General patient representation from Electronic Health Records

Jean Baptiste Escudié, MD, MSc  @APHP, INRIA
Marc Lelarge, PhD, researcher @INRIA
Gerard Weisbuch, PhD, researcher @ENS
Alaa Saade, PhD, senior ML scientist @Snips
Mina He, data scientist @BNP Paribas
Fajwel Fogel, PhD, chief scientist officer @Sancare
Alice Coucke, PhD, senior ML scientist @Snips

February 14, 2018
Datathon for Intensive Care DAT-ICU event
20-21st of January // PARIS, FRANCE

More than 160 participants, 20 teams...
Autorisation par la Commission Nationale Informatique et Libertés de la constitution d’un entrepôt de données de santé à l’AP-HP

L’autorisation de la CNIL ouvre l’accès aux recherches sur données dépassant le cadre de l’équipe de soins

Dans sa délibération du 19 janvier 2017, la Commission Nationale Informatique et Libertés (CNIL) a autorisé l’Assistance Publique – Hôpitaux de Paris à mettre en œuvre un traitement automatisé de données à caractère personnel ayant pour finalité la constitution de l’entrepôt de données de santé (EDS).

https://www.aphp.fr/contenu/autorisation-par-la-commission-nationale-informatique-et-libertes-de-la-constitution-dun
MIMIC is an openly available dataset developed by the MIT Lab for Computational Physiology, comprising deidentified health data associated with ~40,000 critical care patients. It includes demographics, vital signs, laboratory tests, medications, and more.
https://mimic.physionet.org/
Motivation

- Secondary use of electronic health records (EHRs) promises to advance **clinical research** and **precision medicine**:
  ▶ diseases prediction algorithms
  ▶ personalized prescriptions & treatment recommendations
  ▶ patient similarity & clinical trial recruitment

- The success of these applications *depends on feature selection and data representation*:
  ▶ A **domain expert** designate the patterns to look for in the EHR
    • albuminuria is an important factor in **Chronic kidney disease**
  ▶ A **clinical informatician** determines codes and terminologies
    • type 2 diabetes mellitus [hbA1C > 7.0, 250.00 ICD-9 diagnosis code, mention in the clinical notes]
List of ICD-9 codes

From Wikipedia, the free encyclopedia

The following is a list of codes for International Statistical Classification of Diseases and Related Health Problems[^1] [^2].

- List of ICD-9 codes 001-139: infectious and parasitic diseases
- List of ICD-9 codes 140-239: neoplasms
- List of ICD-9 codes 240-279: endocrine, nutritional and metabolic diseases, and immunity disorders
- List of ICD-9 codes 280-289: diseases of the blood and blood-forming organs
- List of ICD-9 codes 290-319: mental disorders
- List of ICD-9 codes 320-389: diseases of the nervous system and sense organs
- List of ICD-9 codes 390-459: diseases of the circulatory system
- List of ICD-9 codes 460-519: diseases of the respiratory system
- List of ICD-9 codes 520-579: diseases of the digestive system
- List of ICD-9 codes 580-629: diseases of the genitourinary system
- List of ICD-9 codes 630-679: complications of pregnancy, childbirth, and the puerperium
- List of ICD-9 codes 680-709: diseases of the skin and subcutaneous tissue
- List of ICD-9 codes 710-739: diseases of the musculoskeletal system and connective tissue
- List of ICD-9 codes 740-759: congenital anomalies
- List of ICD-9 codes 760-779: certain conditions originating in the perinatal period
- List of ICD-9 codes 780-799: symptoms, signs, and ill-defined conditions
- List of ICD-9 codes 800-999: injury and poisoning
- List of ICD-9 codes E and V codes: external causes of injury and supplemental classification
ICD-9-CM Section I

General Coding Guidelines

- Use both the Alphabetic Index and the Tabular List when locating and assigning a code.
- Locate each term in the Alphabetic Index and verify the code selected in the Tabular List.
- For example, code for the condition “combined hyperlipidemia”

Hyperlipidemia 272.4
  carbohydrate-induced 272.1
  combined 272.2

272.2 Mixed hyperlipidemia
  Broad- or floating-betalipoproteinemia
  Combined hyperlipidemia
ICD-9 diagnosis code
Goal: reduce the dimensionality by building a generalist embedding for each stay with deep learning methods, spanning the whole semantic spectrum of healthcare data.
Our idea: **two models** for each type of data **trained jointly** to get one embedding representation of patient stays

- **Convolutional Neural Network**
  - Vocabulary (~100,000 words)
  - Sequence of 2000 words
  - Train on Mini batch SGD
  - Few hours on 1 GPU
  - Predict ICD-9 code

- **Multi Layer Perceptron**
  - Features (~8,000)
  - Train on Mini batch SGD
  - Few hours on 1 GPU

---

Methods: deep learning architecture

Text data

Structured data

Vocabulary (~100,000 words)

Features (~8,000)
Our idea: two models for each type of data trained jointly to get one embedding representation of patient stays.

- **Convolutional Neural Network**
  - Vocabulary (~100,000 words)
  - Sequence of 2000 words
  - Predict ICD-9 code
  - Training
    - Mini batch SGD
    - few hours on 1 GPU

- **Multi Layer Perceptron**
  - Features (~8,000)
  - EMBEDDING
level 1: **19 chapters**
- neoplasms, infectious and parasitic

**Baseline Random Forests**
PPV: 0.706, Sensitivity: 0.433, F1: 0.537

**PPV: 0.861, Sensitivity: 0.782, F1: 0.820**

level 2: **146 sub-chapters**
- digestive neoplasms, neuroendocrine tumors
- mycoses, venereal diseases

**PPV: 0.740, Sensitivity: 0.462, F1: 0.570**
Appendix: ICD-9 Classification

ICD-9 Classification: chapters

Precision: 0.861, Recall: 0.782, F1: 0.820
Results - Similarity

Patient stay similarity

Unsupervised clustering method (k means)

New borns
Almost all stays in cluster contain code « Certain Conditions Originating In The Perinatal Period »

ARDS
Almost all stays in cluster contain code « Acute Respiratory Distress Syndrom »
Results - Medical concept encoded in the embedding

RESISTANT
ENTEROCOCCUS

v1

ENTEROCOCCUS
Results - Medical concept encoded in the embedding

RESISTANT ENTEROCOCCUS

ENTEROCOCCUS

RESISTANT STAPH AUREUS COAG POS

STAPH AUREUS COAG POS

v1

v2
Results - Medical concept encoded in the embedding

RESISTANT ENTEROCOCCUS

\[ \cos(v1, v2) \sim 0.394 \]

>> 0.04

RESISTANT STAPH AUREUS COAG POS

ENTEROCOCCUS

STAPH AUREUS COAG POS

\[ \cos(v1, v2) \sim 0.394 \]
Results - Medical concept encoded in the embedding

\[ \cos(v_1, v_2) \approx 0.394 \]

\[ \gg 0.04 \]
Results - Medical concept encoded in the embedding

\[ \cos(v_1, v_2) \approx 0.706 \]

- RESISTANT E. COLI
- RESISTANT P. AERUGINOSA

E. COLI

P. AERUGINOSA
Results:

- **Generalist clinical fingerprint** giving promising results on similarity-based **stay clustering**
- Encouraging metrics on **ICD-9 classification** proves validity of the architecture
- Learnt representation correctly **encodes medical semantics** such as antibiotic resistance

Limits:

- **Not enough data** (54k stays) to fully benefit from Deep Learning
  - -> 216k stays analyzed by Google, UCSF, Stanford, UCM (2018)
- Temporality dimension aggregated to 1 data point per stay

Perspectives and applications:

- **Deep Learning** works better with more data
  - **APHP database (millions of stays)**
- **Cohort selection** for clinical studies
- A compact representation **available for any predictive model**
Jean-Baptiste Escudié, MD, MSc
@APHP, INRIA

Alaa Saade, PhD
Machine Learning scientist
@Snips

Mina He,
Data scientist
@BNP Paribas

Marc Lelarge, PhD
Researcher @INRIA
marc.lelarge@inria.fr

Alice Coucke, PhD
Machine Learning scientist
@Snips

Gérard Weisbuch, PhD
Researcher @ENS

Fajwel Fogel, PhD
Chief Scientist Officer
@Sancare
Methods: data extraction

MIMIC III database: EHR data from 54,000 ICU stays

Text data from medical reports
- vocabulary size of 119,000 words
- 2M clinical notes
- first 2000 words per stay

Structured medical data:
- medical procedures, drugs, in/out fluids, microbiology, transfers, demographics,
- SNOMED CT concept extractions from text (exact match)
- numerical or categorical (one hot encoded)
- features selection: 44,000 -> 8,000 most discriminant (lowest p-values chi square)