Evaluating the Validity of a Computational Linguistics-Derived Automated Health Literacy Measure across Race/Ethnicity

Dean Schillinger MD
R Balyan PhD
S Crossley PhD
D McNamara PhD
A Karter PhD

UCSF Health
Communications
Science Program

Employing Computational Linguistics to Improve Patient-Provider Secure email Exchanges (The ECLIPPSE Project)

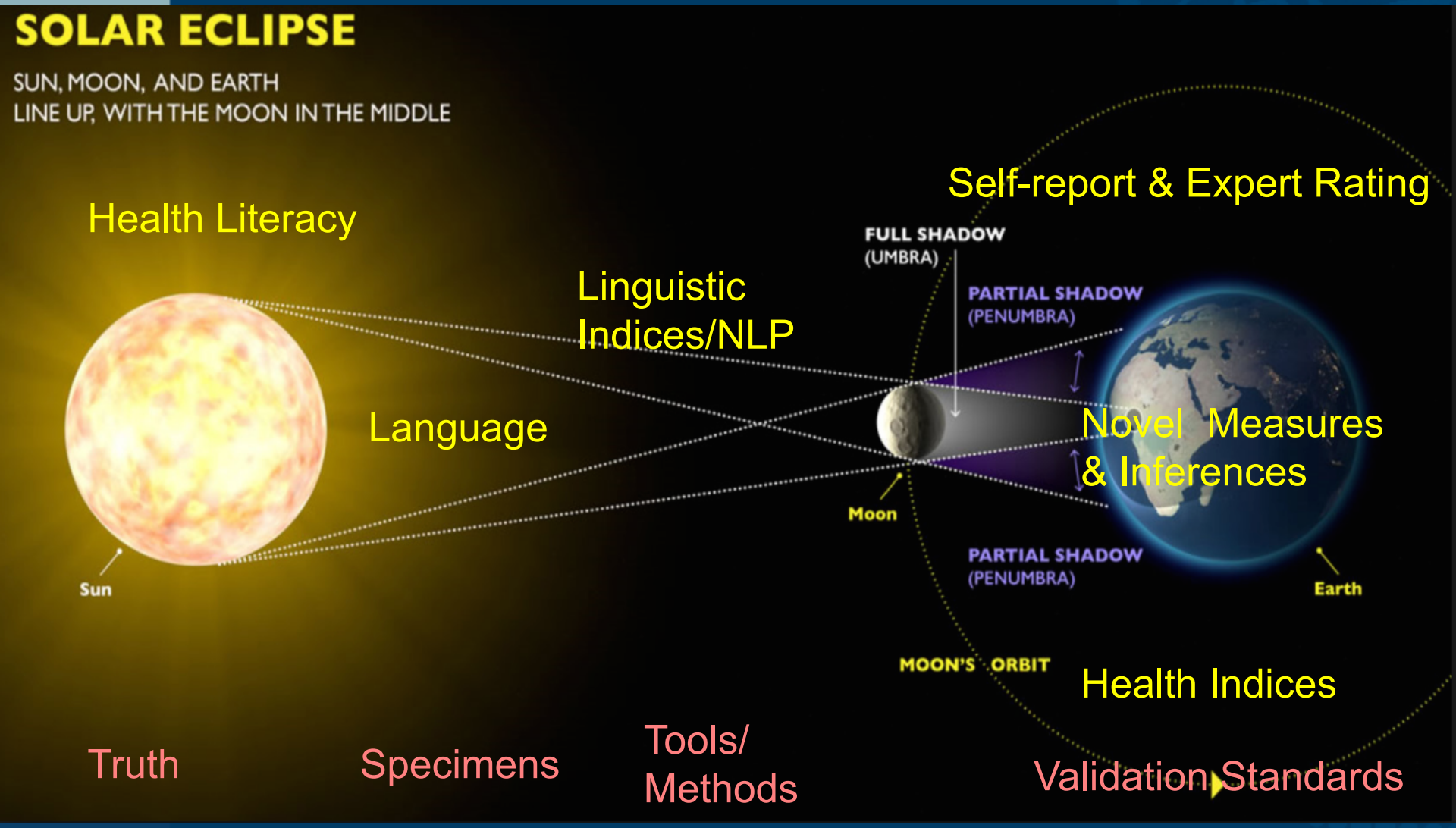
- Funded by NIH (National Library of Medicine, NLM012355)

- Multidisciplinary research collaboration between
  - UCSF
  - Kaiser Division of Research
  - Arizona State University
  - Georgia State University

Background

- Limited health literacy (HL) is associated with worse health and may partially mediate disparities
- HL interventions can improve outcomes and reduce disparities
- Measurement constraints have undermined the field and impeded scaling of interventions:
  - lack of an automated measure of HL so as to overcome administration challenges
  - inadequate ascertainment of validity of HL measures across racial/ethnic groups
- HL researchers need to be aware of the problem of cultural hegemony in literacy assessment, and downstream effects of mis-measurement

Eclipse as Metaphor for Scientific Inquiry



Health Literacy

Self-report & Expert Rating

Linguistic Indices/NLP

Language

Novel Measures & Inferences

Health Indices

Truth

Specimens

Tools/ Methods

Validation Standards

This Research is Challenging

- Health Literacy (HL) is a complex construct, and measurement is inherently challenging
- We are pioneering use of computational linguistics (natural language processing [NLP] and machine learning [ML]) to develop and validate a novel set of methods and measures
- We are harnessing "big data" at a very granular level from an electronic patient portal, requiring large amounts of data transfer, cleaning and confirmatory linguistics and health services research
- We are carrying out both "basic science" as well as translational research, requiring inter-disciplinary collaboration

Methods

- Analyzed language from >300,000 secure messages sent by 9,527 diabetes patients via a patient portal
- Sample: White-NH (N=2797), Black-NH(N=1409), Hispanic (N=1374), and Asian/Pacific Islander (N=2894), Mixed/Other (N=1053)
- Employed computational linguistics to develop a novel HL measure ("Literacy Profile", LP).
- This AI approach relied on gold standard of expert ratings of a purposive sub-sample of messages and applied machine learning to generate LPs on entire sample
- Stratified analyses by race/ethnicity to determine:
  - criterion validity (ROC curves/c-statistic)
  - predictive validity:
    - communication, adherence, glycemia, ED use

Generating Literacy Profiles

- To advance methods for identifying patients' HL in an automated fashion, we developed and compared five (5) automatically generated patient LPs based on distinct theoretical models and associated NLP tools and ML techniques.

Schillinger et al. Health Serv Res 2020

Examples of NLP Tools

- Tool for the Automatic Assessment of Lexical Sophistication (TAALES)
- Tool for the Automatic Analysis of Cohesion (TAACO)
- Tool for the Automatic Assessment of Syntactic Sophistication and Complexity (TAASSC)
- SEntiment ANalysis and Cognition Engine (SÉANCE)
- Writing Assessment Tool (WAT)
- NLP tools used a Stanford Parser, British National Corpus (BNC), MRC psycholinguistic database, CELEX word frequency database and Wordnet.
- We used medical corpora such as HIMERA, i2b2 to generate frequencies of all medical terms in these corpora

Balyan, Schillinger PLOS One 2019

Examples of the >200 Linguistic Indices Used

Linguistic Characteristic	Description
Concreteness	The degree to which a word is thought to be concrete
Imagability	How easy it is to construct the image of a word in one's mind
Familiarity	How familiar a word is thought to be to an adult
Meaningfulness	How strongly words are thought to associate with other words, and how likely words are thought to prime or activate other words
Age of Acquisition	The age at which a word is thought to first appears in a child's vocabulary
WordNet indices	The polysemy (number of word senses) and hypernymy (specificity) values of text using the WordNet database
Lexical Diversity	MTLD (measure of textual, lexical diversity; McCarthy, 2005) and D (Malvern et al., 2004) values.
Syntactic complexity	The mean number of words before the main verb and/or the mean number of higher level constituents per word
Readability Scores	Flesch Reading Ease score, Flesch Kincaid Grade level and Second Language Learners' readability scores.

Balyan, Schillinger PLOS One 2019

Linguistic indices used for Literacy Profiles

Literacy Profile	Linguistic Indices	Description
LP_FK	Readability	The length of words (i.e., number of letters or syllables) and length of sentences (i.e., number of words)
LP_LD	Lexical Diversity	The variety of words used in a text based on D
LP_WQ	Word Frequency	Frequency of word in a reference corpus
	Syntactic Complexity	Number of words before the main verb in a sentence
	Lexical Diversity	The variety of words used in a text based on MTLD
LP_SR	Concreteness	The degree to which a word is concrete
	Lexical diversity	The variety of words used in a text based on two measures of lexical diversity: MTLD, and D
	Present tense	Incidence of present tense
	Determiners	Incidence of determiners (e.g., a, the)
	Adjectives	Incidence of adjectives
	Function words	Incidence of function words such as prepositions, pronouns etc.
LP_Exp	Age of Exposure	The estimated age at which a word first appears in a child's vocabulary
	Lexical decision response time	The time it takes for a human to judge a string of characters as a word
	Attested lemmas	Number of attested lemmas used per verb argument construction
	Determiner per nominal phrase	Number of determiners in each noun phrase
	Dependents per nominal subject	Number of structural dependents for each subject in a noun phrase
	Number of associations	Number of words strongly associated with a single word

"Gold Standard": Expert-Rated HL

- We generated HL scores based on expert ratings of the quality of patients' SMs purposively sampled to represent a balance of self-reported HL, as well as a range of age, race/ethnicity, SES.
- A HL scoring rubric was used to holistically assess the perceived HL of the patients based on the linguistic output found in 512 SMs, adapting an established rubric used to score the quality and proficiency of a written essay
- Did SMs clearly convey health-related content and ideas the patient wanted to express to their physician? (IRR, r > .70)

Crossley, Schillinger. Health Comm 2020

Machine Learning Methods

- Analyses were conducted to develop LPs using several supervised ML algorithms.
- We trained Weka (version 3.8.1) and R (version 3.3.2) implementations for the ML models, including linear discriminant analysis (LDA), support vector machines (SVM), naïve Bayes, random forests, and artificial neural networks.
- These algorithms are some of the simplest and the most commonly used algorithms for classification problems.
- We used 10-fold cross validation approach on 70% of the data for fine-tuning the parameters and validation of the model. The performance of the model was tested and reported on the held-out 30% data.

Schillinger et al. Health Serv Res 2020

ROC curves (c-stat) and test characteristics of LP-Exp are similar across race/ethnicity

Race/ Ethnicity	C-stat (AUC)	Sens	Spec	PPV	NPV	Accuracy	Kappa
Total Sample	0.87						
White-NH	0.88	0.90	0.80	0.92	0.75	0.87	0.68
Black-NH	0.82	0.77	0.81	0.84	0.72	0.79	0.57
Hispanic	0.85	0.74	0.90	0.91	0.72	0.81	0.62
Asian/PI	0.89	0.84	0.92	0.96	0.69	0.86	0.69
Others	0.88	0.79	0.81	0.89	0.67	0.80	0.58

Predictive Validity of LP-Exp in Total Sample N= 9,527

Outcome	High LP (%)	Low LP (%)	p-value
Poor commxn (CAHPS item)	12.7	18.4	<.001
Poor med adherence	38.8	29.8	<.001
Severe hypoglycemia	3.5	4.5	0.024
Poor glycemic control (>9%)	15.0	19.6	<.001
ED visits/year	0.42	0.47	.016

Schillinger, Balyan et al. HSR 2020

Patterns of Concordance in Predictive Validity Fairly Consistent Across Race/Ethnicity

Outcome	White-NH	Black-NH	Hispanic	Asian/ PI	Mixed/ Other
Poor commxn	+	+	+*	+*	=
Poor med adherence	+	+*	+*	+	+
Severe hypo	+	+	+	+	+
Poor glycemia	+*	+	+*	+*	+
ED visits	+*	+**	=	=	+

* p<.05; **p<.10

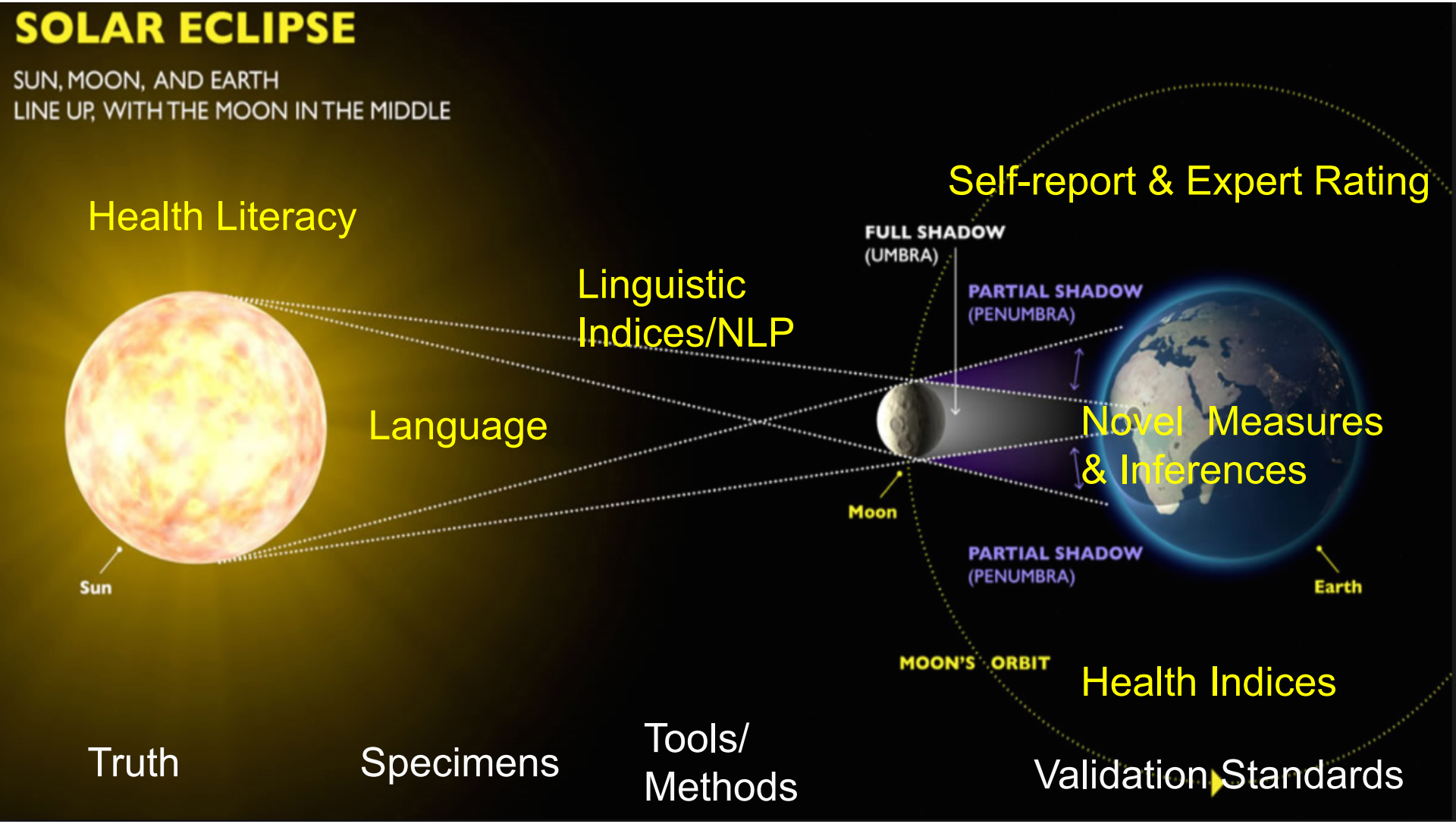
Conclusions

- We developed a novel measure of HL using advanced computational linguistic analyses of secure messages (Literacy Profile)
- We observed robust test characteristics of the LP with respect to the gold standard of expert-rated HL across race/ethnicity (criterion validity)
- We observed expected associations with health outcomes, in patterns that were largely consistent across race/ethnicity
- While concerns have arisen regarding bias in AI, automated Literacy Profiles appear sufficiently valid across race/ethnicity, enabling HL measurement at a scale that could improve clinical care and population health among diverse populations.

Collaborators/ECLIPPSE Team

- Northern CA Kaiser Division of Research
  - Andrew Karter PhD
  - Jennifer Liu PhD
  - Howard Moffet MPH
- Arizona State University
  - Renu Balyan PhD
  - Nicholas Duran PhD
  - Danielle McNamara PhD
- Georgia State University
  - Scott Crossely PhD
- UCSF Health Communications Research Program
  - William Brown PhD
  - Ryane Daniels MPH
  - Courtney Lyles PhD
  - Dean Schillinger MD
  - Wagahta Semere MD MAS

Thank You


SOLAR ECLIPSE
SUN, MOON, AND EARTH
LINE UP, WITH THE MOON IN THE MIDDLE
Health Literacy
Self-report & Expert Rating
FULL SHADOW
(UMBRA)
Linguistic
Indices/NLP
PARTIAL SHADOW
(PENUMBRA)
Language
Novel Measures
& Inferences
Moon
Sun
PARTIAL SHADOW
(PENUMBRA)
Earth
MOON'S ORBIT
Health Indices
Truth
Specimens
Tools/
Methods
Validation Standards

Supplemental Slides

Rubric for Expert Rating of HL of Secure Messages

- The rubric was used to holistically assess the perceived communicative HL of the patients based on the language they produced in their SMs (1-6 scale).
- Highest communicative HL (i.e., a score of 6) was defined as:
  - The patient demonstrates clear and consistent mastery of written English, although the writing may contain a few minor errors. The patient's writing is well organized and accurately focused providing clear access to the content of the message and the ideas that the patient wants to express. The writing demonstrates coherence and smooth progression of ideas, exhibits skillful use of language, using a varied, accurate, and apt vocabulary. The writing demonstrates meaningful variety in sentence structure and is free of most errors in grammar, usage, and mechanics.
- Lowest HL (i.e., a score of 1) was defined as:
  - The patient demonstrates very little or no mastery of written English and the writing is severely flawed by ONE OR MORE of the following weaknesses: disorganized or unfocused writing that results in disjointed or incoherent writing that do not provide access to the content of the messages and the ideas that the patient wants to express. The writing also displays fundamental errors in vocabulary, demonstrates severe flaws in sentence structure, and/or contains pervasive errors in grammar, usage, or mechanics that persistently interfere with meaning.
- Low HL was classified as a score of <4

Crossley, Schillinger. Health Comm 2020

Health Outcome	Health Literacy	White-NH	Black-NH	Hispanic	Asian/PI	Other
Poor Physician Communication (%)	Limited	10.5	14.1	24.4	26.7	13.9
	Adequate	8.4	11.5	16.8	17.2	13.6
	P-value	0.176	0.333	0.033	<0.001	1.00
Poor Medication Adherence (%)	Limited	29.7	51.9	59.7	29.7	41.7
	Adequate	26.1	37.2	36.4	25.5	37.8
	P-value	0.235	0.007	<0.001	0.126	0.550
≥1 Severe Hypoglycemia (%)	Limited	3.9	6.7	5.1	3.2	5.5
	Adequate	2.9	5.2	3.7	2.9	4.5
	P-value	0.261	0.330	0.299	0.709	0.598
Poor glycemic control (A1c ≥9%)	Limited	17.9	22.1	29.6	13.8	22.2
	Adequate	12.1	20.9	22.0	10.7	21.5
	P-value	0.002	0.732	0.032	0.047	0.905
ED use (mean, SD)	Limited	0.51 (1.20)	0.65 (1.48)	0.46 (1.11)	0.33 (0.80)	0.55 (1.28)
	Adequate	0.41 (1.01)	0.53 (1.10)	0.47 (1.07)	0.32 (0.86)	0.50 (1.16)
	P-value	0.029	0.096	0.801	0.790	0.471

NLP

- Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding.
- Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

Machine Learning

- “Machine Learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world.” – Nvidia
- “Machine learning is the science of getting computers to act without being explicitly programmed.” – Stanford
- “Machine learning is based on algorithms that can learn from data without relying on rules-based programming.”- McKinsey & Co.
- “Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.” – University of Washington
- “The field of Machine Learning seeks to answer the question “How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?” – Carnegie Mellon University