A summary of EHR-based phenotyping article annotation

Siyue Yang, Jessica Gronsbell

05/18/2022

Contents

1 Overview **2**

1.1 Traditional ML method 2

1.2 DL method 3

2 Phenotype **4**

3 Data source **5**

3.1 Summary 5

3.2 Structured and unstructured data type 5

3.3 Openly-available data 8

4 NLP software **9**

5 Emebddings **9**

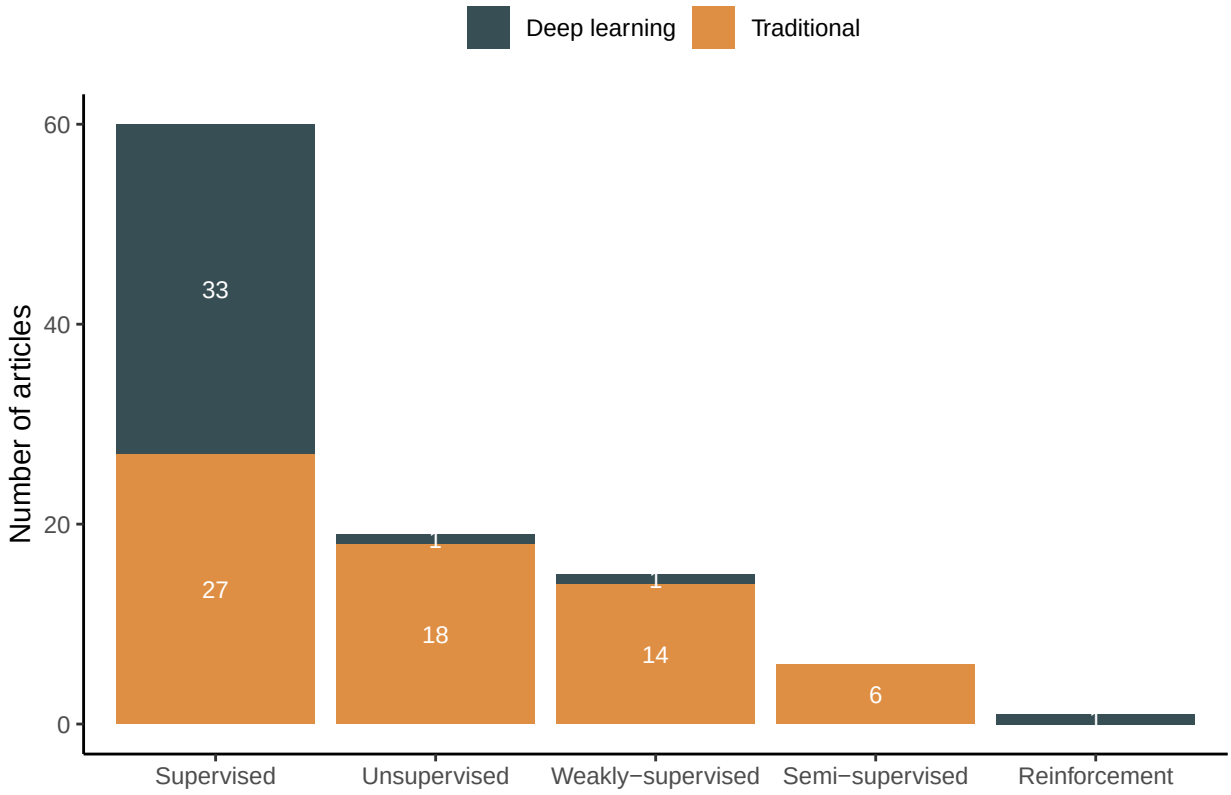
6 Validation and comparison **10**

6.1 Traditonal supervised ML vs. rule-based 10

6.2 Deep supervised ML vs. supervised 11

7 Reporting **11**

1 Overview



1.1 Traditional ML method

Table 1: Common traditional machine learning methods (Count > 1)

ML_type	Traditional_ML_method_unnested	Count
Supervised	Random forest	14
Supervised	Logistic regression	11
Supervised	SVM	11
Supervised	L1 logistic regression	8
Supervised	Decision trees	4
Supervised	XGBoost	4
Supervised	Naive Bayes	3
Unsupervised	LDA	5
Unsupervised	Hierarchical clustering	4
Unsupervised	K-means	4
Weakly-supervised	PheNorm	3
Weakly-supervised	MAP	2
Weakly-supervised	Random forest	2

```
## [1] "There are 18 papers using multiple traditional machine learning methods"
```

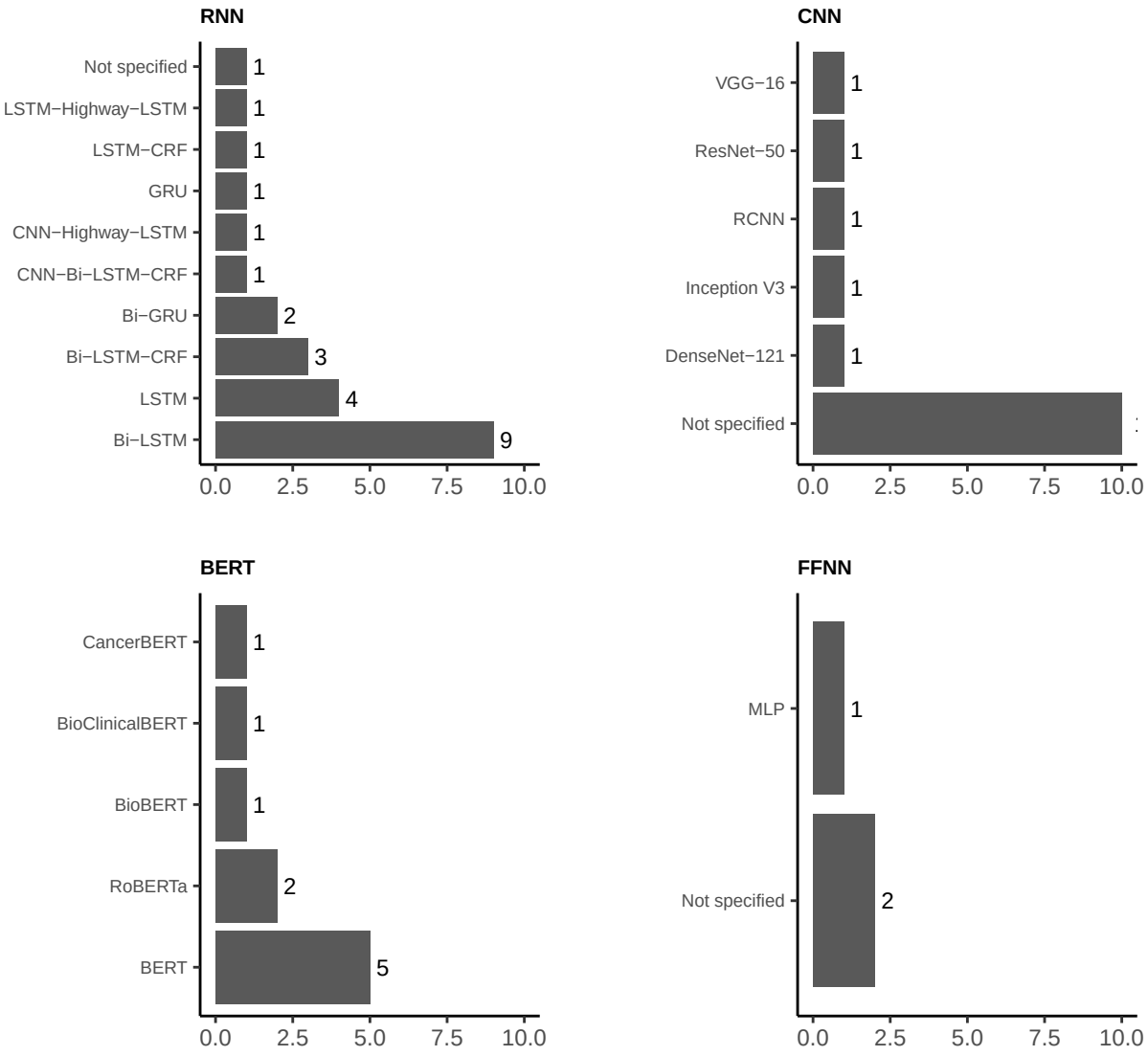
1.2 DL method

Table 2: Deep learning methods

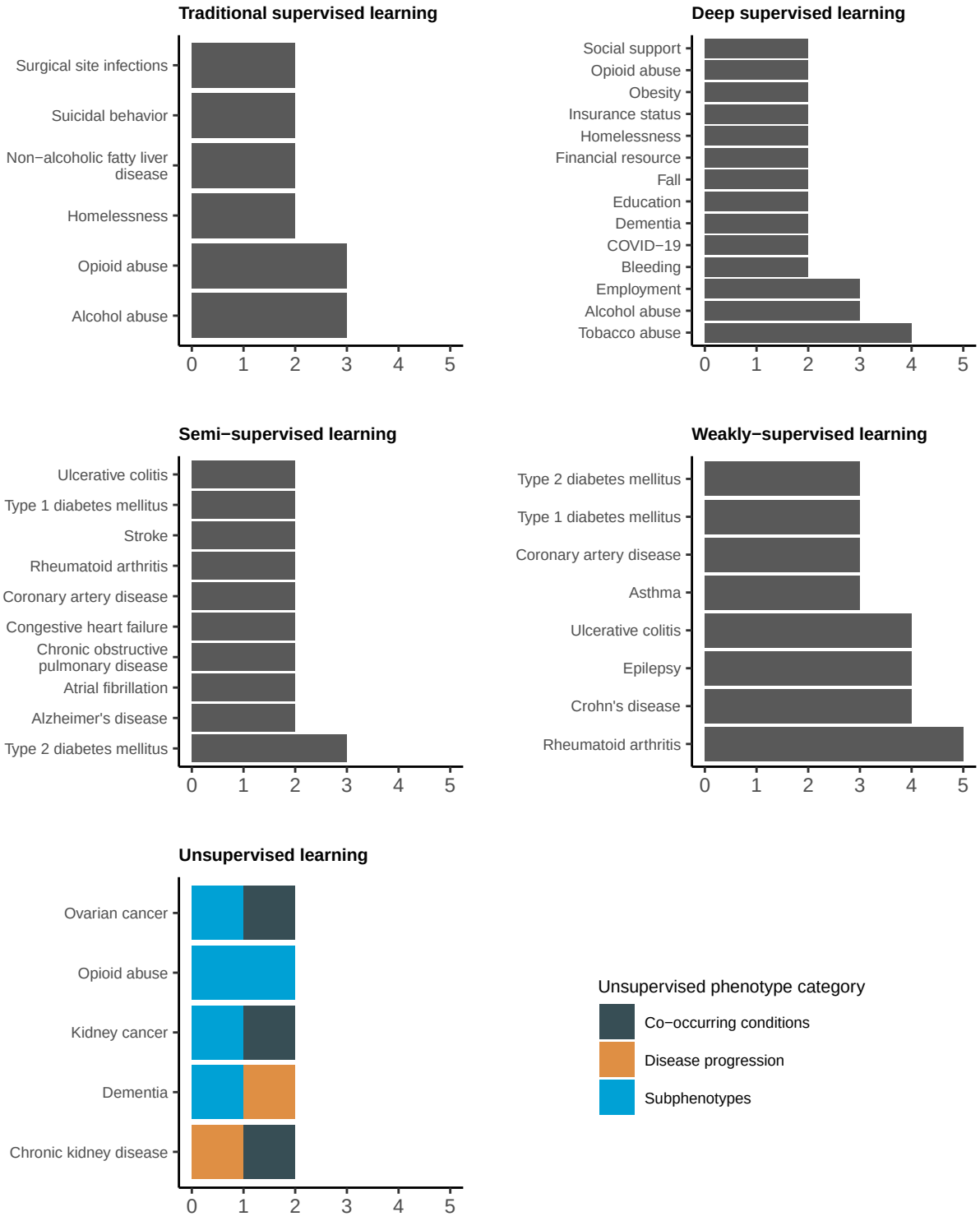
DL_method_unnested	ML_type	Count
BERT	Supervised	7
CNN	Supervised	12
FFNN	Supervised	3
RNN	Supervised	18

```
## [1] "There are 5 papers using multiple deep learning methods"
```

1.2.1 Deep neural network variants

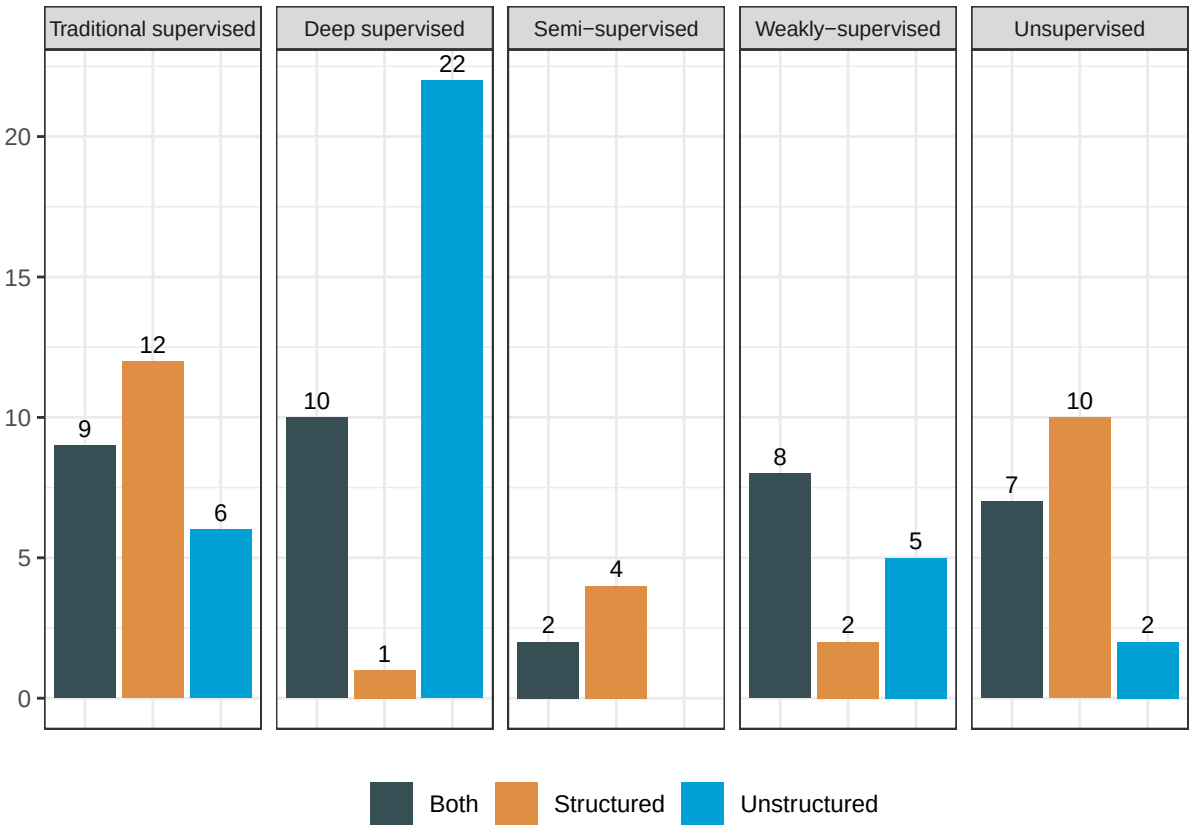


2 Phenotype


Traditional supervised learning
Surgical site infections
Suicidal behavior
Non−alcoholic fatty liver disease
Homelessness
Opioid abuse
Alcohol abuse
0
1
2
3
4
5
Deep supervised learning
Social support
Opioid abuse
Obesity
Insurance status
Homelessness
Financial resource
Fall
Education
Dementia
COVID−19
Bleeding
Employment
Alcohol abuse
Tobacco abuse
Semi−supervised learning
Ulcerative colitis
Type 1 diabetes mellitus
Stroke
Rheumatoid arthritis
Coronary artery disease
Congestive heart failure
Chronic obstructive pulmonary disease
Atrial fibrillation
Alzheimer's disease
Type 2 diabetes mellitus
Weakly−supervised learning
Type 2 diabetes mellitus
Type 1 diabetes mellitus
Coronary artery disease
Asthma
Ulcerative colitis
Epilepsy
Crohn's disease
Rheumatoid arthritis
Unsupervised learning
Ovarian cancer
Opioid abuse
Kidney cancer
Dementia
Chronic kidney disease
Unsupervised phenotype category
Co−occurring conditions
Disease progression
Subphenotypes

3 Data source

3.1 Summary




```
## [1] "There are 101 papers using machine learning models"
## [1] "There are 71 papers using machine learning models with unstructured data"
## [1] "There are 14 papers using machine learning models with competition data"
## [1] "There are 18 papers using machine learning models with data from multiple sites"
## [1] "There are 29 papers using machine learning models with openly available data"
## [1] "There are 64 papers using machine learning models with data from private single site"
## [1] "There are 45 papers reported machine learning models demographics"
## [1] "There are 20 papers released machine learning models source code"
```

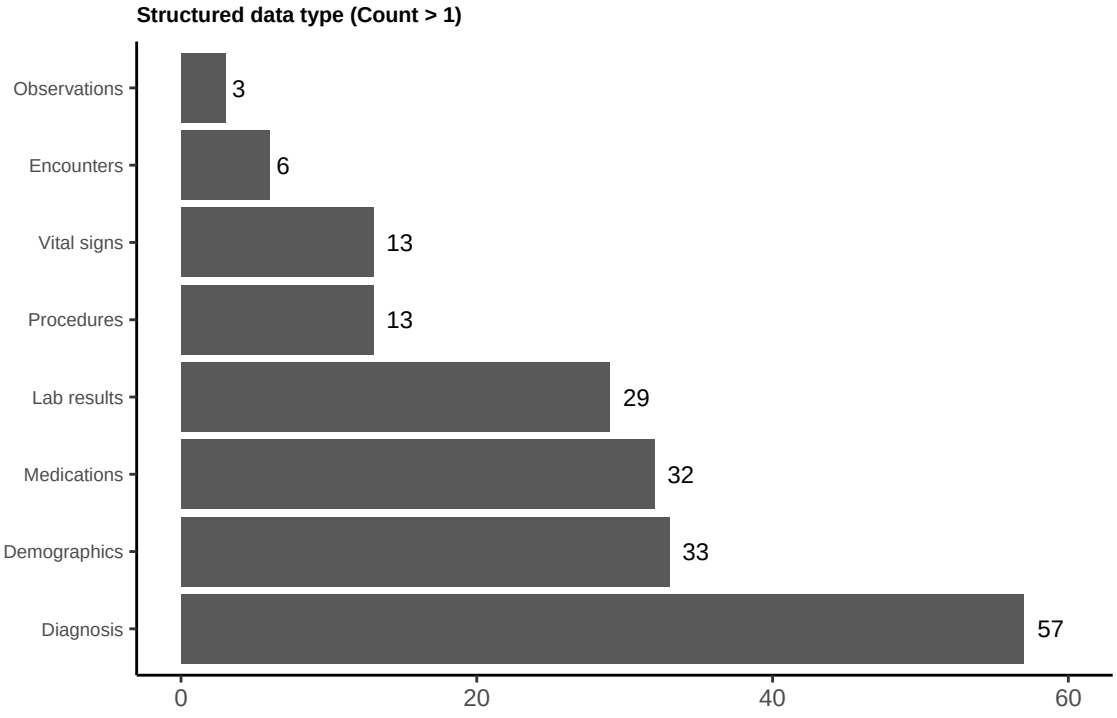
3.2 Structured and unstructured data type

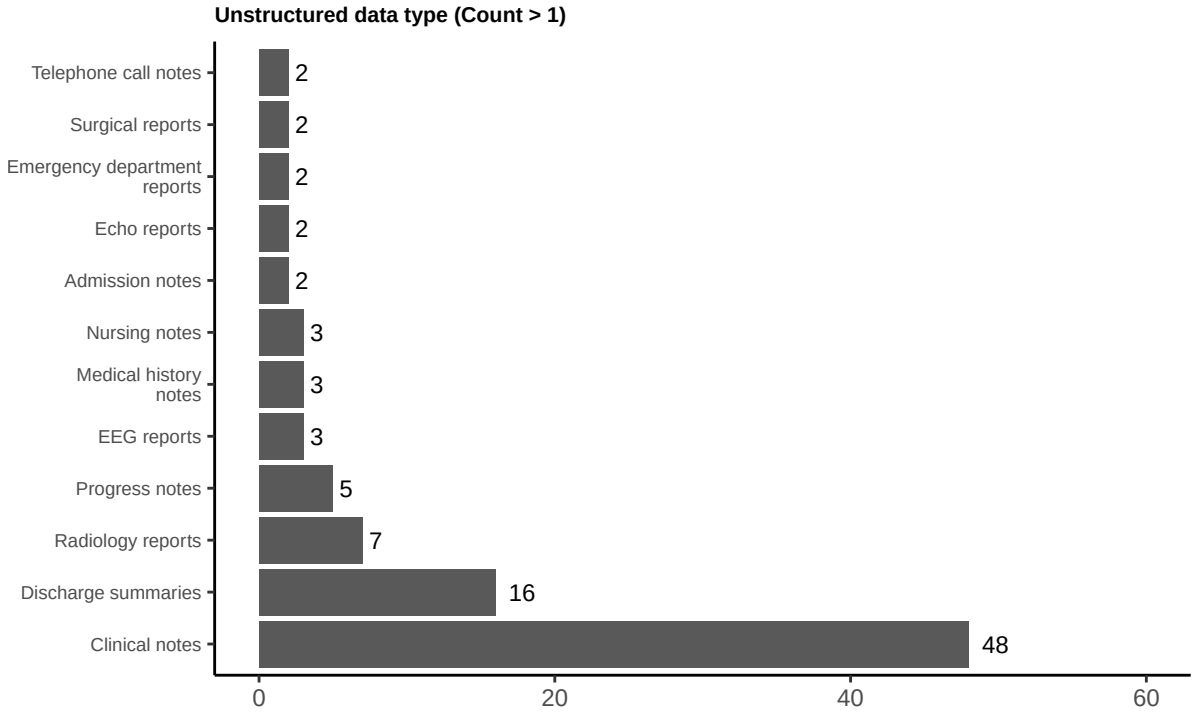
Table 3: Types of data used.

Data_type	Count
Both	36
Unstructured	35
Structured	30

```
## [1] "There are 50 papers using multiple structured data type"
```

```
## [1] "There are 15 papers using multiple unstructured data type"
```





3.2.1 Traditional supervised learning

```
## [1] "There are 27 papers using traditional supervised learning"
## [1] "There are 15 papers using traditional supervised learning with unstructured data"
## [1] "There are 3 papers using traditional supervised learning with competition data"
## [1] "There are 2 papers using traditional supervised learning with data from multiple sites"
## [1] "There are 4 papers using traditional supervised learning with openly available data"
## [1] "There are 22 papers using traditional supervised learning with data from private single site"
## [1] "There are 13 papers reported traditional supervised learning demographics"
## [1] "There are 4 papers released traditional supervised learning source code"
```

3.2.2 Deep supervised learning

```
## [1] "There are 33 papers using deep supervised learning"
## [1] "There are 32 papers using deep supervised learning with unstructured data"
## [1] "There are 11 papers using deep supervised learning with competition data"
## [1] "There are 9 papers using deep supervised learning with data from multiple sites"
## [1] "There are 19 papers using deep supervised learning with openly available data"
## [1] "There are 13 papers using deep supervised learning with data from private single site"
## [1] "There are 9 papers reported deep supervised learning demographics"
## [1] "There are 8 papers released deep supervised learning source code"
```

3.2.3 Semi-supervised learning

```
## [1] "There are 6 papers using semi-supervised learning"
## [1] "There are 2 papers using semi-supervised learning with unstructured data"
## [1] "There are 0 papers using semi-supervised learning with competition data"
## [1] "There are 0 papers using semi-supervised learning with data from multiple sites"
## [1] "There are 0 papers using semi-supervised learning with openly available data"
## [1] "There are 6 papers using semi-supervised learning with data from private single site"
## [1] "There are 3 papers reported semi-supervised learning demographics"
## [1] "There are 0 papers released semi-supervised learning source code"
```

3.2.4 Weakly-supervised learning

```
## [1] "There are 15 papers using weakly-supervised learning"
## [1] "There are 13 papers using weakly-supervised learning with unstructured data"
## [1] "There are 0 papers using weakly-supervised learning with competition data"
## [1] "There are 4 papers using weakly-supervised learning with data from multiple sites"
## [1] "There are 2 papers using weakly-supervised learning with openly available data"
## [1] "There are 10 papers using weakly-supervised learning with data from private single site"
## [1] "There are 4 papers reported weakly-supervised learning demographics"
## [1] "There are 3 papers released weakly-supervised learning source code"
```

3.2.5 Unsupervised learning

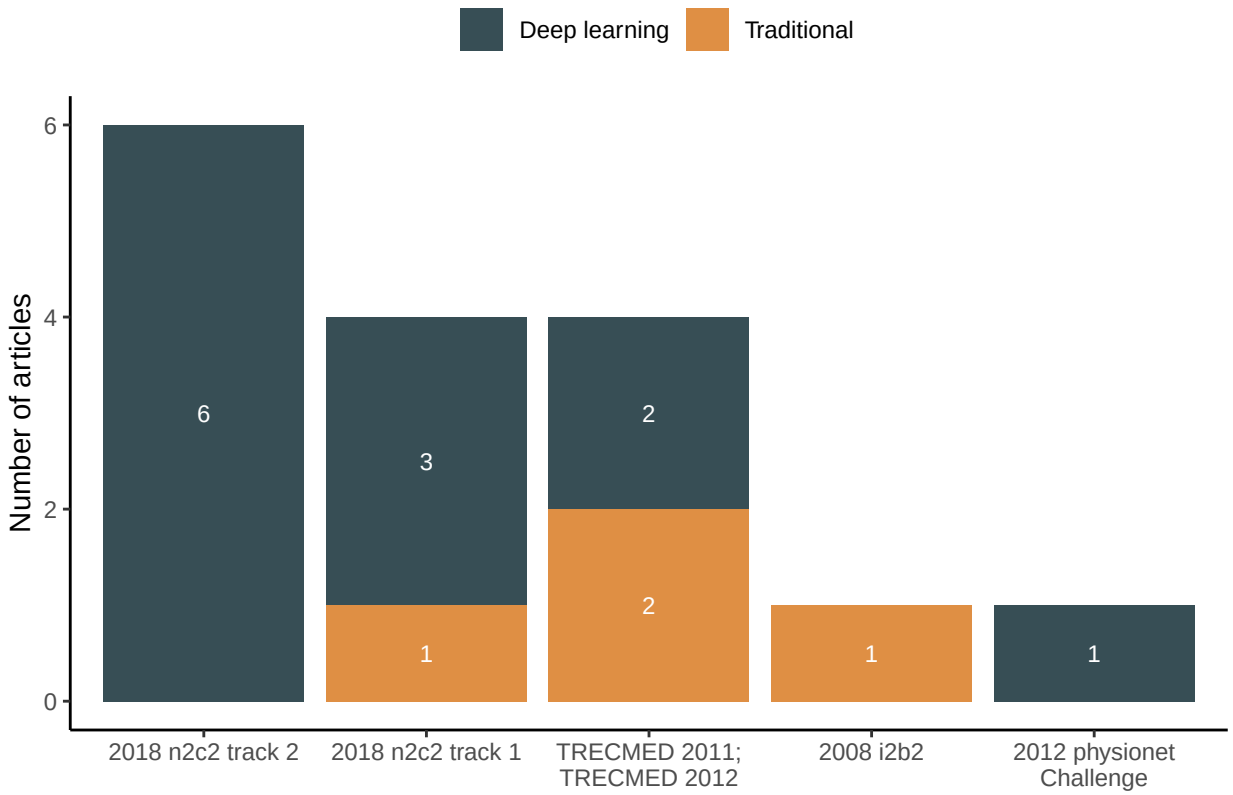
```
## [1] "There are 19 papers using unsupervised learning"
## [1] "There are 9 papers using unsupervised learning with unstructured data"
## [1] "There are 0 papers using unsupervised learning with competition data"
## [1] "There are 3 papers using unsupervised learning with data from multiple sites"
## [1] "There are 3 papers using unsupervised learning with openly available data"
```

```
## [1] "There are 13 papers using unsupervised learning with data from private single site"
## [1] "There are 15 papers reported unsupervised learning demographics"
## [1] "There are 4 papers released unsupervised learning source code"
```

3.3 Openly-available data

Competition_data_name_unnested	Count
2018 n2c2 track 2	6
2018 n2c2 track 1	4
TRECMED 2011	2
TRECMED 2012	2
2008 i2b2	1
2012 physionet Challenge	1

```
## [1] "There are 2 papers using multiple Competition data"
```



Data_source_unnested	Count
MIMIC-III database	21
MTSamples database	1

```
## [1] "There are 1 papers using multiple Openly data"
```

4 NLP software

NLP_software_unnested	Count
cTAKES	19
NegEx	6
NILE	6
NLTK	5
MetaMap	4
Stanford CoreNLP	2

```
## [1] "There are 7 papers using multiple NLP software"
```

5 Emebddings

Embedding_training_data_unnested	Count
Unstructured EHR	13
MIMIC-III database	12
Biomedical literature	10
Wikipedia	6
Structured EHR	2

```
## [1] "There are 7 papers using multiple embedding training data"
```

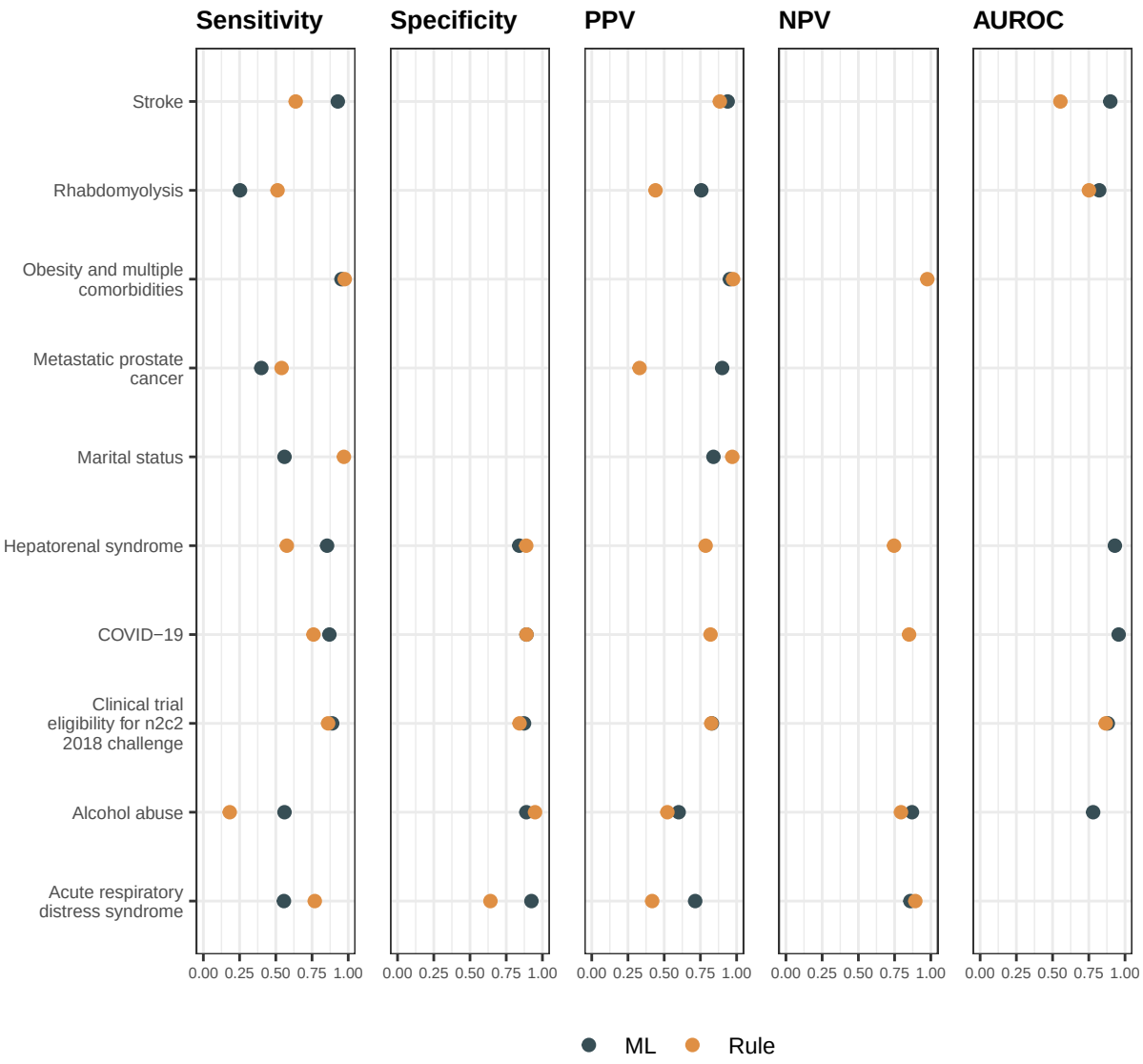
Embedding_unnested	Count
Word2vec	19
GloVe	6
BERT	5
RoBERTa	3
BioBERT	2
BioClinicalBERT	2
FastText	2
Not specified	2

```
## [1] "There are 11 papers using multiple embedding training methods"
```

6 Validation and comparison

Model_performance_metrics_unnested	Count
Precision	61
Recall	59
AUROC	42
F-score	42
Specificity	20
Accuracy	18
NPV	15
AUPRC	9

6.1 Traditonal supervised ML vs. rule-based



6.2 Deep supervised ML vs. supervised

7 Reporting

```
## There are 45 papers reported demographcis 0.4455446
```