Natural language processing models can be trained to accurately recognize the presence of disease within clinical notes

David Vidmar[1], Will Thompson[1], Ruijun Chen[1], Dustin Hartzel[2], Daniel Rocha[2], Joseph Leader[1], Brandon Fornwalt[1], Christopher M Haggerty[2]

[1]Tempus Labs, Chicago, IL, USA, [2]Geisinger, Danville, PA, USA

TEMPUS

INTRODUCTION

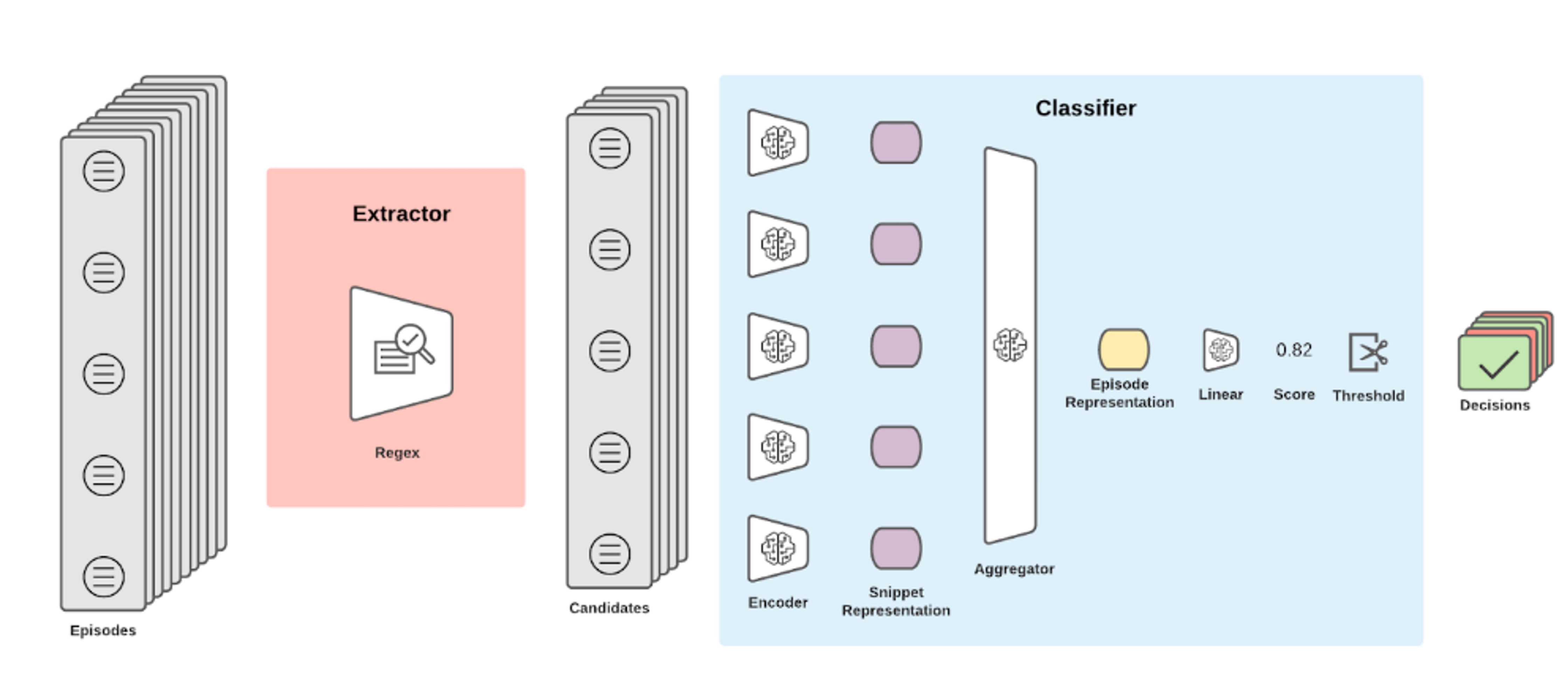
- Automated identification of clinical disease in electronic health record (EHR) data is vital to population health management and machine learning model development
- Rules-based approaches to identifying diagnoses using billing codes suffer from poor generalizability
- We hypothesized that natural language processing (NLP) models could be used to detect atrial fibrillation diagnoses directly from clinical notes
- This approach may facilitate scaling the identification of disease diagnoses across large amounts of clinical data for improved generalizability and also support health record de-identification efforts by automatically extracting important medical concepts

METHODS

- We collected clinical notes from a regional health system into a training set of roughly 29 million code-labeled episodes and a hold-out set of roughly 1.8 million code-labeled episodes
- We trained an NLP model on the training set, consisting of an extractor stage which identifies candidate episodes that are passed to a classifier for adjudication
- Model performance was computed using the code-based labels on the hold-out set, with "un-extracted" episodes scored as zero
- We also performed targeted blinded chart reviews of disagreements between the NLP model output and the code-based labels

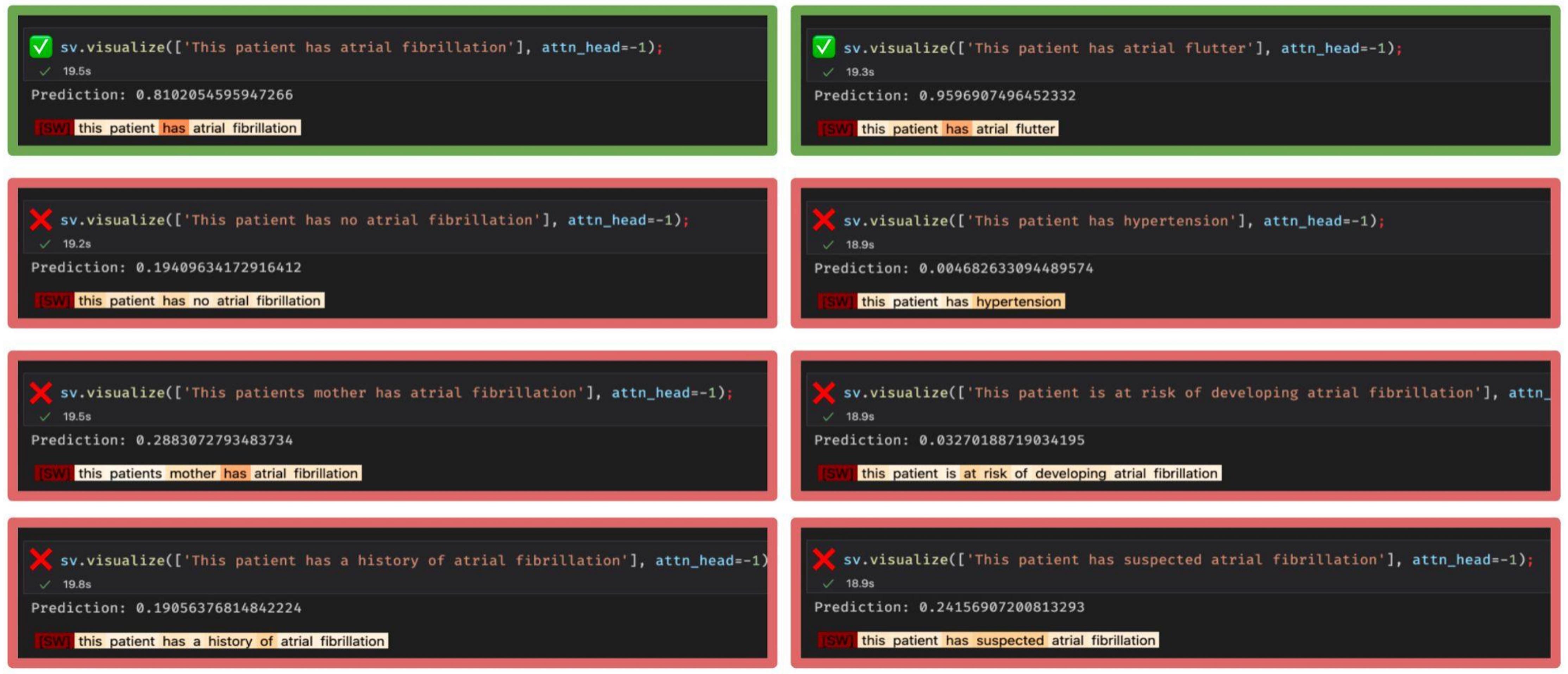
RESULTS

Figure 1. Model diagram



The model follows an extractor-classifier pattern where a regex-based extractor narrows down the universe of episodes to a smaller candidate pool. The classifier is tasked with distinguishing between incidental and positive disease mentions in this candidate pool.

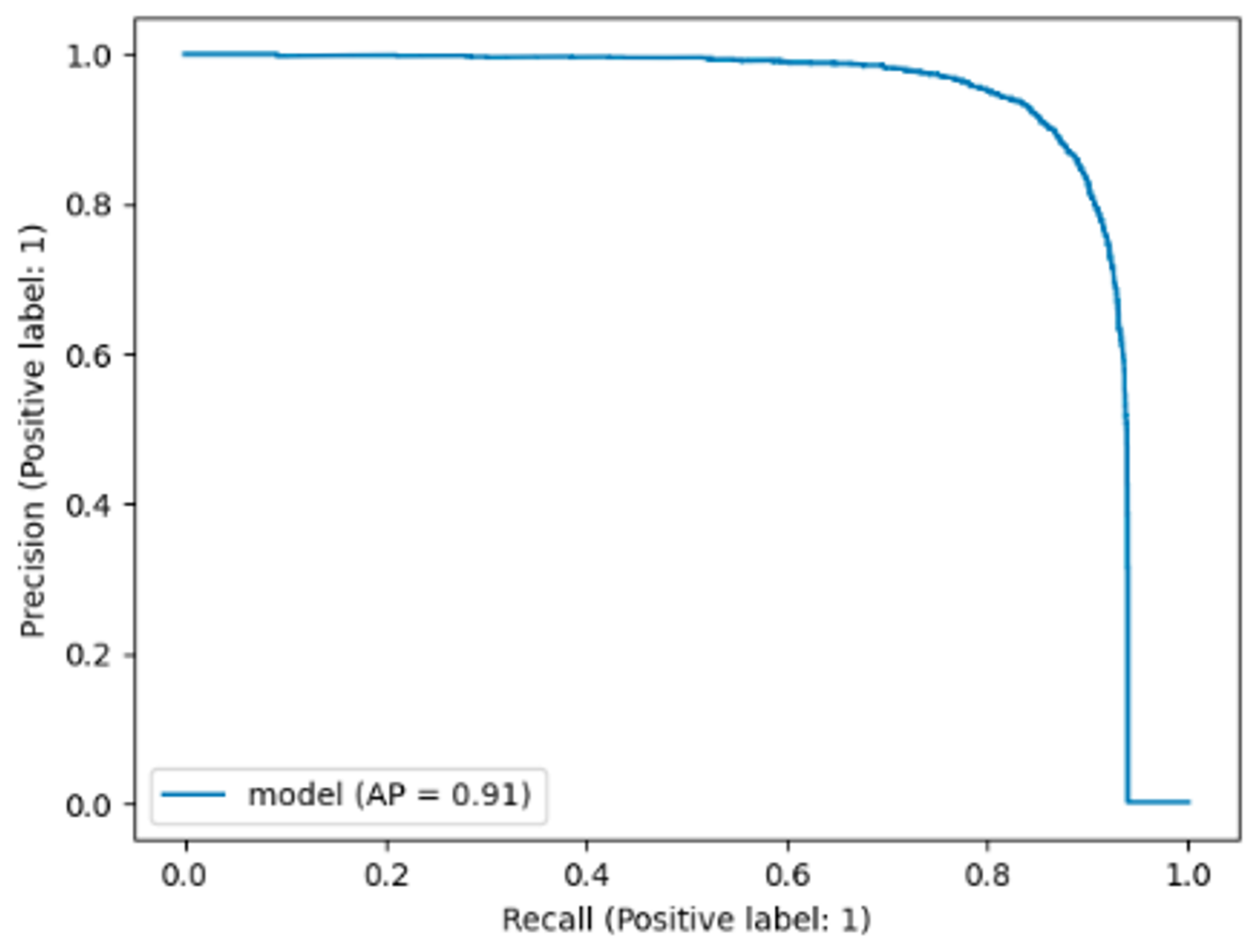
Figure 2. Model stress test



```
✅ sv.visualize(['This patient has atrial fibrillation'], attn_head=-1);
  ✓ 19.5s
Prediction: 0.8102054595947266
[SW] this patient has atrial fibrillation

✅ sv.visualize(['This patient has atrial flutter'], attn_head=-1);
  ✓ 19.3s
Prediction: 0.9596907496452332
[SW] this patient has atrial flutter

❌ sv.visualize(['This patient has no atrial fibrillation'], attn_head=-1);
  ✓ 19.2s
Prediction: 0.19409634172916412
[SW] this patient has no atrial fibrillation

❌ sv.visualize(['This patient has hypertension'], attn_head=-1);
  ✓ 18.9s
Prediction: 0.004682633094489574
[SW] this patient has hypertension

❌ sv.visualize(['This patients mother has atrial fibrillation'], attn_head=-1);
  ✓ 19.5s
Prediction: 0.2883072793483734
[SW] this patients mother has atrial fibrillation

❌ sv.visualize(['This patient is at risk of developing atrial fibrillation'], attn_
  ✓ 18.9s
Prediction: 0.03270188719034195
[SW] this patient is at risk of developing atrial fibrillation

❌ sv.visualize(['This patient has a history of atrial fibrillation'], attn_head=-1)
  ✓ 19.8s
Prediction: 0.19056376814842224
[SW] this patient has a history of atrial fibrillation

❌ sv.visualize(['This patient has suspected atrial fibrillation'], attn_head=-1);
  ✓ 18.9s
Prediction: 0.24156907200813293
[SW] this patient has suspected atrial fibrillation
```

Interpretable model results on hypothetical text snippets demonstrate ability to distinguish between true positive and incidental atrial fibrillation mentions. Snippets outlined in green were labeled positive whereas snippets outlined in red were labeled negative. The heatmap behind each word represents model attention weights, with higher weight correlating with words the model found more important during classification.

Figure 3. Validation set area under the precision-recall curve (AUPRC) for 1.8 million hold-out episodes



- The NLP model achieved an AUPRC of 0.91
- After thresholding, the NLP model achieved 87% recall and 89% precision
- Blinded review of selected episodes showed the NLP model was correct in 90% of disagreements where the code-based approach incorrectly labeled negative

CONCLUSIONS

- NLP models can learn to automatically label the presence or absence of atrial fibrillation within clinical notes
- This may lead to greater accuracy and generalizability relative to code-based labeling methods