Fair Machine Learning in Healthcare

Presenter: Amanda Ng

Supervisor: Professor Jesse Gronsbell

Introduction

🏥 Increase use of ML models for Healthcare

- Disease diagnoses & risk prediction

⚖️ Can perpetuate existing inequalities

📄 Electronic health records (EHR)

- Extract text
- Generate word embeddings
- Identify symptoms
- Diagnosis prediction

Data

MIMIC-III (Medical Information Mart for Intensive Care III) **Clinical Database**

- > 40000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012
- **Billing codes:**
  - Why somebody was in the hospital (diagnosis)
  - ICD => Phecode
- **Clinical notes:**
  - What happened during the hospital visit
  - Free text
  - Use NLP software to extract CUIs
- **Demographics:**
  - Race, insurance type, discharge date

Cleaned Data Sample

Patient ID	Month	Code (CUIs or Phecodes)	Count
1	6	C0011900	2
1	6	187.4	1
2	4	C0233492	1
2	8	C3890715	3
3	11	295.4	2

Number of times a particular medical code was registered for a patient in a given month.

Analysis Overview (Part I)

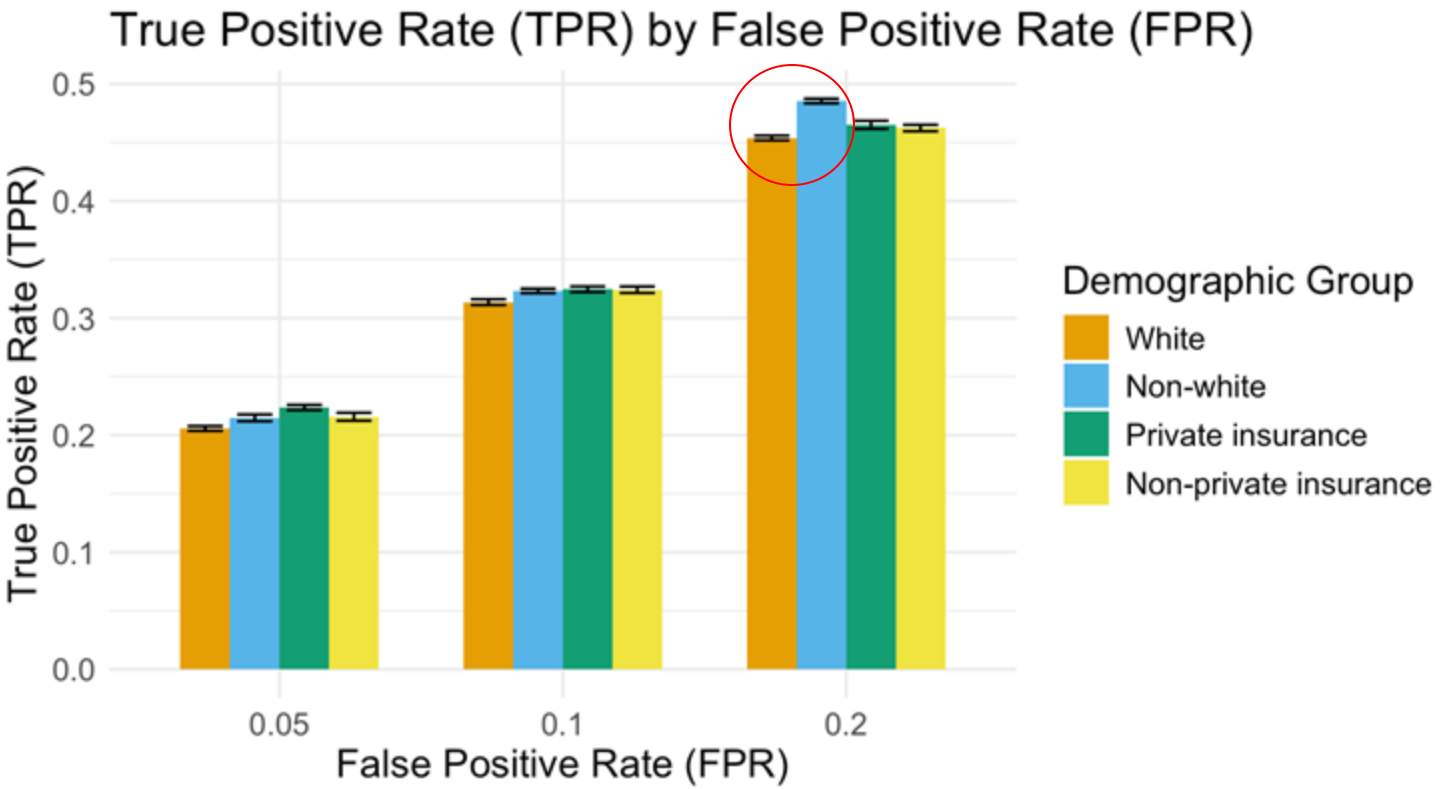
Embedding generation

- How words relate to each other: Co-occurrence
- Represent text a low dimension form: Vectorization
- SVD-PMI algorithm (related to skip-gram)

Embedding evaluation

1. **Known pairs**
   - CUI (2358 pairs)
   - Phecodes: same three-digit parent code (4152 pairs)
2. **Compute Overall AUC**
   - Optimal dimension (1500)
3. **Compute TPR by demographic groups**
   - White, Non-white, Private insurance, Non-private insurance

ROC analysis by Demographic Groups



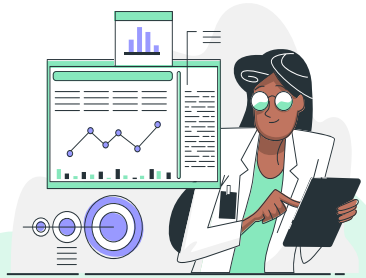
Analysis Overview (Part II)

🎯 Depression Feature Selection: KESER

1. Model the conditional distribution of depression (C4049644) and other codes
2. Impose a Gaussian distribution on embedding vectors
3. Infer dependencies via a Gaussian graphical model
4. Apply elastic net penalized regression

📈 ML Prediction models (70% training; 30% testing)

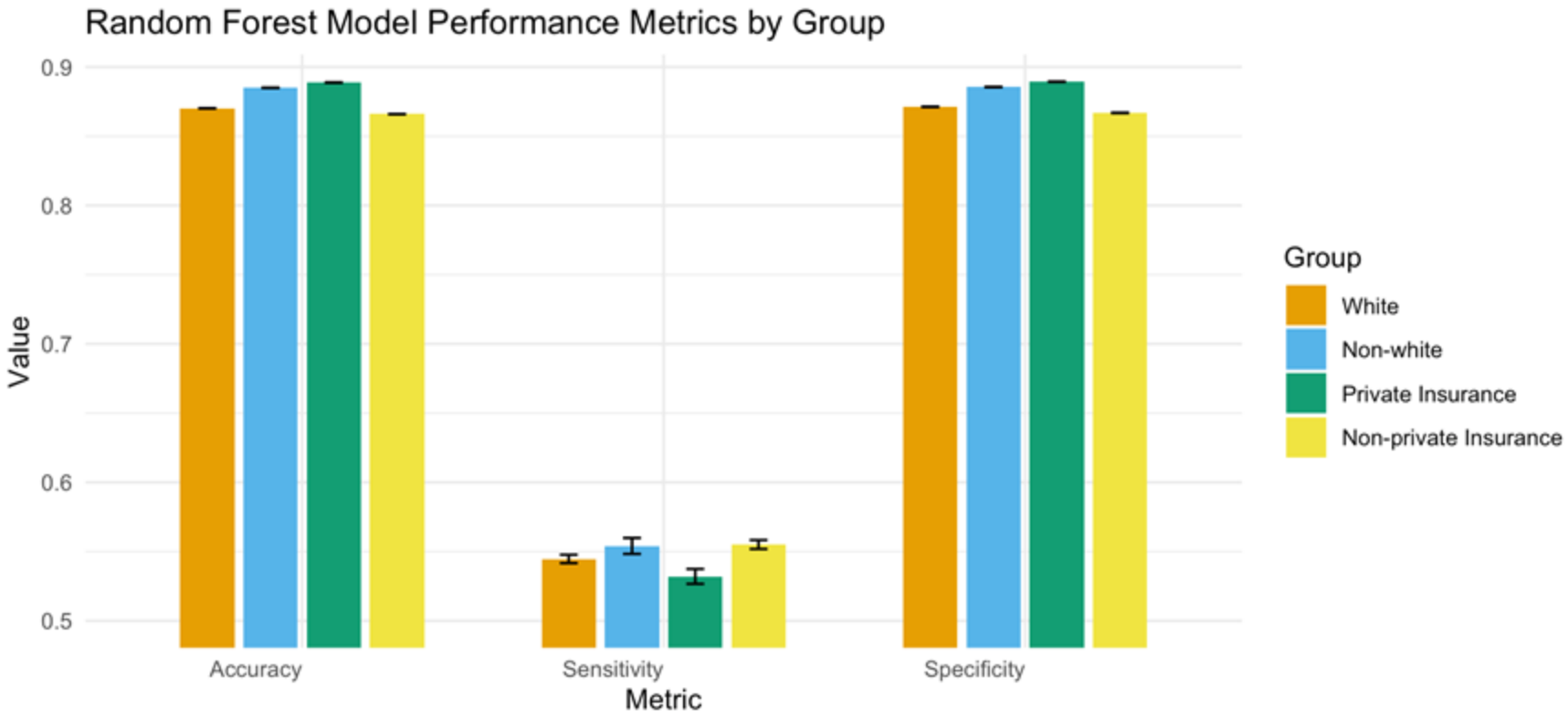
- Logistic Regression, Random Forest
- 20 KESER features + demographic variables as predictors
- Evaluate model metric values on demographic subgroups



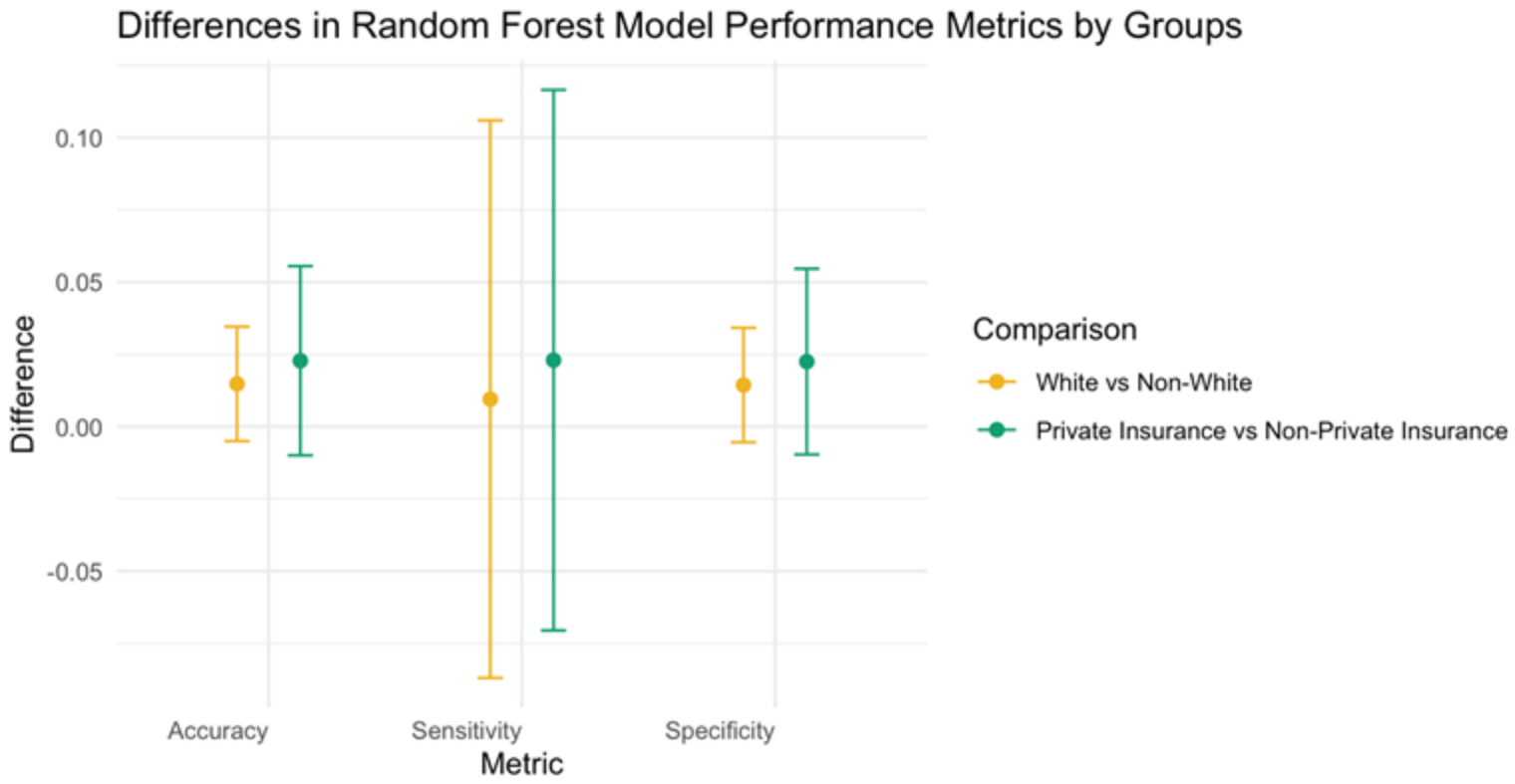
Feature Selection for Depression

- Conducted on embeddings with 1500 dimensions
- In general, selected features make sense
  - Symptoms
  - Related mental health illness
  - Medications
- Some selected features lack clinical relevance or logical connections
  - E.g. EGD (esophagogastroduodenoscopy)

Random Forest Model Results



RF Model Demographic Group Differences



Conclusion

📌 Key Takeaways:

- ROC analysis on embedding reveal demographic disparities
- ML prediction model performance vary across groups
    - Feature Selection does not take account into any demographic information
    - Potentially due to biases that exist in the text or coding

💭 Future Directions:

- Explore alternative embedding techniques for better fairness.
- Improve feature selection to reduce bias.
- Develop bias mitigation strategies in ML models.

References

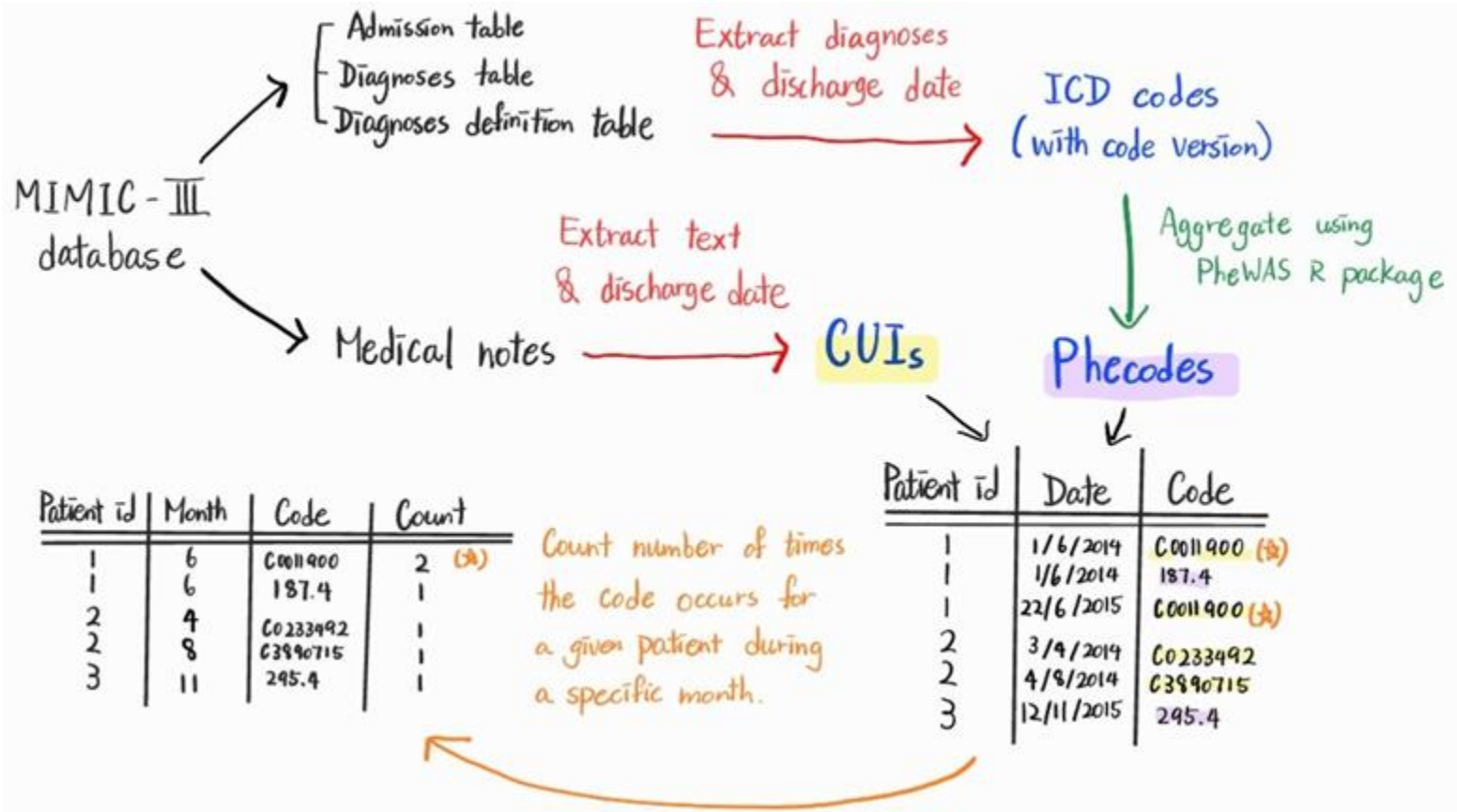
Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E. T., Wu, J. T., Welt, J., Foote, J., Jr, Moseley, E. T., Grant, D. W., Tyler, P. D., & Celi, L. A. (2018). Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS one*, *13*(2), e0192360. https://doi.org/10.1371/journal.pone.0192360

Hong, C., Rush, E., Liu, M., Zhou, D., Sun, J., Sonabend, A., Castro, V. M., Schubert, P., Panickan, V. A., Cai, T., Costa, L., He, Z., Link, N., Hauser, R., Gaziano, J. M., Murphy, S. N., Ostrouchov, G., Ho, Y.-L., Begoli, E., ... Cai, T. (2021). Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *NPJ Digital Medicine*, *4*(1), 151–11. https://doi.org/10.1038/s41746-021-00519-z

Fisher, J., Palfrey, D., Christodoulopoulos, C., & Mittal, A. (2019). *Measuring Social Bias in Knowledge Graph Embeddings*. https://doi.org/10.48550/arxiv.1912.02761

Appendix

Data Pre-processing



Embedding Generation Method

Co-occurrence matrix:

- Each cell contains the count of which the two codes appear together within the same diagnosis month
- First computed for each patient for each month, and sum aggregated across all patients and all months

Pointwise Mutual Information (PMI) matrix:

$$PMI(x,y) = \log \frac{P(x,y)}{P(x)P(y)} = \log \frac{C(x,y) \cdot N}{C(x,\cdot) \cdot C(\cdot,y)}.$$

Singular Value Decomposition (SVD):

- d = Embeddings dimension

$$PMI_d = U_d diag\{\lambda_1, \lambda_2, \ldots\} U_d^T$$

$$V_d = U_d diag\{\sqrt{\lambda_1}, \sqrt{\lambda_2} \ldots\}$$

Bootstrapping Method to Calculate SE

For each demographic group:

- Number of “bootstrap” iterations B = 9
- FPR values of interest = (0.05, 0.1, 0.2)
- Goal: estimate the true positive rate (TPR) at specific false positive rate (FPR) values

For each “bootstrap” iteration:

- Generate a bootstrapped embedding set (dim = 1500) by resampling observations from original dataset
- Calculate TPR values corresponding to the predefined FPRs

After all iterations:

- Original TPR
- Standard error (SE) of 9 bootstrapped TPR estimates

ML Models for Depression Prediction

Groups Analyzed: White / Non-white individuals, Private / Non-private insurance holders

Model Setup: Logistic Regression (LR), Random Forest (RF)

- Dependent Variable: Depression indicator
  - 1 if if any of C4049644, C0011581, 296.2, 296.22 is present
- 20 Independent variables: 20 KESER feature indicators
- Secondary model includes additionally 2 demographic indicators
- Data Split:
  - 70% Training
  - 30% Testing: subsetted into demographic groups

Evaluation Criteria

- LR probability threshold = 0.5
- Compare metric values between complementary demographic groups
- SE calculated by bootstrapping testing dataset