AGENCY FOR HEALTHCARE RESEARCH AND QUALITY

AHRQ

Using Natural Language Processing (NLP) to Code Patient Narratives: Capabilities and Challenges

Steven Martino and Osonde Osoba

October 7, 2021

Acknowledgements and Disclosure

This work was funded by an AHRQ Cooperative Agreement (2 U18 HS025920-01) and developed without commercial support

Our collaborators on this research were Dan Ish, Andy Parker, Marc Elliott, Rachel Grob, Ron Hays, Mark Schlesinger, and Dale Shaller

Drs. Martino and Osoba have no conflicts of interest, financial interests, or sponsorships related to this activity to disclose

The Role of NLP in Analyzing Patient Narratives

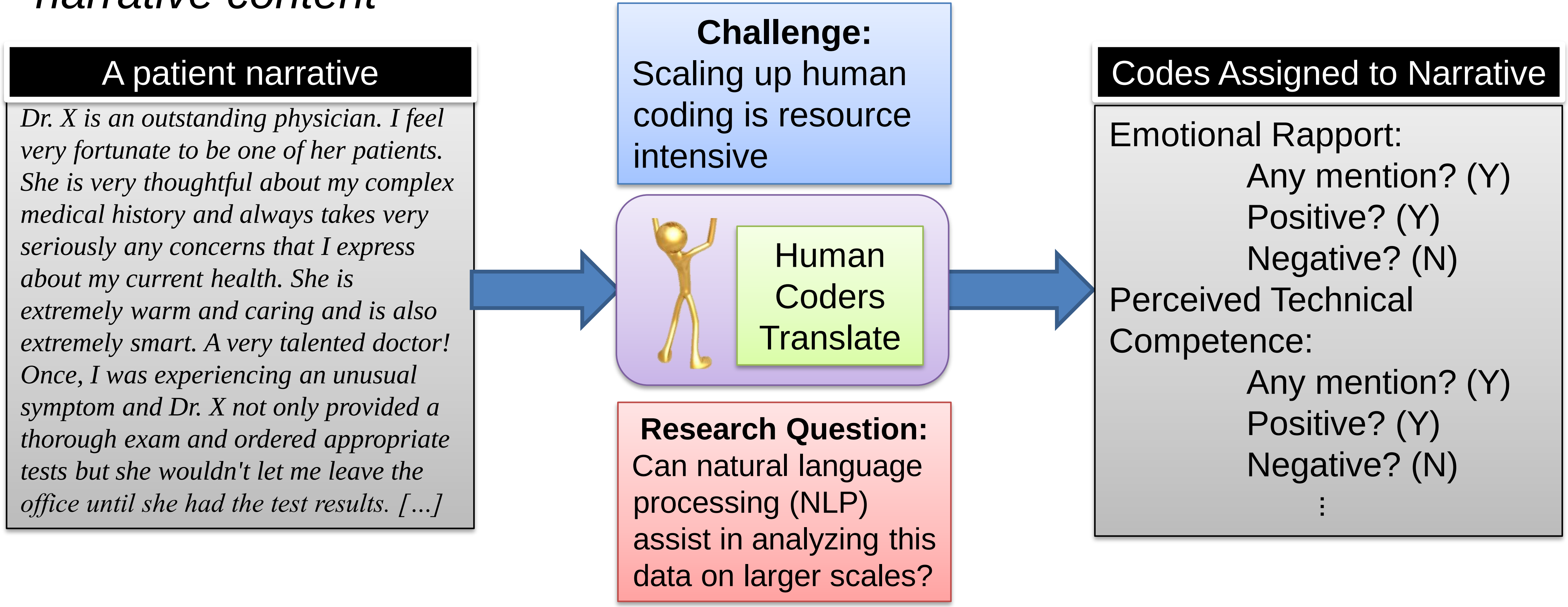
- The CAHPS research team has made great strides in developing rigorous methods for gathering and reporting narratives
- The team has also developed rigorous qualitative approaches to analyzing these data
- Such approaches might be prohibitive or impractical for health systems that collect tens of thousands of comments annually
- One potential solution is the use of computer algorithms to extract meaning from unstructured language, i.e., natural language processing

An Open-Access Demonstration of Capabilities and Challenges

- Use of NLP in this area is a relatively new undertaking; much of the value of this approach for analyzing narratives is yet to be realized
- For-profit firms have made important advances beyond what has been reported in the scientific literature
- These firms provide vital services to health care systems, but their need to protect intellectual property precludes them from being fully transparent about their methods
- Our objective was to provide an open-access demonstration of the feasibility and challenges of extracting key information from a large set of narratives using NLP

Data: patient narratives about healthcare experiences

- This study: responses to a standard set of five questions (CG-CAHPS NIS) administered as a portion of the Massachusetts Health Quality Partners (MHQP) Patient Experience Survey
- One approach to extracting this information: *humans assign codes based on narrative content*



A patient narrative

Dr. X is an outstanding physician. I feel very fortunate to be one of her patients. She is very thoughtful about my complex medical history and always takes very seriously any concerns that I express about my current health. She is extremely warm and caring and is also extremely smart. A very talented doctor! Once, I was experiencing an unusual symptom and Dr. X not only provided a thorough exam and ordered appropriate tests but she wouldn't let me leave the office until she had the test results. […]

Challenge:
Scaling up human coding is resource intensive

Human Coders Translate

Research Question:
Can natural language processing (NLP) assist in analyzing this data on larger scales?

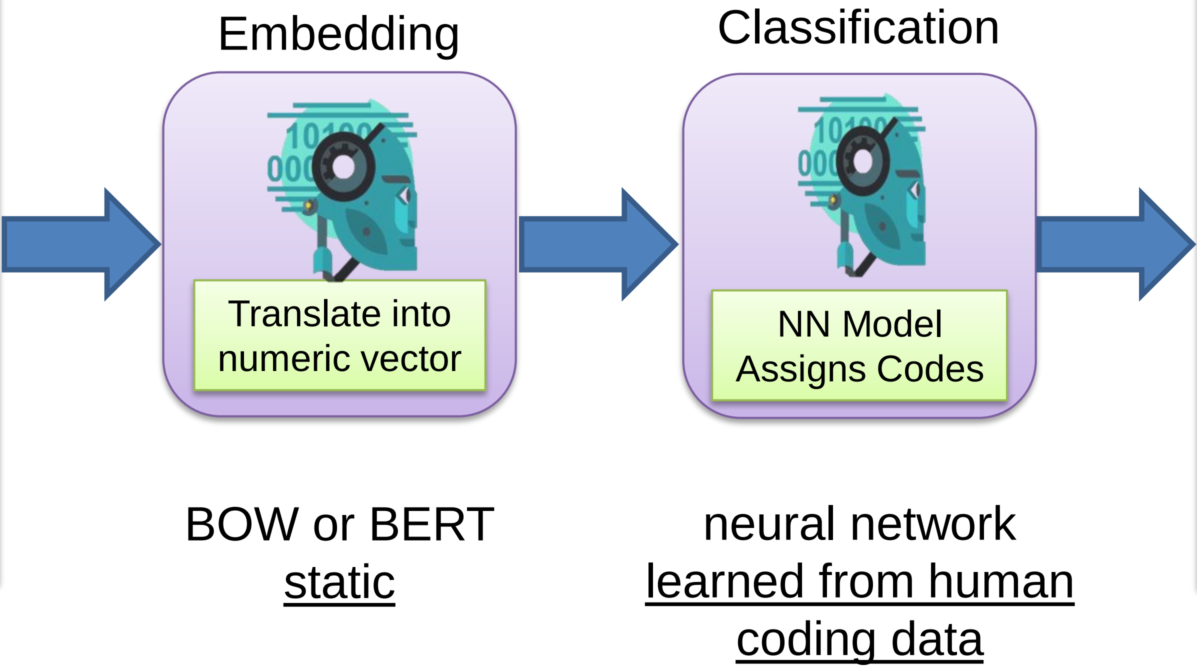
Codes Assigned to Narrative

Emotional Rapport:
Any mention? (Y)
Positive? (Y)
Negative? (N)
Perceived Technical Competence:
Any mention? (Y)
Positive? (Y)
Negative? (N)
⋮

Approach: machine learning (ML) to analyze text data

A patient narrative

Dr. X is an outstanding physician. I feel very fortunate to be one of her patients. She is very thoughtful about my complex medical history and always takes very seriously any concerns that I express about my current health. She is extremely warm and caring and is also extremely smart. A very talented doctor! Once, I was experiencing an unusual symptom and Dr. X not only provided a thorough exam and ordered appropriate tests but she wouldn't let me leave the office until she had the test results. […]



Embedding

Translate into numeric vector

BOW or BERT
static

Classification

NN Model Assigns Codes

neural network
learned from human coding data

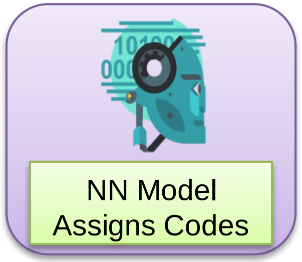
Codes Assigned to Narrative

Emotional Rapport:
Any mention? (Y)
Positive? (Y)
Negative? (N)
Perceived Technical Competence:
Any mention? (Y)
Positive? (Y)
Negative? (N)
Etc….

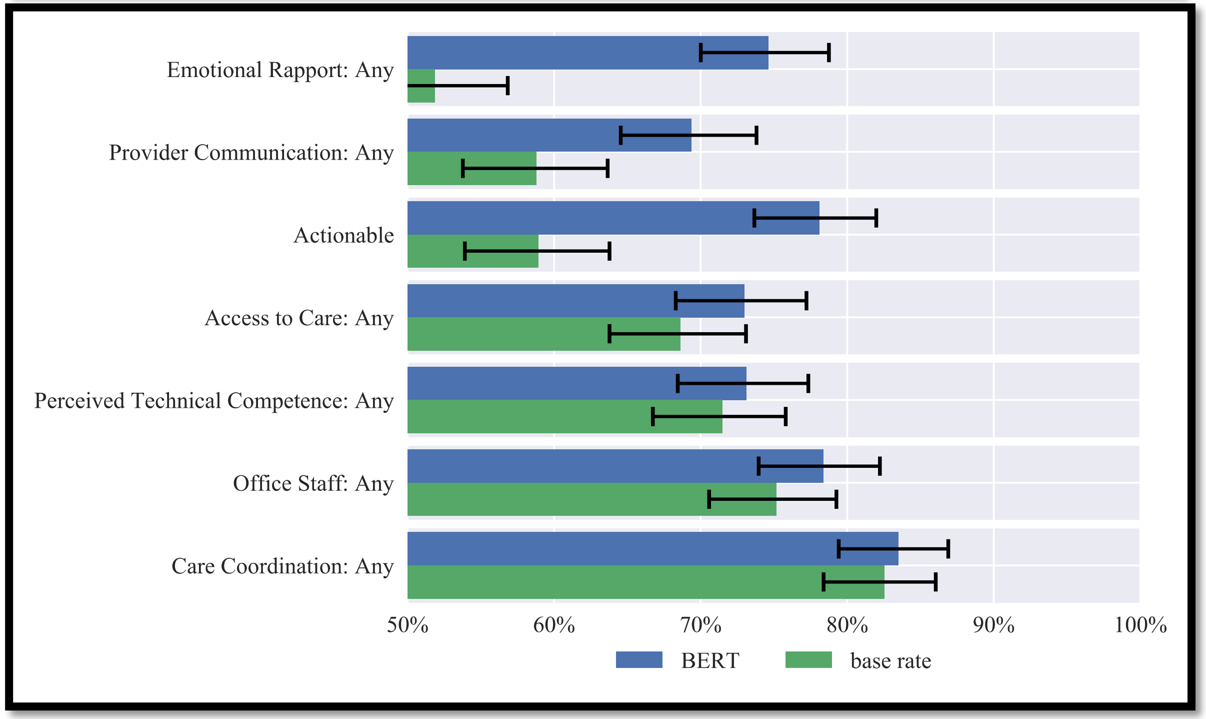
- BOW: simpler “embedding” of narrative text into a numeric vector space for statistical modeling
- BERT: more complex text “embedding” using a higher-powered pre-trained language model
- Neural network (NN): statistical model for predicting codes (y) using vector-representation of text (X)
    - Validation using held-out test set cross-validation

Performance Notes #1: Rarer Codes Harder to Fit

- Measured by accuracy, the performance of BOW and BERT is comparable (N=1,490 narratives)
- Base rate: the percentage of narratives that have the code or don’t have the code, whichever is higher
- If accuracy = base rate, you might as well have just guessed the more common class (judged by accuracy)
- Codes with high base rates (usually where the presence of the code is rare) have model performance indistinguishable from base rate



Accuracy of model using BERT on any mention of code



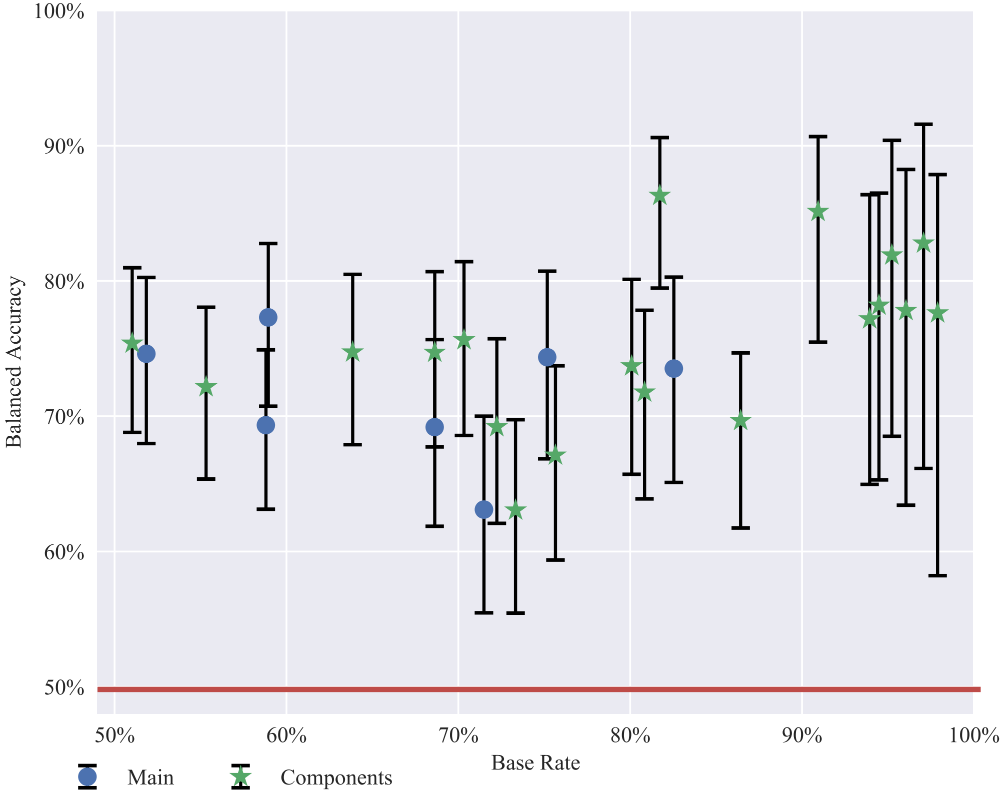
Codes with base rates closer to 100% associated with lower *lift* in model prediction

Notes:
- Oversampling techniques gave no improvement (see report)
- all confidence intervals are Wilson intervals with an overall α=0.05

Performance Notes #2: Adjusting for base rates

- Re-optimize models to optimize *balanced accuracy* instead
  - Penalizes misclassifications proportional to the rarity of the true class
  - “By chance” performance = *50%*, for all codes using balanced accuracy metric
- Result: BERT-based model shows lift in prediction even on rare codes
- So: failure to achieve accuracies above base rate was driven by optimizing for accuracy, not high base rate limiting model power

Balanced Accuracy of model using BERT vs base rate



Optimizing on balanced accuracy metric gives good performance even on rare codes

Performance Notes #3: The Value of Pre-Trained Embeddings

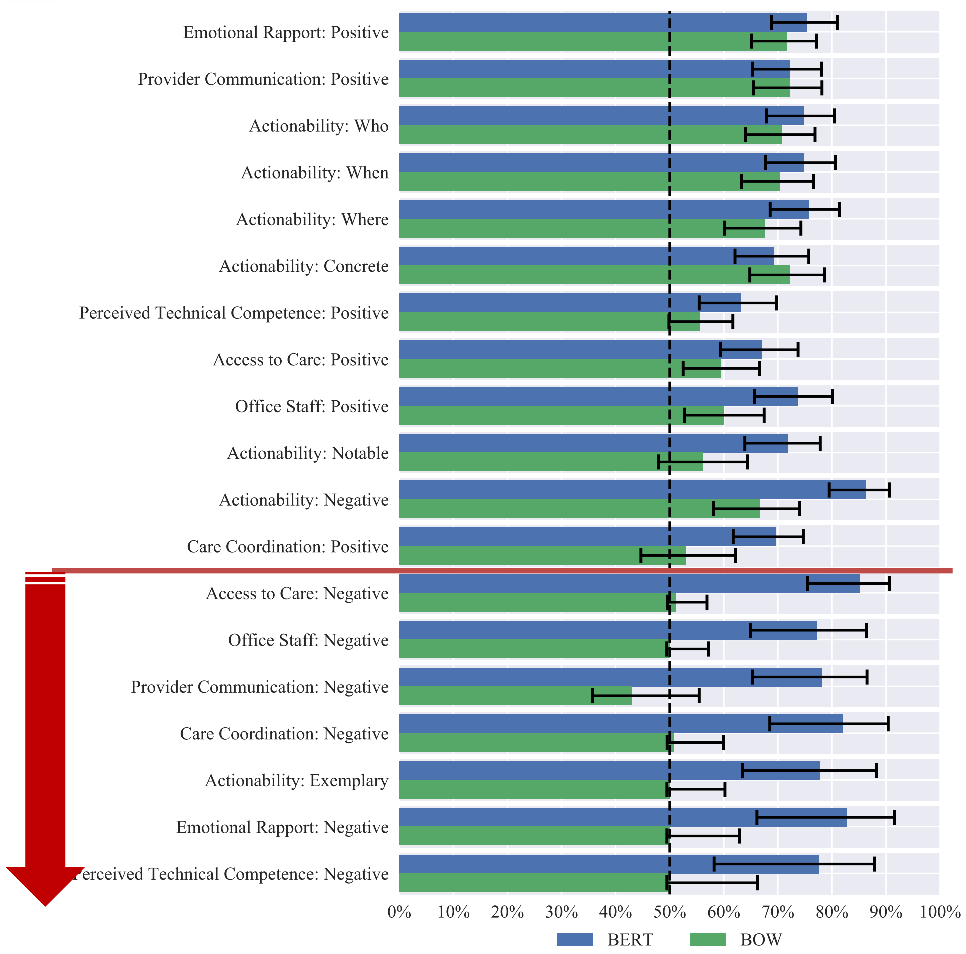
- Intuitively: BERT “knows” a lot about the structure of language. It has “less to learn” from the limited number of examples of rare codes

- BOW is more naïve about context in language use

- BOW-based model underperforms BERT-based model depending on code base-rate. Specifically the BOW-based model:
  - Is predictively useless for base-rates above ~87%
  - Sometimes fails to perform predictively for base-rates between [~74%, 87%]

Translate text to numeric vector

Balanced Accuracy Model Performance: BERT vs BOW on component codes (*Codes in ascending order of base rate*)



BOW not predictively useful below this

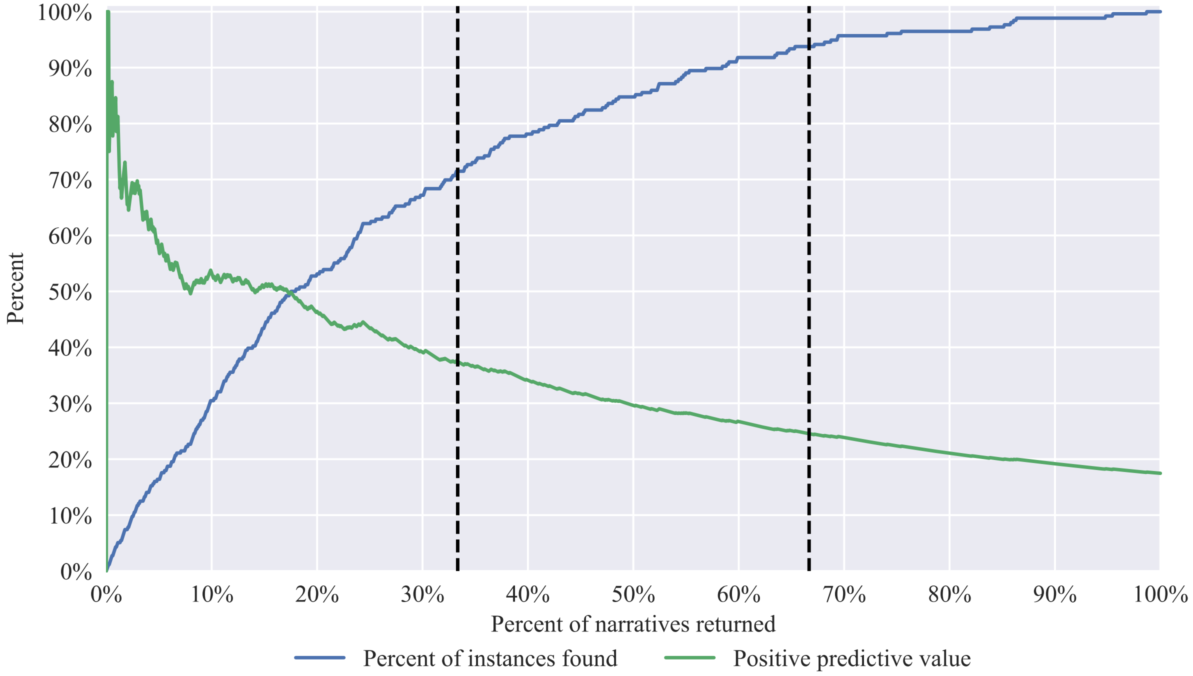
Intervening on the basis of Imperfect Prediction…

- Might be uncomfortable using statistically-significant models with *imperfect* predictive accuracy (e.g. ~80%) to inform real interventions of actions

- Claim: a valid model does not need near-perfect accuracy to be useful
  - This is a question of ***designing*** effective Human-Factors/Human-Computer Interaction practices

- Consider sample use-case: a decent model could be tuned to ***prioritize*** codes for human review
  - If you can only review 33% of the narratives, you can still expect to get 70% of the ones pertaining to care coordination

True Positive Rate vs Positive Rate
For "Care Coordination: Any"



A model attempting to identify the narratives most likely to pertain to "care coordination"

Further Considerations

- This effort was conducted entirely on a standard RAND laptop (No AWS time!)
  - Minimal computing and analyst time required to replicate

- Even if performance is deemed acceptable for standalone use, human coders are needed to build data for ongoing “audits” of model performance

- To increase performance, try:
  - Increasing computational power (fine-tune BERT or descendant)
  - Increase dataset size (more human coder time)
  - Additional pre-training on healthcare specific data (very computationally intensive)

- Big outstanding question: algorithmic equity implications such as
  - Is the narrative data used to train the model demographically representative?
  - Does the model capture variation in language-use across demographics well enough?
  - Can we track variation in patient-reported care quality and break this down along key demographic variables?