AI for Social Good: A Human-in-the-Loop Augmented Intelligence Video Understandability Assessment to Promote Health Literacy

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

December 3, 2022

An estimated 74% of adults in the United States access the Internet, with up to 80% of them looking for health information online. However, only 12% of US adults are assessed to be proficient in health literacy to meaningfully interpret health information. Billions of individuals worldwide access healthcare information on social media platforms, where they may be exposed to misleading, harmful, or irrelevant information. Such lack of gatekeeping by social media platforms has had serious adverse effects on public health. Organizations such as the US National Academy of Medicine (NAM) and the World Health Organization (WHO) have highlighted the importance of prescriptive guidelines for technology-enabled solutions for identifying credible sources of health information on social media. 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 content moderation challenges and prescriptive interventions that harness human intelligence into machine learning. In 2021, the US Surgeon General's advisory identified that social media platforms urgently need to amplify high-quality health 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 the understandability of videos from a patient education perspective. Extracting video features and metadata from YouTube, we develop a human-in-the-loop assessment that explicitly focuses on the human algorithm interaction 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 engagement with videos. Implications for research and practice are discussed.

Acknowledgments:

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, MIT, MSU Outreach, University of Maryland, NYU, Temple University, Texas A&M University and the University of Illinois at Chicago for comments on earlier drafts of the paper. We thank Ernestina Bioh, Sreeja Nair, Mukund Nakhate, and Namrata Navge for their research assistance. We also acknowledge funding from the National Library of Medicine (NIH Grant R01LM013443).

1. Introduction

The World Health Organization (WHO) defines health literacy as "*The cognitive and social skills which determine the motivation and ability of individuals to gain access to understand and use information in ways which promote and maintain good health*." According to the National Assessment of Adult Literacy (Kutner et al. 2006), only 12% are considered proficient in their ability to meaningfully interpret health information. Recently, a large-scale study assessing Covid-19 related health communications from state and federal agencies found that most information, including those from CDC and federal agencies, exceeded recommended reading levels (Mishra & Dexter 2020). In the absence of evidence-based interventions, health consumers may be steered towards dubious medical advice or lack awareness of how to integrate medical advice obtained from social media into their daily routines.

Patients traditionally received health directives from clinicians through verbal advice and printed pamphlets. However, adherence to such sources demands a high level of participation and engagement (Jordan et al. 2008). Educational videos provide both visual and auditory information, have the potential to reach many people, and provide a consistent message in a cost-effective manner (Dahodwala et al. 2018). Research has shown that video-based health education, when presented in an understandable way, is easier to absorb for patients as compared to other patient educational methods (Abed et al. 2014). However, it is prohibitively expensive for healthcare organizations to provide such content on a spectrum of health conditions. Further, patients diagnosed with chronic conditions often need advice on diet, nutrition, exercise, and lifestyle choices in addition to medical advice. The Internet has reduced much of the information asymmetry between healthcare practitioners and consumers by providing multiple avenues whereby patients can educate themselves with user-generated content.

The plethora of user-generated healthcare content on platforms such as YouTube offers the potential to provide relevant and actionable healthcare advice. However, recommending patient education materials on digital platforms is significantly more challenging than, say, recommending news articles or books. The acceptance and usage of recommended videos need to be assessed on multiple criteria, such as understandability, video production quality, and credibility of the video and content creators, among others. While several patient education guidelines promote understandability and clear communication (Baur & Prue 2014; Shoemaker et al. 2014), it is not clear whether digital platforms and health providers are following such guidelines for information dissemination online. The US National Academy of Medicine (NAM) and the World Health Organization (WHO) have called for prescriptive guidelines for identifying credible sources of health information on social media.

With patients regularly turning to YouTube for health information and advice, pointing patients toward understandable video materials is one mechanism to bridge the vast divide in health literacy and to enhance a patient's ability to self-manage their medical conditions. Agency for Healthcare Research and Quality (AHRQ) defines patient educational materials as understandable when consumers of diverse backgrounds and varying levels of health literacy can process and explain key messages. The Patient Education Material Assessment Tool (PEMAT) is a systematic approach developed by the Agency of Healthcare Research and Quality (AHRQ) to evaluate and compare the understandability and actionability of patient educational materials. It is designed to be used by healthcare professionals and help determine whether patients will be able to understand and act on information. PEMAT is the only guideline that includes a measure of objective assessment of audiovisual (A/V) materials (Vishnevetsky et al. 2018) and therefore has been widely adopted in evaluating patient educational materials in video format.

With the vast number of user-generated videos available online, we cannot rely on healthcare professionals to manually annotate all the videos. Machine learning based approaches can reduce the amount of effort from healthcare professionals and perform the evaluation in a scalable manner. We envisage the role of human-in-the-loop artificial intelligence as that of elevating input from healthcare professionals from manual annotators to that where human judgment and domain knowledge from experts can guide the development of algorithms for the delivery of healthcare information on social media platforms. We use a combination of human and algorithmic intelligence in an application area of AI for social good by examining evidence-based guidelines for retrieving health information on social media. Extracting video features and metadata from YouTube, we employ co-training methods to assess the understandability of a video according to PEMAT guidelines. The co-training approach employed in this paper allows us to implement disagreement-based semi-supervised methods and human intelligence into a unified framework.

We make three contributions to the literature in a multi-disciplinary area of digital health literacy and human-algorithm connection. Our first contribution is in using AI methods for social good to posit the understandability of healthcare information as an artifact that bridges a gap between access to information and its understandability (Laurence et al. 2009). Most existing work in health communication and patient education has not examined how health communications need to evolve in the era of algorithmic engagement. We have engaged substantially with agencies and experts in health communication and public health in creating an algorithmic framework that incorporates human intelligence in building an evidence-based guideline that can address challenges in health literacy. Our second contribution is the development of a human-in-the-loop augmented intelligence method that incorporates human judgment and expertise into algorithms.

In characterizing the search for cognitively demanding information on social media, we combine human cognitive capabilities into artificial intelligence (AI) systems by developing an augmented intelligence approach that combines human intelligence with AI. Our approach utilizes inputs from domain experts and PEMAT-based patient education constructs combined with machine learning methods to build an automated tool that analyzes the understandability of YouTube videos from the perspective of patient education. The third contribution is to enable a better understanding of how patients understand healthcare information by assessing the impact of the understandability of user collective engagement on the video. Our findings can offer policy implications for the utilization of healthcare resources and the quality of delivered care.

2. Literature Review and Research Questions

2.1 AI for Social Good

Advances in machine learning and artificial intelligence offer opportunities to develop better tools and solutions to help address some of the most pressing issues in the world and deliver positive social impact in accordance with the priorities outlined in the United Nation's 17 Sustainable Development Goals (SDGs)[1] (Tomašev et al. 2020). Initiatives relying on artificial intelligence (AI) to deliver socially beneficial outcomes – AI for social good (AI4SG) are on the rise. AI4SG facilitates the attainment of socially good outcomes that were previously unfeasible, unaffordable, or unachievable with machine learning and artificial intelligence methods (Cowls et al. 2021). Cowls et al. (2021) formally defined AI4SG as "the design, development, and deployment of AI systems in ways that help to (i) prevent, mitigate and/or resolve problems adversely affecting human life and/or the wellbeing of the world, and/or (ii) enable socially preferable or environmentally sustainable developments, while (iii) not introducing new forms of harm and/or

[1] https://sdgs.un.org/goals

amplifying existing disparities and inequalities. Despite the apparent appeal of using the latest ML methods, which may require a large amount of high-quality training data, AI4SG projects may sometimes benefit from simpler solutions, aiming to solve the problem at hand with minimum overall complexity. Such solutions tend to be faster to implement, easier to maintain, interpret and justify—and are sometimes sufficient to solve valuable practical problems (Tomašev et al. 2020). Ensuring healthy living and promoting well-being for all at all ages are among the SDGs set by the UN. AI can provide much-needed support to improve population health by enhancing the population's health literacy level and providing high-quality health information effectively and efficiently.

2.2 Health Literacy as a Societal Challenge

Health literacy is well recognized as a challenge for both individual and public health, with many adults lacking the requisite skills to engage successfully in the management of their healthcare (Mackert et al. 2016, Moorehead et al. 2013). Providing access to high-quality health information and patient education materials is critical to patient empowerment and building societal resilience. It is estimated that 74.4% of the population in the U.S. seek health information online (Rutten et al. 2019). The availability of health information online is thought to have a positive influence on health literacy. However, the actual impact is more nuanced. For health consumers engaging in searches for health information on digital media platforms, health literacy divides can be exacerbated both by their own lack of knowledge and by algorithmic recommendations, with results that disproportionately impact minorities and low health literacy users (Agarwal et al. 2020). Health consumers with higher health literacy levels were found to seek online health information more frequently than those with lower literacy levels. Besides, consumers with higher health literacy reported fewer difficulties (e.g., less effort to get the information, less concern about

health information quality, and fewer difficulties in understanding health information) (Manganello et al. 2017). To bridge the gap in health literacy and improve health equity, it is important to promote understandable and high-quality health information online.

Prior studies have employed domain experts, such as health professionals, to evaluate the quality of online medical information (C.L. et al. 2011, Dawson et al. 2011). Studies have examined the understandability of 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 the Center for Disease Control and Prevention (CDC)[2] (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) (Salama et al. 2020), Global Quality Score (Kunze et al. 2020), and readability indices (Rooney et al. 2020). Table 1 below summarizes this literature.

Table 1. Summary of studies evaluating online health information

Study	**Data**	**Assessment Tool**	**Focus**	**Finding**
Kang and Lee 2019	85 Videos from hospital websites	PEMAT	Understandability, actionability, and usefulness	The average understandability rating is 49.5% and the actionability rating is 31.4%. The average usefulness score is 4.3 out of 7 point scale.
McClure et al. 2016	9 print and 4 online patient education materials	SMOG, PMOSE/IKIRSCH, PEMAT, CDC Clear Communication Index	The health literacy level of the publicly available patient education materials	Reading levels of available patient education materials exceed the documented average literacy level of the US adult population.
Johnson et al. 2020	4 written patient education materials for sickle cell disease	(1) the Flesch Reading Ease Formula, (2) the Flesch–Kincaid Reading Tool, (3) SMOG Readability Formula, (4) the PEMAT, (5) CDC	Readability, grade level, understandability, and actionability	Literacy levels of the patient education materials were higher than recommended standards.

[2] https://www.cdc.gov/ccindex/index.html

		Clear Communication Index, and (6) PMOSE/IKIRSCH tool.		
Rooney et al. 2020	54 patient education materials from high-performing neurosurgery hospitals and professional societies	6 readability indices	Readability	Publicly available online patient educational materials for stereotactic radiosurgery are written at reading levels above national recommendations. Furthermore, many lack information identified as important by patients.
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 and knowledge increased	79% reported the video was easy to understand, satisfaction scores were high, and knowledge increased significantly.
Salama et al. 2020	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.

Readability, content organization, and presentation are critical to healthcare consumers. These factors impact how patients consume educational materials and whether the medical information can be delivered effectively. There have been a host of studies assessing the readability, suitability or comprehensibility of patient education materials, and the evidence is clear and consistent that most education materials are too complicated for patients to comprehend (Sørensen et al. 2012), especially for those with low health literacy (Smith et al. 2014).

Prior studies have relied on the judgment of domain experts such as health professionals to evaluate medical information online (Backinger et al. 2011; Dawson et al. 2011). Content rated by

an expert (such as medical or health professional staff) is the most common approach to assessing videos focused on health education. Health and medical websites are increasingly encouraged to apply for quality certificate assessments as proof of evidence that they are reliable sources of information. However, as the volume of online health information grows exponentially, using expert evaluation is not a sustainable long-term solution.

2.3 Social Media Analytics for Social Good

Social media analytics, enabled by machine learning and artificial intelligence, have yielded a rich understanding of societal outcomes ranging from health, business, and governance issues (Anderson et al. 2021, Van Den Beemt et al. 2020, Liu et al. 2020, Stier et al. 2020). Research in business disciplines has concentrated on brand-related activities on social media (Ghiassi et al. 2016, Gunarathne et al. 2017, Qiu et al. 2015, Xie and Lee 2015), opinion formation (Munger and Teredesai, Munzert et al. 2021), content contribution (Tang et al. 2012) and diffusion (Susarla et al. 2016, Yoo et al. 2019).

Social computing that facilitates collective action and user-centric interaction has led to enormous changes in the organization of content production industries (Dellarocas et al. 2013). Prior literature on User-Generated Content (UGC) has examined motives for participation and the economic value of such content (Ghose and Ipeirotis 2011). However, the difference is that visual social media is characterized by different patterns of user engagement (Liu et al. 2020) and different pathways by which influence and susceptibility operate (Susarla et al. 2016). There is limited inquiry into evidence-based research on how to make social media relevant and usable for our daily lives, especially in the context of health and wellness.

While people have relied on traditional media (such as the library, books, brochures, and magazines) or healthcare professionals as their primary source of health information (Baker et al.

2003, Dolan et al. 2004, Dutta-Bergman 2004), social media has caused a sea change in providing health information for improving health literacy. The rise of user-generated medical information disseminated through YouTube, which is the largest video-sharing social media platform, can potentially bridge this gap by providing information in a rich, visual format that may be easier to comprehend and adhere to (Backinger et al. 2011; Chatzopoulou et al. 2010). YouTube hosts more than 100 million videos providing information on the pathogenesis, diagnosis, treatments, and prevention of various medical conditions (Madathil et al. 2015). For patients who need complex medical information, healthcare advice in a video format may make it more convenient, understandable, and actionable and improve the efficiency of care. While this plethora of user-generated content can be leveraged by patients to improve health literacy and adherence to treatments, criticisms of social media use in healthcare have also arisen, and the prevalence of misinformation is a risk. Social media analytics can offer a path toward patient education and empowerment and improved health literacy of the population by providing clinicians and patients with the ability to easily retrieve understandable and relevant video-based content.

2.3 Augmented Intelligence and Human in the Loop Training Methods

Evaluation of healthcare video content requires domain expertise. Given the amount of user-generated video content available, it is not feasible to generate a large, labeled dataset for typical standalone machine learning methods. Due to the high level of uncertainty and criticality in healthcare and problem diversity, our objective is to introduce human-like cognitive capabilities into artificial intelligence (AI) systems to develop an augmented intelligence approach. While artificial intelligence, machine learning, and other automation technologies have made significant advances in recent years, many important problems are often solved through the collaboration of human beings and machines together (Jain et al. 2018). Applications of augmented intelligence

and human-in-the-loop methods have started to emerge in a number of domains, such as cybersecurity, counterterrorism, healthcare, fintech, and among others. Augmented intelligence can be divided into two approaches: one is human-in-the-loop augmented intelligence with human-computer collaboration, and the other is cognitive computing based augmented intelligence, in which a cognitive model is embedded in the machine learning system (Zheng et al. 2017). Human-in-the-loop augmented intelligence is defined as an intelligent model that requires human interaction (Williams et al. 2018). In this type of intelligence system, the human is always part of the system and consequently influences the outcome in such a way that the human gives further judgment if a low-confidence result is given by an algorithm. This approach readily allows us to address problems and requirements that may not be easily trained or classified by machine learning (Zheng et al. 2017). Cognitive computing-based augmented intelligence refers to new software and/or hardware that mimics the function of the human brain and improve a computer's capability of perception, reasoning, and decision-making (Kelly 2015). The goal is to create more accurate models that simulate how the human brain/mind senses, reasons, and responds to stimuli.

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). In machine learning, unlabeled data is often substantially cheaper and more plentiful than labeled data. YouTube contains thousands of healthcare related videos. However, annotating these videos requires significant human effort. Therefore, 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 a stopping condition is met. Intuitively, each example contains two "views," and each view contains sufficient 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. This approach has 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 process is viewed as a combinative label propagation over two views. Obtaining labels can be expensive or time-consuming because of the involvement of human experts in this research context. Most learning tasks can be made more efficient in terms of labeling cost by intelligently integrating specific unlabeled instances to be labeled by experts.

2.4 Research Question

Despite the growing attention of policymakers and healthcare providers, it is evident that health educational materials remain too complicated for patients to comprehend. Evaluations of health information online are an urgent issue and are amplified when considering that 80% of Americans search for health information online, and only 12% have proficient health literacy to correctly interpret and use it. In this study, we seek to address the following research question: how can we design a scalable approach to combine human cognitive power and machine learning methods and evaluate understandability, a patient education construct from literature, in user-generated videos?

In this study, we propose 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. A novel feature of our work is in using a human-in-the-loop augmented intelligence tool to develop an AI for social good application: assessing the understandability of medical information in user-generated content.

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 2020[3]). 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, multimedia, 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 1 illustrates our approach, which consists of five components: video collection; video analysis; text analysis; co-training approach for video understandability assessment; and understandability evaluation.

[3] https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

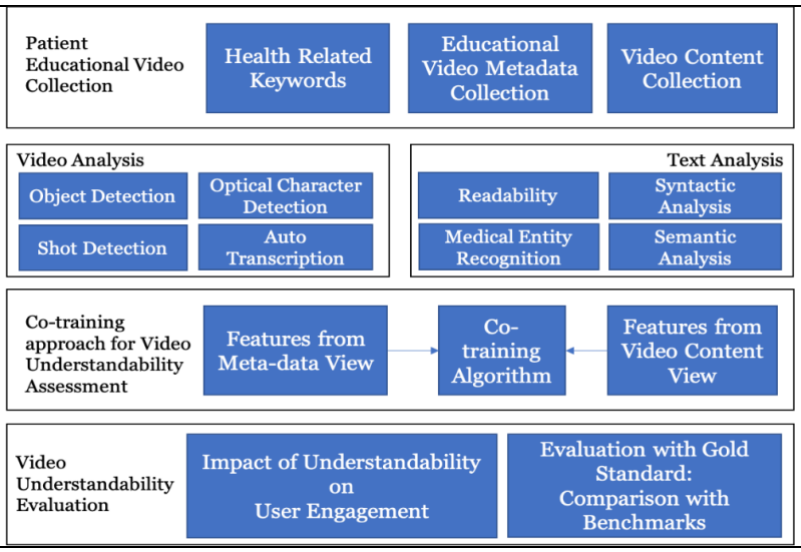


Figure 1. The Research Framework to Assess Healthcare Video Understandability

Combining physician input with a literature review helped us identify 235 search keywords related to diabetes patient education. 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. These keywords are available in Table A1 in the Online Appendix. 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 available in Figure A1 in the online Appendix. Attributes related to video snippets 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 serve as the data for the current study on video understandability.

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-quality content or inaccurate videos produced by both individuals and organizations. This diversity introduces a challenge for annotating the videos. We rely on experts' consensus perspective to evaluate understandability based on the Patient Educational Material Assessment Tool (PEMAT) for audio and video materials (Shoemaker et al. 2014), as it is the only systematic method developed to assess video content. Table 2 lists our adaptation of PEMAT. PEMAT focuses 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.

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

Table 2. Patient Educational Material Assessment Tool for Audio and Video Materials

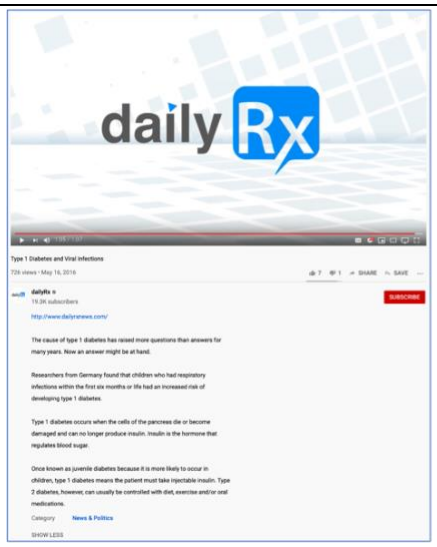
Patient Educational Material Assessment Tool – Video Understandability		
Content		
1	The material makes its purpose completely evident.	0, 1
Word Choice & Style		
2	The material uses common, everyday language.	0, 1
3	Medical terms are used only to familiarize the audience with the terms. When used, medical 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	The 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 2. They watch a video, assess the video according to the criteria within content, word choice and style, organization, layout, and design, and assign them 0, 1, or N/A (not applicable). Figure 2 demonstrates the expert evaluation measures and results[4]. The video in Figure 2 is considered to have high understandability.

[4] Four domain experts watch a video (https://www.youtube.com/watch?v=4JLnkpdjoU8) 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 2. The video in Figure 2 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 2. 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 started working on annotation, all of them carefully studied the PEMAT user guide[5]. 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 a consensus. Then they rated the rest of the videos based on the consensus. We use the intraclass correlation coefficient (Bartko JJ 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 3 below summarizes video understandability scores (according to PEMAT guidelines).

Table 3. Video Understandability Annotation (on a binary scale)

[5] https://www.ahrq.gov/sites/default/files/publications/files/pemat_guide.pdf

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 the 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 2 with video analysis techniques from the Google Cloud platform. Table 4 below summarizes the features we extract from video data processing results.

Table 4. 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 and 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 and 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 each 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 feature 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 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 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. The 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 the 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 the time. To assess this criterion, we use the Flesch-Kincaid readability test to indicate how the description and the transcription of a video are 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 how easy a material is to read (Basch et al. 2017, 2020).

Syntactic analysis

PEMAT assesses whether the material uses active voice. It is often argued that the passive voice will result in a structure that is more verbose than the active voice and, therefore, harder to understand and that the meaning of the 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, 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.

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

Table 5. 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 6 below.

Table 6. Statistics of the Training and Test Sets

	Training Set	**Test Set**
# of sentences	4,000	1,000
# of mentions of body part	227	101
# of mentions of chemicals and drugs	2,181	538
# of mentions of medical devices	545	126
# of mentions of medical events	784	245
# of mentions of medical professional	67	18
# of mentions of 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 A2 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 a review of the key points of the material, either in writing or orally. The summary usually comes at the end of the material and starts with summary words. Therefore, we curated a comprehensive list of summary words and phrases from multiple sources and used them to detect whether a material provides a summary.

PEMAT suggests that information in the material should be presented in an order that makes sense to the user. Main messages or the 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 (Oliu et al. 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 7 lists all the words and phrases we use to identify transitions and summaries.

Table 7. Words and Phrases for Summary and Transition

Category	Expressions
Summary/Conclusion[6]	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
Transition[7]	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,

[6] https://writingcenter.unc.edu/tips-and-tools/transitions/
[7] https://writing.wisc.edu/handbook/style/transitions/

	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 criterion refers to whether the material uses a title or upfront text that tells the reader what the material is about. In the implementation, we implement this criterion by 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

In this study, we define video understandability classification in the context of patient education as a multi-view learning and binary classification problem. Due to the vast amount of user-generated videos available and the cost to annotate the videos manually, it is essential to deploy an augmented intelligence approach in this context. The co-training approach enables us to accomplish this task with limited human effort and incorporate domain experts' assessment when machine learning models are insufficient. Our dataset includes 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. By choosing a feature-based design, we managed to engineer the features based on the evaluation criteria in the PEMAT guideline. The

classification model offers high interpretability, which can benefit healthcare organizations and reputable health content creators in their future content creation process.

3.5.1 Co-training based understandability classification

Figure 3 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 is presented in Table 8.

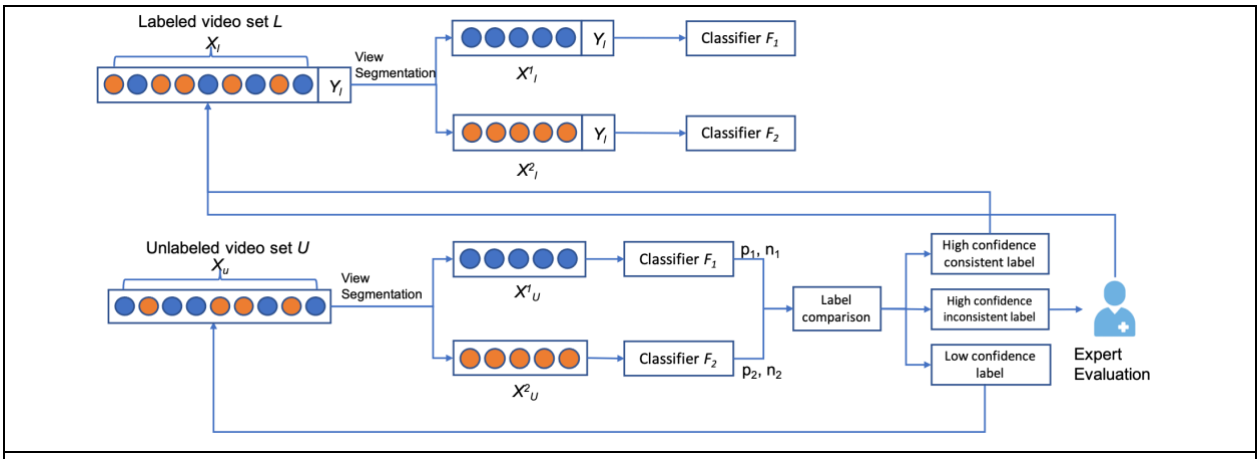


Figure 3. A human-in-the-loop co-training approach to assessing video understandability

Table 8. 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. User 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 the 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 the most confident predictions, specified by a confidence threshold
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 the 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 medical terms (Liu et al. 2020). Table 9 summarizes the features we propose to extract from the video metadata view, the method to derive the measure, and the PEMAT criteria they fall under.

Table 9. 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 the 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
Description word count	The total number of words in the video description	Metadata collection	
Sentence count	The total number of sentences in the 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 the 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 a narrative readability score to examine whether the material uses common everyday language. Part-of-speech tagging is used to extract verbs in the 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 10 summarizes the features we extract from the video content view, the methods to derive the measure, and the PEMAT criteria they fall under.

Table 10. 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 the active voice in the video transcript	Syntactic analysis	The material uses active voice
Summary word count	The number of summary words in the 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 the 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
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

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, and 4,327 of them have text embedded in the videos. We applied text and video analytics techniques to extract metadata view features and video content view features. Tables A3 and A4 in the Online Appendix reported the descriptive statistics of features of all the videos in our data collection and correlations of these features. Our co-training model initially starts with 600 labeled videos for training. The model converged after 12 iterations with a confidence threshold of 0.65. In the co-training process, 305 videos require human annotation. All these hyperparameters are selected based on empirical experiments. The model assigns labels to all the unlabeled data after convergence. Table 11 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 video understandability.

Table 11. Logistic Regression Model Summary

F_1: Video Metadata View			F_2: Video Content View		
Variable Name	Estimate	P-value	Variable Name	Estimate	P-value
Has title	-0.335	0.347	Narrative readability	**0.132**	0.034
Has description	-0.217	0.153	Active word count	**0.017**	0.026
Has tags	-0.184	0.176	Summary word count	**0.117**	0.045
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	**-0.020**	0.067	Video object	**-0.104**	0.055
Constant	-0.319	0.11	Constant	-0.272	0.117

The variables that are most significant are consistent with PEMAT. Low understandability videos are associated with longer duration, lengthier narratives, and greater medical terminologies. For model performance, we compare our predicted results in the 100 videos included in the evaluation set. Although cross-validation is commonly used in evaluating machine learning models, it is not feasible to collect a large repository of labeled data to evaluate co-training models. We therefore adopt a hold-out evaluation that is usually used for co-training methods (Ning et al. 2020). 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 the Random

forest model is achieved by max_features = log2, and N_estimator = 100. Table 12 summarizes the classification performance of our proposed method and benchmarks. Our approach achieved a weighted precision of 0.84, a weighted recall of 0.79, and an F1 score of 0.81 in classifying videos. Figure 4 illustrates the Receiver Operating Characteristic Curves. The results show that the co-training method significantly improved the video understandability classification performance.

Table 12. 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

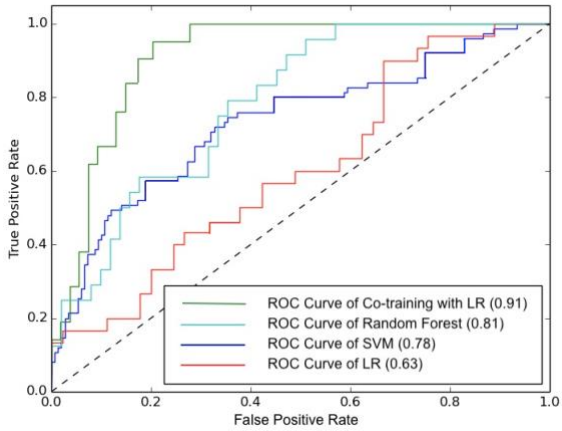


Figure 4. ROC curves of the co-training model and benchmark models

4.2 Impact of the Co-training Process on Classification Performance

The co-training process combines expert effort and machine learning methods to classify the video understandability according to the guidelines in the patient educational domain. One critical issue in the human-algorithm connection is to understand how this collaboration between human experts and machine learning algorithms improves the performance. Figure 5 below shows the classification error rate on the test set over iterations of training. In each iteration, new instances are added to the training process, lowering the classification error of the metadata view classifier,

video content view classifier, and co-training classifier. The reduction in classification error shows that this iterative process improves the overall performance. Furthermore, by combining human intelligence and machine intelligence from classifiers of two different views, the co-training approach achieves the best performance.

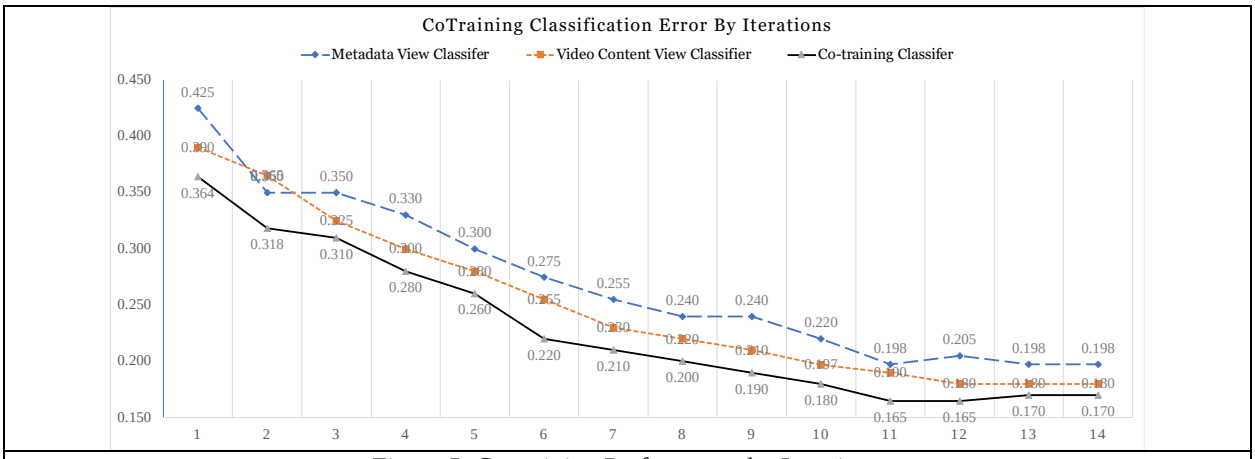


Figure 5. Co-training Performance by Iterations

4.3 Impact of Expert Involvement on Classification Performance

The human algorithm interaction in the co-training process happens in two stages: 1. Medical professionals provide a set of labeled examples to initialize the model training. 2. medical professionals are also involved in the co-training process when there are inconsistent high-confidence labels predicted by two different classifiers. Obtaining inputs from domain experts through a human-in-the-loop algorithm design is essential to our chosen task of assessing the understandability of videos from a patient education perspective. Our design also seeks to minimize human involvement while not compromising performance. To this end, we evaluate the impact of human involvement at different stages of model learning.

Figure 6 below shows the classification error by the number of labeled training examples. Devising a model to classify the understandability of patient educational videos requires high-quality training data. The process of creating training data involves medical professionals

reviewing and categorizing videos based on guidelines. As machine learning algorithms are dependent on the quality and quantity of their training data, we observe that increasing the number of labeled examples leads to an improvement in performance. However, as for co-training, the benefit of adding more training examples diminishes when we accumulate a significant amount of labeled training examples (i.e., 500) for video understandability classification.

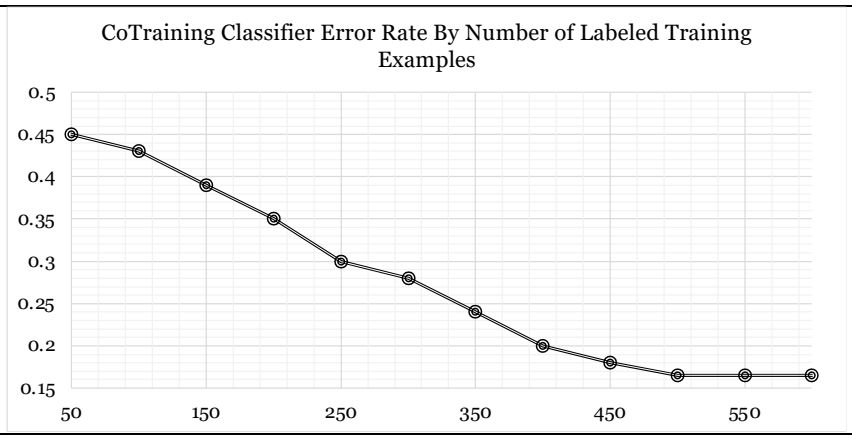


Figure 6. Co-training Performance and Number of Labeled Training Examples

Figure 7 below shows the impact of the confidence threshold on the number of expert interventions needed during the co-training process. The confidence threshold determines how many predicted labels we include in the label comparison. Its purpose is to prevent the unlabeled samples from being labeled with the wrong labels, thus decreasing the ability of the learner. Based on the label confidence threshold, unlabeled data in each iteration will be divided into three categories: videos with low confidence labels, videos with consistent and high confidence labels, and those with inconsistent but high confidence labels. The lower the threshold is, the more videos are compared and evaluated in each iteration.

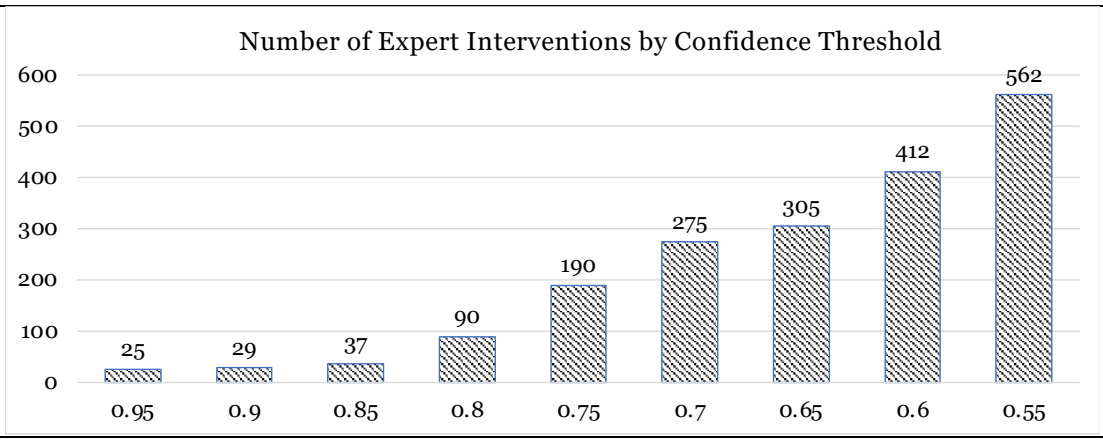


Figure 7. Co-training Confidence Threshold and Number of Expert Interventions

Figure 8 below shows the relationship between classification error rate and confidence threshold. A lower threshold can possibly lead to a faster convergence but, at the same time, result in more human involvement during the co-training process, while a higher confidence threshold may lead to the early stopping of the training because no new labels meet the confidence threshold. When the confidence threshold is high, the training process stops before assigning labels to all the unlabeled data. We follow the majority rule to assign the predicted labels for these unlabeled samples. We observe a negative impact on the classification performance due to early stopping from a high confidence threshold. When the confidence threshold is too low, too many unlabeled examples are misclassified and affect the ability of the co-training model. As a result, we see a performance decrease when the confidence threshold is too low.

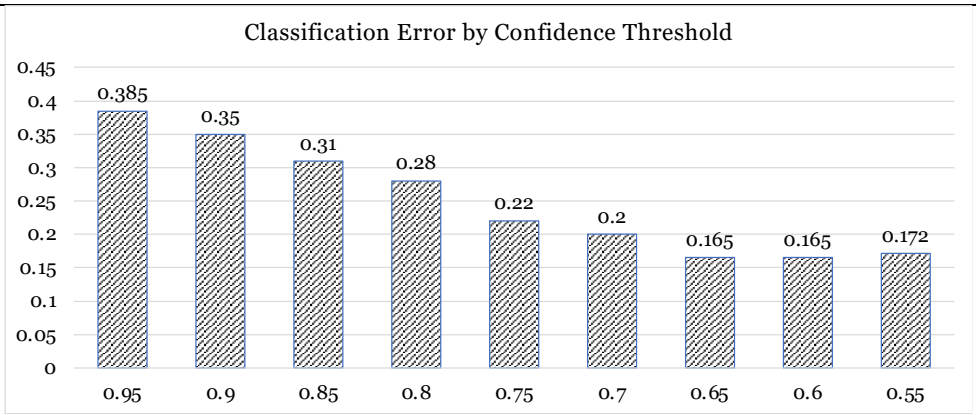


Figure 8. Co-training Confidence Threshold and Classification Error Rate

4.4 Evaluating the Importance of Understandability to Video Recommendations by Experts

We design a small user study 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 the top 10 videos according to our re-ranked results for each query and reported whether they would recommend the videos to patients. 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). We measure the average precision at K with K from 1 to 10 for 20 queries.

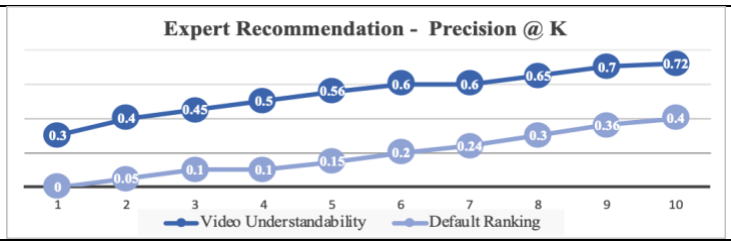


Figure 9. Comparison of Video Understandability and Default YouTube Ranking with Expert Recommendation

Figure 9 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 an expert. None of the videos that were top ranked by YouTube's default ranking received this recommendation. 72% of the top 10 videos are recommended by experts ranking by understandability, while only 40% of the top 10 videos' YouTube ranking are recommended. We conclude that our understandability measure effectively identifies patient education videos.

4.5 Evaluating the Impact of Understandability on Engagement

Building on prior studies that examine collective engagement on YouTube, we identify three measures of user engagement: video view count, comment count, and like count, which can be acquired from publicly available YouTube metadata. We examine how understandability is associated with user engagement with coarsened exact matching. There are concerns about the endogeneity between user engagement and video understandability. First, a video's ability to engage with viewers, as well as the understandability of videos, could be influenced by external factors, such as the credibility of the channel posting the video. Second, there could also be systematic differences in how users engage with understandable or not understandable videos depending on unobserved taste preferences and channel heterogeneity in attracting viewers. Third, a channel may have a higher incentive to create video materials that provide understandable content when they observe greater engagement with the video. The fundamental challenge in identification is that we do not observe the counterfactual, i.e., it is not possible to randomize treatment across a channel by showing viewer videos with high or low understandability and measuring the impact on the engagement of a video. We model a video's likelihood to have high understandability using logistic regression with several observable channels, and video-specific factors to encounter endogeneity and heterogeneity. Since user engagement may be influenced by a whole host of factors external to the content of a video, we control for heuristic measures of video quality, such as the duration of the video (Sood et al. 2011; Steinberg et al. 2010; Pandey et al. 2010), a good description or a comprehensive narrative (Gooding et al. 2011), technical quality (light, sound, resolution; Lim Fat et al. 2011; Steinberg et al. 2010; Gooding et al. 2011), credentials (Gooding et al. 2011), and the number of days since being published. Table 13 below shows the results of the logistic regression.

Table 13. Results of the logistic regression

Coefficients:	**Estimate**	**P-value**

Intercept	-9.7470	< 2e-16 ***
Log(ChannelViewCount+1)	0.5162	< 2e-16 ***
Log(ChannelSubscriberCount+1)	0.2295	2.04e-13 ***
Log(ChannelVideoCount)	-0.7669	< 2e-16 ***
ContentDefinitionSD	-0.4352	2.59e-06 ***
Duration	0.0002	2.23e-11 ***
Description Word Count	-0.0015	0.00875 **
Description Unique Word	0.0069	3.93e-07 ***
Log(publishedDays+1)	0.2874	0.00213 **

We used a coarsened exact matching method to conduct matching between the treatment group of videos that have high understandability and the control group of videos with low understandability. After matching, there are 365 videos in the treated group and 597 videos in the control group. The summary of the balance on the entire dataset and the matched dataset is available in the table below. The differences in the variables between the two groups are significantly reduced after matching.

Table 14. Balance on the Entire Dataset and Matched Dataset

Summary of Balance	**All Data**			**Matched Data**		
Variables	Mean Treated	Mean Control	Std. Mean Diff.	Mean Treated	Mean Control	Std. Mean Diff.
log(channelViewCount + 1)	16.0245	12.1291	1.5060	14.9886	14.9739	0.0057
log(channelSubscriberCount + 1)	10.6054	6.3278	1.3476	9.3214	9.2441	0.0243
log(channelVideoCount + 1)	5.3956	5.1962	0.1114	5.3068	5.3070	-0.0001
contentDefinitionhd	0.7960	0.5683	0.5650	0.6986	0.6986	0
contentDefinitionsd	0.2040	0.4317	-0.5650	0.3014	0.3014	0
Duration	676.7022	413.2235	0.3008	503.0027	395.4571	0.1228
word_count_description	242.9178	146.3287	0.3735	126.7836	129.3711	-0.01
unique_word_description	127.6597	76.5907	0.4522	72.8219	73.8849	-0.0094
log(publishedDays + 1)	7.4026	7.3888	0.0295	7.5097	7.5065	0.0069

The matching provides us a way to identify the impact of understandability on video view count, like count, and comment count. The estimated treatment effect on three measures of collective engagement is shown in Table 15 below. The results suggest that video understandability has a significant and positive impact on all three dimensions of collective

engagement. Increasing the understandability of the videos can bring more viewership, more comments, and more likes. Our results highlight the importance of understandability for healthcare organizations and health practitioners on social media platforms.

Table 15. Estimated Treatment Effect

Measure		**Estimate**	**P-value**
Log(View Count)	Intercept	8.8946	<2e-16 ***
	ATE	2.5523	<2e-16 ***
Log(Like Count)	Intercept	3.2320	<2e-16 ***
	ATE	2.9494	<2e-16 ***
Log(Comment Count)	Intercept	1.6211	<2e-16 ***
	ATE	3.0981	<2e-16 ***

5. Discussions and Conclusions

5.1 Implications for Research

With complex and very large-scale data generated by digital platforms, billions of people worldwide are accessing healthcare information through social media platforms. There is an urgent need for an evidence-based approach with AI-based methods for health literacy promotion. Our research is the first to attempt a guideline-driven consolidation of distinct data sources spanning metadata in text form and video data and adopt a human-in-the-loop learning strategy to address a video understandability assessment problem in the healthcare domain. Expert evaluation is needed when tasks are ambiguous for machine learning models to classify with high confidence.

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 the content found on social media has enabled a variety of applications that were hitherto unthinkable. 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 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 approach is developed within the context of patient educational video design but could be generalized to other settings involving content moderation. 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 the benefits of content and context-based recommendation. Finally, as machine translation methods are gaining popularity in product categorization, our method provides a scalable framework that incorporates the benefits of machine learning 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.

5.2 Implications for Practice

Improving the public's access to health information and their capacity to use it effectively can elevate the population's health literacy, empower patients, and build societal resilience. Advocates of social media in medicine highlight social media's potential to enable patient education and empowerment (Househ et al. 2014), offering the possibility of improving health outcomes (Moorhead et al. 2013). Healthcare organizations lack resources to create video content on a wide range of symptoms and disease progressions, offer easily understandable advice that can be integrated into patients' daily routines, or provide such advice on topics that are outside the physician-patient interaction in a clinical setting. We are working with clinicians to develop a library of user-centric videos tailored for stages of chronic diseases and treatments, allowing clinicians to recommend/prescribe them to patients to watch at home or during clinic visits. Figure 13 provides a summary of the technical components using the human-in-the-loop algorithmic assessment. It will allow clinicians to enter search keywords that are appropriate for the current context and retrieve a small number of videos that match the specifications. These can be viewed and shared with the patient in real-time, as needed, and prescribed as an intervention to facilitate patient education and adherence to clinical recommendations. We will release our technology platform as an open-source tool that can be widely used by the healthcare and technology communities to develop and deliver future innovations.

Currently, digital technologies for public health literacy and individual patient education are limited and not scalable. Nor do they leverage the vast amount of publicly available health-related information online and on social media platforms. Providing a strong open platform will provide a credible alternative to the vested interests of private organizations with proprietary

technologies, which will lead to future innovations in novel data collection devices and technologies in the context of health literacy initiatives.

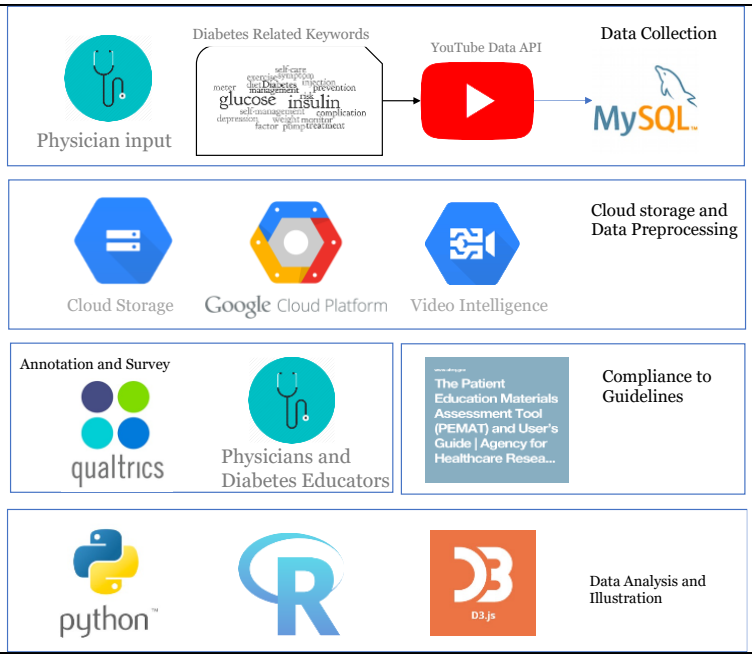


Figure 10. Technical components of our video retrieval platform

Our methodology to develop a patient educational video 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 to a societally challenging problem. Patient empowerment and engagement are essential for appropriate disease management. For health organizations that are producing patient educational materials, our approach could be used as an educational tool for enhancing understandability in patient educational video content design. When designing educational materials, we have the potential to provide best practice guidelines regarding how organizations should engage health consumers with educational videos for varying levels of health literacy. 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 digital therapy. The combination of algorithmic approaches and impact evaluation 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 amplify credible content.

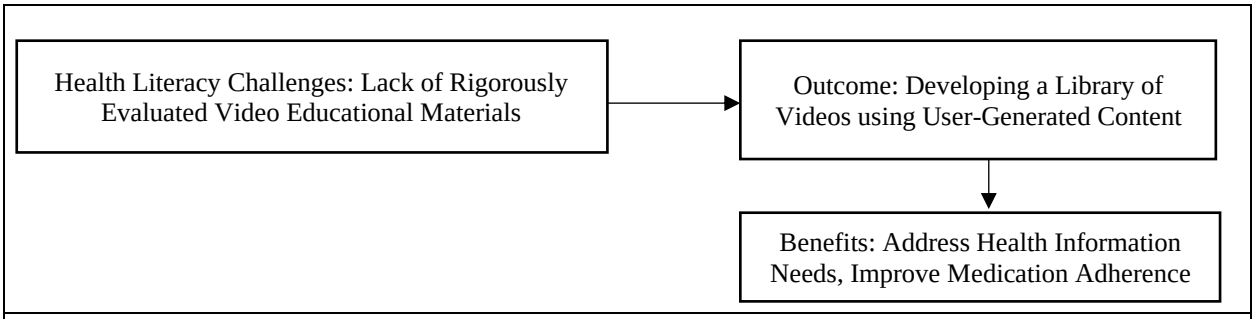


Figure 11. Contributions to Practice

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 for 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 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 also be necessary to examine factors such as concordance, which is the similarity, or shared identity, between physicians 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 develop a human-in-loop, scalable, and multi-modal algorithmic solution to evaluate the understandability of healthcare information on social media platforms. Assessing the educational value of videos in domains ranging from healthcare to education still requires domain expertise to gauge the content. The evaluation results suggest that our method demonstrated an optimal combination of human expert involvement and algorithmic decision. The results show that video understandability in health educational videos is not only critical to engagement on the social media platform but also valuable to medical experts when recommending content for patient education. Our proposed solution can also provide health organizations with actionable guidance in designing and creating patient educational videos.

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 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 metadata and video features from large social media platforms such as YouTube.

References

Abed MA, Himmel W, Vormfelde S, & Koschack J (2014) Video-assisted patient education to modify behavior: a systematic review. *Patient education and counseling* 97(1): 16-22.

Adelman JS, Brown GDA, Quesada JF (2006) Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science* 17(9):814–823.

Anderson J, Rainie L, Nolan H (2021) *The future of digital spaces and their role in democracy*

Backinger C L, Pilsner AM, Augustson EM, Frydl A, Phillips T, Rowden J (2011) YouTube as a source of quitting smoking information. *Tobacco Control* 20(2):119-122.

Baker L, Wagner TH, Singer S, Kate Bundorf M (2003) Use of the Internet and E-mail for Health Care Information: Results from a National Survey. *Journal of the American Medical Association* 289(18):2400–2406.

Bartko JJ (1966) The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 19:3–11.

Basch CH, Hillyer GC, Meleo-Erwin ZC, Jaime C, Mohlman J, Basch CE (2020) Preventive behaviors conveyed on YouTube to mitigate transmission of COVID-19: Cross-sectional study. *JMIR Public Health and Surveillance* 6(2).

Basch E, Deal AM, Dueck AC, Scher HI, Kris MG, Hudis C, Schrag D (2017) Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. *JAMA - Journal of the American Medical Association* 318(2):197–198.

Baur C, & Prue C (2014) The CDC Clear Communication Index is a new evidence-based tool to prepare and review health information. *Health promotion practice* 15(5): 629-637.

Bishop S (2019) Managing visibility on YouTube through algorithmic gossip. *New Media & Society* 21(11–12): 2589–2606.

Van Den Beemt A, Thurlings M, Willems M (2020) Towards an understanding of social media use in the classroom: a literature review. *Technology, Pedagogy and Education* 29(1):35–55.

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 YL (2020) Meeting the Moment: Black Lives Matter, Racial Inequality, Corporate Messaging, and Rebranding. *Advertising & Society Quarterly* 21(3).

Brynjolfsson E, Collis A, Diewert WE, Eggers F, Fox KJ (2019) GDP-B: Accounting for the Value of New and Free Goods in the Digital Economy. *SSRN Electronic Journal*.

Chatzopoulou G, Sheng C, Faloutsos M (2010) A first step towards understanding popularity in YouTube. In *INFOCOM IEEE Conference on Computer Communications Workshops* 2010, pp. 1-6.

C.L. B, A.M. P, E.M. A, A. F, T. P, Rowden J, Backinger CL, et al. (2011) Youtube as a source of quitting smoking information. *Tobacco control* 20(2):119–122.

Cowls J, Tsamados A, Taddeo M, Floridi L (2021) A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence* 3(2): 111-115.

Dawson AL, Hamstra AA, Huff LS, Gamble RG, Howe W, Kane I, Dellavalle RP (2011) Online videos to promote sun safety: Results of a contest. *Dermatology Reports* 3(1).

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.

Dahodwala M, Geransar R, Babion J, de Grood J, & Sargious P. (2018) The impact of the use of video-based educational interventions on patient outcomes in hospital settings: A scoping review. *Patient Education and Counseling*, 101(12), 2116-2124.

Donner A, Rotondi MA (2010) Sample size requirements for interval estimation of the kappa statistic for interobserver agreement studies with a binary outcome and multiple raters. *International Journal of Biostatistics* 6(1).

Dutta-Bergman MJ (2004) Primary sources of health information: Comparisons in the domain of health

attitudes, health cognitions, and health behaviors. *Health Communication* 16(3):273–288.
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.
Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW (2019) Online health information seeking among US adults: Measuring progress toward a healthy people 2020 objective. *Public Health Rep* 34:617–625. doi: 10.1177/0033354919874074
Ghiassi M, Zimbra D, Lee S (2016) Targeted Twitter Sentiment Analysis for Brands Using Supervised Feature Engineering and the Dynamic Architecture for Artificial Neural Networks. *Journal of Management Information Systems* 33(4):1034–1058.
Gillespie T (2019) *Custodians of the Internet*
Gillespie T (2020) Content moderation, AI, and the question of scale. *Big Data and Society* 7(2).
Gillespie T. (2014) The relevance of algorithms. Media technologies: Essays on communication, materiality, and society. 2014 Jan 17.
Gunarathne P, Rui H, Seidmann A (2017) Whose and What Social Media Complaints Have Happier Resolutions? Evidence from Twitter. *Journal of Management Information Systems* 34(2):314–340.
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: Applications and Reviews* 41(6):797–819.
Jain H, Padmanabhan B, Pavlou PA, & Santanam RT (Eds.) (2018) Call for papers—Special issue of information systems research—Humans, algorithms, and augmented intelligence: The future of work, organizations, and society. *Information Systems Research* 29(1): 250-251.
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.).
Kelly JE (2015) Computing, cognition and the future of knowing. *Whitepaper, IBM Research*.
Kincaid J, Fishburne R, Rogers R, Chissom B (1975) Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. :8–75.
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 SD, Loiselle CG (2007) Health information-seeking behavior. *Qualitative Health Research* 17(8):1006–1019.
Laurence A, John V, Bo X, Dov TE, J. H. M. ZS (2009) Current Challenge in Consumer Health Informatics: Bridging the Gap between Access to Information and Information Understanding. *Biomedical Informatics Insights* (2):1–10.
Lazer D (2015) The rise of the social algorithm. *Science* 348(6239):1090–1091.
Lazer DMJ, Pentland A, Watts DJ, Aral S, Athey S, Contractor N, Freelon D, et al. (2020) Computational social science: Obstacles and opportunities. *Science* 369(6507):1060–1062.
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).
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.
Manganello J, Gerstner G, Pergolino K, Graham Y, Falisi A, Strogatz D (2017) The relationship of health literacy with use of digital technology for health information: Implications for public health practice. *J. Public Health Manag. Pract* 23:380–387.
Manning CD, Raghavan P, Schutze H, Manning CD, Raghavan P, Schutze H (2012) Evaluation in information retrieval. *Introduction to Information Retrieval*. 139–161.
McDonald SA, Shillcock RC (2001) Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech* 44(3):295–323.
Millar N, Budgell B, Fuller K (2013) "Use the active voice whenever possible": The impact of style guidelines in medical journals. *Applied Linguistics* 34(4):393–414.
Mills A, Todorova N (2016) An integrated perspective on factors influencing online health-information seeking behaviours. Australasian Conference on Information Systems, 4, 6. *27th Australasian Conference on Information Systems (ACIS 2016)*:83.
Mishra V, Dexter JP (2020) Comparison of Readability of Official Public Health Information about COVID-19 on Websites of International Agencies and the Governments of 15 Countries. *JAMA Network Open* 3(8).
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).
Munger A, Teredesai A Relation Extraction from Biomedical Text Using SVMs With Dependency Tree Kernels. *tacoma.uw.edu*.
Munzert S, BarberÁ P, Guess A, Yang J (2021) Do Online Voter Guides Empower Citizens?Evidence from a Field Experiment with Digital Trace Data. *Public Opinion Quarterly* 84(3):675–698.
Ning X, Wang X, Xu S, Cai W, Zhang L, Yu L, Li W (2020) A review of research on co‐training. *Concurrency and computation: practice and experience*. 2021:e6276.
Oliu,Walter E. Brusaw,Charles T. Alfred GJ (2013) Writing That Works. *IEEE Transactions on Professional Communication*:738.
Pierce D, Cardie C (2001) Limitations of Co-training for Natural Language Learning from Large Datasets. *Proceedings 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.
Qiu L, Tang Q, Whinston AB (2015) Two formulas for success in social media: Learning and network effects. *Journal of Management Information Systems* 32(4):78–108.
Ramagiri R, Kannuri N, Lewis M, Murthy GVS, 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(13):S37–S41.
Rasheed Z, Shah M (2003) Scene detection in Hollywood movies and TV shows. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2.
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.
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.
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. *Genetics in Medicine* 18(5):501–512.
Shahraray B (1995) Scene change detection and content-based sampling of video sequences Behzad Shahraray. *Digital Video Compression: Algorithms and Technologies* 2419:2–13.
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).
Stellefson M, Chaney B, Ochipa K, Chaney D, Haider Z, Hanik B, Chavarria E, Bernhardt JM (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*
Susarla A, Oh JH, Tan Y (2016) Influentials, Imitables, or Susceptibles? Virality and Word-of-Mouth Conversations in Online Social Networks. *Journal of Management Information Systems* 33(1):139–70
T. L, S. D (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2):211–240.
Tang Q, Gu B, Whinston A (2012) Content contribution for revenue sharing and reputation in social media: A dynamic structural model. *Journal of Management Information Systems* 29(2):41–76.
Tomašev N, Cornebise J, Hutter F, Mohamed S, Picciariello A, Connelly B, & Clopath C (2020) AI for social good: unlocking the opportunity for positive impact. *Nature Communications* 11(1): 1-6.
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.
Vishnevetsky J, Walters CB, & Tan KS (2018) Interrater reliability of the patient education materials assessment tool (PEMAT). *Patient education and counseling* 101(3): 490-496.
Williams AM, Muir KW, Rosdahl JA (2016) Readability of patient education materials in ophthalmology: A single-institution study and systematic review. *BMC Ophthalmology* 16(1).
Williams T, Szafir D, Chakraborti T, Ben Amor H (2018) Virtual, augmented, and mixed reality for human-robot interaction. *In Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 403-404). ACM.
Xie K, Lee YJ (2015) Social Media and Brand Purchase: Quantifying the Effects of Exposures to Earned and Owned Social Media Activities in a Two-Stage Decision Making Model. *Journal of Management Information Systems* 32(2):204–238.
Xiong W, Lee JCM (1998) Efficient Scene Change Detection and Camera Motion Annotation for Video Classification. *Computer Vision and Image Understanding* 71(2):166–181.
Yoo E, Gu B, Rabinovich E (2019) Diffusion on Social Media Platforms: A Point Process Model for Interaction among Similar Content. *Journal of Management Information Systems* 36(4):1105–1141.
Zhang M, Tang J, Zhang X, Xue X (2014) Addressing cold start in recommender systems: A semi-supervised co-training algorithm. *SIGIR 2014 - Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 73–82.
Zheng NN, Liu ZY, Ren PJ, Ma YQ, Chen ST, Yu SY, Xue JR, Chen BD, Wang FY (2017) Hybrid-augmented intelligence: collaboration and cognition. *Frontiers of Information Technology & Electronic Engineering* 18(2): 153-179