

# Evaluation of a Rule-Based Phenotyping Algorithm for Endometriosis



Ananya Rajagopalan<sup>1</sup>, Lindsay Guare<sup>1</sup>, Ashley Batugo<sup>2</sup>, Meredith Pollie<sup>3</sup>, Leigh Ann Humphries<sup>3</sup>, Penn Medicine Biobank, Danielle Mowery<sup>2</sup>, Suneeta Senapati<sup>3</sup>, Shefali Setia-Verma<sup>1</sup>

<sup>1</sup>Department of Pathology and Laboratory Medicine, Perelman School of Medicine, Philadelphia, PA, United States

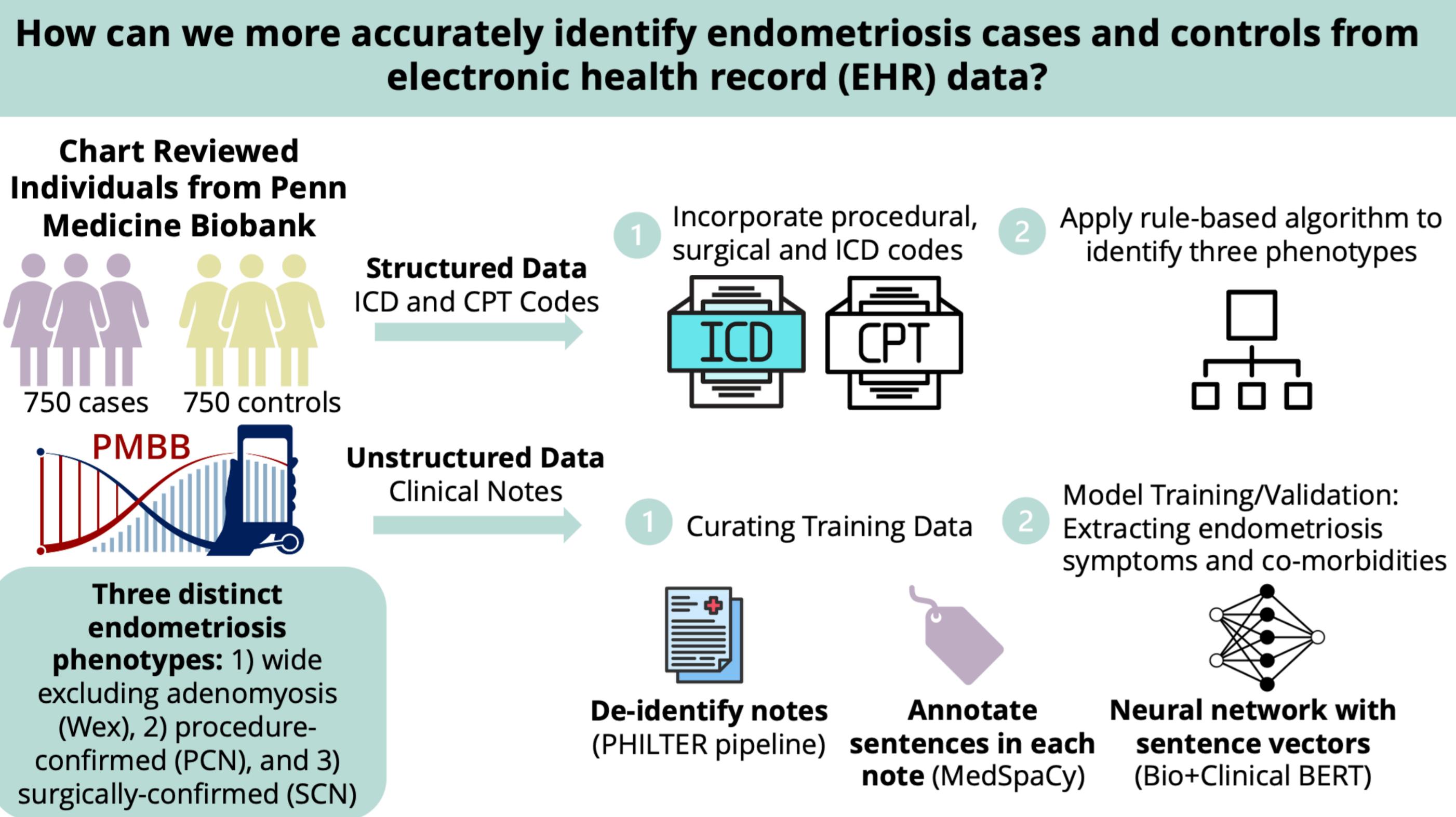
<sup>2</sup>Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, United States

<sup>3</sup>Department of Obstetrics and Gynecology, Hospital of the University of Pennsylvania, Philadelphia, PA

## Background

- Endometriosis is a complex and heterogeneous condition, posing challenges for accurate case identification with a **baseline misclassification rate of 10%**.
- Electronic health record (EHR)-based phenotyping algorithms can facilitate large-scale studies but require rigorous validation.
- Here, we report the performance of a **rule-based algorithm in generating diverse endometriosis phenotypes**, by incorporating procedural and surgical codes along with diagnosis codes.
- We also utilize unstructured patient notes data to extract **sentence-level symptoms and co-morbidities** that can inform more accurate phenotyping.

## Graphical Abstract



## Methods

### Phenotyping

From the Penn Medicine Biobank (PMBB) dataset, we selected **750 women with and 750 women without International Classification of Disease (ICD) codes for endometriosis** (N80.\* for ICD-10 and 617.\* for ICD-9). These individuals were considered in chart reviews to classify their status of endometriosis (case or control).

### Structured Data: ICD and CPT Codes

We evaluated a previously developed, expert-validated phenotyping algorithm that incorporates ICD and CPT codes related to endometriosis diagnosis to identify three distinct phenotypes in this cohort: 1) wide excluding adenomyosis (Wex), 2) procedure-confirmed (PCN), and 3) surgically-confirmed (SCN).

### Unstructured Data: Clinical Notes

For this same set of PMBB participants, we extracted and de-identified clinical notes (e.g., progress notes, discharge summaries, surgical and imaging reports) starting in 2011 using the PHILTER pipeline. Individuals were split 70% training / 10% validation / 20% final test set.

We used MedSpaCy to obtain sentence-level annotations for 14 features (endometriosis symptoms and co-morbidities) labeling positive, negative, or absent status with a 10-token context window to reduce false negatives. These sentence-level labels were checked for accuracy on the patient-level and then used to train a neural network, with 768-dimension sentence vectors from Bio+Clinical BERT. The joint model architecture (Figure 1) was designed to leverage correlations of measured outcomes, and training used a log-loss function with a 1:2 ratio of informative to noisy sentences (Table 1).

Total Unique Sentences	880,978
Noise: Unique Sentences with no relevant information	863,617
Sampled Noise: Unique sentences sampled with no relevant labels	34,722
Signal: Unique sentences with any correct labels	17,361
Total sentences trained on	52,083

Table 1: Sentence Counts from Clinical Notes

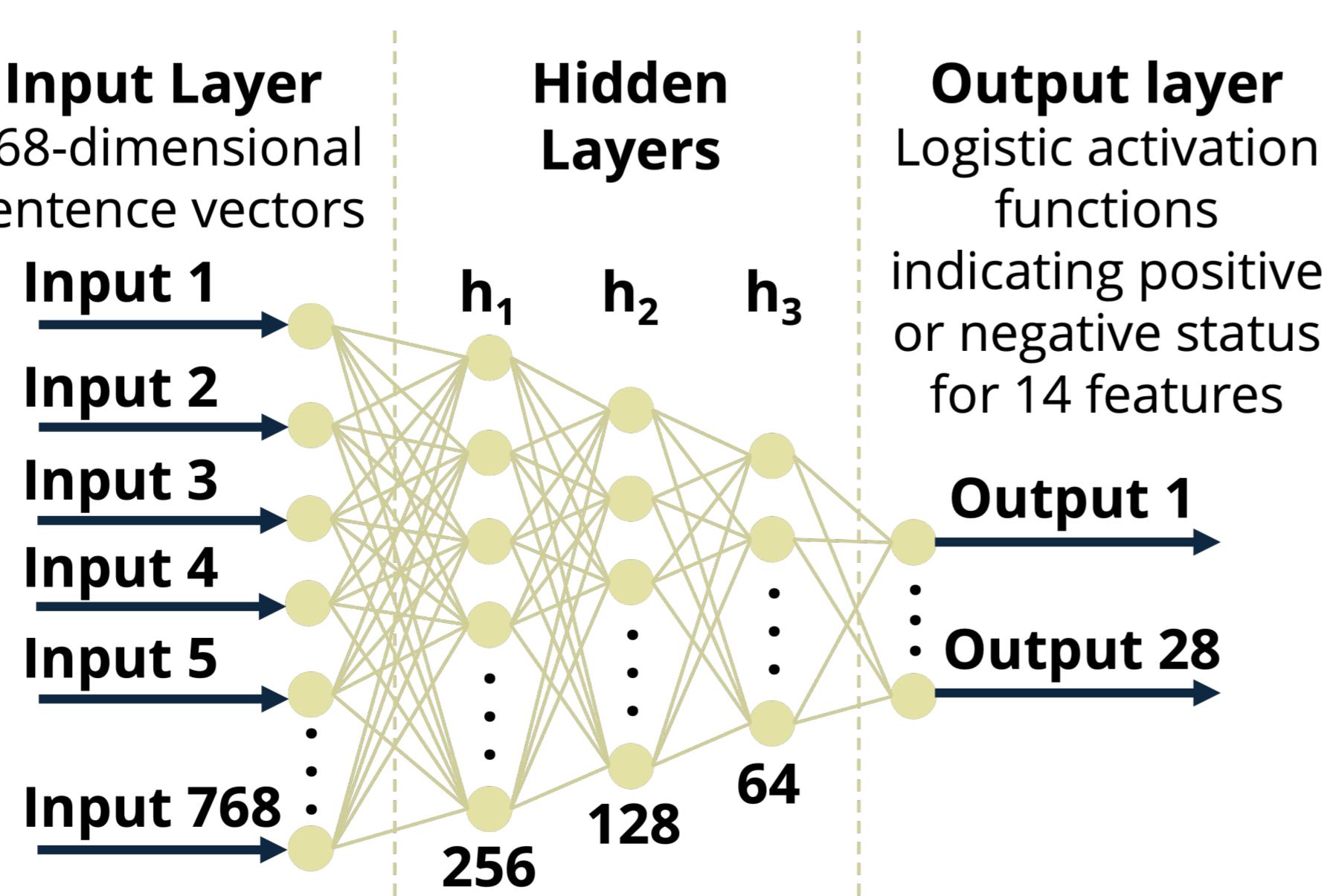
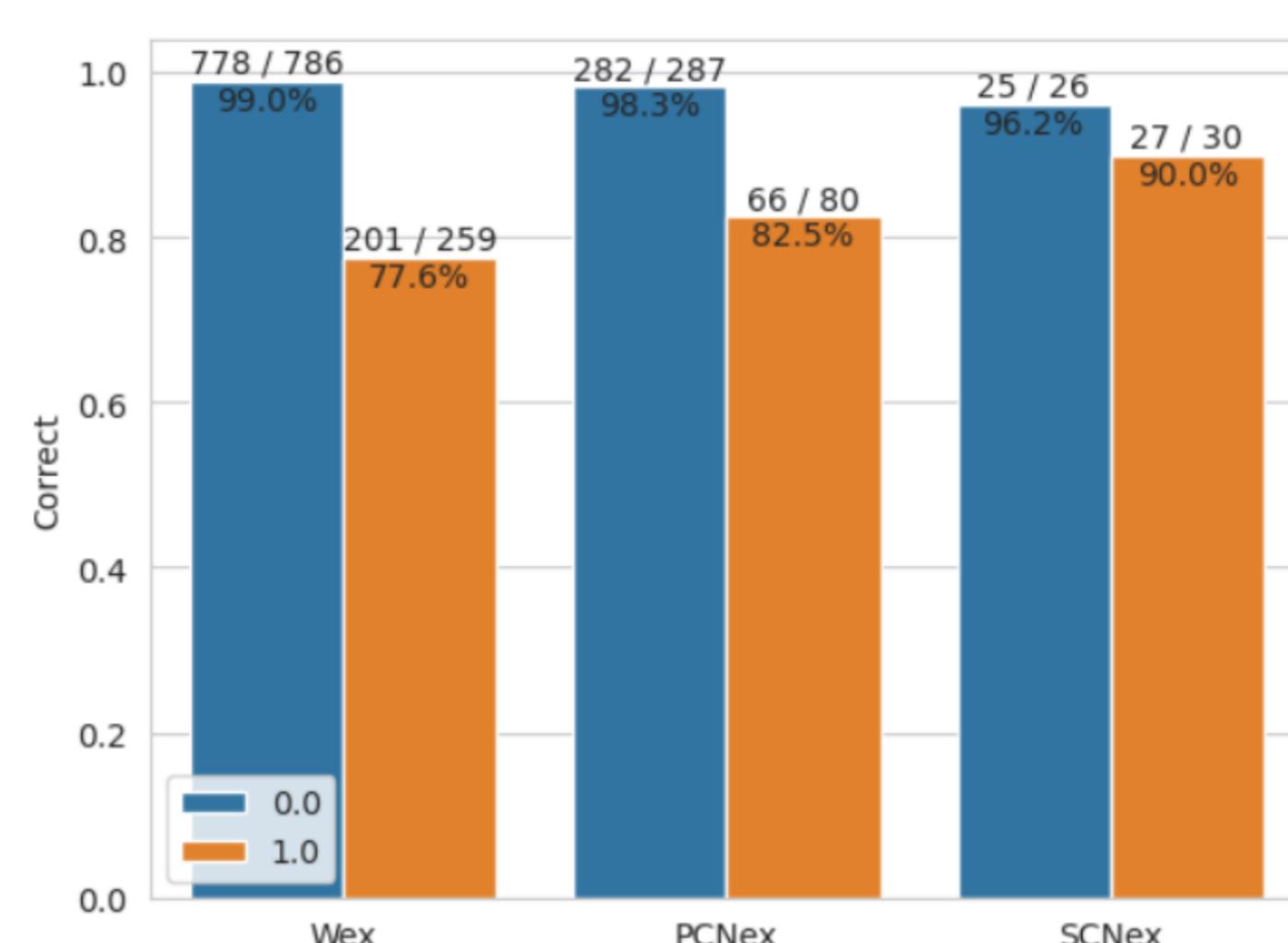


Figure 1: Joint Model Architecture (Multilayer Perceptron)

Model thresholds were fine-tuned with five-fold cross-validation. Sentence-level predictions were evaluated using area under the receiver-operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), single-point precision, sensitivity, and balanced accuracy.

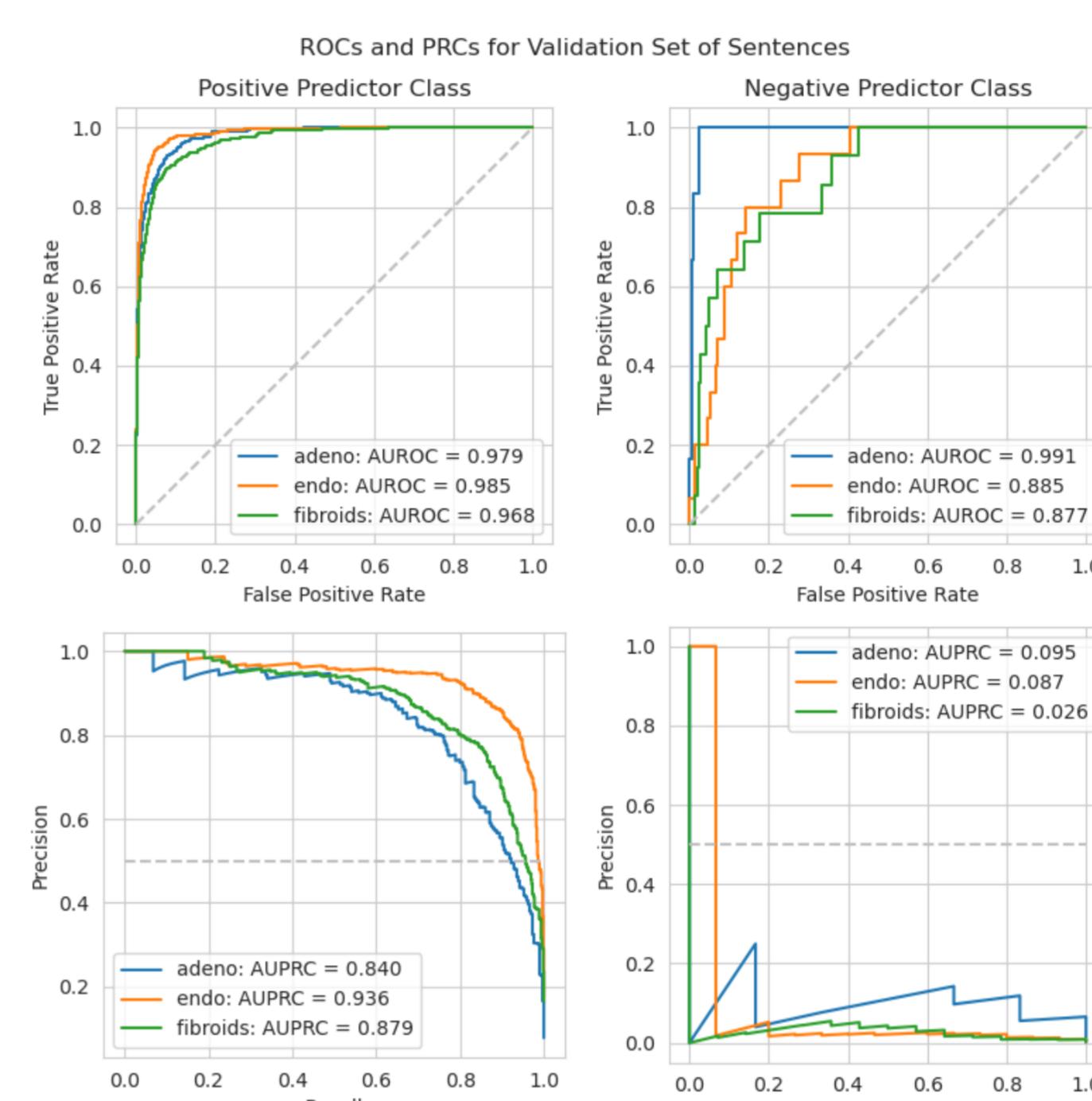
## Results

### Classifier Performance Using Structured Data



This rule-based EHR phenotyping algorithm successfully identified diverse endometriosis phenotypes with high accuracy, as demonstrated by the strong NPV and PPV values. The wide exclusion phenotype can facilitate large-scale genetic and epidemiological studies, while the procedure-confirmed and surgically-confirmed phenotypes offer greater specificity for in-depth biological investigations.

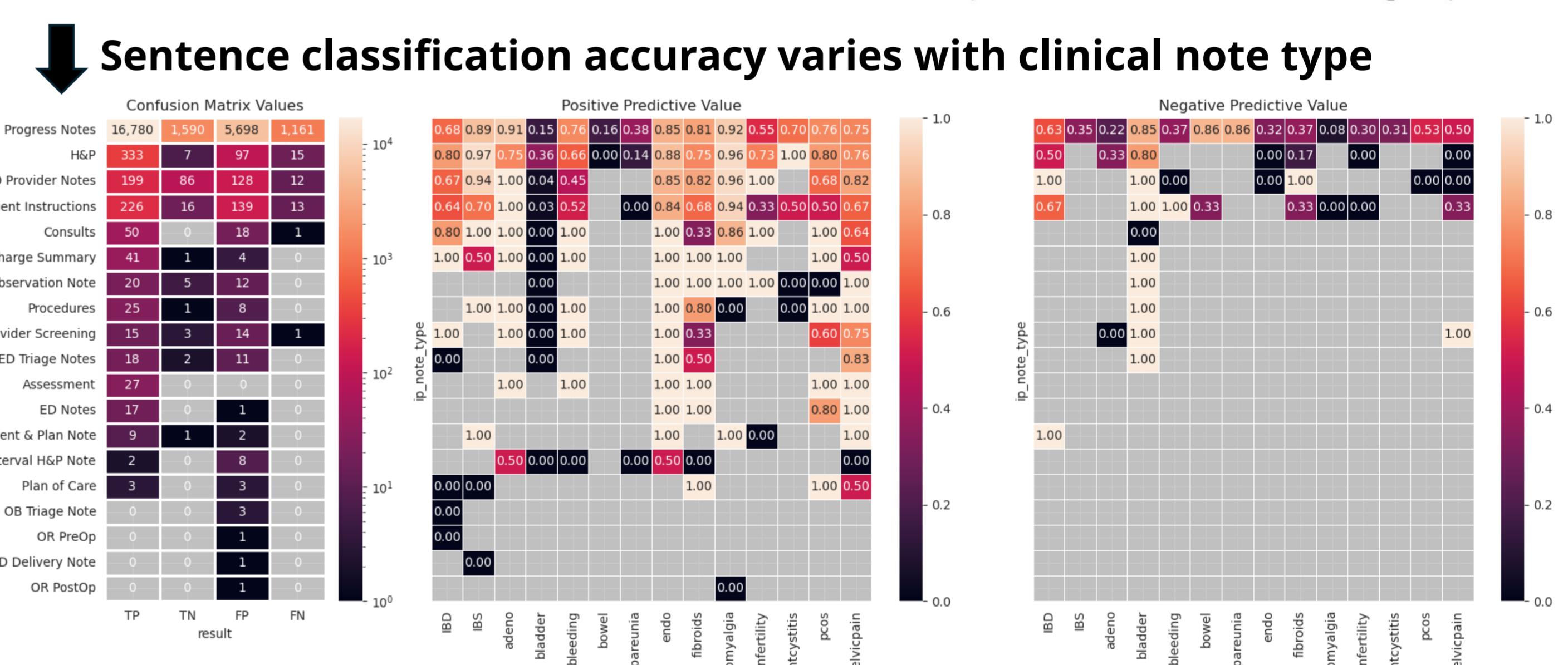
### Clinical Notes: Sentence-based Symptom Extraction



Validation of the neural network demonstrates that positive sentence classification is consistently accurate based on the AUROC and AUPRC scores. On the other hand, negative sentence classification predictions are imprecise (low AUPRC). This may be due to the lower likelihood of negative sentences being documented in a patient's note (e.g., "Patient X does not have uterine fibroids").

Comparison of labeling of symptoms with notes compared to getting labels using diagnosis codes, based on quality metrics of sensitivity, specificity, precision, and balanced accuracy.

**Key Takeaway:** The neural network model underperformed in classifying the endometriosis symptoms and co-morbidities on an individual level compared with diagnosis codes.



## Conclusions

- The rule-based EHR phenotyping algorithm using structured data successfully identified diverse endometriosis phenotypes with high accuracy, as demonstrated by the strong NPV and PPV values.
- While sentence-level classification accuracy is strong, sentence-level noise across clinical notes is high compared to the signal of informative sentences, which may be driving difficulty in generating patient-level endometriosis phenotypes from multiple clinical notes.
- This robust EHR-based approach enables efficient and scalable endometriosis research across diverse clinical settings.

## Future Work

- We plan to use the symptoms and co-morbidities of endometriosis that were extracted from the clinical notes in a rule-based classifier model, which may further improve endometriosis phenotyping accuracy.

## Acknowledgements

- Research reported in this poster was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number R01HD110567-01A0.
- We acknowledge the Penn Medicine Biobank (PMBB) for providing data and thank the patient-participants of Penn Medicine who consented to participate in this research program.
- The PMBB is approved under IRB protocol #813913 and # 852155 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878.