A summary of EHR-based phenotyping article annotation
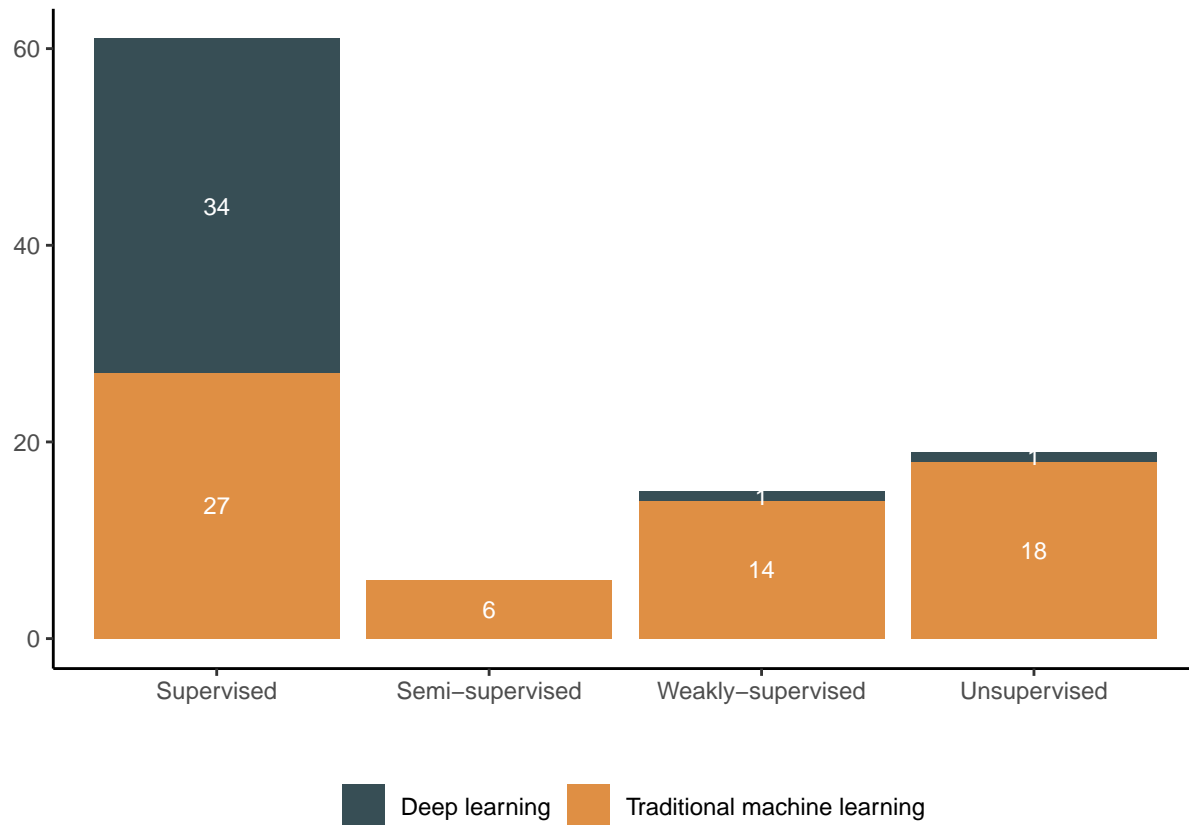
Siyue Yang, Jessica Gronsbell

05/25/2022

Contents

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
Weakly-supervised	PheNorm	3
Weakly-supervised	MAP	2
Weakly-supervised	Random forest	2
Unsupervised	LDA	5
Unsupervised	K-means	4
Unsupervised	UPGMA Hierarchical clustering	2

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

1.2 DL method

Table 2: Deep supervised learning methods

DL method	ML	Count
BERT	Supervised	7
CNN	Supervised	12
FFNN	Supervised	3
RNN	Supervised	19

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

1.2.1 Deep neural network variants

2 Phenotype

(a)

Traditional supervised learning

Category	Value
Surgical site infections	2
Suicidal behavior	2
Non−alcoholic fatty liver disease	2
Homelessness	2
Opioid abuse	3
Alcohol abuse	3

(b)

Deep supervised learning

Category	Value
Social support	2
Respiratory impairment	2
Opioid abuse	2
Obesity	2
Insurance status	2
Homelessness	2
Financial resource	2
Fall	2
Education	2
Dementia	2
COVID−19	2
Bleeding	2
Employment	3
Alcohol abuse	3
Tobacco abuse	4

(c)

Semi−supervised learning

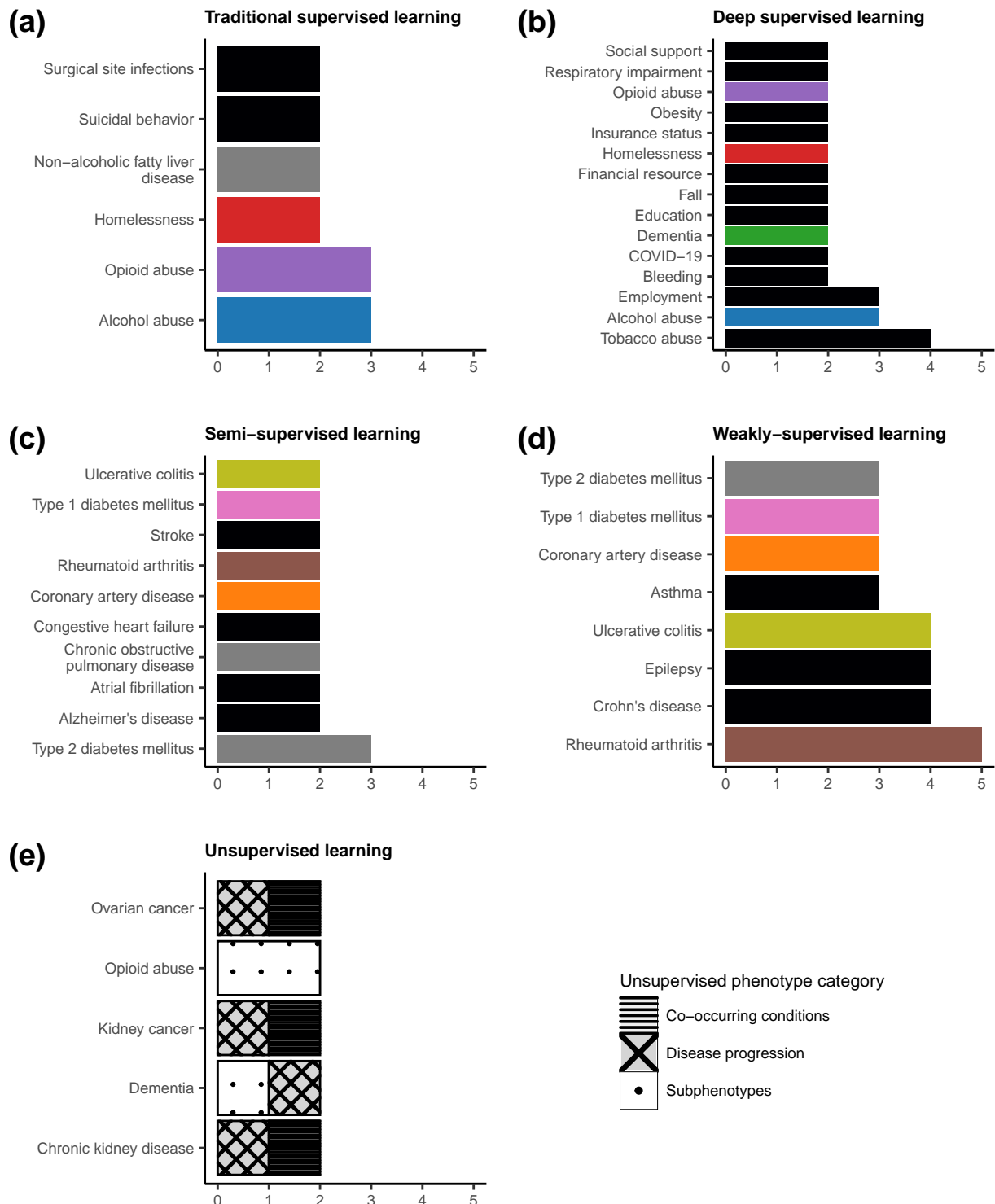| Category	Value
Ulcerative colitis	2
Type 1 diabetes mellitus	2
Stroke	2
Rheumatoid arthritis	2
Coronary artery disease	2
Congestive heart failure	2
Chronic obstructive pulmonary disease	2
Atrial fibrillation	2
Alzheimer's disease	2
Type 2 diabetes mellitus	3

(d)

Weakly−supervised learning

Category	Value
Type 2 diabetes mellitus	3
Type 1 diabetes mellitus	3
Coronary artery disease	3
Asthma	3
Ulcerative colitis	4
Epilepsy	4
Crohn's disease	4
Rheumatoid arthritis	5

(e)

Unsupervised learning

Category	Value
Ovarian cancer	2
Opioid abuse	2
Kidney cancer	2
Dementia	2
Chronic kidney disease	2

Unsupervised phenotype category

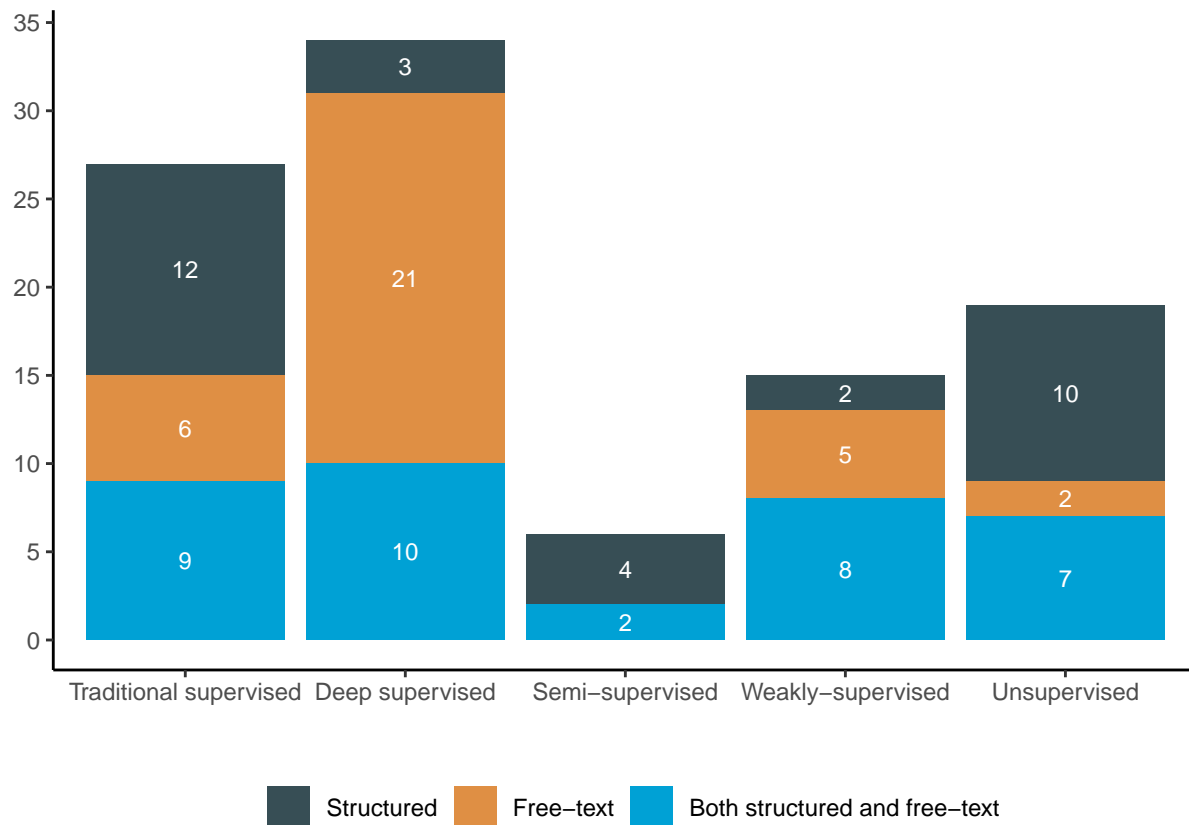- Co−occurring conditions
- Disease progression
- Subphenotypes

2.1 More nuanced phenotype

	Total number of papers	Used free-text	Used NLP software	Used competition data	Used multisite data	Used open data	Used private single-site data	Compared to rule-based algorithms	Comapred to traditional ML	Reported patient demographic	Released open code
TSL	27	15	14	3	2	4	22	10	0	13	4
DSL	34	31	18	11	9	20	13	2	20	10	9
SSL	6	2	1	0	0	0	6	1	0	3	0
WSL	15	13	10	0	4	2	10	8	1	4	3
USL	19	9	4	0	3	3	13	0	0	15	4
Total	101	70	47	14	18	29	64	21	21	45	20

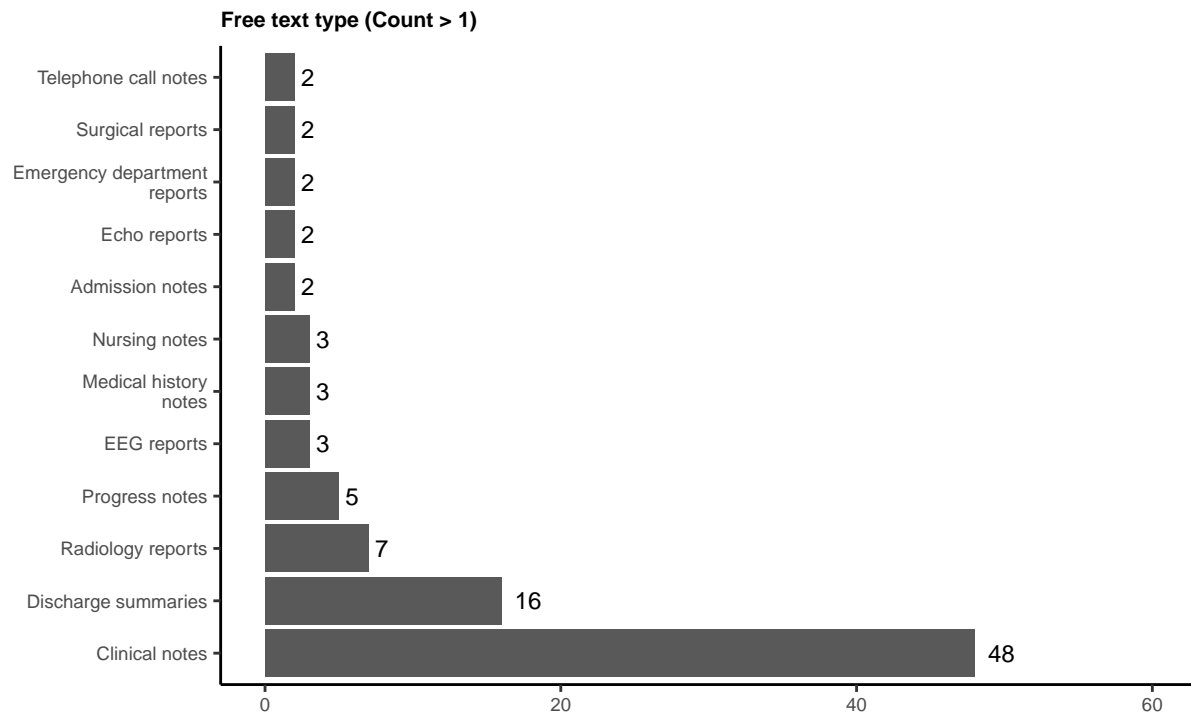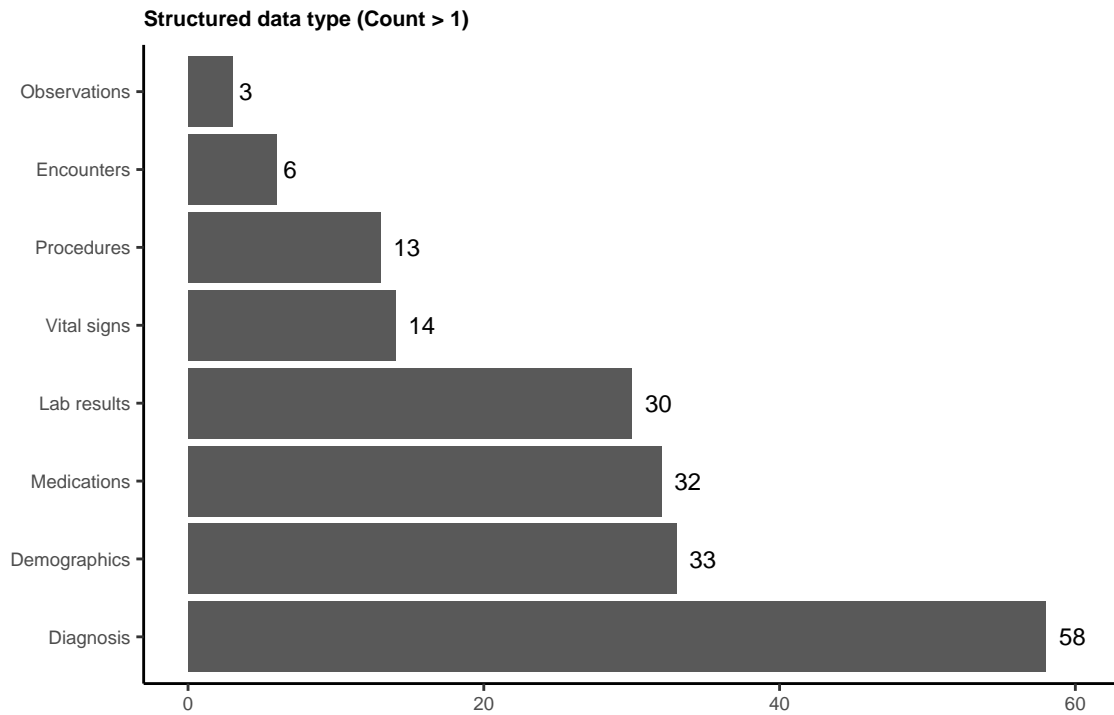# 3 Data source

3.1 Summary

TSL = Traditional supervised learning. DSL = Deep supervised learning. DRL = Reinforcement deep learning. SSL = Semi-supervised learning. WSL = Weakly-supervised learning. US = Unsupervised learning.

3.2 Structured and unstructured data type

```
## [1] "There are 51 papers using multiple structured data type"
```

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

Structured data type (Count > 1)

Data type	Count
Observations	3
Encounters	6
Procedures	13
Vital signs	14
Lab results	30
Medications	32
Demographics	33
Diagnosis	58

Free text type (Count > 1)

Data type	Count
Telephone call notes	2
Surgical reports	2
Emergency department reports	2
Echo reports	2
Admission notes	2
Nursing notes	3
Medical history notes	3
EEG reports	3
Progress notes	5
Radiology reports	7
Discharge summaries	16
Clinical notes	48

Data source	Supervised Deep learning	Supervised Traditional machine learning	Weakly-supervised Deep learning	Weakly-supervised Traditional machine learning	Unsupervised Traditional machine learning	Count
MIMIC-III database	15	1	1	1	3	21
MTSamples database	1	0	0	0	0	1

3.3 Openly-available data

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

Competition data name	Supervised Traditional machine learning	Supervised Deep learning	Count
2018 n2c2 track 2	0	6	6
2018 n2c2 track 1	1	3	4
TRECMED 2011	1	1	2
TRECMED 2012	1	1	2
2008 i2b2	1	0	1
2012 physionet Challenge	0	1	1

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

4 NLP software

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

5 Emebddings

Embeddings were only used in deep supervised articles.

Embedding training data	Count
Unstructured EHR	11
Biomedical literature	10
MIMIC-III database (internal)	7
MIMIC-III database (external)	6
Wikipedia	6
Structured EHR	2

```
## [1] "There are 7 papers using multiple embedding training data"
```
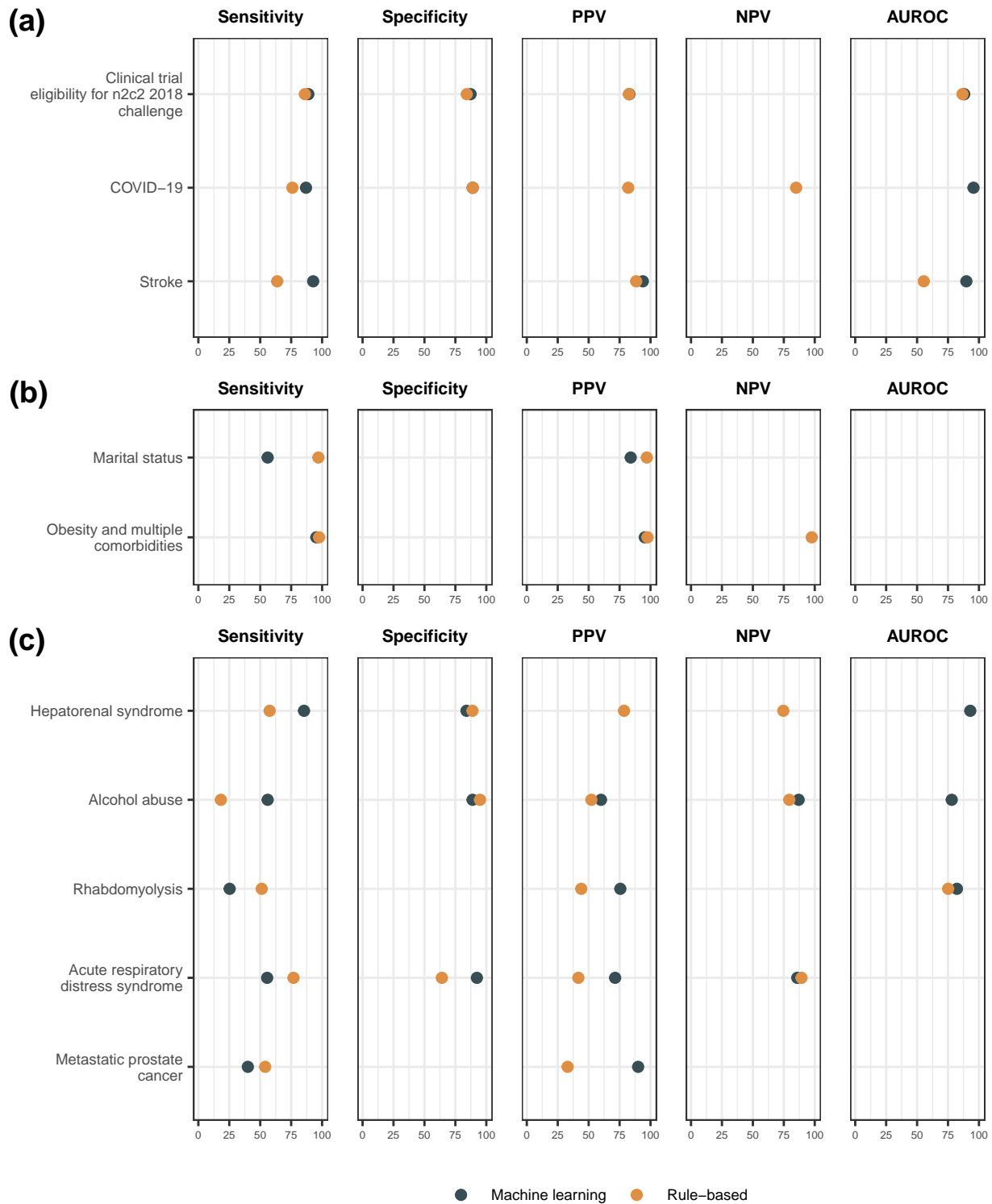
Embedding	Count
Word2vec	19
GloVe	6
BERT	5
RoBERTa	3
BioBERT	2
BioClinicalBERT	2
FastText	2
Not specified	2

```
## [1] "There are 11 papers using multiple embedding training methods"
```
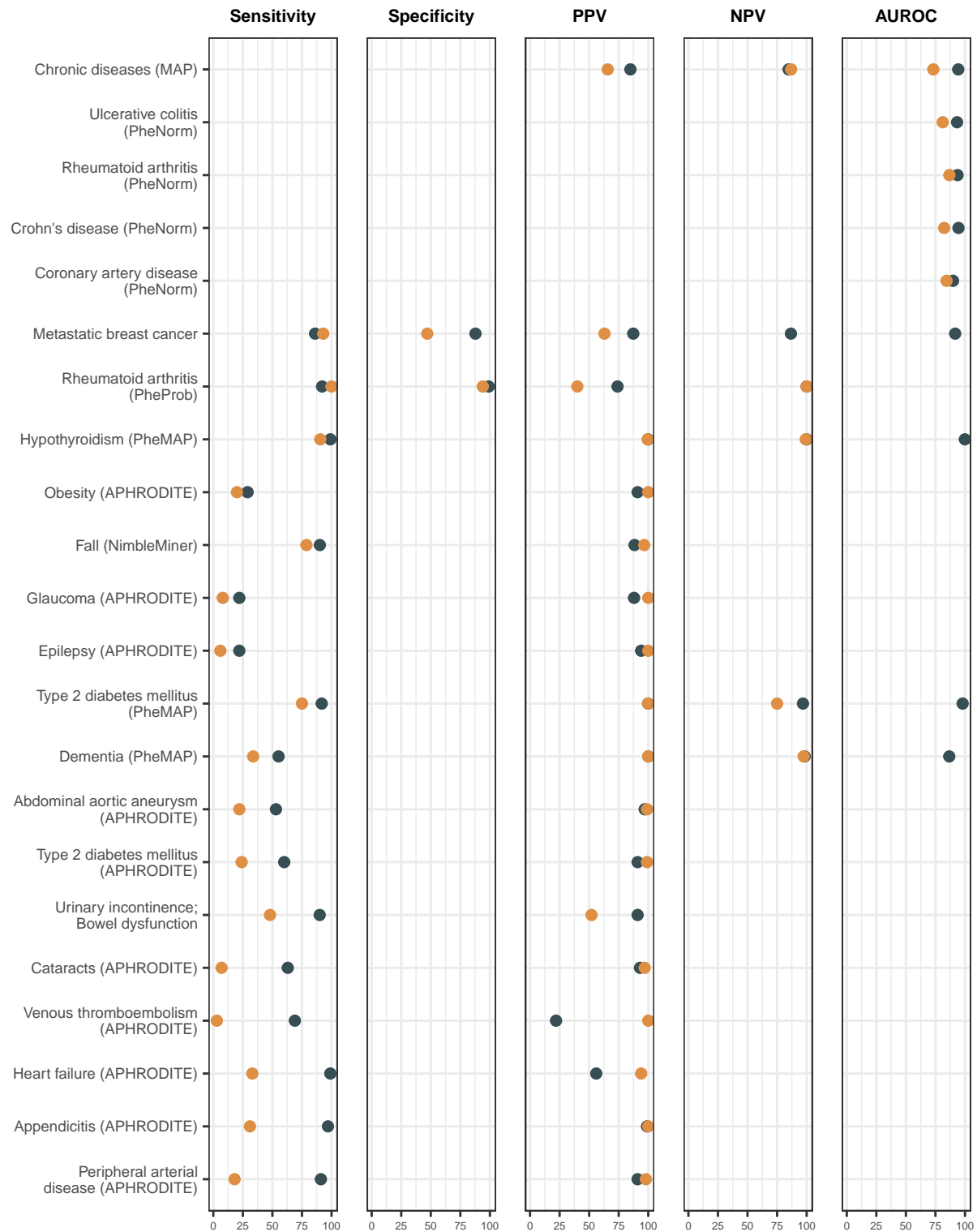
6 Validation and comparison
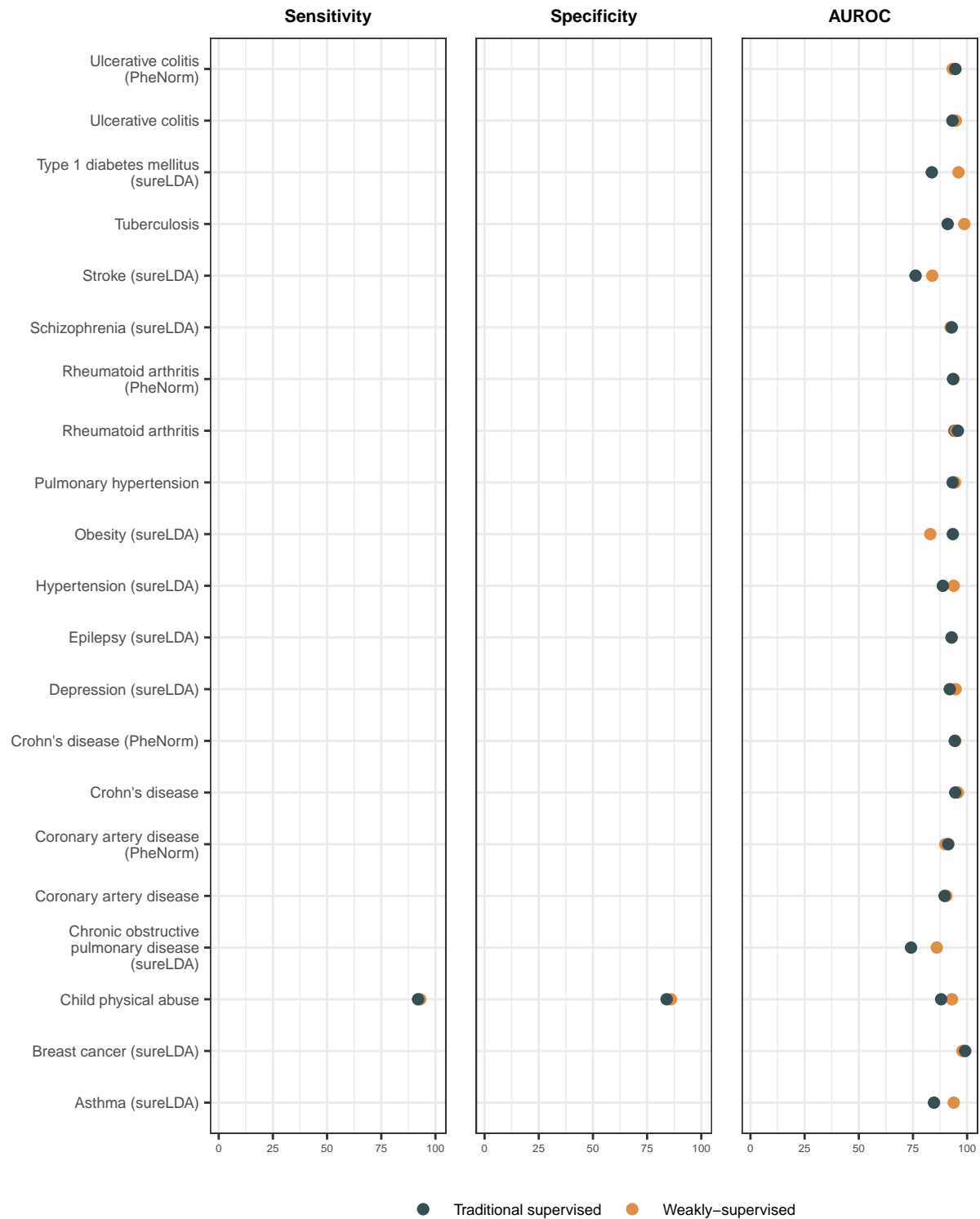
6.1 Traditonal supervised ML vs. rule-based

(a)

(b)

(c)

● Machine learning ● Rule−based

6.2 Deep supervised ML vs. traditional supervised ML

(a)

(b)

(c)

● Deep supervised learning   ● Traditional supervised learning
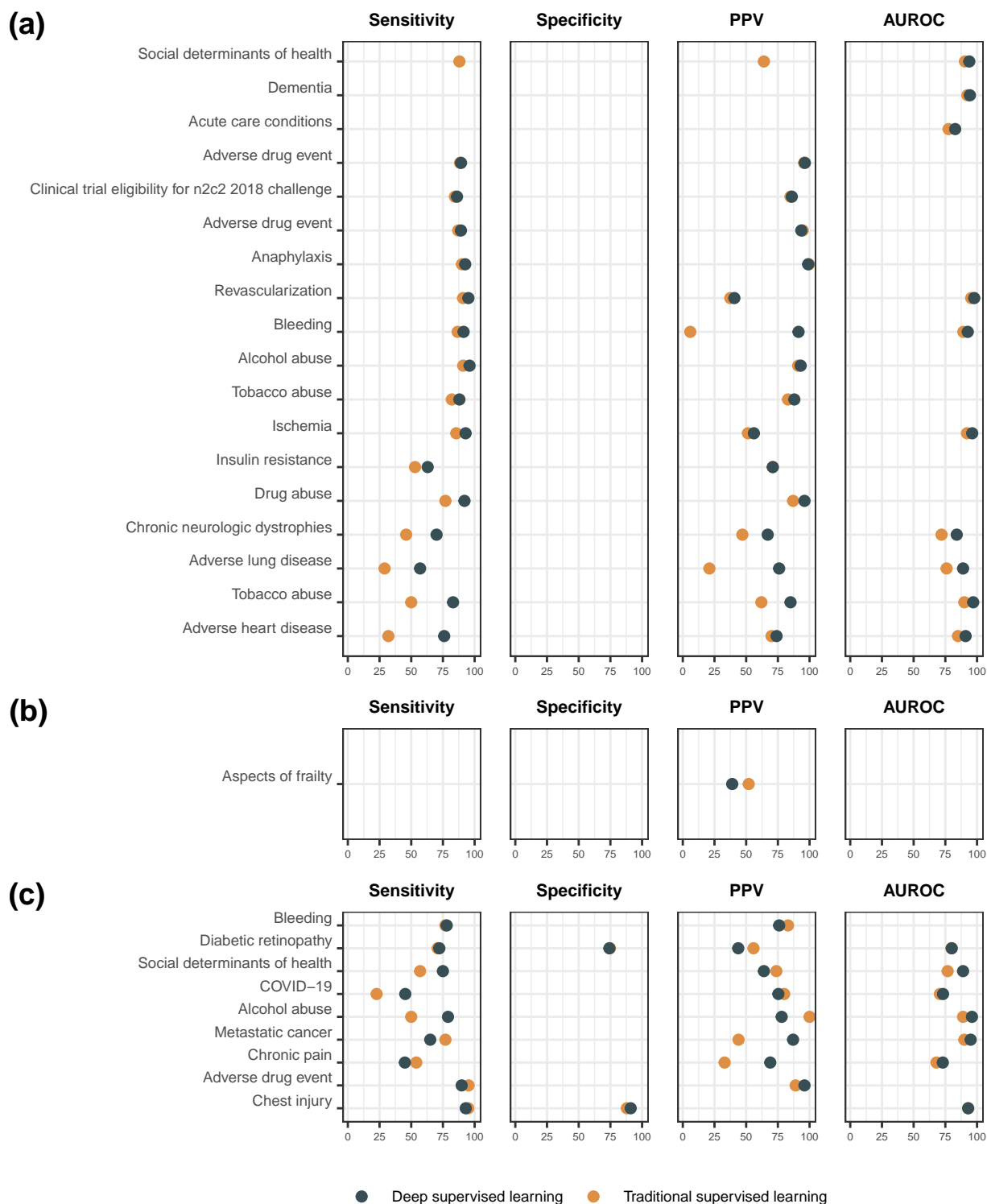
6.3 Weakly-supervised ML vs. rule-based algorithms

6.4   Weakly-supervised ML vs. traditional supervised ML

7   Model performance metric reporting

Model performance metrics	Supervised Deep learning	Supervised Traditional machine learning	Weakly-supervised Deep learning	Weakly-supervised Traditional machine learning	Semi-supervised Traditional machine learning	Count
Precision	26	23	0	8	4	61
Recall	26	23	1	7	2	59
AUROC	11	15	1	10	5	42
F-score	26	9	0	7	0	42
Specificity	7	11	1	1	0	20
Accuracy	5	8	1	4	0	18
NPV	1	7	0	5	2	15
AUPRC	4	2	0	2	0	8
Calibration plots	2	3	0	0	0	5
Log loss	1	1	0	0	1	3
Brier score	1	1	0	0	0	2
Hamming loss	2	0	0	0	0	2
Matthews Correlation Coefficient	1	1	0	0	0	2
Normalized discounted cumulative gain	1	1	0	0	0	2
Binary preference	0	1	0	0	0	1
FPR	0	0	0	1	0	1
Kappa	1	0	0	0	0	1
Label rank loss	0	0	0	1	0	1
Mean absolute error	0	1	0	0	0	1
Mean reciprocal rank	1	0	0	0	0	1
MSE	0	1	0	0	0	1
RMSE	0	1	0	0	0	1