Zeng, Zexian, et al. "**Natural language processing for EHR-based computational phenotyping.**" IEEE/ACM transactions on computational biology and bioinformatics 16.1 (2019): 139-153.

이혜수

SNU
Data**b**ase **S**ystems **L**ab

Contents

1. Purpose of the article
2. NLP-based Computational Phenotyping
3. Motivations for computational phenotyping
4. Applications of computational phenotyping
5. NLP Methods
6. Efforts to improve current methods

1. Purpose of the article

- Summarize the state-of-the-art NLP methods for computational phenotyping.
- Describe the combinations of data modalities, feature learning, and relation extraction that have been used to aid computational phenotyping.
- Discuss challenges and opportunities to NLP methods for computational phenotyping.
- Highlight a few promising future directions.

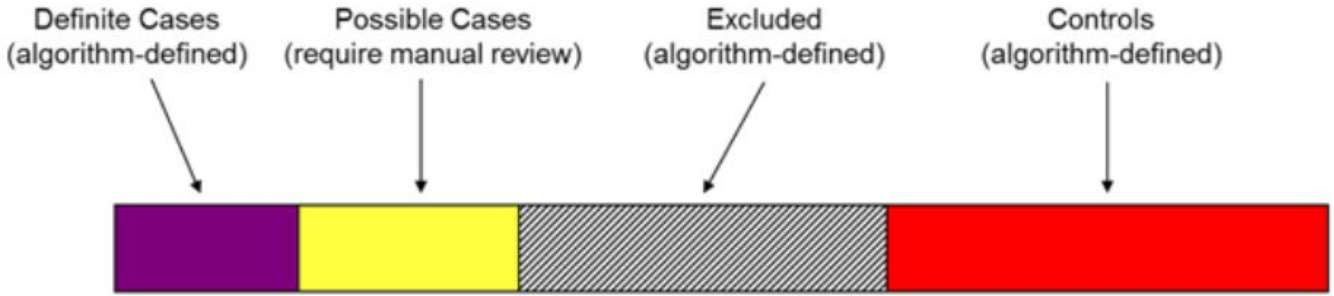
2. NLP-based Computational Phenotyping

- **PHENOTYPE** : an expression of the characteristics that result from genotype variations and an organism's interactions with its environment.
  - physical appearances (e.g., height, weight, BMI), biochemical processes, or behaviors.
  - In the medical domain, phenotypes are often summarized by experts on the basis of clinical observations.

- Computational phenotyping aims to automatically mine or predict clinically significant, or scientifically meaningful, phenotypes from structured EHR data or unstructured clinical narratives.
  - **Clinical narratives** ∋ {clinicians' notes, observations, referring letters, specialists' reports, discharge summaries, a record of communications between doctors and patients}
  - May contain {patients' medical history, diagnoses, medications, immunizations, allergies, radiology images, laboratory test results}

3. Motivations for Computational Phenotyping

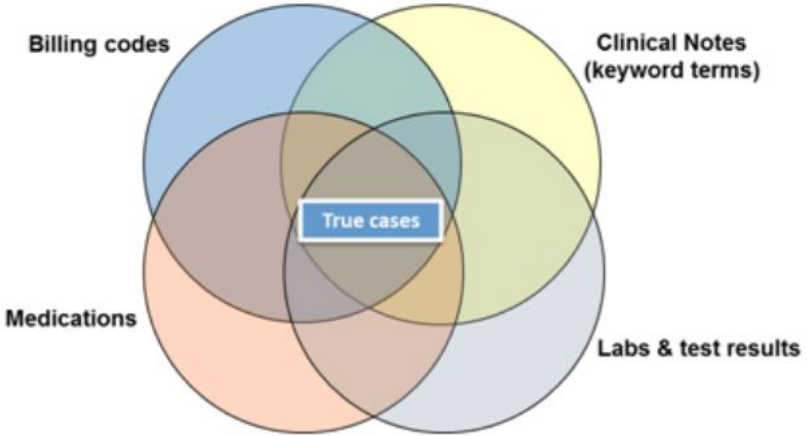
- In EHR, ICD-9 codes are mainly recorded for administrative purposes and are influenced by billing requirements. Consequently, these codes **do not always accurately reflect a patient's underlying physiology**.
  - Outpatient billing limited to 4 diagnoses/visit
  - It takes too long to find the unknown ICD9.
  - May bill for a different condition if it pays for a given treatment.*



*Kirby Jacqueline. Exploring Data and Cohort Discovery in the Synthetic Derivative [PowerPoint slides]. Retrieved May 9, 2019, from https://slideplayer.com/slide/8161934/

3. Motivations for Computational Phenotyping

- Not all patient information is well documented in structured data, such as clinicians' observations and insights. Using structured data alone for phenotype identification often results in low performance.



- The richer features of relations between medical concepts **enables greater expressive power when encoding patient status**, compared to terms and keywords.

*Kirby Jacqueline. Exploring Data and Cohort Discovery in the Synthetic Derivative [PowerPoint slides]. Retrieved May 9, 2019, from https://slideplayer.com/slide/8161934/

4. Applications of computational phenotyping

NLP-based computational phenotyping has numerous applications including :

- **4-1. Diagnosis Categorization,**
- **4-2. Novel Phenotype Discovery,**
- **Clinical Trial Screening,**
- **Pharmacogenomics,**
- **Drug-Drug Interaction(DDI),**
- **Adverse Drug Event (ADE) Detection,**
- **Genome-wide and Phenome-wide Association Studies**

4-1. Diagnosis Categorization

- Enabling the automated and efficient identification of patient cohorts for secondary analysis.
- **Disease identification, disease subtyping, subsequent event detection**
  - suspected tuberculosis (TB) [32], [33]
  - colorectal cancer [34]
  - rheumatoid arthritis [35]
  - diabetes [36]
  - heart failure [37], [38]
  - neuropsychiatric disorders [39].

  - lung cancer stage evaluation [40]
  - breast cancer recurrence detection [41]
  - cancer metastases detection [42]

4-1. Diagnosis Categorization

- Enabling the automated and efficient identification of patient cohorts for secondary analysis.
- **Disease identification, disease subtyping, subsequent event detection**
  - suspected tuberculosis (TB) [32], [33]
  - colorectal cancer [34]
  - rheumatoid arthritis [35]
  - diabetes [36]
  - heart failure [37], [38]
  - neuropsychiatric disorders [39].
  - lung cancer stage evaluation [40]
  - breast cancer recurrence detection [41]
  - cancer metastases detection [42]

Xu, Hua, et al. "Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases." *AMIA Annual Symposium Proceedings*. Vol. 2011. American Medical Informatics Association, 2011.

: Concept Recognition [Extract -> Classify]

- Rule-based
- ML
  - Random Forest (RF)
  - Ripper
  - Support Vector Machine (SVM),
  - Logistic Regression (LR)}
- **Lack of Concept linkage**

4-1. Diagnosis Categorization

- Enabling the automated and efficient identification of patient cohorts for secondary analysis.
- **Disease identification, disease subtyping, subsequent event detection**
  - suspected tuberculosis (TB) [32], [33]
  - colorectal cancer [34]
  - rheumatoid arthritis [35]
  - diabetes [36]
  - heart failure [37], [38]
  - neuropsychiatric disorders [39].

  - lung cancer stage evaluation [40]
  - breast cancer recurrence detection [41]
  - cancer metastases detection [42]

Panahiazar, Maryam, et al. "Using EHRs and machine learning for heart failure survival analysis." *Studies in health technology and informatics* 216 (2015): 40.

: Survival Risk Prediction

- Random Forest (RF)
- Logistic Regression
- Support Vector Machine (SVM),
- Decision Tree
- AdaBoost : 가중치를 부여한 Weak Classifier를 모아서 최종적인 Strong Classifier를 구성

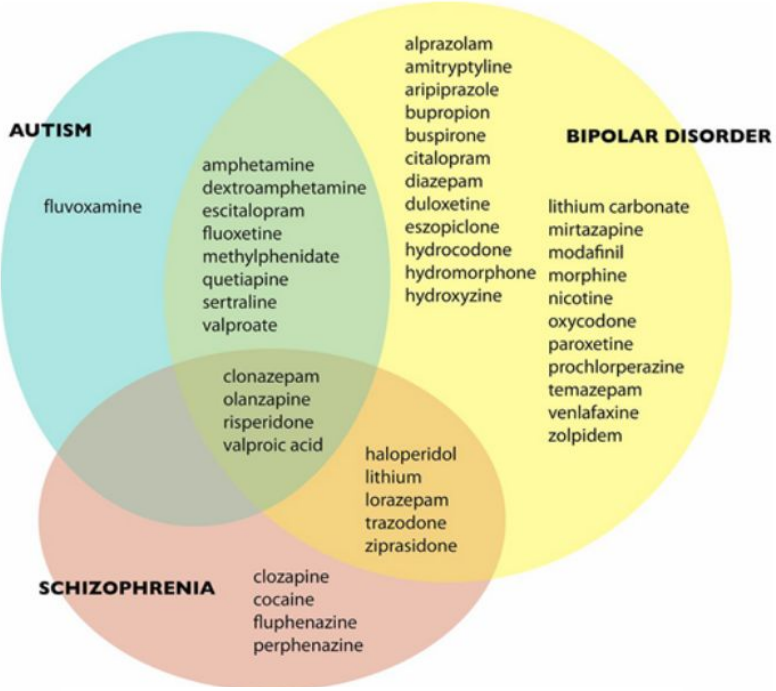
4-1. Diagnosis Categorization

- Enabling the automated and efficient identification of patient cohorts for secondary analysis.
- **Disease identification, disease subtyping, subsequent event detection**
  - suspected tuberculosis (TB) [32], [33]
  - colorectal cancer [34]
  - rheumatoid arthritis [35]
  - diabetes [36]
  - heart failure [37], [38]
  - neuropsychiatric disorders [39].
  - lung cancer stage evaluation [40]
  - breast cancer recurrence detection [41]
  - cancer metastases detection [42]

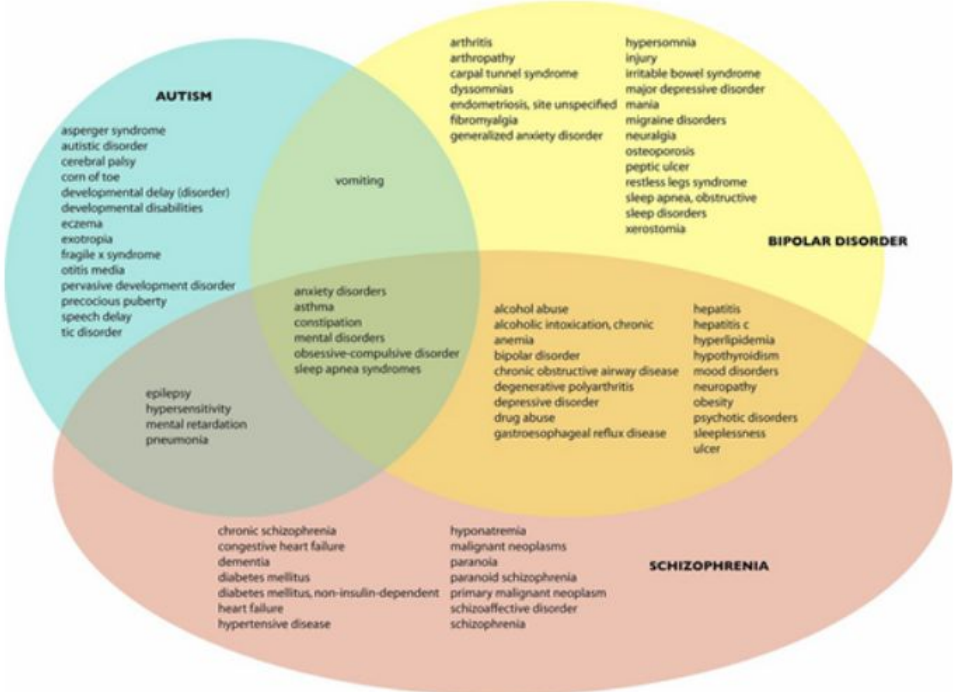
Lyalina, Svetlana, et al. "Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records." *Journal of the American Medical Informatics Association* 20.e2 (2013): e297-e305.

: Find boundaries of autism, bipolar disorder, and schizophrenia

- Search with UMLS codes
- Find associations among codes
- Dimension Reduction to understand phenotypic variation
- Network Analysis



Psychoactive drug terms for each disease, and their overlap



Phenotypes for each disease, and their overlap

*Lyalina, Svetlana, et al. "Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records." *Journal of the American Medical Informatics Association* 20.e2 (2013): e297-e305.

4-2. Novel Phenotype Discovery

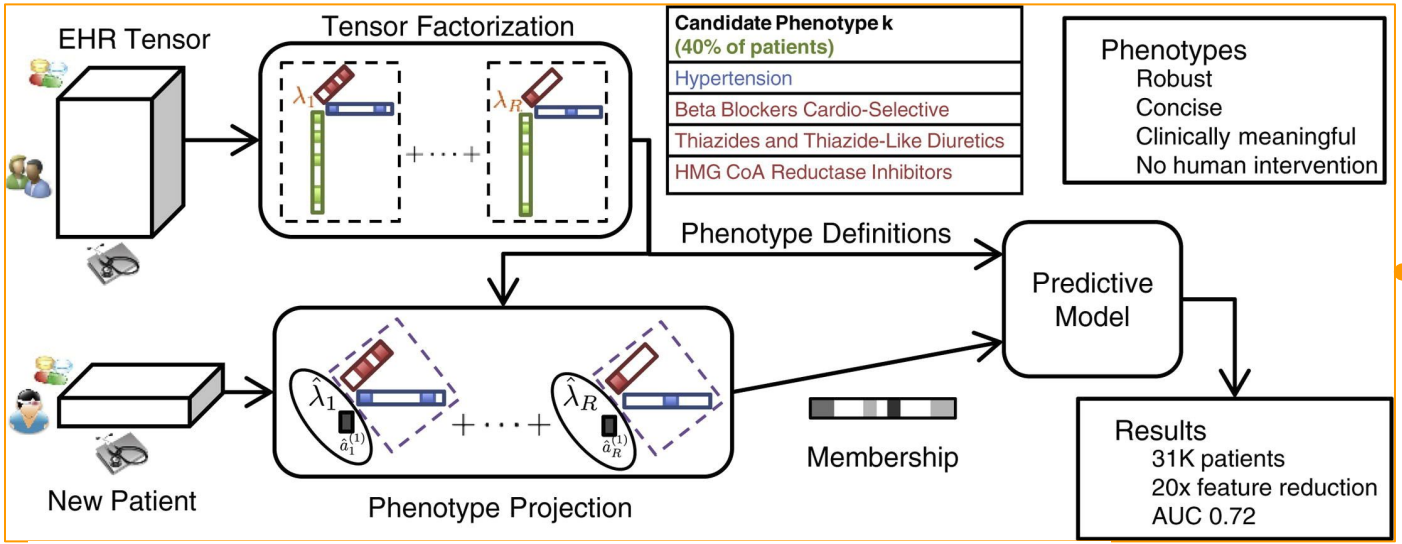
Traditionally, classified by a set of criteria developed by domain experts.

Semi-supervised or **Unsupervised methods** can detect traits based on intrinsic data patterns with moderate or minimal expert guidance, which may promote the discovery of novel phenotypes or sub-phenotypes.

- Tensor factorization on medication orders to generate phenotypes without supervision [46]
- Clustered patients with preserved ejection fraction [49]

4-2. Novel Phenotype Discovery

- Tensor factorization on medication orders to generate phenotypes without supervision [46]
- Clustered patients with preserved ejection fraction [49]



*Ho, Joyce C., et al. "Limestone: High-throughput candidate phenotype generation via tensor factorization." *Journal of biomedical informatics* 52 (2014): 199-211.

5. NLP Methods

5-1. Keyword Search and Rule-Based
5-2. Supervised Statistical Learning
5-3. Unsupervised Learning
5-4. Deep Learning

Study	Assertion	Concept Extraction/Concept Mapping	Data Source	Feature Generation
Aramaki et al. [96]	NA	Self-defined keywords	Narrative	Similarity score between sentences
Chase et al. [92]	NA	MedLEE	Narrative	50 buckets representing pools of synonymous UMLS terms
Chen et al. [99]	NA	KMCI, SecTag, MedLEE, MedEx	Narrative, ICD-9, CPT	ICD-9, CPT, CUIs
DeLisle et al. [101]	Customized rules, NegEx	Examined UMLS-supplied lexical variants/semantic types	Narrative, ICD-9, vital signs and orders for tests, imaging, and medications	186 UMLS associated with phenotype
Gehrmann et al. [104]	cTAKES	cTAKES	Discharge summary	Concepts from cTAKES were transformed to continuous features using the TF-IDF
Pineda et al. [112]	ConText	Topaz pipeline, map to UMLS	Narrative, lab test	Selected UMLS concepts and two lab test concepts

kNN

String matching,
Naive Bayes Clustering

SVM & Active Learning

Logistic Regression

CNN, n-grams,
String matching

ANN, Naive Bayes
and more...

5-1. Keyword Search and Rule-Based

- Looks for keywords, derivations of those keywords or a combination of keywords to extract phenotypes.
- More susceptible to **low accuracy** due to simplicity of features.
- To improve model performance, **supplementary rules** (or other sophisticated criteria) have been added.
- Developing rules is laborious, time-consuming and requires expert knowledge.

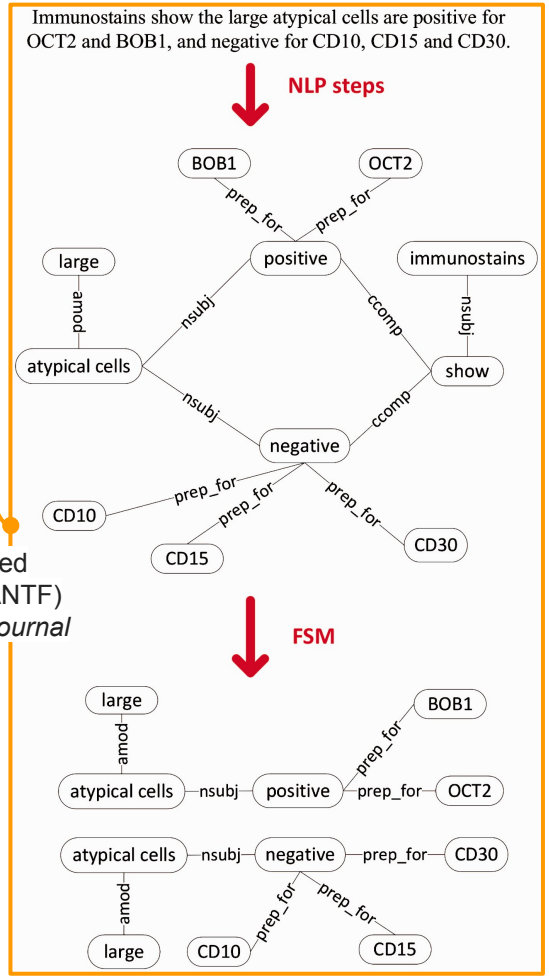
5-2. Supervised Statistical Learning

- To improve upon accuracy and scalability while decreasing domain expert involvement
  - Logistic Regression
  - Bayesian Networks : works well with high-dimensional features
  - Support Vector Machines
  - Decision Trees
  - Random Forests
  - Conditional Random Field

- **Semi-supervised** algorithms : when we have both labeled and unlabeled samples.

5-3. Unsupervised Learning

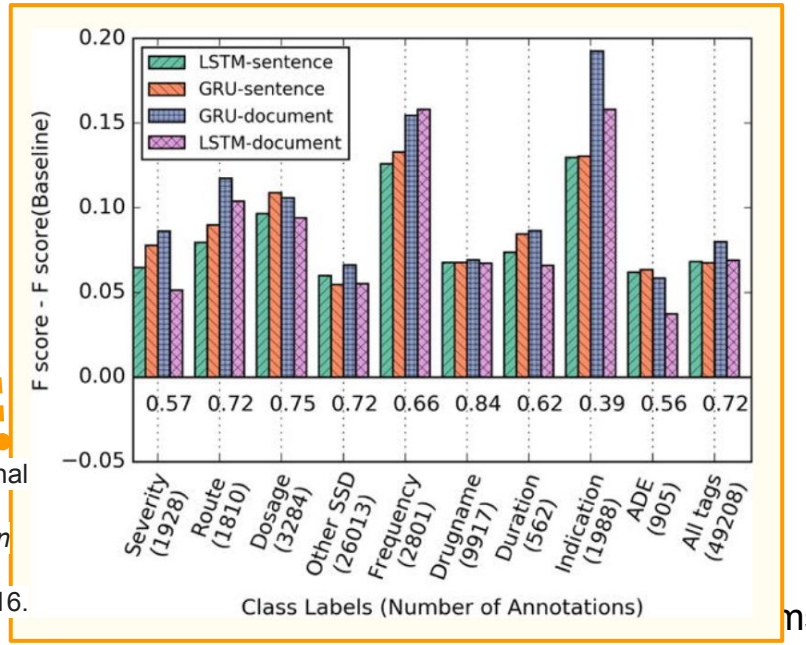
- Automatically classify phenotypes without extra annotations by experts.
  - Subgraph Augmented Non-negative Tensor Factorization (SANTF)
  - Sparse Non-negative Tensor Factorization
  - Hierarchical Clustering
  - Kernel-based Pattern Clustering



Luo, Yuan, et al. "Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text." *Journal of the American Medical Informatics Association* 22.5 (2015): 1009-1019.

5-4. Deep Learning

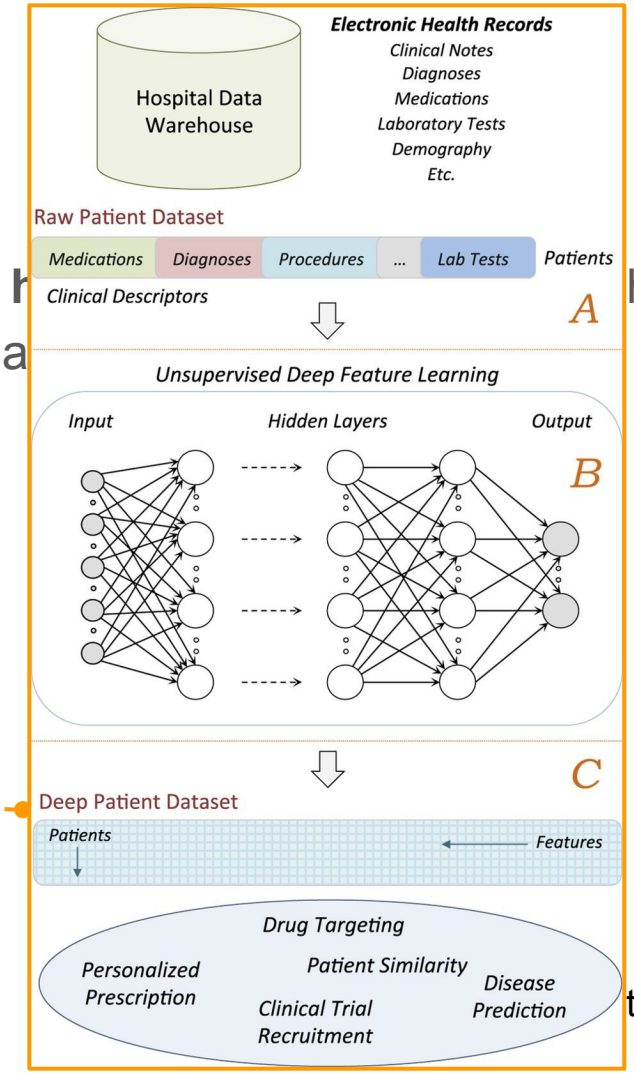
- Good at finding intricate structures in **high-dimensional data** and have demonstrated good performance in natural language.
  - Convolutional Neural Network
  - Recurrent Neural Network
    - LSTM
    - Bidirectional LSTM
    - GRUs
  - Autoencoders



Jagannatha, Abhyuday N., and Hong Yu. "Bidirectional RNN for medical event detection in electronic health records." *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. Vol. 2016. NIH Public Access, 2016.

5-4. Deep Learning

- Good at finding intricate structures in h have demonstrated good performance in na
  - Convolutional Neural Network
  - Recurrent Neural Network
    - LSTM
    - Bidirectional LSTM
    - GRUs
  - Autoencoders : unsupervised pre-training



Miotto, Riccardo, et al. "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records." *Scientific reports* 6 (2016): 26094.

6. Efforts to improve current methods

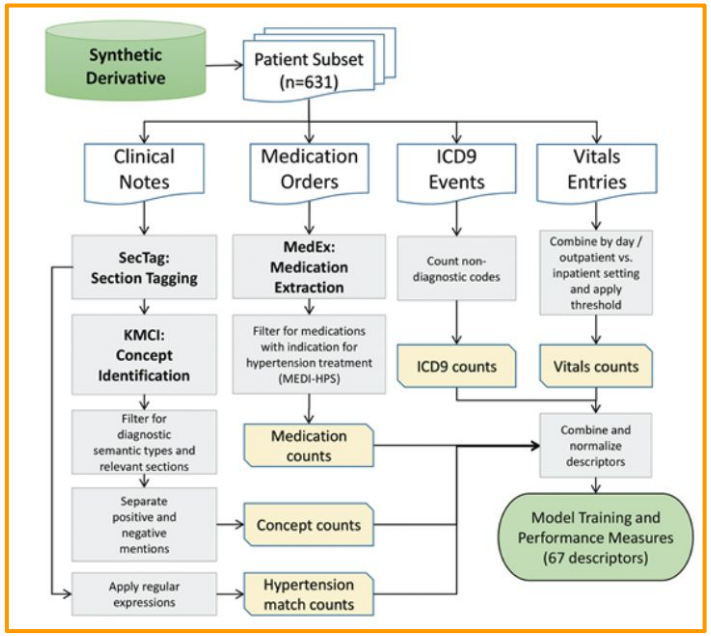
6-1. **Model Comparison** : To explore model performances

a. Algorithm performance differs based on data sources, features, training data sizes, and target phenotypes

6. Efforts to improve current methods

6-2. **Combining Multiple Data Modalities** : Adding heterogeneous data has the benefit of providing complementary perspectives for computational phenotyping models

a. Teixeira et al. [116] showed that **model performance increases for hypertension prediction with the number of data resources** regardless of the method used.



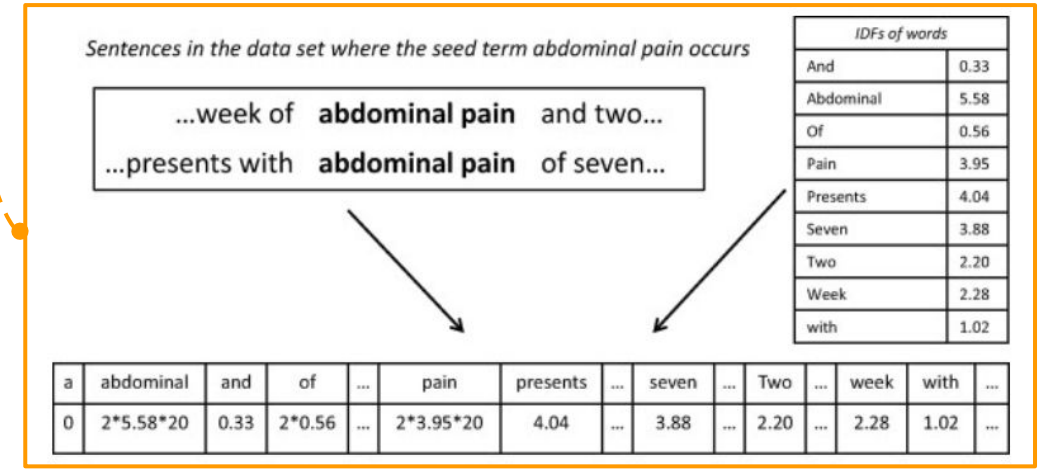
Teixeira, Pedro L., et al. "Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals." *Journal of the American Medical Informatics Association* 24.1 (2016): 162-171.

6. Efforts to improve current methods

6-3. Entity Recognition and Relation Extraction

a. Zhang et al. [181] have applied an unsupervised approach to extract named entities from biomedical text by detecting entity boundaries and classifying entities without pre-defined rules or annotated data.

b. Assume that same class entities tend to have similar vocabulary and context.



Zhang, Shaodian, and Noémie Elhadad. "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts." *Journal of biomedical informatics* 46.6 (2013): 1088-1098.

7. Future Work

7-1. Information Heterogeneity in Clinical Narratives

a. Due to the variance in physicians' expertise and behaviors

b. Clinical narratives are **ungrammatical**, **incomplete** and contain a large number of **abbreviations** and **acronyms**.

7-2. Model Generalizability

a. Expand generalizable algorithms to mine multiple diseases from different narratives.

7-3. Model Interpretability

a. Deep learning models tend to **lack interpretability**.

7-4. Characterizing the Context of Computational Phenotyping

a. Generalized relation and event extraction are expected to be promising.

b. To this end, **graph methods** are a promising class of algorithms.