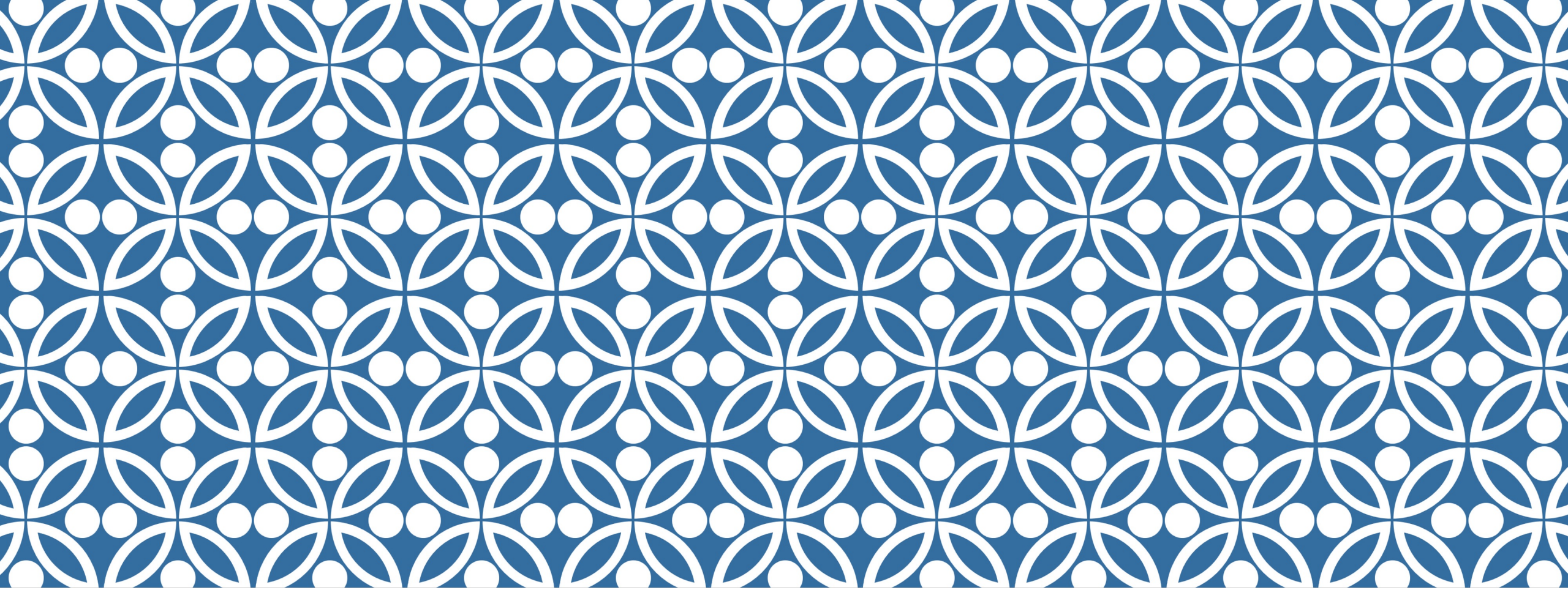


TOWARDS ROBUST NATURAL LANGUAGE PROCESSING TO PROMOTE HEALTH EQUITY

Keith Harrigian

PhD Thesis Defense

July 3, 2024

AGENDA

Introduction

- The Case for AI for Promoting Health Equity

Defensive Tactics for Promoting Health Equity

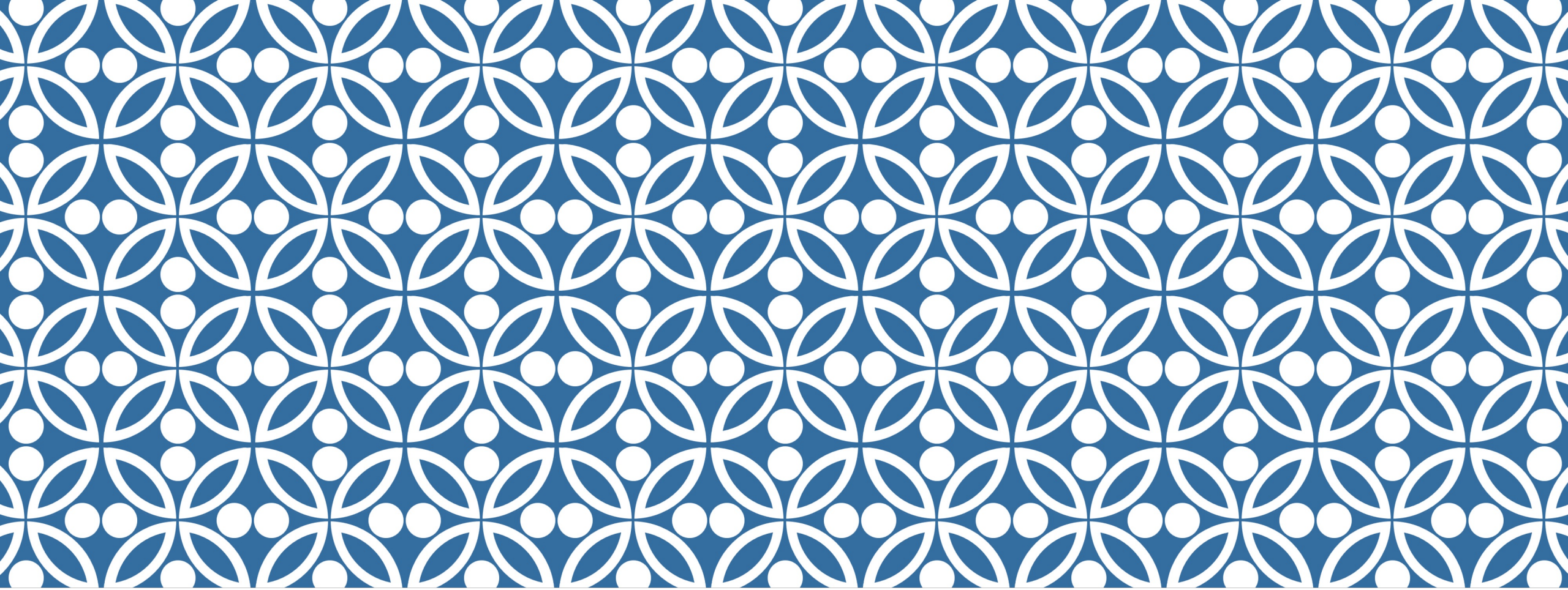
- Measuring & Understanding Selection Bias
- Counteracting the Effects of Selection Bias

Proactive Tactics for Promoting Health Equity

- Characterizing and Measuring Implicit Bias in Medical Records

Conclusion

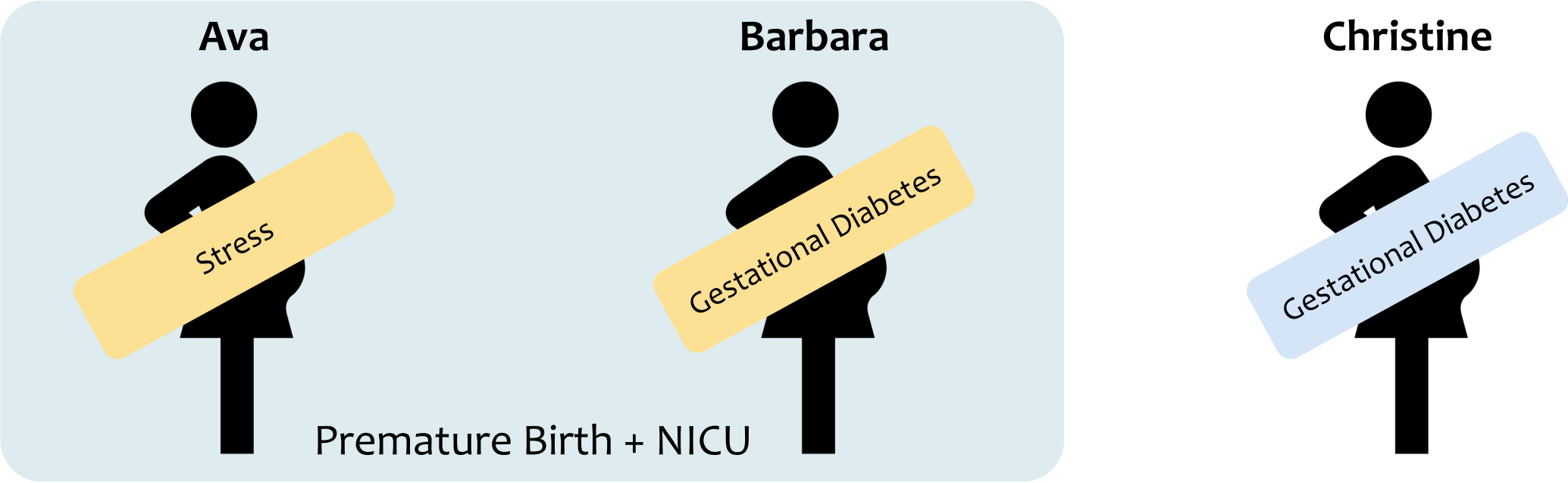
- Future Directions



THE CASE FOR AI IN PROMOTING HEALTH EQUITY

Introduction

LET'S START WITH A SCENARIO



29 Years Old

Queer, African American
Bank Teller 9 to 5pm
Upcoming Layoffs

32 Years Old

Heterosexual, White
AA Degree, Low Income
Self-Pay Health Insurance

31 Years Old

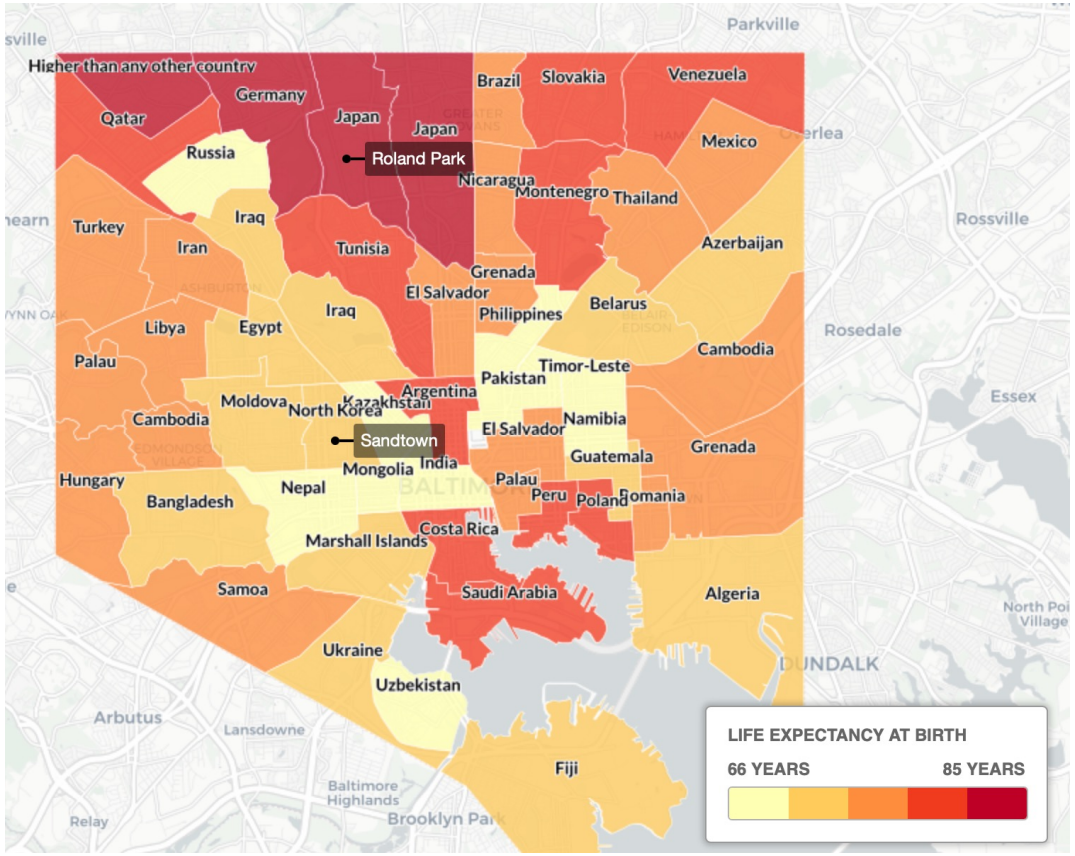
Heterosexual, Native American
Parental Benefits
Employee Health Insurance

HEALTH DISPARITIES & HEALTH EQUITY

"Health equity means that everyone has a **fair and just opportunity** to be as healthy as possible."

"For the purposes of measurement, health equity means reducing and ultimately **eliminating disparities** in health and in the determinants of health that adversely affect **excluded or marginalized groups**."

- "What is Health Equity?" (Braveman et al., 2018)



Disparities caused by **structural and systemic discrimination**

Traditional efforts to reduce disparities are **expensive** and **difficult to scale**

NATURAL LANGUAGE PROCESSING (NLP) IN HEALTH CARE

A Proven Track Record

- Expediting biomedical literature search during COVID-19
- Facilitating cohort selection for clinical trials
- Identifying adverse drug reactions from social media
- Structured data extraction from EHR text

Albeit With A Generic Focus

- Practitioners seek to improve health outcomes generally
- Lesser recognized opportunity to improve outcomes in specific groups

THE ROLE OF NLP IN PROMOTING HEALTH EQUITY

Access to and Quality of Care

Providers working in low-resource settings have:

- Less time and assistance (nursing, residents, administrative, etc.)
- More concentrated types of clinical experience

AI provides an opportunity to augment and expedite care for populations lacking resources

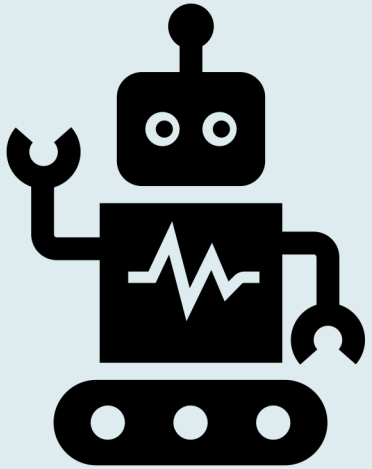
Implicit Bias and Discrimination

Patients from marginalized populations are:

- Treated different by providers (clinically and personally)
- Less trusting of the healthcare system; delay care and treatment

AI provides an opportunity to identify discrimination and modify clinician behavior

Access and Quality of Care



Chatbot for at-home maternal health symptom triage and education

Literature summarization to support maternal care by general practitioners

Implicit Bias

Model that measures doubt in clinical interactions to highlight implicit biases

Ava

Barbara

CONCEPTUAL FRAMEWORK

Defensive Tactics: Mitigating Selection Bias in NLP Models

Systematic error in the outcome of a study due to dataset curation or modeling decisions

Why it Matters

- Datasets are often **not representative** of the population they intends to study
- Models may not **fairly and robustly** characterize their target population

Proactive Tactics: Addressing Social Bias using NLP Models

Human-leveled prejudices and predispositions regarding groups, attributes, or circumstances

Why it Matters

- Healthcare providers may implicitly (unconsciously) **discriminate** against patients
- **Identifying and measuring** these behaviors allows us to correct them

Defensive Tactics

Measuring and Understanding Selection Bias

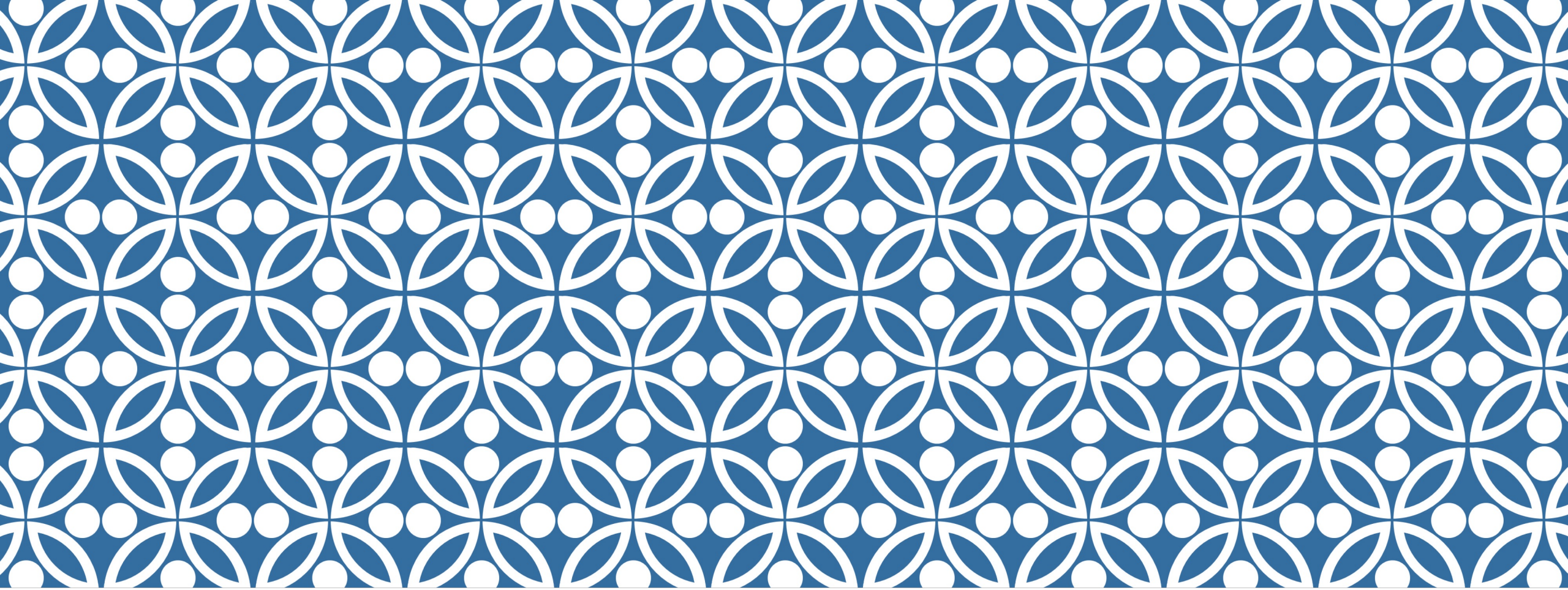
↓

Counteracting the Effects of Selection Bias

Proactive Tactics

Characterizing and Measuring Health Disparities

- On the State of Social Media Data for Mental Health Research (Harrigian, Aguirre, & Dredze, 2021)
- Do Models of Mental Health Based on Social Media Generalize? (Harrigian, Aguirre, & Dredze, 2020)
- Then & Now: Quantifying the Longitudinal Validity of Self-disclosed Depression Diagnoses (Harrigian & Dredze, 2022)
- The Problem of Semantic Shift in Longitudinal Monitoring of Social Media (Harrigian & Dredze, 2022)
- An Eye on Clinical BERT: Investigating Language Model Generalization for Diabetic Eye Disease Phenotyping (Harrigian et al., 2023)
- Characterization of Stigmatizing Language in Medical Records (Harrigian et al., 2023)
- Health disparities in lapses in diabetic retinopathy care (Cai et al., 2023)
- Are Clinical T5 Models Better for Clinical Text? (Li et al., 2024)
- Improving the identification of diabetic retinopathy and related conditions using natural language processing methods (Harrigian et al., 2024)



MEASURING & UNDERSTANDING SELECTION BIAS

Defensive Tactics for Promoting Health Equity

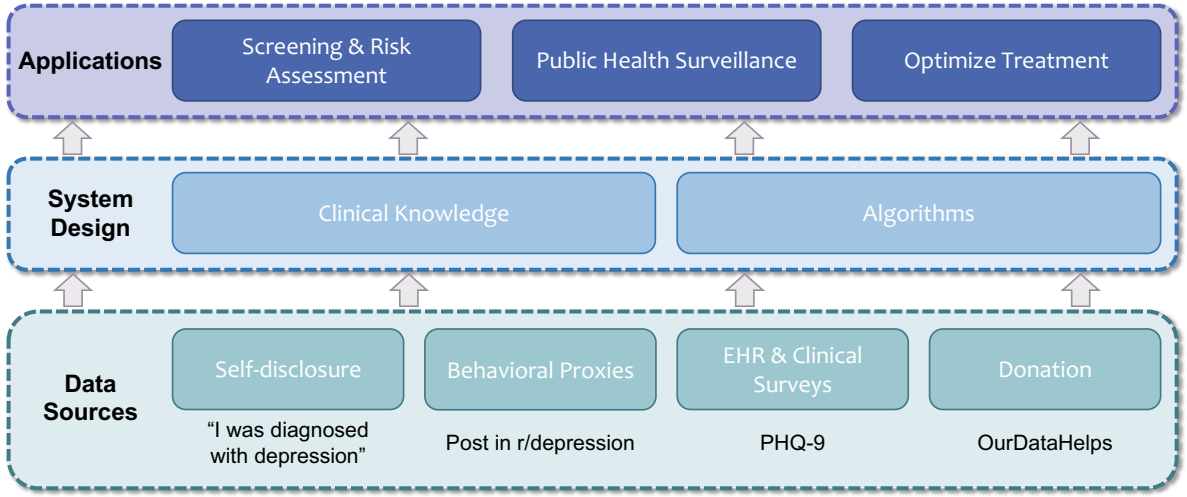
BACKGROUND

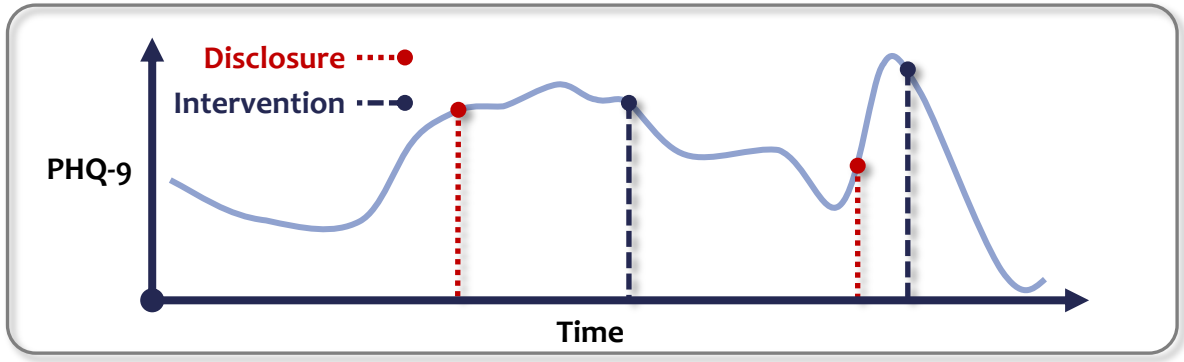
Data Drives Health NLP Applications

- Clinical ground truth preferred, but not always feasible to acquire at scale
- Proxy-based and rule-based methods sacrifice precision for increased recall

Concerns and Criticism

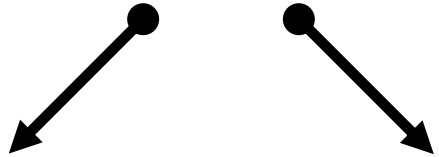
- Construct validity
- Selection bias and representation issues
- Static labels for inherently dynamical latent attributes





SPECIFIC AIMS

To what extent do self-disclosures of a diagnosis remain valid over time as proxies for health status?



How does predictive performance change when training a classifier on new data associated with an old label?

Are changes in predictive performance (or lack thereof) due to condition dynamics or due to sample-related confounds?

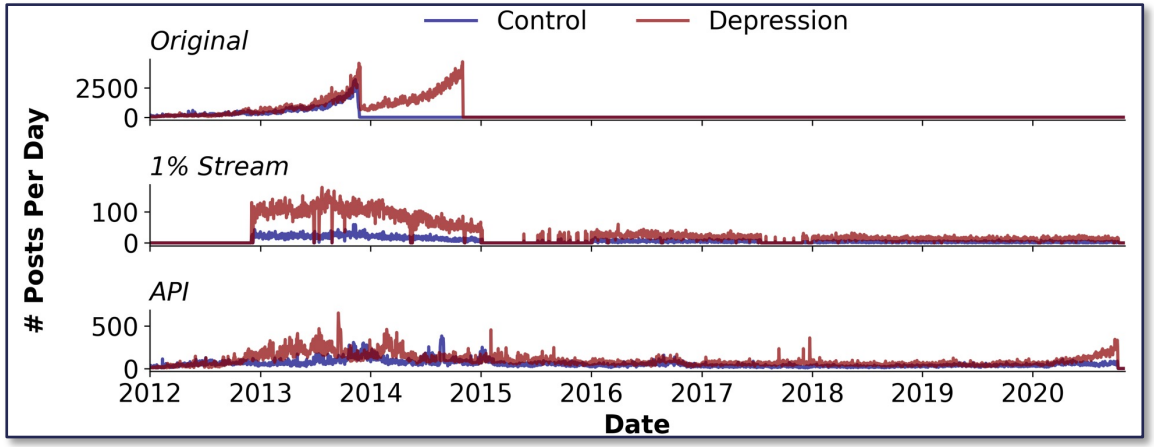
DATA

2015 CLPsych Shared Task

- Regular-expressions and expert verification of diagnoses disclosures

Updating the Dataset

- Account identifiers available w/ explicit permission from Coppersmith et al. (2015)
- Query all available data from Twitter API and institution cache of 1% stream



Dataset	Dates	# Individuals	# Posts
Original	2012 – 2015	D: 477 C: 872	D: 1,121,388 C: 1,907,508
Updated	2012 – 2021	D: 444 C: 172	D: 1,372,868 C: 546,826

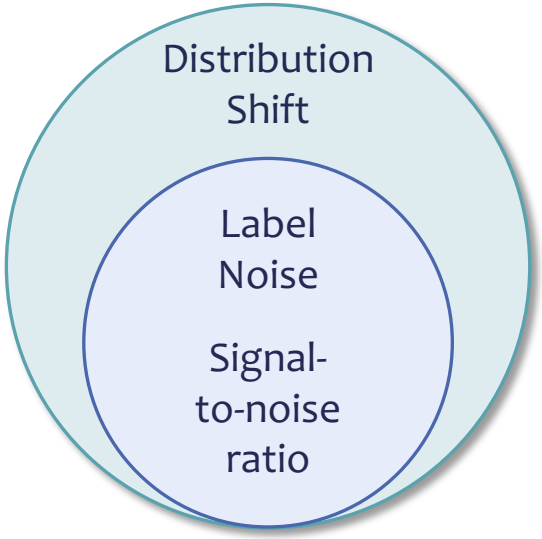
QUANTIFYING LABEL VALIDITY

How does predictive performance change when training a classifier on new data associated with an old label?

Train \ Test	2012 – 2015	2015 – 2018	2018 – 2021
2012 – 2015	**0.71** (0.70, 0.72)	0.66 (0.65, 0.66)	0.69 (0.68, 0.70)
2015 – 2018	0.66 (0.65, 0.67)	0.66 (0.65, 0.66)	0.68 (0.67, 0.69)
2018 – 2021	0.65 (0.65, 0.66)	0.67 (0.66, 0.68)	0.68 (0.67, 0.69)

Within Domain | Between Domain

Causes of Performance Disparities



We know there's a change. What happened?

INTERPRETING LABEL VALIDITY

New measure based on train set debugging*

Average influence of a tweet *x* on user-level inference

$$I(x) = \sum_{k=1}^{K} \underbrace{P_{k,\tau}(y = 1|X_\tau)}_{\text{Probability of outcome given the document history } X} - \underbrace{P_{k,\tau}(y = 1|X_\tau^{\neg x})}_{\text{Probability of outcome given } X \text{ without tweet } x}$$

An annotator is provided up to 30 tweets from each time period with highest $I(x)$

1. Indicate whether there is evidence of depression based on DSM-5 criteria and your prior knowledge regarding presentation of depression in social media
2. (If applicable) Indicate whether the depression appears to be in remission
3. Provide rationale for your decision (e.g., which DSM-5 criteria, topical themes)

* Note: A similar measure was introduced in Ge et al. (2021): "Counterfactual evaluation for explainable AI"

Exemplary tweets for each rationale category identified during train-set debugging procedure

Evidence (Rationale)	Exemplary Tweet
Diagnosis Disclosure	Bipolar disorder and depression. My doctor finally agrees.
Depressed & Irritable Mood	No one ever asks if I'm doing fine.
Loss of Interest/Pleasure/Motivation	…realizing you don't care about the things you used to enjoy
Weight, Body Image, & Nutrition	Not that anyone cares, but I'm almost at my goal weight.
Sleep Disturbance	I CANT SLEEP. PAIN. JUST LIKE ALWAYS.
Fatigue	mentally drained from this pandemic.
Sense of Worthlessness & Guilt	When you let someone do anything to you…
Impaired Thought	I'm failing my classes because I'm depressed.
Death & Self Harm	My scars are faded… unless you care to look close.
Cognitive Distortions	I always think my bf is going to leave me
Treatment	Scared to tell a women that I'm in therapy.
Gatekeeping	depression isn't just a bad day. fuck you all.
Sexuality & Intimacy	Who wants to come take some pics of me for only fans?
Negative Emotions	I feel like no one cares even though I know they do
Coping Strategies	Art is always the easiest way to distract me from my anxiety
Psychiatric Comorbidity & State	I am anorexic and I cut myself
Non-psychiatric Comorbidity	Could use a little bit of aid #DisabilityAid
Substance Use	Weed makes the dreams go away and that's a good thing
Support & Advocacy	RIP Chester. If you're going through pain, please reach out to me.
Personality & Identity	Lol grandma still think I'm bringing a boy home
Music Culture & Lyrics	#FallingInReverse :D
Familial/Romantic Relationships	Mom: You'll never lose weight. Me: is that why dad left?
Hobbies	Missin the old days when everyone played Pokemon yellow
Non-personal Accounts	My life was about to fall apart until I found the Calm app…

Diagnosis Disclosure

DSM-5 Criteria

Empirical Themes

QUANTITATIVE RESULTS

Decrease in Evidence of Depression

- 76% of individuals in original depression group during 2012-2015
- 45% and 39% of individuals during latter two time periods

Label Noise

- 4% of Control group shows strong evidence of depression (Wolohan et al., 2018)
- Non-zero proportion of individuals discussing *prior experience* with depression (remission)

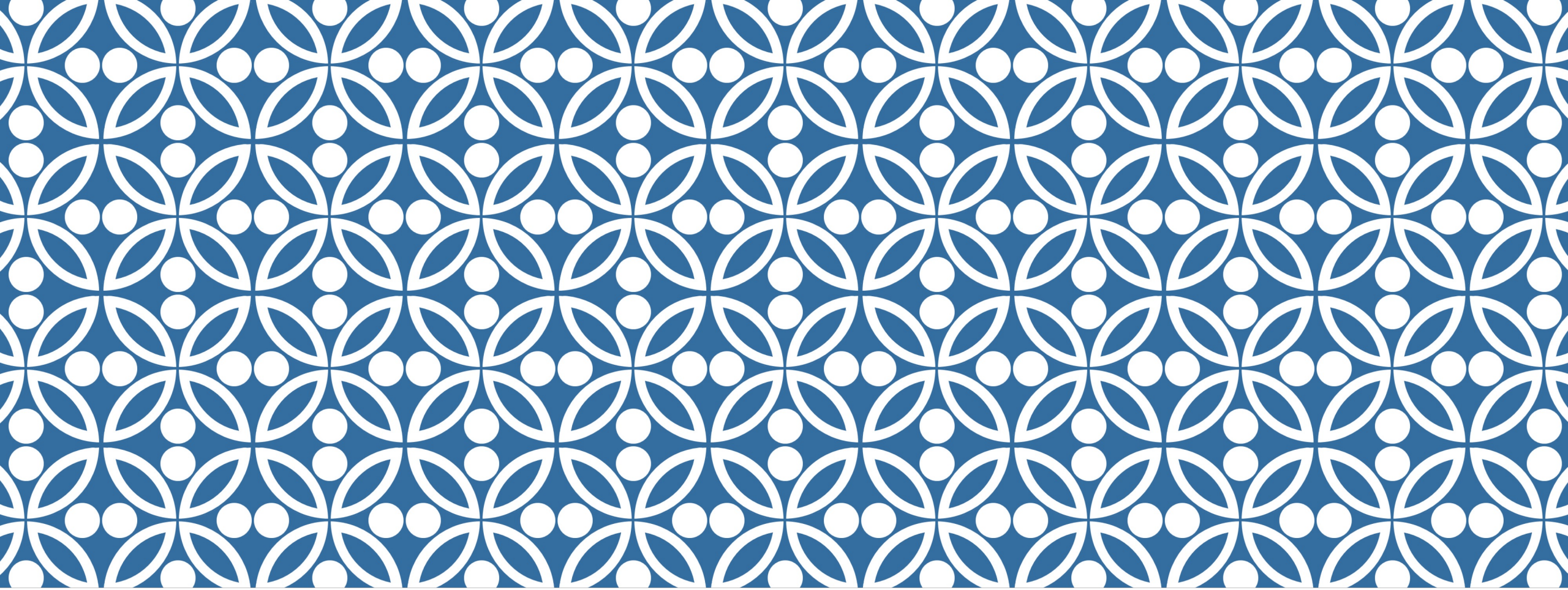
	Dates	Total	Some Evidence	Strong Evidence	Remission
Control	2012 – 2015	83	15	3	1
	2015 – 2018	50	10	2	0
	2018 – 2021	40	5	0	0
Depression	2012 – 2015	215	164	136	10
	2015 – 2018	107	49	28	2
	2018 – 2021	79	31	16	1

Distribution of Annotations (A_1 only)

IMPLICATIONS

- Individuals who disclose a mental health diagnosis systematically differ from the larger population living with that condition (Ernala et al., 2019)
- We need to differentiate between:
  1. Classifiers that estimate whether someone has a mental health condition
  2. Classifiers that estimate whether someone has a mental health condition *AND* discloses their condition online (i.e., selection bias)

Annotate date of diagnosis and comorbidities	Sample control groups using propensity score matching	Identify and filter sample selection biases
"1 year ago", "Depression and Eating Disorder"	Distribution of interests, personality, temporal activity	Fan accounts, non-personal accounts



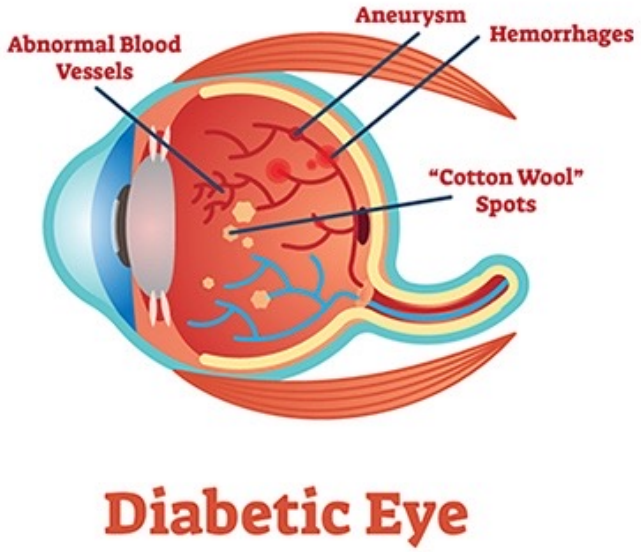
COUNTERACTING THE EFFECTS OF SELECTION BIAS

Defensive Tactics for Promoting Health Equity

BACKGROUND

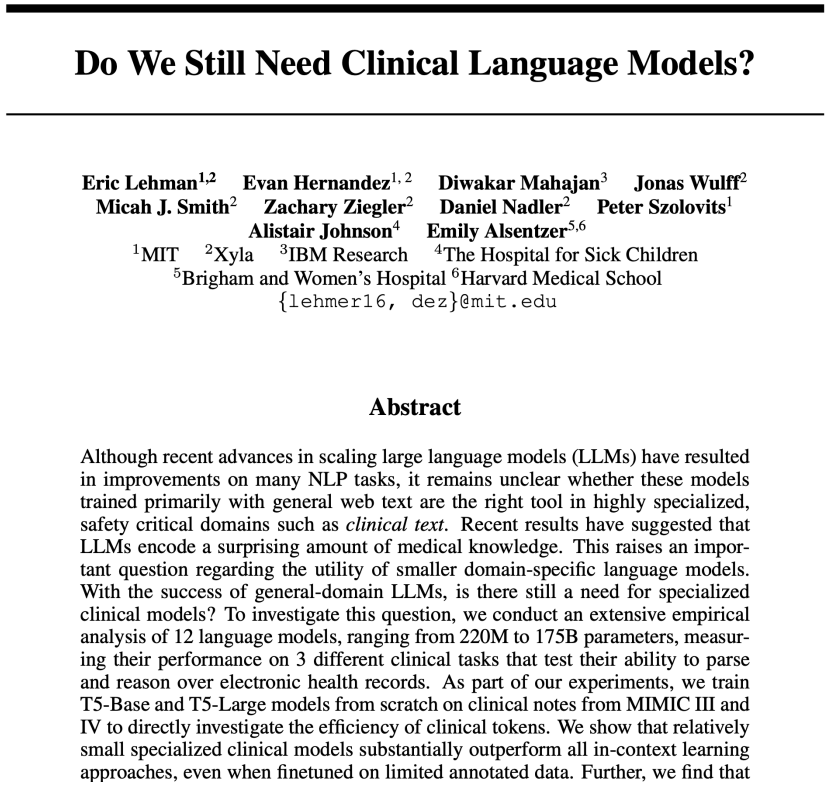
Clinical Problem

- Diabetic eye disease (e.g., diabetic retinopathy) is the leading cause of blindness in people aged 20 to 64 years old
- Stringent screening and management of the disease is essential, as it can prevent vision loss and even return some visual fidelity
- Diagnostic codes (i.e., ICD-10) and other structured fields in the EHR are not sensitive enough for monitoring the condition, unlike free text in clinical notes



Let's build a system to extract and categorize ophthalmic clinical concepts

CONTEMPORARY UNDERSTANDING



Do We Still Need Clinical Language Models?

Eric Lehman[1,2] **Evan Hernandez**[1, 2] **Diwakar Mahajan**[3] **Jonas Wulff**[2]
Micah J. Smith[2] **Zachary Ziegler**[2] **Daniel Nadler**[2] **Peter Szolovits**[1]
Alistair Johnson[4] **Emily Alsentzer**[5,6]
[1]MIT [2]Xyla [3]IBM Research [4]The Hospital for Sick Children
[5]Brigham and Women's Hospital [6]Harvard Medical School
{lehmer16, dez}@mit.edu

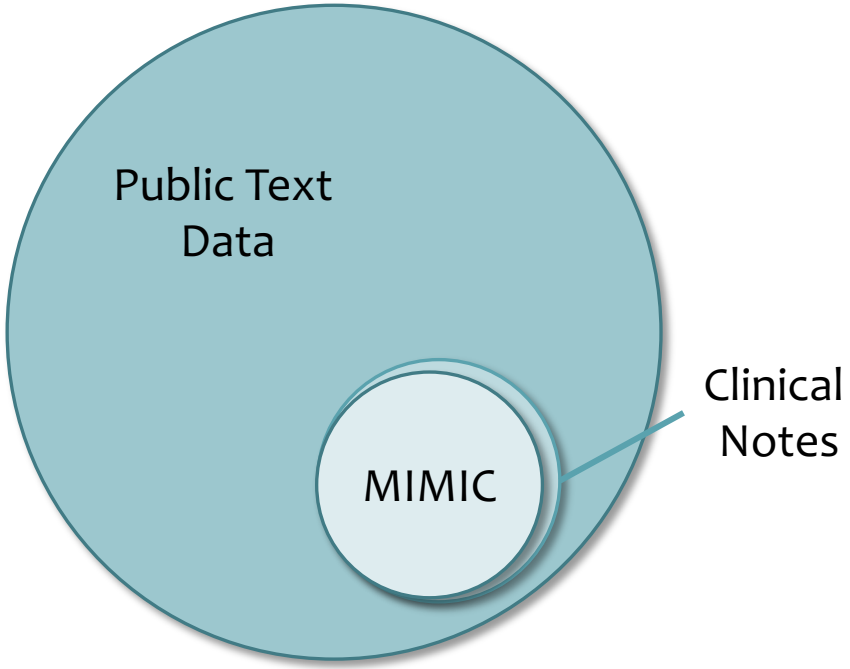
Abstract

Although recent advances in scaling large language models (LLMs) have resulted in improvements on many NLP tasks, it remains unclear whether these models trained primarily with general web text are the right tool in highly specialized, safety critical domains such as *clinical text*. Recent results have suggested that LLMs encode a surprising amount of medical knowledge. This raises an important question regarding the utility of smaller domain-specific language models. With the success of general-domain LLMs, is there still a need for specialized clinical models? To investigate this question, we conduct an extensive empirical analysis of 12 language models, ranging from 220M to 175B parameters, measuring their performance on 3 different clinical tasks that test their ability to parse and reason over electronic health records. As part of our experiments, we train T5-Base and T5-Large models from scratch on clinical notes from MIMIC III and IV to directly investigate the efficiency of clinical tokens. We show that relatively small specialized clinical models substantially outperform all in-context learning approaches, even when finetuned on limited annotated data. Further, we find that

"Do We Still Need Clinical Language Models." (Lehman et al., 2023)

"Further, we find that pretraining on clinical tokens allows for smaller, more parameter-efficient models that either match or outperform much larger language models trained on general text."

AN OPEN QUESTION

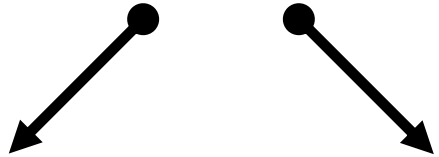


Clinical Language Models
or
MIMC Language Models?

Clinical LMs are often trained and tested on the same data source, limiting our ability to understand how well they **generalize** across domains and patient populations.

SPECIFIC AIMS

How should practitioners approach language modeling in the clinical domain when working on specific targeted problems?



What language model should practitioners use as the foundation?

General vs. Clinical Language Model
General vs. Clinical Vocabulary

How should practitioners train language models to maximize downstream performance?

Task Fine Tuning
Domain Adaptive Pretraining

DATA

Data Acquisition Challenges

- HIPAA-compliant span-level annotation tool not readily available for our team
- Challenging to do span-level labeling at scale

Solution

- Leverage regular expressions to identify clinical concepts related to diabetic eye disease
- Assign match validity and attribute labels (e.g., severity, laterality)

Clinical Concept	ICD-10	∩	Text
A1 - DR (General)	37,743	85,878	55,848
A2 - NPDR	24,970	28,203	14,594
A3 - PDR	14,217	32,212	13,038
A4 - NV	7,624	14,945	58,970
B1 - ME	37,081	83,852	50,719
C1 - VH	3,569	4,882	28,691
C2 - RD	4,337	15,279	70,781
C3 - NVG	155,429	6,127	3,418
D1 - Anti-VEGF	0	0	93,038
D2 - PRP	0	0	41,531
D3 - Focal Grid Laser	0	0	7,339
D4 - Other Injections	0	0	8,230
E1 - Retina Surgery	0	0	90,680
E2 - NVG Surgery	19	11	34,980
F1 - Diabetes Mellitus	44,019	165,813	54,499
G1 - Nephropathy	3,073	33	1,941
G2 - Neuropathy	5,615	363	5,997
G3 - Heart Attack	56	3	4,796
G4 - Stroke	783	921	10,452

notes containing each concept based on different criteria (ICD-10 codes vs. regular expressions)

Document ID: ad1fc53fe509fdea65d2099d8b5b3c57a8b5d1978f9bb8f0fa5eb1c427015aaf
Encounter Date: 2015-06-21

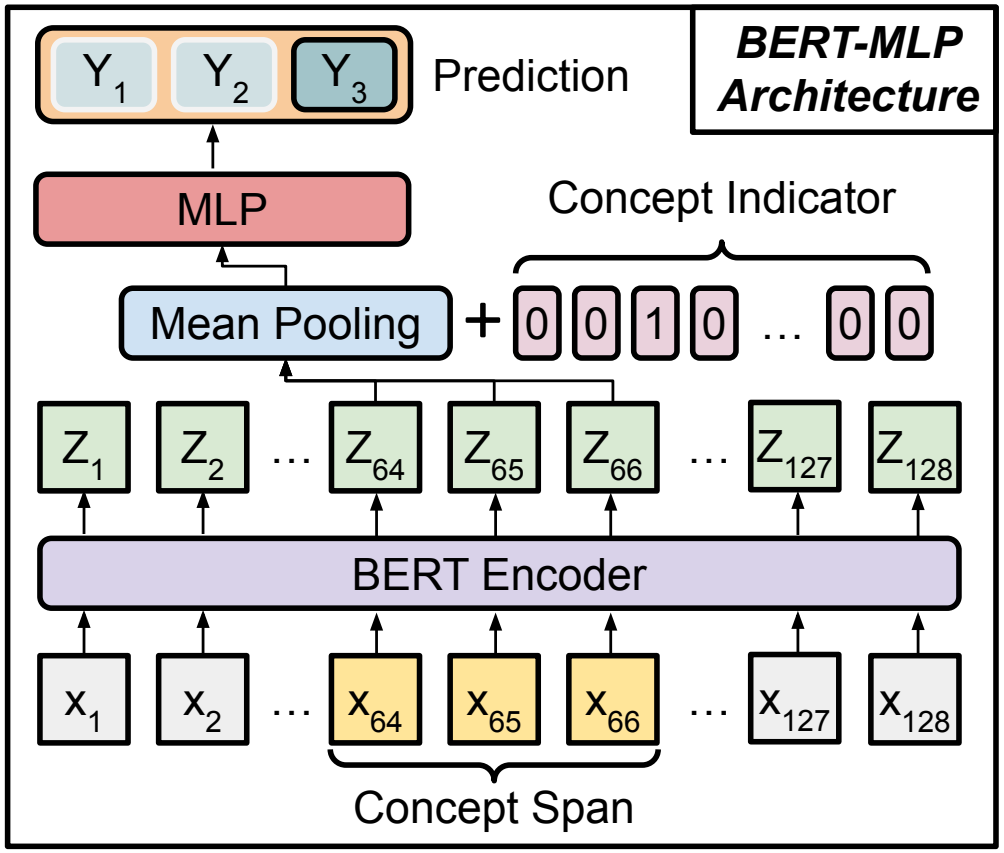
[[[ENCOUNTER ICD-10 CODES]]]
[[E11.319: Diabetic retinopathy]]
[[E11.311: Diabetic macular edema, both eyes]]

[[[PROBLEM LIST]]]
[[E11.3313: Diabetic macular edema of both eyes with moderate nonproliferative diabetic retinopathy associated with Type 2 diabetes mellitus]]
[OVERVIEW]
Eylea initiated right eye 6/2015 and and left eye 5/2014. No progression to PDR.
[ASSESSMENT & PLAN]
Right eye has foveal edema. Eylea #1 given. She will return in 2 weeks for eylea left eye after vacation to MI.

Start	End	Concept	Text Span	Context	Laterality	Severity/Type	Temporality	Negated	Incorrect
31	38	DR (General)	E11.319	[[«E11.319»: Diabetic Retinopathy]]	▼ OU	–	▼ Active	▼	▼
31	38	DM	E11.319	[[«E11.319»: Diabetic Retinopathy]]	–	▼ Type 2	▼ Active	▼	▼
267	270	PDR	PDR	left eye 5/2014. No progression to «PDR».	▼ OU	▼	▼ Active	▼ Negated	▼
				...					
525	537	ME	foveal edema	Right eye has «foveal edema». Eylea #1	▼ OD	▼ CI-DME	▼ Active	▼	▼
601	603	Heart Attack	MI	eylea left eye after vacation to «MI».	–	–	▼	▼	▼ Incorrect

Taxonomy: 19 clinical concepts (14 Classification Tasks)
Annotations: 6,565 spans (12,723 individual attribute labels)

EXPERIMENTAL DESIGN



Architecture

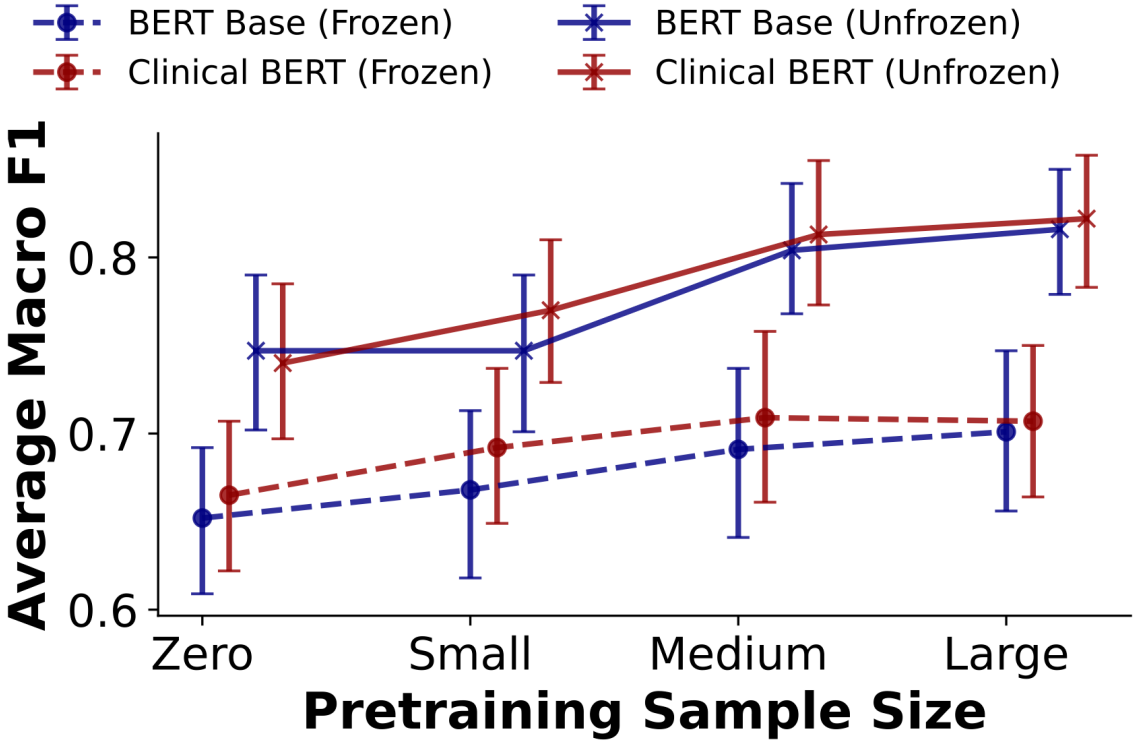
BERT -> Mean Pooling + Concept -> MLP

Ablations

1. Encoder Initialization (pretraining distribution)
2. Task Fine-tuning
3. Continued Pretraining
4. Sample Efficiency
5. Pretraining from Scratch + Vocabulary

RESULTS

- BERT pretrained on out-of-domain clinical data does not offer a significant advantage over BERT pretrained on non-clinical data
- Domain adaptation (task fine-tuning & continued pretraining) are necessary for maximizing downstream task performance
- Domain specific vocabulary significantly improves task performance



Initialization	Tokenizer	Frozen	Unfrozen
BERT Base	BERT Base	70 (66, 75)	82 (78, 85)
Random	BERT Base	71 (66, 75)	77 (73, 81)
Random	Learned	71 (67, 76)	81 (78, 84)

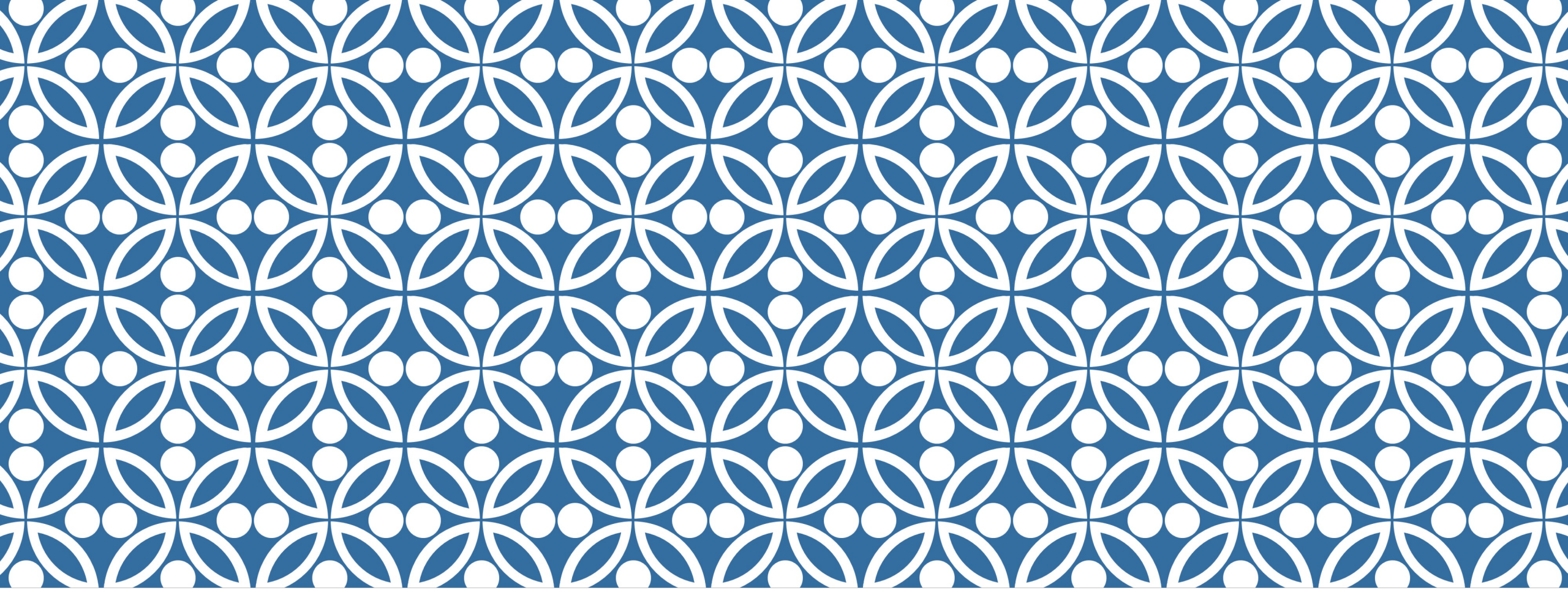
IMPLICATIONS

Not All Clinical Data is the Same

- *Domain-specific* clinical language models still appear to be valuable.
- We cannot assume that clinical LMs pretrained on MIMIC will generalize to new clinical domains.

Prioritize Domain Adaptation Over Domain Generalization

- Non-clinical data will remain orders of magnitude larger than clinical data.
- Non-clinical LLMs have already shown promise for biomedical and clinical tasks.
- Prioritize teaching LMs to understand language, and then to understand nuances of a specific clinical domain.



CHARACTERIZING AND MEASURING IMPLICIT BIAS IN MEDICAL RECORDS

Proactive Tactics for Promoting Health Equity

BACKGROUND

An Equity Problem

- Patients who experience discrimination (more frequently Black patients) have:
  - Lower levels of adherence to treatment plans
  - Lower trust in healthcare providers
  - Increased likelihood to delay care or avoid treatment & screening

Discrimination in Medical Records

- Healthcare providers who read notes containing discriminatory (or generally negative) language are more likely to formulate a less aggressive treatment plan
- 21st Century Cures Act mandates EHRs are readily available to all patients

BACKGROUND

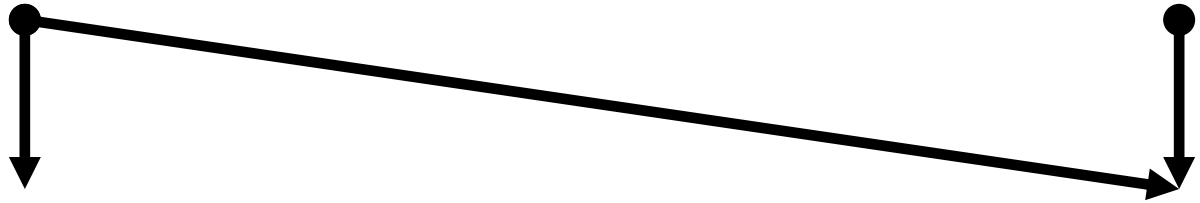
What is Stigmatizing Language?

- Stigmatizing language assigns negative labels, stereotypes, and judgment to certain groups of people.
- In NLP, most focus thus far on mental health and addiction
  - "Addict"
  - "Substance Abuse"
  - "Crazy"
  - "Junkie"

SPECIFIC AIMS

How does stigmatizing language in medical records compare to other harmful language?

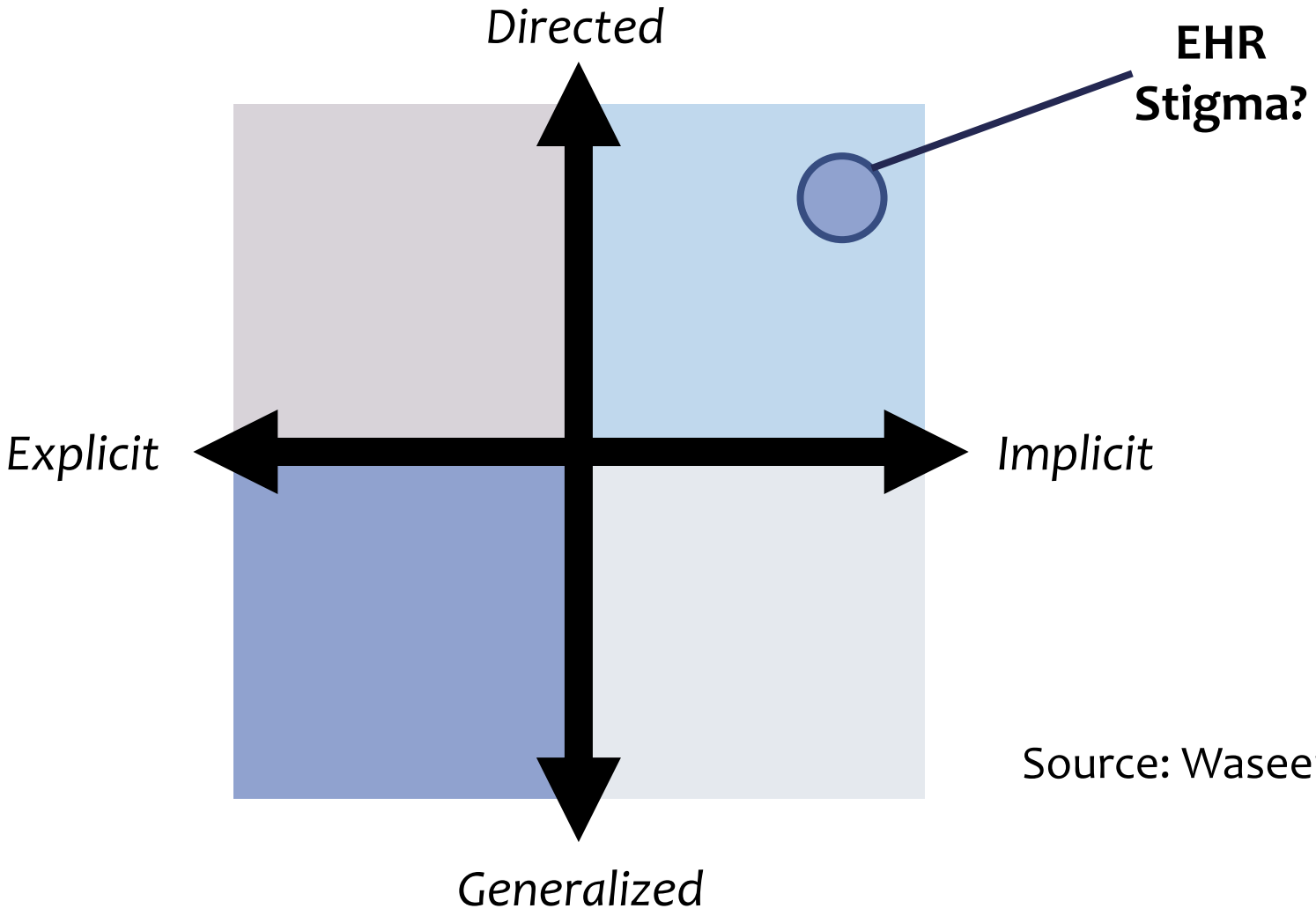
What types of stigmatizing language reflect a health disparity?



What role does context play in characterizing stigmatizing language?

How does stigmatizing language usage differ across patient populations?

ABUSIVE LANGUAGE TAXONOMY



Source: Waseem et al. (2017)

DATA

Johns Hopkins Medicine (Private)

- English-language progress notes
- 5 clinical specialties are represented – internal medicine, emergency medicine, pediatrics, OB-GYN, and general surgery (Baltimore, MD)
- 5,201 labeled instances

MIMIC-IV (Public)

- De-identified, English discharge notes
- Patients admitted to emergency department or an intensive care unit at Beth Israel Deaconess Medical Center (Boston, MA)
- 5,043 labeled instances

STIGMATIZING LANGUAGE TAXONOMY

Credibility & Obstinacy

Class	Definition	Examples
Disbelief	Insinuates doubt about a patient's stated testimony.	*adamant* he doesn't smoke; *claims* to see a therapist
Difficult	Describes patient perspective as inflexible/difficult/entrenched, typically with respect to their intentions.	*insists* on being admitted; *adamantly* opposed to limiting fruit intake
Out of Context	Word/phrase is not used to characterize the patient or describe the patient's behavior; may refer to medical condition or treatment or to another person or context.	patient's friend *insisted* she go to the hospital; test *claims* submitted to insurance

STIGMATIZING LANGUAGE TAXONOMY

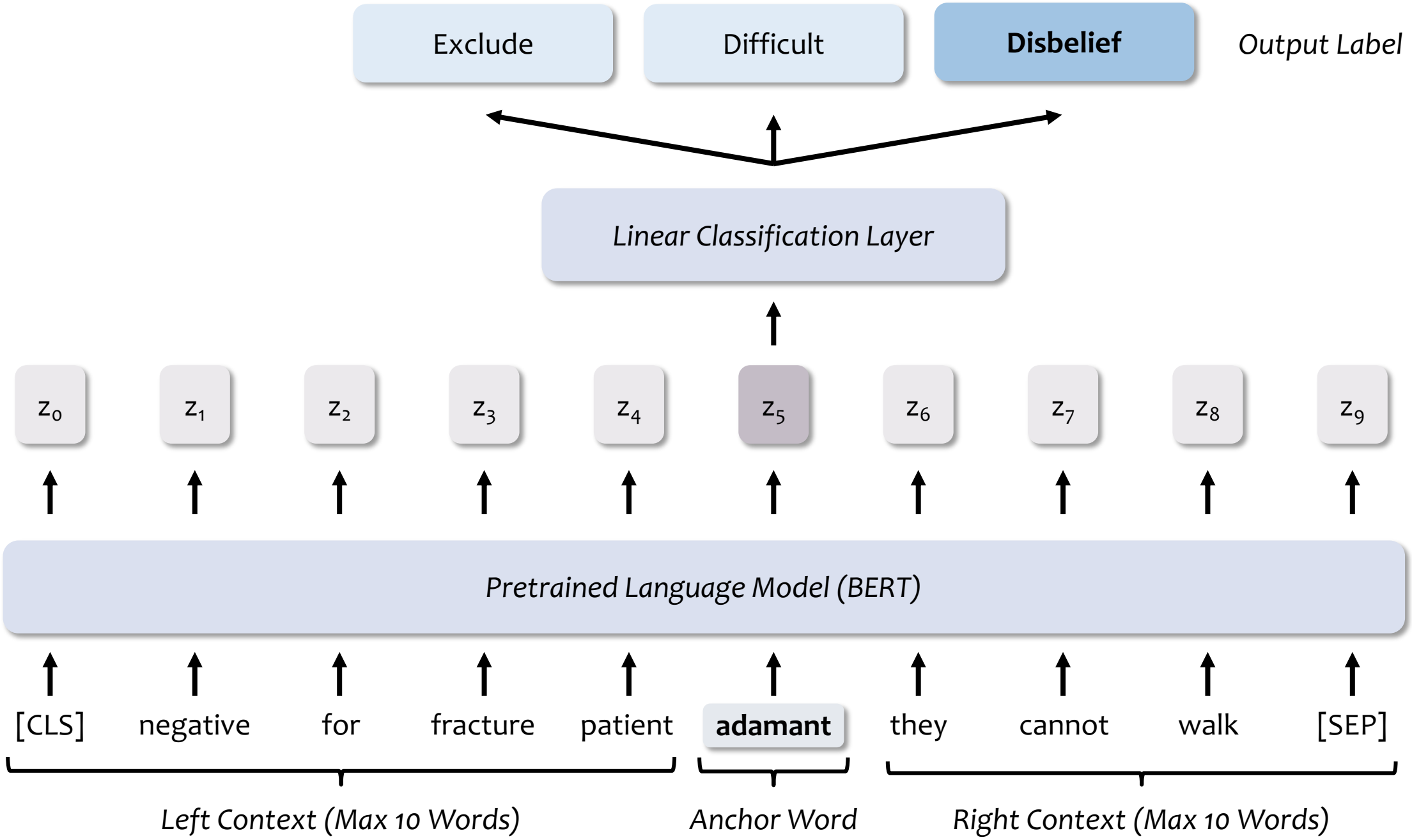
Compliance

Class	Definition	Examples
Negative	Patient not, unlikely to, or questionably following medical advice	*adherence* to therapeutic medication is unclear; mother *declines* vaccines; struggles with medication and follow-up *compliance*
Neutral	Not used to describe whether the patient is not following medical advice or rejecting treatment; often used to describe generically some future plan involving a hypothetical.	discussed medication *compliance*; school *refuses* to provide adequate accommodations; feels that her parents' health has *declined*
Positive	Patient following medical advice.	continues to be *compliant* with aspirin regimen; reports excellent *adherence*

STIGMATIZING LANGUAGE TAXONOMY

Descriptors

Class	Definition	Examples
Negative	Patient's demeanor cast in a negative light; insinuates the patients is not being forthright	concern for *secondary gain*; *unwilling* to meet with case manager; *poorly-groomed* today
Neutral	Negation of negative descriptors; insinuates the patient was expected to have a negative demeanor.	not *combative* or *belligerent*; dad seems *angry* with patient at times
Positive	Patient's demeanor or behavior is described in a positive light; patient is easy to interact with.	*lovely* 80 year old woman; *well-groomed* and holds good eye contact
Out of Context	Patient self-description or description of another individual.	does not want providers to think she's *malingering*; reports feeling *angry*


Exclude
Difficult
Disbelief
Output Label
Linear Classification Layer
z_0
z_1
z_2
z_3
z_4
z_5
z_6
z_7
z_8
z_9
Pretrained Language Model (BERT)
[CLS]
negative
for
fracture
patient
adamant
they
cannot
walk
[SEP]
Left Context (Max 10 Words)
Anchor Word
Right Context (Max 10 Words)

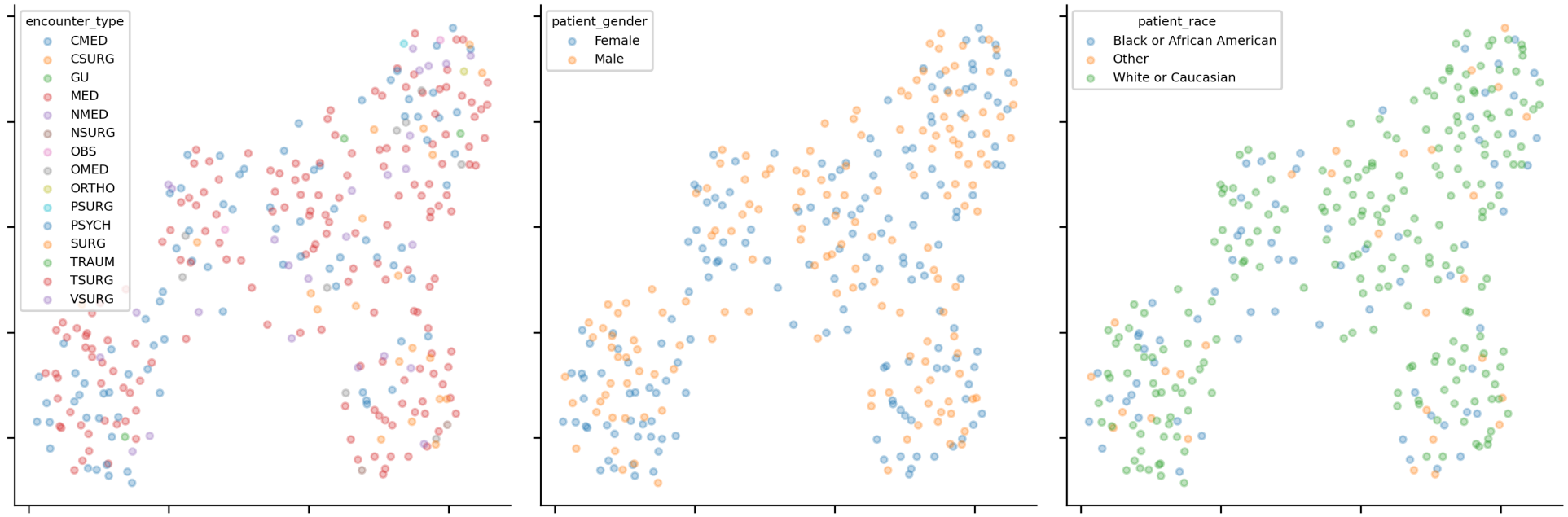
WHAT ROLE DOES CONTEXT PLAY IN CHARACTERIZING STIGMATIZING LANGUAGE?

	Credibility & Obstinacy		Compliance		Descriptors	
Model	**JHM**	**MIMIC**	**JHM**	**MIMIC**	**JHM**	**MIMIC**
Majority Overall	0.21 ± 0.00	0.17 ± 0.00	0.29 ± 0.00	0.24 ± 0.00	0.16 ± 0.00	0.19 ± 0.00
Majority Per Anchor	0.67 ± 0.10	0.55 ± 0.04	0.68 ± 0.04	0.73 ± 0.01	0.82 ± 0.01	0.83 ± 0.00
LR (Context)	0.60 ± 0.05	0.58 ± 0.04	0.55 ± 0.01	0.68 ± 0.02	0.74 ± 0.03	0.60 ± 0.04
LR (Context + Anchor)	0.69 ± 0.02	0.65 ± 0.03	0.68 ± 0.04	0.80 ± 0.02	0.86 ± 0.02	0.76 ± 0.05
Bert (Web)	0.85 ± 0.04	0.76 ± 0.02	**0.86 ± 0.01**	**0.92 ± 0.02**	**0.93 ± 0.01**	**0.86 ± 0.01**
Bert (Clinical)	**0.89 ± 0.03**	**0.78 ± 0.03**	0.85 ± 0.02	**0.92 ± 0.02**	**0.93 ± 0.02**	**0.86 ± 0.01**
– CLS Token	0.89 ± 0.04	0.69 ± 0.03	0.84 ± 0.03	0.92 ± 0.01	0.90 ± 0.01	0.84 ± 0.03
– Sentence Mean	0.85 ± 0.06	0.69 ± 0.06	0.84 ± 0.03	0.92 ± 0.01	0.91 ± 0.01	0.84 ± 0.02
– BERT Pooler	0.83 ± 0.08	0.70 ± 0.07	0.84 ± 0.02	0.91 ± 0.02	0.89 ± 0.03	0.80 ± 0.03

Method: Comparison of architecture and last-layer pooling mechanism

Outcome: Acts like a word-sense-disambiguation task; both context + anchor are critical to performance

IS STIGMA CONVEYED IN THE SAME MANNER TOWARDS DIFFERENT DEMOGRAPHIC GROUPS?



Method: Attempt to infer patient demographics from last embedding layer

Outcome: After controlling for clinical setting and task label, no discernable difference in semantic representations across demographic groups

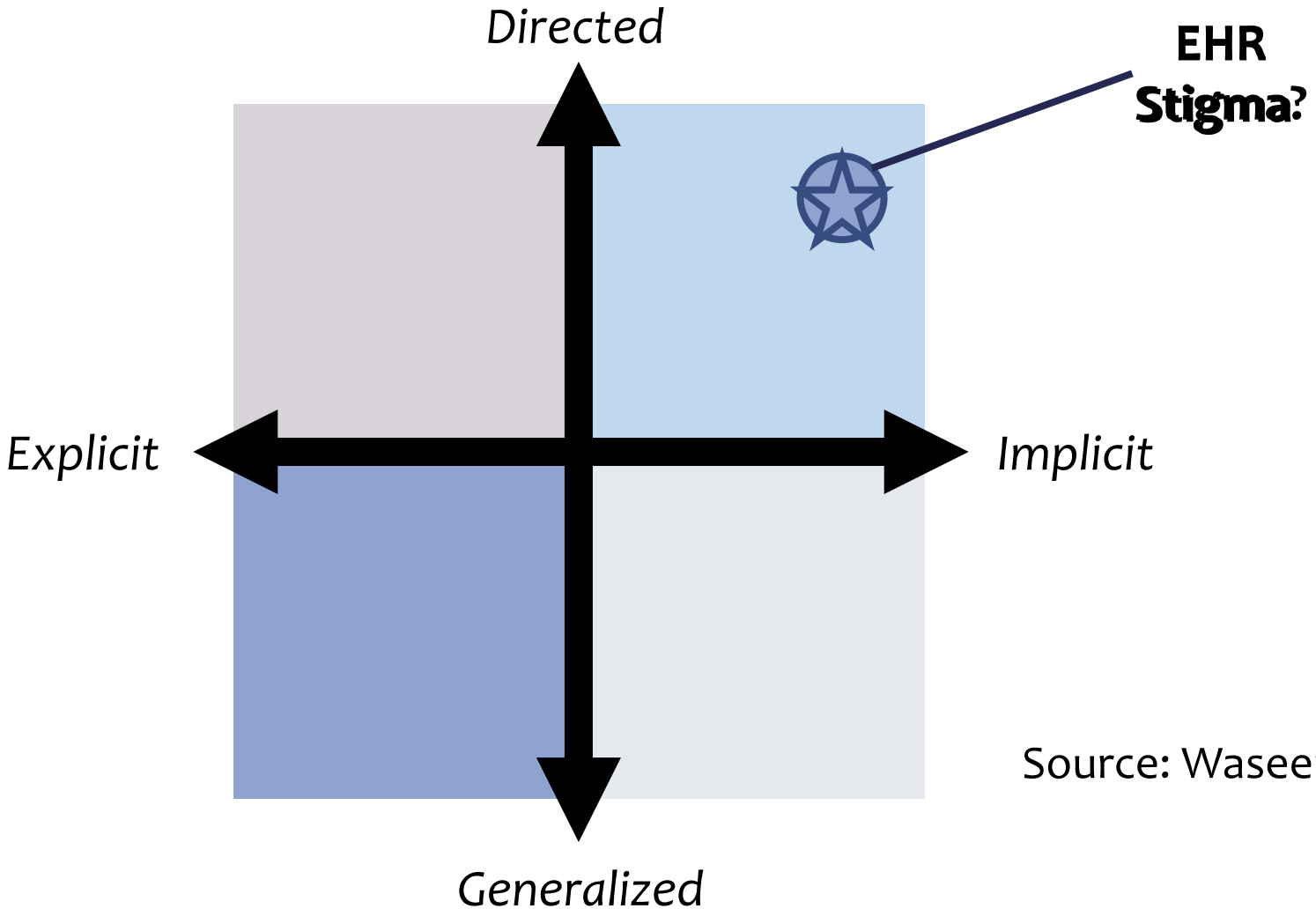
IS STIGMA CONVEYED IN THE SAME MANNER ACROSS DIFFERENT PATIENT POPULATIONS?

		Credibility & Obstinacy		Compliance		Descriptors	
	Target →	**JHM**	**MIMIC**	**JHM**	**MIMIC**	**JHM**	**MIMIC**
Source ↕	**JHM**	**0.89 ± 0.03**	0.70 ± 0.01	**0.85 ± 0.02**	0.86 ± 0.03	**0.93 ± 0.02**	0.81 ± 0.03
	MIMIC	0.81 ± 0.03	**0.78 ± 0.03**	0.82 ± 0.02	**0.92 ± 0.02**	0.89 ± 0.03	**0.86 ± 0.01**

Method: Train on one dataset (e.g., JHM) and test on different dataset (e.g., MIMIC-IV)

Outcome: Significant loss in performance; concept shift (e.g., label + anchor distribution) and clinical setting differences (e.g., family references, psych)

ABUSIVE LANGUAGE TAXONOMY



Source: Waseem et al. (2017)

MEASURING HEALTH DISPARITIES

Exploring Stigmatizing Language in a Public Dataset

- MIMIC-IV models applied to MIMIC-IV discharge notes (280.7k) from patients with known (binary) gender and race of either White, Black, or Hispanic
- Measure rates for conceptual groupings of words
- Linear Mixed Effects Model (repeated measures) to test for statistical significance

Conceptual Groupings

Credibility	Negative Compliance	Negative Appearance
Obstinacy	Negative Compliance (Appropriate)	Positive Appearance
Credibility (w/o Historian)	Negative Compliance (Inappropriate)	Positive Demeanor
Obstinacy (w/o Agitated)	Positive Compliance	

MEASURING HEALTH DISPARITIES

Outcome	Female OR	Hispanic OR	Black OR
Credibility	0.95 (0.83, 1.09)	0.97 (0.71, 1.32)	1.24 (1.04, 1.49) *
Credibility (w/o Historian)	0.93 (0.76, 1.14)	1.14 (0.74, 1.76)	1.18 (0.91, 1.53)
Obstinacy	0.74 (0.67, 0.82) *	0.79 (0.62, 1.00)	0.91 (0.79, 1.06)
Obstinacy (w/o Agitated)	0.77 (0.66, 0.90) *	0.93 (0.65, 1.33)	1.20 (0.98, 1.48)
Negative Compliance	1.01 (0.98, 1.05)	1.04 (0.97, 1.13)	1.79 (1.70, 1.87) *
Negative Compliance (Appropriate)	1.03 (0.94, 1.13)	1.05 (0.85, 1.28)	1.57 (1.39, 1.77) *
Negative Compliance (Inappropriate)	0.98 (0.90, 1.06)	1.07 (0.89, 1.28)	1.73 (1.55, 1.92) *
Positive Compliance	0.85 (0.76, 0.96) *	1.44 (1.13, 1.84) *	1.43 (1.22, 1.67) *
Negative Appearance	0.68 (0.50, 0.93) *	0.91 (0.45, 1.85)	0.85 (0.54, 1.33)
Positive Appearance	1.13 (1.13, 1.13) *	1.13 (0.59, 2.17)	1.29 (0.90, 1.86)
Positive Demeanor	1.10 (1.08, 1.13) *	0.91 (0.86, 0.96) *	1.04 (1.01, 1.08) *

- Male patients more frequently described as difficult
- Black patients more frequently have their testimony questioned
- Positive and negative compliance discussed more frequently for Black patients

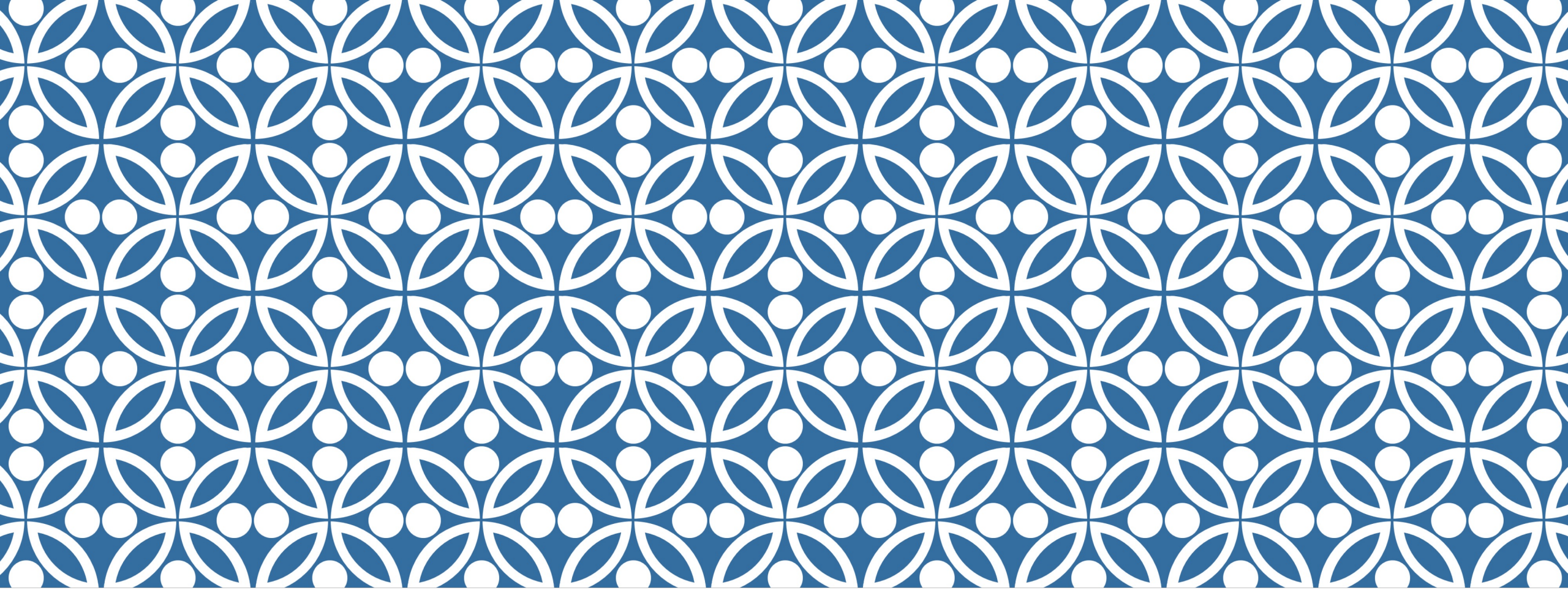
IMPLICATIONS

A Call to Action

- Demonstrated that NLP can be used in a proactive manner to identify disparities that were previously unknown or unable to be measured
- Time to start thinking about stigmatizing language when training LLMs

A Warning

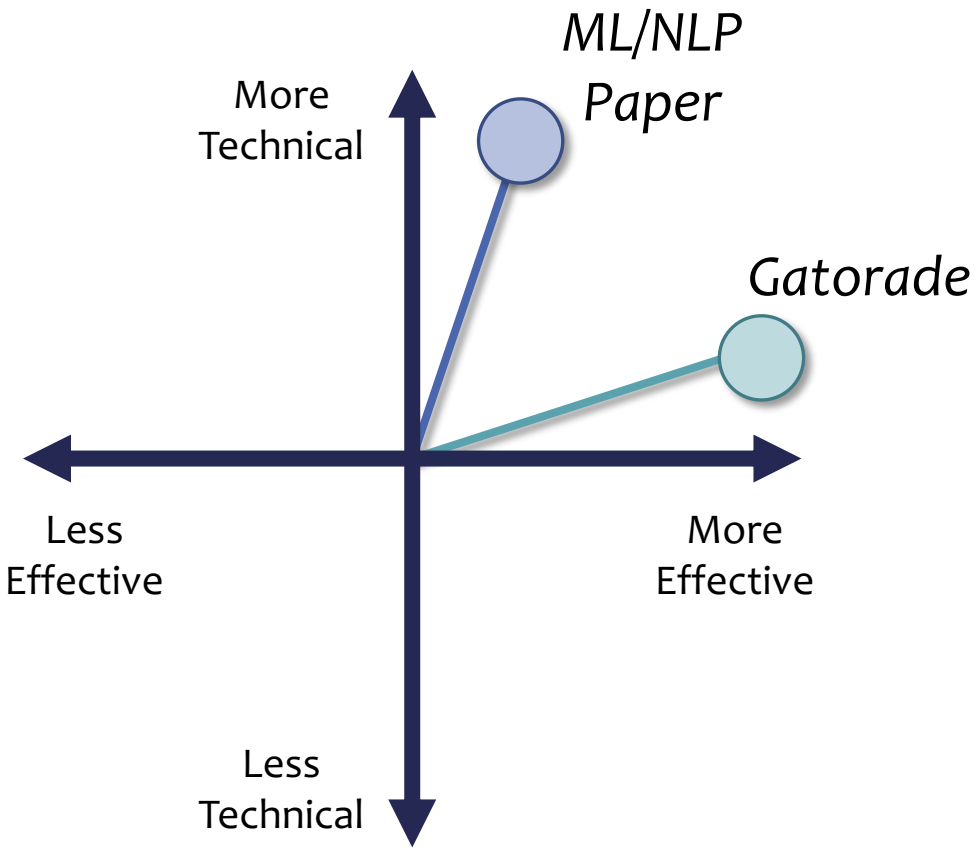
- As before, clinical text is far from homogenous
- Distribution shift can arise in ways that are difficult to understand without in-depth, task-specific analysis



FUTURE DIRECTIONS

Conclusion

THE STATE OF NLP FOR HEALTH CARE



Credit: Charles Delahunt (ML4H 2023)

A Time to Reflect

- What do healthcare providers *actually* need to do their jobs more effectively?
- Focus on underserved and marginalized populations (arguably lower hanging fruit)
- Consider a third-axis (more vs. less targeted)

FUTURE DIRECTIONS

Measuring and Understanding Distribution Shift

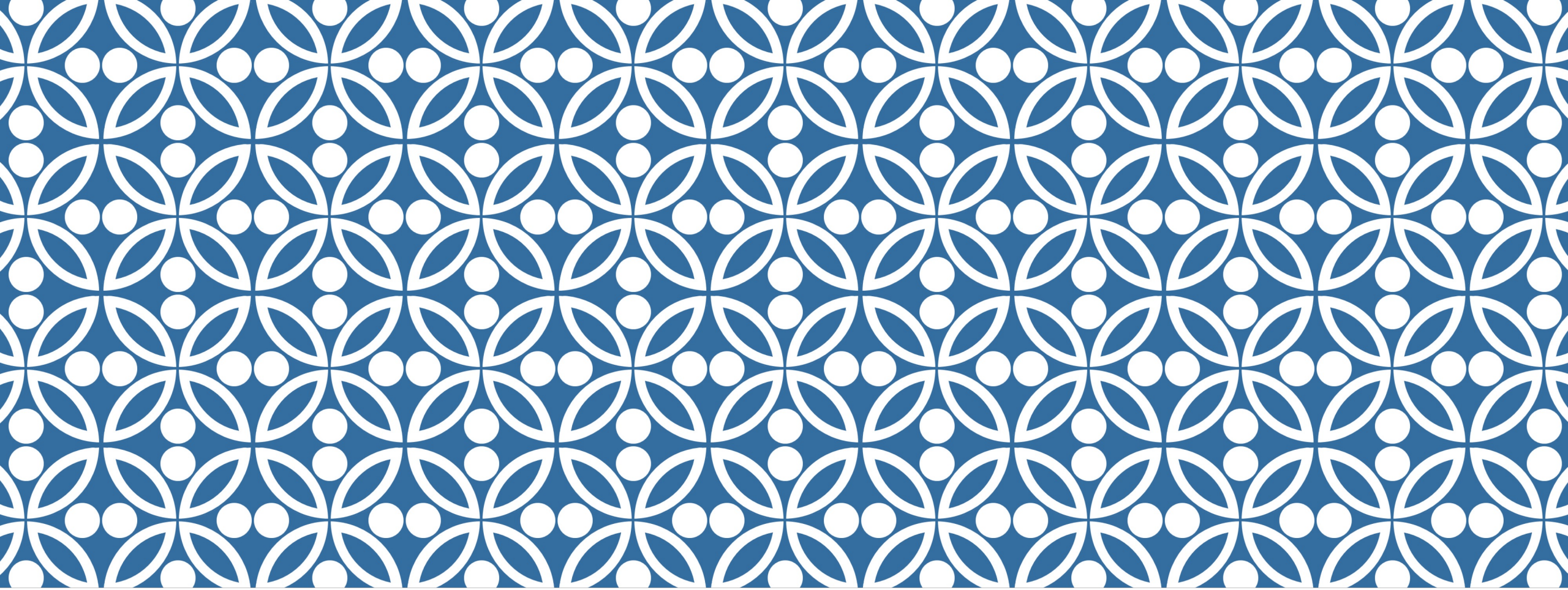
- LLMs for interactive dataset analysis and bias discovery

Promoting Robustness Under Distribution Shift

- Focus on adaptation over generalization
- Low-resource and test-time adaptation strategies

Identifying and Mitigating Health Disparities

- Expanded analysis of stigmatizing language disparities
- Effects on LLMs
- Quotes, complex-forms of stigmatization, personalization



CONTACT INFO

Email: kharrigian@jhu.edu
Website: kharrigian.github.io

INTER-RATER RELIABILITY

Three Annotators (Author A_1; Non-authors B_1, B_2)

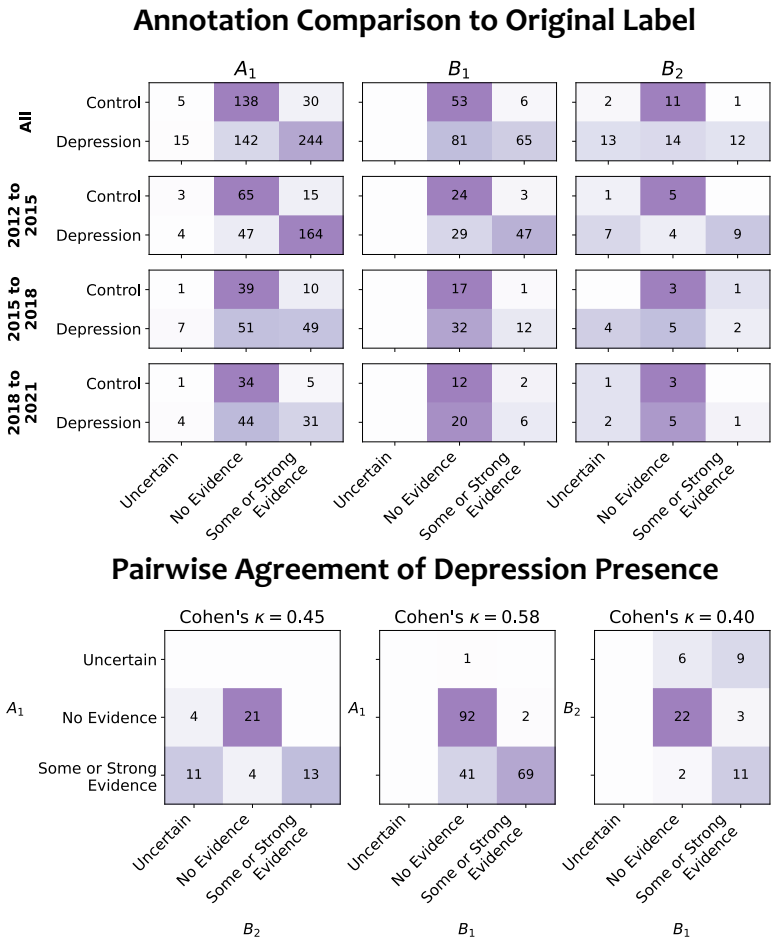
- Several years experience modeling mental health within social media, but not clinical experts

Reliability Measures (Krippendorff α)

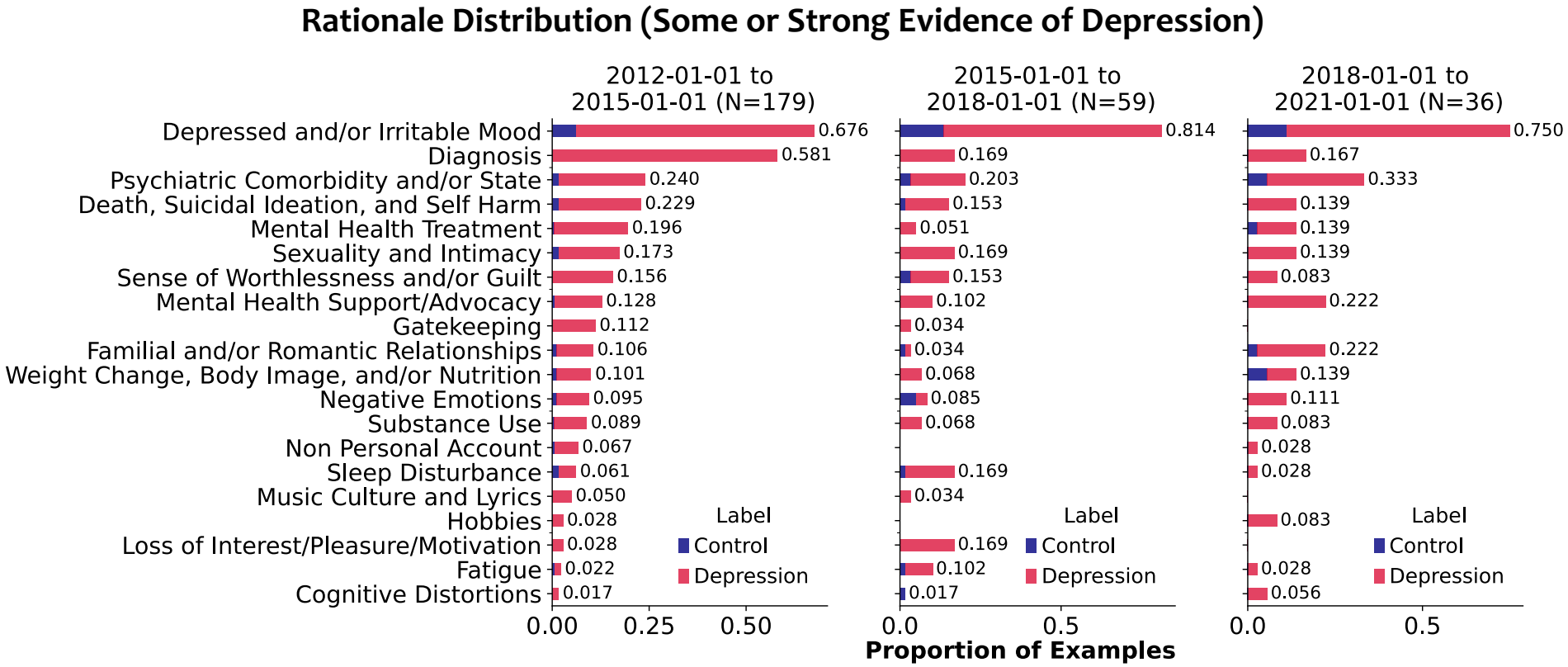
- Evidence of Depression $\alpha = 0.4988$ (Fair to Moderate)
- Remission Status $\alpha = 0.3561$ (Poor to Fair)

Causes of Disagreement

- Prevalence to indicate uncertainty about label
- Exposure bias
- Sensitivity to depressed mood and/or negative emotion



QUALITATIVE RESULTS



Themes: personality (e.g., elevated neuroticism), comorbid conditions, and a propensity for oversharing (e.g., taboo topics)