Exploring Gender Disparities in Time to Diagnosis

Tony Y. Sun[1], Oliver J. Bear Don't Walk IV[1], Jennifer L. Chen[2], Harry Reyes Nieva[1], Noémie Elhadad[1,2]
[1] Department of Biomedical Informatics, Columbia University
[2] Department of Computer Science, Columbia University

Introduction

- Healthcare disparities based on sex and gender [1] contribute to differences in health outcomes. We explore gender differences in time to diagnosis (TTD).
- Two large-scale, complementary analyses
  1) TTD Disparities: we find women are more likely to experience longer TTD than men, even when presenting with similar symptoms
  2) Diagnostic fairness across time: diagnostic process favors men, contradicting observation from previous analysis that women may demonstrate relevant symptoms earlier than men

Data

- Columbia University Irving Medical Center electronic health record (EHR) data
  - Data standardized using the Observational Medical Outcomes Partnership (OMOP) Common Data Model [2]
  - Patients age 13 or older with at least 3 years of continuous observation between 2010 and 2020.
    - 533,566 women (mean age: 49.5 years)
    - 412,498 men (mean age: 49.1 years)
  - 121 EHR phenotype cohort available on OHDSI platform [3]
  - 14,900 unique coded condition occurrences (up to 3 years prior to diagnosis)
    - Recorded date of first occurrence in a patient's longitudinal record converted to binary feature

Time to Diagnosis among Women and Men

- We extended the definition of time-to-diagnosis (TTD) [4] to include all presenting conditions prior to an official diagnosis.
- We computed TTD as the mean time interval between a condition's first occurrence in a patient record and the phenotype diagnosis time (Fig. 1)

Figure 1: TTD computation for each phenotype (e.g., Crohn's disease)

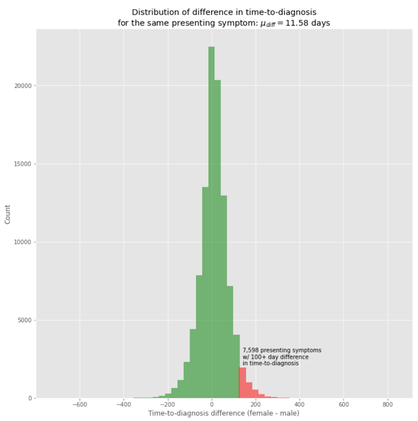

Figure 2: Distribution of TTD differences for the same presenting symptom between women and men at CUIMC

- In 59% of all conditions for patients that go on to develop the same phenotype, women were diagnosed later than men
- Women were diagnosed 11.58 days later than men on average
- 7,598 conditions (7.5%) had a 100+ day TTD difference between women and men
- Using Crohn's disease as an example, given the same symptom presentation, women were diagnosed with an overall TTD delay of 19.54 days.

Time-Variant Model Fairness in Diagnosis Classification

- Does the diagnostic process favor a gender consistently over time?
- For each phenotype, we trained a gender-agnostic disease classifier using clinical observations recorded up to 3 years (1095 days) prior to an official disease diagnosis
- Evaluated classifier performance for each cohort with decreasing levels of right censoring (Fig. 3) up to diagnosis date, mimicking a provider diagnosing patients at increasing time steps with knowledge limited to patient history accumulated thus far
- Since women present longer average TTD than men, we hypothesize women should be diagnosed earlier in the same time window, as women exhibit key symptoms earlier

Gender bias analysis of all 29 phenotypes

- We propose a simple metric, Mean Squared Discrimination (MSD) (Eq. 1), as a proxy for model bias across all windows
- Can be interpreted per phenotype using custom fairness metrics [4]
- As minimizing false negatives is preferred in diagnosis, we use recall gap (Eq. 2)
- Fig. 4 shows marked differences across 29 phenotypes for model gender bias
- Phenotypes are ranked by estimate magnitude per gender (Fig. 4)
- In Crohn's, a small male-biased MSD contradicts our TTD finding. While women present relevant symptoms earlier, they may present many other conditions that confuse the classifier, thus rendering a male-biased recall.

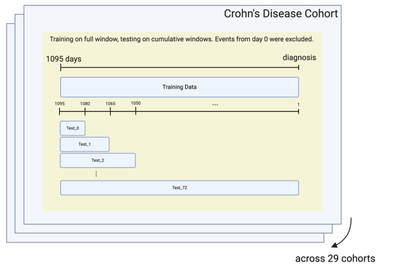

Figure 3: Test sets are generated with varying levels of right censoring, mimicking varying amounts of patient history available to providers

$$\mathrm{MSD}(M_d) = \mathrm{sign}(\tfrac{1}{b}\textstyle\sum_{i=1}^{b} g_i) * (\tfrac{1}{b}\textstyle\sum_{i=1}^{b} g_i^2)$$
$$g_i\text{: fairness gap at window } i.$$

Equation 1: Mean Squared Discrimination is the mean squared error of fairness gaps across all b time windows, multiplied by the sign of the mean fairness gap (e.g., for this experiment, we use recall). Positive MSD indicates model bias towards class 1; negative, bias towards class 2.

$$\text{Recall Gap} \quad \frac{\mathrm{TP}_1}{\mathrm{TP}_1 + \mathrm{FN}_1} - \frac{\mathrm{TP}_2}{\mathrm{TP}_2 + \mathrm{FN}_2}$$

Equation 2: fairness gap for recall between class 1 and class 2; in our case, class 1 is men, class 2 is women

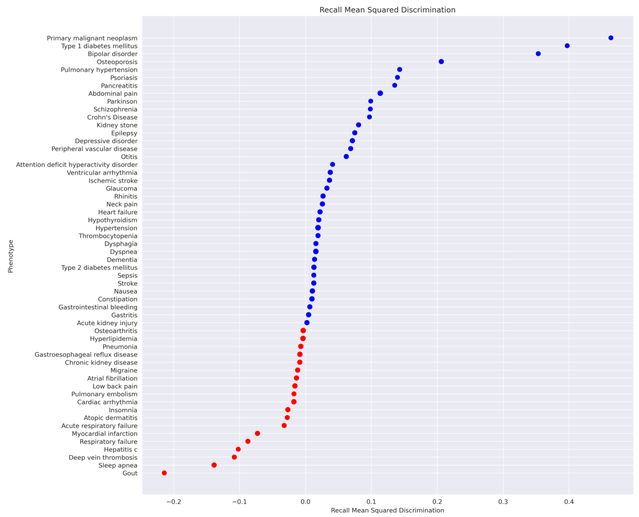

Figure 4: Recall MSD with 95% confidence intervals across 29 phenotypes.
Blue circles indicate better recall for men; pink, for women. Circle size is proportional to the size of the dataset.

Combining Frameworks to Analyze Phenotypes

- Here we present a case study using Crohn's Disease as an example
- Model performance increases as more patient history is used (Fig. 5)
- Recall for men is consistently higher than for women, contradicting our previous finding that women present symptoms earlier for Crohn's (Fig. 6)
- We observe that a male-favored recall gap decreases to zero over windows steps (Fig. 7)

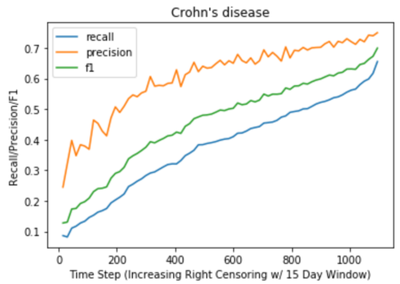

Figure 5: Performance of the Crohn's diagnosis classifier when tested with varying levels of right censoring

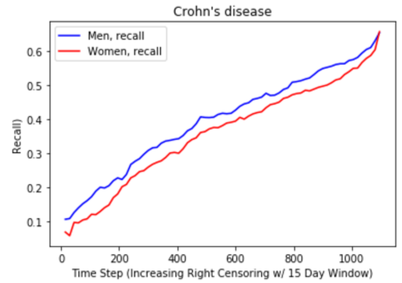

Figure 6: Gender-specific recall at varying levels of right censoring for Crohn's Disease

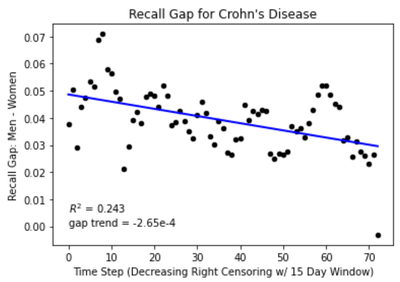

Figure 7: Gender-specific recall gaps at varying levels of right censoring for Crohn's Disease

Conclusion and Future Work

- Our proposed metric, MSD, provides a novel approach to measure complex gender disparities over time
- We propose novel experiments with time-varying windows to investigate gender disparities based on TTD
- Across 121 OHDSI phenotypes, 65 were excluded because of low prevalence.
- We find that women were far more likely to experience a longer TTD than men
- When training gender-agnostic disease classifiers, the majority perform better for men than women and this trend persists across most time windows
- While our analysis is limited to a single clinical site, because we operate in the OHDSI framework, our methodology is extensible to other OHDSI sites and may be applied to other validated disease phenotype definitions

References

[1] *We recognize gender is a social construct while sex is a biological variable. We use the terms "gender," "men," and "women" based on available labels.* World Health Organization. WHO gender policy: Integrating gender perspectives in the work of WHO. Geneva: World Health Organization, page 6, 2002.
[2] Stang, P. E. et al. 2010. "Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership." Ann. Intern. Med. 153 (9): 600–606.
[3] OHDSI. Observational health data sciences and informatics studies. https://github.com/ohdsi-studies/Covid19CharacterizationCharybdis/tree/master/inst/cohort
[4] Haoran Zhang, et al. Hurtful words: quantifying biases in clinical contextual word embeddings. In Proceedings of the ACM Conference on Health, Inference, and Learning, pages 110–120, 2020.
[5] Elise Launay, et al. Reporting studies on time to diagnosis: proposal of a guideline by an international panel (REST). BMC medicine, 14(1):146, 2016.