- Remote monitoring by healthcare professionals for timely and meaningful feedback
Diabetes Prevention

Image Recognition

Knowledge Base

Healthcare Analytics

Social Network

Scan

Diary

Review

Share

Activity

Plan

Recommendation

Healthy Diet + Exercise
Administrator/Dietician Portal

• Dietary Review + Chat
  • Review user’s weekly meal (photo) history

Realtime Chat with Dietician provides instant feedback to users
Foodhealth/Foodlg

**STEP 1**
Collect training images from heterogeneous sources and label them via crowdsourcing

**STEP 2**
Train deep learning models for food recognition

**STEP 3**
Food recognition and health analysis using images and other information from the Foodlg app

**Off-line**

**On-line**
Personalizing and Decentralizing Healthcare
AI + BlockChain + Cloud + big Data

Objectives:
1. Transparency
2. Accountability
3. Auditability
4. Governance
5. Security
6. ...
BlockChain enabled Healthcare

• BlockChain (BC) acts as a tamper-evident storage for archiving Healthcare Records from different healthcare providers
• BlockChain acts a “Central Healthcare Record Repository”
• It enables Data Provenance, Data Analytics, and Medical-care everywhere based on patient’s preference
• It may help transform Healthcare management and research
The MediLOT Solution

1. Holistic
Every patient will have a complete longitudinal health record: their own health story that they can access at any institution.

2. Patient-centric
The patient holds his/her own private key and has fine control over who can view their medical records.

3. Personalised
Using an advanced analytics overlay (GEMINI), MediLOT facilitates personalised treatment strategies.

4. Decentralised
Patients’ data is stored in different locations, eliminating the risk of a single catastrophic breach.
Dual BlockChain Schema

Public (Ethereum)
Allows for transfer and crediting of ERC20 LOT tokens (MediLOT utility token)

Permissioned (Hyperledger++)
Responsible for aggregation of patient EHR

Who will Pay?
On-Chain Scalability

Consensus Layer (PBFT, PoW, PoS, etc.)

Smart Contract Execution Engine
(Virtual Machine, Docker, etc.)

Data Model Layer
(LevelDB, RocksDB, etc.)

MediLOT’s Technologies

**Dual Blockchain**

**Ethereum & Hyperledger++**
- Enhanced Hyperledger with scalable consensus and sharding
- Throughput up by 15x

**Analytics**

**GEMINI**
The underlying healthcare suite that supports big data analytics and personalised medicine

**Data Storage**

**ForkBase**
Proprietary storage with rich semantics, immutability and data sharing, Blockchain optimised native storage system
Conclusions

• Healthcare is a complex but impactful/meaningful Application
  • Domain Knowledge
  • Verification and Validation – a tedious process

• A good (example) application that calls for better integration of AI/ML and Database technologies, and possibly Blockchain technologies

• We have addressed some of the challenges, and have implemented:
  • GEMINI (DICE, CDAS, epiC, Apache SINGA, ForkBase) is being used by 2 major hospitals in Singapore
  • Foodhealth (foodlg) is used by 3 hospitals in Singapore
  • MediLOT is in testnet phase and used by hospitals in China

• Objectives:
  • To predict, prevent/pre-empt, personalize for more effective healthcare

• Be Good. If you can’t, be Safe. Live well ...
Acknowledgements

• Collaborators: Gang Chen, H.V. Jagadish, Kee Yuan Ngiam, James Yip++
• Collaborators (ex-students): Meihui Zhang, Wei Wang, Jinyang Gao, Chang Yao
• Visitors: Divy Agrawal, H.V. Jagadish, Dave Maier, Renée Miller, Tamer Özsü, Amit Sheth, Wang-Chien Lee, Wang-Chew Tan, Ju Fan, ++
• Current set of 6-10-10 bosses: Zhaojing Luo, Kaiping Zheng, Jian Dai, Sheng Wang, Shaofeng Cai, Lei Zhu, Qian Lin, Pingcheng Ruan, Qingchao Cai, Anh Dinh, Zhongle Xie, Piaopiao Feng ++
• Ex-Research Fellows and RAs/Engineers/Students: ....
Healthcare AI Success Factors

**Foundational factors:**
- Funding
- Ethics
- Trusted custodian
- Central governance
- Freedom to innovate and implement

**Clinic problems and clinician drivers**

- Data, data, data
- Data scientists
- Scalable, secure hardware
- Clinical trials and Clinicians
- Deployment Platforms/ Productisation

**Clinical problems and clinician drivers**

- Foundational factors:
  - Funding
  - Ethics
  - Trusted custodian
  - Central governance
  - Freedom to innovate and implement