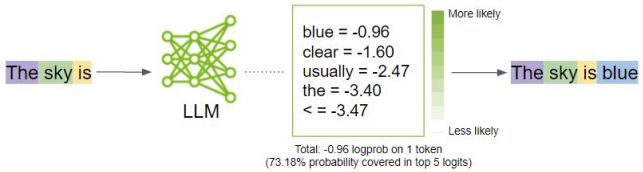
Jak porozumět “jazyku” longitudinálních dat ve zdravotnictví?

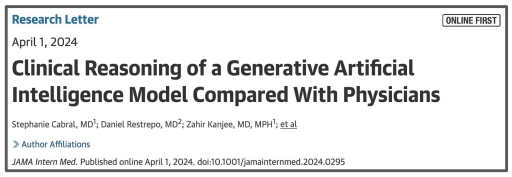


klempond@fbmi.cvut.cz

Ondřej Klempíř, Ph.D.
Katedra biomedicínské informatiky
FBMI ČVUT

Personalizovaná medicína
Analýza řeči
Neurozobrazování

IBM, MSD, DNAnexus


Research Letter
ONLINE FIRST
April 1, 2024
Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians
Stephanie Cabral, MD[1]; Daniel Restrepo, MD[2]; Zahir Kanjee, MD, MPH[1]; et al
Author Affiliations
JAMA Intern Med. Published online April 1, 2024. doi:10.1001/jamainternmed.2024.0295

- ❑ Role “AI” a velká data
  - ❑ UK Biobank
  - ❑ Our Future Health
- ❑ Úspěchy “AI” v medicíně
  - ❑ AlphaFold 2
  - ❑ Radiodiagnostika
  - ❑ GPT
  - ❑ ***Fenotypování***
- ❑ Metody
- ❑ Příklady modelování

Computational (EHR-based) Phenotyping

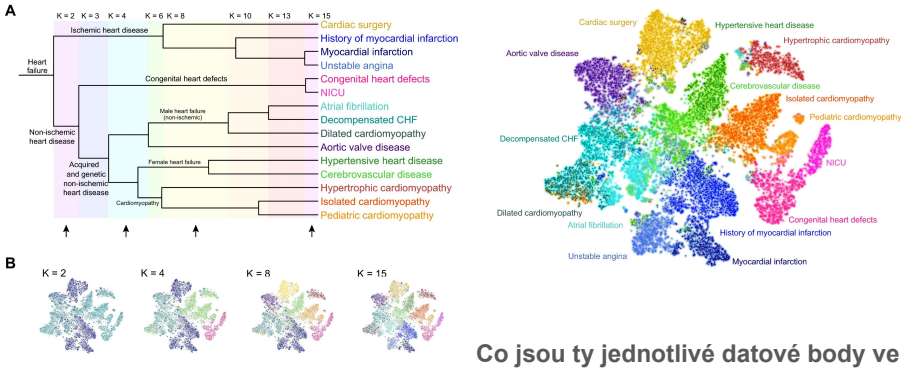
Fenotyp onemocnění = **POZOROVATELNÉ CHARAKTERISTIKY ČI ZNAKY ONEMOCNĚNÍ**

Data-driven phenotype

PROJEVUJE SE SEKUNDÁRNĚ V DATECH

Fenotypy → Symptomy → Břemeno → Čerpání zdrav. péče → Charakteristický “podpis” v datech o vykázané péči

(„datový fenotyp nemoci“)



Co jsou ty jednotlivé datové body ve shlucích?

Nagamine, T., Gillette, B., Pakhomov, A. et al. Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. Sci Rep 10, 21340 (2020).

Generátory velkých dat ve zdravotnictví

- EHR/EMR
- genomika
- PACS
- ***administrativní data***
- laboratoře
- domácí monitoring

<u>Úhly pohledu na longitudinální cestu pacienta</u>

- Co se vyskytuje?
- Co se vyskytuje a kolikrát?
- Jaká je sekvence?
- Jaká je sekvence a časová delta mezi událostmi?

Code:	57243	87433	87223	87223	96167
ICD10 Code:	J44.8	C45.2	C34.3	C34.3	J44.8
Day:	1	2	5	5	8

temporal dimension

Metody pro stanovení/hledání fenotypů

Nekontextové

- Pravidlové systémy
- Velká řídká matice (one-hot encoding)
  - Binární
  - Četnost (kolikrát)
- tf-idf
  - částečně kontext pomocí n-gramů

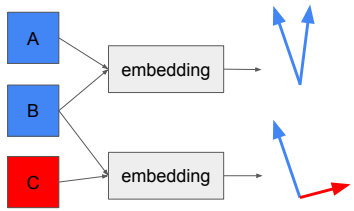
vs.

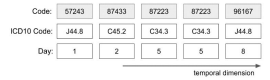
S pamětí (kontextem)

- Skryté Markovovy modely (HMM)
- word2vec
- doc2vec
- LSTM
- Transformery
  - BERT
  - GPT

Vektorová reprezentace sekvence

- S číselnými vektory se “dobře” počítá
  - Lineární algebra
  - Vhodný vstup pro neuronovou síť (hluboké učení)
  - Urychlení výpočtu na grafické kartě
  - Paralelizace
- Vektory umí zachovat různé vlastnosti





- Tohle mohou být i objekty jiného typu
  - Medicínské obrazy
  - Audio
  - EKG
  - …
  - multimodální…

f()

[6.3, 2.5, 6.2, 4.4 …]

- je kompaktní
- uchovává kontext
- dimenze bývá důležitá

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Hledání fenotypů “s učitelem” vs. “bez učitele”

Známe rozdělení do skupin

- Překlasifikování chybně stanovených diagnóz
- Subtypování a kvantifikace podobnosti mezi diagnózami
- Stanovení pravděpodobnosti výskytu prodromální fáze
- Predikce (identifikace rizikových faktorů) pro určitou kondici, např. COVID

Neznáme přiřazení

- Detekce shluků, které vykazují podobné vlastnosti
- Detekce netypických cest pacienta

Datasety, nástroje

EHRAPY

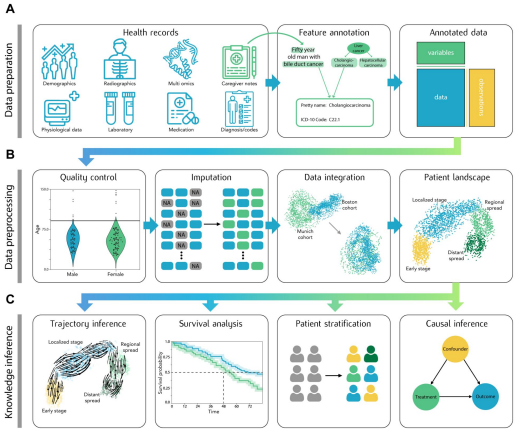
- Longitudinální data

BERTopic

MIMIC-III

Simulovaná data

- Atributy a stejná struktura
- Pokročilé metody pro generování simulovaných dat při zachování charakteristik

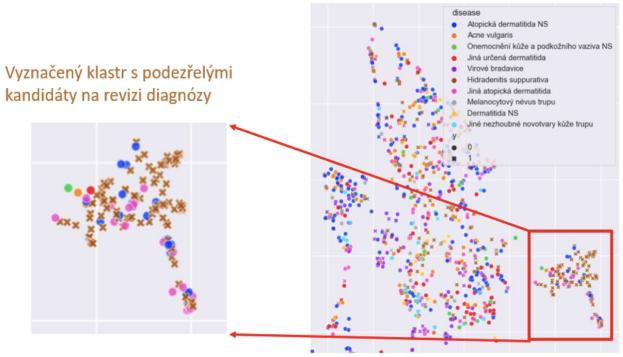


https://ehrapy.readthedocs.io/en/development/tutorials/notebooks/ehrapy_introduction.html

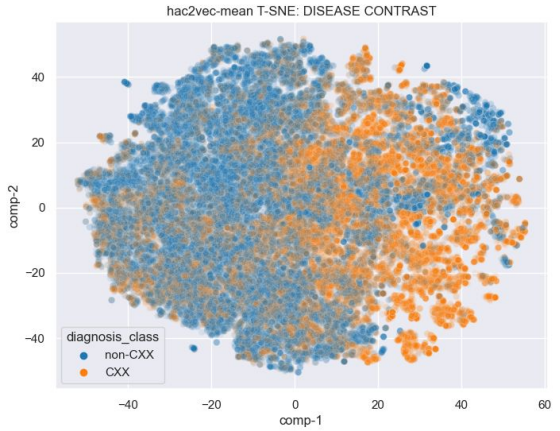
Zpracování jazyka - transformerové předtrénované modely

- BERT (baseline)
- ClinicalBERT
  - Předtrénovaný na MIMIC-III
- PubmedBERT
  - Předtrénovaný na PubMed
- BioBERT
  - Předtrénovaný na PubMed abstraktech a PMC full-textech
- BlueBERT
  - Předtrénovaný na kombinaci MIMIC-III a PubMed

“BERTopic cluster finder” = over-represented



word2vec embedding - obarven podle nejčastěji vykazované dermatologické choroby



Ales Tichopad, Gleb Donin and Ondrej Klempir. Identifying Hidden Patterns from Health Administrative Claims by means of “HAC2Vec” Embedding. EHB 2023.

- ❏ Fenotypování
- ❏ Vybrané metody zpracování jazyka
- ❏ Co je to embedding
- ❏ Příklady modelů na longitudinálních datech
- ❏ Co dál? LLMs?
  - ❏ ChatGPT vs. lokální

FAKULTA BIOMEDICÍNSKÉHO INŽENÝRSTVÍ ČVUT V PRAZE

DĚKUJI MOC, ŽE JSTE PŘIŠLI!