Oregon Health & Science University
School of Medicine

Scholarly Projects Final Report

Title *(Must match poster title; include key words in the title to improve electronic search capabilities.)*

Toward and Automated Method for Assessing Jargon in Medical Encounters

Student Investigator's Name

Katherine (Katie) King

Date of Submission *(mm/dd/yyyy)*

03/17/2023

Graduation Year

2023

Project Course *(Indicate whether the project was conducted in the Scholarly Projects Curriculum; Physician Scientist Experience; Combined Degree Program [MD/MPH, MD/PhD]; or other course.)*

Scholarly Projects Curriculum

Co-Investigators *(Names, departments; institution if not OHSU)*

Cliff Coleman MD MPH, Department of Family Medicine / Center for Ethics in Health Care

Mentor's Name

Cliff Coleman, MD, MPH

Mentor's Department

Department of Family Medicine, Center for Ethics in Health Care

Scholarly Project Final Report

Concentration Lead's Name

David Buckley, MD, MPH

Project/Research Question

Can automated transcription and medical jargon detection software be used to transcribe and identify medical jargon during clinical encounters?

Type of Project *(Best description of your project; e.g., research study, quality improvement project, engineering project, etc.)*

Research study

Key words *(4-10 words describing key aspects of your project)*

Medical jargon, health literacy, medical education, automation

Meeting Presentations

If your project was presented at a meeting besides the OHSU Capstone, please provide the meeting(s) name, location, date, and presentation format below (poster vs. podium presentation or other).

Poster accepted to 22nd Annual Institute for Healthcare Advancement Health Literacy Conference (virtual), May 9-11, 2023

Publications *(Abstract, article, other)*

If your project was published, please provide reference(s) below in JAMA style.

N/A

Submission to Archive

Final reports will be archived in a central library to benefit other students and colleagues. Describe any restrictions below (e.g., hold until publication of article on a specific date).

N/A

Scholarly Project Final Report

Next Steps
What are possible next steps that would build upon the results of this project? Could any data or tools resulting from the project have the potential to be used to answer new research questions by future medical students?

Use methodology established by this study to compare jargon use in clinical encounters between cohorts of students completing different clinical communication curriculums, or to track jargon use in individuals over time. Curate dictionary of medical jargon to improve automated jargon detection.

Please follow the link below and complete the archival process for your Project in addition to submitting your final report.

https://ohsu.ca1.qualtrics.com/jfe/form/SV_3ls2z8V0goKiHZP

Student's Signature/Date *(Electronic signatures on this form are acceptable.)*
This report describes work that I conducted in the Scholarly Projects Curriculum or alternative academic program at the OHSU School of Medicine. By typing my signature below, I attest to its authenticity and originality and agree to submit it to the Archive.

Katherine Alexis King / 17 March 2023

Mentor's Approval *(Signature/date)*

Scholarly Project Final Report

Report: *Information in the report should be consistent with the poster, but could include additional material. Insert text in the following sections targeting 1500-3000 words overall; include key figures and tables. Use Calibri 11-point font, single spaced and 1-inch margin; follow JAMA style conventions as detailed in the full instructions.*

Introduction *(≥250 words)*

Health literacy is defined as "one's ability to obtain, process, communicate and understand basic health information needed to make appropriate health decisions."[1] In the United States, less than 15% of adults have proficient health literacy.[2] Health literacy impacts patient care[3]--those with higher health literacy are more likely to adhere to treatment recommendations,[4] while those with lower health literacy are more likely to experience poor outcomes and decreased connection to health resources.[5] Fortunately, provider-based interventions and attention allow barriers related to low health literacy to be overcome,[4,5,6] and health literacy and clear communication best practices for providers have emerged, including the use of plain non-medical language.[7]

Physicians have been documented utilizing language that is not understandable to patients, leading to poor patient comprehension.[8] In one study, jargon terms were identified in 81% of encounters with an average of four jargon terms per encounter, often at critical points such as providing education or recommendation.[8] The frequent use of jargon and its potential to limit understanding and adversely impact patient care has made medical jargon spoken with patients a key concern for health professions educators, health system managers, and administrators. Unfortunately, assessment of health professionals' jargon use in spoken communication as well as the efficacy of interventions to reduce jargon use by health professionals in clinical encounters is limited by current jargon identification methods which are subjective, labor-intensive, and expensive. Current jargon detection methods involve one or more observers manually coding language felt to represent jargon[8,9] without validated tools allowing for easy reproducibility, reliability or validity across cohorts, disciplines, observers, or institutions.

The goal of this study is to validate a relatively fast, inexpensive, low-tech, automated, and widely available method to identify medical jargon in clinical encounters. We hypothesized that video-recorded clinical encounters can be converted into written transcripts using free software from YouTube.com and analyzed for medical jargon using Health Literacy Advisor Online™ (HLA) proprietary software.

Methods *(≥250 words)*

Eight randomly selected video-recorded Observed Structured Clinical Examinations (OSCEs) of medical students during their first clinical rotation were matched by case topic to eight OSCEs of similar students who completed a redesigned preclinical communication skills curriculum. Each OSCE was transcribed twice; once manually by a professional transcriptionist, and once by YouTube software. Both the manually transcribed and YouTube-transcribed OSCEs were anonymized by a member of the research team not involved in assessing jargon content and edited to produce two versions: one version containing both patient and student utterances during the encounter portion of the OSCE, and one version containing only the student utterances during the patient encounter portion of the OSCE. All transcripts were analyzed for jargon in two ways: once manually and once by the "health words" filter of a medical jargon detection software called Health Literacy Advisor™ (HLA). Manual jargon counting was completed by two researchers who were blinded to which cohort the transcript had come from and who independently counted jargon

using the technical terminology, medical vernacular, unnecessary synonyms, and acronyms/abbreviations ("alphabet soup") categories from a previously published taxonomy.[9,10] The two researchers then compared their jargon assessments and resolved discrepancies by consensus. At this point, researchers were unblinded to the transcript source to review videos and correct for any transcription errors on the manual transcripts. Words or phrases that could not be deciphered were recorded as "inaudible" and a second round of consensus discussion was completed to address new jargon terms that arose from the process of correcting errors in transcription. Jargon counts were then recorded and averaged for transcripts including both student and patient utterances during the clinical encounter, transcripts with only student utterances during the encounter, and for discrete jargon words spoken by the student during the encounter (e.g. emphysema said six times by the student is counted once). Jargon counts were compared in Excel with two-tailed t-tests assuming equal variance at the 0.05 α level.

The sensitivity and specificity of HLA "health words" filter alone as well as the HLA "health words" filter combined with the "non-health words" filter on transcripts containing only the student words spoken during the encounter were calculated using the manually transcribed and manually counted student-only transcripts as the gold standard. False positives represented words that HLA flagged as jargon but were not counted as jargon by manual counters. False negatives represented words counted as jargon by manual counters but missed by HLA. True positives represented words that both HLA and manual counters recognized as jargon. True negatives represented words that both HLA and manual counters recognized as non-jargon. The true negative count for each OSCE was derived by subtracting the sum of true positives, false positives, false negatives, and compound jargon term correction factor from the total word count of the YouTube transcript containing only the student words spoken during the encounter. Since HLA ascribes a value of one to all jargon terms (regardless of how many words are in the term), the compound jargon term correction factor was needed to account for jargon terms involving more than one word. This correction factor was determined by identifying jargon terms longer than one word and summing any additional words. For example, the correction factor for the phrase 'squamous cell carcinoma' was two. This ensured that all jargon words, including those that were part of larger phrases, were excluded from the true negative count.

Additionally, number of cases completed by men and women (binary gender assumed by researcher), mean duration of encounter, mean word count of encounter (student and patient utterances), and mean student-only word count during an encounter were compared by cohort (students completing old clinical skills curriculum vs redesigned clinical communication skills curriculum) in Excel with two-tailed t-tests assuming equal variance at the 0.05 α level. Word counts were also compared in Excel by type of transcription (manual vs YouTube) with a two-tailed t-test assuming equal variance at the 0.05 α level for both transcripts containing student and patient utterances during the clinical encounter as well as transcripts containing only student utterances during the clinical encounter.

The sample size for this study was derived from an *a priori* decision rule by Castro and colleagues[8] using the results of a power analysis for a study designed to detect differences in jargon use amongst two cohorts of learners. This study was thus not powered to detect differences in jargon use by exposure to redesigned clinical communication skills curriculum, but rather to validate the proposed methodology for automated jargon detection to be used in a future study comparing jargon use in the cohorts of learners. However, using a manually transcribed and counted OSCE which only counted discrete jargon terms that were used by the student during the patient encounter, not introduced by the patient, and not defined by the student at the student's first use of the term, average jargon terms were compared by cohort with a two-tailed t-test assuming equal variance at the 0.05 α level in Excel. This study was also not specifically powered to detect differences in jargon use by gender. However, binary gender of participants assumed by researcher

was recorded and a two-tailed t-test assuming equal variance at the 0.05 α level in Excel was used to evaluate for differences in jargon use by gender using a manually transcribed and counted OSCE which only counted discrete jargon terms that were used by the student during the patient encounter, not introduced by the patient, and not defined by the student at the student's first use of the term. For these by cohort and by gender analyses, terms were counted as defined if the student made any attempt to explain the term in another way or utilized gestures/movements to explain the word. This version of counting jargon by tracking if the jargon was defined and/or introduced by the patient reflects what has been previously done in the field but was not utilized throughout the study given its goal to establish a methodology that is automated (i.e. does not require transcript review to determine who introduced terms and if they were defined).

Results *(≥500 words)*

Of the eight cases selected from the cohort of students completing the old clinical communication skills curriculum, 50% were completed by men and 50% were completed by women. Of the eight cases selected from the cohort of students completing the redesigned clinical communication skills curriculum, 40% were completed by men and 60% were completed by women. There were no differences in number of cases completing by binary gender between cohorts ($p = 0.95$).

The mean duration of the patient encounter for the cohort of students completing the old clinical communication skills curriculum was 11.1 minutes. The mean duration of the patient encounter for the cohort of students completing the redesigned clinical communication skills curriculum was 10.3 minutes. There was no difference in the mean duration of the patient encounter between cohorts ($p = 0.053$).

The mean word count of the patient encounter for the cohort of students completing the old clinical communication skills curriculum was 1851.0. The mean word count of the patient encounter for the cohort of students completing the redesigned clinical communication skills curriculum was 1870.8. There was no difference in the mean word count of the patient encounters between cohorts ($p = 0.86$).

The mean student-only word count during the patient encounter for the cohort of students completing the old clinical communication skills curriculum was 1344.4. The mean student-only word count during the patient encounter for the cohort of students completing the redesigned clinical communication skills curriculum was 1220.3. There was no difference in the mean student-only word count during the patient encounter between cohorts ($p = 0.13$).

The mean number of discrete jargon terms not defined by the student at first use or introduced by the patient for the cohort of students completing the old clinical communication skills curriculum was 20.1. The mean number of discrete jargon terms not defined by the student at first use or introduced by the patient for the cohort of students completing the redesigned clinical communication skills curriculum was 19.9. There was no difference in the mean number of discrete jargon terms not defined by the student at first use or introduced by the patient between cohorts ($p = 0.96$).

The mean number of discrete jargon terms not defined by the student at first use or introduced by the patient for woman participants was 16.5. The mean number of discrete jargon terms not defined by the student at first use or introduced by the patient for woman participants was 25.8. There was no difference in the mean number of discrete jargon terms not defined by the student at first use or introduced by the student between genders ($p = 0.06$).

YouTube transcribed fewer patient and student words compared to manual transcription (p = 0.029). However, when transcribing only student words, YouTube and manual transcriptions do not differ in their word count (p = 0.25) (Table 1).

Table 1. Word Count by Manually and Automatically Transcribed Medical Encounters.

	Manual Transcription	YouTube Transcription	p
Mean Encounter (Patient and Student) Word Count	1860	1691	0.029*
Mean Student-Only Word Count	1280	1215	0.25

Less than half as many jargon words are detected when using "health words" automated jargon detection software compared to manually counting (p < 0.05) across all transcript versions for both manually and YouTube transcribed encounters. Manual and YouTube transcriptions do not differ across any transcript versions in number of jargon terms detected by the HLA "health words" filter (Figure 1).

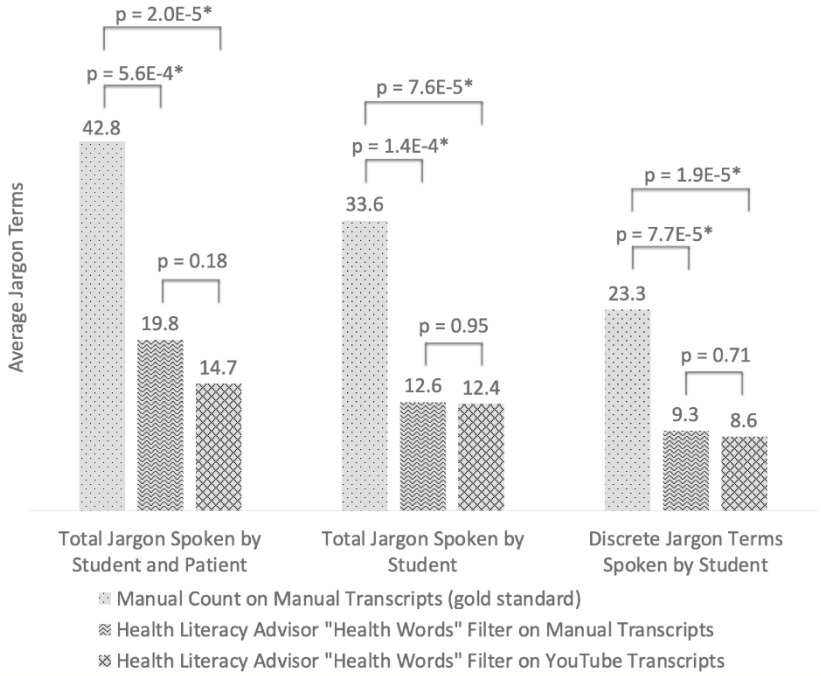

Figure 1. Jargon Detection in Medical Encounters by Manual and Automated Methods.

The specificity of the HLA "health words" filter is 99.7% for both manual and YouTube transcriptions. The sensitivity of the HLA "health words" filter is 45.5% for manual transcriptions and 25.8% for YouTube transcriptions. The specificity of the HLA "Health Words" filter combined with the "non-health words" filter is 95.9% for manual transcriptions, 96.2% for YouTube transcriptions. The sensitivity of the HLA "health

words" filter combined with the "non-health words" filter is 75.8% for manual transcriptions, 53.1% for YouTube transcriptions (Table 2).

Table 2. Sensitivity and Specificity of Health Literacy Advisor™ on Manual and Automated Transcripts.

	Manual Count on Manual Transcripts (gold standard)	Health Literacy Advisor™ on Manual Transcripts		Health Literacy Advisor™ on YouTube Transcripts	
		"Health Words" Filter	"Non-Health Words" Filter	"Health Words" Filter	"Non-Health Words" Filter
Sensitivity	100%	45.8%	75.8%	25.8%	53.1%
Specificity	100%	99.7%	95.9%	99.7%	96.2%

Discussion *(≥500 words)*

The HLA "health words" filter identifies half as many jargon terms compared to manual counting and produces the same jargon counts ($p > 0.05$) on YouTube and manual transcripts (Figure 1). The HLA "health words" filter has a high specificity on both manual and YouTube transcriptions (Table 2). However, the HLA "health words" sensitivity is low for both YouTube and manual transcriptions, with the sensitivity on YouTube transcriptions 20% worse than manual transcriptions (Table 2). The sensitivity of HLA improves when the "non-health words" filter is added, though a 22.7% difference favoring manual over YouTube transcriptions remains (Table 2). There is little loss in specificity (under 4%) for both manual and YouTube transcriptions when the HLA "non-health words" is added (Table 2).

The use of automated transcription plus jargon detection software has the potential to provide a low-cost, easy-to-use, and accessible methodology to evaluate for differences in medical jargon use during clinical encounters. The removal of cost and time barriers will allow for further clinical communication research to improve patient care and outcomes. The data presented here suggest that when analyzing a clinical encounter for instances of jargon with the HLA "health words" filter, YouTube and manual transcriptions can be used interchangeably (Figure 1). The ability to transcribe clinical encounters via software offers a time- and money-saving alternative to manual transcription when investigating the jargon frequency in clinical encounters. While manual and YouTube transcription appear equal when assessing jargon count with the HLA "health words" filter, the twenty percent difference in sensitivity between the two modes of transcription suggests that manual transcription is superior to YouTube transcription if using HLA to explore transcript content or what jargon words are being used. However, the low sensitivity of the HLA "health words" filter, limits its use to scenarios where rough estimations of the frequency of jargon use or a high false negative rate are acceptable. The improved sensitivity with minimal compromise of specificity when adding the "non-health words" filter to the "health words" filter, offers a promising methodology to potentially allow for greater utility of HLA to detect medical jargon. However, further work is needed to validate this method, including the expansion of the Figure 1 analyses to include data derived from HLA's "health words" and "non-health words" filters when used together.

This study has multiple limitations. First, the quality of automated transcription is dependent on audio quality and conversation characteristics. As seen in Table 1, YouTube transcription captures fewer words than manual transcription for encounters containing both patient and student utterances. However, this difference is not seen when comparing YouTube and manual transcriptions for encounters containing only student utterances. This indicates that for the encounters used in this study, YouTube had difficulty transcribing patient utterances, potentially due to microphone placement (i.e. microphones recording the

encounter were further from the patient, making patients more difficult to hear such that YouTube software did not recognize their speech). The impact of audio quality on transcription is important to note because our results may not be reproduced in files with audio quality that differs from the videos used in this study. Importantly, manual transcription is also susceptible to errors related to audio quality and conversation characteristics. Portions of manual transcripts in this study in were labeled "inaudible," often due to low audio volume or crosstalk amongst participants. This work is also limited in its ability to be used in real-word clinical settings. YouTube is not HIPAA compliant. Thus, the automated transcription method described in this study cannot be used in encounters with real (non-standardized) patients without permission. A third limitation of this study is inaccuracies in the data inherent to a manual counting process. The process of identifying jargon manually was guided by a previously published taxonomy but remained subjective, making reproducibility unlikely. Furthermore, the process of manual counting thousands of words is prone to simple counting errors. This both creates noise in the data while simultaneously highlighting the need for an objective, standardized, and automated jargon detection process. Next steps toward this study's goal of validating such a process for use in clinical communication research and education include additional analyses using HLA's "health words" and "non-health words" filters together, using YouTube transcription and HLA to evaluate for differences in jargon use amongst learners completing different clinical communication curriculums, and curating a dictionary of health jargon terms to improve the sensitivity (and consequently utility) of jargon detection by software.

Conclusions *(2-3 summary sentences)*

There is no difference in the number of jargon terms counted by the HLA "health words" filter when comparing YouTube and manual transcriptions. However, the HLA "health words" filter provides an underestimate of total jargon use and has low sensitivity which limits its utility. The improved sensitivity seen when using the HLA "health words" and "non-health words" filters together requires further investigation but may offer a superior automated jargon detection method than use of the HLA "health words" filter alone.

References *(JAMA style format)*

1. Somers SA, Mahadevan R. Health literacy implications of the Affordable Care Act. Center for Health Care Strategies, Inc, November 2010. http://www.mffh.org/mm/files/Summit_ Mahadevan_handout.pdf. Accessed September 26, 2014. Quoted by: Coleman CA. Introduction to Health Literacy and Clear Communication. *OHSU Student Guide*. 2019.
2. Kutner M, Greenberg E, Jin Y, Paulsen C. The Health Literacy of America's Adults: Results From the 2003 National Assessment of Adult Literacy (NCES 2006–483). U.S. Department of Education. Washington, DC: National Center for Education Statistics.
3. Kilfoyle KA, Vitko M, O'Conor R, Bailey SC. Health Literacy and Women's Reproductive Health: A Systematic Review. *J Womens Health (Larchmt)*. 2016;25(12):1237-1255. doi:10.1089/jwh.2016.5810.
4. Miller TA. Health literacy and adherence to medical treatment in chronic and acute illness: A meta-analysis. *Patient Educ Couns*. 2016;99(7):1079-1086. doi:10.1016/j.pec.2016.01.020.
5. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. *Ann Intern Med*. 2011;155(2):97-107. doi:10.7326/0003-4819-155-2-201107190-00005.
6. Barrett SE, Puryear JS, Westpheling K. Health literacy practices in primary care settings: Examples from the field. The Commonwealth Fund; 2008. https://www.commonwealthfund.org/publications/fund-reports/2008/jan/health-literacy-practices-primary-care-settings-examples-field?redirect_source=/publications/fund-reports/2008/jan/health-literacy-practices-in-primary-care-settings--examples-from-the-field.

7. Coleman C, Hudson S, Pederson B. Prioritized Health Literacy and Clear Communication Practices For Health Care Professionals. *Health Lit Res Pract*. 2017;1(3):e91-e99. Published 2017 Jul 10. doi:10.3928/24748307-20170503-01.
8. Castro CM, Wilson C, Wang F, Schillinger D. Babel babble: physicians' use of unclarified medical jargon with patients. *Am J Health Behav*. 2007;31 Suppl 1:S85-S95. doi:10.5555/ajhb.2007.31.supp.S85.
9. Miller AN, Bharathan A, Duvuuri VNS, et al. Use of seven types of medical jargon by male and female primary care providers at a university health center. *Patient Educ Couns*. 2022;105(5):1261-1267. doi:10.1016/j.pec.2021.08.018.
10. Pitt MB, Hendrickson MA. Eradicating Jargon-Oblivion-A Proposed Classification System of Medical Jargon. *J Gen Intern Med*. 2020;35(6):1861-1864. doi:10.1007/s11606-019-05526-1.