

A summary of EHR-based phenotyping article annotation

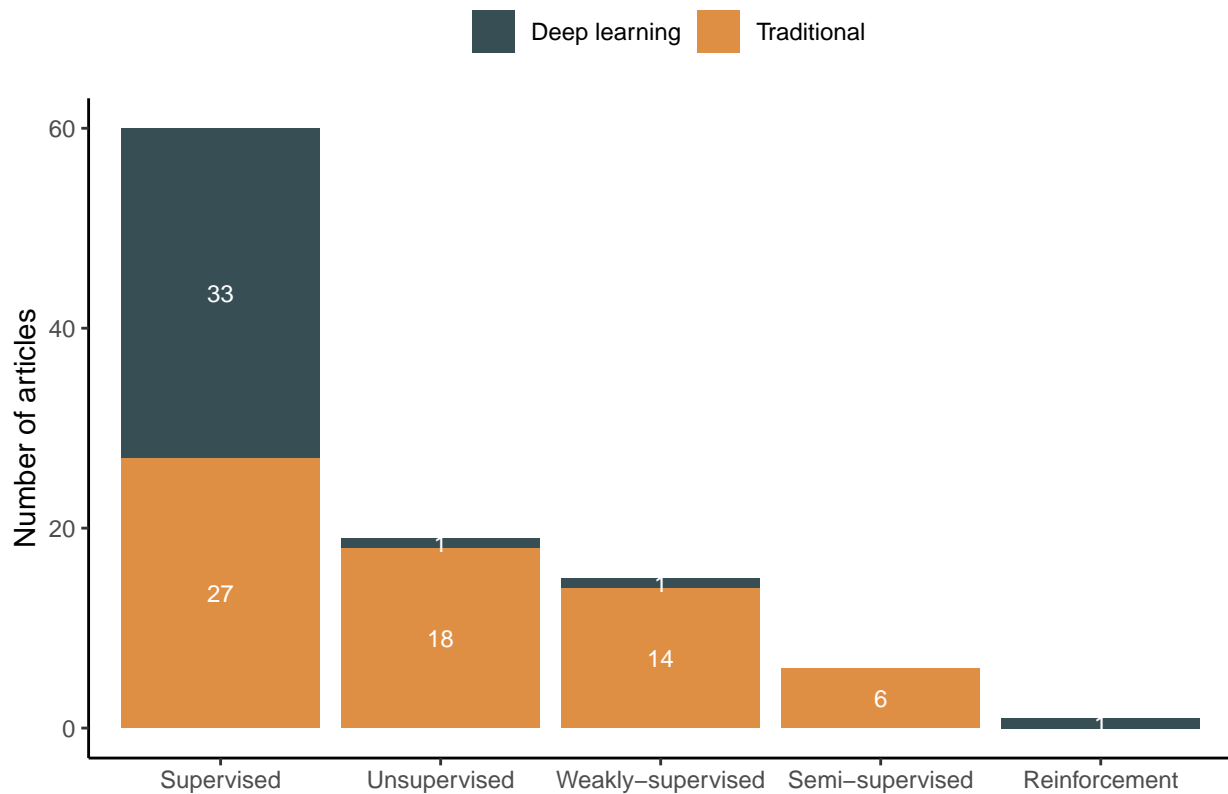
Siyue Yang, Jessica Gronsbell

05/25/2022

Contents

1	Overview	2
1.1	Traditional ML method	2
1.2	DL method	3
2	Phenotype	4
2.1	More nuanced phenotype	5
3	Data source	6
3.1	Summary	6
3.2	Structured and unstructured data type	6
3.3	Openly-available data	9
4	NLP software	10
5	Emebddings	10
6	Validation and comparison	11
6.1	Traditonal supervised ML vs. rule-based	11
6.2	Deep supervised ML vs. traditional supervised ML	12
7	Reporting	13

1 Overview



1.1 Traditional ML method

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

ML	Traditional ML method	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"

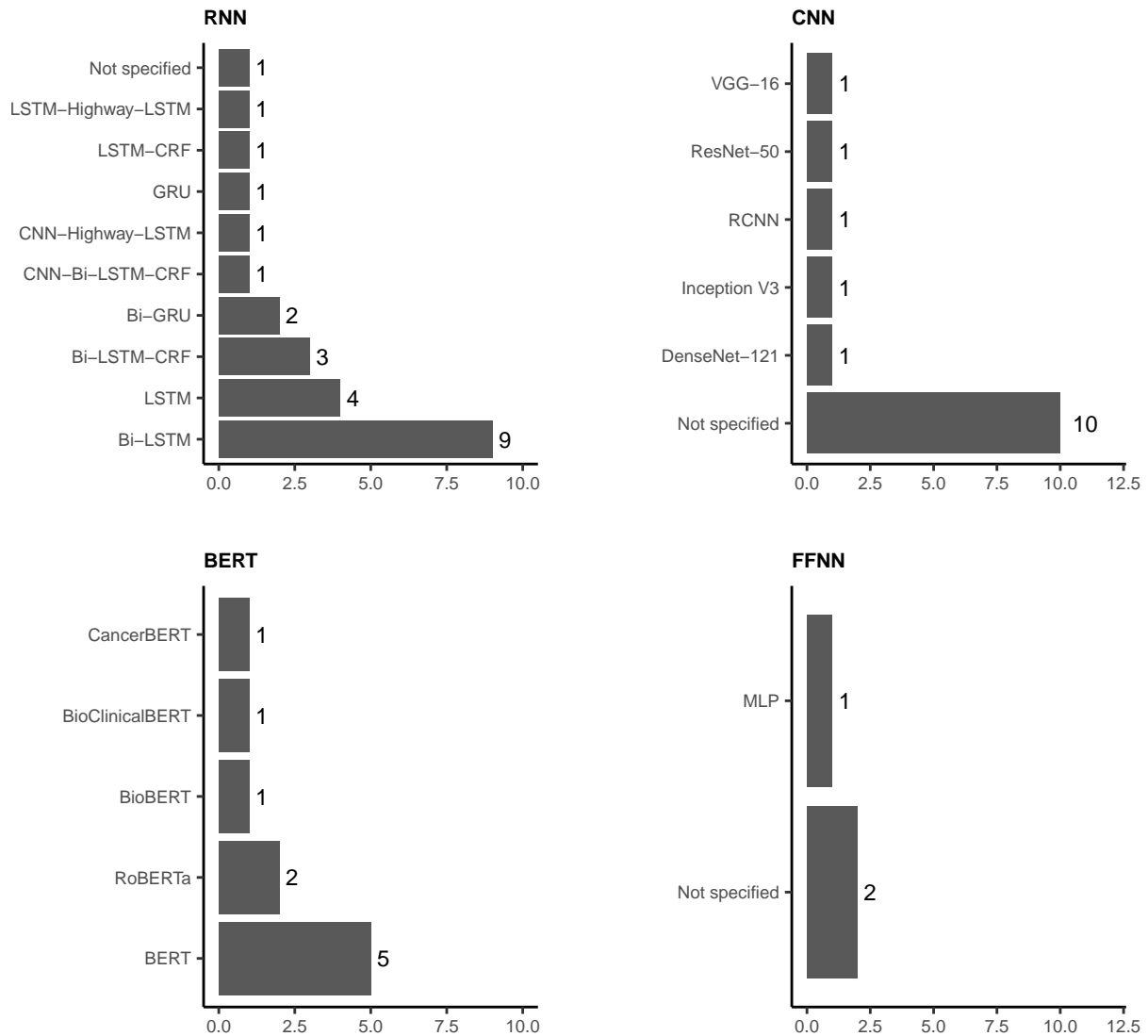
1.2 DL method

Table 2: Deep learning methods

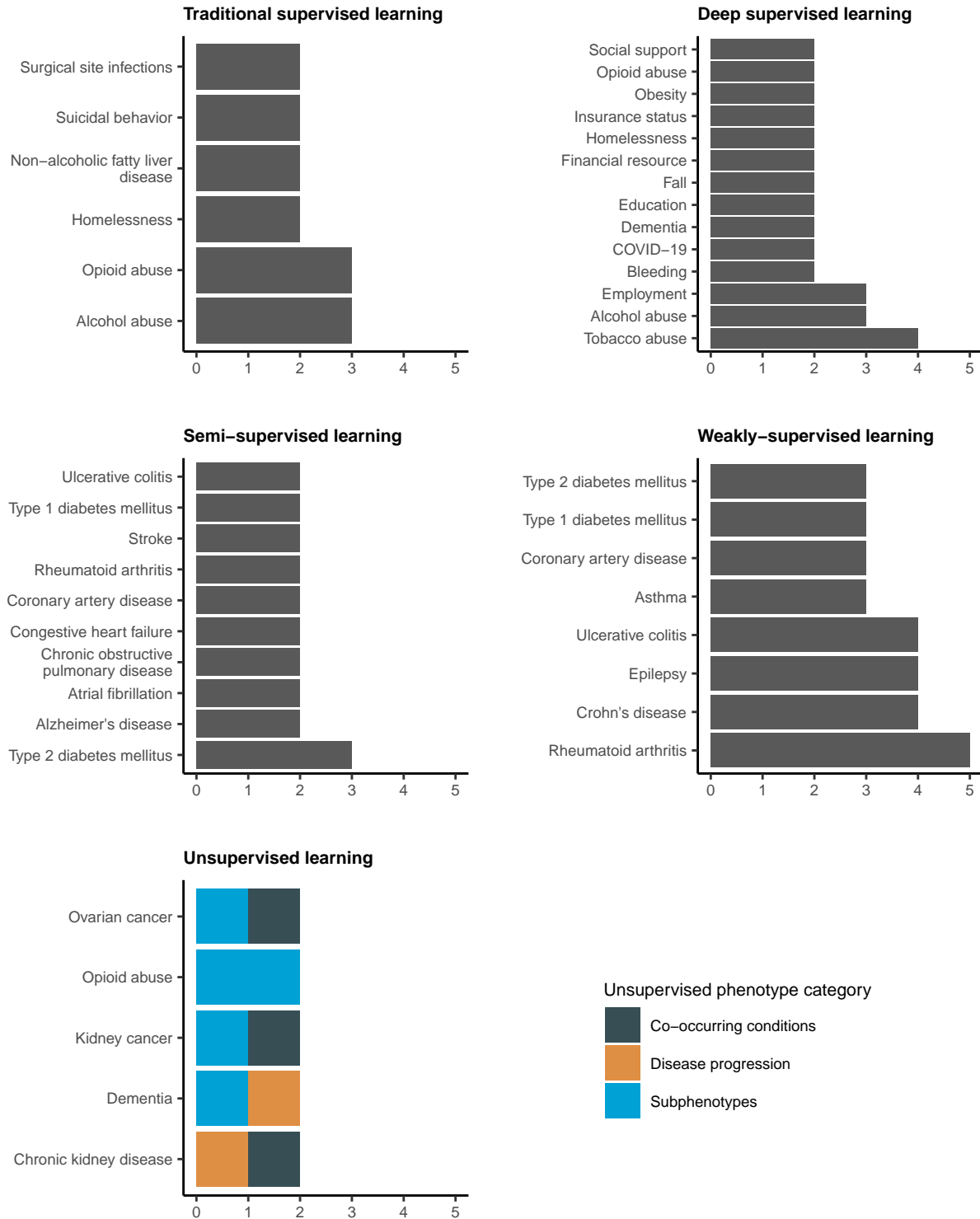
DL method	ML	Count
BERT	Supervised	7
CNN	Supervised	12
FFNN	Supervised	3
RNN	Supervised	18

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

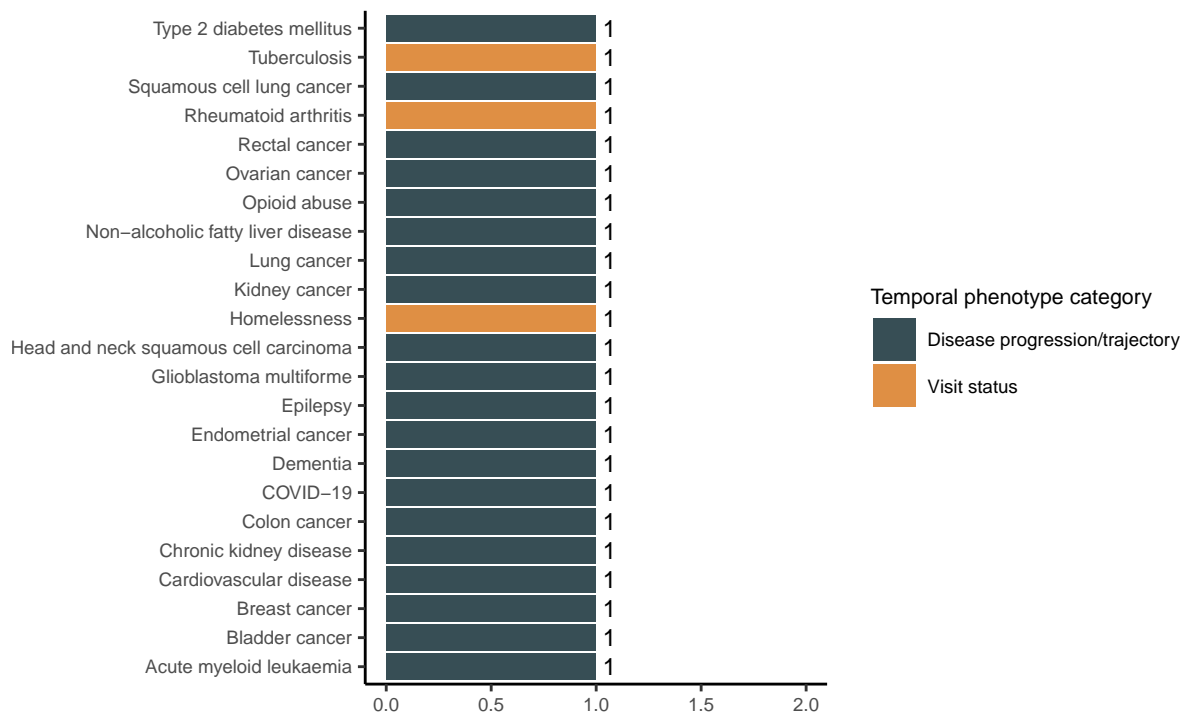
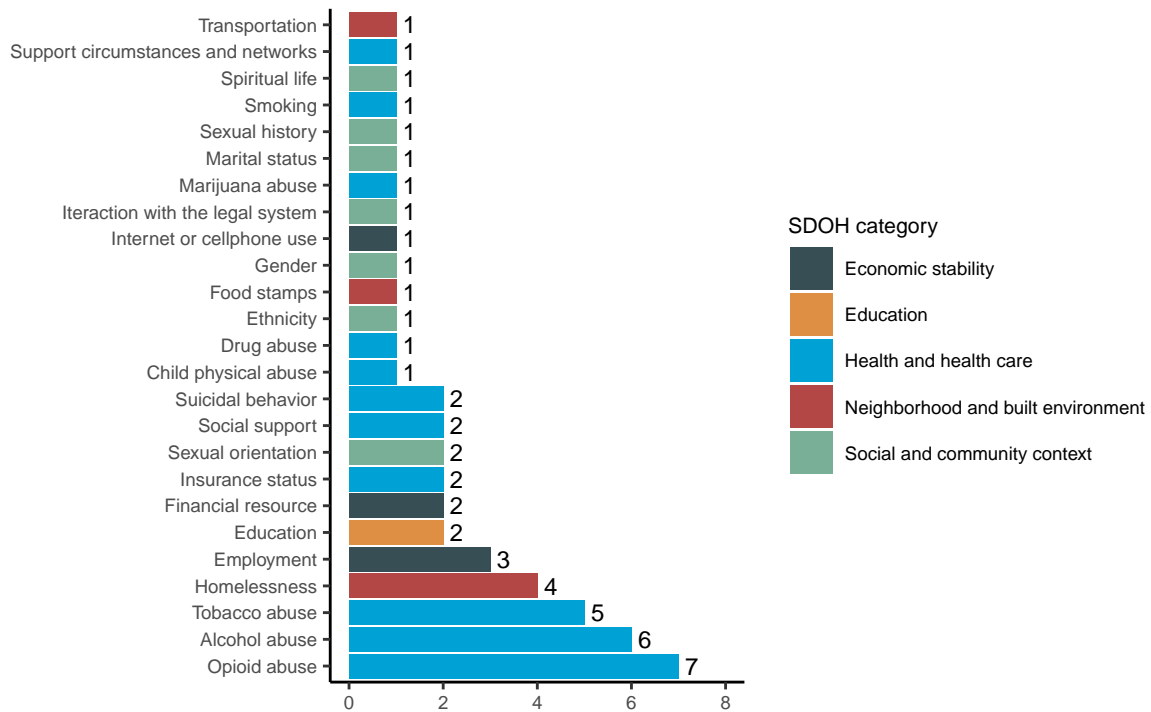
1.2.1 Deep neural network variants



2 Phenotype

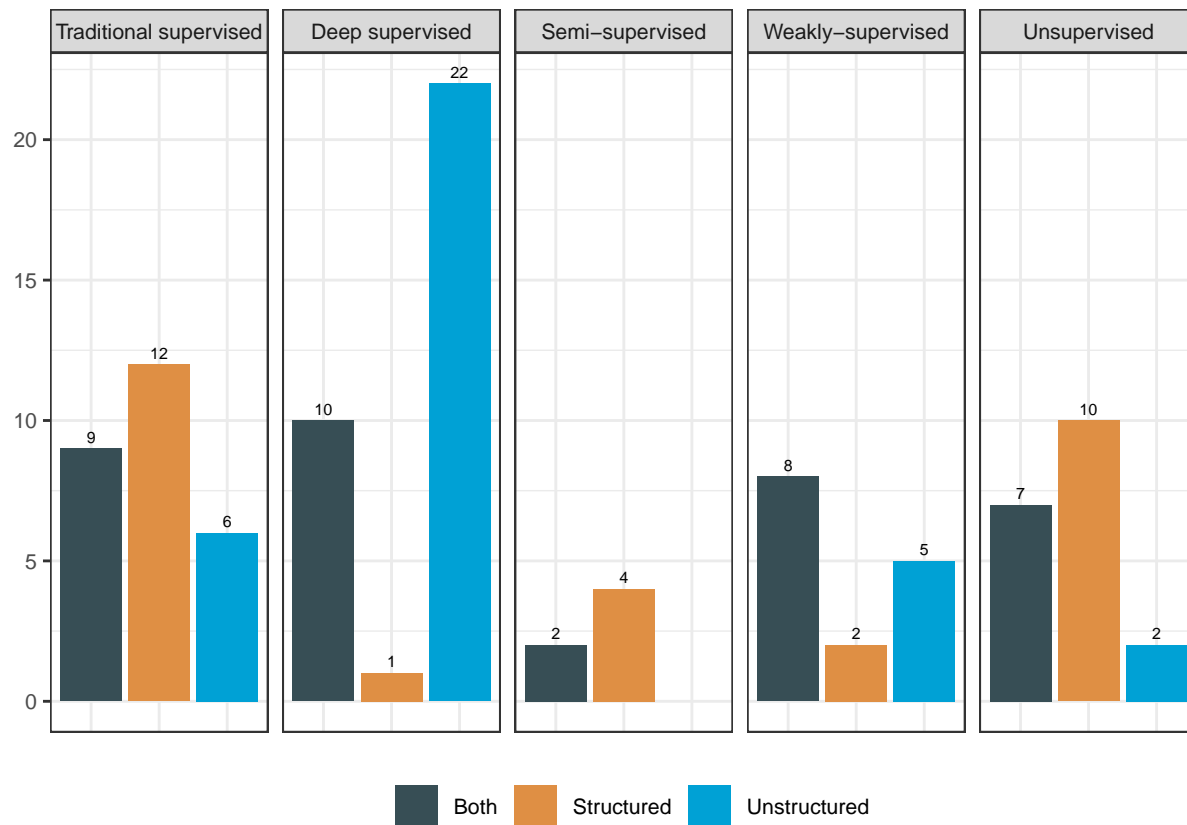


2.1 More nuanced phenotype



3 Data source

3.1 Summary

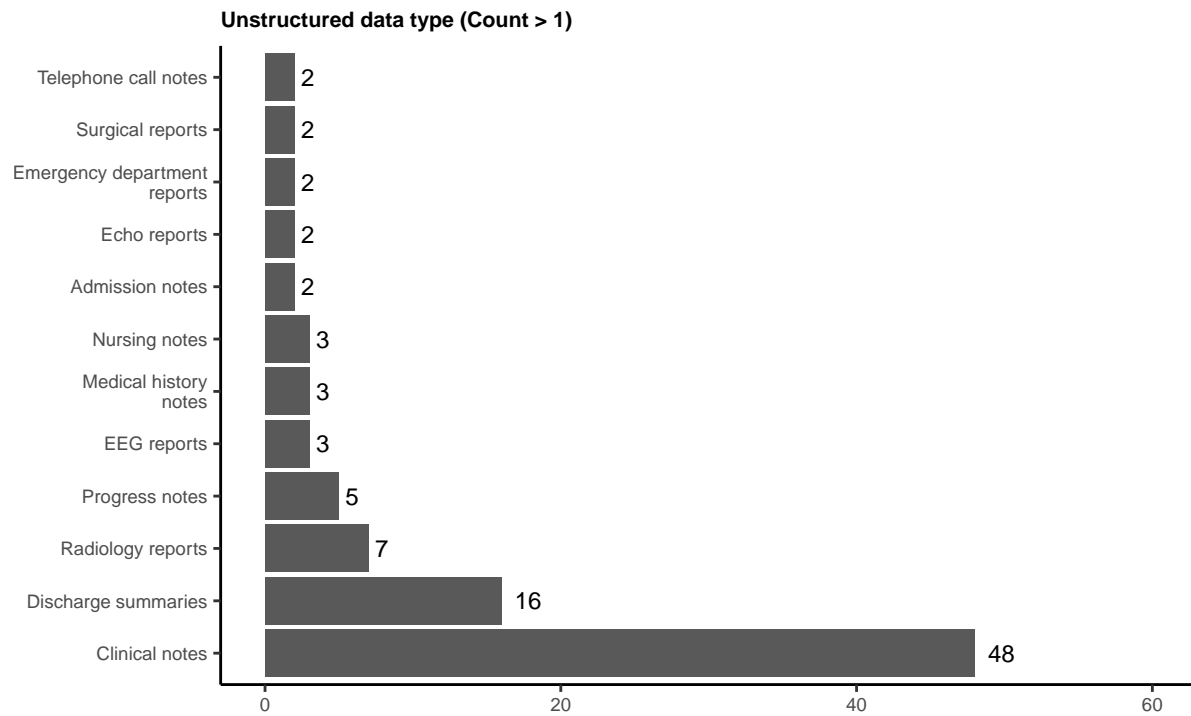
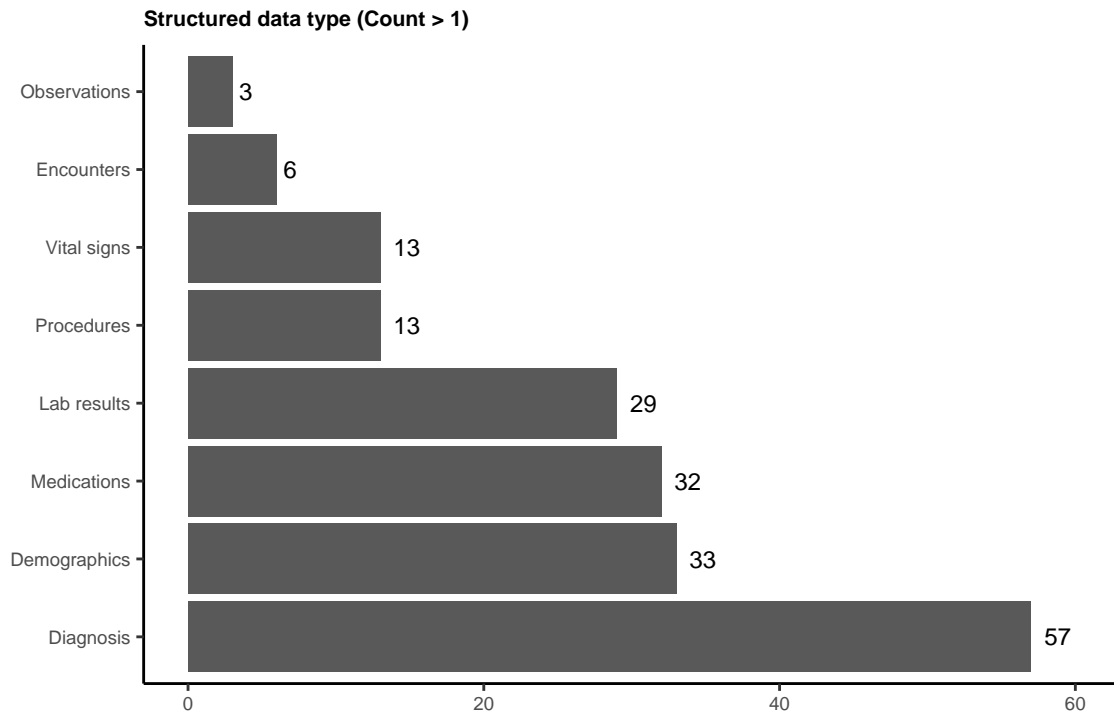


```
## [1] "There are 101 papers using machine learning models"
## [1] "There are 71 papers using machine learning models with unstructured data"
## [1] "There are 47 papers using machine learning models with NLP software"
## [1] "There are 14 papers using machine learning models with competition data"
## [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] "-----"
## [1] "There are 20 papers machine learning models compared with rule-based algorithms"
## [1] "There are 21 papers machine learning models compared with traditional ML algorithms"
## [1] "-----"
## [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

```
## [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 14 papers using traditional supervised learning with NLP software"
## [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] "-----"
## [1] "There are 10 papers traditional supervised learning compared with rule-based algorithms"
## [1] "There are 0 papers traditional supervised learning compared with traditional ML algorithms"
## [1] "-----"
## [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 18 papers using deep supervised learning with NLP software"
## [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] "-----"
## [1] "There are 2 papers deep supervised learning compared with rule-based algorithms"
## [1] "There are 19 papers deep supervised learning compared with traditional ML algorithms"
## [1] "-----"
## [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 1 papers using semi-supervised learning with NLP software"
## [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] "-----"
## [1] "There are 1 papers semi-supervised learning compared with rule-based algorithms"
## [1] "There are 0 papers semi-supervised learning compared with traditional ML algorithms"
## [1] "-----"
## [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 10 papers using weakly-supervised learning with NLP software"
## [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] "-----"
## [1] "There are 7 papers weakly-supervised learning compared with rule-based algorithms"
## [1] "There are 1 papers weakly-supervised learning compared with traditional ML algorithms"
## [1] "-----"
## [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 4 papers using unsupervised learning with NLP software"
## [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] "-----"
## [1] "There are 0 papers unsupervised learning compared with rule-based algorithms"
## [1] "There are 0 papers unsupervised learning compared with traditional ML algorithms"
## [1] "-----"
## [1] "There are 15 papers reported unsupervised learning demographics"
## [1] "There are 4 papers released unsupervised learning source code"

```

3.3 Openly-available data

```

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

```

Competition data name	Supervised Traditional	Supervised Deep learning	Count
2018 n2c2 track 1	1	3	4
2018 n2c2 track 2	0	6	6
TRECMED 2011	1	1	2
TRECMED 2012	1	1	2

Data source	Count
MIMIC-III database	21
MTSamples database	1

```

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

```

Data source	Reinforcement Deep learning	Supervised Deep learning	Supervised Traditional	Unsupervised Traditional	Weakly- supervised Deep learning	Weakly- supervised Traditional	Count
MIMIC-III database	1	14	1	3	1	1	21

4 NLP software

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

NLP software	Supervised Deep learning	Weakly- supervised Traditional	Semi- supervised Traditional	Supervised Traditional	Unsupervised Traditional	Count
cTAKES	8	0	1	8	2	19
MetaMap	1	0	0	3	0	4
NegEx	0	2	0	3	1	6
NILE	0	5	0	1	0	6
NLTK	4	0	0	0	1	5
Stanford CoreNLP	2	0	0	0	0	2

5 Emebddings

Embeddings were only used in deep supervised articles.

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

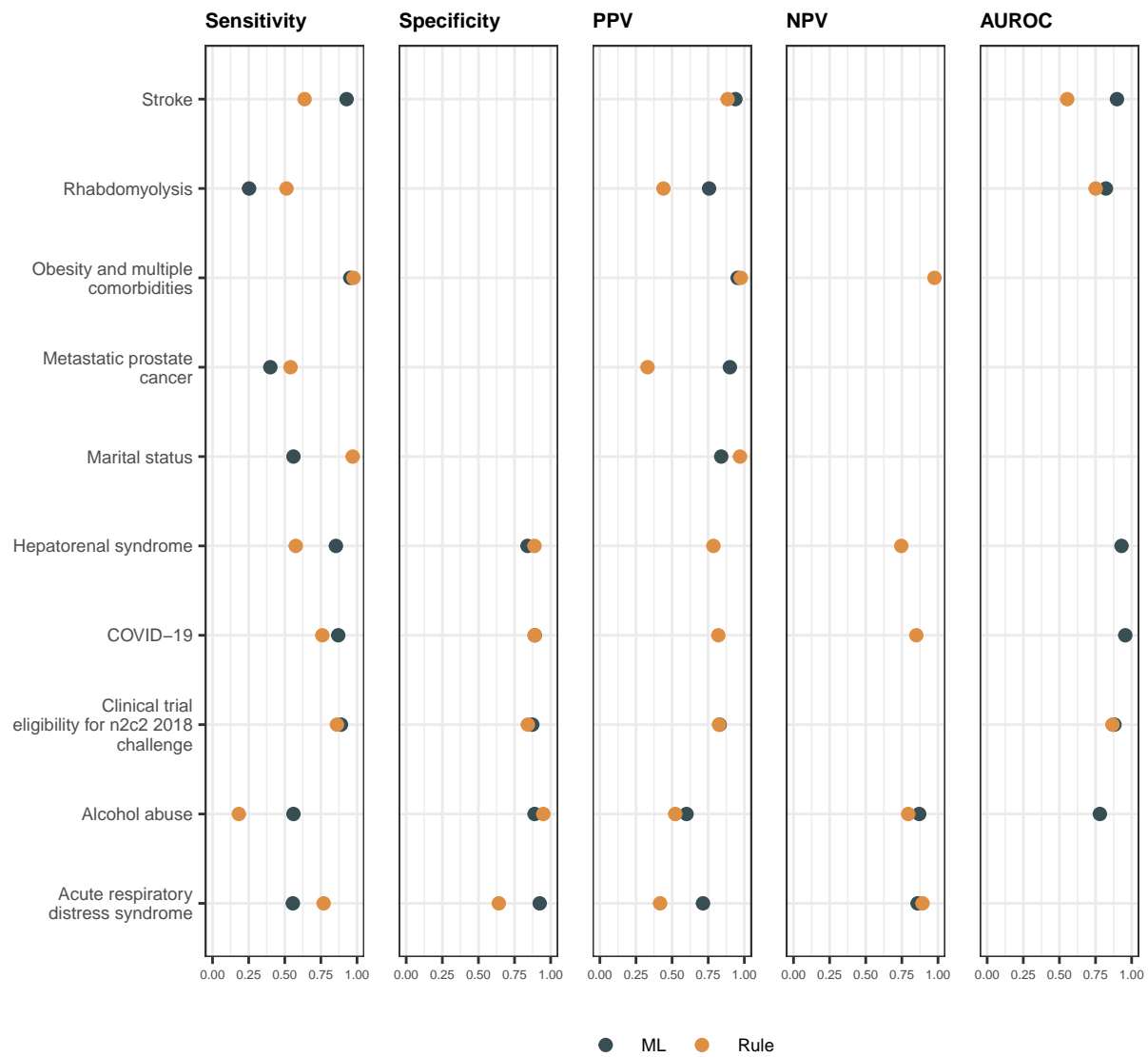
[1] "There are 7 papers using multiple embedding training data"

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

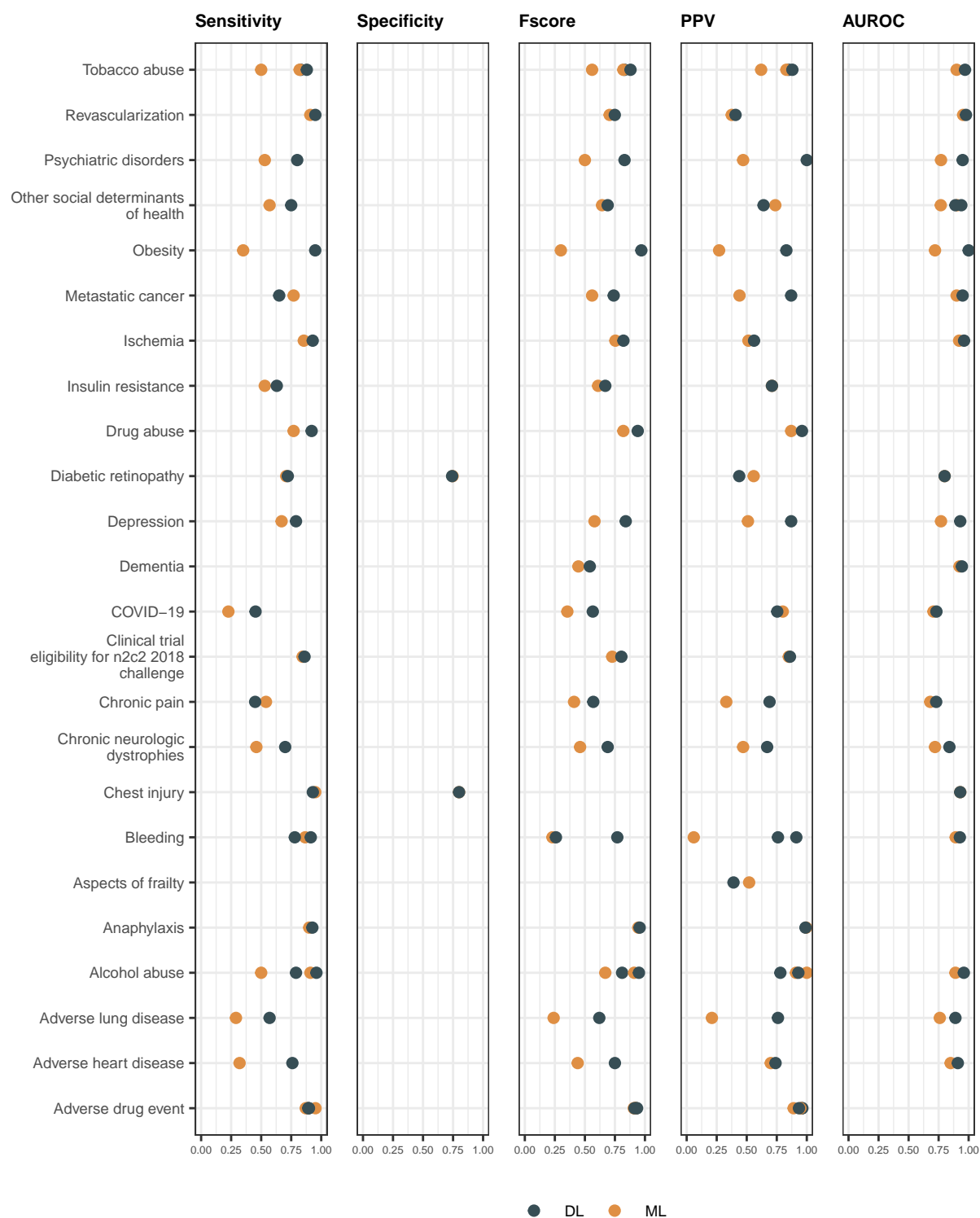
[1] "There are 11 papers using multiple embedding training methods"

6 Validation and comparison

6.1 Traditional supervised ML vs. rule-based



6.2 Deep supervised ML vs. traditional supervised ML



Model perfor- mance metrics	Supervised Deep learning	Supervised Tradi- tional	Weakly- supervised Deep learning	Weakly- supervised Tradi- tional	Reinforcemen Deep learning	Unsupervised Tradi- tional	Semi- supervised Tradi- tional	Count
Accuracy	5	8	1	4	0	0	0	18
AUPRC	3	2	0	2	1	1	0	9
AUROC	10	15	1	10	1	0	5	42
F-score	26	9	0	7	0	0	0	42
NPV	1	7	0	5	0	0	2	15
Precision	26	23	0	8	0	0	4	61
Recall	26	23	1	7	0	0	2	59
Specificity	7	11	1	1	0	0	0	20

7 Reporting

There are 45 papers reported demographcis, 0.4455

There are 20 papers reported demographcis, 0.198