POTENTIAL USE OF ARTIFICIAL INTELLIGENCE TO ANALYZE DATA EXTRACTED FROM ELECTRONIC HEALTH RECORDS FOR DECISION ANALYTIC MODELS

Sandor Kovacs[1], Gabor Vincze[1], Dalma Erdosi[1], Antal Zemplenyi PhD[1]

[1]University of Pécs, Chancellery, Health Technology Assessment Center, Pécs, Hungary

Background

Data collected in electronic medical records (EMR) or electronic health records (EHR) are subject to various limitations for the assessment of clinical effectiveness and cost-effectiveness, as they are sparse, highly fragmented and require breakdown of temporal events to provide valuable insight in the history of the disease and the treatment.

Objective

Our aim was to review the methods used to process and analyze routinely collected EMR/EHR data to create model input variables for decision-analytic modeling in the field of oncology.

Methodology

A systematic literature review was conducted and reported in compliance with the PRISMA Statement.

The search resulted in 112 publications, which amount was reduced to 38 during the title/abstract screening. Finally, 28 publications were retrieved for detailed review after the full text screening. *[Fig. 1]*

The relevant publications were analyzed and categorized, as **information was systematically extracted,** based on the following criteria: the aim, methods, results and conclusions of research, diseases in question, patient population, patient selection and exclusion, type of collected data, data analysis methods, applied software and number of citations.

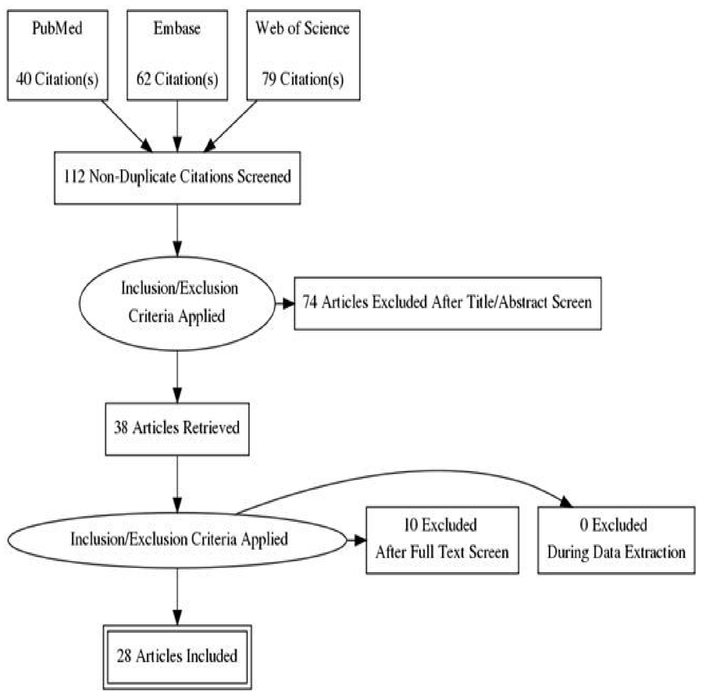

Figure 1

Results

The **number of relevant articles** in this topic has shown a growing tendency during the past 5 years *[Fig. 2]*.

Concerning their **purpose of data processing**, the publications were classified as follows: determining parameters to describe health states and assess transition probability, providing clinical decision support, predicting mortality or risk and focusing exclusively on feature selection, standardization, and temporal sequence alignment. *[Fig. 3]*.

Our results suggest, that **the most common types of data sources** were final reports, text reports of imaging diagnostics results, pathological reports, oncoteam reports, outpatient discharge reports, nursing diaries, referrals and prescriptions. The **most frequently collected types of data** were patient demographics (age, gender, ethnicity), BMI, medication, interventions, general lab results, specific variables of cancer types (i.e. HER-2, PSA, etc.), TNM and staging.

Some of the studies involved multiple **types of cancer** (10), while a larger number of them foucsed on one specific type, such as breast (8), colorectal (5), lung (2), prostate (2) or ovarian cancer (1).

The **population of patients** in the studies varied from 20 to 1.2 million, the population median was 4117.

The use of machine learning (as a generic concept) was present in all of the researches we studied, while more resource-intensive technologies such as neural networks and deep learning were less frequent. For the unstructured data natural language processing (NLP) techniques were used in almost every case, while data pipelines were created to overcome the unique structure of EHR databases. In most researches, **validation** was conducted to test feasibility of the applied methods *[Table 1]*.

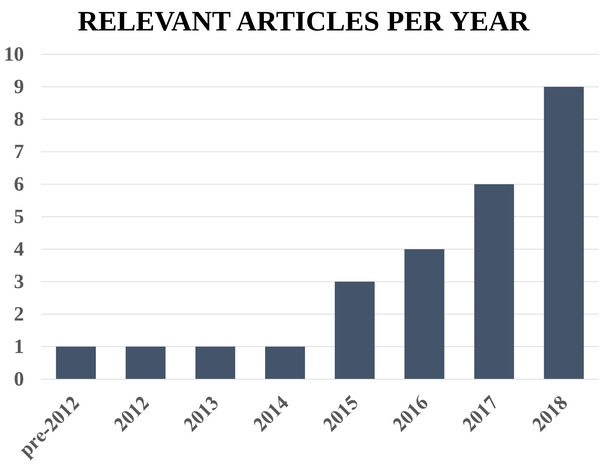

Figure 2

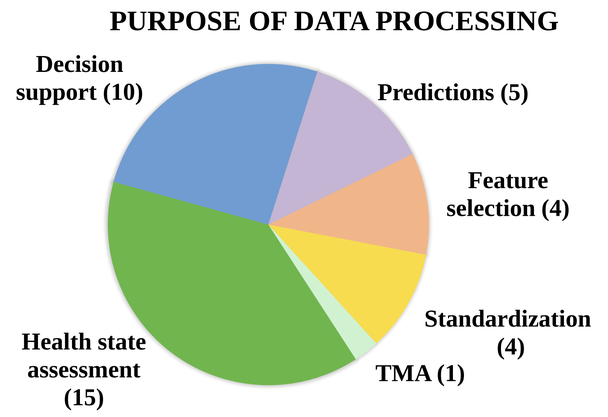

Figure 3

Publication	Country	Applied AI methods	Validation	Times cited in Google Scholar
Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records	USA	**multi-layer neural network, deep learning**: random forest		373
Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning	UK	**machine learning:** novel algorithm, the Semi-supervised Set Covering Machine' (S3CM)		64
Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management	USA	**machine learning:** Semi-supervied, support vector machine (SVM), Laplacian SVM	✓	52
Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry	AUS	**machine learning:** SVM	✓	37
Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text	USA	**machine learning:** Subgraph augmented non-negative tensor factorization (SANTF)	✓	31
Support Vector Feature Selection for Early Detection of Anastomosis Leakage From Bag-of-Words in Electronic Health Records	NO	**machine learning:** Bag-of-Words model, SVM	✓	28
Using machine learning to parse breast pathology reports	USA	**machine learning:** Learning curve analysis, non-linear classification	✓	21
Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer	NL	**machine learning:** NLP**:** Bag-of-Words model, topic modeling	✓	15
Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records	NL	**machine learning**: CART, random forest, logistic regeression	✓	13
Predictive Modeling of Therapy Decisions in Metastatic Breast Cancer with Recurrent Neural Network Encoder and Multinomial Hierarchical Regression Decoder	GER	**deep learning**: encoder-decoder framework, NLP	✓	11
On the Advantage of Using Dedicated Data Mining Techniques to Predict Colorectal Cancer	NL	**machine learning**: CART, logistic regression, RF	✓	7
Machine Learning Methods to Extract Documentation of Breast Cancer Symptoms From Electronic Health Records	USA	**machine learning:** CRF model, NLP	✓	7
Development and Application of a Machine Learning Approach to Assess Short-term Mortality Risk Among Patients With Cancer Starting Chemotherapy	USA	**machine learning**: Boosted classifier, NLP	✓	4
Artificial Intelligence Learning Semantics via External Resources for Classifying Diagnosis Codes in Discharge Notes	TW	**deep learning:** random forest, SVM, feature matrix, NLP		4
Machine learning to parse breast pathology reports in Chinese	CN	**machine learning**		4
Temporal Needleman–Wunsch	USA	**Needleman-Wunsch algorythm**	✓	3
Predicting Sentinel Node Status in Melanoma from a Real-World EHR Dataset	USA	**deep learning:** logistic regression, decision tree, random forest model	✓	2
Using natural language processing and machine learning to identify breast cancer local recurrence	USA	**machine learning:** SVM, NLP	✓	2
Applied Informatics Decision Support Tool for Mortality Predictions in Patients With Cancer	USA	**machine learning:** self-developed software for decision support	✓	2
Automated Extraction and Classification of Cancer Stage Mentions from Unstructured Text Fields in a Central Cancer Registry	USA	**machine learning: NLP:** clinical Text Analysis and Knowledge Extraction System (cTAKES**)**		2
Prediction of transition sequence of diseases' severity levels using clinical datasets with data mining approaches	USA	**machine learning:** Phyton-based data mining tool	✓	1
Predicting Emergency Visits and Hospital Admissions During Radiation and Chemoradiation: An Internally Validated Pretreatment Machine Learning Algorithm	USA	**deep learning:** random forest, logistic regression, SVM	✓	1
Applying Artificial Intelligence to Address the Knowledge Gaps in Cancer Care	USA	**machine learning:** Oncology Expert Advisor		1
Artificial neural networks are highly predictive for hepatocellular carcinoma in patients with cirrhosis	USA	**neural network, TM-GTP, machine learning**	✓	0
Natural language processing and recurrent network models for identifying genomic mutation-associated cancer treatment change from patient progress notes.	USA	**neural network, deep learning:** NLP: named entity recognition		0
Identifying Cases of Metastatic Prostate Cancer Using Machine Learning on Electronic Health Records	USA	**machine learning:** APHRODITE package: feature vector, LASSO	✓	0
Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches	CN	**neural network, deep learning, NLP**	✓	0
Breast Cancer Prognosis Using a Machine Learning Approach	I	**machine learning:** kernel-based learning platform (KeLP), multiple kernel learning (MKL), SVM		0

Table 1

Conclusion

On one hand, our analysis suggests that the use of AI in the analysis of oncological EHRs is becoming more and more common. Based on the 28 articles we reviewed the most commonly reported difficulties related to EMR like incompleteness, sparseness and high dimensionality can be overcome by the application of feature selection and a well-defined data pipeline. While the precision and recall of the subsequently generated predictive models employing supervised or unsupervised machine and deep learning processes are way beyond the expectations compared to traditional modelling. On the other hand, none of these analytic methods had any clinical implementation at the time of our research. In spite of the outstanding academic results provided by these processes and the increasing demand for using EHRs for real-world data analysis and for model input generation the practical value and the applicability of AI in healthcare decision support is still ambiguous.

PCN 494 **W 6**

Corresponding author:

Sandor Kovacs, M.Sc.
University of Pécs, Chancellery
Health Technology Assessment Centre
E-mail: kovacs.sandor@pte.hu

Scan me

PÉCSI TUDOMÁNYEGYETEM
UNIVERSITY OF PÉCS

ISPOR
Improving healthcare decisions

SZÉCHENYI 2020
HUNGARIAN GOVERNMENT
European Union
European Social Fund
INVESTING IN YOUR FUTURE