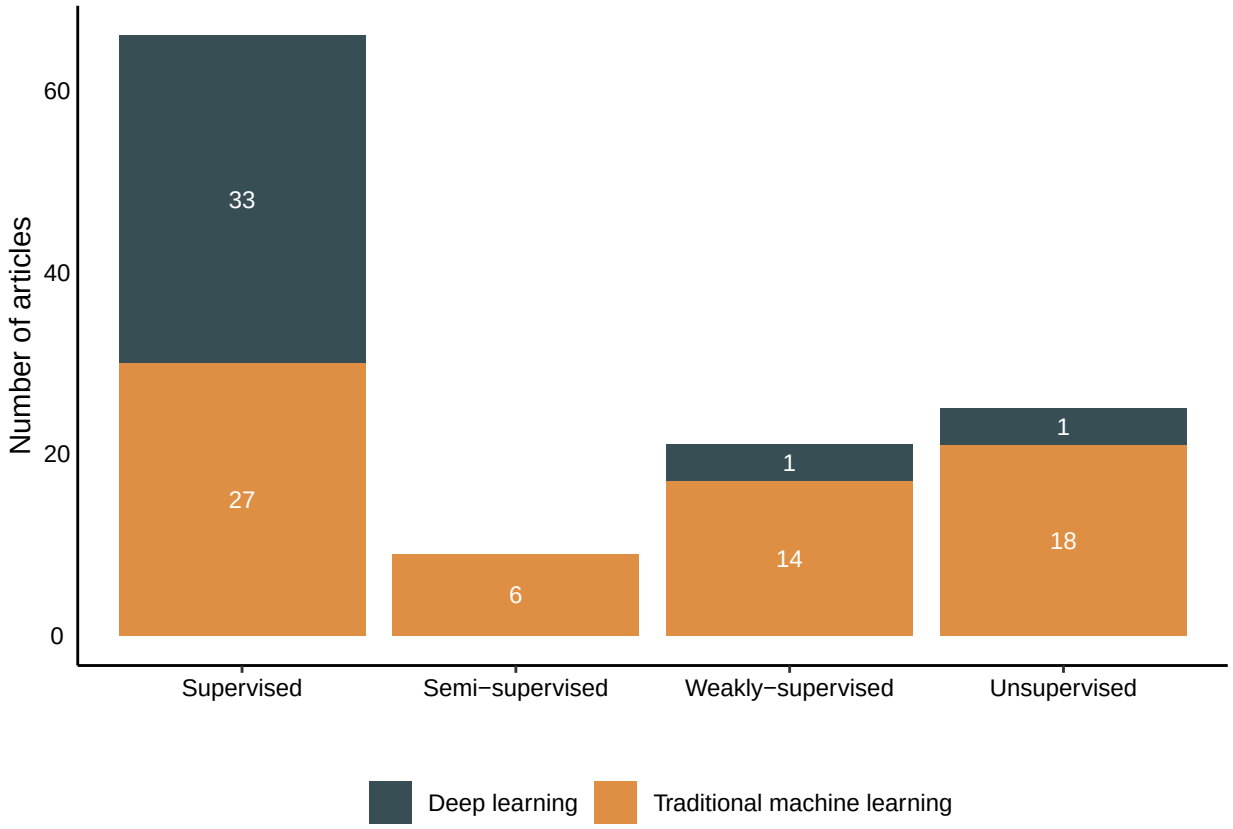
A summary of EHR-based phenotyping article annotation

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"
```

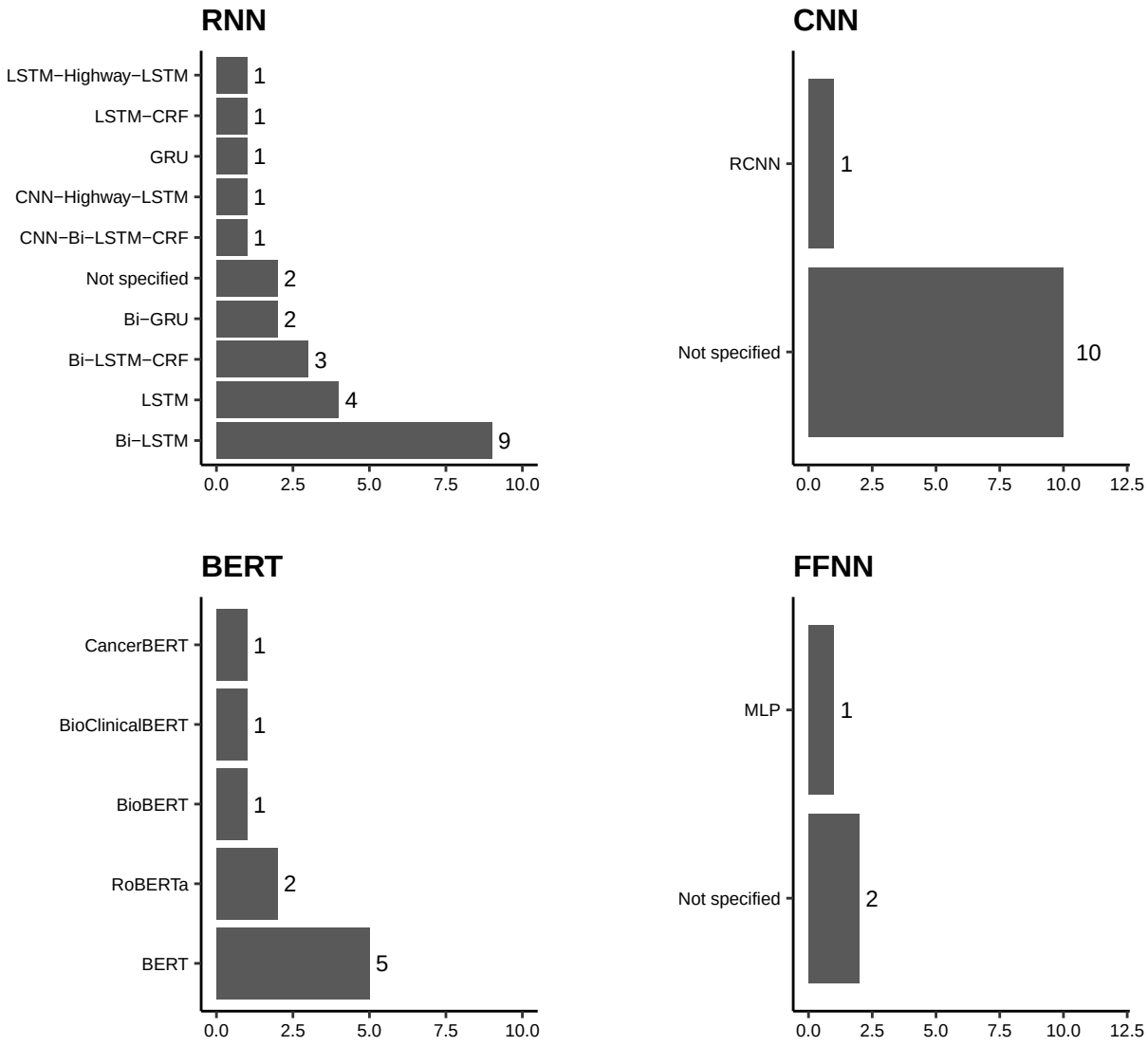
1.2 DL method

Table 2: Deep supervised learning methods

DL method	ML	Count
BERT	Supervised	7
CNN	Supervised	11
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

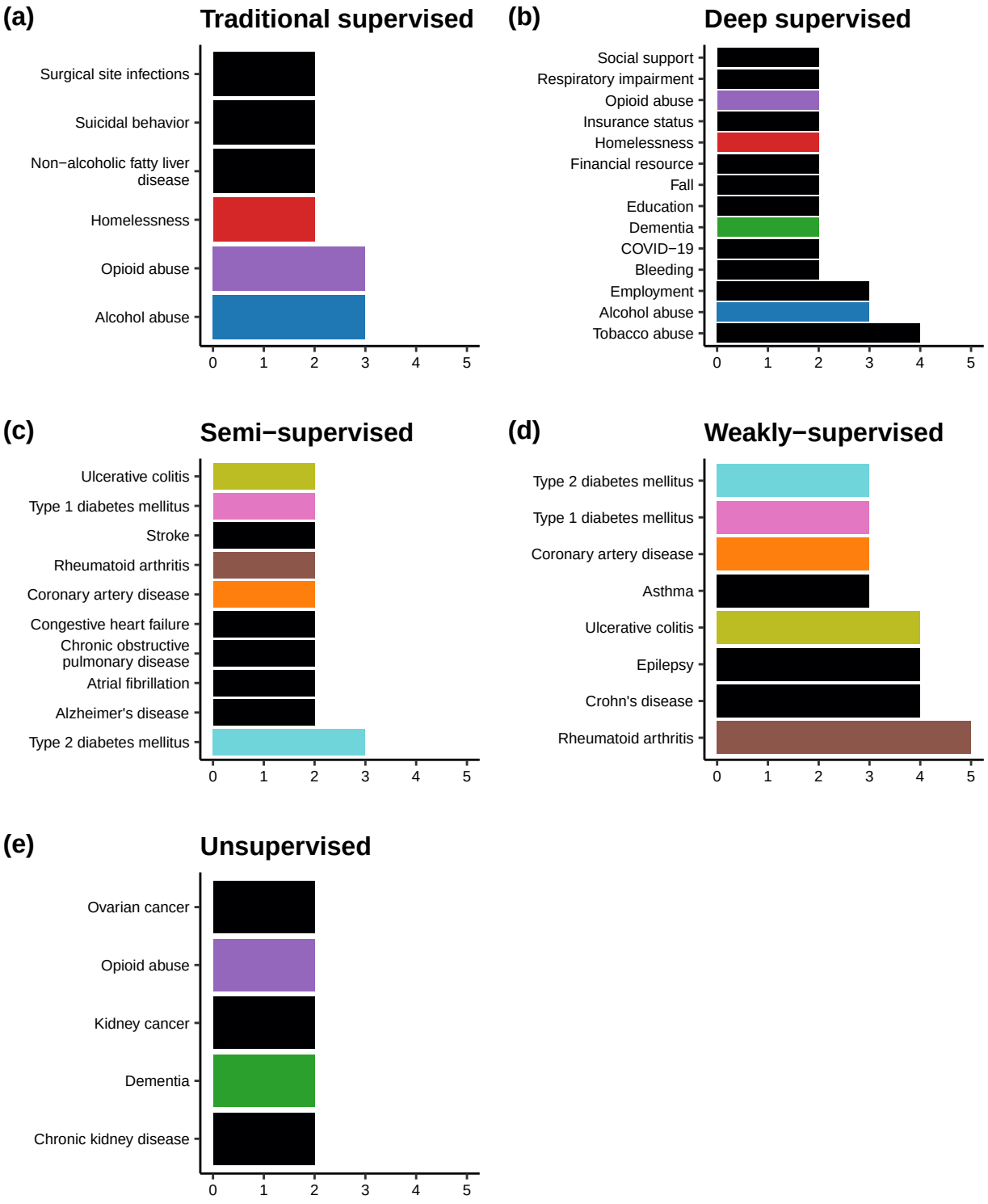
```
## [1] 156
```

```
## [1] 40
```

```
## [1] 69
```

```
## [1] 4
```

```
## [1] 11
```

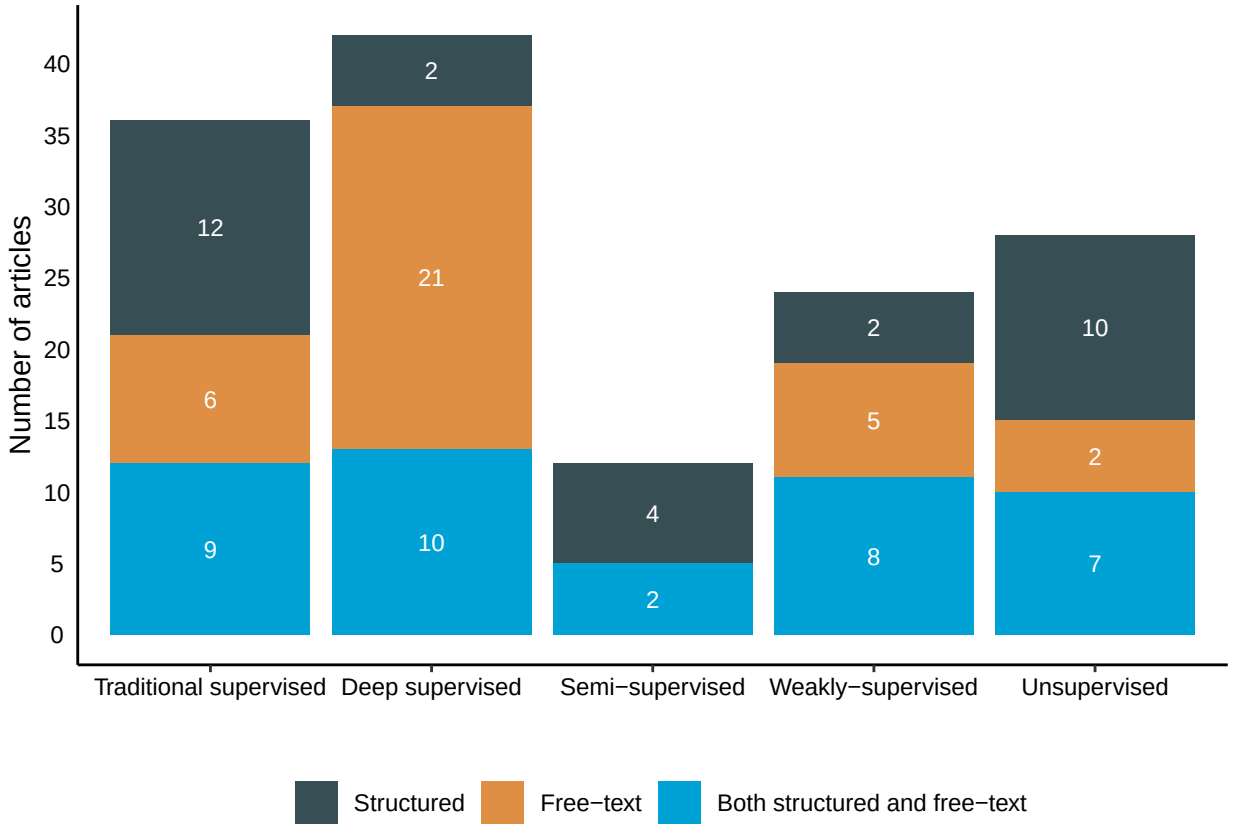


2.1 More nuanced phenotype

	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	1	1	22	10	0	13	4
DSL	33	31	18	11	1	9	12	2	20	9	9
SSL	6	2	1	0	0	0	6	1	0	3	0
WSL	15	13	10	0	3	2	10	8	1	4	3
USL	19	9	4	0	3	3	13	0	0	15	4
Total	100	70	47	14	8	15	63	21	21	44	20

3 Data source

3.1 Summary

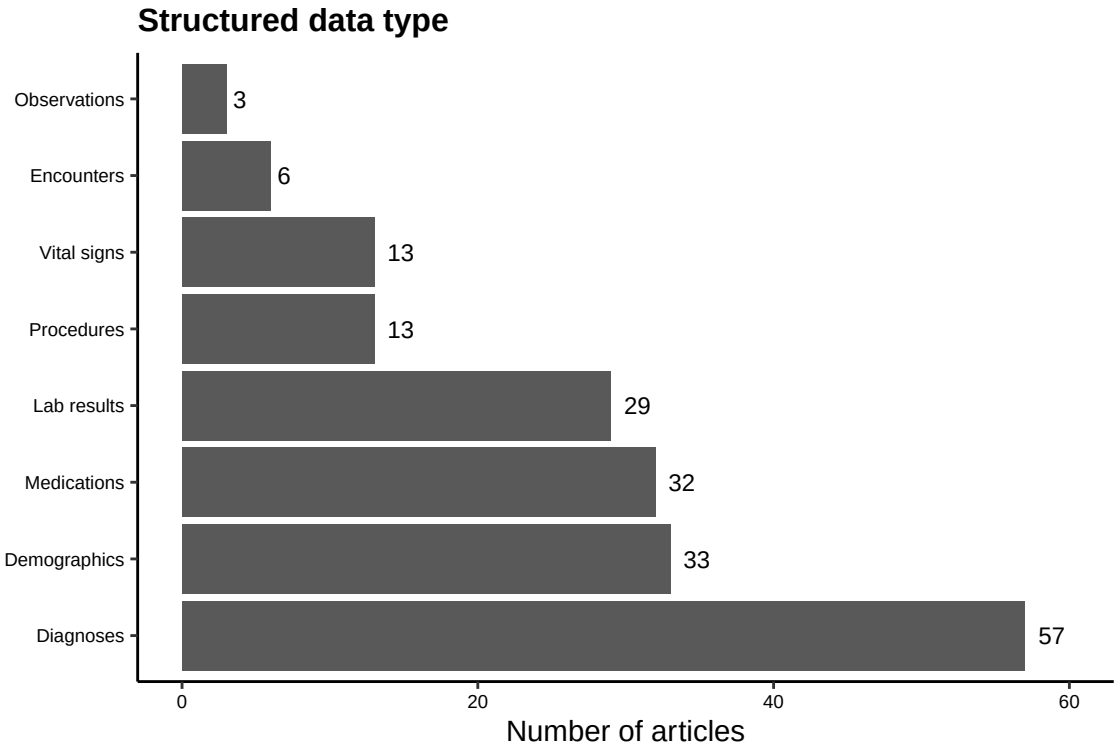


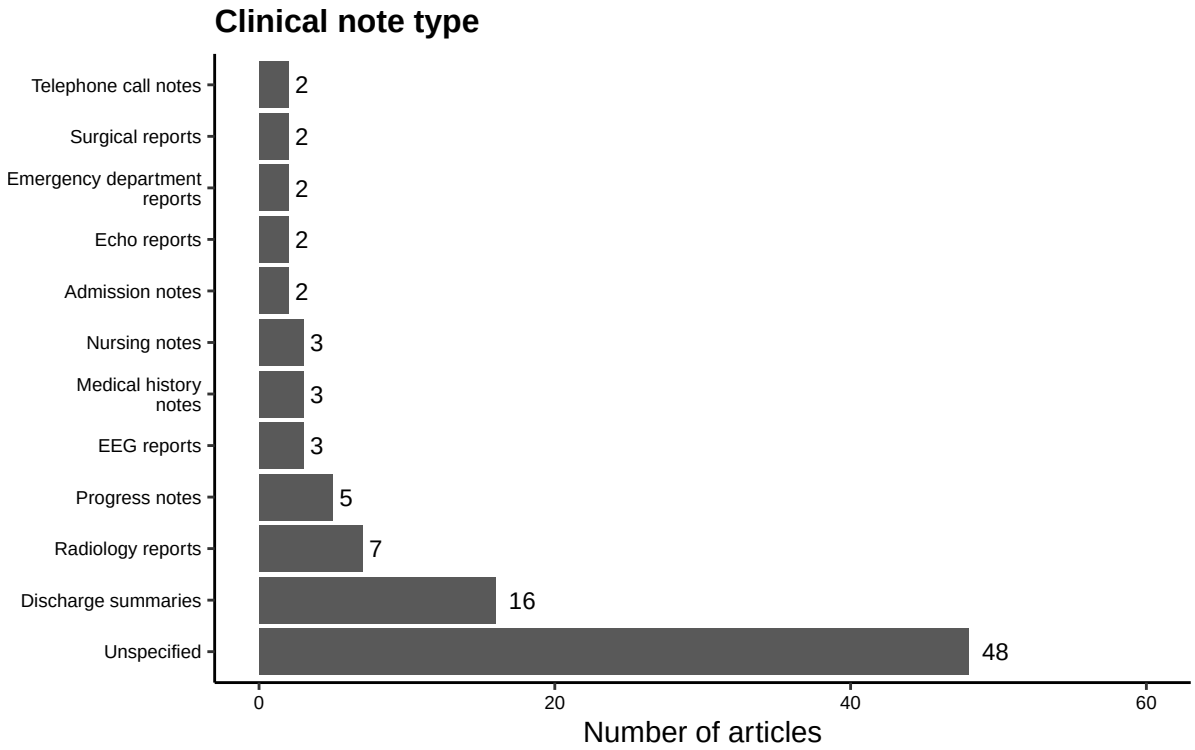
TSL = Traditional supervised learning. DSL = Deep supervised learning. DRL = Reinforcement deep learning. SSL = Semi-supervised learning. WSL = Weakly-supervised learning. US = Unsupervised learning.

3.2 Structured and unstructured data type

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

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


Structured data type
Observations
Encounters
Vital signs
Procedures
Lab results
Medications
Demographics
Diagnoses
3
6
13
13
29
32
33
57
0
20
40
60
Number of articles


Clinical note type
Telephone call notes
Surgical reports
Emergency department reports
Echo reports
Admission notes
Nursing notes
Medical history notes
EEG reports
Progress notes
Radiology reports
Discharge summaries
Unspecified
2
2
2
2
2
3
3
3
5
7
16
48
0
20
40
60
Number of articles

3.3 Institutions

Country	Count
US	94
France	2
Cancada	1
China	1
Germany	1
Israel	1
Italy	1
Korean	1
Netherland	1
Singapore	1
Spain	1

3.4 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

```
## [1] 14
```

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	9	1	1	1	3	15
MTSamples database	1	0	0	0	0	1

Terminology unnested	Supervised Traditional machine learning	Unsupervised Traditional machine learning	Supervised Deep learning	Weakly-supervised Traditional machine learning	Semi-supervised Traditional machine learning	Count
ICD-9	17	6	7	4	4	38
UMLS	11	3	8	8	1	31
ICD-10	11	1	4	1	3	20
SNOMED-CT	2	3	4	3	0	12
RxNorm	3	1	2	2	1	9
CPT	2	0	3	2	0	7
Phecode	0	2	0	3	2	7
ICD	0	1	0	4	0	5
ICD-9-CM	1	2	0	1	0	4
LOINC	3	0	0	1	0	4
ATC (Anatomical therapeutic chemical)	2	0	0	0	0	2
NDC (National drug codes)	2	0	0	0	0	2

4 Terminology

```
## [1] "There are 43 papers using multiple terminologies"
```

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

5 NLP software

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

6 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

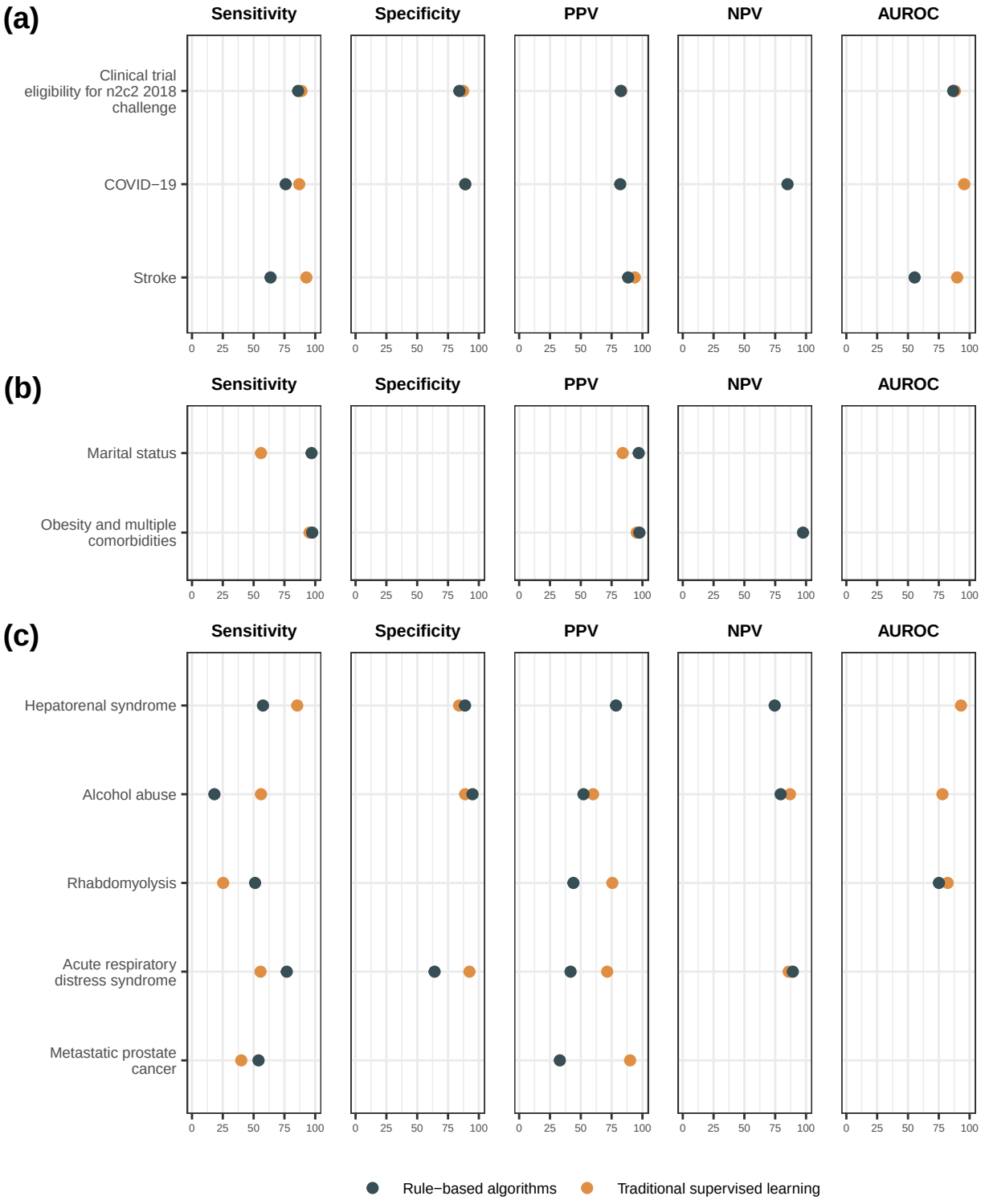
```
## [1] "There are 7 papers using multiple embedding training data"
```

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

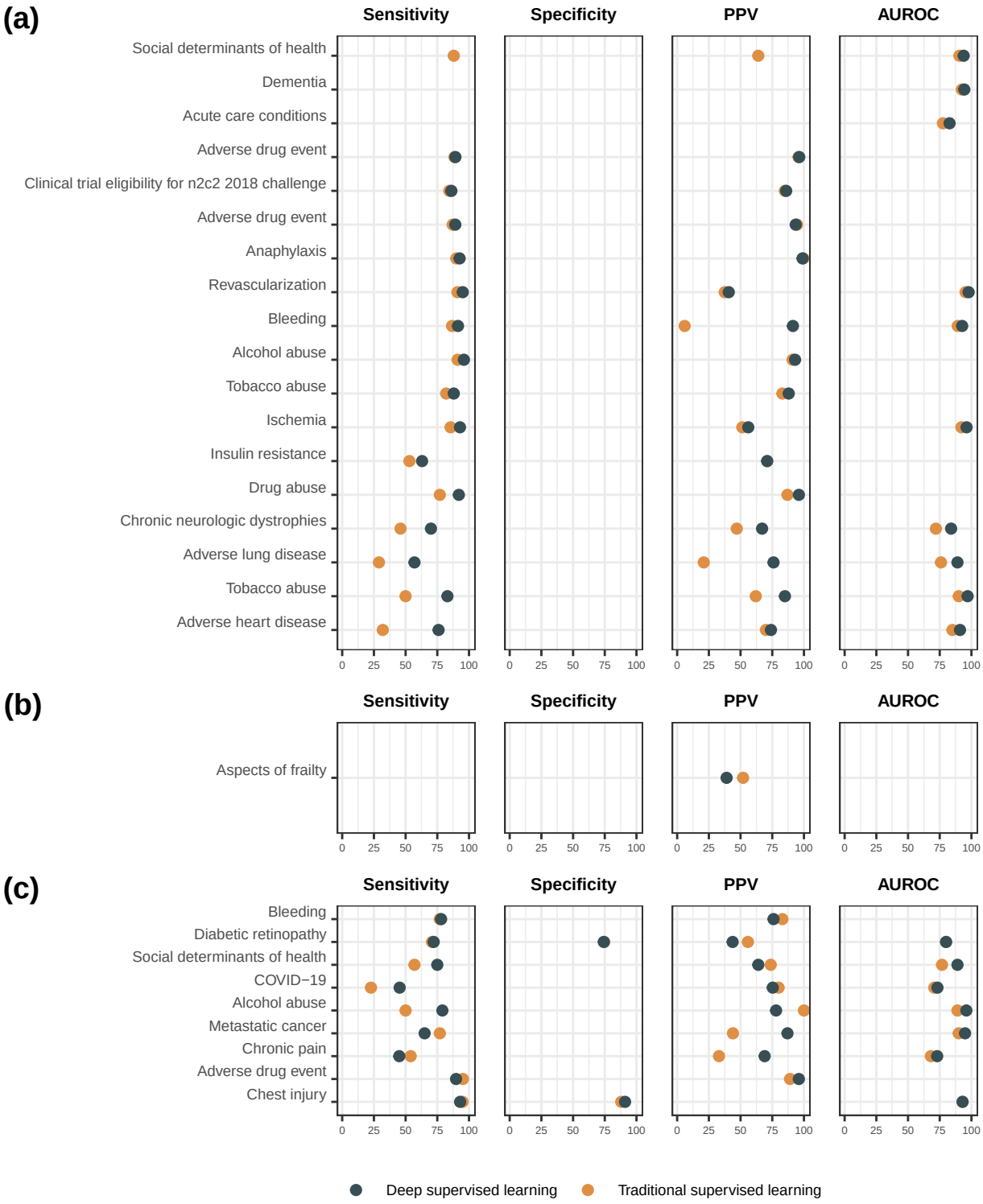
```
## [1] "There are 11 papers using multiple embedding training methods"
```

7 Validation and comparison

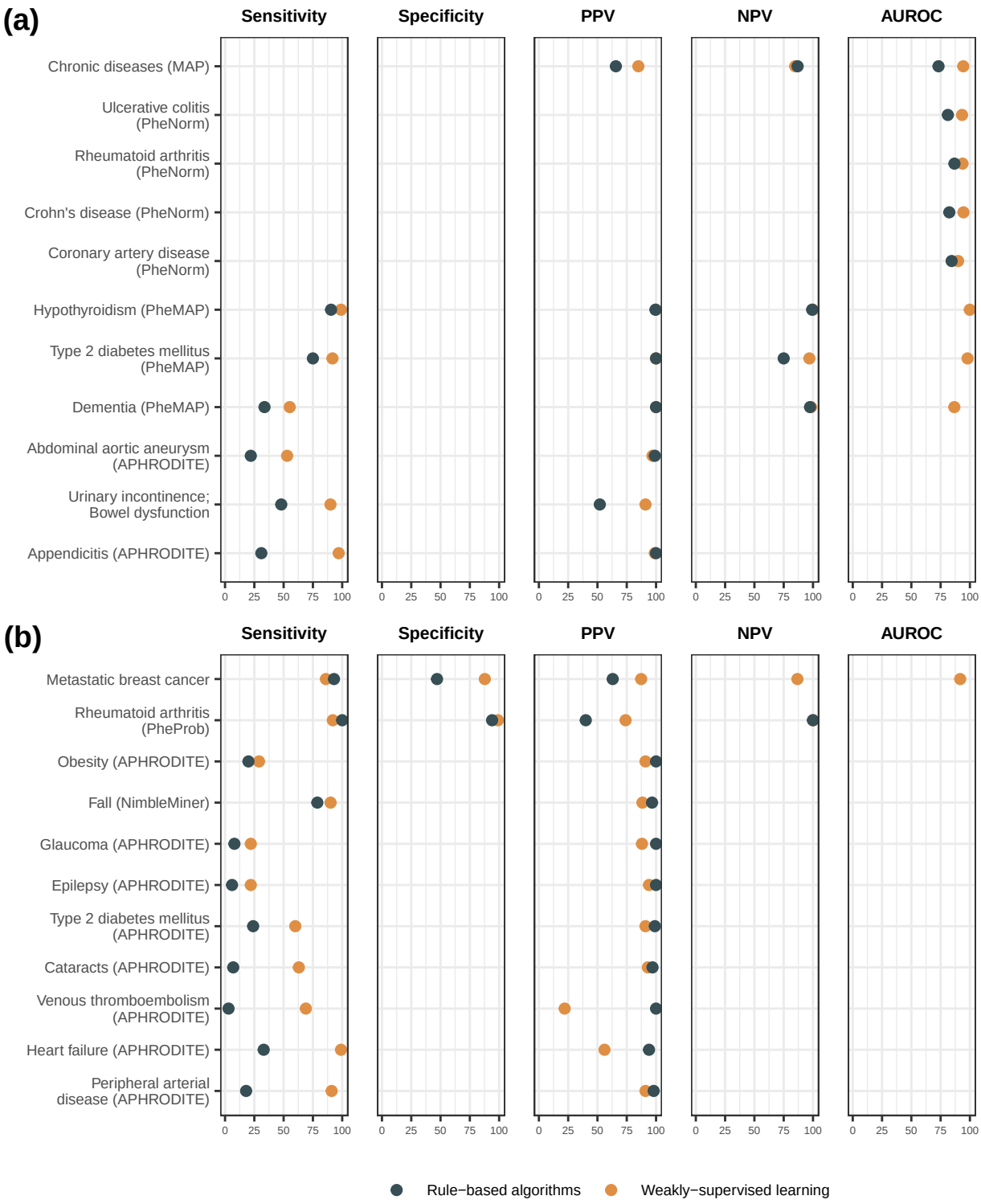
7.1 Traditonal supervised ML vs. rule-based



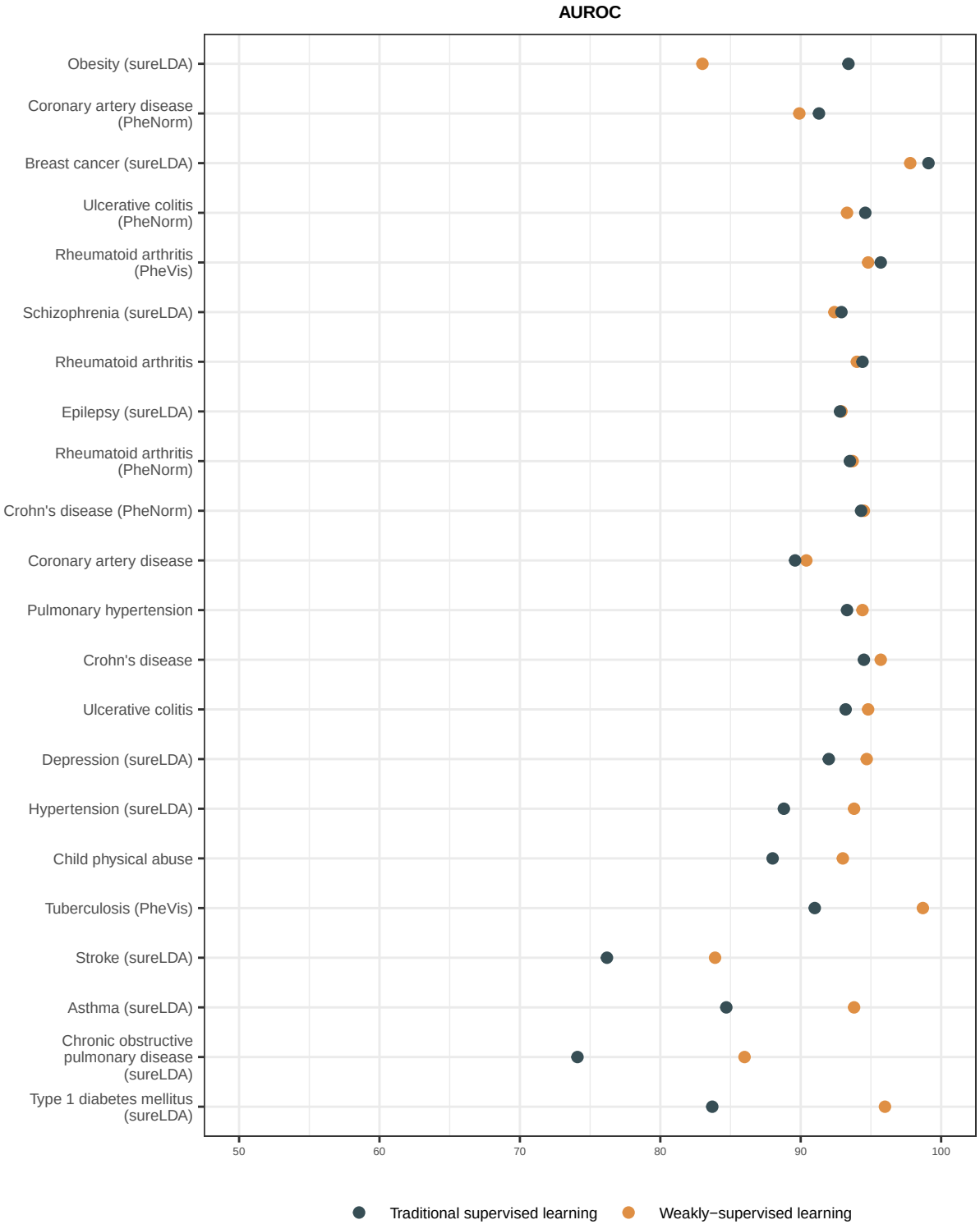
7.2 Deep supervised ML vs. traditional supervised ML



7.3 Weakly-supervised ML vs. rule-based algorithms



7.4 Weakly-supervised ML vs. traditional supervised ML



8 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	25	23	1	7	2	58
AUROC	11	15	1	10	5	42
F-score	26	9	0	7	0	42
Specificity	6	11	1	1	0	19
Accuracy	4	8	1	4	0	17
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