

Ask Your Doctor to Prescribe a YouTube Video: Health Information Seeking and Algorithmic Artifacts on Digital Platforms

Xiao Liu, Anjana Susarla, Rema Padman
(xiao.liu.10@asu.edu, asusarla@msu.edu, rpadman@cmu.edu)

November 2021

Health information seeking behavior (HISB) refers to the ways in which individuals seek information about their health, risks, illnesses, and health-protective behaviors. In conceptualizing such artifacts to the context of large social media platforms, we also need computational methods to address the complex and very large-scale data generated by technological capabilities, such as multi-media rich videos, and human behaviors, such as viewership and engagement. We integrate multiple theories and domains of inquiry such as user acceptance and self-efficacy research in the information systems (IS) domain and literature on health education and information retrieval to posit two novel algorithmic artefacts that encapsulate HISB: understandability and encoded information. We draw on the Patient Education Material Assessment Tool (PEMAT), a systematic approach for audio-visual educational materials assessment, to develop a method to assess understandability of videos from a patient education perspective. Extracting video features and metadata from YouTube, we develop a human in the loop algorithmic assessment combining PEMAT-based patient education constructs, annotations from domain experts and co-training methods from machine learning to assess the understandability of diabetes videos. We further examine the impact of understandability on several dimensions of collective engagement with videos. A challenge in evaluating collective engagement with understandable videos is that there could be content that is not medically validated but engages users. We synthesize multiple machine learning methods to design a causal framework that permits us to first understand the process of information seeking behavior and then employ the methods of causal inference to understand the impact of algorithmic artefacts such as understandability and encoded medical information on collective engagement. We consider the simultaneous impact of understandability and encoded medical information in a video on collective engagement by conducting a multiple-treatment propensity score based matching approach that allows us to implement a quasi-randomization research design. The availability of digital trace data collected from social media usage and the ubiquity of information search mediated by algorithms necessitates a better understanding of digital and algorithmic artefacts such as those highlighted in the current study. Implications for research and practice are discussed.

Acknowledgements:

The authors thank participants at the AIDR Workshop, AMIA, Conference on Machine Learning, Optimization, and Data Science, Informs Data Science Workshop, VIDE Seminar Series, Neurips MLPH, WITS, WISE and SCECR Conference for comments on earlier drafts of the paper, and seminar participants at Boston University, McGill University, MSU Outreach, University of Maryland, Temple University, Texas A&M University and University of Illinois at Chicago for comments on earlier drafts of the paper. We thank Ernestina Bioh, Sreeja Nair, Mukund Nakhate and Namrata Navge and for their research assistance. We also acknowledge funding from National Library of Medicine (NIH Grant R01LM013443).

1. Introduction

Digital traces from social media have yielded a rich understanding of societal outcomes ranging from health, education, and governance issues (Liu et al. 2020, Pew 2021, Stier et al. 2020, Van den Beem et al. 2020). Corporate messaging on social media likewise emphasizes broader social outcomes such as equity and fairness (Bonaparte 2020). Yet research in business disciplines has largely ignored how individuals engage with social media for non-monetary purposes. Motives for participation on social media platforms are substantially different from motives such as technology acceptance (Venkatesh 2000) and internet self-efficacy posited by prior research in the information systems (IS) discipline. These studies considered information systems usage within an organization or personal use of information technology for a limited range of tasks. Individuals' increased dependence on social media and digital artefacts for so many aspects of life, heightened during the pandemic, necessitates an inquiry into the context of information seeking. This is the first gap in the literature we seek to uncover. Another gap we explore is there is limited inquiry into what makes the information obtained from social media relevant and usable for our daily lives. Aside from opinion formation (Munger 2019, Munzert et al. 2020), literature has not examined the why, or the range of motives characterizing participation and information seeking on digital platforms. The availability of digital trace data, or rich multi-media data collected from social media usage (Freelon 2014, Munzert et al. 2020, Stier et al. 2020) and the ubiquity of information search mediated by algorithms (Lazer 2015, Lazer et al. 2020) necessitates a better understanding of digital and algorithmic artefacts.

Studies have estimated that only 12 percent of US adults have proficient health literacy, according to the National Assessment of Adult Literacy (Kutner et al. 2006), while more than 80 million have low health literacy and lack the ability to understand basic medical information to engage in effective self-care and chronic disease management. Such limited health literacy results in poorer health outcomes and higher healthcare costs (Adams 2010). Patients traditionally received health directives from clinicians through verbal advice and printed pamphlets. However, for patients to benefit from such educational materials requires a high level of participation and engagement (Jordan et al. 2008). Most education materials are too complex for patients to comprehend, particularly those with low health literacy (Johnson et al. 2020,

Rooney et al. 2020). Platforms such as YouTube potentially offer a novel pathway to enhance patient education (Liu et al. 2020). Healthcare information in video format disseminated through social media could transform health communication and patient education (O’Neill et al. 2014), improving patient-physician interactions and patient self-care (Adams 2010). Health information seeking behavior (HISB) refers to the ways in which individuals seek information about their health, risks, illnesses, and health-protective behaviors (Lambert and Loiselle, 2007; Mills and Todorova, 2016). Since users are heterogeneous in their health information needs as well as in their levels of health literacy, we integrate multiple theories and domains of inquiry such as user acceptance and self-efficacy research in IS and literature on health education and information retrieval to posit two novel algorithmic artefacts that encapsulate HISB: understandability and encoded information.

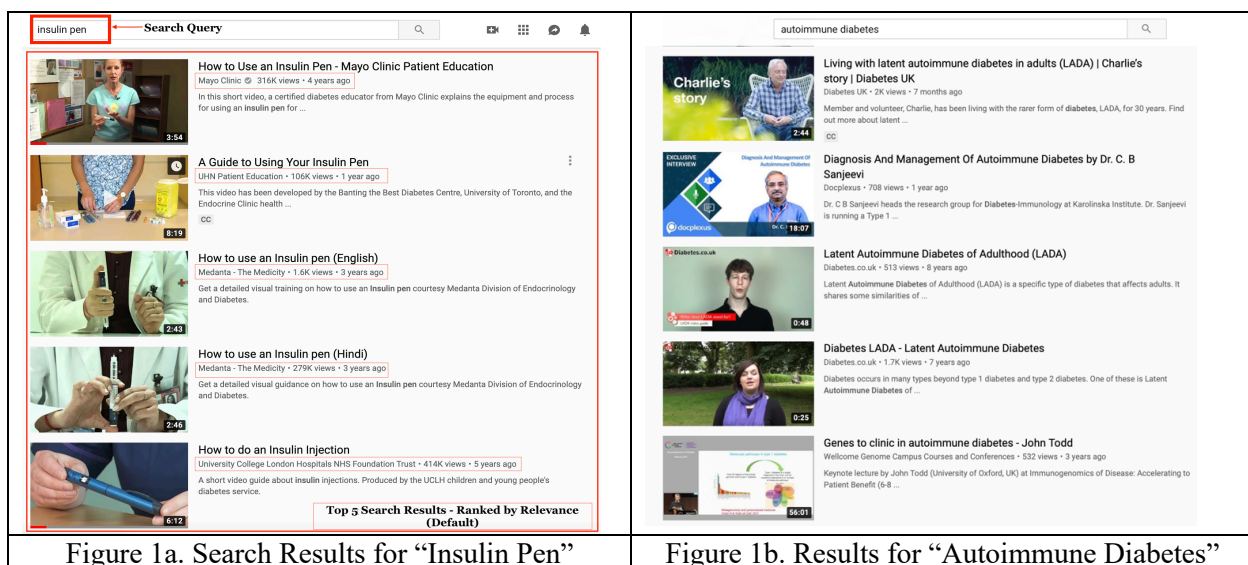
Most existing work in health communication and patient education have evaluated videos manually which is not scalable, replicable, or efficient given the sheer volume and variety of content created, curated and favorited on social media (Gillespie 2018, Gillespie 2020). We develop a scalable, human-in-the-loop automated method based on the Patient Education Material Assessment Tool (PEMAT), a systematic approach developed by the Agency of Healthcare Research and Quality (AHRQ), to assess understandability. Our algorithmic artifact of medical information encoded in the videos builds on an exhaustive lexicon of medical terms (e.g., Liu et al. 2020). The collective engagement with videos serves as a proxy for popularity, relevance of content to users and collective attention attracted by the videos (e.g., Liu et al. 2020). We examine the impact of encoded medical information and understandability of a video on three dimensions of collective engagement: disengagement, sustained attention driven engagement, and selective attention driven engagement. To address causal identification challenges, we conduct a multiple-treatment propensity score based matching approach that allows us to implement a quasi-randomization research design. This is the first study that looks at algorithmic artifacts that capture HISB and collective attention with digital traces from viewers of healthcare videos.

We make three contributions to the literature. Our first contribution is to posit understandability of healthcare information as a theoretical and algorithmic artefact that bridges a gap between access to

information and its understandability (Alpay et al. 2009). Our second contribution is the human-in-the-loop augmented intelligence method to characterize understandability. In characterizing search for cognitively demanding information on social media, we combine human cognitive capabilities into artificial intelligence (AI) systems by developing an augmented intelligence approach. A third contribution is to translate intentions to usage by looking at the link between information seeking behavior and collective engagement of viewers. We synthesize multiple machine learning methods to design a causal framework that permits us to first understand the process of information seeking behavior and then employ the methods of causal inference to understand the impact of algorithmic artefacts such as understandability and encoded information on collective engagement.

2. Literature Review and Research Questions

2.1 Health Information Seeking Behavior



Users typically encounter videos on healthcare conditions through keyword searches on YouTube. Figure 1 below shows the search results for “Insulin Pen”. YouTube provides links to authenticated medical information, such as the top ranked results in Figure 1a, several of which are produced by reputable health organizations. However, such links are available only for a handful of highly visible and popular topics but not for most healthcare related queries on social media, as Figure 1b illustrates. It is prohibitively expensive for healthcare organizations to create medically validated and understandable video content encompassing

the entire range of symptoms, treatment, and disease progression. It is an equally daunting challenge for both patients and clinicians to search for appropriate videos, retrieve them for each care delivery context, and use them in the form of a just-in-time, contextualized, prescriptive digital therapeutic intervention.

For patients engaging in searches for health information on digital media platforms, the gap in health literacy is exacerbated both by their own lack of knowledge and by algorithmic recommendations on digital platforms. Studies suggest that one in three US adults use the internet to diagnose or learn about a health concern (Fox and Duggan, 2013). Such extensive search for health information through social media platforms could exacerbate the disparities in health information availability and use (Percheski and Hargittai, 2011). While there is educational value from healthcare information in video format (Wood et al. 2017), it is a daunting challenge for both patients and clinicians to search for appropriate videos, retrieve them for each care delivery context, and use them in the form of a contextualized digital therapeutic intervention. Limited research exists on developing automated retrieval techniques to find helpful, validated health information on YouTube. Our research aims to address that gap. We employ a human-in-the-loop algorithmic approach to examine the vast corpus of video content created by individuals including laypeople, and organizations such as health care providers, public health agencies and professional medical societies.

2.2 Algorithmic Artefacts Underlying HISB

Information seeking through technological artefacts, whether on social media, or in earlier generations of computing, is conceptualized to start with the intention of information seeking and can result in various actions or utilities. A considerable literature in IS has examined self-efficacy, including computer self-efficacy and internet self-efficacy and technology acceptance, as motives for information seeking and posited various impacts on the resultant information usage. Venkatesh et al. (2003) posit a unified theory of acceptance and use of information technology. Agarwal and Karahanna (2000) posit instrumental beliefs drive individual usage intention. Scholars of information quality highlight contextuality of information retrieved, but from a service quality perspective (Kettinger and Lee 1997). While people have relied on traditional media (such as library, books, brochure, magazines) or healthcare professionals as their primary

source of health information (Baker et al. 2003; Dolan et al. 2004; Dutta-Bergman 2004), the advent of the Internet and social media has witnessed a sea change in HISB. Prior research on HISB has proposed constructs such as contextual distinctiveness (McDonald and Shillcock 2001), semantic diversity (Landauer and Dumais 1997) and contextual diversity (Adelman et al. 2006) in HISB and characterized emotional support, information seeking about health needs (such as disease prevention, treatment of specific conditions and management of various health conditions and chronic illnesses) and patient education as different motives for HISB. However, these constructs are not tailored to the sheer complexity of HISB in very large-scale social media platforms. In conceptualizing such artifacts to the context of large social media platforms, we also need computational methods to address the complex and very large-scale data generated by technological capabilities, such as multi-media rich videos, and human behaviors, such as viewership and engagement (e.g., Lazer et al. 2020). Table 1 highlights the gaps in the literature.

Table 1. Prior Streams of Literature

Prior Streams of Literature	Intention in information seeking	Dimensions Studied	Link to understanding usage
Technology Acceptance	Theory of Planned Behavior	Ease of Use and Usefulness	Does not address the diversity of usage motives or context
Information Quality	Dimensions of quality such as contextuality	Service Quality	Does not incorporate semantic or video understanding
Cognitive Absorption	Social cognitive theory	Instrumental beliefs drive usage	Investigates antecedents of absorption
Health Information Seeking	Contextual distinctiveness and diversity	Does not examine usage	Does not connect intentions of usage to actual usage
Current Study	Health information seeking	Encoded medical information and its understandability	Impact on collective engagement (contribution of current study)

HISB on social media platforms changes the direction of information seeking from top-down medical directives to a multi-faceted curation of information from experts other than physicians, such as nutritionists and patient educators and peer groups of patients (Madathil et al. 2015). Second, another characteristic of social media platforms is that users are heterogenous in their health information needs and in their levels of health literacy. We therefore conceptualize two distinct facets of HISB that can be constructed as algorithmic artefacts: understandability and encoded medical information. Figure 2 presents a conceptual model and serves as a roadmap for hypotheses development. HISB on social media platforms is shaped by the sociomaterial context (Orlikowski 2010) of the interplay between digital patterns and

human behavior that influences both facets of HISB and the relationship between HISB and usage of HISB. We therefore consider the collective engagement of users, which is another algorithmic artefact that was not studied in prior generations of technology. We consider the link from intention to actual usage by considering the link between dimensions of HISB artifact and their collective engagement on YouTube.

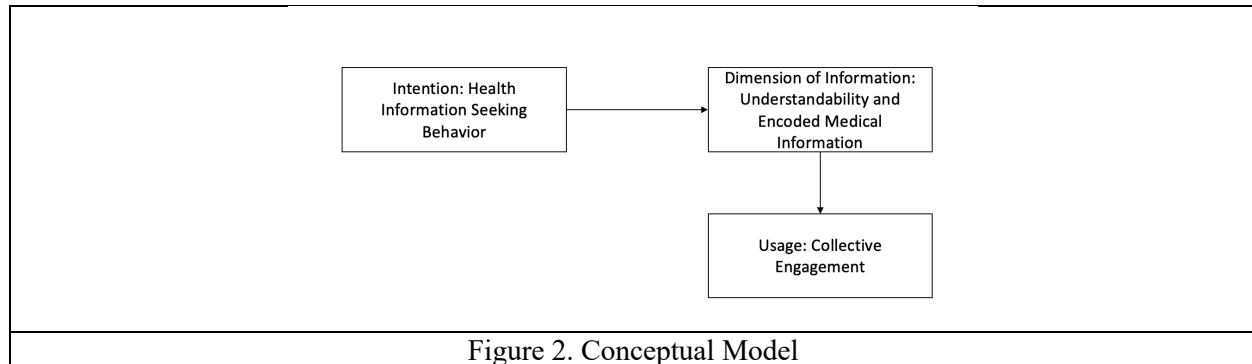


Figure 2. Conceptual Model

Prior studies have employed domain experts, such as health professionals, to evaluate online medical information (Backinger et al. 2011; Dawson et al. 2011). To translate this to the context of an algorithmic artefact, we build on prior studies of HISB that looked at the volume of medical information and the complexity of medical information provided (Stellefson et al. 2014). We consider prior research in IS that has identified encoded medical information to assess HISB (Liu et al. 2020). Wyatt (2002) defines codification as a means of identifying, capturing, indexing and making available explicit knowledge. Encoded medical information reduces information asymmetry between healthcare consumers and providers (Eysenbach and Jadad 2001). Medical information in online videos is conveyed through medical terminology, i.e., healthcare related words such as diseases, conditions, procedures, symptoms, and treatments (Fernandez-Llatas et al. 2017). The Unified Medical Language System (UMLS) developed by the National Library of Medicine (NLM) maps an exhaustive lexicon of medical terms as concepts (Bodenreider 2004). We use UMLS as our framework to define the construct of encoded medical information and employ deep learning methods to extract encoded medical information. Deep learning allows computational models, composed of multiple processing layers based on neural networks, to learn representations of data with multiple levels of abstraction (LeCun et al. 2015).

Studies have examined understandability in online patient educational materials from those created by professional societies to user generated content (Kunze et al. 2020, Rooney et al. 2020), using guidelines such as the Clear Communication Index from Center of Disease Control and Prevention (CDC)¹ (Johnson et al. 2020), PEMAT from Agency of Health Research and Quality (Salama et al. 2020), Benchmark criteria from Journal of American Medical Association (Kunze et al. 2020), Suitability Assessment of Materials (SAM) (Desai et al. 2013), Global Quality Score (Kunze et al. 2020), and readability indices (Rooney et al. 2020). Table 2 below summarizes some of this literature. Studies on the readability, suitability or comprehensibility of patient education materials find that most education materials are too complex for patients with low health literacy (Smith et al. 2014, Mishra and Dexter, 2020). Research suggests that well-made health promotion videos act as a low-cost intervention to improve knowledge, self-efficacy, and attitudes to prevent chronic illness (Latif et al. 2013), and improved adherence especially in low literacy, vulnerable populations (Ramagiri et al. 2020). We build on the Patient Education Materials Assessment Tool (PEMAT), which is a systematic method to evaluate and compare the understandability and actionability of patient education materials in written, audio or video format. PEMAT posits that: “patient education materials are *understandable* when consumers of diverse backgrounds and varying levels of health literacy can process and explain key messages” (Shoemaker et al. 2014). We therefore consider an algorithmic assessment of understandability of videos as a facet of the HISB artifact.

Table 2. Summary of studies evaluating videos for patient education

Study	Data	Guidelines for patient education	Criteria	Findings
Williams et al. 2016	950 written patient educational materials	Three guidelines from AMA, CDC, and NIH for written materials	Readability, structure, and presentation	Materials are consistently written at a readability level that is poorly suited for patients with low health literacy.
Kunze et al. 2020	50 YouTube videos of meniscus	JAMA benchmark criteria, and Global Quality Score	Quality and reliability	Information on the meniscus found in YouTube videos is of low quality and reliability.
Sanderson et al. 2016	One YouTube video about Genome Sequencing	N/A	Understandability, knowledge increased	79% reported the video was easy to understand, satisfaction scores were high, and knowledge increased significantly.

¹ <https://www.cdc.gov/ccindex/index.html>

Salama et al. 2019	53 YouTube videos about hypospadias	PEMAT	Understandability and actionability	Only 5.6% of videos are understandable and 15.1% are actionable. The vast majority of hypospadias-related YouTube content is not appropriate for users with low health literacy
Desai et al. 2013	607 videos from Mayo Clinic's social media health network	Suitability Assessment of Materials (SAM)	Suitability and user engagement	Healthcare organizations produce very few videos with high SAM scores. An optimal video is no more likely to engage users than less optimal videos.

2.4 Hypotheses

While prior research on engagement with social media has primarily examined individual level engagement, YouTube offers a unique context to study aggregate engagement across the population of users (Wu and Huberman 2007). The collective engagement with videos provides us with proxies for popularity, relevance of content to users as well as the collective attention attracted by the videos (Liu et al. 2020). There is tremendous heterogeneity in content consumption or experience of content on YouTube. A substantial fraction of YouTube videos are only viewed once, while the top 20% of videos gather more than half of all views (Peck and Mullen 2008). Given the fleeting popularity of videos (Wu and Huberman 2007) and power laws in online attention (Keselman et al. 2008), the HISB process on YouTube can be noisy. Viewers searching for widely prevalent medical symptoms could easily be led to a popular but irrelevant video. By contrast, videos that contain valuable medical information (and those made by reputable medical institutions) may not be as engaging if their content is not easily understandable from a patient education perspective. We therefore need to consider the simultaneous impacts of understandability (from a patient education perspective) and encoded medical information on the collective engagement with a video.

We consider three dimensions of engagement that builds on prior work (Wu and Huberman 2007; Mocanu et al. 2015; Liu et al. 2020). The process of non-engagement occurs when participants are not engaged either by their cognitive mental models or when they seek information that is not of interest to them. Selective attention drives a basic form of engagement as a user quickly glancing at a video that has the potential of relevant interest. Selective attention is required to trigger a point of engagement (O'Brien and Toms 2010). Sustained attention generates a more elaborate form of engagement and allows the possibility of showing affect, feedback, and interaction (Peters et al. 2009). Sustained engagement is marked by participants' attention being maintained in the interactions.

Theories of health literacy may explain how users engage with understandable information in videos related to healthcare. While users with greater health literacy may value the depth of medical information encoded in a video, those who have low health literacy may encounter significant difficulties in interpreting complex medical information. Desai et al. (2013) found that although authoritative/credible healthcare organizations produce highly educational and/or suitable medical videos, user engagement was greater for inaccurate videos. Patients who are motivated to learn more about their condition might be highly engaged by relevant videos on YouTube, resulting in self-empowerment, greater confidence in their treatment, and acceptance of their illness. When users are actively seeking to manage their conditions using *understandable* medical information obtained from YouTube, they are less likely to disengage with the videos. Therefore, we posit:

H1: *High understandability of a video, when coupled with high encoded medical information, is negatively associated with user disengagement.*

As users are interested in information related to health conditions and disease management, understandability of information in the videos can match with users' interest and lead to points of engagement. Hence, we posit:

H2: *Higher understandability of a video, when coupled with high encoded medical information, is positively associated with selective attention driven user engagement.*

Theories of health literacy may explain how users engage with understandable information in videos related to healthcare. Users with greater health literacy may value the depth of medical information encoded in a video, leading to greater sustained attention. Thus:

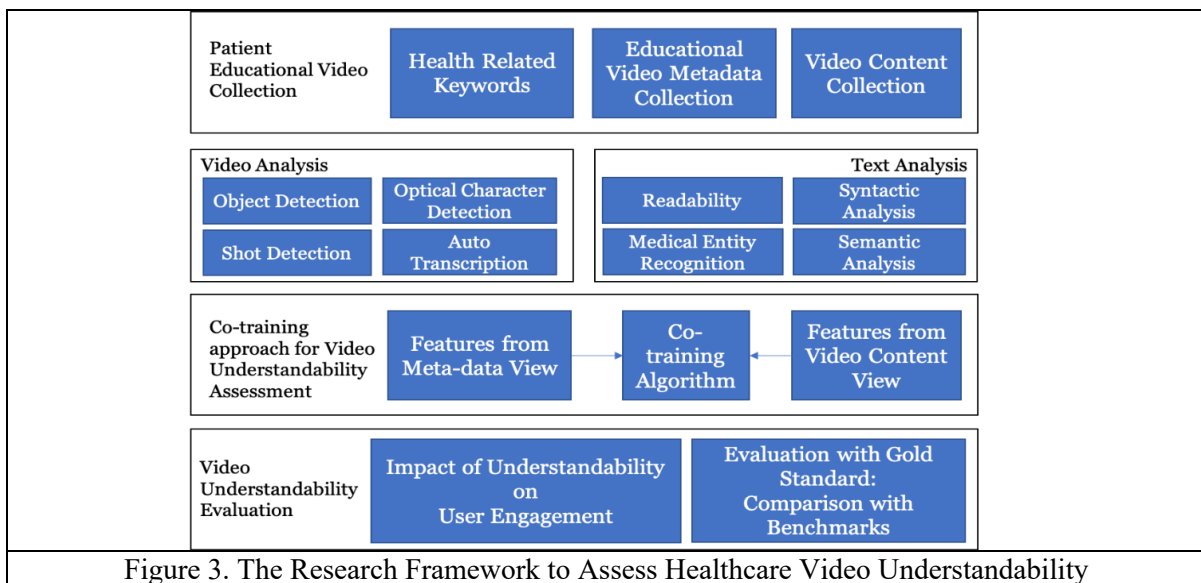
H3: *High understandability of a video, when coupled with high encoded medical information, is positively associated with sustained attention driven user engagement.*

We validate the above hypotheses using algorithmic approaches to assess encoded medical information in a video and its understandability and their impact on collective engagement.

3. Data and Research Methods

3.1 Video Collection

More than 100 million U. S. adults are now living with diabetes or prediabetes according to the Centers for Disease Control and Prevention (CDC 2020²). As of 2018 it was reported that 34.2 million Americans – 10.5 percent of the U.S. population – have diabetes. Another 88 million have prediabetes, a condition that if not treated often leads to type 2 diabetes within 5 years. Type 2 diabetes is increasing in the population and is proving difficult to control with conventional therapy. Diabetes is a contributing factor to many other serious health conditions, such as heart disease, stroke, nerve and kidney diseases, and vision loss. To reduce the impact of prediabetes and type 2 diabetes, healthcare institutions and medical professionals are applying a multi-pronged approach to increase awareness of diabetes and promote patient education on self-management and lifestyle behavior change programs to improve healthy eating habits and increase physical activity (Shrivastava et al. 2013). Understandable, multi-media and content rich videos can complement and support clinician and public health efforts. We develop a scalable, generalizable, augmented-intelligence-based, co-training approach to assess the understandability of YouTube videos. Figure 3 illustrates our approach, which consists of five components: video collection; video analysis; text analysis; co-training approach for video understandability assessment; and understandability evaluation.



² <https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf>

The video collection should represent what YouTube searches return to patients. Combining physician input with literature review helped us identify 235 search keywords related to diabetes patient education to investigate medical information encoded in the videos. These keywords cover various aspects of diabetes patient education including general information about the disease, treatments, lab tests, prevention, self-management procedures, and lifestyle management. We collected the top 50 videos for each search term with YouTube Data API and stored the videos, their rankings, and metadata in a database for further analysis. The attributes we collected from each video are grouped and illustrated in Figure 4 below. Attributes related to video snippet and content details are generated at the time of video upload while video usage is generated by user engagement over time and the statistics are from the day of video data collection. In total, we collected 9,873 unique videos using over 200 search terms, which will serve as the data for the current study on video understandability. These terms are available in Table A2 in the Online Appendix.

Video Snippet	Content Details	Usage Statistics
<ul style="list-style-type: none"> • PublishedAt • Title • Description • Thumbnails • Tags • CategoryID • DefaultLanguage • ChannelID • ChannelTitle 	<ul style="list-style-type: none"> • Duration • Dimension • Definition • CaptionAvailability • LicensedContent • RegionRestriction 	<ul style="list-style-type: none"> • ViewCount • LikeCount • DislikeCount • CommentCount • CommentText

Figure 4: Data Attributes from YouTube Data API

3.2 Video Annotation

YouTube offers a diverse range of content and perspectives from professionals in healthcare organizations to patients, caregivers and the general public. Our dataset includes professionally produced videos by healthcare organizations and individuals on the basics of diabetes, its complications and treatments. It also contains research presentations from renowned researchers and medical experts on the latest research development and scientific findings about the disease as well as low content or inaccurate videos produced by both individuals and organizations. This diversity introduces a challenge for annotating the videos. Video understandability is a critical requirement of patient education videos. We rely on experts' consensus perspective to evaluate this requirement using a widely adopted standard, The Patient Educational Material

Assessment Tool (PEMAT) for audio and video materials (Shoemaker et al. 2014). Table 3 lists our adaptation of PEMAT's focus on four aspects of video materials, specifically: content, word choice and style, organization, and layout and design, with multiple criteria within each aspect. The understandability score of a video is calculated based on the scores for each criterion with the following equation. When a video is scored above 50%, it is considered to have high understandability.

$$Understandability = \frac{\text{The total number of 1's in PEMAT result}}{12 - \text{the total number of NA's in the PEMAT result}} \times 100\% \quad (1)$$

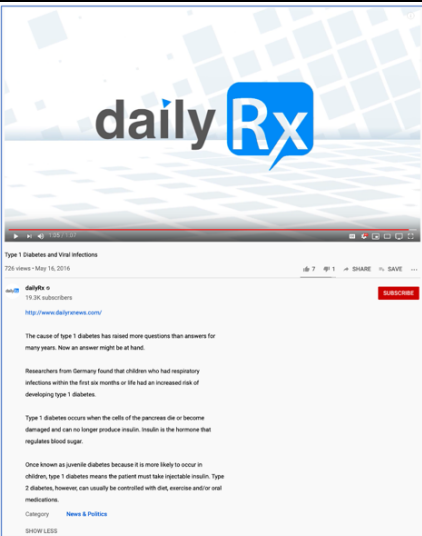
Table 3. Patient Educational Material Assessment Tool for Audio and Video Materials

Patient Educational Material Assessment Tool – Video Understandability		
Content		
1	The material makes its purpose completely evident.	0, 1
Word Choice & Style		
2	The material uses common, everyday language.	0, 1
3	Medical terms are used only to familiarize audience with the terms. When used, medical terms are defined.	0, 1
4	The material uses the active voice.	0, 1
Organization		
5	The material breaks or “chunks” information into short sections.	0, 1, N/A
6	The material's sections have informative headers.	0, 1, N/A
7	The material presents information in a logical sequence.	0, 1
8	The material provides a summary.	0, 1, N/A
Layout & Design		
9	Text on the screen is easy to read.	0, 1, N/A
10	The material allows the user to hear the words clearly (e.g., not too fast, not garbled).	0, 1, N/A
11	The material uses illustrations and photographs that are clear and uncluttered.	0, 1, N/A
12	The material uses simple tables with short and clear row and column headings.	0, 1, N/A

*Scoring: 0 = disagree, 1 = Agree, N/A = Not applicable

Given the volume and scope of healthcare videos on YouTube, manual evaluation as well as annotation of a large number of videos by domain experts can be time consuming and costly, hence impractical. We develop an automated approach that employs a semi-supervised method called co-training which not only learns from the labeled observations but also leverages the unlabeled instances to improve model performance. 600 diabetes related videos are randomly selected from our corpus of 9,873 unique videos as the initial labeled dataset for co-training. Another 100 videos are sampled for evaluation. Sample size calculation indicates that less than 500 videos are needed to achieve high inter-rater reliability ($\kappa > 0.80$) with multiple raters (Donner and Rotondi 2010). The remaining videos are used as unlabeled data to

evaluate the effectiveness of co-training for semi-supervision. When the machine learning models yield inconsistent results, the medical experts will review the videos and provide supervision according to PEMAT. Four physicians, trained to use these guidelines, labeled these videos for video understandability according to the PEMAT guideline in Table 3. They watch a video, assess the video according to the criteria within content, word choice and style, organization, layout and design, and assigns them 0, 1, or N/A (not applicable). Figure 5 demonstrates the expert evaluation measures and result³. The video in Figure 5 is considered to have high understandability.



Patient Educational Material Assessment Tool Video Understandability		
Content		
1	The material makes its purpose completely evident.	1
Word Choice & Style		
2	The material uses common, everyday language.	1
3	Medical terms are used only to familiarize audience with the terms. When used, medical terms are defined.	1
4	The material uses the active voice.	1
Organization		
5	The material breaks or “chunks” information into short sections.	0
6	The material’s sections have informative headers.	0
7	The material presents information in a logical sequence.	1
8	The material provides a summary.	0
Layout & Design		
9	Text on the screen is easy to read.	1
10	The material allows the user to hear the words clearly (e.g., not too fast, not garbled).	1
11	The material uses illustrations and photographs that are clear and uncluttered.	1
12	The material uses simple tables with short and clear row and column headings.	N/A
Video Understandability Score = 8/11 High Understandability		

Figure 5. An illustrative example of PEMAT annotation by domain experts

The PEMAT is designed to be completed by healthcare professionals, including healthcare providers, health librarians, and other clinical practitioners. The selected raters fall into the targeted user group who are qualified to use the PEMAT tool to rate the videos. Before they start working on annotation, all of them have carefully studied the PEMAT user guide⁴. To maximize the consistency among these raters, we had each rater independently rate the same ten videos. A study session was held with these four raters to discuss items with discrepancies. Each rater provided his/ her rationale for the rating provided. The group reviewed the PEMAT user guide to clarify how each item was intended to be rated and come to consensus. Then they

³ Four domain experts watch a video (<https://www.youtube.com/watch?v=4JLnkpdpjoU8>) and assess the video according to its content, word choice and style, organization, layout and design. They assign scores from 0, 1, or N/A (not applicable) to items in Table 1. The video in Figure 3 is considered to have high understandability.

⁴ https://www.ahrq.gov/sites/default/files/publications/files/pemat_guide.pdf

rated the rest of the videos based on the consensus. We use intraclass correlation coefficient (Bartko 1966) to assess the interrater reliability of the annotation at the video level. To ensure there is an agreement on every video, we have a fifth rater to review and consolidate the videos with discrepancies. Each video takes approximately 10 minutes to review. The inter-rater reliability of the video understandability score is 87%. Table 4 below summarizes video understandability scores (according to PEMAT guidelines).

Table 4. Video Understandability Annotation (on a binary scale)

Variables	# of 0 (no)	# of 1 (yes)	# of N/A
The material makes its purpose completely evident	175	525	0
The material uses common, everyday language	183	517	0
Medical terms are used only to familiarize audience with the terms. When used, medical terms are defined	241	459	0
The material uses the active voice	174	626	0
The material breaks or “chunks” information into short sections	548	143	9
The material’s sections have informative headers	601	90	9
The material presents information in a logical sequence	164	536	0
The material provides a summary	458	233	9
Text on the screen is easy to read	137	294	269
The material allows the user to hear the words clearly	97	539	64
The material uses illustrations and photographs that are clear and uncluttered	111	338	251
The material uses simple tables with short and clear row and column headings	192	57	451
Understandability	315	385	0

3.3 Video Analysis

Video data analysis forms the building blocks for designing our machine learning approach to evaluating patient educational videos. A large body of research has examined object detection within image frames. State-of-the-art-performance for video mining tasks is often achieved by using one of many large open-source datasets or pre-trained models (Lee et al. 2019). In processing the video data, we extract the features according to PEMAT criteria in Table 3 with video analysis techniques from the Google Cloud platform. Table 5 below summarizes the features we extract from video data processing results.

Table 5. Features from Google Cloud Video Intelligence API

Tasks	Features	Description
Detect shot changes	# of scenes in a video	The total number of scenes throughout the video
Optical character recognition	Text on screen	A string of text detected in the video
	Text confidence score	The confidence score of a detected text
Video transcription	Transcribed text	The automated video transcription results
	Transcription confidence score	The confidence score of a transcribed text

PEMAT guideline suggests that breaking the information into small chunks or sections is positively related to video understandability. We utilize scene detection methods to detect the number of scenes in a

video as an indicator of whether the videos are organized in small sections. We build on prior work that has defined a scene as one of the subdivisions of a movie or a play, in which the setting is fixed or when it presents continuous action in one place (Rasheed & Shah 2003). A scene comprises a single, complete and unified event, or segment of a movie. A scene normally occurs in one location and deals with one action; the end of a scene is often indicated by a change in time, action and/or location. Scene detection is a widely adopted method in computer vision and video analytics for video classification, video understanding and management (Xiong & Lee 1998). Video content analysis relies on scene detection to extract story units and segments. We utilize scene detection methods to detect the number of scenes in a given video as an indicator of whether the videos are organized in small sections. Scene change detection estimates the sub-sections in a video (Shahraray 1995). Scene change detection is important in a number of video applications, including video indexing, semantic features extraction, and, in general, pre- and post-processing operations (Hu et al. 2011). Video content analysis relies on scene detection to extract story units and segments.

A video transcript is a text version of a video's audio track. Video transcription techniques can extract the video narratives, which convey a significant portion of the information in the videos (Liao et al. 2013). The quality of narratives also may affect viewers' understanding of the videos. PEMAT evaluates if the material allows the user to hear the words clearly. We conduct video transcription to perform in-depth content analysis and assess the clarity of narratives. Optical character recognition is used to detect and extract text, tables, or illustrations in the videos. Layout and design are an important aspect of patient education video evaluation. Text on the screen should be easy to read. The illustrations and tables should have clear headings. Optical character recognition can extract features related to the clarity of text, tables, and illustrations and enable the evaluation of video layout and design.

PEMAT expects a video or a multimedia material with narration to allow the viewer to hear the words clearly. The narrator or voiceover should not be speaking too fast nor should the speech be garbled or hard to understand. Video transcription algorithm returns not only the transcript but also the confidence score of the predicted transcript. The confidence score reflects whether the speech is clear. We also assessed whether the text on screen is easy to read with optical character recognition (OCR). Audiovisual materials

that are overcrowded with words or have text that flashes briefly on the screen are difficult to read and understand. This item is not applicable (N/A) if no text appears in the material or a narrator reads all of the text out loud, because the material is not relying on the viewer to read the text. We use OCR to detect text in the videos and the confidence score of OCR as a proxy for whether the text is easy to recognize.

3.4 Text Analysis for Videos

Readability

We assess whether the material uses common everyday language with readability analysis. PEMAT requires that the material should use common, everyday language that would be easy to understand for most consumers or patients nearly all of the time. To assess this criterion, we use the Flesch-Kincaid readability test to indicate how the description and the transcription of a video is to understand. The Flesch-Kincaid readability test was developed under the contract to the U.S. Navy research in 1975 (Kincaid et al. 1975). It has been widely adopted especially in the public health domain to assess the how easy a material is to read (Paasche-Orlow et al. 2003; Si & Callan 2001; Basch et al. 2020; Basch et al. 2017).

Syntactic analysis

PEMAT assesses whether the material uses active voice. It is often argued that passive voice will result in a structure which is more verbose than active voice and therefore harder to understand and that the meaning of passive voice is indirect and/or less forceful than an active voice (Millar et al. 2013). Therefore, the use of active voice is highly advocated in patient educational materials, medical writings, and other areas. According to PEMAT, if the material overall uses active voice, this criterion is met. To automatically assess this criterion, we use part-of-speech tagging (Voutilainen 2003), a common linguistic technique, to detect the category of verbs in the video description and narratives, and compute the number of verbs in active voices. The number of verbs in active voice is extracted with Part-of-speech tagging. The verbs in active voices belong to the tag set: VB, VBD, VBG, VBP and VBZ (Marcus et al. 1993).

Medical entity recognition

We adopted a Bidirectional Long-Short Term Memory model from prior work to extract six types of medical terms from the text data (Liu et al. 2020). Table 6 lists the medical term categories and provides

explanations. These six categories cover the majority of medical terminologies used in the patient educational materials and communications (Fage-Butler and Nisbeth Jensen 2016).

Table 6. Medical Terminologies Used in Patient Educational Materials and Communications

Medical Term Category	UMLS Semantic Type	Examples
Body part	bdsy (Body System), blor (Body Location or Region), bpoc (Body Part, Organ, or Organ Component)	Liver, foot, pancreas
Chemicals or Drugs	chem (Chemical), chvf (Chemical Viewed Functionally), chvs (Chemical Viewed Structurally), clnd (Clinical Drug), elii (Element, Ion, or Isotope), enzy (Enzyme), hops (Hazardous or Poisonous Substance), inch (Inorganic Chemical), orch (Organic Chemical), phsu (Pharmacologic Substance)	Insulin, Metformin, Lantus
Medical devices	drdd (Drug Delivery Device), medd (Medical Device)	Insulin pen, glucometer
Medical events	acab (Acquired Abnormality), dsyn (Disease or Syndrome), inpo (Injury or Poisoning), mobd (Mental or Behavioral Dysfunction), patf (Pathologic Function), sosy (Sign or Symptom)	Nausea, ketosis, diabetes
Medical professionals	humn (Human), famg (Family Group)	physician, diabetes educators, nurses
Medical procedures	lbpr (Laboratory Procedure), lbtr (Laboratory or Test Result), topp (Therapeutic or Preventive Procedure)	HbA1C, Creatinine.

Five thousand sentences were randomly selected from the video description and transcription test bed with 4,000 in the training set and 1,000 in the test set. Two expert annotators independently labeled the sentences for semantic types. We used Cohen’s kappa to measure inter-annotator reliability. The kappa value is 0.90 for the medical terminology annotation. A third annotator reviewed the disagreements and made the final judgments. Finally, the ground truth was generated, containing 4,000 training sentences and 1,000 test sentences. The statistics of the training and test sets are shown in Table 7 below.

Table 7. Statistics of the Training and Test Sets

	Training Set	Test Set
# of sentences	4,000	1,000
# of mentions for body part	227	101
# of mentions for chemical and drugs	2,181	538
# of mentions for medical devices	545	126
# of mentions for medical events	784	245
# of mentions for medical professional	67	18
# of mentions for medical procedures	197	53

We train an embedding model using the Skip-gram method in Word2vec and devise a Bidirectional Long Short-term Memory model to extract medical terms from video descriptions and transcriptions at the sentence level. Overall, the model achieves a precision of 87.4%, a recall of 87.8%, and an f-measure of 87.3%. We have also provided several experiments to evaluate the classification of our method in

comparison to dictionary based approaches and state of the art methods such as conditional random fields (CRF). Performance is reported in Table A1 in the Online Appendix. We then extract medical terms from video descriptions and transcriptions using the model.

Semantic analysis

PEMAT expects materials to have a summary of the key points or review of the key points at the material, either in writing or orally. The summary usually comes at the end of material and starts with summary words. Therefore, we curated a comprehensive list of summary words and phrases from multiple sources and use them to detect whether a material provides a summary.

PEMAT suggests that information in a material should be presented in an order that makes sense to the user. Main messages or most important ideas should be at the beginning of sections or in lists because users tend to pay more attention to them. To measure whether the material presents information in a logical sequence, we evaluate the use of transitional words and phrases in the material. A transition is a change from one idea to another in writing or speaking and can be achieved using transition terms or phrases, which are most often placed at the beginning of sentences, independent clauses, and paragraphs and thus establish a specific relationship between ideas or groups of ideas. Transitions are used to create “flow” in writing or speaking and make its logical development clearer to the audience. The use of transition words and phrases can improve the logical connections in writing and speech (Harbaugh 2013).

Transition words and phrases can be grouped into categories such as causation, chronology, combinations, contrast, example, clarification, summary and more. We collect common transitional terms and phrases under these categories as a proxy to measure whether the material presents information in a logical sequence. Table 8 lists all the words and phrases we use to identify transition and summaries.

Table 8. Words and Phrases for Summary and Transition

Category	Expressions
Summary/Conclusion ⁵	finally, in a word, in brief, briefly, in conclusion, in the end, in the final analysis, on the whole, thus, to conclude, to summarize, in sum, to sum up, in summary, lastly

⁵ <https://writingcenter.unc.edu/tips-and-tools/transitions/>

Transition ⁶	Accordingly, as a result, and so, because, consequently, for that reason, hence, on account of, since, therefore, thus, after, afterwards, always, at length, during, earlier, following, immediately, in the meantime, later, never, next, once, simultaneously, so far, sometimes, soon, subsequently, then, this time, until now, when, whenever, while, additionally, again, also, and, or, not, besides, even more, finally, first, firstly, further, furthermore, in addition, in the first place, in the second place, last, lastly, moreover, next, second, secondly, after all, although, and yet, at the same time, but, despite, however, in contrast, nevertheless, notwithstanding, on the contrary, on the other hand, otherwise, thought, yet, as an illustration, e.g., for example, for instance, specifically, to demonstrate, to illustrate, briefly, critically, foundationally, more importantly, of less importance, primarily, above, centrally, opposite to, adjacent to, below, peripherally, below, nearby, beyond, in similar fashion, in the same way, likewise, in like manner, i.e., in other word, that is, to clarify, to explain, in fact, of course, undoubtedly, without doubt, surely, indeed, for this purpose, so that, to this end, in order that, to that end.
-------------------------	--

We evaluate whether the material makes its purpose evident. According to the PEMAT User’s Guide, this criteria refers to whether the material uses a title or upfront text that tells the reader what the materials about. In the implementation, we implement this criterion with checking whether this video has a title, tags, and description. YouTube suggests that tags are descriptive keywords content creators can add to the video to help viewers find the content. Your video’s title, tags, and description are important pieces of metadata for the video’s discovery. These main pieces of information should provide important information about the purpose of the video so that viewers can find the video and decide whether to watch it.

3.5 Co-training Approach for Video Understandability Assessment

Co-training is a multi-view learning paradigm that exploits unlabeled data in addition to labeled data to improve learning performance (Blum and Mitchell 1998). It is not feasible to obtain a large amount of annotated video data given the domain expertise required. Co-training trains two learners, respectively, from two different views and lets the learners label the most confident unlabeled instances to enlarge the training set of the other learner (Platanios et al. 2017). When the two learners are inconsistent, a human expert will evaluate the performance and decide on the label. Such a process is repeated until some stopping condition is met. Intuitively, each example contains two “views,” and each view contains sufficient

⁶ <https://writing.wisc.edu/handbook/style/transitions/>

information to determine the label of the example. This redundancy implies an underlying structure of the unlabeled data (since they need to be “consistent”), and this structure makes the unlabeled data informative. Variants of this approach have been used for a variety of learning problems, including recommender systems (Zhang et al. 2014), text classification (Ma et al. 2017), natural language processing (Pierce and Cardie 2001), and image recognition (Ma et al. 2017). The co-training approach allows us to reconcile the disagreement between learning methods using a human-in-the-loop augmented intelligence mechanism. We define classification of video understandability in the context of patient education as a multi-view learning and binary classification problem. Our dataset includes the video metadata and video data. We develop classifiers from two sufficient and conditionally independent views (i.e., video metadata and video content) to assess the video understandability. Feature based classification methods are adopted for both video metadata view and video content view. Features from videos are selected and extracted according to PEMAT guidelines.

3.5.1 Co-training based understandability classification

Figure 6 illustrates the procedures of the co-training approach for video understandability classification. This consists of the following: a set of L labeled videos and a set of U unlabeled videos, classifier F_1 trained with features from video metadata view, classifier F_2 trained with features from video content view, and a hyper parameter confidence threshold. The video metadata contains the video title, video description, video tags, and video usage information. It represents how the content creators would like the viewers to perceive the video. The video content view captures the information delivered by the video. Combining video content and video metadata gives us a comprehensive view of the videos on YouTube. An initial labeled dataset L , and an unlabeled dataset U are given. The co-training process in this study can be summarized as: (1) The video metadata view of L is used to initialize the classifier F_1 , and the video content view of L is used to initialize the classifier F_2 . (2) F_1 and F_2 are used to make predictions on the unlabeled data U . (3) A certain number of the most confident newly labeled videos are selected with a confidence threshold. p_1 and n_1 are the most confident positive and negative examples predicted by F_1 , while p_2 and n_2 are those predicted by F_2 . (4) When a video falls within the high confidence positive

examples (p_1 and p_2) or the high confidence negative examples (n_1 and n_2), it indicates that this video is classified consistently positive or negative with high confidence. The video will be added to the labeled dataset L . (5) When a video gets inconsistent but high confidence predictions (i.e., belonging to p_1 and n_2 or n_1 and p_2), the medical experts will review the videos and provide supervision. The video with its expert label will be added to L . (6) Videos low confidence predictions remain in the unlabeled set. (7) Train F_1 and F_2 based on updated L . The process is repeated until the unlabeled dataset is depleted, or no new videos are added to the labeled set⁷. The detailed algorithm is presented in Table 9.

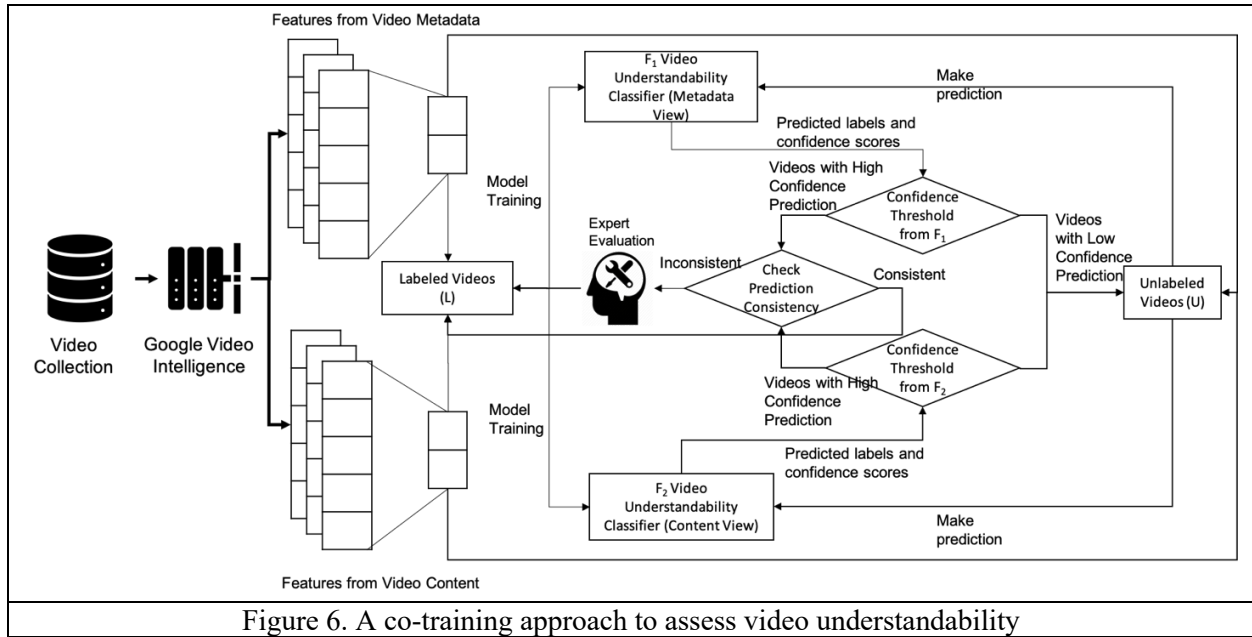


Figure 6. A co-training approach to assess video understandability

Table 9. Pseudocode for co-training algorithm

Input:
A set L of labeled video examples
A set U of unlabeled video examples
Output:
A set L' of labeled video examples
Procedures:
Loop for K iterations
1. Train a classifier F_1 on L that considers only video metadata features; train a classifier F_2 on L that considers only features from video content features
2. Use the trained classifiers to make predictions on videos in U
3. Extract p_1 positive and n_1 negative examples from U on which F_1 has most confident predictions, specified by a confidence threshold
4. Extract p_2 positive and n_2 negative examples from U on which F_2 has most confident predictions, specified by a confidence threshold

⁷ After the classifiers F_1 and F_2 converge, we discard videos in the unlabeled set.

-
5. Compare p_1 with p_2 , and n_1 with n_2
 6. If a video appears in both p_1 and p_2 or both in n_1 and n_2 , move the video and its label from U to L
 7. If a video appears in both p_1 and n_2 or both p_2 and n_1 , expert reviewers annotate inconsistent labels and add the final label of this video to L
 8. Halt when U is empty or no new videos are added to the labeled set
-

3.5.2 Features from Video Metadata View

In the video metadata view classifier, we leverage the features generated from video metadata to classify video understandability. Each video's metadata contains video title, description, and tags, which are submitted by the content creator. These elements suggest the purpose of a given video. Therefore, we can use them to represent whether this video makes its purpose evident. A video with good understandability to patients uses common daily language. Text preprocessing techniques are used to identify the total number of words, sentences, and unique words from the video description. A Bidirectional Long-Short Term Memory named entity recognition model is used to extract the number of the medical terms (Liu et al. 2020). Table 10 summarizes the features we propose to extract from video metadata view, the method to derive the measure, and PEMAT criteria they fall under.

Table 10. Features for Video Understandability Classification from Video Metadata View

Feature name	Feature description	Method	PEMAT Criterion
Has title	Whether the video has a title	Metadata collection	The material makes its purpose evident
Has description	Whether the video has a text description	Metadata collection	
Has tags	Whether the video has tags	Metadata collection	
Description Readability	The automated readability index of the video description	Readability analysis	The material uses common everyday language
Active word count	The number of verbs in active voice in the video description	Syntactic analysis	The material uses active voice
Summary word count	The number of summary words in the video description	Semantic analysis	The material provides a summary
Transition word count	The number of transition words in the video description	Semantic analysis	The material presents information in a logical sequence.
Video duration	The total length of the video in seconds	Metadata collection	Medical information encoded in the video (Liu et al. 2019)
Description word count	The total number of words in the video description	Metadata collection	
Sentence count	The total number of sentences in video description	Metadata collection	
Description unique words	The total number of unique words in the video description	Metadata collection	
Description medical term count	The total number of medical terms in video description	Medical entity recognition	

3.5.3 Features from Video Content View

In the video content view, we derive features from the video narratives, video shots and associated confidence scores. We generate narrative readability score to examine whether the material uses common everyday language. Part-of-speech tagging is used to extract verbs in active voice in the transcript. The numbers of transition words and summary words are identified according to the transition word list. We use the video transcription confidence score as a proxy for whether the users can hear the words in narratives clearly. Videos are often broken into different chunks by scenes. We use Google Video Intelligence to detect the number of scenes in the video as an indicator if the video has short sections and employ text processing methods to generate features from the transcript. Table 11 summarizes the features we extract from video content view, the methods to derive the measure, and PEMAT criteria they fall under.

Table 11. Features for Video Understandability Classification from Video Content View

Feature Name	Feature Description	Method	PEMAT Criterion
Narrative readability	The automated readability index for narrative	Readability analysis	The material uses common everyday language
Active word count	The number of verbs in active voice in the video transcript	Syntactic analysis	The material uses active voice
Summary word count	The number of summary words in video transcript	Semantic analysis	The material provides a summary
Transition word count	The number of transition words in the video transcript	Semantic analysis	The material presents information in a logical sequence.
Video transcription confidence	The video transcription confidence score	Auto transcription	The material allows users to hear the words clearly
Text detection confidence	Text recognition confidence score	Optical character recognition	The text on screen is easy to read
Scene count	The total number of scenes in the video	Scene detection	The material breaks or “chunks” information into short sections.
Transcript word count	The total number of words in the video transcript	Auto transcription	Medical information encoded in the video (Liu et al. 2019)
Transcript unique word	The total number of unique words in the transcript	Auto transcription	
Transcript sentence count	The number of unique words in a video	Auto transcription	
Transcript medical term	The total number of medical terms in the video	Medical entity recognition	
Video object	The total number of unique objects in the video.	Object detection	

4. Evaluating Video Understandability Classification Performance

4.1 Video Understandability Classification

Our co-training model initially starts with 600 labeled videos for training. The model converged after 12 iterations and 305 video annotations. Table 12 shows the coefficients of the logistic regression classifiers for each view. The active word count and summary count have a significant and positive impact on understandability. The transition word count in narratives is significant but that of description is not. Transcription confidence and text detection confidence have a positive impact on video understandability. Video duration, medical terms count in descriptions and transcriptions negatively affected the video understandability. The readability scores of the description and narratives have a significant and positive impact on the video understandability.

Table 12. Logistic Regression Model Summary

F1: Video Metadata View			F2: Video Content View		
Variable Name	Estimate	P-value	Variable Name	Estimate	P-value
Has title	-0.335	0.347	Narrative readability	0.132	0.054
Has description	-0.217	0.153	Active word count	0.017	0.056
Has tags	-0.184	0.176	Summary word count	0.117	0.095
Description readability	0.367	0.073	Transition word count	0.045	0.087
Active word count	0.029	0.088	Transcription confidence	0.028	0.043
Summary word count	0.152	0.049	Text detection confidence	0.021	0.039
Transition word count	0.096	0.104	Shot count	-0.254	0.203
Video duration	-0.071	0.086	Transcript word count	-0.036	0.141
Description word count	0.038	0.144	Transcript unique word	-0.085	0.072
Sentence count	0.157	0.121	Transcript sentence count	-0.074	0.143
Description unique words	0.085	0.144	Transcript medical term	-0.009	0.045
Description medical term count	-0.020	0.067	Video object	-0.104	0.055
Constant	-0.319	0.11	Constant	-0.272	0.117

Most significant variables are consistent with PEMAT. Low understandability videos are associated with longer duration, lengthier narratives, and more medical terminologies. For model performance, we compare our predicted results in the 100 videos included in the evaluation set. We compare our model with three benchmark models: logistic regression, Support Vector Machines, and Random Forest. To ensure a fair comparison, we have carefully tuned the model hyperparameters to get the best performance of the benchmark models and proposed method. For logistic regression, we have experimented with difference solvers and regularization methods. Our best performance model utilizes liblinear solver and L2 regularization. The best performance of SVM is achieved by RBF kernel and penalty score of 0.1. The best performance of Random forest model is achieved by $\text{max_features} = \log_2$, and $\text{N_estimator} = 100$.

Table 13 summarizes the classification performance of our proposed method and benchmarks. Our approach achieved a weighted precision of 0.84, weighted recall of 0.79, and F1 score of 0.81 in classifying videos. Figure 7 illustrates the Receiver Operating Characteristic Curves. The results show that co-training method significantly improved the video understandability classification performance.

Table 13. Video Understandability Classification Results

	Precision	Recall	F1 score
Co-training with logistic regression	0.84	0.79	0.81
Logistic regression	0.63	0.60	0.61
Support Vector Machines	0.77	0.75	0.76
Random forest	0.80	0.74	0.77

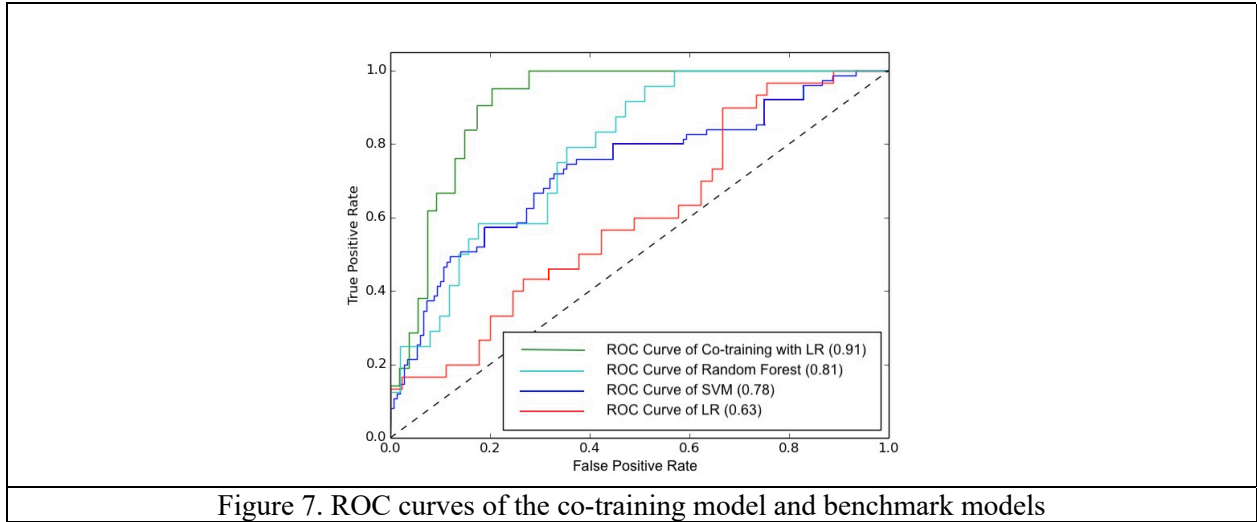


Figure 7. ROC curves of the co-training model and benchmark models

4.2 Comparing Video Understandability Classification with Expert Opinions

Precision at K is a common information retrieval measure used in modern (web-scale) information retrieval systems (Manning et al. 2012). In web-scale retrieval, queries have thousands of relevant documents, and few users will be interested in reading all of them. Precision at K ($P@K$) assesses how many of the top K results are relevant (e.g., $P@10$ or "Precision at 10" corresponds to the number of relevant results among the top 10 documents). To evaluate how significant video understandability is to experts' decision to recommend a YouTube video for patient education, we re-rank the search results from 20 randomly selected queries according to video understandability. Four medical experts reviewed top 10 videos according to our re-ranked results for each query and reported whether they would recommend the videos to patients. We measure the average precision at K with K from 1 to 10 for 20 queries.

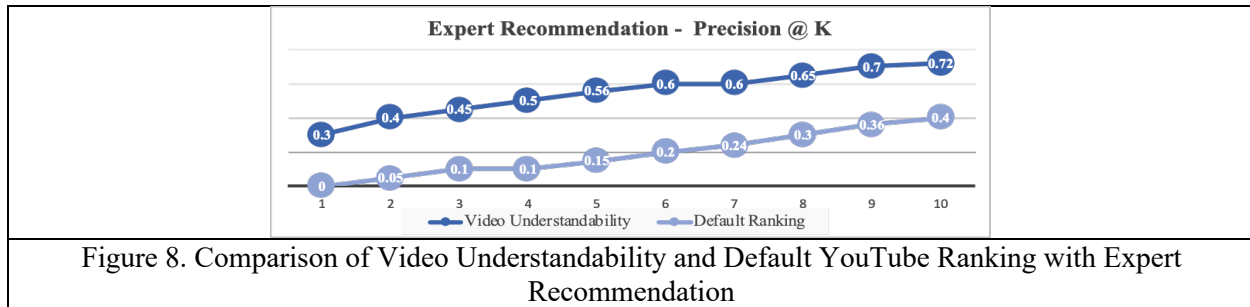


Figure 8 shows a chart comparing the significance of video understandability ranking. 30% of the top ranked videos (videos ranked 1 or Top 1) in understandability are recommended by the expert. None of the top ranked videos according to the default ranking on YouTube received this recommendation. 72% of top 10 videos are recommended by experts ranking by understandability while only 40% of top 10 videos with the default ranking on YouTube are recommended. We conclude that our video understandability measure identifies patient education videos effectively.

4.3 Causal Relationship between Video Understandability and Collective Engagement

Following Liu et al. (2020), we identify three dimensions of collective engagement by conducting a principal component analysis from the video metadata: disengagement, sustained attention driven engagement, and selective attention driven engagement (Peters et al. 2009). Table A3 in the Online Appendix provides the descriptive statistics of engagement measures. Principal component loadings are presented in Table A4 in the Online Appendix. First, we identify disengagement with the video, which could be the result of users cognitively dis-engaging with the content or the topic of the video. The second principal component is labeled sustained attention driven engagement. When users have more sustained engagement with the videos, they interact more with the videos by leaving comments, likes, and dislikes. The third principal component loads strongly with measures of relevance, which reveals the extent to which the content of a video matches the interests of users. We label this dimension as selective attention driven engagement.

A challenge in estimating the impact of understandability and encoded medical information on collective engagement is the endogeneity between collective engagement and dimensions of information in the video. A video's ability to inform viewers about medical terms, as well as the understandability of the

video, could be influenced by external factors, such as the visibility of the channel posting the video. Further, there could be systematic differences in how users engage with understandable vs. non-understandable videos depending on unobserved preferences of users and a channel's propensity to engage viewers. Third, a channel may be more likely to create understandable video materials when it enjoys greater engagement from its users. We cannot observe the counterfactual as we cannot randomize treatment across a channel by showing videos with high/low understandability to viewers and measuring the impact on collective engagement. Given the difficulty in identifying whether it is the understandability of a video that triggers collective engagement, matching on propensity scores (PSM) offers the best means in this context to identify the causal impact (Dehejia and Wahba 2002, Rosenbaum and Rubin 1984). However, videos that are understandable, but lack validated medical information (e.g., Liu et al. 2020), may be more engaging to users if they are cognitively less demanding.

Our constructs of encoded medical information and understandability use features that are a function of channel and video characteristics, which avoids simultaneity with the measures of engagement obtained from YouTube analytics. Further, our method of data gathering considers a temporal separation between video data used to create the measures of encoded medical information and understandability. This data is from the time the video is created. The data on engagement is at a later data and comes from YouTube analytics. Our method of identification essentially considers factors that influence the creation of video (from the choices made by content creators) influencing the aggregate usage statistics of the video (measured through engagement metrics). This is in keeping with an enormous literature on user-generated content as well as borne out with our interviews in practitioners from Facebook AI, YouTube Analytics etc.

The endogeneity issue in this study occurs in the following way: first, the encoded medical information in the video is a result of being created by a reputed actor or unobserved motivations of content creators to post highly informative medical videos on YouTube (likewise for understandability). Second, the collective engagement with the video could arise from unobserved user preferences or results from video-specific factors. Our multi-treatment propensity score matching approach considers the action of a content creator creating an understandable video and creating a video high in encoded medical information

as two distinct treatments. We therefore looked at videos with a similar propensity to be high in encoded medical information and understandability, then conduct a matched sample analysis using propensity scores. Identification is possible since our method of data gathering lends itself to a distinct categorization - as (i) pre-treatment set of channel and video characteristics, (ii) a treatment period where we observe engagement behavior from aggregate users on YouTube, and (iii) post treatment engagement.

In our propensity score matching estimation, we can identify that videos and channels form groups depending on the quality of medical information and understandability that they provide. We conduct multiple-treatment propensity score matching (PSM) to examine this multi-criteria requirement, which extends the single treatment propensity score methods to construct counterfactual groups across multiple treatments. The causal effects involve summary statistics of the individual effects across populations (or sub-populations) of interest. Most causal evaluations, either through randomized trials or observational studies, try to obtain an estimate of the average treatment effect (ATE), which is the difference in means between the observations assigned to the treatment group and those that are in the control group. Causal evaluations also estimate the average treatment effect on the treated (ATT), which is the result of the treatment on the group of individuals that receive the treatment. Propensity score is the conditional probability of receiving a treatment assignment with given covariates, wherein the propensity score can be estimated using methods such as logistic regression. The propensity score allows us to construct a counterfactual group of individuals that did not receive a treatment. PSM is based on the idea that, conditional on observables and assuming that observations in one group are unaffected by the treatment in the other group, matching on the propensity score provides us an estimate of the ATE and the ATT.

In our study, we have videos classified with both high and low understandability and high and low medical information. We build upon Liu et al. (2020) to build a measure of high and low medical information encoded in a video. This gives us four possible treatment conditions to characterize each video, on both dimensions of understandability and medical information. These characteristics of videos are not exogenous, rather, they are determined by video level features, content creator features and the reputation of the content provider etc., which makes the dimensions of medical information and understandability

endogenous, lending itself to a treatment effects type of causal estimation. That is, since we only observe videos that exhibit one of these types (high medical information and high understandability, low medical information and high understandability, high medical information and low understandability, low medical information and low understandability) we only know what happens to the collective engagement for pre-determined medical information and understandability. With multiple treatment propensity score methods, it is as if we can randomize the quality of a video meant for patient education (the level of understandability and medical information in a video) across different content providers, conditional on observables. This allows us to draw inferences about how collective engagement on YouTube with patient education videos varies across groups of content creators.

Table 14. Propensity Score Matching Estimates

PC	Treatment	Estimate	Std. Error	Pr ($> t $)
User Disengagement	(Intercept)	-1.701	0.490	0.001
	Low medical information High understandability	-0.839	0.507	0.099
	High medical information Low understandability	1.136	0.631	0.072
	High medical information High understandability	-0.102	0.052	0.050
	(Intercept)	0.169	0.107	0.101
Sustained Attention Driven Engagement	Low medical information High understandability	0.492	0.203	0.015
	High medical information Low understandability	-0.073	0.288	0.801
	High medical information High understandability	0.286	0.157	0.069
	(Intercept)	-0.262	0.116	0.025
	Low medical information High understandability	0.208	0.150	0.167
Selective Attention Driven Engagement	High medical information Low understandability	0.005	0.170	0.979
	High medical information High understandability	0.181	0.123	0.142

PSM results confirm common assessments of the relationship between user engagement and understandability of education materials (Desai et al. 2013). Our analysis, however, quantify these effects using actual usage data, not survey data, in the specific context of understandability of complex medical information encoded in patient education videos found on YouTube. We find that video understandability has a negative impact on disengagement. A video with high understandability usually attracts more views,

likes, and comments, reducing user disengagement. Moreover, high understandability can help high medical information videos become more engaging. On the other hand, high medical information videos with low understandability are the least engaging. We find support for H1, that high encoded medical information and high understandability are associated with lower user disengagement. Encountering medical videos that are not understandable (from a patient education perspective) could make viewers unwilling to engage with the video. Users encountering complex information in a video could be bewildered by the complexity of medical terminology and abstract concepts, which may lead to lower engagement with the video. When a video is not very understandable, a casual viewer of healthcare information may be too daunted by the content. Videos with high medical information may be intimidating if not explained well. We observe from the results that high understandability can mitigate this issue. We find that high encoded medical information and high understandability are significantly associated with both sustained attention driven engagement and selective attention driven engagement, although the relationship is not significant for selective attention driven engagement, indicating support for H3 but no adequate support for H2.

The video ranking in search results and video occurrence through the YouTube recommendation system are major sources of selective attention driven engagement (Zhou et al. 2010). Such divergence in attention is likely to be exacerbated by the heterogeneity in health literacy among YouTube viewers. On the one hand, users with greater health literacy may value the depth of health advice delivered in an understandable manner and develop a stronger connection with the video. On the other hand, those who lack health literacy cannot interpret the instructions provided by a video. Thus, their subsequent behavior such as their emotional connection towards the video and further interaction could be to disengage with the video altogether. Carefully designed experiments will inform these effects in future studies.

5. Discussions and Conclusions

5.1 Implications for Research

With complex and very large-scale data generated by digital platforms and the range of human behaviors enabled by the ubiquity of algorithmic and digital artefacts in everyday life, the range of motives of information seeking and the outcomes of such information seeking behavior need to encompass a larger set

of phenomena beyond technology acceptance, information quality and cognitive absorption considered in prior IS literature. We extend the literature on personal IT usage to the context of the very large social media platforms. Our proposed evaluation components may also provide new insights on collective user engagement categories learned from aggregated video usage data that can be extended to understand the dynamics of collective engagement with user-generated content or video content. Our research methods can be applied in financial markets where users obtain information about investing through social media, in marketing where brand identities depend on two-way interactions between brands and consumers and in politics and communication studies where individuals study political persuasion and dynamics of attitude formation.

Our method is an alternative to recommendation systems based on content or collaborative filtering-based approaches. As we demonstrate, identifying and recommending relevant materials leveraging the vast corpora of publicly available user-generated content is a feasible way to deliver personalized and contextualized information, be it for healthcare, for do-it-yourself (DIY) projects or even to leverage (UGC) for education. The adaptability of content found on social media has enabled a variety of applications that were hitherto unthinkable. For instance, applications such as PatientsLikeMe, as well as the use of Yelp data in public health and Twitter data in predicting disease outbreaks and progression has parallels to the YouTube context. Well-designed UGC videos, in tandem with evidence from rigorous field experiments, could serve as part of a holistic system of care encompassing disease prevention and lifestyle changes along with resources for emotional support, better patient-physician interactions, and providing current and scientifically valid medical information to patients. Methods such as the ones we developed help us harness the power of free Internet goods that provide broad public benefits (Brynjolfsson et al. 2019) by involving multiple stakeholders in information dissemination.

Our method can be scaled to other settings such as journalism where there are significant concerns about fake news. Other applications include product classification in e-commerce where e-commerce websites typically employ editors and crowdsourcing platforms such as Amazon Mechanical Turk (AMT) to classify products. Integrating text and images from customer reviews into automated machine learning

based classification approaches using co-training algorithms might improve classification accuracy and enable customers to obtain benefits of a content and context-based recommendation. Finally, as machine translation methods are gaining popularity in product categorization, our method provides a scalable framework that incorporates both the elements of deep learning in classification while also allowing for input from domain experts. The methods developed here can be incorporated by video search engines like YouTube, Dailymotion and other popular video archives to improve the quality of their search ranking by accounting for content, structure, vocabulary and other constructs.

Our approach is a first step in extending prior literature on health literacy to an algorithmic context. Health literacy is well recognized as a challenge for public health, with many adults lacking the requisite skills to engage successfully in the management of their healthcare. Our research is also a first attempt at a guideline-driven consolidation of distinct data sources spanning metadata in text form and video data and the first study in an IS context to evaluate video understandability. Our study is among the first in the Information Systems domain to adopt a human in the loop learning strategy to address a video analysis problem for healthcare domain. Expert evaluation is needed when tasks are ambiguous for machine learning models to classify with high confidence. Our approach is developed within the context of patient educational video design but could be generalized to content design for different purposes.

5.2 Implications for Practice

Advocates of social media in medicine highlight social media's potential in enabling patient education and empowerment (Househ et al. 2014), offering the possibility of improving health outcomes (Moorhead et al. 2013). It has been posited that purely text-based medical instructions result in poor patient attention, comprehension, recall and adherence, especially those with low literacy levels. The plethora of user generated content and the visual nature of video content could medical knowledge gaps. For instance, a patient with diabetes can turn to YouTube to find advice about healthy recipes, guidance on how to manage blood sugar levels, information about how to use diabetes related equipment such as insulin pumps etc. Viewing videos specific to managing glycemic indices made by dieticians enables a diabetic patient to better manage her health condition and make daily insulin and medication adjustments that consider diet

and nutrition information. Healthcare organizations lack resources to create video content on such a range of symptoms and disease progressions, as well as offering easily understandable advice that can be integrated into patients' daily routines. Healthcare organizations also may not have the time and resources to provide such advice on topics that are outside the physician-patient interaction in a clinical setting.

By improving public's access to health information and their capacity to use it effectively, improved health literacy through education is critical to empowerment and building societal resilience. In this project, we address these recommendations by proposing the development, implementation, and preliminary evaluation of efficient automated methods for the identification of appropriate user generated content (UGC) in the form of YouTube videos for public education and health promotion. Our method will develop a library of user-centric videos for patients with diabetes, and tailored for stages of disease and treatments, allowing clinicians to recommend/prescribe them to patients to watch at home or during clinic visits. Automating the easy retrieval of understandable patient education videos for clinicians to recommend to their patients offers the potential for significant impact due to scalability and generalizability of the approach. Leveraging advanced machine learning methods for automation combined with social science methods for impact evaluation allows the incorporation of critical elements for improved patient-physician communication such as transparency, credibility, and relevance of the recommended materials.

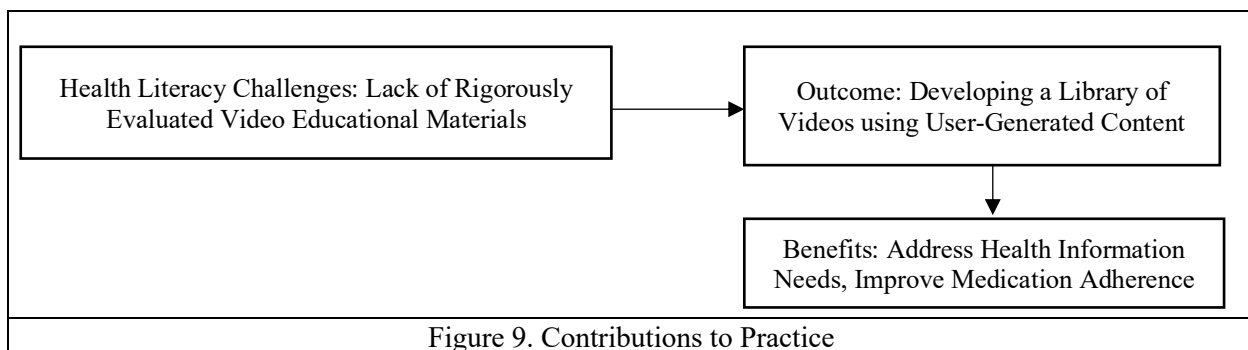
These methods can be refined further and evaluated via randomized clinical trials to improve patient education and societal health and literacy. Health literacy is well recognized as a challenge for public health, with many adults lacking the requisite skills to engage successfully in the management of their healthcare. Our proposed methodology to develop a patient educational video classification system for understandability by integrating human efforts, i.e., the perspectives of clinical practitioners and healthcare consumers, with machine learning algorithms is an innovative approach for a very challenging problem. Patient empowerment and engagement is essential for appropriate disease management. For health organizations who are producing patient educational materials, our approach could be used as an educational tool for enhancing understandability in patient educational video content design.

Table 15. A typology of videos according to understandability and medical information

	Understandability -Low	Understandability-High
Medical Information-Low	1. These are low information and low quality videos	2. These videos have the potential to mislead or misinform users
Medical Information - High	3. These are the videos that are likely to be biased against users with low health literacy	4. Best practice for healthcare organizations

When designing educational materials, the proposed study has the potential to provide best practice guidelines regarding how organizations should engage health consumers with educational videos for chronic care. Given recent concerns in the public policy about promoting popularity-driven engagement, our method highlights different dimensions of engagement that social media platforms can consider in designing recommendations. Videos can present health information using visual aids and cues. Understandability can be further improved with the use of visual aids, summaries, and tangible tools such as personalized charts. The individual level assessment can help identify highly understandable videos.

Our study could add to patient communication and education literature and practice by enabling clinical practitioners to identify the most understandable, medically informative, and engaging videos for their patients as a digital therapy. The combination of algorithmic approaches and causal inference methods aims to find the right intervention methods that can allow both platform designers and clinicians the ability to retrieve the appropriate videos (as a digital therapeutic tool). Our methods parallel recent efforts by digital platforms to identify authoritative sources and to amplify credible content.



5.3 Limitations and future directions

This study has some limitations. Our study is built on the PEMAT developed by AHRQ. Although it is the most prevalent evaluation tool on patient education materials, PEMAT is not designed for user-generated content but materials produced by healthcare organizations. The PEMAT criteria may need to be adapted/extended to YouTube videos in evaluating sub-criteria such as whether the materials used for

illustration were uncluttered, etc. In future work, we would like to explore alternative assessment tools or develop one which is more suitable for user-generated videos. PEMAT does not define numerical values for “good” or “bad” scores, rather it serves as a comparison between materials. Therefore, its interpretation can be subjective. We also relied heavily on the evaluation of patient education materials from four physician evaluators, which poses the risk of evaluator bias. The calculated kappa score indicates that there was variability in the reviewers’ use of tools. However, we minimized this limitation by using the adjudication process for each item with a discrepancy which is the accepted method to achieve consensus scores. Additional video features that focus on the aesthetics, production qualities, whether the video contains a human and so on are not used in this study due to our restricted definition of video understandability following the guideline of AHRQ. In addition to patient educational guidelines, it may be also necessary to examine factors such as concordance which is the similarity, or shared identity, between physician and patients based on a demographic attribute, such as race, gender, or age (Street et al. 2008). Future work may incorporate these features to potentially improve the value of the recommended videos.

5.4 Conclusions

In this paper, we posit two novel algorithmic artefacts that encapsulate HISB: understandability and encoded information. We develop an automated, scalable, and multi-modal algorithmic solution to evaluate HISB. We combine healthcare informatics, machine learning and causal inference methods to assess the potential impact on collective engagement. Existing machine learning algorithms are not sophisticated to understand the semantic meaning of videos, especially in medical domain. Assessing the educational value of videos in domains ranging from healthcare to education still requires domain expertise to gauge the content. Future work can build on our approach to create a method of automated video retrieval that would accommodate users’ varying levels of both literacy and engagement. Future work can also build on the methods and causal inference frameworks developed in this paper to develop multi-criteria recommendations for a range of video content on topics such as education, investing and virtual communities based on the metadata and video features.

References

- Adams RJ (2010) Improving health outcomes with better patient understanding and education. *Risk Management and Healthcare Policy* 3:61–72.
- Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological science*, 17(9), 814-823.
- Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly*, 665-694.
- Alpay, L., Verhoef, J., Xie, B., Te'eni, D., & Zwetsloot-Schonk, J. H. M. (2009). Current challenge in consumer health informatics: Bridging the gap between access to information and information understanding. *Biomedical Informatics Insights*, 2, 1–10. <https://doi.org/10.4137/BII.S2223>
- Backinger, C. L., Pilsner, A. M., Augustson, E. M., Frydl, A., Phillips, T., & Rowden, J. (2011). YouTube as a source of quitting smoking information. *Tobacco Control*, 20(2), 119-122.
- Baker, L., Wagner, T. H., Singer, S., & Bundorf, M. K. (2003). Use of the Internet and e-mail for health care information: results from a national survey. *Jama*, 289(18), 2400-2406.
- Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. *Proceedings of the Annual ACM Conference on Computational Learning Theory*. 92–100.
- Bonaparte, Y. L. (2020). Meeting the Moment: Black Lives Matter, Racial Inequality, Corporate Messaging, and Rebranding. *Advertising & Society Quarterly*, 21(3).
- Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1), D267-D270.
- Dawson, A. L., Hamstra, A. A., Huff, L. S., Gamble, R. G., Howe, W., Kane, I., & Dellavalle, R. P. (2011). Online videos to promote sun safety: results of a contest. *Dermatology Reports*, 3(1).
- Dehejia RH, Wahba S (2002) Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 84(1):151–161.
- Desai T, Shariff A, Dhingra V, Minhas D, Eure M, Kats M (2013) Is content really king? An objective analysis of the public's response to medical videos on YouTube. *PLoS ONE* 8(12).
- Dolan, G., Iredale, R., Williams, R., & Ameen, J. (2004). Consumer use of the internet for health information: a survey of primary care patients. *International Journal of Consumer Studies*, 28(2), 147-153.
- Dutta-Bergman, M. J. (2004). Primary sources of health information: Comparisons in the domain of health attitudes, health cognitions, and health behaviors. *Health communication*, 16(3), 273-288.
- Eysenbach, G., & Jadad, A. R. (2001). Evidence-based patient choice and consumer health informatics in the Internet age. *Journal of medical Internet research*, 3(2), e19.
- Fage-Butler AM, Nisbeth Jensen M (2016) Medical terminology in online patient-patient communication: Evidence of high health literacy? *Health Expectations* 19(3):643–653.
- Fernandez-Llatas, C., Traver, V., Borrás-Morell, J. E., Martínez-Millana, A., & Karlsen, R. (2017). Are health videos from hospitals, health organizations, and active users available to health consumers? An analysis of diabetes health video ranking in YouTube. *Computational and mathematical methods in medicine*, 2017.
- Fox, S., & Duggan, M. (2013). Health online 2013. *Health*, 2013, 1-55.
- Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting & Electronic Media*, 58(1), 59-75.
- Gillespie, T. (2018). *Custodians of the Internet*. Yale University Press.
- Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2), 2053951720943234.
- Harbaugh AD (2013) Writing that works — Walter E. Oliu, Charles T. Brusaw, and Gerald J. Alred. *IEEE Transactions on Professional Communication* PC-23(4):202–202.
- Househ M, Borycki E, Kushniruk A (2014) Empowering patients through social media: The benefits and challenges. *Health Informatics Journal* 20(1):50–58.
- Hu W, Xie N, Li L, Zeng X, Maybank S (2011) A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man and Cybernetics Part C* 41(6):797–819.
- Hui SK, Huang Y, Suher J, Jeffrey Inman J (2013) Deconstructing the “First Moment of Truth”:

- Understanding Unplanned Consideration and Purchase Conversion Using In-Store Video Tracking. *Journal of Marketing Research* 50(4):445–462.
- Johnson R, Edwards R, Rivers A, Patil C, Walsh S (2020) Evaluating literacy levels of patient education materials for a sickle cell transition group education programme. *Health Education Journal* 79(3):253–265.
- Jordan JE, Briggs AM, Brand CA, Osborne RH (2008) Enhancing patient engagement in chronic disease self-management support initiatives in Australia: The need for an integrated approach. *Medical Journal of Australia* 189(10 SUPPL.).
- Kang SJ, Lee MS (2019) Assessing of the audiovisual patient educational materials on diabetes care with PEMAT. *Public Health Nurs.* 36(3):379–387.
- Keselman A, Logan R, Smith C (2008) Developing informatics tools and strategies for consumer-centered health communication. *Journal of American Medical Informatics Association* 15(4):473–483.
- Kettinger, W. J., & Lee, C. C. (1997). Pragmatic perspectives on the measurement of information systems service quality. *MIS quarterly*, 223-240.
- Kunze KN, Krivicich LM, Verma NN, Chahla J (2020) Quality of Online Video Resources Concerning Patient Education for the Meniscus: A YouTube-Based Quality-Control Study. *Arthroscopy - Journal of Arthroscopic and Related Surgery* 36(1):233–238.
- Kutner M, Greenberg E, Jin Y, Paulsen C (2006) The health literacy of America's adults: results from the 2003 National Assessment of Adult Literacy. *Education* 6:1–59.
- Lambert, S. D., & Loiselle, C. G. (2007). Health information—seeking behavior. *Qualitative health research*, 17(8), 1006-1019.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.
- Lazer, D. (2015). The rise of the social algorithm. *Science*, 348(6239), 1090-1091.
- Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507), 1060-1062.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444.
- Lee J, Natsev AP, Reade W, Sukthankar R, Toderici G (2019) The 2nd youtube-8M large-scale video understanding challenge. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 193–205.
- Liao H, McDermott E, Senior A (2013) Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*. 368–373.
- Liu X, Zhang B, Susarla A, Padman R (2020) Go to youtube and call me in the morning: Use of social media for chronic conditions. *MIS Quarterly: Management Information Systems* 44(1):257–283.
- Lu S, Xiao L, Ding M (2016) A video-based automated recommender (VAR) system for garments. *Marketing Science* 35(3):484–510.
- Ma F, Meng D, Xie Q, Li Z, Dong X (2017) Self-paced co-training. *34th International Conference on Machine Learning, ICML 2017*.
- Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK (2015) Healthcare information on YouTube: A systematic review. *Health Informatics Journal* 21(3):173–194.
- Manning CD, Raghavan P, Schutze H, Manning CD, Raghavan P, Schutze H (2012) Evaluation in information retrieval. *Introduction to Information Retrieval*. 139–161.
- Marcus M, Santorini, Marcinkiewicz M (1993) Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics* 19(2):313.
- McClure E, Ng J, Vitzthum K, Rudd R (2016) A mismatch between patient education materials about sickle cell disease and the literacy level of their intended audience. *Preventing Chronic Disease* 13(5).
- McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44(3), 295-322.
- Mills, A., & Todorova, N. (2016). An integrated perspective on factors influencing online health-information seeking behaviours.

- Mishra V, Dexter JP. Comparison of Readability of Official Public Health Information About COVID-19 on Websites of International Agencies and the Governments of 15 Countries. *JAMA Netw Open*. 2020;3(8):e2018033. doi:10.1001/jamanetworkopen.2020.18033
- Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C (2013) A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of Medical Internet Research* 15(4).
- Munezero M, Montero CS, Sutinen E, Pajunen J (2014) Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Trans. Affect. Comput.* 5(2):101–111.
- Munger, K. (2019). The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media+ Society*, 5(3), 2056305119859294.
- Munzert, S., BarberÁ, P., Guess, A., & Yang, J. (2020). Do online voter guides empower citizens? Evidence from a field experiment with digital trace data. *Public opinion quarterly*, 84(3), 675-698.
- O'Brien HL, Toms EG (2010) The development and evaluation of a survey to measure user engagement. *J. Am. Soc. Inf. Sci. Technol.* 61(1):50–69.
- O'Neill B, Ziebland S, Valderas J, Lupiáñez-Villanueva F (2014) User-generated online health content: A survey of internet users in the United Kingdom. *Journal of Medical Internet Research* 16(4).
- Orlikowski, W. J. (2010). The Sociomateriality of Organizational Life: Considering Technology in Management Research. *Cambridge Journal of Economics* (34)1, pp. 125-141.
- Papageorgiou CP, Oren M, Poggio T (1998) General framework for object detection. *Proceedings of the IEEE International Conference on Computer Vision*. 555–562.
- Peck CM, Mullen CA (2008) New Media, New Voices: A Complex School Public Relations and Human Resources Challenge. *Journal of School Public Relations* 29(3):401–424.
- Percheski, C., & Hargittai, E. (2011). Health information-seeking in the digital age. *Journal of American College Health*, 59(5), 379-386.
- Peters C, Castellano G, De Freitas S (2009) An exploration of user engagement in HCI. *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots, AFFINE '09, held during the ICMI-MLMI'09 Conference*.
- Pew Research Center 2021. The Future of Digital Spaces and Their Role in Democracy
- Pierce D, Cardie C (2001) Limitations of Co-training for Natural Language Learning from Large Datasets. *Proc of the Conference on Empirical Methods in Natural Language Processing (EMNLP'01)*:1–9.
- Platanios EA, Poon H, Mitchell TM, Horvitz E (2017) Estimating accuracy from unlabeled data: A probabilistic logic approach. *Advances in Neural Information Processing Systems*. 4362–4371.
- Ramagiri, R., Kannuri, N. K., Lewis, M. G., Murthy, G. V. S., & Gilbert, C. (2020). Evaluation of whether health education using video technology increases the uptake of screening for diabetic retinopathy among individuals with diabetes in a slum population in Hyderabad. *Indian journal of ophthalmology*, 68(Suppl 1), S37.
- Rooney MK, Rooney MK, Golden DW, Byun J, Lukas R V., Lukas R V., Sonabend AM, Lesniak MS, Sachdev S (2020) Evaluation of patient education materials for stereotactic radiosurgery from high-performing neurosurgery hospitals and professional societies. *Neuro-Oncology Practice* 7(1):59–67.
- Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79(387):516–524.
- Ruppert L, Køster B, Siegert AM, Cop C, Boyers L, Karimkhani C, Winston H, et al. (2017) YouTube as a source of health information: Analysis of sun protection and skin cancer prevention related issues. *Dermatology online journal* 23(1).
- Sanderson SC, Suckiel SA, Zweig M, Bottinger EP, Jabs EW, Richardson LD (2016) Development and preliminary evaluation of an online educational video about whole-genome sequencing for research participants, patients, and the general public. *Genet. Med.* 18(5):501–512.
- Salama A, Panoch J, Bandali E, Carroll A, Wiehe S, Downs S, Cain MP, Frankel R, Chan KH (2020) Consulting “Dr. YouTube”: an objective evaluation of hypospadias videos on a popular video-sharing website. *Journal of Pediatric Urology* 16(1):70.e1-70.e9.
- Shahraray B (1995) Scene change detection and content-based sampling of video sequences</title>. *Digital*

- Video Compression: Algorithms and Technologies 1995*. 2–13.
- Shoemaker SJ, Wolf MS, Brach C (2014) Development of the Patient Education Materials Assessment Tool (PEMAT): A new measure of understandability and actionability for print and audiovisual patient information. *Patient Education and Counseling* 96(3):395–403.
- Shrivastava SRBL, Shrivastava PS, Ramasamy J (2013) Role of self-care in management of diabetes mellitus. *Journal of Diabetes and Metabolic Disorders* 12(1).
- Smith, S. K. (2014, September). Decision-making and participation in bowel cancer screening: Challenges and interventions for low health literacy. In *12th International Conference on Communication in Healthcare (ICCH)*. Sept 2014. Amsterdam, the Netherlands.
- Sørensen, K., Van den Broucke, S., Fullam, J., Doyle, G., Pelikan, J., Slonska, Z., & Brand, H. (2012). Health literacy and public health: a systematic review and integration of definitions and models. *BMC public health*, 12(1), 1-13.
- Stellefson, M., Chaney, B., Ochipa, K., Chaney, D., Haider, Z., Hanik, B., ... & Bernhardt, J. M. (2014). YouTube as a source of chronic obstructive pulmonary disease patient education: a social media content analysis. *Chronic respiratory disease*, 11(2), 61-71.
- Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field.
- Street, R. L., O'Malley, K. J., Cooper, L. A., & Haidet, P. (2008). Understanding concordance in patient-physician relationships: personal and ethnic dimensions of shared identity. *The Annals of Family Medicine*, 6(3), 198-205.
- Van Den Beemt A, Thurlings M, Willems M. Towards an understanding of social media use in the classroom: a literature review. *Technol Pedagog Educ*. 2020;29:35–55.
- Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information systems research*, 11(4), 342-365.
- Williams AM, Muir KW, Rosdahl JA (2016) Readability of patient education materials in ophthalmology: A single-institution study and systematic review. *BMC Ophthalmol*. 16(1).
- Wood EB, Harrison G, Trickey A, Friesen MA, Stinson S, Rovelli E, McReynolds S, Presgrave K (2017) Evidence-Based Practice: Video-Discharge Instructions in the Pediatric Emergency Department. *Journal of Emergency Nursing* 43(4):316–321.
- Wu F, Huberman BA (2007) Novelty and collective attention. *Proceedings of the National Academy of Sciences of the United States of America* 104(45):17599–17601.
- Wyatt, J. C., & Liu, J. L. (2002). Basic concepts in medical informatics. *Journal of Epidemiology & Community Health*, 56(11), 808-812.
- Zhang M, Tang J, Zhang X, Xue X (2014) Addressing cold start in recommender systems: A semi-supervised co-training algorithm. *SIGIR 2014 - Proc. of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 73–82.
- Zhou R, Khemmarat S. and Gao L (2010, November). The impact of YouTube recommendation system on video views. In *Proc of the 10th ACM SIGCOMM conference on Internet measurement* (pp. 404-410).

Online Appendix

1. Co-training

Most successful techniques in machine learning, such as deep learning, require ground-truth labels to be given for a big training data set; in many tasks, however, it can be difficult to attain strong supervision information due to the high cost of the data-labeling process. Hence, the success of supervised learning is often limited by the quality and quantity of available data (Gennatas et al. 2020). Weakly supervised learning is concerned with the situation in which there are a small amount of labeled data, insufficient to train a good learner, while abundant unlabeled data are available. Formally, the task is to learn $f: X \rightarrow Y$ from a training data set $D = \{(x_1, y_1), \dots, (x_l, y_l), (x_{l+1}, y_{l+1}), \dots, (x_m, y_m)\}$, where there are l number of labeled training examples and $u = m - l$ number of unlabeled instances, X is the feature space, $Y = \{0,1\}$, $x_i \in X$ and $y_i \in Y$. For convenience of discussion, we call the l labeled examples “labeled data” and the u unlabeled instances “unlabeled data.”

We employ semi-supervised learning (Chapelle et al. 2009) that exploit unlabeled data in addition to labeled data to improve learning performance, where no human intervention is assumed. There are two widely used selection criteria, i.e. informativeness and representativeness (Huang et al. 2014). Informativeness measures how well an unlabeled instance helps reduce the uncertainty of a statistical model, whereas representativeness measures how well an instance helps represent the structure of input patterns. Approaches based on representativeness exploit the cluster structure of unlabeled data, usually by a clustering method (Dasgupta and Hsu 2008).

2. Medical Entity Extraction Performance

We devise a Bidirectional Long Short-term Memory (LSTM) model to extract medical terms from the user-generated video descriptions at sentence level. The video descriptions are first parsed into individual sentences. The model consists of two LSTMs that run in parallel: one on the input sequence and the other on the reverse of the input sequence. At each time step, the hidden state of the BiLSTM is the concatenation of the forward and backward hidden states. This setup allows the hidden state to capture both the past and the future information. To reduce computational complexity, we train a 300-dimensional Glove embedding

model, meaning each word is converted to a 300-dimensional semantic vector. Then the word sequence is represented as an embedding sequence, which is passed to the BiLSTM layer. Instead of using a large hidden layer size, we use 150 neurons in the BiLSTM layer to avoid over-fitting. This hidden layer size setup has also been successfully tested in other studies. The outputs of the BiLSTM layers are then processed by a CRF classifier, which predicts the semantic type of each word in the input sentence. The BiLSTM model was trained on 4,000 annotated sentences, with 1,000 sentences as the validation set (for cross validation). Another 1,000 annotated sentences were used as the test set. Table A1 below shows the performance of the BiLSTM model by category.

Table A1. Medical Term Extraction Performance By Category

Semantic Group	MetaMap+ CHV			CRF			BiLSTM RNN		
	P	R	F	P	R	F	P	R	F
All	58.94%	36.28%	44.19%	97.32%	60.09%	73.29%	87.43%	87.81%	87.32%
Body part	75.40%	55.30%	63.80%	95.20%	67.00%	78.60%	93.90%	89.80%	91.80%
Chemicals or Drugs	67.80%	35.20%	46.34%	98.90%	65.90%	79.10%	82.10%	91.50%	86.50%
Medical devices	14.90%	20.70%	17.33%	99.10%	79.00%	87.90%	94.40%	91.90%	93.10%
Medical events	65.60%	45.60%	53.80%	99.50%	38.60%	55.60%	91.90%	77.30%	84.00%
Medical professionals	57.90%	12.30%	20.29%	99.80%	91.10%	95.30%	98.60%	96.70%	97.60%
Medical procedures	12.00%	13.00%	12.48%	70.30%	31.80%	43.80%	88.20%	82.50%	85.30%

Prior work (Liu et al. 2020) assessed the volume of medical information in YouTube videos and examined users' collective engagement with them. Medical information is often conveyed in the video description text and narratives in the form of medical terms and the underlying semantic associations between them. Two graduate research associates reviewed 600 videos, the descriptions, and their caption and categorized their encoded medical information into high or low. The inter-rater reliability for classifying encoded medical information in videos is 0.92. A medical doctor consolidated the disagreement and makes the final decision. The annotated data is leveraged in the propensity score matching process in this study.

3. Search Terms for Creating a Corpus of Videos

we collected 9,873 unique videos using over 200 search terms, which will serve as the data for the current study on video understandability. These terms were collected with help from physicians and diabetes educators. These terms appear below:

Table A2. Video Search Keywords

Diabetes	diabetes supplements	insulin resistance	HemoglobinA1c	insulin infusion pump	Toujeo
diabetes causes	diabetes type	insulin resistance syndrome	high glucose	insulin injection	Tradjenta
diabetes complication	diabetes and depression	insulin secretion	iv glucose tolerance test	insulin needles	Tresiba
diabetes cure	diabetes impotence	insulin sensitive	lifescan	insulin pen injector	Welchol
diabetes depression	about diabetes	islet amyloid polypeptide	normal blood glucose levels	insulin pump	artificial sweetener
diabetes diagnosis	diabetes recipes	Islet cell antibodies	normal glucose tolerance	insulin syringe	diabetes diet
diabetes exercise	diabetes lifestyle	islets of Langerhans	oral glucose challenge	insulin therapy	diabetes prevention
diabetes high risk	diabetes medication side effects	lipodystrophy	Oral glucose tolerance test	intermediate acting insulin	diabetes prevention program
diabetes information	abnormal fat metabolism	obesity	plasma concentration	Invokana	diabetes target range
diabetes medication	abnormal glucose homeostasis	pancreas	plasma glucose	islet cell transplantation	diabetic diet
Diabetes Mellitus	abnormal insulin	pancreatic exocrine disease	post glucose tolerance test	Januvia	glycemic index
diabetes self care	abnormal insulin secretion	pancreatic islet	post prandial glucose	Lantus	physical activity
diabetes self management	adipose tissue	reduced insulin	Post-prandial glucose	Levemir	adult blindness
diabetes supplies	adult-onset diabetes	syndrome X	PPG	linagliptin	Autonomic Neuropathy
diabetes surgery	amylin	Acanthosis nigricans	urine glucose	Lispro	cellulitis
diabetes symptoms	B Cell	body mass index	Acarbose	long acting insulin	diabetes and anorexia

diabetes testing	beta cell failure	body weight	ACTOS	meglitinides	diabetes and female sexual dysfunction
diabetes treatment	beta cell of pancreatic islets	diabetes risk factor	alogliptin	metformin	diabetes and nausea
diabetes type 1 symptoms	diabetes genetic predisposition	diabetes risk factors	alpha-glucosidase inhibitors	miglitol	diabetes and reduced sexual desire women
diabetes type 2 symptoms	diabetes genetics	glycosuria	Apidra	Nateglinide	diabetes and vomiting
diets for diabetes	glucagon	hirsutism	artificial pancreas	Nesina	diabetes infections
DM	glucagon like peptide	hyperandrogenism	Aspart	Novolin	diabetes myocardial infarct
Gestational diabetes	Glucose	increased urination	Avandia	Novolin N	diabetes staph infection
gestational DM	glucose intolerance	increased water intake	bariatric surgery	Novolog	diabetes urinary tract infections
increased risk of diabetes	glucose metabolism	intermediate hyperglycemia	Biguanides	NPH insulin	diabetes UTI
insulin dependent diabetes mellitus	glucose tolerance	PCOS	bolus insulin	Onglyza	diabetic coma
juvenile diabetes	glucose transport	polycystic ovarian syndrome	Canagliflozin	oral hypoglycemic agents	diabetic dermopathy
Type 1	glucose uptake	polydipsia	carbohydrate counting	pancreatic transplantation	diabetic foot
Type 1 diabetes	glycemic control	polyuria	colesevelam	parenteral glucagon emergency	diabetic heart disease
type 1 DM	glycogenolysis	weight loss	Cycloset	pioglitazone	diabetic keto acidosis
Type 2	hepatic gluconeogenesis	weight gain	Detemir	Pramlintide	diabetic mononeuropathy
Type 2 diabetes	HLA complex	blood glucose	diabetes clinical trials	prandin	diabetic nephropathy
Type 2 DM	hyperglycemia	blood pressure	Diabetes medicine	precose	diabetic neuropathy

type I diabetes	hyperinsulinemia	continuous glucose monitoring systems	DPP-4 Inhibitors	rapid acting insulin	diabetic peripheral vascular disease
type I DM	IAPP	c-peptide test	FlexPen	Regular insulin	diabetic polyneuropathy
Type II	IFG	fasting glucose	Glargine	Repaglinide	diabetic retinopathy
Type II diabetes	IGT	fasting plasma glucose	Glucophage	Rosiglitazone	diabetic skin spots
Type II DM	impaired fasting glucose	fingerstick glucose test	Glulisine	Ryzodeg	End stage renal disease
www diabetes org	impaired glucose tolerance	FPG	Glumetza	saxagliptin	ESRD
diabetes lose weight	impaired glycemic control	glucometers	glyset	SGLT2 inhibitors	foot ulcers
borderline diabetes	impaired insulin secretion	glucose meter	Humalog	short acting insulin	gangrene
diabetes kit	increased glucose production	glucose monitor	Humulin	Sitagliptin	gastroparesis
onset diabetes	insulin	glucose test strip	Humulin N	starlix	hypoglycemia
diabetes magazines	insulin deficiency	glucose tolerance test	inhaled insulin	SymLinPen	insulin reaction
herbal treatment for diabetes	insulin receptor	HbA1c	injection site rotation	thiazolidinediones	insulin shock
neuropathy	nocturnal diarrhea	peripheral neuropathy			

4. Video Engagement Measures

Prior studies of individual user engagement have posited factors associated with user engagement such as attention, affect, aesthetics, novelty, motivation, interest, feedback and control (O'Brien and Toms 2010). Feedback, affect, motivation, and interest are measurable with YouTube metadata. We consider the number of likes, number of dislikes, and number of comments as measures of users' feedback. Affect has been conceptualized in the literature as encompassing feelings and sentiments (Fleckenstein 1991), or the

positive and negative evaluations of an object, behavior, or idea with intensity and activity dimensions (Munezero et al. 2014). We therefore examine the sentiment in comments with Valence Aware Dictionary and Sentiment Reasoner (VADER)¹. We use VADER to classify the sentiment polarity of each comment and compute the total number of positive, negative, and neutral comments for each video. We also consider two measures of relevance building on Liu et al. (2020) to assess motivation; the cosine similarity between search keywords and video title, and the cosine similarity between search keywords and video description. Prior research has used document relevance score to evaluate the quality of comments (Diakopoulos 2015) while the video description relevance score can be considered as a reflection of user interest. The result of principal component analysis is available in Table A4.

Table A3. Descriptive Statistics of Collective Engagement about the Videos

Numeric Variables	Min	Q1	Median	Mean	Q3	Max
# of views	0	228	2,118	197,845	23,988	369,970,829
# of dislikes	0	0	1	95.51	8	212,994
# of neutral comments	0	0	0	4.0	2	100
sim(comment, description)	0	0	0	0.008	0.005	0.15
# of likes	0	1	9	1,538	116	2,422,615
# of comments	0	0	1	143	15	319,810
sim(keyword, title)	0	0.08	0.26	0.27	0.41	1
sim(keyword, description)	0	0	0.04	0.09	0.14	1
sim(comment, title)	0	0	0	0.005	0.003	0.14
# of negative comments	0	0	0	2.65	1	73
# of positive comments	0	0	3	6.67	5	93

Table A4. Principal Component Loadings

Component & Loadings	1	2	3
	Non engagement	Sustained Attention	Selective Attention
sim(keyword, title)		0.159	0.476
sim(keyword, description)		0.142	0.679
sim(comment, description)	-0.303	-0.225	-0.116
sim(comment, title)	-0.307	0.179	
Total number of words in comments	-0.312	0.302	-0.177
Unique number of words in comments	-0.320	0.284	-0.166
# of comment	-0.329	0.125	

¹ <https://github.com/cjhutto/vaderSentiment>

# of positive comment	-0.308	0.151	
# of negative comments	-0.299	0.194	
# of neutral comment	-0.250	-0.214	0.178
# of views	-0.267	0.362	-0.272
# of likes	-0.304	0.297	-0.214
# of dislikes	-0.299	0.282	-0.240

5. Weighted Precision, Recall, and F₁

c is a class and $Classes$ is a set of all possible classes.

$$Weight(c) = Prevalance(c) = \frac{Actual(c)}{T}, \text{ percentage of instances in class } c$$

$$Weighted\ Precision = \sum_{c \in Classes} Weight(c) \times Precision(c)$$

$$Weighted\ Recall = \sum_{c \in Classes} Weight(c) \times Recall(c)$$

$$Weighted\ F_1 = \sum_{c \in Classes} Weight(c) \times F_1(c)$$

6. Descriptive Statistics of Features for Video Understandability Classification

In total, we collected 9,873 videos using the search keywords identified by a medical expert. Among these videos, 8,963 videos have descriptions, 8,719 have narratives, 4,327 of them have text embedded in the videos. We applied text and video analytics techniques Table A5 reported the descriptive statistics of features all the videos in our data collection.

Table A5. Descriptive Statistics of Features for Video Understandability Classification

View	Variable Name	N	Yes			No		
Video meta data	Has title	9873	9873			0		
	Has description	9873	8963			910		
	Has tags	9873	9873			0		
			Min	Q1	Mean	Median	Q3	Max
	Description readability	8963	0	6.6	9.9	9.4	13.1	18
	Active verb count	8963	0	1	16.7	4	18	170
	Summary word count	8963	0	0	0.02	0	0	11
	Transition word count	8963	0	0	2.5	1	3	39
	Video duration	9873	1	67	387.6	168	388	26156
	Description word count	8963	1	18	153.8	58	197	1118
	Sentence count	8963	1	1	6.9	2	7	61
	Description unique word count	8963	1	11	126.2	38	108	388

	Description medical term Count	8963	0	0	3.7	3	7	125
Video content	Narrative readability	8719	0	7.8	10.5	10	14	19
	Active verb count	8719	0	1	20.2	6	25	205
	Summary word count	8719	0	0	0.1	0	2	25
	Transition word count	8719	0	0	3.5	2	5	47
	Transcription confidence	8719	0	0.37	0.73	0.65	0.84	0.99
	Text detection confidence	4327	0	0.24	0.63	0.57	0.74	0.99
	Shot count	9873	1	4	12.6	11	28	141
	Transcript word count	8719	1	131	526	312	754	49312
	Transcript unique word count	8719	1	26	215.7	107	389	614
	Transcript sentence count	8719	1	6	32.6	24	67	4123
	Transcript medical term count	8719	0	5	17.4	13	35	135
	Video object count	9873	1	11	43.6	38	79	127

We examine the correlations of the numeric variables among the videos with description, narratives, and text on the screen. The correlation analysis results are reported in Table A6.

Table A6. Correlation Analysis of Features for Understandability Classification

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1 Description readability	1																				
2 Description active verb count	0.07	1																			
3 Description summary word count	0.03	0.18	1																		
4 Description transition word count	0.07	0.87	0.21	1																	
5 Video duration	0.03	-0.04	0.02	-0.03	1																
6 Description word count	0.1	0.95	0.21	0.86	-0.02	1															
7 Description sentence count	0.04	0.36	0.07	0.34	-0.01	0.38	1														
8 Description unique word count	0.25	0.38	0.08	0.37	0	0.41	0.91	1													
9 Description medical term Count	-0.04	0.03	0.04	0	0.1	0.17	0.08	0.12	1												
10 Narrative readability	0.33	0.07	0.01	0.08	0.09	0.14	0.13	0.04	-0.05	1											
11 Narrative active verb count	0.02	0.09	0.02	-0.01	0.12	0.09	0.1	0.07	0.06	0.12	1										
12 Narrative summary word count	0.04	0.03	0.08	0.06	0.08	0.03	0.06	0.08	0.04	0.08	0.07	1									
13 Narrative transition word count	0.03	0.02	0.03	0.04	0.14	0.07	0.08	0.04	0.07	0.15	0.09	0.15	1								
14 Transcription confidence	0.06	0.08	0.06	0.02	0.06	0.05	0.04	0.07	0	0.18	0.05	0.06	0.04	1							
15 Text detection confidence	0.07	0.01	0.07	0.05	0.04	0.04	0.07	0.08	0.04	0.07	0.04	0.04	0.01	0.05	1						
16 Shot count	-0.05	0	-0.04	0.05	0.15	-0.06	-0.03	0.04	0.05	-0.12	-0.07	-0.08	0.06	-0.04	-0.1	1					
17 Transcript word count	0.05	0.06	0.01	0.07	0.31	0.07	0.1	0.08	0.04	-0.07	0.15	0.12	0.22	0.08	0.05	0.1	1				
18 Transcript unique word count	0.01	0.04	0	0.03	0.21	0.05	0.05	0.1	0.06	-0.11	0.17	0.08	0.12	0.06	0.04	0.18	0.3	1			
19 Transcript sentence count	0.06	0.07	0.03	0.01	0.27	0.11	0.09	0.08	0.13	-0.06	0.24	0.36	0.14	0.12	0.03	0.28	0.86	0.89	1		
20 Transcript medical term count	-0.01	0.02	0.05	0.04	0.13	0.07	0.06	0.12	0.16	-0.13	0.08	0.18	0.15	-0.09	0.07	0.13	0.27	0.22	0.67	1	
21 Video object count	0.03	0.05	0.07	0.05	0.26	0.06	0.04	0.09	0.12	-0.08	0.14	0.21	0.19	0.03	0.12	0.37	0.09	0.24	0.34	0.16	1

We select videos with both descriptions and narratives to evaluate the co-training approach. In total, there are 8,616 videos containing both descriptions and narratives. 600 videos are selected to initialize the co-training process. 100 videos are reserved for evaluation. Among them, 714 videos have obtained classification results with low confidence. These videos are in general shorter videos with high variances in video features and meta data features. 305 videos obtained inconsistent predictions with high confidence and required human labeling. 6,897 videos yield consistent predictions with high confidence in the co-training process. The descriptive statistics of these three types of videos are reported in Tables A7, A8, and A9.

Table A7. Descriptive Statistics of Videos with Low Confidence Predictions

View	Variable Name	N	Yes			No		
Video meta data	Has title	714	714			0		
	Has description	714	714			0		
	Has tags	714	714			0		
			Min	Q1	Mean	Median	Q3	Max
	Description readability	714	1	5.5	8.9	9.2	14.6	18
	Active verb count	714	0	1	7.8	4	13	43
	Summary word count	714	0	0	0.01	0	1	5
	Transition word count	714	0	0	1.7	1	3	13
	Video duration	714	1	66	137.7	130	370	784
	Description word count	714	1	12	35.3	29	69	230
	Sentence count	714	1	1	4.4	3	7	19
	Description unique word count	714	1	8	16.7	15	31	137
	Description medical term Count	714	0	0	2.3	3	6	12
Video content	Narrative readability	714	0	6.3	9.6	10.7	15	18
	Active verb count	714	0	1	7.7	5	13	19
	Summary word count	714	0	0	0.1	0	2	6
	Transition word count	714	0	0	2.4	2	4	7
	Transcription confidence	714	0	0.32	0.73	0.67	0.78	0.93
	Text detection confidence	714	0	0.31	0.58	0.54	0.68	0.89
	Shot count	714	1	3	12.6	6	11	15
	Transcript word count	714	1	117	265.3	204	569	1105
	Transcript unique word count	714	1	74	149.7	127	352	390
	Transcript sentence count	714	1	7	21.4	16	33	89
	Transcript medical term count	714	0	3	10.3	13	27	48
	Video object count	714	1	7	16.4	12	37	46

Table A8. Descriptive Statistics of Videos with High Confidence Inconsistent Predictions

View	Variable Name	N	Yes			No		
Video meta data	Has title	305	305			0		
	Has description	305	305			0		
	Has tags	305	305			0		
			Min	Q1	Mean	Median	Q3	Max
	Description readability	305	0	5.8	9.5	9	12.8	18
	Active verb count	305	0	1	8.4	6	16	27
	Summary word count	305	0	0	0.02	0	0	7
	Transition word count	305	0	0	1.7	1	4	13
	Video duration	305	37	155	530.4	315	580	6570
	Description word count	305	1	65	253.4	166	370	1008
	Sentence count	305	1	5	21.7	12	29	48
	Description unique word count	305	1	50	133.4	111	198	418
	Description medical term Count	305	0	1	5.7	4	6	37
Video content	Narrative readability	305	0	7.8	10.5	10	14	18
	Active verb count	305	0	1	20.2	6	25	205
	Summary word count	305	0	0	0.1	0	2	25
	Transition word count	305	0	0	3.5	2	5	47
	Transcription confidence	305	0	0.37	0.73	0.65	0.84	0.99
	Text detection confidence	305	0	0.24	0.63	0.57	0.74	0.99
	Shot count	305	1	4	12.6	11	28	141
	Transcript word count	305	1	203	847.3	530	984	10349
	Transcript unique word count	305	1	137	368.4	289	394	472
	Transcript sentence count	305	1	14	59.7	39	76	890
	Transcript medical term count	305	0	8	18	15	22	47
	Video object count	305	1	10	45.4	38	71	103

Table A9. Descriptive Statistics of Videos with High Confidence Consistent Predictions

View	Variable Name	N	Yes			No		
Video meta data	Has title	6897	6897			0		
	Has description	6897	6897			0		
	Has tags	6897	6897			0		
			Min	Q1	Mean	Median	Q3	Max
	Description readability	6897	0	6.6	10.1	9.4	13.1	18
	Active verb count	6897	0	1	16.3	4	18	170
	Summary word count	6897	0	0	0.02	0	0	11
	Transition word count	6897	0	0	2.6	1	3	39

	Video duration	6897	1	67	379.4	168	388	26156
	Description word count	6897	1	18	148.7	58	197	1118
	Sentence count	6897	1	1	6.4	2	7	61
	Description unique word count	6897	1	11	123.3	38	108	388
	Description medical term Count	6897	0	0	3.7	3	7	125
Video content	Narrative readability	6897	0	8.1	9.8	10.2	14.5	18
	Active verb count	6897	0	1	21.2	7	23	205
	Summary word count	6897	0	0	0.1	0	2	25
	Transition word count	6897	0	0	3.3	2	5	47
	Transcription confidence	6897	0	0.34	0.71	0.69	0.79	0.99
	Text detection confidence	6897	0	0.25	0.67	0.61	0.71	0.99
	Shot count	6897	1	4	13.4	12	26	141
	Transcript word count	6897	1	129	510.3	330	747	49312
	Transcript unique word count	6897	1	24	207.3	117	374	614
	Transcript sentence count	6897	1	7	33	24	66	4123
	Transcript medical term count	6897	0	6	16.4	14	36	135
	Video object count	6897	1	10	45.3	35	81	127

References

- Chapelle O, Scholkopf B, Zien, Eds. A (2009) Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]. *IEEE Trans. Neural Networks* 20(3):542–542.
- Dasgupta S, Hsu D (2008) Hierarchical sampling for active learning. *Proc. 25th Int. Conf. Mach. Learn.* 208–215.
- Desai T, Shariff A, Dhingra V, Minhas D, Eure M, Kats M (2013) Is content really king? An objective analysis of the public’s response to medical videos on YouTube. *PLoS One* 8(12).
- Diakopoulos N (2015) The editor’s eye: Curation and comment relevance on the New York Times. *CSCW 2015 - Proc. 2015 ACM Int. Conf. Comput. Coop. Work Soc. Comput.* 1153–1157.
- Fleckenstein KS (1991) Defining Affect in Relation to Cognition: A Response to Susan McLeod. *Defin. Affect Relat. to Cogn. A Response to Susan McLeod.* 11(2).
- Gennatas ED, Friedman JH, Ungar LH, Pirracchio R, Eaton E, Reichmann LG, Interian Y, et al. (2020) Expert-augmented machine learning. *Proc. Natl. Acad. Sci. U. S. A.* 117(9):4571–4577.
- Huang SJ, Jin R, Zhou ZH (2014) Active Learning by Querying Informative and Representative Examples. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(10):1936–1949.
- Johnson R, Edwards R, Rivers A, Patil C, Walsh S (2020) Evaluating literacy levels of patient education materials for a sickle cell transition group education programme. *Health Educ. J.* 79(3):253–265.
- Kang SJ, Lee MS (2019) Assessing of the audiovisual patient educational materials on diabetes care with PEMAT. *Public Health Nurs.* 36(3):379–387.
- Kunze KN, Krivicich LM, Verma NN, Chahla J (2020) Quality of Online Video Resources Concerning Patient Education for the Meniscus: A YouTube-Based Quality-Control Study. *Arthrosc. - J. Arthrosc. Relat. Surg.* 36(1):233–238.
- McClure E, Ng J, Vitzthum K, Rudd R (2016) A mismatch between patient education materials about sickle cell disease and the literacy level of their intended audience. *Prev. Chronic Dis.* 13(5).
- Munezero M, Montero CS, Sutinen E, Pajunen J (2014) Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Trans. Affect. Comput.* 5(2):101–111.
- O’Brien HL, Toms EG (2010) The development and evaluation of a survey to measure user engagement.

- J. Am. Soc. Inf. Sci. Technol.* 61(1):50–69.
- Rooney MK, Rooney MK, Golden DW, Byun J, , Lukas RV, Sonabend AM, Lesniak MS, Sachdev S (2020) Evaluation of patient education materials for stereotactic radiosurgery from high-performing neurosurgery hospitals and professional societies. *Neuro-Oncology Pract.* 7(1):59–67.
- Salama A, Panoch J, Bandali E, Carroll A, Wiehe S, Downs S, Cain MP, Frankel R, Chan KH (2020) Consulting “Dr. YouTube”: an objective evaluation of hypospadias videos on a popular video-sharing website. *J. Pediatr. Urol.* 16(1):70.e1-70.e9.
- Sanderson SC, Suckiel SA, Zweig M, Bottinger EP, Jabs EW, Richardson LD (2016) Development and preliminary evaluation of an online educational video about whole-genome sequencing for research participants, patients, and the general public. *Genet. Med.* 18(5):501–512.
- Williams AM, Muir KW, Rosdahl JA (2016) Readability of patient education materials in ophthalmology: A single-institution study and systematic review. *BMC Ophthalmol.* 16(1).