## Zucker Hillside Hospital Northwell Health

### Introduction

- **Background:** Language models can detect psychosis-related language disturbances by measuring semantic similarity between speech units, but no standardized approach exists for model selection or transcript preprocessing.
- □ **Models Compared:** Static models (GloVe, LSA, word2vec) vs. contextual models (GPT-2, RoBERTa, LLaMA), which differ in their ability to capture semantic and grammatical context.
- □ **Preprocessing Levels:** Verbatim (Level 1), removal of disfluencies/repetitions (Level 2), and additional stop-word removal (Level 3).
- **Evaluation:** Models are assessed using established semantic similarity metrics and correlated with categorical SSD diagnosis and severity of disturbance three language inefficient speech, expressivity, dimensions (impaired incoherence).

	Participant Cl	naracteristics	
	Healthy Control	SSD	p Value
n (%)	76 (33%)	152 (67%)	
Age mean (SD)	29 (7)	28 (7)	0.285
Sex			0.016*
Female (%)	41 (54%)	55 (36%)	
Male (%)	35 (46%)	97 (64%)	
Race			0.141
Asian	12 (16%)	19 (12%)	
Black	27 (36%)	58 (38%)	
White	30 (39%)	45 (30%)	
Other	7 (9%)	30 (20%)	
TLC Total Score	2 (3)	17 (14)	0.000***
	Average Fac	ctor Scores	
Impaired Expressivity	-0.30	0.28	0.000***
Inefficiency	-0.58	0.59	0.000***
Incoherence	-0.38	0.53	0.000***

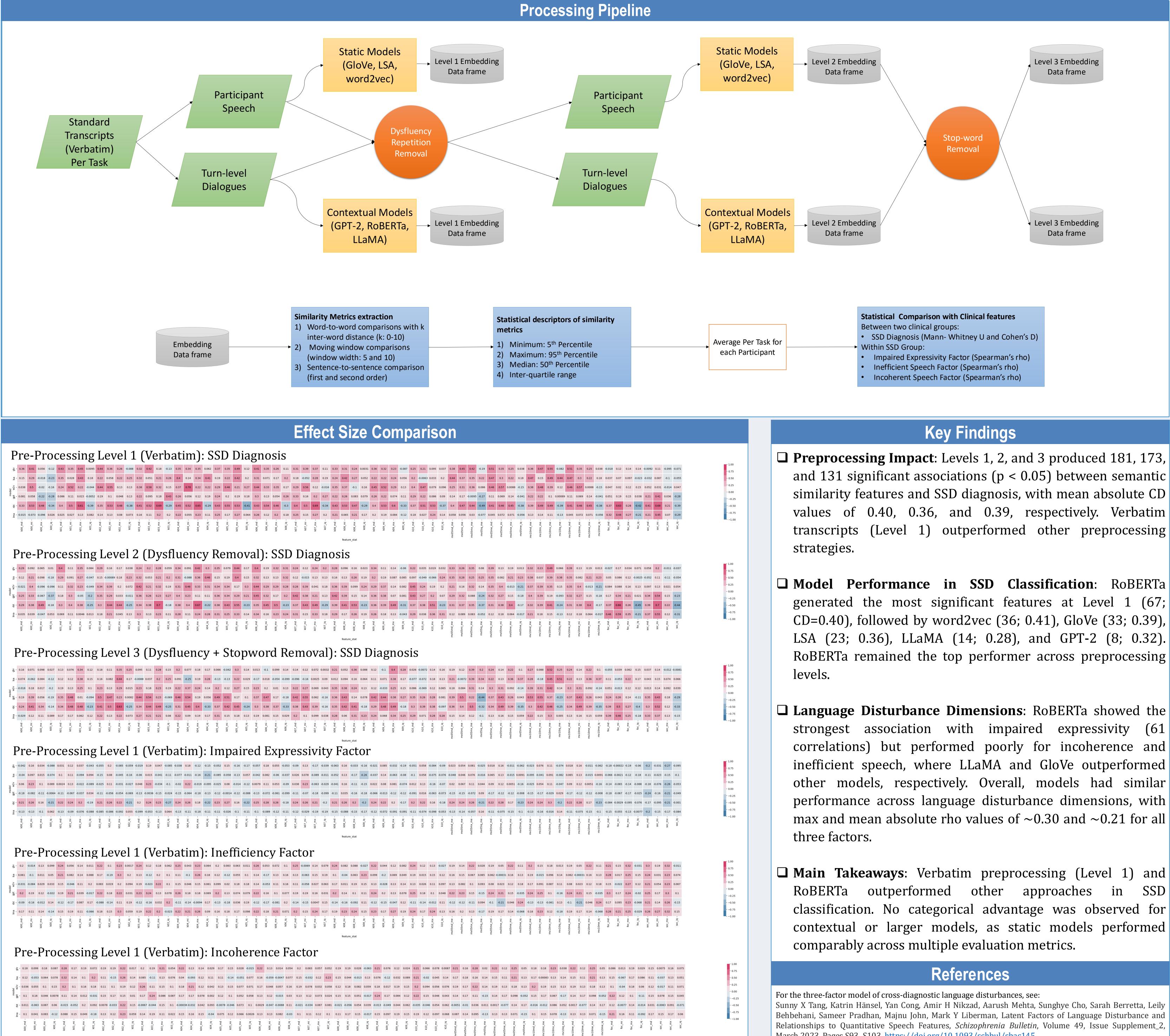
## **Feature Description and Nomenclature**

- **K-inter-word distances:** Compares word-to-word similarity at 0–10-word intervals. For example, k00 Compares each word to the next word (immediate neighbor, 0-word interval), whereas k01 compares each word to the word after the next (skipping one word, 1-word interval).
- □ **Moving window similarity:** Calculates average similarity of all word pairs within a fixed window size of 5 (mv05) or 10 (mv10).
- **Sentence similarity:** Compares sentence to sentence similarity between adjacent sentences (first–order coherence: foc) or between each sentence and the sentence after the next (second-order coherence: soc)
- **Statistical descriptors:** Comparisons across a sample generate a set of similarity metrics, which are summarized using minimum (mi: 5<sup>th</sup> percentile), median (md), maximum (mx: 95th percentile), and inter-quartile range (iq). For example, k03\_mx represents the 95th percentile similarity value for inter-word comparisons at a 3-word interval.

## **Evaluating Language Models and Preprocessing Strategies** for Detecting Psychosis-Related Language Disturbances

# Amir Nikzad<sup>1</sup>, Yan Cong<sup>2</sup>, Sunghye Cho<sup>3</sup>, Ryan Partlan<sup>1</sup>, Jiefei Li<sup>1</sup>, Sunny Tang<sup>1</sup>

1. Zucker Hillside Hospital, Northwell Health. 2. School of Languages and Cultures, Purdue University. 3. Department of Linguistics, University of Pennsylvania.



	feature_stat
Pre-Processing Level 2 (	Dysfluency Removal), SSD Diagnosis

	ΓΙ	e-			CE	:5:	511	ΠĘ	ζL	Je.	ve	:1 4	Z	L	<b>y</b>	511	lu	er	IC	У	Πt	211	10		al <sub>.</sub>	J	22		' L	710	ag	110	05	15								
Ť	- 0.29	0.092	0.065	0.01	0.4	0.11	0.35	0.084	0.29	0.16	0.17	0.038	0.24	0.2	0.28	0.059	0.34	0.091	0.42	0.3	0.35	0.079	0.44	0.17	0.4	0.19	0.32	0.31	0.24	0.12	0.24	0.2	0.28	0.096	0.16	0.023	0.34	0.11	0.14	-0.06	0.22	0.0
1	2 - 0.12	0.21	0.098	-0.18	0.29	0.091	0.27	-0.047	0.15	-0.00089	9 0.18	0.23	0.32	0.053	0.21	0.2	0.31	-0.088	0.36	0.46	0.15	0.19	0.4	0.15	0.32	0.13	0.13	0.32	0.12	-0.015	0.13	0.13	0.16	0.13	0.26	0.19	0.2	0.19	0.087	0.085	0.097	-0.0
del	0.021	0.4	-0.096	-0.096	0.11	0.32	0.23	-0.049	0.34	0.39	0.2	0.072	0.42	0.21	0.32	0.19	0.31	0.46	0.33	0.31	0.34	0.34	0.17	0.3	0.44	0.29	0.29	0.28	0.28	0.39	0.041	0.18	0.36	0.39	0.099	0.24	0.29	0.37	0.14	0.082	0.45	0.2
ou -	<u>6</u> - 0.25	0.33	-0.067	-0.37	0.18	0.3	-0.05	-0.2	0.35	0.29	0.033	-0.011	0.36	0.26	0.23	0.27	0.4	0.23	0.11	0.11	0.36	0.34	0.29	0.21	0.45	0.32	0.17	0.2	0.42	0.38	0.21	0.13	0.42	0.39	0.15	0.14	0.36	0.38	0.07	0.081	0.43	0.2
ł	<u>-</u> 0.29	0.38	0.49	-0.18	0.3	0.4	0.38	-0.25	0.3	0.44	0.44	-0.25	0.34	0.38	0.7	-0.18	0.38	0.4	0.67	-0.22	0.38	0.43	0.55	-0.23	0.35	0.45	0.5	-0.23	0.37	0.43	0.49	-0.29	0.39	0.41	0.53	-0.23	0.36	0.39	0.49	-0.31	0.37	0.3
	- 0.035	0.039	-0.047	0.053	0.069	0.11	0.0046	0.013	0.18	0.21	0.045	0.13	0.3	0.13	0.23	0.11	0.28	0.11	0.24	0.28	0.31	0.25	0.33	0.14	0.34	0.16	0.23	0.24	0.31	0.15	0.33	0.18	0.29	0.17	0.26	0.19	0.26	0.18	0.13	-0.043	0.29	0.0
	- bm_00Å	k00_mi -	- xm_00Å	koo_iq -	- pm_roy	ko1_mi -	k01_mx -	k01_iq -	k02_md -	k02_mi -	k02_mx -	k02_iq -	k03_md -	k03_mi -	k03_mx -	k03_iq -	k04_md -	k04_mi -	k04_mx -	k04_iq -	k05_md -	k05_mi -	k05_mx -	k05_iq -	k06_md -	k06_mi -	k06_mx -	k06_iq -	k07_md -	k07_mi -	k07_mx -	k07_iq -	- bm_80Å	k08_mi -	k08_mx -	k08_iq -	- pm <sup>-</sup> 604	k09_mi -	- xm_eox	- pi_eox	k10_md -	k10 mi -
																																		featur	e_stat							

glv	- 0.16	0.071	0.098	0.027	0.13	0.076	0.34	0.12	0.18	0.11	0.35	0.25	0.095	0.11	0.28	0.15	0.2	0.077	0.18	0.17	0.066	-0.042	0.3	0.14	0.013	-0.1	0.099	0.14	0.14	0.12	0.072	0.0032	0.21	0.052	0.36	0.088	0.12	-0.1	0.4	0.28	0.026	-0
lsa	- 0.074	-0.062	0.084	-0.12	0.12	0.12	0.38	0.15	0.16	0.082	0.44	0.17	-0.0088	0.037	0.2	0.25	0.091	-0.25	0.19	0.28	-0.13	-0.13	0.22	0.029	-0.17	0.018	-0.054	-0.099	-0.096	-0.16	0.0025	0.09	0.012	0.094	0.16	0.064	0.11	0.071	0.38	0.17	-0.077	-0
del w2v	0.018	0.18	0.017	-0.2	0.19	0.13	0.25	0.1	0.23	0.13	0.29	0.015	0.23	0.18	0.23	0.19	0.22	0.37	0.24	0.14	0.2	0.12	0.27	0.15	0.23	0.2	0.01	0.13	0.22	0.27	0.069	0.043	0.35	0.38	0.24	0.13	0.12	-0.033	0.25	0.15	0.086	-0
gpt gpt	- 0.19	0.39	0.056	-0.19	0.35	0.48	0.01	-0.094	0.5	0.47	0.23	0.0082	0.44	0.54	0.23	-0.069	0.46	0.54	0.19	0.056	0.49	0.51	0.17	0.1	0.37	0.47	0.17	-0.18	0.42	0.55	0.062	-0.16	0.36	0.43	0.14	0.078	0.42	0.44	0.38	0.27	0.35	(
rbt	- 0.24	0.41	0.34	-0.14	0.34	0.48	0.48	-0.23	0.41	0.5	0.63	-0.25	0.34	0.44	0.49	-0.25	0.31	0.45	0.4	-0.33	0.37	0.42	0.45	-0.24	0.3	0.38	0.37	-0.33	0.38	0.43	0.39	-0.16	0.35	0.42	0.41	-0.18	0.29	0.48	0.49	-0.18	0.3	(
lma	0.029	0.12	0.11	0.089	0.17	0.17	0.062	0.12	0.22	0.13	0.22	0.073	0.27	0.21	0.21	0.08	0.22	0.09	0.19	0.17	0.31	0.15	0.18	0.13	0.19	0.081	0.15	0.029	0.2	0.1	0.099	0.038	0.28	0.06	0.31	0.23	0.24	0.068	0.34	0.25	0.29	0
	-	-	, ,	<u>.</u>	-	.=	×		-	-	, ,		-		, ,		-		, ×		-	-	' ×	÷	-		×		-		, ,		-		' ×		-	-	' ×		-	
	k00_mc	k00_m	koo_m	koo_ic	k01_m(	k01_m	k01_m;	k01_id	k02_m(	k02_m	k02_m;	k02_ic	k03_mc	k03_m	k03_m;	k03_ic	k04_m(	k04_m	k04_m)	k04_ic	k05_m(	k05_m	k05_m;	k05_ic	k06_m(	k06_m	k06_m	k06_id	k07_m(	k07_m	k07_m	k07_ic	k08_m(	k08_m	k08_m;	k08_id	k09_m(	m_604	(m_60)	k09_i	k10_m	

2	0.042	2 0.16	0.034	-0.088	0.031	0.12	0.037	-0.043	-0.055	0.2	-0.085	-0.059	-0.019	0.19	0.047	-0.085	-0.038	0.18	-0.12	-0.15	-0.052	0.15	-0.16	-0.17	-0.057	0.18	0.055	-0.053	-0.09	0.13	-0.17	-0.039	-0.063	0.16	-0.033	-0.16	-0.021	0.085	-0.032	-0.19	-0.051	0
<u>-</u>	0.04	0.097	0.015	-0.074	0.1	0.11	-0.094	0.094	-0.15	0.08	-0.045	-0.16	-0.06	0.015	-0.041	-0.11	-0.077	-0.011	-0.16	-0.21	-0.085	-0.058	-0.13	0.057	-0.042	0.082	-0.06	-0.037	0.026	0.078	-0.089	-0.011	-0.052	0.13	-0.17	-0.26	-0.037	0.14	-0.063	-0.08	-0.1	0.
del w2v	- 0.06	0.23	0.1	0.089	0.0024	0.13	-0.022	-0.089	-0.025	0.11	-0.031	-0.027	0.048	0.23	-0.034	-0.1	-0.02	0.22	-0.019	-0.095	-0.025	0.08	-0.014	-0.12	0.0079	0.11	0.053	-0.09	0.038	0.23	-0.063	-0.029	-0.041	0.16	-0.11	-0.15	0.022	0.08	0.081	-0.074	0.012	C
ant ant	0.18	-0.082	-0.11	-0.0064	-0.11	-0.067	-0.037	0.034	-0.11	-0.056	-0.054	-0.069	-0.13	-0.0036	-0.15	-0.024	-0.15	-0.064	-0.18	-0.13	-0.12	-0.0014	-0.12	-0.088	-0.13	-0.072	-0.061	-0.099	-0.11	-0.07	-0.18	-0.099	-0.11	0.035	-0.16	-0.18	-0.066	-0.013	-0.12	-0.12	-0.091	0
ŧ	- 0.21	0.26	0.16	-0.21	0.22	0.24	0.2	-0.19	0.21	0.26	0.22	-0.21	0.2	0.24	0.23	-0.27	0.24	0.26	0.18	-0.22	0.23	0.27	0.16	-0.22	0.25	0.28	0.26	-0.18	0.24	0.26	0.21	-0.2	0.21	0.26	0.2	-0.2	0.24	0.22	0.2	-0.17	0.2	C
em	0.13	-0.13	-0.1	0.062	-0.13	-0.09	-0.076	-0.088	-0.085	-0.086	-0.092	0.055	-0.099	-0.053	-0.13	0.064	-0.13	-0.11	-0.18	-0.11	-0.11	-0.026	-0.1	-0.11	-0.1	-0.089	-0.12	-0.11	-0.12	-0.029	-0.19	-0.19	-0.15	-0.088	-0.15	-0.17	-0.12	-0.072	-0.091	-0.091	-0.11	-0
	- bm_00Å	koo_mi -	- xm_00Å	k00_iq -	k01_md -	k01_mi -	k01_mx -	k01_iq -	k02_md -	k02_mi -	k02_mx -	k02_iq -	k03_md -	k03_mi -	k03_mx -	ko3_iq -	k04_md -	k04_mi -	k04_mx -	k04_iq -	- k05_md	k05_mi -	k05_mx -	k05_iq -	k06_md -	k06_mi -	- xm_ð0	k06_iq -	k07_md -	k07_mi -	k07_mx -	k07_iq -	k08_md -	k08_mi -	k08_mx -	k08_iq -	- bm_e0å	- im_609	- xm_60Å	- pi_eo	k10_md -	

مار ار	0.2	-0.014	0.13	0.099	0.24	0.056	0.14	0.011	0.22	0.1	0.23	0.0017	0.24	0.12	0.18	0.062	0.25	0.043	0.23	0.064	0.2	0.083	0.063	0.011	0.26	0.053	0.072	0.1	0.25	-0.0069	0.14	0.079	0.24	0.082	0.088	-0.027	0.22	0.044	0.12	0.082	0.24	1
lsa -	0.061	-0.1	0.011	0.05	0.21	0.082	0.14	0.088	0.17	-0.19	0.3	0.2	0.13	-0.12	0.2	0.1	0.11	-0.1	0.26	0.18	0.12	-0.12	0.055	0.1	0.14	-0.17	0.13	0.16	0.13	-0.063	0.15	0.19	0.1	-0.04	0.063	0.23	0.099	-0.2	0.069	0.049	0.19	c
lel w2v	-0.031	-0.084	0.029	0.033	0.15	-0.046	0.11	0.2	0.083	0.019	0.2	0.094	0.19	-0.023	0.22	0.1	0.15	0.046	0.15	0.061	0.099	0.02	0.18	0.18	0.14	-0.053	0.11	0.16	0.11	-0.058	0.027	0.063	0.17	0.011	0.19	0.15	0.13	-0.028	0.13	0.14	0.13	c
gpt	0.2	0.19	0.12	-0.022	0.19	0.21	0.039	-0.017	0.22	0.18	0.22	0.031	0.23	0.24	0.13	0.078	0.26	0.18	0.18	0.069	0.2	0.13	0.074	0.079	0.22	0.16	0.1	0.077	0.19	0.19	0.16	0.031	0.2	0.14	0.1	0.11	0.24	0.2	0.13	0.078	0.25	ſ
Ę.	-0.09	-0.16	-0.012	0.14	-0.12	-0.17	0.087	0.17	-0.088	-0.14	0.11	0.19	-0.12	-0.16	0.032	0.2	-0.11	-0.14	-0.0094	0.17	-0.13	-0.18	0.036	0.19	-0.12	-0.17	-0.081	0.2	-0.14	-0.15	0.0047	0.15	-0.14	-0.16	-0.092	0.11	-0.12	-0.15	-0.047	0.12	-0.11	-
ma	0.17	0.11	0.14	-0.14	0.15	0.19	0.11	-0.066	0.18	0.15	0.3	0.059	0.19	0.22	0.2	-0.023	0.22	0.21	0.26	0.09	0.16	0.18	0.17	0.098	0.22	0.18	0.21	0.071	0.2	0.15	0.24	0.17	0.19	0.23	0.24	0.15	0.23	0.17	0.27	0.19	0.24	ſ
	koo_md -	k00_mi -	- xm_00Å	koo_iq -	k01_md -	k01_mi -	k01_mx -	kol_iq -	k02_md -	k02_mi -	k02_mx -	k02_iq -	k03_md -	k03_mi -	k03_mx -	k03_iq -	k04_md -	k04_mi -	k04_mx -	k04_iq -	k05_md -	k05_mi -	k05_mx -	k05_iq -	k06_md -	k06_mi -	k06_mx -	k06_iq -	k07_md -	k07_mi -	- xm_70Å	k07_iq -	k08_md -	k08_mi -	k08_mx -	k08_iq -	- pm_60Å	k09_mi -	- xm_e0y	- pi_09	k10_md -	

۸lp	- 0.18	0.099	0.18	0.087	0.28	0.17	0.19	0.072	0.19	0.19	0.22	0.017	0.2	0.19	0.21	0.054	0.22	0.13	0.14	0.029	0.17	0.15	0.028	-0.015	0.22	0.13	0.014	0.054	0.2	0.083	0.057	0.052	0.19	0.19	0.028	-0.063	0.21	0.076	0.12	0.024	0.21
lsa	- 0.12	-0.053	0.064	0.078	0.32	0.14	0.1	0.2	0.11	-0.15	0.26	0.14	0.085	-0.11	0.13	0.076	0.04	-0.093	0.12	0.11	0.11	-0.14	-0.051	0.077	0.16	-0.059	-0.0047	0.077	0.15	-0.032	0.12	0.23	0.15	0.044	-0.013	0.13	0.076	-0.12	0.032	0.089	0.21
iel w2v	- 0.036	0.055	0.1	0.15	0.2	0.1	0.16	0.18	0.11	0.1	0.19	0.12	0.26	0.11	0.15	0.1	0.18	0.21	0.12	0.042	0.13	0.15	0.077	0.071	0.17	0.048	0.057	0.16	0.19	0.078	0.032	0.058	0.12	0.18	0.082	0.059	0.18	0.017	0.19	0.15	0.2
gpt	- 0.1	0.16	0.046	0.0078	0.11	0.14	0.012	-0.031	0.15	0.17	0.15	0.01	0.17	0.24	0.086	0.087	0.17	0.17	0.078	0.002	0.12	0.1	0.052	0.058	0.13	0.12	-0.015	0.03	0.13	0.12	0.073	0.024	0.15	0.15	0.051	-0.017	0.24	0.17	0.084	0.12	0.22
rbt	- 0.011	-0.063	0.087	0.06	-0.015	-0.052	0.2	0.092	0.0078	-0.033	0.22	0.15	-0.0067	-0.044	0.15	0.1 -	0.00038	3-0.032	0.042	0.093	-0.0078	-0.046	0.073	0.1	0.0029	-0.047	-0.0088	0.11	-0.021	-0.034	0.087	0.081	-0.021	-0.056	0.054	0.039	-0.013	-0.049	0.044	0.062	-0.035
ma	- 0.1	0.041	0.083	-0.12	0.088	0.15	0.049	-0.18	0.13	0.12	0.23	0.059	0.14	0.19	0.11	0.022	0.15	0.16	0.15	-0.04	0.075	0.12	0.066	0.0026	0.13	0.13	0.082	-0.03	0.1	0.11	0.12	0.1	0.11	0.17	0.15	-0.017	0.15	0.097	0.19	0.15	0.19
	- pu	'a	- xu	.' .'	- pu	, a	- xu	_id.	- pu	'a	- xu		- pu	'n.	- xu	-id	- pu	, a	- xu	- iq	- pu	'a	- xu	- iq	- pu	'a	- xu	_id	- pu	, E	- xu	- iq	- pu	Ē	- xu	.e	- pu	Ë.	- xu	- iq	- pu
	koo_r	koo	koo_r	koo	k01_r	k01_	k01_r	k01_	k02_r	k02_	k02_r	k02_	k03_r	k03_	k03_r	k03	k04_r	k04_	k04_r	k04_	k05_r	k05_	k05_r	k05_	k06_r	k06_	k06_m	k06_	k07_r	k07_	k07_r	k07_	k08_r	k08_	k08_r	k08_	k09_r	609_	r_004	609	k10_r

feature\_stat



### DONALD AND BARBARA ZUCKER SCHOOL of medicine AT HOFSTRA/NORTHWELL

March 2023, Pages S93–S103, https://doi.org/10.1093/schbul/sbac145